1.5.2.2 Asymptotic properties within a single basin—three examples (advanced)

In an attempt to establish consistency and asymptotic normality for the SPARROW model parameter estimates, we immediately encounter a technical problem. The problem arises in the assumptions needed to extend the number of observations to an infinitely large size. In many applications of a SPARROW model, the researcher will have a prescribed watershed containing a defined reach network. Because the watershed represents a bounded region, the only way in which observations can be extended to infinity is by infilling—that is, increasing the density of measurements within the study area. A hydrologic system is by nature bounded and accumulative, however, meaning the contaminant flux at the outflow of the basin is an accumulation of individual processes within a bounded watershed. It is also true that not all uncertainty in the description of the basin can be resolved at the basin outlet, even if the uncertainty is independent at the smallest scale. As is shown below, the existence of error at the aggregate scale implies asymptotic theory commonly used to justify finite sample estimates is not valid in the context of a finite watershed.

To better understand the limitations imposed by a finite basin with aggregate error, we present three examples. Each example is built from a stochastic process that is well defined and statistically independent at the smallest scale, yet leads to non-degenerate stochastic behavior at the aggregate scale. The examples demonstrate that model estimates from finite basins do not converge in probability to a constant, implying the asymptotic properties of consistency and normality do not necessarily hold. The utility of this result is technical; however, the examples serve another purpose—they demonstrate how a SPARROW model arises from a fundamental description of hydrologic stochastic processes. In this way, some light is shed on the somewhat 'black box' nature of large-scale hydrologic models.

The first example, conceptually depicted in figure 1.20, considers a simple SPARROW model for a single reach of length d_T. In this example, we assume that there are no incremental additions to stream flux along the reach. The only hydrologic process acting on flux is in-stream attenuation, governed by the decay parameter δ and the length of the reach to which it is applied. Let there be *n* monitoring stations along this reach, sequentially indexed by *i* beginning with the station located at the furthest upstream location (fig. 1.20). Let d_i represent the length of stream between station $i - 1$ and station *i*. Let the decay process operating on the section of stream between station $i - 1$ and station *i* be subject to error u_i. This error is assumed to be continuous, independent between any two non-overlapping segments of the stream, and have a mean of zero and a variance that is proportional to the stream length d_i. An example of such a random process taken from the stochastic calculus literature is the Brownian motion process derived from the Weiner process (see Malliaris and Brock, 1982). It has been shown that any process exhibiting continuity and having stationary, independent and identically distributed increments must be normally distributed (Breiman, 1968, proposition 12.4). Finally, let y_i represent the log of flux measured at location *i*, and let y_0, the log of flux at the upstream end of the reach, be defined and known.

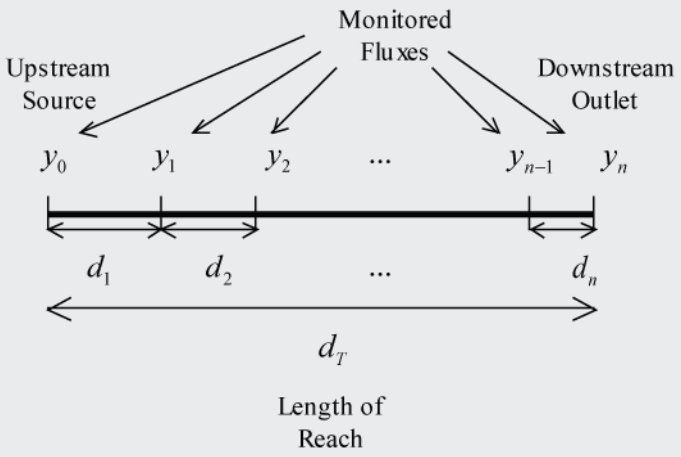

Figure 1.20. A graphical depiction of a simple hydrologic system consisting of a single reach segment of length d_T divided into *n* sub-segments of length d_i, each having a measured logarithm of flux y_i.

The SPARROW model equation in this case for any measured segment *i* is

$$y_i = \ln\left(e^{y_{i-1}} e^{-\delta d_i + u_i}\right) = y_{i-1} - \delta d_i + u_i, \tag{1.68}$$

where $V(u_i) = \sigma_v^2 d_i$, σ_v^2 being the variance of the decay process per unit distance. From equation (1.68), we see that this simple case leads to a SPARROW model that is linear with heteroscedastic errors; a model that is efficiently estimated using standard weighted least squares.

To facilitate exposition, we write the simple model for all *n* observations in vector notation

$$\Delta\mathbf{y} = -\delta\mathbf{d} + \mathbf{u}, \tag{1.69}$$

where $\Delta\mathbf{y} = \{y_1 - y_0, \ldots, y_n - y_{n-1}\}'$, $\mathbf{d} = \{d_1, \ldots, d_n\}'$, and $\mathbf{u} = \{u_1, \ldots, u_n\}'$. Using this notation, the covariance matrix for the errors is $E(\mathbf{u}\mathbf{u}') = \sigma_v^2 \mathbf{D}(\mathbf{d})$, where $\mathbf{D}(\cdot)$ is the diagonal matrix operator that creates an $n \times n$ diagonal matrix from its *n*-element vector argument. The weighted least squares estimate of the coefficient δ and its variance are given by

$$\begin{aligned} \hat{\delta} &= \left(\mathbf{d}'\mathbf{D}(\mathbf{d})^{-1}\mathbf{d}\right)^{-1}\mathbf{d}'\mathbf{D}(\mathbf{d})^{-1}\Delta\mathbf{y} = \left(\mathbf{i}'\mathbf{d}\right)^{-1}\mathbf{i}'\Delta\mathbf{y} = \frac{y_n - y_0}{d_T}, \\ V(\hat{\delta}) &= \sigma_v^2\left(\mathbf{d}'\mathbf{D}(\mathbf{d})^{-1}\mathbf{d}\right)^{-1} = \frac{\sigma_v^2}{d_T}. \end{aligned} \tag{1.70}$$

Notice that the optimal weighted least squares estimate of δ depends only on the change in flux over the entire pathway, $y_n - y_0$; measurements of flux made at intermediate locations along the pathway have no bearing on the optimal estimate. Consequently, the variance of $\hat{\delta}$ does not depend on *n*, implying that no amount of infill sampling can improve its estimate. In this case, the estimate $\hat{\delta}$ is unbiased; it has an expectation of δ. $\hat{\delta}$ is not consistent, however, because its variance does not go to zero asymptotically (this is a consequence of $\hat{\delta}$ being

normally distributed; see theorem 18.14 in Davidson, 1994). Here, because u_i are derived from a Wiener process, the distribution of $\hat{\delta}$ will be normal. It is possible, however, to construct other examples in which the underlying process is not continuous, and therefore not Wiener and not normally distributed. Consequently, the asymptotic distribution of the estimated decay rate need not be normal.

A trivial extension of this simple example can be used to show that the asymptotic limitations of the hydrologic model cannot be overcome by simply appealing to a higher dimension. Consider a single dendritic reach network consisting of an infinite number of reach segments indexed by i, $i = 1, \ldots, \infty$, contained within a bounded two-dimensional watershed of area A (see figure 1.21). For each reach segment i, define a corresponding sub-basin of area A_i that represents the incremental drainage of the segment, so that $\sum_{i=1}^{\infty} A_i = A$. Assume there exists some finite constant b such that $d_i \leq bA_i$ for all i. This assumption implies that no segment can be infinitely long, a reasonable assertion if flux is to accumulate at the outlet of the watershed in finite time. Finally, as in the previous example, we make the simplifying assumption that sources are located only at the upstream ends of each segment.

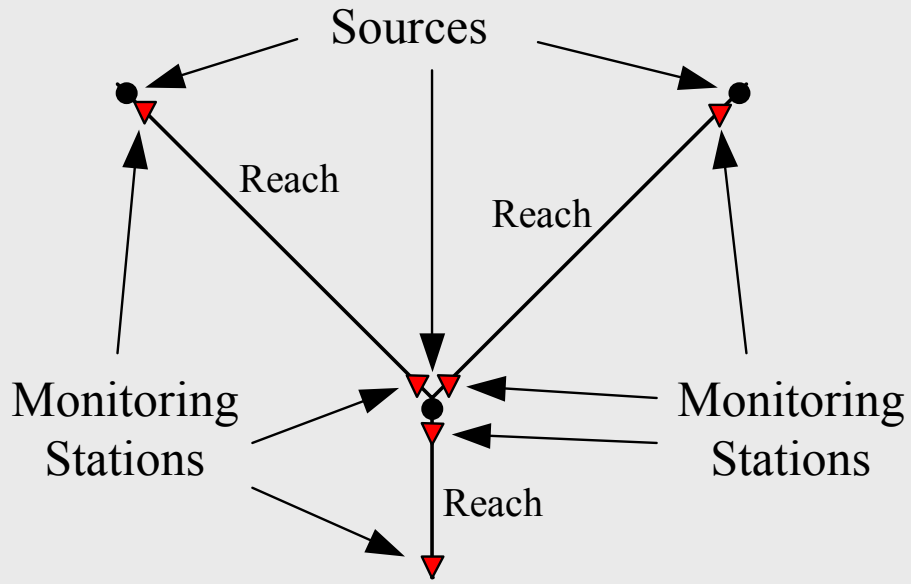

Figure 1.21. The arrangement of reaches, sources and monitoring stations in a two-dimensional hydrologic model.

The previous example shows that the most efficient way to monitor a reach is to monitor the endpoints; monitoring intermediate locations along a reach has no bearing on the estimate of decay. Here it is assumed that the monitoring of reach i represents two measurements—a downstream measurement at the reach outlet and an upstream measurement just below the introduction of the reach's source. In a sample of N independent reaches, the estimate of decay is again efficiently estimated by weighted least squares. The first two equalities for $\hat{\delta}$ in equation (1.70) remain applicable, resulting in

$$\hat{\delta} = \left(\mathbf{d}'\mathbf{D}(\mathbf{d})^{-1}\mathbf{d}\right)^{-1}\mathbf{d}'\mathbf{D}(\mathbf{d})^{-1}\Delta\mathbf{y} = \left(\mathbf{i}'\mathbf{d}\right)^{-1}\mathbf{i}'\Delta\mathbf{y} = \frac{\sum_{i=1}^{N}\Delta y_i}{\sum_{i=1}^{N} d_i},$$

(1.71)

$$V\left(\hat{\delta}\right) = \sigma_v^2\left(\mathbf{d}'\mathbf{D}(\mathbf{d})^{-1}\mathbf{d}\right)^{-1} = \frac{\sigma_v^2}{\sum_{i=1}^{N} d_i},$$

where, in this context, the i^{th} element of $\Delta\mathbf{y}$ pertains to the difference between the downstream and upstream measurements of reach *i*. Since $d_i \le bA_i$ for each *i*, we have

(1.72) $$V\left(\hat{\delta}\right) > \frac{\sigma_v^2}{b\sum_{i=1}^{N} A_i} \ge \frac{\sigma_v^2}{bA}.$$

The variance of the estimator is again bounded away from zero, regardless of the number of observations. As *N* goes to infinity, and every segment of the reach network becomes monitored, the variance of $\hat{\delta}$ does not go to zero. Consequently, the conditions for consistency are not met.

The last example demonstrates that the limitations of asymptotic analysis within a finite basin are not restricted to the estimation of the decay rate, but also pertain to the estimation of source coefficients. Consider again the simple case of a single reach of length d_T. For this example, it is assumed that the in-stream decay rate is zero throughout the full reach segment. Arrayed along the reach segment are sources, defined continuously by the function $S(t)$. Associated with each source is a source coefficient that determines the amount of source $S(t)$ that is delivered to the stream. The source coefficient is assumed to be stochastic and is given by $a\left(dq(t)\right)$, where *a* is a constant and *dq*(*t*) is a Poisson jump process defined over continuous distance *t*, where *dq*(*t*) equals 1 with probability $\lambda\, dt$ and equals zero with probability $1-\lambda dt$ (see Malliaris and Brock, 1982). Thus, the expectation of *dq*(*t*) is $\lambda\, dt$, and the variance is (ignoring terms smaller than *dt*) also $\lambda\, dt$. The adoption of a Poisson process to define the source coefficient implies sources are effectively distributed discretely over the length of the reach but can occur at any location with equal probability. Because *q*(*t*) has jumps, it is not a continuous process, as was the case for the Wiener process used in the examples involving stream decay. Like a Wiener process, however, the Poisson process *q*(*t*) has the Markov property that the probability distribution for all downstream values of the *q*(*t* + *s*) conditioned on all information available at location *t* depends only on the local value of *q*(*t*) and not on any upstream values. This implies the intervals *dq*(*t*) and *dq*(*s*) are independent for $s \ne t$.

Assume monitoring stations are positioned at locations t_i, $i = 1,\ldots,n$, with spacing $d_i = t_i - t_{i-1}$. The flux measured at station *i* is given by

(1.73) $$Y_i = Y_{i-1} + a\int_{t_{i-1}}^{t_{i-1}+d_i} S(t)\, dq(t).$$

The mean and variance of $Y_i - Y_{i-1}$ are given by

(1.74) $$E\left(Y_i - Y_{i-1}\right) = a\lambda\int_{t_{i-1}}^{t_{i-1}+d_i} S(t)dt, \text{ and}$$

(1.75) $$V\left(Y_i - Y_{i-1}\right) = a^2\lambda\int_{t_{i-1}}^{t_{i-1}+d_i} S^2(t)\, dt.$$

Due to the assumptions associated with *q*(*t*), the covariance between $Y_i - Y_{i-1}$ and $Y_j - Y_{j-1}$ is zero for $i \ne j$.

Estimates of the source coefficient *a* and Poisson parameter λ can be obtained from a simple linear model having the form

(1.76) $$\Delta\mathbf{Y} = b\mathbf{X} + \mathbf{Z},$$

where $\mathbf{\Delta Y} = \{Y_1, Y_2 - Y_1, \ldots, Y_n - Y_{n-1}\}'$, $\mathbf{X} = \left\{\int_0^{d_1} S(t)\,dt, \int_{t_1}^{t_1+d_2} S(t)\,dt, \ldots, \int_{t_{n-1}}^{t_{n-1}+d_n} S(t)\,dt\right\}'$, $b = a\lambda$, and error vector $\mathbf{Z} = \{Z_1, \ldots, Z_n\}'$ has zero mean and diagonal covariance matrix $a^2\lambda\mathbf{D}(\mathbf{g})$, with **D**(**g**) being a diagonal matrix having diagonal elements **g** given by the vector $\mathbf{g} = \left\{\int_0^{d_1} S^2(t)\,dt, \int_{t_1}^{t_1+d_2} S^2(t)\,dt, \ldots, \int_{t_{n-1}}^{t_{n-1}+d_n} S^2(t)\,dt\right\}'$. The model given in equation (1.76) is not technically a SPARROW model because it is estimated in real space as opposed to logarithm space, but it is a valid model and will suffice to make the necessary point concerning asymptotic properties of estimators based on infinitely dense monitoring stations.

Equation (1.76) can be estimated using linear weighted least squares, with the weight of observation *i* set to $1\Big/\int_{t_{i-1}}^{t_{i-1}+d_i} S^2(t)\,dt$. The expectation of the mean squared weighted residual is given by

$$E\left(\sum_{i=1}^{n} w_i Z_i^2 \Big/ n\right) = a^2\lambda. \tag{1.77}$$

The weighted least squares estimate of the slope coefficient *b* is

$$\hat{b} = \left(\mathbf{X}'\mathbf{D}(\mathbf{g})^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}(\mathbf{g})^{-1}\mathbf{\Delta Y}, \tag{1.78}$$

and the variance of this estimate is given by

$$V(\hat{b}) = a^2\lambda\left(\mathbf{X}'\mathbf{D}(\mathbf{g})^{-1}\mathbf{X}\right)^{-1} = a^2\lambda\left\{\sum_{i=1}^{n}\frac{\left(\int_{t_{i-1}}^{t_{i-1}+d_i} S(t)\,dt\right)^2}{\int_{t_{i-1}}^{t_{i-1}+d_i} S^2(t)\,dt}\right\}^{-1}. \tag{1.79}$$

Estimates of the slope coefficient *b* and the mean squared weighted residual suffice to identify the source coefficient *a* and Poisson scaling factor λ; that is, the estimated slope coefficient is an estimate of the product $a\lambda$, and the estimated mean squared weighted residual is an estimate of the product $a^2\lambda$. The ratio of mean squared weighted residual to the coefficient *b* provides an estimate of *a* and the ratio of the squared coefficient estimate to the mean squared weighted residual gives an estimate of λ.

From the inequality

$$0 \le \int_{t_{i-1}}^{t_{i-1}+d_i}\left(S(t) - \frac{1}{d_i}\int_{t_{i-1}}^{t_{i-1}+d_i} S(s)\,ds\right)^2 dt = \int_{t_{i-1}}^{t_{i-1}+d_i} S^2(t)\,dt - \frac{1}{d_i}\left(\int_{t_{i-1}}^{t_{i-1}+d_i} S(t)\,dt\right)^2, \tag{1.80}$$

we have $d_i \ge \left(\int_{t_{i-1}}^{t_{i-1}+d_i} S(t)\,dt\right)^2 \Big/ \int_{t_{i-1}}^{t_{i-1}+d_i} S^2(t)\,dt$, and

$$d_T = \sum_{i=1}^{n} d_i \ge \sum_{i=1}^{n}\frac{\left(\int_{t_{i-1}}^{t_{i-1}+d_i} S(t)\,dt\right)^2}{\int_{t_{i-1}}^{t_{i-1}+d_i} S^2(t)\,dt}, \tag{1.81}$$

which, via equation (1.79), leads directly to the lower bound on the variance of $\hat{b}$,

$$V\left(\hat{b}\right) \geq \frac{a^2\lambda}{d_T}. \tag{1.82}$$

As with the previous examples, a finite basin, here represented by a finite value for the length of the reach, d_T, places a lower bound on the variance of the estimated slope coefficient $\hat{b}$, implying $\hat{b}$ is not consistent.

The above examples illustrate that the conditions required to apply large sample theory in a hydrologic model can be met only by expanding the analysis to non-nested basins. In some sense this limitation is technical and refers only to the theoretical justification of certain statistical results. The practical implication, however, is that large sample theory cannot be applied in the context of a small basin in which additional observations are generated by increasing the density of the sampling network. If the choice is between expanding a sampling network by including other basins or by concentrating more samples within a given basin, large sample theory suggests the former would have a larger statistical payoff. There are, of course, other reasons for adopting this protocol; statistical inference is always improved the greater the variability in conditions expressed by the explanatory variables of a model. The consideration of large sample properties addressed here marginally adds to the considerable weight of these arguments.

It is important to recognize that the failure of the model to yield consistent estimates within a finite basin is a direct consequence of the hydrologic system and is not due to any assumptions used to define the SPARROW model. The statistical analysis of a fixed basin using any model faces the same limitations described above. As long as basins are finite and uncertainty accumulates in them, it is not possible to satisfy the conditions needed to apply asymptotic properties to the model estimates. An alternative to the static models described above would be to consider data collection in the context of a dynamic model. A dynamic model implies data can be accumulated along a temporal dimension, in addition to the spatial dimension exploited by SPARROW. If the underlying error processes are dynamic, meaning, for example, the Brownian motion process *u* used in the first example varied randomly with time, then repeated sampling of a fixed basin through time would yield consistent estimates. Consequently, a dynamic model may display large sample behavior that cannot be obtained by a purely spatial analysis. If any of the underlying stochastic processes are static, however, varying only over space and not time, the statistical description of these processes by a dynamic model confronts the same asymptotic limitations as a strictly spatial analysis, such as SPARROW.

1.5.3 Coefficient bias and uncertainty—additional issues

The methods described in the previous sections pertain to large sample properties of the estimators. In finite samples, parameter estimators may be biased and may not be normally distributed; consequently, standard methods for testing the statistical significance of parameters could be invalid. Explicit knowledge of the distributions of estimators would correct this deficiency, but these distributions are typically unknown. An alternative approach, known as *bootstrapping*, is to infer the distributions of parameter estimators by assessing their empirical distributions, the distributions implied by the available sample data (as opposed to the population of all possible data). The idea is to generate all possible *N*-element combinations of the *N* observations, allowing repetitions of observations, with a set of coefficient estimates obtained for each combination. The distribution of these sets of estimates forms the empirical distribution of the coefficients. With *N* observations in a sample, there are $\binom{2N-1}{N}$ possible unique combinations of the observations on which to base the empirical distribution, a prohibitive number for even modest sample sizes. An alternative approach is to build the empirical distribution from *R* random draws of the $\binom{2N-1}{N}$ possible combinations. SPARROW implements such an approach, which is called *Monte Carlo resampling*, or simply *resampling*.

The bootstrap paradigm is this: the relation between the population distribution and the true moments of the population is assumed to be the same as the relation between the empirical distribution and the estimated moments, as obtained via minimization of some objective function (nonlinear least squares, for example). The

practical implication of this paradigm is that if the computation of some statistic of interest requires knowledge of the relation between the population distribution and the true moments, the relation between the empirical distribution and the empirical moments can be used in its place. This paradigm is later shown in section 1.6 to be most useful for the assessment of bias and uncertainty in predictions, but is shown here to also be useful for assessing small sample properties of the coefficient estimates.

1.5.3.1 Bootstrap estimate of coefficient bias (advanced)

The additive bias of a coefficient estimate, say $\hat{\beta}_k$, is given by

$$B\left(\hat{\beta}_k\right) = E\left(\hat{\beta}_k\right) - \beta_k . \tag{1.83}$$

Both terms in the right-hand side of this expression are unknown. The bootstrap paradigm tells us to use the empirical distribution relative to the empirical estimate $\hat{\beta}_k$ to assess the bias. That is, random sets of *N* observations, drawn from the original set of *N* observations with replacement, are used to generate alternative estimates of the coefficients, each using the same methodology that was used to compute $\hat{\beta}_k$. Let there be *R* such random re-samples drawn from the original sample, with *R* corresponding estimates of the coefficient vector $\hat{\boldsymbol{\beta}}_r$. The bootstrap paradigm says that the relation between the true value β_k and the population distribution of $\hat{\beta}_k$ is the same as the relation between the empirical estimate $\hat{\beta}_k$ and the *R* coefficients $\hat{\beta}_{k,r}$ derived from the randomly drawn samples.

The implementation of the bootstrap procedure used in SPARROW can be described in terms of repetitive application of random weights to the model observations, following each reweighting with a re-estimation of the coefficients. For each bootstrap repetition $r = 1,...,R$, randomly generate *N* observation indices with replacement $\vartheta_j^r,\ j = 1,\ldots,N:\ \vartheta_j^r = \max\left(1, \text{ceil}\left(N\xi_j^r\right)\right)$, where $\xi_j^r,\ j = 1,\ldots,N$ is drawn from a uniform [0,1] distribution. Let n_i^r be the number of times observation *i* is selected in repetition *r* (*i.e.*, the number of times over all *j* that ϑ_j^r equals *i*). Then $\hat{\boldsymbol{\beta}}_r$ is the value of *K*-element coefficient vector $\boldsymbol{\beta}$ that minimizes $Q^r = \sum_{i=1}^{N} w_i^r n_i^r \left(f_i^M - f_i^*\left(\boldsymbol{\beta}\right)\right)^2$, where w_i^r is the standard weight for the i^{th} observation and r^{th} bootstrap repetition as determined using the methods described in section 1.5.3.1.

The bootstrap estimate of bias mirrors equation (1.83) and is given by

$$B\left(\hat{\beta}_k\right)^R = \overline{\beta}_k^R - \hat{\beta}_k , \tag{1.84}$$

where $\overline{\beta}_k^R = R^{-1}\sum_{r=1}^{R}\hat{\beta}_{k,r}$. Consequently, the bootstrap bias-corrected estimate of β_k is given by

$$\tilde{\beta}_k^R = \hat{\beta}_k - B\left(\hat{\beta}_k\right)^R = 2\hat{\beta}_k - \overline{\beta}_k^R . \tag{1.85}$$

$\tilde{\beta}_k^R$ represents an estimate of β_k approximately corrected for first-order bias. That is, for any $P < 2$, N^P times the remaining bias (after bootstrap bias correction) goes to zero as *N* goes to infinity, a limit that has the mathematical notation $O\left(N^{-2}\right)$ (Davison and Hinkley, 1997; Shao and Tu, 1995). The correction is assessed as approximate because a formal proof of the limit pertains to the assumption that $\hat{\beta}_k$ is a quadratic statistic, which is only approximately true in the case of nonlinear least squares (Davison and Hinkley, 1997).

The difference between the average of the bootstrap estimates and the parametric estimate indicates the degree to which the estimation methodology can recover the original parameters that underlie the data generating process. In large samples, given the standard assumptions described above, the coefficient estimates are consistent and the *t*-statistics have a standard normal distribution. If the bootstrap estimate of bias, which is sensitive to sample size, were large, then this would indicate the assumption of large sample properties is not appropriate.

1.5.3.2 Bootstrap estimate of the coefficient covariance matrix (advanced)

The *R* estimates of the coefficient vectors $\hat{\boldsymbol{\beta}}_r$ also can be used to derive the bootstrap estimate of the covariance matrix of the coefficient estimates (Efron and Tishirani, 1993)

$$\mathbf{V}\left(\hat{\boldsymbol{\beta}}\right)^R = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}}^R\right)\left(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}}^R\right)', \tag{1.86}$$

where $\hat{\boldsymbol{\beta}}^R = R^{-1}\sum_{r=1}^{R}\hat{\boldsymbol{\beta}}_r$. The bootstrap estimates of the variances of the coefficients are given by the diagonal elements of this matrix. Efron and Tishirani (1993) show that this estimate has a variance (that is, the variance of the variance) of order $O\left(N^{-2}\right)$, meaning that for any $P < 1$ the variance estimate $N^P\mathbf{V}\left(\hat{\boldsymbol{\beta}}\right)^R$ goes to zero as *N* goes to infinity. This is the same accuracy as the asymptotic covariance matrix given in equation (1.57), so there is no advantage in using the bootstrap estimate of the covariance matrix as compared to the parametric (that is, asymptotic) estimate.

1.5.3.3 Bootstrap coefficient confidence interval (advanced)

The standard confidence interval given above in equation (1.67) requires the assumption that the coefficient estimates have an underlying normal distribution. Although the large sample distribution of the coefficient estimates approaches normal, there is no assurance that the normal approximation is valid in finite samples. Bootstrap analysis has been used to derive a more refined estimate of the confidence interval in these cases.

One bootstrap approach, called the *hybrid* approach, uses the quantiles of the empirical distribution for $\hat{\beta}_{k,r} - \hat{\beta}_k$ in place of the standard normal quantiles appearing in equation (1.67). Let $H_{k,R}(x)$ represent the empirical distribution of $\hat{\beta}_{k,r} - \hat{\beta}_k$; that is, $H_{k,R}(x)$ is the share of the *R* bootstrap estimates of $\hat{\beta}_{k,r} - \hat{\beta}_k$ that are less than or equal to *x*. The inverse of the empirical distribution, denoted $H_{k,R}^{-1}(p)$, represents the empirical quantile associated with the cumulative probability *p*. The hybrid bootstrap equal-tail two-sided confidence interval lower and upper bounds are

$$\underline{\hat{\beta}}_k^R = \hat{\beta}_k - H_{k,R}^{-1}\left((1+P_c)/2\right), \text{ and } \overline{\hat{\beta}}_k^R = \hat{\beta}_k - H_{k,R}^{-1}\left((1-P_c)/2\right). \tag{1.87}$$

Note that a standard error term, comparable to the $\sqrt{V_{kk}\left(\hat{\boldsymbol{\beta}}\right)}$ term in equation (1.67), is absent from (1.87). This is because the empirical distribution pertains to $\hat{\beta}_{k,r} - \hat{\beta}_k$, which is not normalized by its standard deviation. Note also that it is not necessary to apply bias correction to the estimates in order to obtain valid confidence intervals. This follows from the assumption that bias is additive and constant in the sense that the entire distribution of $\hat{\beta}_k$ is shifted with respect to β_k by the same amount, as is the distribution of $\hat{\beta}_{k,r}$ with respect to $\hat{\beta}_k^*$. In this case, as long as the bias in the bootstrap estimates equals the bias in the parametric coefficient

estimate $\hat{\beta}_k$, the bias in the derivation of the quantile $H_{k,R}^{-1}(p)$ is negated by the bias in $\hat{\beta}_k$ resulting in an unbiased interval. Further remarks regarding this property of the hybrid interval are included in the discussion of prediction intervals in section 1.6.5.

In practice, the quantiles are determined by ordering the *R* estimates of $\hat{\beta}_{k,r} - \hat{\beta}_k$ in ascending order, with $q_k(s)$ representing the s^{th} value from this list. Then

$$
\begin{aligned}
H_{k,R}^{-1}\left((1-P_c)/2\right) &= q_k\left(\lfloor R(1-P_c)/2 \rfloor + 1\right), \text{ and} \\
H_{k,R}^{-1}\left((1+P_c)/2\right) &= q_k\left(\lceil P_c R \rceil + \lfloor R(1-P_c)/2 \rfloor\right),
\end{aligned}
\tag{1.88}
$$

where $\lfloor z \rfloor$ is the floor function (round to the next lowest integer), and $\lceil z \rceil$ is the ceiling function (round to the next highest integer) (see appendix A for a derivation).

Shao and Tu (1995) show that the hybrid bootstrap equal-tail two-sided confidence interval is second-order accurate (meaning that for all $P < 1$, N^P times the difference between the hybrid confidence interval coverage probability and the stated confidence level goes to zero as *N* goes to infinity)—the same as the normal approximation for the parametric method. Therefore, there is no statistical advantage to using bootstrap methods for estimating equal-tailed two-sided confidence intervals for parameters. [Note that second-order accuracy for confidence intervals means something different from removal of second-order bias, which explains why the criterion for *P* here is $P < 1$ and was $P < 2$ above in reference to bias.]

Shao and Tu (1995) also show that for one-sided confidence intervals, accuracy can be improved by expressing the desired coefficient in its pivoted form—that is, $\hat{\beta}_{k,r} - \hat{\beta}_k$ is divided by a bootstrap estimate of its standard error. The accuracy of the one-sided confidence interval in this case is greater than the accuracy obtained with the one-sided normal approximation or the hybrid bootstrap described above. Unfortunately, the method requires a double bootstrap whereby an additional set of bootstrap estimates is required for each original bootstrap repetition in order to estimate the variance. Given the high computational costs required to obtain a single set of bootstrap estimates in SPARROW, performing a double bootstrap is infeasible and the more accurate pivot form of the confidence interval is not implemented.

1.5.3.4 Discussion of bootstrap methods for coefficient estimation

Shao and Tu (1995) point out that for any given bootstrap replication it is possible the resampled data may be collinear. This would occur if a large number of draws from the *N* observations happened by chance to come from only a small number of observations. They suggest a filter be placed on the execution of each bootstrap iteration such that the iteration's coefficient estimates are set to the parametric estimates $\hat{\beta}_k$ if the smallest eigenvalue used to evaluate multicollinearity (see the discussion of eigenvalues in section 1.5.4.3) is below some specified threshold. In practice, even with a modest sample size, this is a highly unlikely outcome unless the sample itself, without resampling, is already highly multicollinear. SPARROW currently does not check the eigenvalues of the individual bootstrap iterations in order to prevent the inclusion of highly multicollinear coefficient estimates in the bootstrap analysis.

The bootstrap methods described above are useful for assessing small sample bias in the nonlinear weighted least squares estimated coefficients. The methods are less useful for testing hypotheses. As explained above in section 1.5.3.3, the bootstrap estimate of the confidence interval is of the same order of accuracy as the standard normal assumption. Thus, with regard to evaluation of model specification and reporting of the estimation results, it is reasonable to limit the analysis to the parametric estimates—the estimates obtained without resampling that are justified on the basis of asymptotic behavior. This is a practical observation as well for it means much of the hard work required to specify a model can be completed without the need for the computationally expensive bootstrap analysis. A useful estimation strategy, therefore, is one that applies bootstrap analysis only after a satisfactory model specification has been achieved. The estimate of bias in the coefficient estimates revealed by that analysis demonstrates the reasonableness of the assumption of asymptotic conditions for the evaluation of the parametric coefficient estimates. Of greater utility, however, as shown in

sections 1.6.3-5, will be the application of the empirical distribution of the coefficient estimates to the evaluation of bias and uncertainty in model predictions.

1.5.3.5 Measurement error (advanced)

We conclude this section with a discussion of the effects of measurement error on the analysis of bias and uncertainty. Measurement error can arise in the model in either the explanatory variables or the dependent variable. In linear models, it can be shown that the presence of measurement error in explanatory variables tends to bias coefficients towards zero. The intuitive understanding of this bias is that greater noise in a predictor makes it more difficult to detect a causal relation with the dependent variable, causing a reduction in the absolute value of the correlation between the dependent variable and the predictor measured with noise. A technical explanation shows the bias to arise due to correlation between the measured values of the explanatory variable and the error terms that, under conditions of measurement error in a predictor, incorporate some of the error associated with that predictor. In the limit, as the variance of the measurement error goes to infinity, it will not be possible to discern any relation between the predictor and the dependent variable, and the correlation becomes zero. The introduction of measurement error in one of the predictors has the potential of biasing the coefficient estimates for other predictors if the covariance between these predictors and the true value of the noisy predictor is non-zero. Unfortunately, the direction of this "collateral" bias cannot be predicted without knowledge of this covariance structure.

It is important to understand that the effect of measurement error in the predictors, although leading to biased coefficient estimates, does not necessarily imply bias in the model predictions. For linear models, in fact, the best prediction of the dependent variable is obtained using the coefficient estimates from standard least squares methods, without adjustment for measurement error bias. It is not immediately clear whether this assessment carries over to nonlinear models because the measurement error creates error in the model that is non-additive with respect to the dependent variable.

The ability to detect a relation between the dependent variable and the predictors may also be impeded if there is large measurement error in the dependent variable. This may be of particular concern because the dependent variable, flux, is not typically observed but is estimated from a separate relation involving streamflow (see section 1.3.1 above). As usually formulated, measurement error in the dependent variable does not result in a bias in coefficient estimates; rather, measurement error increases the mean squared error of the model, thereby proportionately inflating the standard error of all model coefficients. The measurement error introduced by the estimation of flux, however, is not the standard measurement error. The usual definition of measurement error expresses error as orthogonal to the true variable, implying the error is correlated with the measured variable. But for flux estimation, which is an expectation of true flux conditioned on streamflow and other variables, the error is orthogonal to the measurement. This implies a potential bias is introduced in the coefficient estimates if the SPARROW predictor variables are correlated with the unobserved error in flux.

To understand the nature of the bias, consider a simple analysis of such bias arising in a linear model. Let $\mathbf{y}$ be a $N \times 1$ vector of the true dependent variable and let $\tilde{\mathbf{y}}$ be its measured value. Because $\tilde{\mathbf{y}}$ is a conditional expectation of $\mathbf{y}$, we have

$$\mathbf{y} = \tilde{\mathbf{y}} + \mathbf{u}\,, \tag{1.89}$$

where $\mathbf{u}$ is a $N \times 1$ vector of error terms orthogonal to $\tilde{\mathbf{y}}$ with mean zero and variance σ_u^2. Consider a regression of $\mathbf{y}$ on a set of K predictors, denoted by the $N \times K$ matrix $\mathbf{X}$. The estimated coefficients, which are best linear unbiased, are given by $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$. The coefficients estimated with the measured dependent variable is

$$\hat{\boldsymbol{\beta}}^M = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\tilde{\mathbf{y}} = \hat{\boldsymbol{\beta}} - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}\,. \tag{1.90}$$

Thus, if the predictors are correlated with the orthogonal component $\mathbf{u}$, the estimated coefficients using the measured flux are biased relative to the true coefficients.

An upper bound on the absolute magnitude of the bias can be obtained by noting that the absolute value of the correlation between **u** and any of the predictors is bounded by 1. Equivalently, without centering of the variables, consider the regression of **u** on $\tilde{\mathbf{X}}_k$, the k^{th} column of the transformed explanatory variable $\tilde{\mathbf{X}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$. The sum of squared errors of that regression must satisfy

$$\mathbf{u}'\mathbf{u} - \mathbf{u}'\tilde{\mathbf{X}}_k \left(\tilde{\mathbf{X}}_k'\tilde{\mathbf{X}}_k\right)^{-1} \tilde{\mathbf{X}}_k'\mathbf{u} > 0\,, \tag{1.91}$$

or $\left|\tilde{\mathbf{X}}_k'\mathbf{u}\right| < \sqrt{\left(\mathbf{u}'\mathbf{u}\right)\left(\tilde{\mathbf{X}}_k'\tilde{\mathbf{X}}_k\right)}$. Let $\mathbf{D}\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)$ represent the diagonal matrix composed of the diagonal elements of the square matrix $\tilde{\mathbf{X}}'\tilde{\mathbf{X}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}$, let $\sigma_u \equiv \sqrt{\mathbf{u}'\mathbf{u}/N}$, and let the bias in $\hat{\boldsymbol{\beta}}^M$ be given by $\boldsymbol{\beta}^{\Delta} \equiv \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u} = \tilde{\mathbf{X}}'\mathbf{u}$. The bound in equation (1.91) implies a bound on the absolute value of the bias given by

$$\left|\boldsymbol{\beta}^{\Delta}\right| \le \sqrt{N}\sigma_u \mathbf{D}\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)^{1/2} \mathbf{i}\,, \tag{1.92}$$

where **i** is a $K \times 1$ vector of ones. It is obvious from equation (1.92) that the upper bound on bias goes to zero as the measurement error in the dependent variable, σ_u, goes to zero. The bias bound is also smaller the larger is the variation in the predictors; however, because of the $\sqrt{N}$ term, the bias does not go to zero as sample size goes to infinity.

All terms on the right-hand side of the inequality in equation (1.92) can be computed from information on the standard error of the flux estimates, σ_u, obtained from output of the flux estimation model, and the *K*-element vector of the standard errors of the $\hat{\boldsymbol{\beta}}^M$ coefficient estimates, $SE\left(\hat{\boldsymbol{\beta}}^M\right) = \mathbf{D}\left(\sigma_e^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)^{1/2}\mathbf{i}$, and root mean squared error of the regression model, σ_e, both obtained from regression model output. The bound given in equation (1.92) expressed in these terms is

$$\left|\boldsymbol{\beta}^{\Delta}\right| \le \sqrt{N}\,\frac{\sigma_u}{\sigma_e}\,SE\left(\hat{\boldsymbol{\beta}}^M\right). \tag{1.93}$$

Although the analysis used to obtain equation (1.93) is based on the assumption of a linear model, the bound is equally valid, in an asymptotic sense, for coefficients estimated from a nonlinear model. Note, however, that the standard error of flux, σ_u, for a SPARROW analysis would need to be in logarithm units. An approximation of this standard error can be made by taking the average across monitoring stations of the ratio of standard error of the mean flux estimate, in mass units, to the estimate of mean flux.

The primary protection against bias arising from dependent variable measurement error is to exclude stations from the analysis that have a large standard error for their flux estimate. The weighting of observations according to the standard error of the flux estimate may be another, less drastic option, although it should be noted that the problem of bias cannot be eliminated by weighting alone.

The nature of the measurement error in the dependent variable removes a potential concern in models that include nested stations (models in which some monitoring stations are located upstream of other monitoring stations). For these models, the dependent variable is also a predictor, and the measurement error in the dependent variable would seem to induce coefficient bias for the same reasons, remarked above, that predictor variable error causes bias. Because the error in this case is not correlated with the dependent variable, however, no bias will arise.

1.5.4 Evaluation of the model parameters

Parameter evaluation in SPARROW modeling has the objectives of determining whether a converged model gives *statistically sound* and *physically interpretable* coefficient values. The process of parameter evaluation commonly becomes a delicate balance, with allowances being made for one consideration in order to accommodate strong evidence or beliefs from the other. If after completing this section the reader retains a view that statistics is best practiced as an art, a proper understanding of this process will have been achieved.

1.5.4.1 Statistical evaluations

The first objective in parameter evaluation entails the appraisal of model parameters for statistical significance and the quantification of uncertainty (i.e., the range of probable values of the parameters). This provides important information for identifying unique model specifications (i.e., parameters and values for which the model predictions are sensitive) and determining the level of model complexity (i.e., number and types of explanatory variables and model functions) that can be empirically supported by the stream monitoring data. The emphasis on parameter estimation in SPARROW models has the objective of identifying the important contaminant sources and factors affecting mean-annual contaminant transport over large spatial scales in soils and in ground and surface waters.

The key parameter statistics that a user should examine include the estimated mean values of the coefficients, estimates of the variance of these coefficient estimators based on the standard error estimate, and measures of statistical significance based on statistical evaluations of the *t* statistics (ratio of the coefficient value to its standard error) (see table 1.5). These statistics are biased in finite samples but consistent as sample size goes to infinity; the *t* statistics are asymptotically distributed as a standard normal. The *p*-values are based on a two-tailed probability from a Student's *t* distribution. The *p*-values can be used to identify statistically significant model coefficients—i.e., those that are statistically distinguishable from zero—and can be used to refine the parameter set to identify parsimonious SPARROW models. The derivation of these statistics for nonlinear optimization procedures is shown in section 1.5.2.1.

Evaluations of the statistical significance of SPARROW model coefficients allow a user to determine whether the coefficients are statistically distinguishable from zero. The results of a *two-sided hypothesis test* are routinely reported in the SPARROW software. The null hypothesis (H_0) of this test is $\beta_1 = 0$ versus an alternative hypothesis (H_a) $\beta_1 \neq 0$. The reported *p* statistic is the probability that the absolute value of a statistic drawn from a Students *t* distribution, with degrees of freedom equal to the number of observations minus the number of estimated coefficients (that is, the number of coefficients not determined by prior constraints), equals or exceeds the absolute value of the computed *t* statistic for the estimated coefficient. Large absolute values of *t* are less frequently observed in the Students *t* distribution and thus are indicative of model coefficients that are more statistically distinguishable from zero. This implies the confidence intervals of statistically significant coefficients are not likely to include zero.

Because the distribution of the *t* statistic is valid only asymptotically (see section 1.5.1.3), it would be equally valid to base the *p* statistic on a Students *t* distribution having infinite degrees of freedom—that is, the standard normal distribution. Note also that if the alternative hypothesis restricts the value of the coefficient to be either positive or negative, as would be the case if the model specifies either a lower or upper bound of zero for the coefficient, it is appropriate to use a one-sided *p* statistic. One-sided *p* statistics are not reported by SPARROW, but can be easily calculated by dividing the reported two-sided *p* statistic by two.

Upon examination of the *p*-values reported in table 1.5, we determine that all but four coefficients (point-source effluent, grass land, shrub land, and large stream reach decay) are statistically significant at the 5 percent level (*p*-value < 0.05). The three source coefficients for which the null hypothesis of $\boldsymbol{\beta} = 0$ is not rejected at the 5 percent level would be considered statistically significant at the 10 percent level under the restriction that the coefficients must be positive. In that case, the null hypothesis is rejected if (*p*-value / 2) < 0.10, which is the case for these three coefficients.

Table 1.5. SPARROW estimates of model statistics for the United States national total nitrogen illustration data set.

[The land-to-water delivery variables are expressed as deviations from their national means, thereby standardizing the source coefficients to reflect the mean rate of delivery of nitrogen from each source to aquatic systems; $\overline{Q}$ is the stream reach mean-annual streamflow; "N.A." indicates not applicable; kg, kilograms; ha, hectares; yr, year; hr, hours; cm, centimeter; km, kilometers; °F, degrees Fahrenheit; ft, feet; sec, second; m, meters; <, less than; and > greater than]

Parameter	Coefficient units	Mean Coeff.	Std. Error	*t* statistic	*p*-value	Literature / expected range
Sources						
Point-source effluent	dimensionless	0.1340	0.0893	1.50	0.1343	1.0
Wet-nitrate atmospheric deposition	dimensionless	1.406	0.3762	3.74	0.0002	0 – 3#
Fertilizer use	dimensionless	0.1882	0.0433	4.35	<0.0001	0 – 1
Livestock waste	dimensionless	0.2136	0.0814	2.62	0.0090	0 – 1
Forest land	kg ha^{-1} yr^{-1}	2.82	0.8354	3.39	0.0008	0.3 – 12*
Grass land	kg ha^{-1} yr^{-1}	1.42	0.9515	1.50	0.1354	0.5 – 25*
Shrub land	kg ha^{-1} yr^{-1}	1.02	0.6565	1.55	0.1216	
Transitional land (forest-agriculture)	kg ha^{-1} yr^{-1}	75.10	18.43	4.07	<0.0001	0.3 – 40*
Urban land	kg ha^{-1} yr^{-1}	64.60	16.30	3.96	<0.0001	3 – 40*
Land-to-Water Delivery						
Permeability	hr cm^{-1}	-0.1177	0.0158	-7.42	<0.0001	N.A.
Drainage density	km^{-1}	1.575	0.4124	3.81	0.0002	N.A.
Temperature	$°F^{-1}$	-0.0331	0.0069	-4.81	<0.0001	N.A.
Reach decay						
Small streams $\overline{Q}$ <500 ft^3sec^{-1}	day^{-1}	0.3676	0.046	7.98	<0.0001	
Intermediate streams 500< $\overline{Q}$ <10,000 ft^3sec^{-1}	day^{-1}	0.1029	0.0245	4.20	<0.0001	0.005 – 2
Large streams $\overline{Q}$ >10,000 ft^3sec^{-1}	day^{-1}	-0.0003	0.0294	-0.0099	0.9921	
Reservoir decay	m yr^{-1}	7.34	1.91	3.85	0.0001	< 10
Mean square error	0.337					
Root mean square error	0.581					
Number of observations	379					
R-squared	0.910					

The land-to-water delivery of wet nitrate deposition may exceed unity because of additional contributions from wet deposition of ammonium and organic nitrogen and dry deposition of inorganic nitrogen (Alexander and others, 2001)

* Literature ranges from Jordan and Weller (1996) and Beaulac and Reckhow (1982)

The two-sided *t* statistic reported in SPARROW is equivalent to a *partial F test* (i.e., $F = t^2$) in which the test evaluates the statistical significance of a *complex* model that results from the addition of one additional explanatory variable to a *simple* model that has all of the other variables present. The simple model is therefore *nested* within the more complex model and differs by only one explanatory variable. By contrast, cases may exist in which a *nested F test* needs to be applied to determine whether the addition of more than one explanatory variable (e.g., the collection of aquatic decay variables or land-to-water delivery variables) results in a significant improvement in the performance of the model (i.e., improved explanation of the variability in the response variable). This test is not calculated as part of the SPARROW software, but can be manually calculated by the user. The nested *F* statistic is expressed as

$$F = \frac{\frac{(SSE_S - SSE_C)}{(df_S - df_C)}}{\frac{SSE_C}{df_C}}, \tag{1.94}$$

where SSE_S is the sum of squares of error of the simple model with degrees of freedom, df_S; and SSE_C is the sum of squares of error of the complex model with degrees of freedom, df_C (degrees of freedom equal the difference between the number of observations and the number of estimated parameters—excluding parameters determined by a prior constraint). The test provides a measure of the tradeoff between the reduction in error (i.e., improved explanatory power) that results from a more complex model and the estimation penalty that results from the addition of parameters and the corresponding reduction in the model degrees of freedom. The test, therefore, assesses whether the reduction in error is statistically worth the loss of information for estimating the model as measured by the degrees of freedom. As with the *t* test, the *F* test is valid only asymptotically, implying it could be replaced by a chi-square test with degrees of freedom equal to *dfs – dfc*.

One example use of a nested *F* test in SPARROW is the evaluation of a hypothesis concerning whether the addition of aquatic decay parameters to a model collectively results in a statistically significant improvement in the overall model performance. In this test, we compare the more complex model containing aquatic decay variables as given in table 1.5 (MSE equals 0.337) with a simple model wherein both the in-stream and reservoir decay coefficients are removed (MSE equals 0.767). In this case, an *F* statistic is computed such that

$$F = \frac{\frac{(283.6 - 122.4)}{(379 - 12) - (379 - 16)}}{\frac{122.4}{(379 - 16)}} = 119.5. \tag{1.95}$$

The *p*-value (less than 0.00001) associated with this F statistic is highly significant, and indicates that the addition of the aquatic decay coefficients provides a significant improvement in the explanatory power of the model. Note that this test does not indicate that all of the aquatic coefficients are significantly distinguishable from zero, but only that at least one of the coefficients is. The results of a partial *F* test (i.e., the individual coefficient *t* statistics) must be examined to determine the significance of individual coefficients.

1.5.4.2 Physical interpretations

A *second* complementary objective in assessing SPARROW model parameters is the evaluation of the parameters for their physical interpretability. This objective entails the evaluation of the sign and magnitude of model coefficients to test hypotheses about the importance of different contaminant sources and the hydrologic and biogeochemical processes that are represented by the explanatory variables of the model. The interpretability of the parameters and their relation to specific processes is enhanced in SPARROW by the use of a mass balance, mechanistic structure that explicitly separates the terrestrial and aquatic properties of watersheds and accounts for nonlinear interactions among watershed properties (see section 1.2.2), together with an emphasis on the statistical estimation of parameter values. As discussed in section 1.2.3, the SPARROW model

parameters reflect the net effects over large spatial scales of an aggregate set of hydrologic and biogeochemical processes and human-related activities.

The sign of SPARROW model coefficients can be evaluated to determine the direction of the relation of any explanatory variable to the in-stream estimates of the mean-annual flux (i.e., the model response variable). The direction of the relation should be assessed for consistency with the anticipated response based on available theoretical or empirical information about processes that may be related to individual explanatory factors. For example, for the model results shown in table 1.5, a negative sign on the soil permeability coefficient indicates that total nitrogen loads in streams are inversely related to permeability—i.e., in-stream loads of nitrogen are generally lower in watersheds with highly permeable soils. This relation is frequently found in SPARROW nitrogen models and is consistent with the storage or permanent removal (i.e., denitrification) of nitrogen in soils and the subsurface. The relation indicates that nitrogen losses are larger (and in-stream nitrogen flux smaller) in watersheds where water and nitrogen are more readily routed through permeable soils. The sign of the coefficient is also important in estimating physically meaningful contaminant source terms in SPARROW. Interpretable sources within the model are generally expected to contribute positive mass to the watersheds. In fact, we often constrain the sources to be positive; thus, a one-sided hypothesis test is frequently of interest in evaluating the statistical evidence of the importance of source inputs in the model. Constraints on the coefficient sign are generally not applied to land-to-water delivery factors as there is commonly no compelling prior expectation as to the nature of the physical relation to flux. Constraints on the aquatic decay factors are also generally unnecessary; however, there may be a need to constrain the "large" river decay rates (mean rates are frequently near zero with a considerable fraction of the parameter distribution below zero) and reservoir decay rates to positive values in bootstrap executions of final SPARROW models to obtain a more physically realistic simulation of contaminant transport in rivers (i.e., negative portions of the parameter distribution may unrealistically skew the estimates of the mean; e.g., see discussion in Alexander, Elliott, and others, 2002).

The values of selected source and aquatic decay coefficients should also be evaluated to determine whether they are consistent with the range of values expected on the basis of literature studies and the prevailing information on experimental reaction rates. For the source coefficients to be easily interpreted, they must be standardized for mean levels of the land-to-water delivery variables (see section 1.4.3), such as those shown in table 1.5. It is important to note that the coefficients of the land-to-water factors cannot be interpreted individually in terms of a contaminant transport rate that is specific to the landscape property, but must be combined with individual sources to quantify an aggregate delivery of the contaminant mass to streams. By contrast, the aquatic decay coefficients can be directly interpreted without any standardization. For example, the rates of nitrogen removal in streams (ranging from near zero to 0.37 day^{-1}) and reservoirs (7.3 meters yr^{-1}) reported in table 1.5 can be directly compared to literature rates, as illustrated in previous sections of this report.

Source-related coefficients that are based on source inputs expressed in areal units, such as the land-use source terms (forest, grass, shrub, urban) in table 1.5, describe the mass per unit area delivered to streams from these land areas. These areal expressions of contaminant transport or "export" can be directly compared with ranges of export coefficients that are frequently reported in the literature (e.g., Beaulac and Reckhow, 1982). Coefficients reported for different land uses such as those in table 1.5 generally compare favorably with export coefficients reported in the literature. The SPARROW estimated export coefficients in table 1.5 are standardized to reflect the supply and delivery of nitrogen to aquatic systems under the mean levels of the landscape delivery factors in the model. Of course, one complicating aspect of such a comparison is that the literature export coefficients implicitly include the effects of watershed properties (e.g., soils, climate, in-stream processes) on nutrient transport that likely differ from those in the SPARROW model. Nevertheless, SPARROW estimates and export coefficients reported in the literature are consistent in indicating that the nutrient supply and delivery to streams and reservoirs is generally larger in urban and agricultural watersheds; much lower export coefficients are found in forests and in grass and shrub lands, where relatively small natural sources (e.g., nitrogen fixation by vegetation) of nitrogen predominate.

Other source coefficients that are expressed in dimensionless units provide a measure of the fraction of the contaminant that is delivered from each source to streams, rivers, and reservoirs. These coefficients can be evaluated to determine how reasonably they reflect the net mean rates of contaminant removal by a source as part of the delivery to aquatic systems. For example, about 18 percent of the fertilizer inputs of nitrogen are delivered to streams based on the model results reported in table 1.5. Such large losses of fertilizer inputs are generally expected and reflect the numerous processes and activities that remove nitrogen from agricultural lands and along subsurface flow paths. The estimated fertilizer coefficient reflects the aggregate effects of these

factors and may include the volatilization of ammonia fertilizer forms, the removal of nitrogen in harvested crops, and long-term immobilization of nitrogen and denitrification in soils and ground waters. In the case of atmospheric deposition, the greater than unity coefficient of 1.4 is consistent with additional contributions from wet deposition of organic nitrogen and dry deposition of inorganic nitrogen, which are not included in the wet nitrate measurements input to the model. This result would be expected, provided that these unmeasured quantities are correlated with the measured wet deposition, which is commonly the case (Alexander and others, 2001).

When direct measures of point source loadings (e.g., municipal wastewater effluent) are used in the model (and the response variable has identical units), the point-source coefficient estimated in SPARROW is expected to be close to 1.0 (i.e., the confidence interval should contain 1.0). A significant deviation from 1.0 for the estimated point-source coefficient may indicate a poor specification of the model or inaccurately measured point-source effluent data. Point-source coefficients for the national and selected regional total nitrogen models are shown in figure 1.22. The confidence interval for most of the regional models contains 1.0, and many mean estimates are also very close to 1.0 in value. By contrast, the national model displays a very low coefficient (less than 0.20) that suggests appreciable bias in the point-source coefficient estimate. Because of the known poor quality of the wastewater treatment plant estimates of nitrogen loads in the national dataset, it seems likely that the bias reflects point-source data quality problems rather than a misspecification of the model.

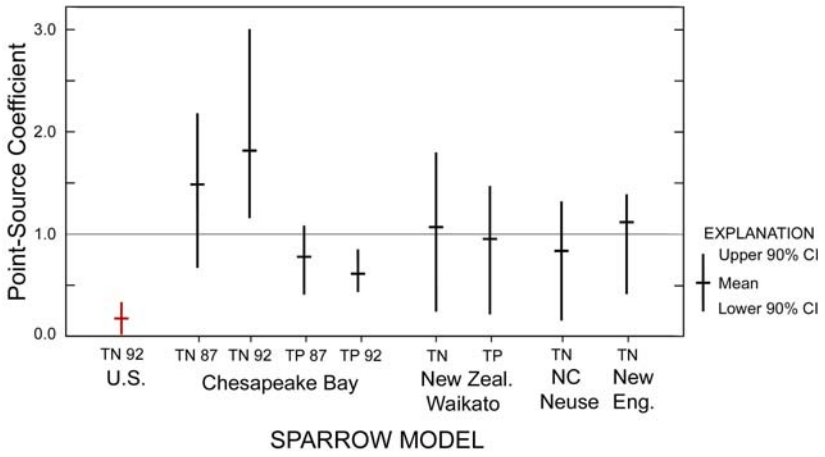

Figure 1.22. Estimated municipal wastewater treatment coefficient in the national and regional SPARROW models. [The United States (U.S.) 1992 model is based on estimates of municipal/industrial nitrogen loads from a 1992 data retrieval from the U.S. Environmental Protection Agency Permit Compliance System (PCS); TN, total nitrogen; TP, total phosphorus; CI, confidence interval.]

1.5.4.3 Statistical insignificance and multicollinearity

A SPARROW model coefficient that is statistically insignificant (e.g., p equals 0.50) indicates that the estimated mean value of the coefficient is statistically indistinguishable from zero and that a large proportion of the parameter distribution would lie below zero (i.e., the confidence interval includes zero). This implies that the associated explanatory variable is relatively uncorrelated with the response variable. It is important to recognize that this outcome of the statistical evaluation of the coefficient does not necessarily indicate that the watershed properties represented by this variable are *intrinsically* unimportant in affecting the supply and transport of contaminants in the modeled region. Several possible statistical factors may explain the occurrence of statistically insignificant coefficients that should be considered in evaluating the model fit and coefficient estimates, including the number of observations (i.e., station mean loads) in the regression (i.e., *quantity* of the information), the amount of variability in the explanatory factors (i.e., *quality* of the information content), and the level of collinear variability in the explanatory factors (i.e., *multicollinearity*).

One cause of a statistically insignificant coefficient is the lack of a sufficient *quantity* of stream monitoring data. The statistical power to detect the effects of explanatory factors on stream contaminant loads in a SPARROW model is dependent on the number of observations (i.e., monitoring station mean flux

measurements) used in the nonlinear regression. As discussed in section 1.2.4, the number of stream monitoring stations influences the level of complexity (i.e., number of explanatory variables) that can be supported in SPARROW models. For example, we find that fewer explanatory variables—typically from six to eight—are statistically significant in many of the regional models as compared to upwards of 18 or more variables in the national model. Therefore, models with fewer station flux measurements are generally more limited in their ability to identify statistically significant explanatory variables.

A second cause of a statistically insignificant coefficient is the lack of sufficient spatial variability in an explanatory factor (in the introduction, we cited this as related to issues of the *quality* of the data). The effect of explanatory factors on stream contaminant flux can be difficult to detect in SPARROW models if the spatial variability in the factor is relatively small over the modeled region. For example, precipitation is clearly an important contributor in determining the magnitude of stream contaminant flux at regional and national spatial scales. In many of the regional SPARROW models, however, variability in mean-annual precipitation is small across the regions (i.e., variations that are less than an order of magnitude) and this factor is rarely found to be statistically significant as a land-to-water delivery factor. By contrast, mean-annual precipitation varies by several orders of magnitude in the national SPARROW model and in the New Zealand national model (Elliott and others, 2005) and has been found to be highly significant as a delivery factor in recent versions of these models. Given the level of statistical power for many of the regional models, the spatial variability in the regional measures of precipitation in comparison to that of other controlling factors in the models is typically insufficient to support the estimation of an explicit precipitation term in the models. It is important to note that this does not imply that a model without precipitation data as input is invalid as a prediction tool. Indeed, such a model can be reliably used to predict in-stream flux and the contributions of pollutant sources to streams. The model does not, however, provide an explicit description of how precipitation influences pollutant flux, and therefore could not be used to assess climate-related effects on stream water quality.

It is also noteworthy that the source coefficient for a relatively small contaminant source (e.g., natural or background inputs of nitrogen) may be difficult to estimate with a high degree of statistical significance because the true numerical value of the coefficient is small, especially relative to its level of precision (i.e., standard error of estimate). The detection of only weak statistical significance for such a variable does not necessarily provide sufficient cause to exclude it, especially if the intent in using the model is to provide a comprehensive understanding of contaminant sources. For example, the grass and shrub land export coefficients are only weakly significant in the national SPARROW model illustrated in table 1.5, although the level of precision of these terms is equal to or even surpasses the precision associated with the more highly significant forest export coefficient. Nitrogen from natural fixation on grass and shrub lands is generally smaller in comparison to nitrogen generated from fixation and other sources in forests (Jordan and Weller, 1996; nitrogen export from forested land may also include some contributions from atmospheric deposition). This is a likely explanation for why the estimated mean nitrogen export from grass and shrub land (table 1.5) is only about one half of that estimated for forested lands.

Finally, another potential explanation for the lack of statistical significance in two or more explanatory variables is the effect of *multicollinearity* on the variance of the model parameters. Multicollinearity describes the presence of high levels of correlation between two or more explanatory variables in a regression model that cause all of the correlated variables to have statistically insignificant coefficients. SPARROW provides several statistics and matrices that are useful for evaluating the presence and causes of multicollinearity.

The problem of mulicollinearity is one of model interpretability rather than model validity. The presence of multicollinearity does not imply the model coefficients or their standard errors are estimated with bias. Moreover, the predictions from the model are asymptotically minimum variance unbiased. The most serious consequence of multicollinearity is that coefficients associated with collinear variables (or, in the case of the nonlinear SPARROW model, collinear gradients) are imprecisely estimated. This lack of precision is reflected in large standard errors and a tendency for coefficients of collinear variables to be individually insignificant. Thus, in cases of multicollinearity, a coefficient estimated as statistically insignificant may in fact represent an important process in the model, but its incremental contribution to model fit is masked by other collinear processes. Indicators of multicollinearity are useful therefore in distinguishing coefficients that have potential significance and coefficients that truly should be dropped from the model.

An interesting situation arises if two coefficients are collinear but one coefficient is statistically significant and the other is not. It can be shown that collinearity does not affect the ratio of variance between two coefficients or, therefore, the ratio of *t*-statistics. It can be concluded therefore that the signal being

transmitted through the significant coefficient from its associated predictor is quantitatively more important than the signal transmitted through the insignificant coefficient. In other words, collinearity in this case is not so strong that it masks the contribution of a quantitatively important predictor. For example, as illustration of this, we modified the total nitrogen model described in section 1.4.4 so two highly spatially correlated atmospheric deposition sources, wet nitrate and ammonia, are included in the model. In the resulting model, we find that only the nitrate deposition coefficient is statistically significant (p equals 0.0007), whereas the ammonia deposition coefficient is negative and statistically insignificant (p equals 0.90). This result suggests that the strongest atmospheric deposition effect on in-stream nitrogen flux is apparent from the wet nitrate deposition source in the model.

The *variance inflation factor* (VIF) is a commonly used statistic for determining the importance of multicollinearity. Under linear least squares, the variance inflation factor for coefficient k, VIF_k, is given by the k^{th} diagonal element of the $\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}$ matrix, where $\tilde{\mathbf{X}}$ is the $N \times (K-1)$ matrix of predictor variables, excluding the intercept, centered and scaled to unit length (Montgomery and Peck, 1982). That is, observation i for predictor variable k has been transformed according to $\tilde{X}_{ik} = \left(X_{ik} - \bar{X}_k\right) \Big/ S_k^{1/2}$, where $\bar{X}_k$ is the mean of the N observations of the k^{th} variable and $S_k = \sum_{i=1}^{N} \left(X_{ik} - \bar{X}_k\right)^2$. It can be shown (Montgomery and Peck, 1982) that

$$VIF_k = \frac{1}{1 - R_k^2}, \tag{1.96}$$

where R_k^2 is the coefficient of multiple determination from the regression of $\mathbf{X}_k$ on the remaining $K-1$ predictor variables, including an intercept. If there is a close relation between variable k and the remaining variables, then R_k^2 is near one and the variance inflation factor is large. Conversely, if the kth variable is independent of the other variables, then R_k^2 is near zero and the variance inflation factor is near its lower bound of one.

Another useful interpretation of the variance inflation factor relates to the effect that collinearity of the predictors has on coefficient variance, t-statistics, and confidence intervals (Montgomery and Peck, 1982). The square root of the k^{th} coefficient's variance, given by the model root mean squared error times the k^{th} diagonal element of the inverse of the $\mathbf{X}'\mathbf{X}$ matrix, is proportional to the length of the k^{th} coefficient's symmetric confidence interval and is inversely proportional to the magnitude of the k^{th} coefficient's t-statistic. Suppose observations could be chosen in such a way that each predictor is independent of all others but retains the predictor variances exhibited in the original sample. Such a sampling scheme, called an orthogonal design, has no collinearity and results in the smallest possible coefficient variances—that is, the smallest possible values along the diagonal of the $\mathbf{X}'\mathbf{X}$ matrix. Consequently, orthogonal design sampling results in the smallest possible (symmetric) confidence intervals and largest possible t-statistics for the estimated coefficients. It can be shown that the variance inflation factor for a coefficient is equal to the ratio of that coefficient's variance to the coefficient's variance that would be possible under orthogonal design. The square root of the variance inflation factor, therefore, represents the proportion by which the t-statistic could be increased if multicollinearity were eliminated. This insight provides a useful interpretation of the variance inflation factor. If a coefficient is insignificant, and inflating the coefficient's t-statistic by the square root of its variance inflation factor fails to make the coefficient significant, then multicollinearity is an unlikely explanation of the coefficient's insignificance. Conversely, if applying the inflation factor makes the coefficient significant then it is possible that multicollinearity is masking the significance of the coefficient.

To apply the variance inflation factor to a nonlinear model, and thus provide for interpretation of collinear coefficients as described above, the gradients (see section 1.5.1.2) evaluated at the final coefficient estimates, $\mathbf{f}_{\boldsymbol{\beta}}^{*}\left(\hat{\boldsymbol{\beta}}\right)$, are substituted for the predictor variables, $\mathbf{X}$. Because a SPARROW model typically has no intercept, however, it is inappropriate to center the gradients prior to normalizing to unit length. This is because the R^2 statistic implied by a variance inflation factor computed from centered predictors is a valid indicator of

explanatory power only if the set of predictors includes an intercept; the relation between the variance inflation factor computed using centered predictors and coefficient variance does not hold absent the intercept term. SPARROW automatically tests the gradient vectors to determine if they include an intercept term. If no intercept is present, the normalization of the gradient vectors is performed without centering, resulting in the computation of an uncentered variance inflation factor, $\overline{\text{VIF}}$. This factor can be used to determine the potential effect collinearity has on the coefficient *t*-statistics in exactly the same way the standard variance inflation factor is used if an intercept is present.

The uncentered variance inflation factor also bears a relation to a fit statistic. That is, $\overline{\text{VIF}}_k = 1/\left(1-\overline{R}_k^2\right)$, where $\overline{R}_k^2$ is the uncentered *r*-square statistic formed by regressing the k^{th} gradient on the remaining $K-1$ gradients. The uncentered *r*-square statistic, defined as the ratio of the sum of squares of the regression predicted values to the sum of squares of the regression dependent variable, is commonly used in place of the normal *r*-square if the regression does not include an intercept. The uncentered *r*-square statistic is bounded between 0 and 1 and will always exceed the standard *r*-square. This implies that the uncentered variance inflation factor always exceeds the standard variance inflation factor, the relation between them being

$$\overline{\text{VIF}}_k = \text{VIF}_k\left(1+\frac{1}{\text{CV}_k^2}\right), \tag{1.97}$$

where CV_k is the coefficient of variation for the k^{th} gradient.

As an illustration of the effects of multicollinearity in a SPARROW model, we modified the total nitrogen model described in section 1.4.4 by adding a new explanatory variable that is the square of an existing variable (see results in table 1.6). In this example, the new variable (TEMP2) is the square of temperature (TEMP), which is a statistically significant (*p* less than 0.0001) land-to-water delivery factor in the original model. The new estimated coefficients are both statistically insignificant with, $\overline{\text{VIF}}$ values of about 30. The magnitude of the variance inflation factor indicates that the variance of the coefficients has been inflated by about a factor of five (i.e., $\sqrt{\overline{\text{VIF}}}$ equals 5.5). At least one of the coefficients (TEMP2) would be statistically significant (*t* equals 3.7; *p* equals 0.0001) if this effect were accounted for and suggests that multicollinearity could be masking the significance of the coefficient.

Table 1.6. Model coefficient results for two correlated temperature land-to-water delivery factors in the national total nitrogen model.

[The total nitrogen model contains source, land-to-water delivery, and discrete reach-decay variables as described in Alexander and others (2000), a reservoir removal rate specified according to equation (1.35); the model, applied to the Enhanced Reach File 1 (ERF1) version 2.0 infrastructure as described in Nolan and others (2002), was modified by adding a new temperature variable (TEMP2) that is equal to the square of a land-to-water temperature variable (TEMP) already in the model; VIF is the variance inflation factor]

Parameter	Coefficient	Standard Error	*t* statistic	*p*-value	$\overline{\text{VIF}}$	Eigenvector term
Temperature (TEMP)	-0.0046	0.0382	-0.1197	0.9048	30.42	0.7072
Temperature squared (TEMP2)	-0.0010	0.0015	-0.6769	0.4989	29.51	-0.6957
Eigenvalue Spread	312.4					

Another statistic reported by SPARROW to help identify multicollinearity is the *eigenvalue spread*. The eigenvalue spread is computed from the eigenvalues of the $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ matrix of normalized gradients. If an intercept is absent from the model then the normalized gradients are uncentered prior to normalization. The eigenvalues of the $K \times K$ matrix $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ represent the K roots, denoted λ, of the equation $\det\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} - \lambda\mathbf{I}\right) = 0$. Because $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ is a positive semi-definite matrix, all of its eigenvalues must be greater than or equal to zero. The eigenvalue spread is defined as

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}. \tag{1.98}$$

If the $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ matrix is nearly singular, an implication of multicollinearity among the gradients, then one or more eigenvalues will be near zero; a large value for eigenvalue spread is therefore evidence of multicollinearity. In practice, if the eigenvalue spread is less than 100 there is no serious problem with multicollinearity (Montgomery and Peck, 1982). As discussed above, however, issues of collinearity make sense only in the context of determining coefficient significance. The fact that a general model statistic like the eigenvalue spread is large does not necessarily imply any of the coefficients are insignificant or help identify which coefficients have statistical significance that is sensitive to collinearity. According to our illustration model results in table 1.6, the eigenvalue spread was reported as being well above 100.

Perhaps the most useful interpretation derived from the $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ matrix is the use of its *eigensystem* for determining which coefficients are related to each other through collinear gradients. Inference on this issue can be ascertained by looking at the eigenvectors corresponding to very small eigenvalues. The $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ matrix can be factored into the following eigensystem

$$\tilde{\mathbf{X}}'\tilde{\mathbf{X}} = \mathbf{C}\boldsymbol{\Lambda}\mathbf{C}', \tag{1.99}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix whose elements are the eigenvalues, λ, of $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$, and $\mathbf{C}$ is a $K \times K$ orthogonal matrix having the property that $\mathbf{C}'\mathbf{C} = \mathbf{I}$. The k^{th} column of $\mathbf{C}$ is called the k^{th} eigenvector corresponding to eigenvalue λ_k. Pre- and post-multiplying $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ by $\mathbf{C}'$ and $\mathbf{C}$ results in the relation $\mathbf{C}'\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\mathbf{C} = \boldsymbol{\Lambda}$. Define $\mathbf{Z} \equiv \tilde{\mathbf{X}}\mathbf{C}$. Then, for each k,

$$\lambda_k = \sum_{i=1}^{N} Z_{k,i}^2. \tag{1.100}$$

Suppose the k^{th} eigenvalue is nearly zero—indicating collinearity. Then equation (1.100) implies that for each observation i, $Z_{k,i}$ is nearly zero which, through the definition of $\mathbf{Z}$, implies

$$\sum_{j=1}^{K} C_{j,k}\tilde{X}_{i,j} \approx 0. \tag{1.101}$$

That is, the k^{th} eigenvector represents the coefficients that define a collinear grouping of the normalized gradients. Because the normalized gradients are unitless, so too are the elements of the eigenvector, implying the values of individual terms are comparable. Therefore, the largest absolute value elements of the k^{th} eigenvector effectively define the group of gradients that are collinear.

A table of eigenvalues and eigenvectors is reported in the SPARROW software output that lists the eigensystem of the $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ matrix (see figure 1.23). The first row of the eigensystem output gives the K eigenvalues, and the column beneath each eigenvalue represents the associated eigenvector. Insight into the collinear structure of the model is obtained by first looking across the first row to determine if there are any eigenvalues near zero. If an element in the first row is near zero, then the largest absolute value elements in the column below it correspond to the predictors that form a set of collinear gradients. According to our illustration results for the model given in table 1.6, the largest absolute value eigenvector elements in the column corresponding to the smallest eigenvalue appear for the two temperature variables and have values of 0.7072

and -0.6956 for TEMP and TEMP2, respectively. All other eigenvector elements—those associated with all the other variables in this column—are near zero, between -0.05 and 0.05. The high values of the eigenvector elements for TEMP and TEMP2 indicate the presence of two collinear variables in the model.

X'X Eigenvalues and Eigenvectors (NC)

EigVal	5.1884968	1.9777856	1.1533937	0.9799855	0.7553549	0.7217894	0.6113136
BSEWER	0.2254973	0.1021533	0.3594944	-0.168021	-0.159884	0.7694809	-0.332118
BATMDEP	0.3747209	-0.101274	0.1471633	-0.024958	0.0814361	-0.019034	0.0203806
BFERTILIZER	0.351207	0.0645221	-0.09454	0.1400642	0.2064662	0.0307999	0.2205545
BWASTE	0.387925	-0.023012	-0.102404	0.1151898	0.1702441	0.0476112	0.1495758
BNONAGR	0.3479573	-0.231675	0.0378305	0.0116486	-0.050623	-0.124095	-0.149289
BPERM	0.0495452	0.1043679	0.7957816	-0.270159	0.0986155	-0.253465	0.3789936
BDRAINDEN	-0.294857	0.0172986	-0.112527	-0.019004	0.0937108	0.5043171	0.3704468
BTEMP	0.1330067	0.6614915	-0.136807	-0.00562	-0.034329	-0.035845	-0.006599
BTEMP2	0.1303199	0.6629731	-0.08332	-0.019791	-0.037106	-0.091304	-0.012972
BRCHDECAY1	-0.352947	0.0844163	-0.000113	-0.144732	-0.013334	0.091986	0.4104304
BRCHDECAY2	-0.325864	0.1055234	0.1651185	-0.116162	-0.239863	-0.220172	-0.474135
BRCHDECAY3	-0.083178	0.0639104	0.3138155	0.8779462	-0.319343	0.0412347	0.1054922
BRESDECAY	-0.228215	0.1014894	0.1758348	0.2431518	0.8443204	0.0456366	-0.332436

X'X Eigenvalues and Eigenvectors (NC)

EigVal	0.564629	0.4542545	0.2790165	0.1773496	0.1200222	0.0166086
BSEWER	-0.190897	-0.071404	-0.04427	-0.038761	0.0907486	-0.006677
BATMDEP	0.209349	-0.277887	0.6694236	0.3077171	-0.391057	-0.053355
BFERTILIZER	-0.429877	0.4639021	0.0943289	0.4926532	0.3162263	0.0104697
BWASTE	0.030653	0.1975653	0.3035251	-0.783853	0.1555691	-0.029668
BNONAGR	0.5300649	-0.19756	-0.172262	0.1551304	0.633239	0.093305
BPERM	0.065174	0.1623297	-0.158336	-0.067655	-0.021207	0.0457924
BDRAINDEN	0.5888851	0.3599148	0.0771978	0.1278079	-0.018887	-0.020327
BTEMP	0.1223889	-0.067793	0.0108967	0.0057323	-0.060361	0.7071872
BTEMP2	0.1542281	-0.087312	-0.067181	0.0354731	0.0715686	-0.695676
BRCHDECAY1	-0.239253	-0.50032	0.3572757	-0.019365	0.4838683	0.0234888
BRCHDECAY2	0.0643169	0.4335056	0.5046819	-0.001859	0.2534186	0.01067
BRCHDECAY3	0.0230054	-0.054657	0.0346015	0.0014566	0.0139829	0.0025489
BRESDECAY	0.0363556	-0.12276	0.0019742	0.0050876	0.0743614	0.0123512

Figure 1.23. SAS output showing the eigensystem from an example SPARROW model.

In the event that multicollinearity is identified as a problem for a particular model specification, the following corrective actions are suggested (although none of these are completely satisfying and/or consistently successful). The *first* defense against multicollinearity is to collect more monitoring data. The standard errors of coefficients are inversely proportional to the square root of the number of observations. Therefore, increasing the number of observations has the effect of enlarging the magnitude of *t*-statistics, making it more likely that a given value of a coefficient is significant. A *second* approach is to simply remove one of the coefficients associated with a collinear set of gradients. Although this approach could lead to a misspecified model and thereby bias the estimates of coefficients, it could also improve the accuracy with which other collinear coefficients are estimated, thereby increasing their significance. *Finally*, if it is suspected that collinearity is causing a group of coefficients to be individually insignificant, it is possible to form a statistical test, an *F* test, that jointly evaluates their significance (see the previous discussion of the *F* test in this section; the *F* statistic is given in equation (1.94)). A significant *F* statistic is evidence that the collinear coefficients are jointly significant and that collinearity is masking the significance of individual coefficients. Unfortunately, it is not possible to know if one or all of the collinear coefficients belong in the model. Moreover, the *F* test is not appropriate when one of the collinear coefficients is individually significant.

The explanatory variable *covariance* and *correlation matrices* provide additional information about collinear relations between the variables; however, this evaluation is less useful if the collinearity involves more than two predictors. The covariance matrix describes the covariances between the estimated coefficients that

arises from the particular finite sample used to estimate the model. The reported covariances for the nonlinear SPARROW model are asymptotically valid, meaning that the estimates are valid in large samples but are only suggestive in small samples. The $K \times K$ covariance matrix is computed as the mean squared error of the model times the inverse of the $\mathbf{X}'\mathbf{X}$ matrix, where $\mathbf{X}$ in the nonlinear context is the $N \times K$ matrix of gradients corresponding to each of the K parameters. The matrix is symmetric with the coefficient variances along the diagonal. Because covariances are somewhat difficult to interpret, and depend on the units of the underlying variables, the associated correlation matrix provides a more readily interpreted metric for examination. An element in the correlation matrix represents the correlation between two estimated coefficients. The element given in the *i*th row and *j*th column is computed by taking the covariance between the *i*th and *j*th coefficients and dividing by the square root of the product of the variances for the *i*th and *j*th coefficients. As with correlations in general, the elements of this matrix must lie between –1 and 1. Because a coefficient estimate is perfectly correlated with itself, the elements along the diagonal are set to one. As previously explained in this section, collinear predictors tend to have coefficients with large standard errors—the large standard errors arising from large covariance among the collinear coefficients. The correlation matrix can be useful in identifying the bivariate case of multicollinearity—that is, collinearity between only two predictors; in this case the coefficients estimated for the two predictors will have high variance and a large mutual correlation. A simple way to evaluate suspected collinearity between two statistically insignificant coefficients is therefore to check the correlation matrix for a large value of correlation between these two coefficients.

1.5.5 Evaluation of model errors

The estimated residuals from the model contain a great deal of information for evaluating model specification. The assumptions of the model (see section 1.5.1.2) require the weighted residuals to be identically distributed (homoscedastic), independent across observations, and uncorrelated with the explanatory variables. In this section, we describe various statistics and graphical procedures that are useful for evaluating the reasonableness of these assumptions for a given SPARROW application.

1.5.5.1 Heteroscedasticity

Estimation of a SPARROW model, based on nonlinear least squares methodology, requires that the model residuals be independent and identically distributed. The residuals are not required to be normally distributed; however, certain types of departures of the residuals from normality are also indicative of cases where the residuals are *heteroscedastic*—that is, not identically distributed. Heteroscedastic residuals may present problems for the interpretation of coefficient test statistics, which are inconsistent (biased in large samples) if the variance of the residuals is systematically related in some way to the predictors or, for the nonlinear model, to the gradients (White, 1980). Heteroscedastic residuals also cause the estimated model to be inefficient (Judge and others, 1985).

Departure of the residuals' distribution from normality does not necessarily invalidate the SPARROW model. The test statistics used for validating coefficient significance are based on large sample properties that assure normality regardless of the underlying form of the residual distribution. With regard to prediction, departures of the residuals' distribution from normality can affect the validity of certain methods used for transformation of predictions from logarithm space to real space. However, this concern does not apply to the Smearing estimator used for SPARROW transformations; the Smearing estimator is consistent regardless of the error distribution (see section 1.6.2).

Evidence of problems related to heteroscedasticity can be obtained primarily by inspection of a set of four diagnostic graphs shown in figure 1.24; the graphs are generated using the example nitrogen model described above in table 1.6. The first plot is of the observed versus predicted flux in log units (figure 1.24a). The graphed points should exhibit an even spread about the one-to-one line (the straight line in figure 1.24a) with no outliers. A common pattern expressed in this graph for SPARROW nutrient models is the tendency for larger scatter among observations with smaller predicted flux—a pattern of heteroscedasticity. One possible cause for this pattern is greater error in the measurement of flux in small basins due to greater variability in flow or to greater relative inhomogeneity of contaminant sources within small basins. If the heteroscedasticity is caused by measurement error, then appropriate assignment of weights reflecting the relative measurement error in each observation (plus an additional common model error) can improve the coefficient estimates and correct the inference of coefficient error. If the heteroscedasticity is due to structural features of the model, the

observations can be weighted to improve the coefficient estimates and correct their estimates of error. Alternatively, the heteroscedasticity observed in this graph could be caused by structural processes that are not yet included in the model.

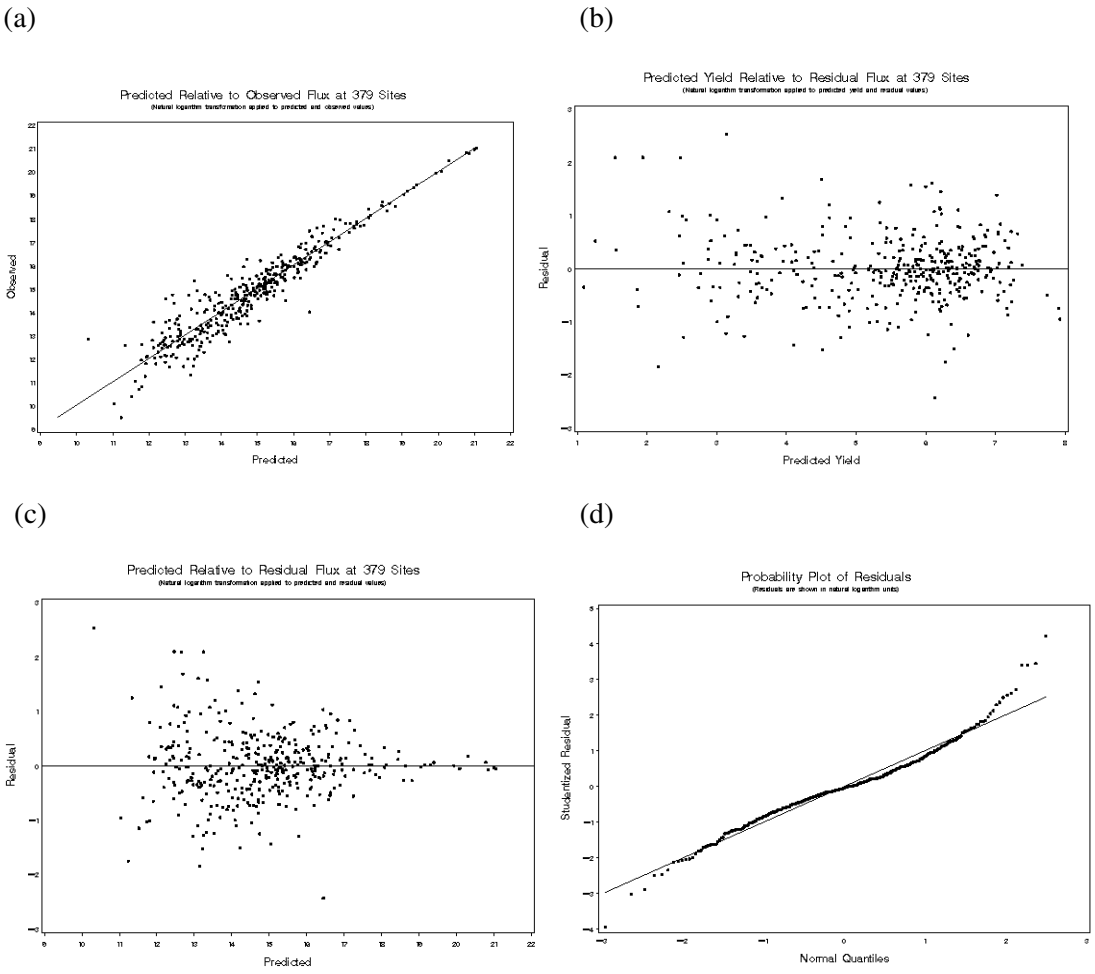

Figure 1.24. Diagnostic plots for evaluating SPARROW model errors and adherence of the residuals to the model assumptions: (a) predicted and observed flux; (b) residuals and predicted yield; (c) residuals and predicted flux; and (d) a probability plot of residuals.

The pattern of predicted versus observed logarithm of flux may also indicate systematic bias in the model. A significant deviation of the plotted points from the one-to-one line in a particular region of the graph indicates the model is structurally biased. Structural bias of this kind implies the residuals of the model are likely to be correlated with the predictors (another example of failure of the third assumption that residuals are independent of predictors) and may result in biased coefficient estimates. Such bias is generally not eliminated by including additional observations; rather, it is likely that an important predictor—one associated with basin scale—is absent from the model. Identifying such a predictor will usually correct the problem and remove the region-specific bias of residuals from the one-to-one line.

The plot of log residuals versus predicted yield (i.e., mass per unit of drainage area), as shown in figure 1.24b, is also useful for validating the model fit. The graphed points once again should exhibit an even spread about the one-to-one line, with no outliers. The graph is useful for identifying and diagnosing bias and heteroscedasticity in much the same way as the graph of predicted versus observed log of flux (fig. 1.24a). The conversion to yield units, however, tends to remove scale effects, such as those related to drainage area. Deviations from the one-to-one line in this graph are indicative of a systematic bias or misspecification of the model at the watershed scale related to specific land-to-water or in-stream processes, such as reservoir

attenuation. In this case, including an additional process or modifying the functional form of an existing process may solve the problem.

A plot of log residuals versus predicted flux, as shown in figure 1.24c, provides a third check of whether residuals meet the assumptions of the least squares methodology: the residuals should not vary systematically either in terms of spread or bias with the predictions. The plotted residuals are the weighted residuals, $\hat{e}_i\sqrt{w_i}$, where $\hat{e}_i$ is the estimated residual from the fitted model and w_i is the associated weight assigned to each observation. Under heteroscedasticity, unweighted residuals may exhibit varying levels of spread across the range of predictions. If a proper weighting of the observations has been applied, so that the heteroscedasticity is removed, the residuals in figure 1.24c will show a common spread that is centered near zero throughout the range of predictions (homoscedasticity). A user may thus test various assignments of weights by comparing figures 1.24a and 1.24c: weights are optimal if the systematic pattern of heteroscedasticity in figure 1.24a is absent from figure 1.24c.

A fourth type of graph that is indicative of cases of heteroscedasticity, but is most commonly used to identify non-normally distributed residuals, is a probability plot of the model residuals, as shown in figure 1.24d. The probability plot depicts the relation between the empirical distribution of the residuals and the normal distribution: specifically, it is the scatter plot relating the ordered standardized weighted residuals, e_i^* and the quantiles of the adjusted ranks q_i. The standardized weighted residuals have the form

$$e_i^* = \hat{e}_i\sqrt{\frac{w_i}{s^2\left(1-h_i\right)}}, \tag{1.102}$$

where $\hat{e}_i$ is the estimated least-squares residual, w_i is the weight for the ith observation, s^2 is the mean squared weighted error of the model, and h_i is the leverage of observation *i*,

$$h_i = \mathbf{f}^*_{\boldsymbol{\beta}',i}\left(\hat{\boldsymbol{\beta}}\right)\left(\mathbf{f}^*_{\boldsymbol{\beta}'}\left(\hat{\boldsymbol{\beta}}\right)'\mathbf{f}^*_{\boldsymbol{\beta}'}\left(\hat{\boldsymbol{\beta}}\right)\right)^{-1}\mathbf{f}^*_{\boldsymbol{\beta}',i}\left(\hat{\boldsymbol{\beta}}\right)', \tag{1.103}$$

where $\mathbf{f}^*_{\boldsymbol{\beta}'}\left(\hat{\boldsymbol{\beta}}\right)$ is the $N \times K$ matrix of gradients (see section 1.5.1.4) and $\mathbf{f}^*_{\boldsymbol{\beta}',i}\left(\hat{\boldsymbol{\beta}}\right)$ is the $1 \times K$ row vector of gradients for observation *i*. The intended effect of weighting the residuals is to make the variance of e_i^* (the standardized weighted error) the same for each observation. The user-supplied weights, if appropriately specified, should correct for structural heteroscedasticity in the distribution of model residuals and the correction for leverage removes small sample effects on the accuracy with which specific errors are estimated. [Note that the standardized form of the residual in (1.102) is also known as an internally studentized residual, as distinct from an externally studentized residual, which is based on a mean squared error in the denominator that omits the i^{th} observation.] The quantiles of the standard normal distribution are generated from *N* values of Cunnanne (1978) adjusted ranks (Helsel and Hirsch, 1992, p. 27). The *N* quantiles take the form

$$q_i = \Phi^{-1}\left(\frac{i-a}{N+1-2a}\right), \tag{1.104}$$

where $\Phi^{-1}\left(\cdot\right)$ is the inverse of the standard normal cumulative probability distribution, *i* is the rank of e_i^*, and *a* is the rank offset. For the Cunnane adjustment, *a* is set to 0.4.

The empirical distribution will plot along the reference line in figure 1.24d if the standardized weighted residuals are normally distributed. Conversely, if the empirical distribution plot is a convex shape (that is, the steepness of the graph is greater than the one-to-one line for the lower portion and less than one-to-one for the upper portion), then the residuals are skewed to the left (negative skew), implying there are more small residuals

and fewer large residuals compared to a normal distribution. If the empirical distribution is a concave shape (that is, the steepness of the graph is less than the one-to-one line for the lower portion and greater than one-to-one for the upper portion), then the residuals are skewed to the right (positive skew), implying there are more large residuals and fewer small residuals compared to a normal distribution. If the empirical distribution generally plots along the one-to-one line in the middle section of the graph but the tails of the figure show points consistently above or below the line, then there is more or less probability in the tails as compared to a normal distribution. For example, a group of points falling below the one-to-one line at the low end of the graph is indicative of an empirical distribution having a fatter left tail than the normal distribution. A group of points lying above the one-to-one line on the upper end of the graph is indicative of an empirical distribution that is fatter than the normal distribution in the right tail.

Because departure of the residuals distribution from normality does not necessarily invalidate the SPARROW model results, departures of the empirical distribution from the one-to-one line is not necessarily of concern. Failure to meet the three assumptions of the nonlinear least squares methodology (that residuals are mutually independent, identically distributed, and independent of the predictor variables) is, however, sometimes associated with deviations from the one-to-one line in the normal probability plot. For example, heteroscedasticity of the residuals (failure of the second assumption) causes the tails of the empirical distribution to be fatter than the normal distribution, which is expressed on the probability plot as points at the low end of the probability plot lying below the one-to-one line and points at the high end lying above the one-to-one line (as is the case for the residuals of the example nitrogen model shown in figure 1.24d). It is stressed, however, that heteroscedasticity represents a problem for model estimation and interpretation only if the heteroscedasticity is caused by failure of the third assumption, that is, if the residuals are related to the predictors (i.e., gradients). This particular cause of heteroscedasticity can be detected by interpreting the graph of predicted and observed flux in figure 1.24a.

The *probability plot correlation coefficient* provides a measure of the linear correlation between the ordered, standardized weighted residuals (e_i^*), obtained from the estimated parametric model, and the quantiles of the standard normal distribution (q_i). A value of the correlation coefficient near one is evidence that the residuals are from a normal distribution, whereas a value below 0.98 is generally indicative of non-normal residuals. A table of critical values for the normal probability plot correlation coefficient is given in Vogel (1986).

A formal test of the normality assumption is provided by the *Shapiro-Wilk's test statistic*. The Shapiro-Wilks statistic takes the form

$$W = \mathbf{v}'\mathbf{x} / \mathbf{x}'\mathbf{M}\mathbf{x}\,, \tag{1.105}$$

where **x** is a vector of the weighted residuals $\hat{e}_i\sqrt{w_i}$ ordered from low to high; **M** is the idempotent matrix calculated as $\mathbf{I} - \mathbf{i}\left(\mathbf{i}'\mathbf{i}\right)\mathbf{i}'$, where **i** is a vector of ones; and **v** is a vector of factors representing the normalized values of the order statistics derived from a standard normal distribution

$$\mathbf{v} = \mathbf{V}_\mathbf{m}^{-1}\mathbf{m} \Big/ \sqrt{\mathbf{m}'\mathbf{V}_\mathbf{m}^{-2}\mathbf{m}}\,, \tag{1.106}$$

where **m** is the vector of order statistics from a standard normal distribution and $\mathbf{V}_\mathbf{m}$ is the covariance matrix of **m**. The *W* statistic is essentially the squared value of the correlation coefficient between the residuals and the expected values of the normal order statistics. Because **v** is approximately proportional to the normal scores, *W* is a measure of the straightness of the normal probability plot. Probability values for evaluating the statistical significance of *W* are numerically estimated in SPARROW using an algorithm by Royston (1982). The standard SPARROW output includes the normal distribution probability plot correlation coefficient, the Shapiro-Wilks normality test statistic, and the probability value of the Shapiro-Wilks test statistic.

1.5.5.2 Spatial biases

An important additional assessment concerns the spatial distribution of the prediction errors to determine if the model systematically under- or over-predicts in certain regions of the modeled basin. Evidence of a regional bias in the prediction errors suggests that the errors are geographically correlated and may indicate a misspecification of the model. In this case, the prediction errors are likely to be associated with some underlying property of the watershed that has important large-scale effects on stream contaminant flux, but is not accounted for in the model.

One example of a spatial bias in model prediction errors was indicated in the North Carolina coastal SPARROW total nitrogen model (McMahon and others, 2003). A map of the prediction errors for the stream monitoring sites (see figure 1.25) indicated that the model over-predicted stream nitrogen flux in the headwater, Piedmont portions of the watersheds and generally under-predicted in the lowland, coastal areas. One possible explanation is that physiographic differences in the land-to-water factors related to soil properties may not be properly accounted for in the model. Additionally, the types and intensity of cultivation (e.g., fertilizer use) is generally greater in the coastal plain. The effects of these practices on stream nitrogen flux are unlikely to be accurately reflected in a model in which agricultural land area is used to predict agricultural nitrogen sources (McMahon and others, 2003).

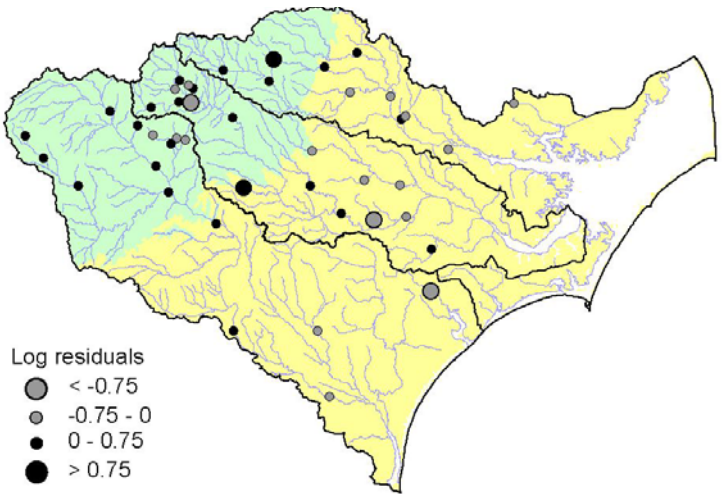

Figure 1.25. Map of prediction errors for the SPARROW total nitrogen model plotted for the stream monitoring stations in the North Carolina coastal watersheds. The prediction errors are expressed as log values of the residuals, computed as the difference between the predicted and actual log of flux. [From McMahon and others (2003).]

The SPARROW software contains a mapping routine that allows users to plot standardized model errors at station locations (see figure 1.26) and visually determine whether regional patterns are present in the model residuals. A standardized residual has unit variance, making it interpretable in absolute terms (see equation (1.102)). The mapping of standardized residuals can be helpful in evaluating whether the model provides similar predictive capability over different regions of the modeled drainage basin. In the example in figure 1.26, there is an indication that the model specification slightly over-predicts stream nitrogen flux in the Ohio Valley and Upper Mississippi, as evidenced by the large number of green triangles that are associated with small (0 to 1.5), positive residuals. Regions of slight under-prediction are evident in the Columbia and Upper Missouri Rivers as well as along the southeastern Atlantic coast. Areas of large over- (less than -1.5) or under-prediction (greater than 1.5) are not as evident, but include some monitoring stations in the drainages of California and the southern Plains states.

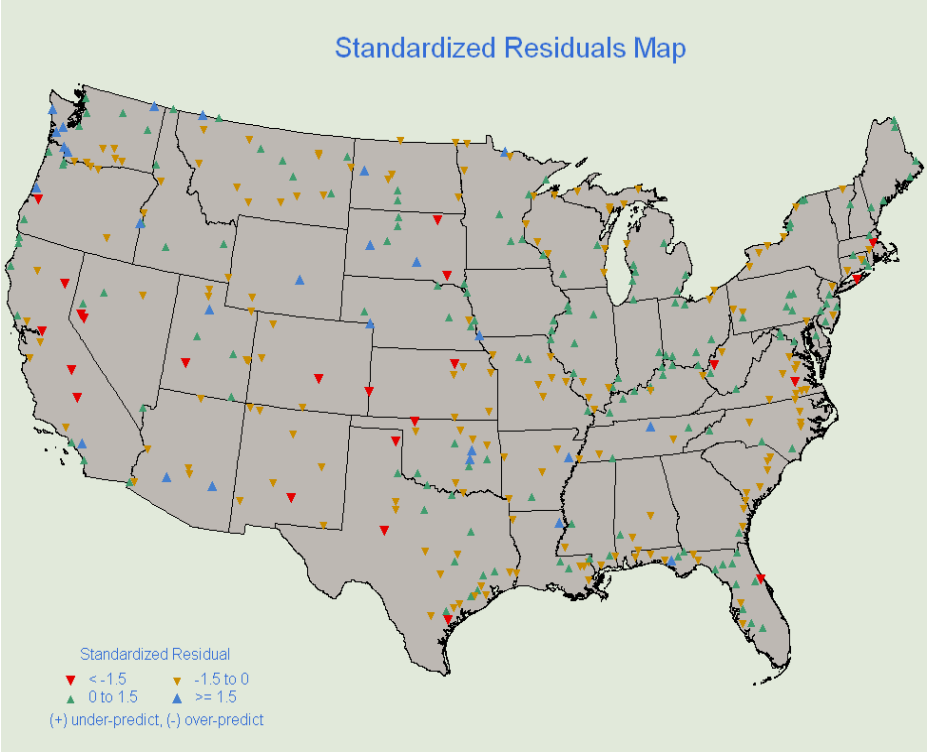

Figure 1.26. Map of prediction errors for the SPARROW national total nitrogen model plotted for the stream monitoring stations, as displayed by the SAS-based SPARROW modeling software. Residuals are expressed in standardized form.

1.5.5.3 Statistical outliers

A final consideration of the validity of the model fit is whether there is evidence of *outlier predictions*—i.e., predictions that deviate considerably from the overall distribution of the observations. The graphs of predicted versus observed flux, predicted versus residual flux, and predicted versus residual yield (figures 1.24a-c) are useful for evaluating the presence of unusual or "outlier" predictions of flux from the model. Outliers may be indicative of model misspecification (i.e., failure to include a variable that affects stream flux) or may be caused by errors in the station flux estimates or explanatory factor data. Past experiences with SPARROW models show that many of the outlier observations are likely to be caused by problems with the data rather than problems with model specification, and correction of these data problems commonly leads to an improved model fit. One illustration of the former problem is apparent from the regional SPARROW model applied to the Waikato River Basin in New Zealand, as shown in figure 1.27. Here, an outlier is evident in the predicted versus observed yield plot (figure 1.27b)—the model greatly underpredicts the amounts of nitrogen loading measured at the watershed outlet. This underprediction may be explained by the predominance of horticulture or market gardening operations within this particular watershed, a source that was not explicitly included in the land-use based source model because of the unavailability of data on fertilizer use. Agricultural sources of nitrogen in the Waikato Basin, and those reflected in the SPARROW model of this basin, originate primarily from livestock wastes, especially sheep and dairy cows. Row crop agriculture is limited entirely to the one watershed for which the model provided a poor fit (Alexander, Elliott, and others, 2002); the inclusion of fertilizer data in the model would likely provide a solution to the problem.

(a) (b)

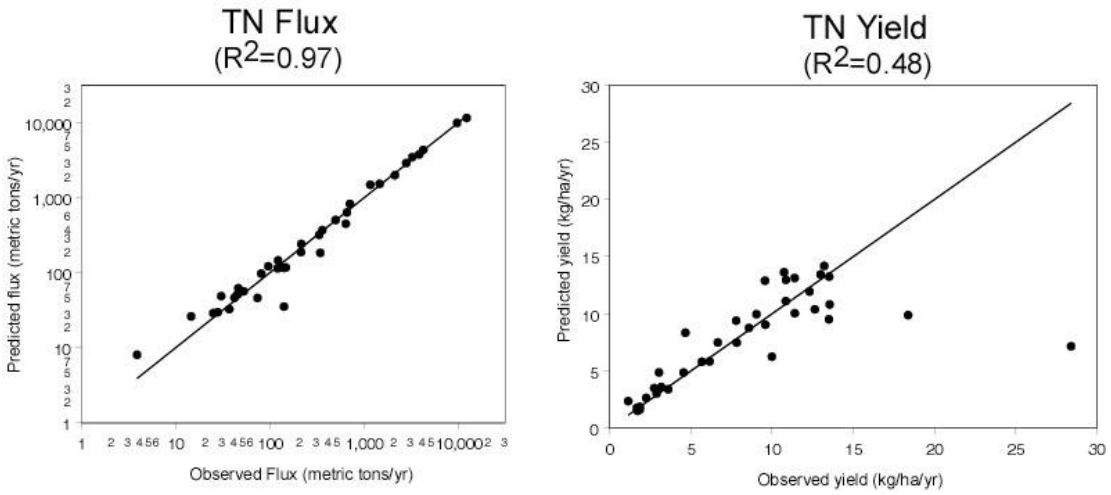

Figure 1.27. Plots of predicted and observed total nitrogen (TN) (a) flux and (b) yield for the New Zealand Waikato River Basin SPARROW model. [From Alexander, Elliott, and others (2002).]

Outlier observations may also exist for one or more of the explanatory variables and may exert considerable influence or leverage on the model fit. *Leverage* statistics provide the most efficient method to identify these observations. The calculation of the leverage statistic is given in equation (1.103) above and is also discussed earlier in section 1.5.1.1. Observations with a high degree of leverage can be determined for a given model based on the number of estimated parameters, K, and the number of observations in the model, N. Leverage statistics that exceed $3K/N$ are those that may exert considerable influence on the model fit; these observations should be examined to determine whether any data errors might explain their values and to improve understanding of the sensitivity of the model fit to these observations.

1.5.6 Measures of model performance and fit

A number of standard summary statistics are reported by the SPARROW software to describe the absolute and relative performance of the models in explaining variablity in the response variable. The Sum of Squared Errors (SSE) statistic is the squared value of the estimated residual, $\hat{e}$, times its weight, w, and summed over all N monitored reaches

$$SSE = \sum_{i \in I} w_i \hat{e}_i^2 \; . \tag{1.107}$$

The Mean Squared Error (MSE) is equal to the SSE divided by ($N - K$), the number of degrees of freedom for the error (DF Error). The "DF Error" statistic pertains to the difference between the number of observations and the number of degrees of freedom used in model estimation ($N - K$). This statistic represents the number of degrees of freedom used to estimate the residuals of the model.

The root mean squared error (Root MSE or RMSE) is the square root of the mean squared error. A rule-of-thumb for interpreting its value in relation to percent error is as follows. Let F and $\hat{F}$ denote the actual and predicted flux at a given location, in real space, where the predicted flux is assumed to include the adjustment for retransformation bias. If the residual in the model is assumed to be normally distributed with mean zero and RMSE σ, then the percent error in the prediction is given by

$$100\frac{\left(F-\hat{F}\right)}{\hat{F}}=100\left(e^{\varepsilon-\sigma^2/2}-1\right). \tag{1.108}$$

Consider the percent error in flux associated with a one standard deviation error, $PE(\sigma)$. We have,

$$PE(\sigma)=100\left(e^{\sigma-\sigma^2/2}-1\right)\approx 100\sigma\,, \tag{1.109}$$

where the approximation corresponds to a second-order Taylor expansion of the exponential term with respect to σ, which is approximately valid for all σ between 0 and 0.6. Thus, 100 times the RMSE approximately equals the percent error in the flux estimate, for any given reach, associated with a one standard deviation error. For σ greater than 0.6, the approximation is less precise and results in an overestimate of the percent error.

The remaining fit statistics reported by the SPARROW software are *R*-square, adjusted *R*-square, and *R*-square of the logarithm of contaminant yield. The *R*-square statistic (denoted R^2) is given by (Judge and others, 1985, p. 30)

$$R^2=1-\frac{\sum_{i=1}^{N}e_i^2}{\sum_{i=1}^{N}\left(f_i^*-\overline{f}^*\right)^2}, \tag{1.110}$$

where e_i is the model residual, in log space, for the *i*th observation, f_i^* is the logarithm of measured flux for the i^{th} observation, and $\overline{f}^*$ is the average of the f_i^* over all *N* observations. Adjusted *R*-square applies a degrees of freedom adjustment to the *R*-square statistic (Judge and others, 1985, p. 30)

$$\text{Adjusted } R^2=1-\left(\frac{N-1}{N-K}\right)\left(1-R^2\right). \tag{1.111}$$

The *R*-square and Adjusted *R*-Square statistics for a SPARROW model tend to be large (greater than 0.6). Large values for these statistics result partly from the fact that much of the variation in the dependent variable is associated with the size (drainage area) of the basin upstream from the monitored reach, and drainage area in turn is typically highly correlated with contaminant source variables. A high *R*-square, therefore, does not necessarily indicate the strength of the model within a smaller basin. Goodness of model fit for small basins might be better described by *R*-square of the logarithm of contaminant yield, denoted R^2_{Yield}. This statistic is defined as

$$R^2_{Yield}=1-\frac{\sum_{i=1}^{N}e_i^2}{\sum_{i=1}^{N}\left(\left(f_i^*-\overline{f}^*\right)-\left(d_i-\overline{d}\right)\right)^2}, \tag{1.112}$$

where d_i is the log of drainage area for the ith observation and $\overline{d}$ is the mean of d_i over all *N* observations. Because the log of drainage area is highly positively correlated with the log of flux, the denominator of the ratio term will be smaller than the corresponding term in the R-square equation, implying a smaller value for yield *R*-square.