Chapter 8

Snowmelt runoff problem

8.1 Introduction

Design of water resource systems generally relies on historical data such as streamflow records or records of extreme floods. Probability and statistics have been used extensively to characterize the large uncertainties in such hydrologic phenomena. Valencia and Schaake (1972) and Reese and Krzysztofowicz (1989) discuss a number of stochastic models in hydrology.

The purpose of disaggregation models in particular is to estimate a quantity of interest (e.g., total precipitation) over short periods of time (e.g., quarterly or monthly) given the total for a longer time period (e.g., total annual precipitation). Such estimates are often useful in forecasting. For example, if w_i is the total precipitation during month *i*, then $\mathbf{w} = (w_1,\ldots,w_{n+1})$ is a vector of monthly precipitation amounts. A stochastic disaggregation model for this vector amounts to a multivariate conditional probability density function of the form $f(w_1,\ldots,w_{n+1} \mid w)$, where

$$w_1+\ldots+w_{n+1} = w, \qquad 0 < w_i < \infty, \quad i = 1,\ldots,n+1. \tag{8.1}$$

Obtaining a conditional density $f(w_1,\ldots,w_{n+1} \mid w)$ that yields the desired conditional means, variances, and covariances of the w_i is a difficult and as yet unsolved problem (Krzysztofowicz and Reese, 1991). Stedinger, Pei, and Cohn (1985) and Krzysztofowicz and Reese (1991a) discuss this problem in more detail.

Krzysztofowicz and Reese (1991) developed a new approach for addressing the disaggregation problem. Their approach concentrates on developing a density function for the seasonal pattern $(w_1/w,\ldots,w_{n+1}/w)$, rather than for the unnormalized vector $(w_1,\ldots,w_{n+1})$. In particular, they use stochastic bifurcation processes (section 3.3) to model the disaggregation process, and attempt to select particular bifurcations that match the observed conditional dependence structures where possible. Krzysztofowicz and Reese noted that the resulting densities satisfy the required balance equation (8.1), and allow a wide variety of distribution shapes.

Reese and Krzysztofowicz (1991) applied the adaptive Dirichlet distribution to data on snowmelt runoff. In particular, they analyzed monthly runoff data for 14 western United States rivers. They were interested particularly in the fraction of total annual runoff occurring in each month. In describing snowmelt runoff patterns $(w_1/w,\ldots,w_{n+1}/w)$, they found that the correlations among these fractions were a reasonable way to characterize the seasonal patterns at particular rivers. Reese and Krzysztofowicz hypothesized that the vector of fractions at a given river was generated from an adaptive Dirichlet distribution, and tested that assumption. They found the model valid for about one-half to two-thirds of the rivers tested, and therefore concluded that the adaptive Dirichlet distribution is useful, but not sufficiently general, because it is restricted by two structural assumptions. First, the composition is assumed to be stochastically independent of the size of the basis (i.e., the total seasonal runoff); second, the ratios of the fractions (defined in chapter 3) are assumed to be mutually independent. Some rivers violated one or both of these assumptions. Thus, Krzysztofowicz and Reese pointed out the need to construct multivariate distributions on the

simplex that can accommodate the remaining rivers whose seasonal runoff patterns exhibit dependence structures cannot be represented using their model.

We propose using adaptive Dirichlet distributions with dependent ratios (chapter 5) to analyze the same problem. We will apply our model to all rivers considered by Reese and Krzysztofowicz, even the ones where their model performed well, to see whether our model performs "better" in some of those cases. In fact, our model has more degrees of freedom than theirs, so we would expect our model to perform at least somewhat better than theirs in general; the question of how much better will be addressed in the next sections. In particular, we will compare the means, variances, and correlation matrices obtained from their model and ours with the empirical data as a reference point. Through this comparison, we hope to find that our model preserves the empirical correlation sign structures for most of the 14 rivers, even those where the model of Reese and Krzysztofowicz did not preserve the signs. As proposed by Krzysztofowicz and Reese, one possible approach to measure the "closeness" of the model correlations to the empirical correlations is to compute the Euclidean distance between the correlation matrix of our model (or the adaptive Dirichlet model without dependent ratios) and the empirical correlation matrix. This Euclidean distance is given by

$$\Delta = \left\{ \frac{2}{n(n+1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n+1} [\mathrm{cor}(w_i / w, w_j / w) - \hat{\mathrm{cor}}(w_i / w, w_j / w)]^2 \right\}^{\frac{1}{2}} \qquad (8.2)$$

where $\mathrm{cor}(w_i / w, w_j / w)$ is the empirical correlation between w_i / w and w_j / w, and $\hat{\mathrm{cor}}(w_i / w, w_j / w)$ is the corresponding correlation resulting from the adaptive Dirichlet distribution with (or without) dependent ratios.

Our work in this chapter will concentrate on the seasonal patterns $(w_1 / w,\ldots, w_{n+1} / w)$. We hypothesize that this vector of fractions was generated from an adaptive Dirichlet distribution in which two of the ratios are dependent, as defined in chapter 5. With the exception of this assumption, the development of the multivariate conditional densities of $(w_1,\ldots, w_{n+1} \mid w)$ will follow the same steps as used by Krzysztofowicz and Reese.

8.2 Stochastic disaggregation methodology

Let w denote the total seasonal runoff at a particular location. Let n+1 be the number of months in a season, and let $\mathbf{w} = (w_1,\ldots, w_{n+1})$ be a vector of monthly runoff amounts such that $\sum_{i=1}^{n+1} w_i = w$. Dividing by w yields a vector $\mathbf{x} = (x_1,\ldots, x_{n+1})$ describing the seasonal pattern of runoff, where

$$x_i = \frac{w_i}{w}, \qquad i = 1,\ldots, n+1, \tag{8.3}$$

The set of all possible vectors **x** is an n-dimensional simplex as defined in section 2.3. In chapter 5, we showed how to construct a joint density of the fractions vector **x** under the assumption that some of the ratios of these fractions are correlated.

As Krzysztofowicz and Reese point out, for each value of n (n=2,3,….) there exists a set of possible bifurcation topologies. Furthermore, each topology is associated with a set of possible permutations of fractions. Each combination of a basic topology with a permutation of fractions generates a family of multivariate densities on the n-dimensional simplex. By considering this rich family of structural densities, Krzysztofowicz and Reese were able to

identify density functions with correlation structures “close” to the empirical correlation structures evidenced by the data for most rivers.

In comparing the results of the two models, our work will consider the data, topologies, and permutations that were chosen by Reese and Krzysztofowicz for each river, but will allow the two most strongly correlated ratios to be dependent. (We could, of course, allow more than two ratios to be dependent, but for simplicity in the multivariate density g of the fractions we will restrict ourselves to just two dependent ratios.)

Reese and Krzysztofowicz studied 14 rivers. Due to short runoff seasons, eight of these rivers have only four fractions; i.e., $\mathbf{x} = (x_1, x_2, x_3, x_4)$. They used topology 2-2 (given by figure 8.1) for four of these eight, and topology 1-3 (i.e., the Connor-Mosimann distribution defined by figure 8.2) for the other four. For the remaining fourteen rivers, they used the topologies shown in Table 2. These particular topologies were selected in order to match the model correlation structure to the empirical structure evidenced by the data.

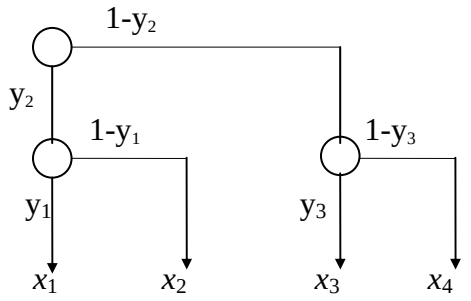


Figure 8.1 Double-cascaded bifurcation topology of four fractions (taken from Krzysztofowicz and Reese, 1991).

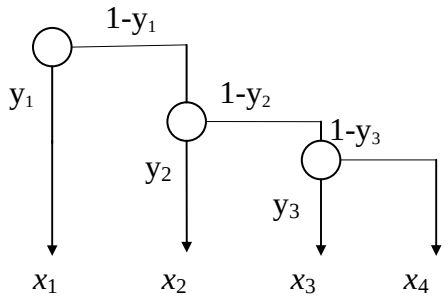


Figure 8.2 Cascaded bifurcation topology of four fractions (taken from Krzysztofowicz and Reese, 1991).

8.3 Expressions for moments and density

Expressions for the means, variances, and covariances of the fractions x_i are derived from the mixed moments of the ratios y_i, denoted

$$\mu_{\boldsymbol{m}} = \mathrm{E}\left(\prod_{i=1}^{n} y_i^{m_i}\right), \tag{8.4}$$

where $\boldsymbol{m} = (m_1,...,m_n)$ and $m_i = 0,1,2,3,\cdots$ for all $i = 1,...,n.$ When these mixed moments of ratios are replaced by their empirical estimates $\hat{\mu}_{\boldsymbol{m}}$, the resulting expressions estimate the moments of the fractions x_i. These estimators are distribution-free, since they do not depend upon the functional form of the density of the ratios.

For example, for the double-cascaded topology shown in figure 8.1, we have

$x_1 = y_1 y_2$,

$x_2 = (1 - y_1) y_2$,

$x_3 = (1 - y_2) y_3$, and

$x_4 = (1 - y_2)(1 - y_3)$. (8.5)

This system of four equations has a unique inverse given by

$$y_1 = \frac{x_1}{x_1 + x_2},$$

$$y_2 = x_1 + x_2,$$

$$y_3 = \frac{x_3}{x_3 + x_4}, \text{ and}$$

$$x_1 + x_2 + x_3 + x_4 = 1. \tag{8.6}$$

The Jacobian of this transformation is $\left(\frac{1}{x_1 + x_2}\right)\left(\frac{1}{x_3 + x_4}\right)$. Note that there are three ratios (i.e., y_1, y_2, and y_3) defined by the four fractions x_i.

Mean Values. The mean values of the x_i are given by

$$\begin{aligned} &\mathrm{E}(x_1) = \mu_{(1,1,0)}, \\ &\mathrm{E}(x_2) = \mu_{(0,1,0)} - \mu_{(1,1,0)}, \\ &\mathrm{E}(x_3) = \mu_{(0,0,1)} - \mu_{(0,1,1)}, \text{ and} \\ &\mathrm{E}(x_4) = 1 - \mu_{(0,0,1)} - \mu_{(0,1,0)} + \mu_{(0,1,1)}. \end{aligned} \tag{8.7}$$

Variances. The variances of the x_i are given by

$$\begin{aligned} &\mathrm{var}(x_1) = \mu_{(2,2,0)} - (\mu_{(1,1,0)})^2, \\ &\mathrm{var}(x_2) = \mu_{(0,2,0)} - 2\mu_{(1,2,0)} + \mu_{(2,2,0)} - (\mu_{(0,1,0)} - \mu_{(1,1,0)})^2, \\ &\mathrm{var}(x_3) = \mu_{(0,0,2)} - 2\mu_{(0,1,2)} + \mu_{(0,2,2)} - (\mu_{(0,0,1)} - \mu_{(0,1,1)})^2, \text{ and} \\ &\mathrm{var}(x_4) = 1 - 2\mu_{(0,1,0)} - 2\mu_{(0,0,1)} + \mu_{(0,2,0)} + \mu_{(0,0,2)} + 4\mu_{(0,1,1)} - \\ &\quad 2\mu_{(0,1,2)} - 2\mu_{(0,1,1)} + \mu_{(0,2,2)} - (1 - \mu_{(0,1,0)} - \mu_{(0,0,1)} + \mu_{(0,1,1)})^2 \end{aligned} \tag{8.8}$$

Covariances. The covariances of x_i and x_j $(i \neq j)$ are given by

$$\begin{aligned}
\operatorname{cov}(x_1, x_2) &= \mu_{(1,2,0)} - \mu_{(2,2,0)} - \mu_{(1,1,0)}\mu_{(0,1,0)} + (\mu_{(1,1,0)})^2, \\
\operatorname{cov}(x_1, x_3) &= \mu_{(1,1,1)} - \mu_{(1,2,1)} - \mu_{(1,1,0)}\mu_{(0,0,1)} + \mu_{(1,1,0)}\mu_{(0,1,1)}, \\
\operatorname{cov}(x_1, x_4) &= -[\mu_{(1,2,0)} - \mu_{(1,1,0)}\mu_{(0,1,0)} + \mu_{(1,1,1)} - \mu_{(1,2,1)} - \\
&\qquad \mu_{(1,1,0)}\mu_{(0,0,1)} + \mu_{(1,1,0)}\mu_{(0,1,1)}], \\
\operatorname{cov}(x_2, x_3) &= \mu_{(0,1,1)} - \mu_{(0,2,1)} - \mu_{(1,1,1)} + \mu_{(1,2,1)} - \\
&\qquad \mu_{(0,1,0)}\mu_{(0,0,1)} + \mu_{(0,1,0)}\mu_{(0,1,1)} + \mu_{(1,1,0)}\mu_{(0,0,1)} - \mu_{(1,1,0)}\mu_{(0,1,1)}, \\
\operatorname{cov}(x_2, x_4) &= -[\operatorname{var}(x_2) + \operatorname{cov}(x_1, x_2) + \operatorname{cov}(x_2, x_3)], \text{ and} \\
\operatorname{cov}(x_3, x_4) &= -[\operatorname{var}(x_3) + \operatorname{cov}(x_1, x_3) + \operatorname{cov}(x_2, x_3)].
\end{aligned} \tag{8.9}$$

Equations for other topologies could be derived in the same manner.

Remark 8.1

1) None of the above expressions depend on the functional form of the density function of the ratios. Also note that we have not yet imposed any restriction on the ratios y_i. If all ratios are independent, then we have $\mu_{(1,1,1)} = \mu_{(1,0,0)}\mu_{(0,1,0)}\mu_{(0,0,1)}$. Similarly, if y_2 is independent of (y_1, y_3), but y_1 and y_3 are correlated, then we have $\mu_{(1,1,1)} = \mu_{(0,1,0)}\mu_{(1,0,1)}$ and $\mu_{(2,2,0)} = \mu_{(0,2,0)}\mu_{(2,0,0)}$, and so on.
2) If all three ratios are independent, then our model is simply the adaptive Dirichlet with independent ratios. If two of the ratios are correlated, then the resulting model is the adaptive Dirichlet with dependent ratios.

3) Equations (8.8) and (8.9) above determine the correlation matrix of the x_i. This means that neither the correlation matrix nor the mean values of the x_i depend on the functional form of $g(\mathbf{x})$.

As an example, if y_1 and y_3 are assumed to be correlated and both are independent of y_2, $g(\mathbf{x})$ has the general form

$$g(\mathbf{x}) = g_2(y_2)g_{13}(y_1, y_3)J(\mathbf{y} \to \mathbf{x}), \tag{8.10}$$

where g_{13} could for example be a copula density function as given in equation (5.5), and $J(\mathbf{y} \to \mathbf{x})$ is the Jacobian of the transformation from $\mathbf{y} \to \mathbf{x}$.

Letting $g_{13}(y_1,y_3)$ be one of Frank's family of distributions, then we have

$$g_{13}(y_1, y_3) = \frac{(\delta - 1)\log(\delta)\delta^{G_1(y_1)+G_3(y_3)}}{\{(\delta - 1) + (\delta^{G_1(y_1)} - 1)(\delta^{G_3(y_3)} - 1)\}^2} g_1(y_1)g_3(y_3), \tag{8.11}$$

where δ is a parameter expressing the degree of correlation between y_1 and y_3. Also, if we let $y_i \sim Be(\alpha_i, \beta_i)$ for $i = 1,2,3$, then by equation (8.10) the joint density of $\mathbf{x}$ is given by

$$g(\mathbf{x}) = \left[\prod_{i=1}^{3} \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)}\right] \frac{(\delta - 1)\log(\delta)\delta^{G_1(\frac{x_1}{x_1+x_2})+G_3(\frac{x_3}{x_3+x_4})}}{\{(\delta - 1) + (\delta^{G_1(\frac{x_1}{x_1+x_2})} - 1)(\delta^{G_3(\frac{x_3}{x_3+x_4})} - 1)\}^2}$$

$$x_1^{\alpha_1-1}x_2^{\beta_1-1}x_3^{\alpha_3-1}x_4^{\beta_3-1}(x_1 + x_2)^{\alpha_2-\alpha_1-\beta_1}(x_3 + x_4)^{\beta_2-\alpha_3-\beta_3}, \tag{8.12}$$

where the marginals $G_1(y_1)$ and $G_3(y_3)$ are the cumulative distribution functions (cdf's) of the beta marginal distributions for y_1 and y_3. (Note that we use Frank's copula because it has the ability to describe strong positive and negative correlations, as discussed in chapter 5.)

The parameters (α_i,β_i) of the beta distributions for the y_i (i=1,…,n) can be estimated using the equations

$$\alpha_i = \frac{\hat{\mu}_i[(1-\hat{\mu}_i)\hat{\mu}_i - \hat{\sigma}_i^2]}{\hat{\sigma}_i^2}, \text{ and}$$

$$\beta_i = \frac{(1-\hat{\mu}_i)[(1-\hat{\mu}_i)\hat{\mu}_i - \hat{\sigma}_i^2]}{\hat{\sigma}_i^2}, \tag{8.13}$$

where $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ are empirical estimates of the mean and variance, respectively, of the ratio y_i (Krzysztofowicz and Reese, 1991).

The strength-of-dependence parameter δ appearing in equation (8.12) can be estimated using trial and error, as follows: Let $\gamma = \text{cov}(y_1, y_3)$ be estimated from the data. Then we wish to choose the density $g_{13}(y_1, y_3)$ to satisfy

$$\gamma = \text{E}(y_1 y_3) - \text{E}(y_1)\text{E}(y_3), \tag{8.14}$$

where $\text{E}(y_1 y_3)$ is the mean of $y_1 y_3$ computed according to the joint density function $g_{13}(y_1, y_3)$. The R.H.S. of this equation is a function of δ, and the L.H.S. is computed from the data. By trying different values of δ, it is possible to match the L.H.S. to any desired degree of precision, because Frank's copula can model any correlation in the interval [-1,1].

8.4 Strategy of analysis

As we said earlier, in comparing the results of the two models, our work will consider the same topologies, permutations of fractions, and data that were used by Reese and Krzysztofowicz for each river (as displayed in table 2 and appendix 1), but will allow the two most strongly correlated ratios to be dependent. For the rivers with short runoff seasons, we also consider models in which the two ratios assumed to be dependent do not have the strongest correlation; this provides a basis for comparison, especially in cases where the

correlation sign structure is not preserved by the model with the strongest correlation. We now describe the procedure we use to analyze the data:

1. First, we order the monthly runoffs **w** to match the permutation of $x_1,\ldots,x_{n+1}$ chosen by Reese and Krzysztofowicz for each river.
2. We next transform each vector of monthly runoffs **w** into a vector of monthly fractions **x**.
3. Considering the topology chosen by Reese and Krzysztofowicz for the river that we are dealing with, we then define the ratios y_i of fractions associated with this topology, and transform each vector of fractions **x** into a vector of ratios **y**. From the resulting set of these annual vectors, we estimate the empirical correlation matrix of ratios, and determine the two most strongly correlated ratios for each river.
4. All empirical correlations of the y_i are tested to see whether they are statistically significant; the null hypothesis is that a given correlation coefficient ρ is equal to zero. The two-tailed test uses the Student-Fisher t distribution with *k*-2 degrees of freedom:

$$t = |\hat{\rho}|\left(\frac{k-2}{1-\hat{\rho}^2}\right), \tag{8.15}$$

where $\hat{\rho}$ is the sample correlation coefficient and *k* is the sample size (i.e, the number of years for which runoff data is available). The hypothesis that $\rho = 0$ can be rejected at the significance level *p* if $t > t(p, k-2)$, a tabulated critical value (Hoshmand, 1988). We test this null hypothesis at level of significance .05; .01; and .001.

5. The t-test described in step 4 above is also applied to the correlation coefficients of any fractions x_i affected by sign reversals.
6. From the data on the observed vectors of ratios **y**, we estimate the mixed moments needed to compute the means, variances, and covariances of the fractions x_i according to equations (8.7-8.9) (or the corresponding set of equations for the selected topology). These moments are used to compute the correlations of the fractions for both the case with independent ratios and the case with dependent ratios.
7. Finally, for each river, we compute the Euclidean distance, as given by (8.2), between the correlation matrix achieved by our model (or the adaptive Dirichlet distribution with independent ratios, respectively) and the empirical correlation matrix.

8.5 Analysis and results

Below we list the questions to be answered by our analysis of each river:

1. Do the empirical means match the empirical means given by Reese and Krzysztofowicz (1989)?
2. What are the two most strongly correlated ratios?
3. Are the signs of the correlations preserved under the independent ratios model?
4. Are the signs of the correlations preserved under the dependent ratios model(s)?
5. Are the correlations with unpreserved signs statistically significant under the t-test described above?
6. Is the Euclidean distance between the empirical correlations and the dependent ratios model(s) less than the distance between the empirical correlations and the independent ratios model?

Boise (Table A-1)

1. The average discrepancy between our empirical means and those of Reese and Krzysztofowicz's empirical means is 0.2%. The reasons for this discrepancy are unclear.
2. y_1 and y_3 are the two most strongly correlated ratios.
3. The signs of the correlations are preserved under the independent ratios model.
4. The signs of the correlations are also preserved under all three dependent ratios models considered for this river.
5. Not applicable.
6. The Euclidean distance between the empirical correlations and the dependent ratios model is at least as small as the distance between the empirical correlations and the independent ratios model, for all three dependent ratios models considered for this river.

Note: As illustrated in table 3, the Euclidean distance when y_1 and y_2 are assumed to be dependent is actually smaller than the Euclidean distance when y_1 and y_3 are assumed to be dependent, even though y_1 and y_3 are more strongly correlated than y_1 and y_2.

Weiser (Table A-2)

1. The empirical means match the corresponding values given by Reese and Krzysztofowicz.
2. y_2 and y_3 are the two most strongly correlated ratios.
3. The signs of the correlations are preserved under the independent ratios model.
4. The signs of the correlations are also preserved under all three dependent ratios models considered for this river.
5. Not applicable.

6. The Euclidean distance between the empirical correlations and the dependent ratios model is less than the distance between the empirical correlations and the independent ratios model, for all three dependent ratios models considered for this river.

Little Truckee (Table A-3)

1. The average disecrepancy between our empirical means and those of Reese and Krzysztofowicz is 1.5%. The reasons for this discrepancy are unclear.
2. y_1 and y_2 are the two most strongly correlated ratios.
3. The signs of the correlations are preserved under the independent ratios model.
4. The signs of the correlations are also preserved under all three dependent ratios models considered for this river.
5. Not applicable.
6. The Euclidean distance between the empirical correlations and the dependent ratios model is less than the distance between the empirical correlations and the independent ratios model, for all three dependent ratios models considered for this river.

Gila (Table A-4)

1. The average discrepancy between our empirical means and those of Reese and Krzysztofowicz's empirical means is 1.0%. The reasons for this discrepancy are unclear.
2. y_1 and y_2 are the two most strongly correlated ratios.
3. The sign of $\text{corr}(x_2, x_4)$ is not preserved under the independent ratios model.
4. The sign of $\text{corr}(x_2, x_4)$ is not preserved under any of the three dependent ratios models considered for this river.

5. corr(x_2, x_4) is not statistically significant at the 0.05 level under the t-test described above.
6. The Euclidean distance between the empirical correlations and the dependent ratios model is less than the distance between the empirical correlations and the independent ratios model, for all three dependent ratios models considered for this river.

Salmon (Table A-5)

1. The empirical means match the corresponding values given by Reese and Krzysztofowicz.
2. y_2 and y_3 are the two most strongly correlated ratios.
3. The signs of the correlations are preserved under the independent ratios model.
4. The signs of the correlations are also preserved under all three dependent ratios models considered for this river.
5. Not applicable.
6. The Euclidean distance between the empirical correlations and the dependent ratios model is less than the distance between the empirical correlations and the independent ratios model, for all three dependent ratios models considered for this river.

Falls (Table A-6)

1. The empirical means match the corresponding values given by Reese and Krzysztofowicz.
2. y_1 and y_2 are the two most strongly correlated ratios.
3. The signs of the correlations are preserved under the independent ratios model.

4. The signs of the correlations are also preserved under all three dependent ratios models considered for this river.
5. Not applicable.
6. The Euclidean distance between the empirical correlations and the dependent ratios model is less than the distance between the empirical correlations and the independent ratios model, for all three dependent ratios models considered for this.

Humboldt (Table A-7)

1. The average discrepancy between our empirical means and those of Reese and Krzysztofowicz is 0.5%. The reasons for this discrepancy are unclear.
2. y_1 and y_2 are the two most strongly correlated ratios.
3. The signs of the correlations are preserved under the independent ratios model.
4. The signs of the correlations are also preserved under all three dependent ratios models considered for this river.
5. Not applicable.
6. The Euclidean distance between the empirical correlations and the dependent ratios model is less than the distance between the empirical correlations and the independent ratios model, for all three dependent ratios models considered for this river.

Sevier (Table A-8)

1. The average discrepancy between our empirical means and those of Reese and Krzysztofowicz is 0.06%. The reasons for this discrepancy are unclear.
2. y_1 and y_3 are the two most strongly correlated ratios.

3. The sign of $\text{corr}(x_1, x_2)$ and $\text{corr}(x_1, x_4)$ are not preserved under the independent ratios model.
4. None of the dependent ratios models preserve the sign of $\text{corr}(x_1, x_4)$. The sign of $\text{corr}(x_1, x_2)$ is preserved under the dependent ratios model where y_1 and y_2 are assumed to be dependent, but not under the other models.
5. Neither $\text{corr}(x_1, x_2)$ nor $\text{corr}(x_1, x_4)$ is statistically significant at the 0.05 level under the t-test described above.
6. The Euclidean distance between the empirical correlations and the dependent ratios models is at least as small as the distance between the empirical correlations and the independent ratios model, except for the dependent ratio model where y_1 and y_2 considered to be dependent.

Note 1: For this river, Reese and Krzysztofowicz (1989) apparently did not recognize that the signs of $\text{corr}(x_1, x_2)$ and $\text{corr}(x_1, x_4)$ were not preserved under their model, although it is clear from the table given on page 111 of their report.

Note 2: The only model under which the sign of $\text{corr}(x_1, x_2)$ is preserved (namely, the model in which y_1 and y_2 are assumed to be dependent) is also the only dependent ratios model that yields a Euclidean distance larger than that for the independent ratios model. Thus, the criteria of sign preservation and minimization of Euclidean distance appear to be at odds for this river.

Verde (Table A-9)

1. The empirical means match the corresponding values given by Reese and Krzysztofowicz.
2. y_1 and y_4 are the two most strongly correlated ratios.
3. The sign of $\text{corr}(x_1, x_4)$ is not preserved under the independent ratios model.
4. The sign of $\text{corr}(x_1, x_4)$ is also not preserved under the dependent ratios model.
5. $\text{corr}(x_1, x_4)$ is not statistically significant at the 0.05 level under the t-test described above.
6. The Euclidean distance between the empirical correlations and the dependent ratios model is less than the distance between the empirical correlations and the independent ratios model.

Salt (Table A-10)

1. The empirical means match the corresponding values given by Reese and Krzysztofowicz.
2. y_2 and y_4 are the two most strongly correlated ratios.
3. The signs of the correlations are preserved under the independent ratios model.
4. The signs of the correlations are also preserved under the dependent ratios model.
5. Not applicable.
6. The Euclidean distance between the empirical correlations and the dependent ratios model is less than the distance between the empirical correlations and the independent ratios model.

Little Colorado (Table A-11)

1. The average discrepancy between our empirical means and those of Reese and Krzysztofowicz is 1.0%. The reasons for this discrepancy are unclear.
2. y_1 and y_3 are the two most strongly correlated ratios.
3. The sign of $\text{corr}(x_2, x_4)$ is not preserved under the independent ratios model.
4. Under the dependent ratios model, the sign of $\text{corr}(x_2, x_4)$ is preserved, but the sign of $\text{corr}(x_2, x_5)$ is no longer preserved.
5. Neither $\text{corr}(x_2, x_4)$ nor $\text{corr}(x_2, x_5)$ is statistically significant at the 0.05 level under the t-test described above.
6. The Euclidean distance between the empirical correlations and the dependent ratios model is less than the distance between the empirical correlations and the independent ratios model.

Yellowstone (Table A-12)

1. The average discrepancy between our empirical means and those of Reese and Krzysztofowicz is 0.6%. The reasons for this discrepancy are unclear.
2. y_4 and y_5 are the two most strongly correlated ratios.
3. The signs of $\text{corr}(x_1, x_4)$ and $\text{corr}(x_4, x_5)$ are not preserved under the independent ratios model.
4. Under the dependent ratios model, the sign of $\text{corr}(x_4, x_5)$ is preserved, but the sign of $\text{corr}(x_1, x_4)$ is still not preserved.
5. Neither $\text{corr}(x_1, x_4)$ nor $\text{corr}(x_4, x_5)$ is statistically significant at the 0.05 level under the t-test described above.

6. The Euclidean distance between the empirical correlations and the dependent ratios model is less than the distance between the empirical correlations and the independent ratios model.

Payette (Table A-13)

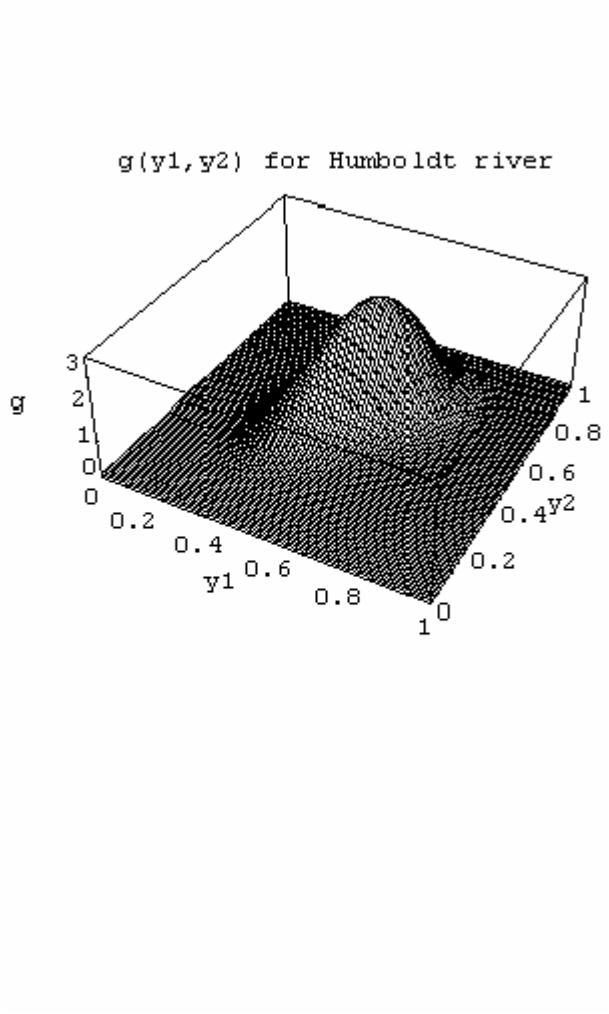
1. The empirical means match the corresponding values given by Reese and Krzysztofowicz.
2. y_2 and y_4 are the two most strongly correlated ratios.
3. The sign of $\text{corr}(x_1, x_6)$ is not preserved under the independent ratios model.
4. The sign of $\text{corr}(x_1, x_6)$ is also not preserved under the dependent ratios model.
5. $\text{corr}(x_1, x_6)$ is not statistically significant at the 0.05 level under the t-test described above.
6. The Euclidean distance between the empirical correlations and the dependent ratios model is less than the distance between the empirical correlations and the independent ratios model.

Rio Grande (Table A-14)

1. The empirical means match the corresponding values given by Reese and Krzysztofowicz.
2. y_2 and y_4 are the two most strongly correlated ratios.
3. The sign of $\text{corr}(x_4, x_6)$ is not preserved under the independent ratios model.
4. The sign of $\text{corr}(x_4, x_6)$ is preserved under the dependent ratios model, but the sign of $\text{corr}(x_4, x_5)$ is no longer preserved.

5. Neither corr(x_4, x_5) nor corr(x_4, x_6) is statistically significant at the 0.05 level under the t-test described above.
6. The Euclidean distance between the empirical correlations and the dependent ratios model is less than the distance between the empirical correlations and the independent ratios model.

Most of the analysis presented in this chapter is distribution-free. However, Reese and Krzysztofowicz hypothesized that each ratio y_i is beta distributed. They used the Kolmogorov-Smirnov test to check this hypothesis for the ratios of all rivers. They stated that "In 49 out of 51 tests, the beta distributions could not be rejected at a significance level of 0.20", and that "the beta distribution seems to be a remarkably versatile model of ratios, having a strong empirical support." Therefore, we now assume that the y_i are beta distributed, and use Frank's copula to illustrate the dependent ratios model for the Humboldt river. Using trial and error, we estimated the strength-of-dependence parameter δ appearing in equation (8.12). With $\delta = 0.01$, the model covariance of (y_1, y_2) is equal to 0.0103, while the empirical covariance of (y_1, y_2) is equal to 0.0109. Therefore, δ=0.01 appears to be a reasonable value for this parameter. The Figure below shows the resulting joint distribution for y_1 and y_2.


g(y1,y2) for Humboldt river
g
3
2
1
0
0
0.2
0.4
y1
0.6
0.8
1
0
0.2
0.4
y2
0.6
0.8
1

8.6 Discussion

To summarize, we applied two models, the adaptive Dirichlet distribution with independent ratios (i.e., Krzysztofowicz and Reese's model) and the adaptive Dirichlet distribution with dependent ratios (i.e., our model) to data on snowmelt runoff. In particular, we reanalyzed monthly runoff data for 14 western United States rivers. We were interested particularly in the fraction of total annual runoff occurring in each month. In describing snowmelt runoff patterns $(w_1/w,...,w_{n+1}/w)$, Reese and Krzysztofowicz (1991) found that the correlations among these fractions were a reasonable way to characterize the seasonal patterns at particular rivers.

Our model seems to performs "better" for all of those rivers, at least as measured by Euclidean distance. In particular, for all rivers, the Euclidean distance between the empirical correlations and those of a suitably chosen dependent ratios model is less than the distance between the empirical correlations and those of independent ratios model. We compared the means, variances, and correlation matrices obtained from the two models. Through this comparison, we found that our model preserves the empirical correlation sign structures for most of the 14 rivers. For the Gila, Sevier, Verde, Little Colorado, Yellowstone, Payette, and Rio Grande Rivers, the sign structure of the correlation matrix is not preserved by the independent ratios model. Unfortunately, our model is still not able to preserve the correlation sign structures of these rivers. However, in some cases (such as the Sevier and Yellowstone rivers), the number of unpreserved signs is reduced by a suitably chosen dependent ratios model. Also, none of the correlations whose signs are not preserved is statistically significant at the 0.05 level.

It is interesting to note that in some cases, the Euclidean distance between the empirical correlations and the dependent ratios model with the most strongly correlated ratios is actually larger than the Euclidean distance for the dependent ratios model with less strongly correlated ratios. This is illustrated in table 3 for the Boise River, and in table 10 for the Sevier River. Therefore, care must be taken in fitting the adaptive Dirichlet distribution with dependent ratios to observed data.

Over all, the adaptive Dirichlet distribution with dependent ratios (i.e., our model) appears to be an improvement over the adaptive Dirichlet distribution with independent ratios for two reasons: 1) for all rivers, the Euclidean distance achievable between the empirical correlations and a suitable dependent ratios model is less than the distance between the empirical correlations and the independent ratios model; and 2) as shown in chapter 5, the adaptive Dirichlet distribution with dependent ratios has a much broader range of correlation sign structures than the independent ratios model. However, this added flexibility appears to be of limited value in this particular case study, since when correlation signs were not preserved under the independent ratios model, they were also not preserved under the corresponding dependent ratios model(s).

Table 1. River gauging stations analyzed by Reese and Krzysztofowicz.

Gauging Station	Runoff Season	Length of Record	Time Period
Boise River near Twin Springs, Idaho	April-July	35	1951-1985
Weiser River near Weiser, Idaho	April-July	35	1951-1985
Little Truckee River above Boca Reservoir, California	April-July	28	1958-1985
Gila River at Calva, Arizona	February-May	25	1961-1985
Salmon River at Salmon, Idaho	April-July	35	1951-1985
Falls River near Squirrel, Idaho	April-July	35	1951-1985
Humboldt River at Palisade, Nevada	April-July	28	1958-1987
Sevier River at Hatch, Utah	April-July	58	1915-1984
Verde River above Horseshoe Dam, Arizona	January-May	47	1939-1985
Salt River near Roosevelt, Arizona	January-May	73	1913-1985
Little Colorado River above Lyman Lake, Arizona	February -June	46	1941-1986
Yellowstone River at Billings, Montana	April-September	57	1929-1985
Payette River near Horseshoe, Bend, Idaho	April-September	33	1953-1985
Rio Grande River near Del Norte, Colorado	April-September	27	1958-1985

Table 2. Topologies and permutations of fractions chosen by Reese and Krzysztofowicz.

River Name	Bifurcation Topology	Permutation of Fractions
Boise	1-3	(2,1,3,4)
Weiser	1-3	(2,1,3,4)
Little Truckee	1-3	(2,1,3,4)
Gila	1-3	(1,2,3,4)
Salmon	2-2	(1,2,3,4)
Falls	2-2	(1,2,3,4)
Humboldt	2-2	(1,2,3,4)
Sevier	2-2	(1,2,3,4)
Verde	1-4(1-3)	(4,3,2,5,1)
Salt	1-4(1-3)	(1,2,3,4,5)
Little Colorado	2-3	(4,5,3,1,2)
Yellowstone	2-4(1-3)	(1,2,3,6,4,5)
Payette	2-4(2-2)	(1,2,3,4,5,6)
Rio Grande	2-4(2-2)	(3,4,5,6,1,2)