Mixture model of generalized chain-dependent processes and its application to simulation of interannual variability of daily rainfall

Xiaogu Zheng a,*, Richard W. Katz b

a National Institute of Water and Atmospheric Research, Wellington, New Zealand
b Institute for Study of Society and Environment, National Center for Atmospheric Research, Boulder, CO, United States

Received 11 April 2007; received in revised form 26 October 2007; accepted 29 October 2007

Summary This study is devoted to simulation of the interannual variability of daily rainfall, such as the mean and variance of seasonal rainfall totals and the correlation between seasonal rainfall totals and a seasonal climate variable (e.g., an index of atmospheric or oceanic circulation). Prior work established that a mixture of chain-dependent processes, involving conditioning on an observed seasonal circulation index, results in improved representation of the interannual variance of seasonal rainfall totals, as well as of the correlation between rainfall and this index. The present study extends this work through consideration of a mixture model of generalized chain-dependent processes. Here "generalized" refers to the inclusion of a generalized linear model in accounting for the relationship between daily rainfall statistics and the circulation index. It is shown that this extension results in further improvements in the simulation of the interannual variability of seasonal rainfall and, especially, in its correlation with the circulation index.

Introduction

Daily rainfall is a very important environmental variable for hydrological research. Particularly, it is the most important input for hydrological models, such as TOPNET (Bandaragoda et al., 2004). Such hydrological models can simulate inflows and outflows of reservoirs, which are vital in the analysis of hydroelectricity generation capacity in New Zealand (Zheng and Thompson, 2007).

A chain-dependent process (i.e. Katz and Parlange, 1998) is a popular stochastic rainfall generator to simulate daily rainfall. It can simulate the intraseasonal variability of daily rainfall, such as the distributions of daily rainfall amount and wet–dry durations, reasonably well. However, in order to study the response of hydrological states to climate variability and change, it is also important to correctly simulate the interannual variability of rainfall, such as the variance of seasonal rainfall totals and the correlation between

* Corresponding author. Tel.: +64 4 386 0499; fax: +64 4 386 2153. E-mail address: x.zheng@niwa.co.nz (X. Zheng).
seasonal rainfall totals and a climate variable (e.g., a circulation index) (Zheng and Thompson, 2007).

Numerous studies indicate that a chain-dependent process tends to underestimate the variance of seasonal rainfall total, a phenomenon called "overdispersion" (Katz and Parlange, 1998; Katz and Zheng, 1999; Zheng and Thompson, 2007). Overdispersed models tend to underestimate the risks of drought and flood to hydropower generation capacity. Katz and Zheng (1999) proposed a mixture of chain-dependent processes to reduce this overdispersion. Despite only two possible states, they found that the hidden index involved in the mixture model is closely related to observed information about atmospheric circulation patterns.

Observed climate variables, such as Southern Oscillation indexes, tropical sea surface temperature (SST indexes), and other indexes of global and regional mean circulations, are closely related to rainfall. Many of these climate variables possess considerable predictability at seasonal to decadal time scales. Correctly simulating the relation between rainfall and a climate variable will help to understand the variability of rainfall in response to climate variability and change and, therefore, to project future scenarios of hydropower generation capacity in New Zealand. As an extension of Katz and Zheng (1999), Zheng and Thompson (2007) proposed a mixture of chain-dependent process, conditional on an observed circulation index, to reflect the correlation between rainfall and this index and to reduce the overdispersion. Although the model simulates the correlation fairly well, it is still not entirely satisfactory. In particular, for some climates considerable overdispersion remains.

Generalized linear modeling is another approach to rainfall generation to simulate the relation between rainfall and a climate variable. Katz and Parlange (1998). Let \( Y_t \) denote the sequence of daily precipitation occurrences in year \( y \) (i.e., \( Y_{y,t} = 1 \) indicates a "wet day" and \( Y_{y,t} = 0 \) a "dry day"). It is assumed that this process is a first-order Markov chain: a model completely characterized by the transition probabilities

\[
P_{jk} = \Pr(Y_{y,t+1} = k|Y_{y,t} = j), \quad j, k = 0, 1.
\]

Note that \( P_{00} = 1 - P_{11}, J = 0, 1.\)

| Table 1 | Statistical models for daily rainfall and IPO |

<table>
<thead>
<tr>
<th>Model</th>
<th>Name of model</th>
<th>Number of parameters</th>
<th>Constraints on parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Chain-dependent process</td>
<td>6</td>
<td>( \alpha_0 = 1, \alpha_1 = 0, \beta_0 = 1, \beta_1 = 0 )</td>
</tr>
<tr>
<td>2</td>
<td>Generalized chain-dependent process</td>
<td>8</td>
<td>( \alpha_0 = 1, \alpha_1 = \alpha_2 = 1, \beta_1 = 0 )</td>
</tr>
<tr>
<td>3</td>
<td>Mixture of chain-dependent process</td>
<td>11</td>
<td>( \alpha_0 = 1, \alpha_1 = \alpha_2 = 1, \beta_1 = 0 )</td>
</tr>
<tr>
<td>4</td>
<td>Mixture of chain-dependent process conditional on climate variable</td>
<td>12</td>
<td>( \alpha_0 = 1, \alpha_1 = \alpha_2 = 1, \beta_1 = 0 )</td>
</tr>
<tr>
<td>5</td>
<td>Mixture of generalized chain-dependent process</td>
<td>15</td>
<td>( \alpha_0 = 1, \alpha_1 = \alpha_2 = 1, \beta_1 = 0 )</td>
</tr>
<tr>
<td>6</td>
<td>Mixture of generalized chain-dependent process conditional on climate variable</td>
<td>16</td>
<td>( \alpha_0 = 1, \alpha_1 = \alpha_2 = 1, \beta_1 = 0 )</td>
</tr>
</tbody>
</table>

Left-subscript indicates the state of the two-state index \( l \). \( \alpha_0, \alpha_1, \beta_0 \) and \( \beta_1 \) are the GLM parameters for the transition probabilities (see Eq. (2)), \( \mu \) and \( \sigma \) are the mean and standard deviation of power transformed daily rainfall intensity (see Eq. (A-1)), and \( m \) is the mean of climate variable \( X \) (see Eq. (4)).
Let \( R = \{R_{y,t}, t=1, \ldots, T\} \) denote the time series of daily precipitation amounts in year \( y \). The "intensity" \( R_{y,t} \) (i.e., days for which \( J_{y,t} = 1 \)) is taken to be conditionally independent and identically distributed. It is further assumed that daily precipitation intensity has a power transform distribution. That is, the transformed variable \( R_{y,t}^p \) has a Gaussian distribution, say, with mean \( \mu \) and standard deviation \( \sigma \). For example, the values of \( p = 1/2, 1/3, \) or \( 1/4 \) are commonly employed to account for the high degree of positive skewness in the distribution of daily precipitation amounts. Alternatively, a positively skewed distribution, such as the gamma or mixed-exponential, may be fitted directly to the transformed data.

Recently, generalized linear models (GLM, McCullagh and Nelder, 1989) were applied to model daily rainfall, including its relationship with a circulation index (see, for example, Chandler and Wheater, 2002; Furrer and Katz, 2007). In this study, we will investigate a simple GLM rainfall generator, in which the rainfall occurrence is related to a seasonal climate variable, whose state in year \( y \) is denoted by \( X_y \) (i.e. an SST index), by logistic regression

\[
\Pr(y_{t-1} = j | y_t = 1) = 1 - 1/[1 + \exp(x_0 + \beta_1 X_y + j(\beta_0 + \beta_1 X_y))].
\]

If \( \beta_0 = 0 \) and \( \beta_1 = 0 \), the model degenerates to a chain-dependent process.

**Mixture of generalized chain-dependent processes conditional on a climate variable**

Let \( I_y \) (\( I_y = i, i = 0, 1 \)) denote a two-state index, usually hidden or unobserved or only partially observed, that remains constant over the time period of \( T \) days (i.e., a season). The annual time series of index states is assumed to be identically distributed, with a common distribution

\[
w = \Pr(I_y = 1) = 1 - \Pr(I_y = 0).
\]

For a given year, the time series of daily precipitation amounts \( \{R_{y,t}, t=1, \ldots, T\} \) is a mixture of generalized chain-dependent processes; i.e., given \( I_y = i \), it is conditionally a generalized chain-dependent process with parameters \( \rho_0, \rho_1, \beta_0, \beta_1, \mu_i, \sigma_i \) and \( p \) (i.e., for simplicity, the power transform parameter \( p \) remains fixed). Moreover, if the rainfall depends not only on the hidden index \( I_y \), but also on a seasonal climate variable \( X_y \), then \( \{R_{y,t}, t=1, \ldots, T\} \) is a mixture of generalized chain-dependent processes conditioned on a climate variable.

The seasonal climate variable \( X_y \) is assumed to be a mixture of two Gaussian random variables with mean \( m \) that depends on \( I_y = 1 \) and a common standard deviation \( s \). Specifically,

\[
\Pr(X_y | I_y = 1) = \phi(X_y | m, s).
\]

where "\( \phi \)" is being used as a generic "probability function" (i.e., both for discrete probabilities and for density functions) and \( \phi (x; m, s) \) denotes the Gaussian density function with mean \( m \) and standard deviation \( s \). By Bayes theorem,

\[
w_y = \Pr(I_y = 1 | X_y) = \frac{\Pr(X_y | I_y = 1) \Pr(I_y = 1)}{\Pr(X_y | I_y = 0) \Pr(I_y = 0) + \Pr(X_y | I_y = 1) \Pr(I_y = 1)}
= \frac{[\phi(X_y | m, s)w]/[\phi(X_y | 0, m, s)(1 - w)]}{\phi(X_y | m, s)}. \quad (5)
\]

If \( \rho m = \rho m, \) then \( \Pr(I_y = 1 | X_y) = w = Pr(I_y = 1) \) by Eq. (5).

Therefore, if \( X_y \) is independent of \( I_y \), then the mixture of generalized chain-dependent processes conditioned on \( X_y \) degenerates to a mixture of generalized chain-dependent processes.

**Parameter estimation**

Estimation of the parameters is undertaken by maximum likelihood. If the state index \( I_y \) (0 or 1) were actually known, the complete-data log likelihood function for the precipitation and the climate variable would be

\[
\ln \Pr(R_{y,T}, X_y, y = 1, \ldots, Y; \eta) = \sum_y (1 - I_y) \ln[\Pr(R_y | X_y, I_y = 0) \Pr(X_y | I_y = 0) \Pr(I_y = 0)]
+ \sum_y I_y \ln[\Pr(R_y | X_y, I_y = 1) \Pr(X_y | I_y = 1) \Pr(I_y = 1)]
= \sum_y (1 - I_y) \ln[\Pr(R_y | X_y, 0\mu)(\phi(X_y | 0, m, s)(1 - w)]
+ \sum_y I_y \ln[\Pr(R_y | X_y, 1\mu)(\phi(X_y | 1, m, s)w)], \quad (6)
\]

where \( R_y = \{R_{y,t}, t=1, \ldots, T\}, \eta = \{w, \theta, m, s; i = 0, 1\}, \theta = \{\rho_0, \rho_1, \beta_0, \beta_1, \mu_i, \sigma_i\} \), and an expression for \( \Pr(R_y | X_y, \theta) = \Pr(R_y | X_y, I_y = i) \) is given in Appendix A. However, the state index \( I_y \) is not observable, we have to apply the popular expectation-maximization (EM) algorithm (see Dempster et al., 1977; McLachlan and Krishnan, 1997 for the theory; and Hughes and Guttorp, 1994; Sansom and Thomson, 1998; Katz and Zheng, 1999; Bellone et al., 2000 for applications to rainfall modeling).

The fitting procedure comprises an E-step and a M-step. The E-step involves constructing the incomplete-data log likelihood function. Because the complete-data log likelihood function is a linear function of the unobservable state \( I_y \), the incomplete-data log likelihood can be obtained by replacing \( I_y \) in Eq. (6) with its conditional expectation \( \hat{I}_y \equiv \Pr(I_y = 1 | R_y, X_y) \) (see p. 18 of McLachlan and Krishnan, 1997). That is,

\[
\ln \Pr_e(R_y, X_y, y = 1, \ldots, Y; \eta)
= \sum_y (1 - I_y) \ln[\Pr(R_y | X_y, 0\mu)(\phi(X_y | 0, m, s)(1 - w)]
+ \sum_y I_y \ln[\Pr(R_y | X_y, 1\mu)(\phi(X_y | 1, m, s)w)]. \quad (7)
\]

By Bayes theorem,

\[
i_y = \Pr(I_y = 1 | R_y, X_y) = \Pr(R_y | I_y = 1, X_y) \Pr(I_y = 1 | X_y)
= \sum_i \Pr(R_y | I_y = i, X_y) \Pr(I_y = i | X_y)
= \Pr(R_y | X_y, 0\mu)w_y / [\Pr(R_y | X_y, 0\mu)(1 - w_y)]
+ \Pr(R_y | X_y, 1\mu)w_y.
\]

The M-step involves estimating the parameters \( \eta \) by maximizing the incomplete-data likelihood function (7). For example, from Eq. (2), the parameters \( \rho_0 \) and \( \rho_1 \) are estimated by maximizing
Here the transition count $n_{jk,y}$ denotes the number of times that the Markov chain for precipitation occurrence makes a transition from state $j$ to state $k$ in the sample in year $y$ (e.g. $n_{01,y}$ is the number of times that a dry day is followed by a wet day in year $y$). Given the estimates of $\lambda_0$ and $\lambda_1$, and $\beta_0$ and $\beta_1$ are estimated by maximizing

$$\sum_y i_y \ln \left( \prod_{t=0}^{n_{01,y}} \Pr(Y_{t+1} \mid Y_{t-1}, X_t, l_y = 1) \right)$$

$$= \sum_y i_y (n_{01,y}(\lambda_0 + \lambda_1 X_t)) - (n_{01,y} + n_{11,y}) \ln [1 + \exp ((\lambda_0 + \lambda_1 X_t))] \right).$$

(9)

Similarly, $\mu$ and $\sigma$ are estimated by minimizing

$$\sum_y i_y \left\{ - (n_{10,y} + n_{11,y}) \ln [2\pi(\sigma)^2] / \left[ 2 - \sum_{t=0}^{n_{11,y}} (R_{y,t+1}^2 - \mu)^2 / 2(\sigma)^2 \right] \right\}$$

(10)

to get

$$\hat{\mu} = \sum_y i_y s_{1,y} / \sum_y i_y (n_{10,y} + n_{11,y})$$

(12)

and

$$\hat{\sigma}^2 = \sum_y i_y (s_{2,y} - 2s_{1,y}(\hat{\mu}) + (\hat{\mu})^2) / \sum_y i_y (n_{10,y} + n_{11,y}).$$

(13)

where

$$s_{1,y} = \sum_t R_{y,t}^0, \quad s_{2,y} = \sum_t (R_{y,t}^0)^2.$$  

(14)

The expressions for the estimation of the parameters $\{\lambda_0, \lambda_1, \beta_0, \beta_1, \sigma, \mu\}$ are identical in form to those for the corresponding parameters $\{\lambda_0, \lambda_1, \beta_0, \beta_1, \sigma, \mu\}$, except that $i_y$ should be replaced by $1 - i_y$ in Eqs. (9) – (13).

Finally, from the expression of $\mu, s$ and $w$ are estimated by maximizing

$$\sum_y \{(1 - i_y) \ln [\phi(X_y; \beta_3, m, s)(1 - w)] + i_y \ln [\phi(X_y; \beta_3, m, s)w]\}$$

(15)

to get

$$\hat{\mu} = \sum_y (1 - i_y) X_y / \sum_y (1 - i_y), \quad \hat{\mu} = \sum_y i_y X_y / \sum_y i_y,$$

(16)

$$\hat{\sigma}^2 = \frac{1}{Y} \sum_{y=1}^{Y} [(1 - i_y)(X_y - \hat{\mu})^2 + i_y(X_y - \hat{\mu})^2]$$

(17)

and

$$\hat{\omega} = \frac{1}{Y} \sum_{y=1}^{Y} i_y.$$  

(18)

The EM algorithm requires a set of initial values for the model parameters, and the E- and M-steps are repeated alternately until $\ln \Pr(Y, X, l_y = 1, \ldots, Y, \eta)$ (i.e. Eq. (7)) converges. As the initial parameters, we use the fitted parameters from the mixture of chain-dependent processes (Zheng and Thompson, 2007).

Having estimated the parameters, Monte Carlo techniques can be applied to simulate realizations (see Appendix B).

A simulation study

Background

Lake Pukaki (Fig. 1) is one of the most important lakes in New Zealand to supply water for hydroelectric power generation. It is crucial for New Zealand to understand the variability of the outflow from the lake and to estimate this variability for the next two to three decades. To address this issue, in 2003 the National Institute of Water and Atmospheric Research (NIWA) of New Zealand launched a project "Climate-related Risks for Energy Supply and Demand" (Zheng and Thompson, 2007).

In this NIWA project, the hydrological catchment model TOPNET (Bandaragoda et al., 2004) will be used to simulate the inflow into the lake and then the outflow from the lake. The input variables of the catchment model are daily precipitation amount, daily maximum and minimum temperatures and dewpoint temperature. Since rainfall explains more than 70 percent of the annual variation in inflow into the lake, it is the most important forcing of the four weather elements for TOPNET.

In order to estimate the rainfall variability over the next two to three decades, a climate variable is needed that is both predictable and significantly associated with rainfall.
on a decadal time scale. Fortunately, the Interdecadal Pacific Oscillation (IPO) may be such a climate variable. The IPO has significant impacts on rainfall and river flows in certain regions of New Zealand. In the west and south of the South Island, the negative IPO phase is generally associated with lower rainfall and inflows and vice versa for the positive IPO phase (Salinger et al., 2001; McFarlane and Henderson, 2003; Zheng and Frederiksen, 2006; Zheng and Thompson, 2007). An aspect of this NIWA project is to assess how precipitation in the Lake Pukaki area, on interannual time scales, is linked to the IPO, and in particular the correlation between seasonal rainfall and the IPO.

Data

As a seasonal climate variable, we use the “low-frequency” IPO index, provided by the Hadley Centre of the United Kingdom Meteorological Office. It is derived from the third empirical orthogonal function (EOF) pattern of 13-year low-pass filtered global SST (see Fig. 2 in Zheng and Thompson, 2007). There are three major phases during the 20th century: positive phases from 1920–1945 and 1978–1999, and a negative phase between 1946 and 1977. The IPO appears to modulate the frequency and intensity of the El Niño-Southern Oscillation (ENSO) phenomenon: during the last positive phase of the IPO there were more El Niño episodes and fewer La Niña ones than during the preceding negative IPO phase (Folland et al., 1999; Power et al., 2005).

There are only four rainfall stations, Lake Tekapo, Lake Ohau and Mt Cook, and Franz Josef (see Fig. 1 for locations), with long record lengths covering the period 1950–2000, which roughly spans one complete cycle of the IPO, i.e. one positive and negative phase. Only these four stations are used in this study, since one of our major objectives is to model the relationship between precipitation and the IPO. There are, however, several other rainfall stations near the lake, but these cannot be used owing to their shorter record lengths. A wet day, in the context of this study, occurs when at least 1 mm of precipitation was recorded by the rain gauge; otherwise, the day is treated as dry.

Results

The six rainfall generators: (1) The chain-dependent process; (2) the generalized chain-dependent process; (3) the mixture of chain-dependent processes; (4) the mixture of chain-dependent processes conditioned on the IPO; (5) the mixture of generalized chain-dependent process; and (6) the mixture of generalized chain-dependent processes conditioned on the IPO are fitted to the austral summer season daily precipitation for the four long-term rainfall stations near Lake Pukaki. The names, the number of parameters and the constraints on the parameters of the general process to obtain these models are listed in Table 1. We will investigate models (4)–(6) in the simulation of interannual variability, while models (1)–(3) are alternatives for comparison sake. For simplicity, we assume that the parameters of these models are constant over the entire summer season (i.e., ignoring any annual cycles).

The daily precipitation has been power transformed. Values for p of 1/3, 2/5, 1/2 and 4/7 are adopted for Lake Tekapo, Lake Ohau, Mt Cook and Franz Josef, respectively. These values of p were empirically chosen to make the simulated seasonal mean totals close to the observed. The climate variable is the observed mean IPO index over the summer season. Based on the fitted parameters and the observed seasonal IPO index, 100 independent simulations of the DJF daily precipitation over a 50-year period are generated for the six models. Although quite rare, theoretically it is possible that a negative value of $R_p$ could be generated on wet days. In this case, the simulated daily precipitation amount is set to 1 mm, the lowest rainfall measurement on wet days.

For each simulation, the mean and variance of the summer seasonal total, and the correlation between the simulated seasonal total and the IPO index are calculated. The ensemble means of these simulated statistics and the corresponding observed statistics are listed in Tables 2–5 for the four stations near Lake Pukaki.

The mean seasonal precipitation totals are simulated quite well by all six rainfall generators. At all four sites, the simulated statistics are virtually equivalent to the observed. The observations are well within the 90% two-sided confidence interval for the simulated means. The simulated seasonal mean totals are sensitive to the power transformation parameter p. Our experience is that when p increases, the simulated seasonal mean total increases, presumably because of a reduction in the bias introduced by applying such a nonlinear transformation.

For Lake Tekapo, Lake Ohau and Mt Cook, models (1)–(2) underestimate the observed standard deviations by more than 30%. The observed standard deviation falls outside of the 90% confidence interval of the simulated statistics, indicating overdispersion in the data relative to these models. The simulated standard deviations for models (3)–(6) are about 10% less than those observed, the observed standard deviations being well within the 90% confidence interval of the simulated statistics. This suggests that the overdispersion...
The correlation coefficients between the IPO index and seasonal precipitation total (mm); and Cor[(S(T))] denotes the correlation coefficient between the IPO index and seasonal precipitation total. Numbers in italics indicate that the observation is outside of the 90% two-sided confidence interval from the simulated variance. An asterisk indicates the optimal model.

Table 2 Fitted and observed statistics of the Lake Tekapo (p = 1/3) December—February daily precipitation over 1950–2000

<table>
<thead>
<tr>
<th>Model</th>
<th>E[S(T)]</th>
<th>SD[S(T)]</th>
<th>Cor[S(T),IPO]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>126</td>
<td>44</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>127</td>
<td>45</td>
<td>0.14</td>
</tr>
<tr>
<td>3</td>
<td>127</td>
<td>58</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>127</td>
<td>58</td>
<td>0.13</td>
</tr>
<tr>
<td>5</td>
<td>128</td>
<td>59</td>
<td>0.08</td>
</tr>
<tr>
<td>6</td>
<td>127</td>
<td>58</td>
<td>0.17</td>
</tr>
<tr>
<td>Observation</td>
<td>130</td>
<td>62</td>
<td>0.08</td>
</tr>
</tbody>
</table>

S(T) denotes the precipitation total within a season of T days; E[S(T)] denotes the mean seasonal precipitation total (mm); SD[S(T)] denotes the standard deviation of the seasonal precipitation total (mm); and Cor[S(T),IPO] is the correlation coefficient between the IPO index and seasonal precipitation total. Numbers in italics indicate that the observation is outside of the 90% two-sided confidence interval from the simulated variance. An asterisk indicates the optimal model.

Table 3 Fitted and observed statistics of the Lake Ohau (p = 2/5) December—February daily precipitation over 1950–2000

<table>
<thead>
<tr>
<th>Model</th>
<th>E[S(T)]</th>
<th>SD[S(T)]</th>
<th>Cor[S(T),IPO]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>260</td>
<td>78</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>260</td>
<td>78</td>
<td>0.01</td>
</tr>
<tr>
<td>3</td>
<td>262</td>
<td>102</td>
<td>0.01</td>
</tr>
<tr>
<td>4</td>
<td>262</td>
<td>103</td>
<td>0.09</td>
</tr>
<tr>
<td>5</td>
<td>263</td>
<td>103</td>
<td>0.02</td>
</tr>
<tr>
<td>6</td>
<td>262</td>
<td>102</td>
<td>0.08</td>
</tr>
<tr>
<td>Observation</td>
<td>264</td>
<td>112</td>
<td>0.14</td>
</tr>
</tbody>
</table>

S(T) denotes the precipitation total within a season of T days; E[S(T)] denotes the mean seasonal precipitation total (mm); SD[S(T)] denotes the standard deviation of the seasonal precipitation total (mm); and Cor[S(T),IPO] is the correlation coefficient between the IPO index and seasonal precipitation total. Numbers in italics indicate that the observation is outside of the 90% two-sided confidence interval from the simulated variance. An asterisk indicates the optimal model.

Table 4 Fitted and observed statistics of the Mt Cook (p = 1/2) December—February daily precipitation over 1950–2000

<table>
<thead>
<tr>
<th>Model</th>
<th>E[S(T)]</th>
<th>SD[S(T)]</th>
<th>Cor[S(T),IPO]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1089</td>
<td>263</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>1089</td>
<td>278</td>
<td>0.33</td>
</tr>
<tr>
<td>3</td>
<td>1096</td>
<td>460</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>1098</td>
<td>465</td>
<td>0.21</td>
</tr>
<tr>
<td>5</td>
<td>1091</td>
<td>443</td>
<td>0.21</td>
</tr>
<tr>
<td>6*</td>
<td>1098</td>
<td>460</td>
<td>0.32</td>
</tr>
<tr>
<td>Observation</td>
<td>1104</td>
<td>493</td>
<td>0.31</td>
</tr>
</tbody>
</table>

S(T) denotes the precipitation total within a season of T days; E[S(T)] denotes the mean seasonal precipitation total (mm); SD[S(T)] denotes the standard deviation of the seasonal precipitation total (mm); and Cor[S(T),IPO] is the correlation coefficient between the IPO index and seasonal precipitation total. Numbers in italics indicate that the observation is outside of the 90% two-sided confidence interval from the simulated variance. An asterisk indicates the optimal model.

Table 5 Fitted and observed statistics of the Franz Josef (p = 4/7) December—February daily precipitation over 1950–2000

<table>
<thead>
<tr>
<th>Model</th>
<th>E[S(T)]</th>
<th>SD[S(T)]</th>
<th>Cor[S(T),IPO]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1444</td>
<td>302</td>
<td>0.01</td>
</tr>
<tr>
<td>2</td>
<td>1432</td>
<td>321</td>
<td>0.33</td>
</tr>
<tr>
<td>3</td>
<td>1448</td>
<td>445</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>1433</td>
<td>437</td>
<td>0.24</td>
</tr>
<tr>
<td>5</td>
<td>1448</td>
<td>413</td>
<td>0.23</td>
</tr>
<tr>
<td>6*</td>
<td>1428</td>
<td>436</td>
<td>0.38</td>
</tr>
<tr>
<td>Observation</td>
<td>1435</td>
<td>526</td>
<td>0.40</td>
</tr>
</tbody>
</table>

S(T) denotes the precipitation total within a season of T days; E[S(T)] denotes the mean seasonal precipitation total (mm); SD[S(T)] denotes the standard deviation of the seasonal precipitation total (mm); and Cor[S(T),IPO] is the correlation coefficient between the IPO index and seasonal precipitation total. Numbers in italics indicate that the observation is outside of the 90% two-sided confidence interval from the simulated variance. An asterisk indicates the optimal model.

For Mt Cook and Franz Josef, although models (4)–(5) rep-
resent some of the correlation between the simulated seasonal means and the IPO index, it appears that the simulated correlations of model (6) are the closest to the observed value. So model (6) appears preferable.

The estimated probability densities of the seasonal rainfall total for the observations are reasonably close to those simulated by the optimal models, with the degree of positive skewness being captured well (Fig. 2). To further investigate the quality of simulations, the lower, middle and upper quartiles of the seasonal rainfall totals observed and simulated by the optimal models are listed in Table 7. All simulated interquartile ranges are reasonably close to those observed. This coincides with the fact that the overdispersion for Lake Tekapo, Lake Ohau and Mt Cook is effectively eliminated. The overdispersion for Franz Josef could be due to the underestimation of large rainfall as shown in Fig. 2.

### Discussion and conclusions

From the simulation study in Section "A simulation study", the following approach to best capturing the interannual variability of rainfall emerges.

The correlation between seasonal mean rainfall and a climate variable may be modeled by conditioning the hidden index and/or the transition probabilities of rainfall occurrence on the seasonal climate variable. The former approach was studied by Zheng and Thompson (2007) and the latter approach, in combination with mixture model, has been studied in the present paper. The covariability modeled by each approach seems to reflect independent sources. In fact, from Tables 2–5, the sum of squared correlations for models 4 and 5 is close to the squared correlation for model 6. This indicates that both approaches (model 4 and model 5) are important and that the combination of the two approaches (model 6) can lead to higher estimated correlation. This is indeed the case for rainfall at both Mt Cook and Franz Josef, where the correlations with the IPO index are statistically significant and the correlations estimated by the model 6 are very close to those observed. However, in some situations, especially when the observed correlation is not statistically significant, model 6 overestimates the correlation, such as for Lake Tekapo rainfall. In this case, a simpler approach may be preferable.

The mixture of generalized chain-dependent processes conditional on the IPO seems the most effective way to eliminate overdispersion in seasonal rainfall. However, there are some exceptions, such as at Franz Josef, where the mixture model reduces the overdispersion, but fails to completely eliminate it.

Adjustment of the power transformation $p$ seems a useful approach to simulate the correct seasonal mean rainfall total. As an example, power transformations with different values of $p$ were applied to the Mt Cook rainfall data and the results are listed in Table 6. As $p$ decreases, the simulated mean seasonal precipitation total decreases. But the change in the power transformation $p$ may cause a change in the fitted distribution of rainfall intensity as well. As an example, Table 6 shows that as $p$ decreases, the simulated extremes (characterized by the simulated maximum and 99th percentile) increase. Therefore, caution must be taken. Nevertheless, except for the tail of the distribution, the histograms of observed daily rainfall are reasonably close to the simulated distributions (not shown here, but very similar to those shown in Figure 4 of Zheng and Thompson (2007)).

The gamma is a popular distribution for rainfall intensity (e.g. Hughes and Guttorp, 1994). For investigating the per-

<table>
<thead>
<tr>
<th>Simulation model for rainfall amount</th>
<th>$E[S(T)]$</th>
<th>$SD[S(T)]$</th>
<th>90th percentile</th>
<th>95th percentile</th>
<th>99th percentile</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = 4/7$</td>
<td>1115</td>
<td>467</td>
<td>75</td>
<td>98</td>
<td>151</td>
<td>360</td>
</tr>
<tr>
<td>$p = 1/2$</td>
<td>1096</td>
<td>460</td>
<td>73</td>
<td>97</td>
<td>156</td>
<td>501</td>
</tr>
<tr>
<td>$p = 2/5$</td>
<td>1055</td>
<td>456</td>
<td>71</td>
<td>97</td>
<td>161</td>
<td>508</td>
</tr>
<tr>
<td>$p = 1/3$</td>
<td>1035</td>
<td>450</td>
<td>70</td>
<td>97</td>
<td>169</td>
<td>630</td>
</tr>
<tr>
<td>Gamma</td>
<td>1092</td>
<td>417</td>
<td>75</td>
<td>103</td>
<td>171</td>
<td>558</td>
</tr>
<tr>
<td>Observation</td>
<td>1104</td>
<td>494</td>
<td>78</td>
<td>121</td>
<td>218</td>
<td>537</td>
</tr>
</tbody>
</table>

The percentiles and maximum are for daily rainfall, the other statistics for seasonal rainfall.

### Table 6 Simulated statistics (mm) for Mt Cook

<table>
<thead>
<tr>
<th>Data</th>
<th>Low quartile</th>
<th>Median</th>
<th>Upper quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lake Tekapo</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observation</td>
<td>84</td>
<td>118</td>
<td>166</td>
</tr>
<tr>
<td>Simulation $p = 1/3$</td>
<td>89</td>
<td>116</td>
<td>162</td>
</tr>
<tr>
<td>Lake Ohau</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observation</td>
<td>189</td>
<td>243</td>
<td>313</td>
</tr>
<tr>
<td>Simulation $p = 2/5$</td>
<td>183</td>
<td>250</td>
<td>313</td>
</tr>
<tr>
<td>Mt Cook</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observation</td>
<td>743</td>
<td>975</td>
<td>1411</td>
</tr>
<tr>
<td>Simulation $p = 1/2$</td>
<td>770</td>
<td>963</td>
<td>1418</td>
</tr>
<tr>
<td>Franz Josef</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observation</td>
<td>1105</td>
<td>1449</td>
<td>1740</td>
</tr>
<tr>
<td>Simulation $p = 4/7$</td>
<td>1051</td>
<td>1454</td>
<td>1710</td>
</tr>
</tbody>
</table>
formance of the gamma distribution in eliminating overdispersion, the mixture model of chain-dependent processes with daily rainfall intensity distributed as the gamma was fitted to the Mt Cook daily rainfall. The results in Table 6 show that the simulated percentiles for the gamma distribution are marginally closer to the observed than are those for the power transformed Gaussian. However, the simulated mean standard deviation of seasonal mean total rainfall is less than that for the power transformed Gaussian distribution. For this reason, in this study we did not use the gamma distribution to model daily rainfall intensity.

A mixed exponential distribution is another popular model for rainfall intensity (e.g. Wilks and Wilby, 1999). However, when we used this distribution instead of the transformed normal distribution, overdispersion is not reduced at all because the probability of state \( I_y = 1 \) is always estimated either as being zero or one. This may be due to the hidden state index for light or heavy precipitation involvement in the mixed exponential distribution being the dominant effect, completely masking any influence of low-frequency variability (Zheng and Thompson, 2007). For this reason, in this study we did not use the mixed exponential distribution to model daily rainfall intensity.

An annual cycle can be introduced in modeling both rainfall occurrence and rainfall intensity (e.g. Furrer and Katz, 2007). Using the Bayesian Information Criterion (BIC), Furrer and Katz (2007) found that the model with annual cycles in both occurrence and intensity is preferred when fit to an annual time series of daily rainfall. As a preliminary study, we fitted a chain-dependent process with annual cycles in both occurrence and intensity, following Furrer and Katz (2007). However the annual cycle model is not preferred by the BIC. This indicates that the intra-seasonal variability within DJF is not statistically significant. For this reason, in this study we did not use models with seasonal cycles.

Overall, the mean and variance of seasonal rainfall totals may be best captured by the combination of the power transformation parameter \( \rho \) and the introduction of the mixture of two stochastic models for daily rainfall. The correlation between the seasonal rainfall totals and a seasonal climate variable is best captured by introducing dependence between the climate variable and the hidden index or the transition probabilities of the rainfall occurrence. However, in practice simulations should be performed to determine the optimal model for a particular location.

**Acknowledgements**

This work was initiated while R.W. Katz was a visiting scientist supported by the New Zealand Foundation for Research, Science and Technology (Contract C01X0302) and the Hidden Markov Models and Complex Systems research programme of the New Zealand Institute of Mathematics and its Applications. We thank the two anonymous reviewers and Drs. Ross Woods, Murray H. Smith and C.S. Thompson for useful comments. The National Center for Atmospheric Research is sponsored by the National Science Foundation.

**Appendix A. Expression for \( Pr(R_y|X_y,i) \)**

Since the rainfall occurrence is assumed to be related to a seasonal climate variable \( X_y \) by logistic regression (see Eq. (2)),

\[
Pr(J_y = 1|J_{y-1} = j; X_y, I_y = i) = 1 - \frac{1}{1 + \exp((i - \beta_0 + \beta_i X_y) \cdot j)}
\]

(A-1)

Since the daily rainfall intensity is also assumed to be a transformed Gaussian distribution,

\[
Pr(R_y|J_y = 1; X_y, I_y = i) = (2\pi(\sigma)^2)^{-1/2} \times \exp \left\{ - \sum_{t,J_y = 1} (R_y - i\mu)^2 / 2(\sigma)^2 \right\}
\]

(A-2)

Therefore,

\[
Pr(R_y|X_y,i) = Pr(R_y|X_y, I_y = i) \prod_{t} Pr(J_y|I_y = 1)
\]

\[
\prod_{t,J_y = 1} Pr(R_y|J_y = 1; X_y, I_y = i)
\]

\[
= Pr(J_y = 0|J_{y-1} = 0; X_y, I_y = i)^{n_{00}}
\]

\[
Pr(J_y = 1|J_{y-1} = 0; X_y, I_y = i)^{n_{10}}
\]

\[
Pr(J_y = 0|J_{y-1} = 1; X_y, I_y = i)^{n_{01}}
\]

\[
Pr(J_y = 1|J_{y-1} = 1; X_y, I_y = i)^{n_{11}}
\]

\[
\exp \left\{ - \sum_{t,J_y = 1} (R_y - i\mu)^2 / 2(\sigma)^2 \right\}
\]

\[
= \frac{\exp((x_0 + y_1 X_y))^{n_{01}}}{[1 + \exp((x_0 + y_1 X_y))^{n_{01}}]}
\]

\[
\times \frac{\exp((x_0 + y_1 X_y))^{n_{11}}}{[1 + \exp((x_0 + y_1 X_y))^{n_{11}}]}
\]

\[
(2\pi(\sigma)^2)^{-1/2} \exp \left\{ - \sum_{t,J_y = 1} (R_y - i\mu)^2 / 2(\sigma)^2 \right\}
\]

(A-3)

**Appendix B. Monte Carlo simulation**

For an initial occurrence state \( J_{y_0} \), a seasonal climate variable \( X_y \), and a given parameter set \{\( w_1, x_0, \beta_1, \beta_0, \beta, \sigma, i, m, s, p, t \mid t = 0, 1 \}\}, we can generate by Monte Carlo simulation a rainfall time series from a mixture of chain-dependent process conditional on \( X_y \). The procedure is as follows.

First, independent uniform random variables (over the range \((0, 1)\) \( U_y \) and \( \{X_y|t, t = 1, \ldots, T\} \), and independent standard Gaussian random variables \( \{Z_y|t, t = 1, \ldots, T\} \) are generated. The state index for the year \( y \) is constructed by \( I_y = [U_y < w_y] \).

(B-1)

where \([\cdot]\) is the indicator function and \( w_y = Pr(I_y = 1|X_y) \) (see Eq. (5)).

Second, the precipitation occurrences are recursively constructed by
\[ J_{y,t} = \{ \mathbb{I}_{Y_{y,t} \leq \mathbb{I}} | J_{y,t-1} = j; X_{y,t} = i \} \]  
(B-2)

where \( \mathbb{I}_{Y_{y,t} = 1 | J_{y,t-1} = j; X_{y,t} = i} \) is defined by Eq. (A-1).

Finally, the daily precipitation amount is simulated as
\[ R_{y,t}^{d} = J_{y,t} \left( \alpha Z_{y,t} + \mu \right) \]  
(B-3)

References


