# A multisite rainfall generation model applied to New Zealand data


Craig Thompson

*National Institute of Water and Atmospheric Research*

*New Zealand*

Peter Thomson

*Statistics Research Associates Ltd*

*New Zealand*

Xiaogu Zheng

*National Institute of Water and Atmospheric Research*

*New Zealand*

## Outline

1. Background
2. HMM model framework
3. Model fitting
4. Results
5. Conclusions

## 1. Background

Part of a research programme on *Climate-related risks for energy supply and demand*.

Aims: To construct suitable forecasting models of rainfall in hydro catchments which

- reliably estimate rainfall–related risk over forecast horizons of months to years;
- provide realistic scenarios of future rainfall variability over diverse spatial and temporal scales;
- account for seasonality, ENSO, IPO and other external forcings.

## Starting point

Wilks (1998) multisite daily rainfall generation model used within NIWA on an operational basis over the last 5 years.

This model has been

- reformulated as a (partially) hidden Markov model (HMM), rather than a simulation model;
- embedded within a more general HMM framework and its stochastic properties determined;
- fitted to selected New Zealand rainfall data using suitable statistical estimation procedures;
- evaluated and further potential improvements identified.

## 2. HMM model framework

Consider a small network of $K$ rainfall stations and observations

$$R_t(k) = \textit{accumulated rainfall over day } t$$

at rainfall station $k$.

Associate a local rainfall state $S_t(k)$ with each measurement $R_t(k)$ where

- $S_t(k) = \begin{cases} 0 & (\text{Dry at time } t) \\ 1 & (\text{Light rain at time } t) \\ 2 & (\text{Heavy rain at time } t) \end{cases}$

- the $S_t(k)$ are hidden with the exception of the dry state.

Only the rainfall amounts $R_t(k)$ are observed.

Key assumptions:

- Rainfall $\mathbf{R}_t = (R_t(1), \ldots, R_t(K))$ on day $t$ depends only on the hidden states $\mathbf{S}_t = (S_t(1), \ldots, S_t(K))$ for that day; i.e.

$$P(\mathbf{r} \leq \mathbf{R}_t < \mathbf{r} + \mathbf{dr} | \mathbf{S}) = P(\mathbf{r} \leq \mathbf{R}_t < \mathbf{r} + \mathbf{dr} | \mathbf{S}_t)$$

  where $\mathbf{S} = (\mathbf{S}_1, \ldots, \mathbf{S}_T)$.

- The $\mathbf{R}_t$ are independent given $\mathbf{S}$.

Commonly used (e.g. Katz (1977), Zucchini and Guttorp (1991), Wilks (1998) etc).

Advantages: Can separately model the

- distribution of rainfall amounts $R_t(k)$ within rainfall states;

- dynamics of rainfall patterns (persistence) through the $S_t(k)$.

## 2.1 A conditional model for rainfall $R_t(k)$

If $S_t(k)$ is known, assume that

$$R_t(k) = \beta_{S_t(k)}(k) X_t(k)$$

where the $X_t(k)$ are temporally independent exponentials with $E(X_t(k)) = 1$ and $\beta_0(k) = 0 < \beta_1(k) < \beta_2(k)$.

If $S_t(k)$ is unknown, the unconditional distribution of $R_t(k)$ is a mixture of two exponentials with a point mass at 0.

This simple parsimonious specification was adopted by Wilks (1998), but other distributions could be used (e.g. lognormal, gamma).

Model contemporaneous spatial dependence of the $X_t(k)$ by

$$X_t(k) = -\log(\mathbf{\Phi}(V_t(k)))$$

where $\mathbf{\Phi}(.)$ is the standard Gaussian cdf, the $\mathbf{V}_t = (V_t(1), \ldots, V_t(K))$ are iid and

$$\mathbf{V}_t \sim N(\mathbf{0}, \mathbf{\Psi}).$$

The correlation matrix $\mathbf{\Psi}$ determines the degree of spatial dependence between rainfall amounts. It does not depend on the local weather state $S_t(k)$.

This specification

- was proposed by Wilks (1998);
- is consistent with exponentials at each location;
- builds a relatively flexible joint distribution from the exponential marginals using a meta–Gaussian copula.

## 2.2 A model for rainfall states $S_t(k)$

At each location $S_t(k)$ is assumed to follow a stationary 3–state Markov chain with

$$P(S_t(k) = j | S_{t-1}(k) = i) = P_{ij}(k) \qquad (i, j = 0, 1, 2).$$

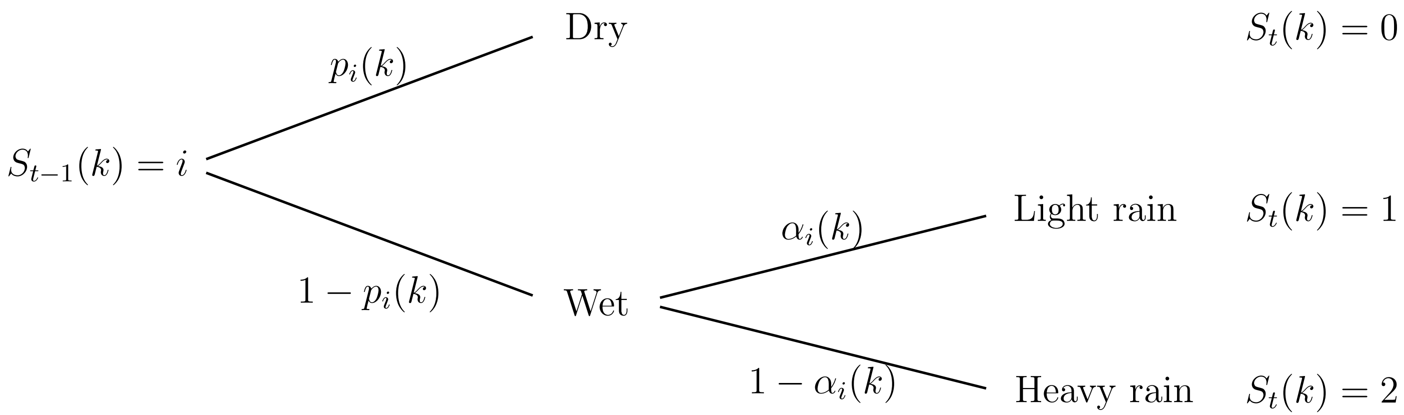
The transition probability matrix $\mathbf{P}(k)$ is parameterised as

$$\mathbf{P}(k) = \begin{bmatrix} p_0(k) & \alpha_0(k)(1 - p_0(k)) & (1 - \alpha_0(k))(1 - p_0(k)) \\ p_1(k) & \alpha_1(k)(1 - p_1(k)) & (1 - \alpha_1(k))(1 - p_1(k)) \\ p_2(k) & \alpha_2(k)(1 - p_2(k)) & (1 - \alpha_2(k))(1 - p_2(k)) \end{bmatrix}$$

where the probabilities $p_i(k)$, $\alpha_i(k)$ satisfy

$$\alpha_i(k) = P(S_t(k) = 1 | S_t(k) > 0, S_{t-1}(k) = i)$$
$$p_i(k) = P(S_t(k) = 0 | S_{t-1}(k) = i).$$

If $S_{t-1}(k) = i$ then the outcome of $S_t(k)$ can be represented by




For each location there are 6 parameters $p_i(k)$, $\alpha_i(k)$ $(i = 0, 1, 2)$.

This structural model is more general than the Wilks model where

$$p_1(k) = p_2(k), \quad \alpha_0(k) = \alpha_1(k) = \alpha_2(k)$$

and only 3 parameters are needed for each location.

Model contemporaneous spatial dependence of the $S_t(k)$ by

$$S_t(k) = \begin{cases} 0 & U_t(k) \in (-\infty, a_i(k)] \\ 1 & U_t(k) \in (a_i(k), b_i(k)] \\ 2 & U_t(k) \in (b_i(k), \infty) \end{cases}$$

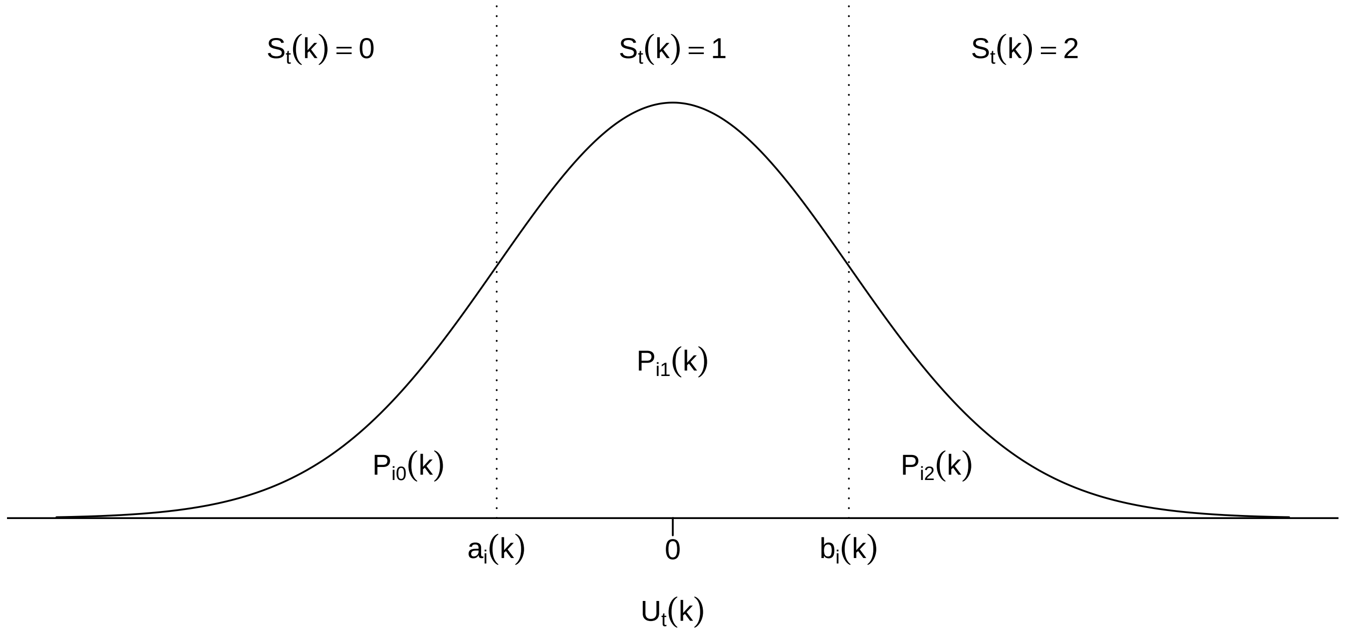
when $S_{t-1}(k) = i$ with

$$a_i(k) = \Phi^{-1}(p_i(k)), \qquad b_i(k) = \Phi^{-1}(p_i(k) + \alpha_i(k)(1 - p_i(k))).$$

Here the $\mathbf{U}_t = (U_t(1), \ldots, U_t(K))$ are iid $N(\mathbf{0}, \mathbf{\Omega})$, independent of the $\mathbf{V}_t$, and the correlation matrix $\mathbf{\Omega}$ determines the degree of spatial dependence between the $S_t(k)$.

This specification, as before,

- is consistent with the (marginal) Markov chain specification;
- builds a relatively flexible joint distribution from given marginals.




Readily simulated: If $S_{t-1}(k) = i$ the $R_t(k)$ are obtained by:

- generating the Gaussian $U_t(k)$ and corresponding $S_t(k)$;
- independently generating the Gaussian $V_t(k)$ and amounts

$$R_t(k) = -\beta_{S_t(k)}(k) \log(\Phi(V_t(k))).$$

## 2.3 Wilks model

This is the special case where

$$p_1(k) = p_2(k), \quad \alpha_0(k) = \alpha_1(k) = \alpha_2(k).$$

These imply that $S_{t-1}(k)$ and $S_t(k)$ are independent when

- $S_{t-1}(k) > 0$ (yesterday is wet);
- $S_t(k) > 0$ (today is wet).

Possibly too restrictive.

Wilks model is a simple Markov chain for wet and dry occurrences, with amounts modelled as a mixture of two exponentials.

Calibration of the Wilks model begins by estimating

- $p_0(k)$, $p_1(k)$ directly from observed wet and dry transitions;

- $\alpha_0(k)$, $\beta_1(k)$, $\beta_2(k)$ by fitting an exponential mixture to amounts.

Given these estimates, the spatial correlations

$$\omega_{jk} = \mathrm{cor}(U_t(j), U_t(k)), \qquad \psi_{jk} = \mathrm{cor}(V_t(j), V_t(k))$$

are backed out from

$$\mathrm{cor}(I_t(j), I_t(k)), \qquad \mathrm{cor}(R_t(j), R_t(k)).$$

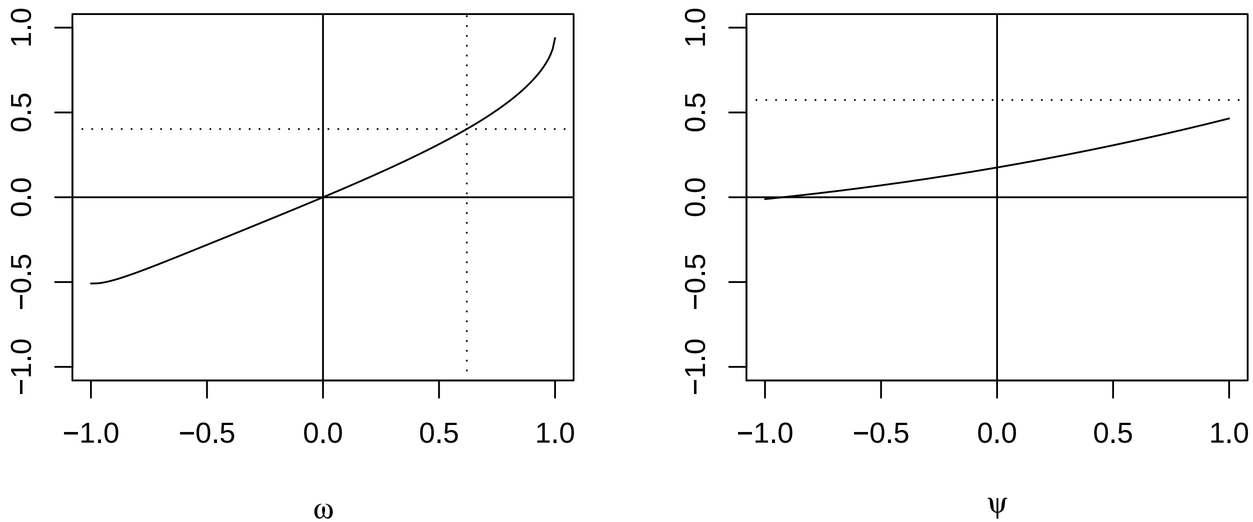
where

$$I_t(k) = \begin{cases} 0 & (R_t(k) = 0) \\ 1 & (R_t(k) > 0) \end{cases} = \text{rainfall occurrence}$$

and correlations are estimated from simulated $R_t(j)$, $R_t(k)$.

Computationally very intensive.

## Calibration of Wilks model for January rainfall at Coleridge and Rangiora




Left plot: $\mathrm{cor}(I_t(j), I_t(k))$ as function of $\omega$ with other parameters held fixed.

Right plot: $\mathrm{cor}(R_t(j), R_t(k))$ as function of $\psi$ with other parameters held fixed.

Horizontal lines show observed sample correlations.

Restricted ML estimates were $\tilde{\omega} = 0.66$ and $\tilde{\psi} = 0.52$.

## 3. Model fitting

Model fitted using maximum likelihood.

Strategy adopted.

- Use EM algorithm to explore marginal log–likelihoods (spatial independence) to obtain a range of initial estimates.
- Starting from initial estimates, use numerical optimisation to directly maximise the full log–likelihood.
- Fit a range of reduced dynamic models to the states $S_t(k)$.
- Examine the resulting estimates, AIC values, graphical diagnostics, etc to assess goodness of fit.

## Comments

- Takes advantage of EM's robustness to choice of initial values.

- Likelihood values and EM depend on

$$\gamma_t(\mathbf{s}) = P(\mathbf{S}_t = \mathbf{s}|\mathbf{R})$$

  where $\mathbf{R}$ denotes available observations. Calculated using computationally efficient recursions.

- The $\gamma_t(\mathbf{s})$ are also used to identify likely rainfall states and to estimate hidden quantities such as the stochastic mean

$$E(\beta_{S_t}|\mathbf{R})$$

  and to forecast risk parameters such as

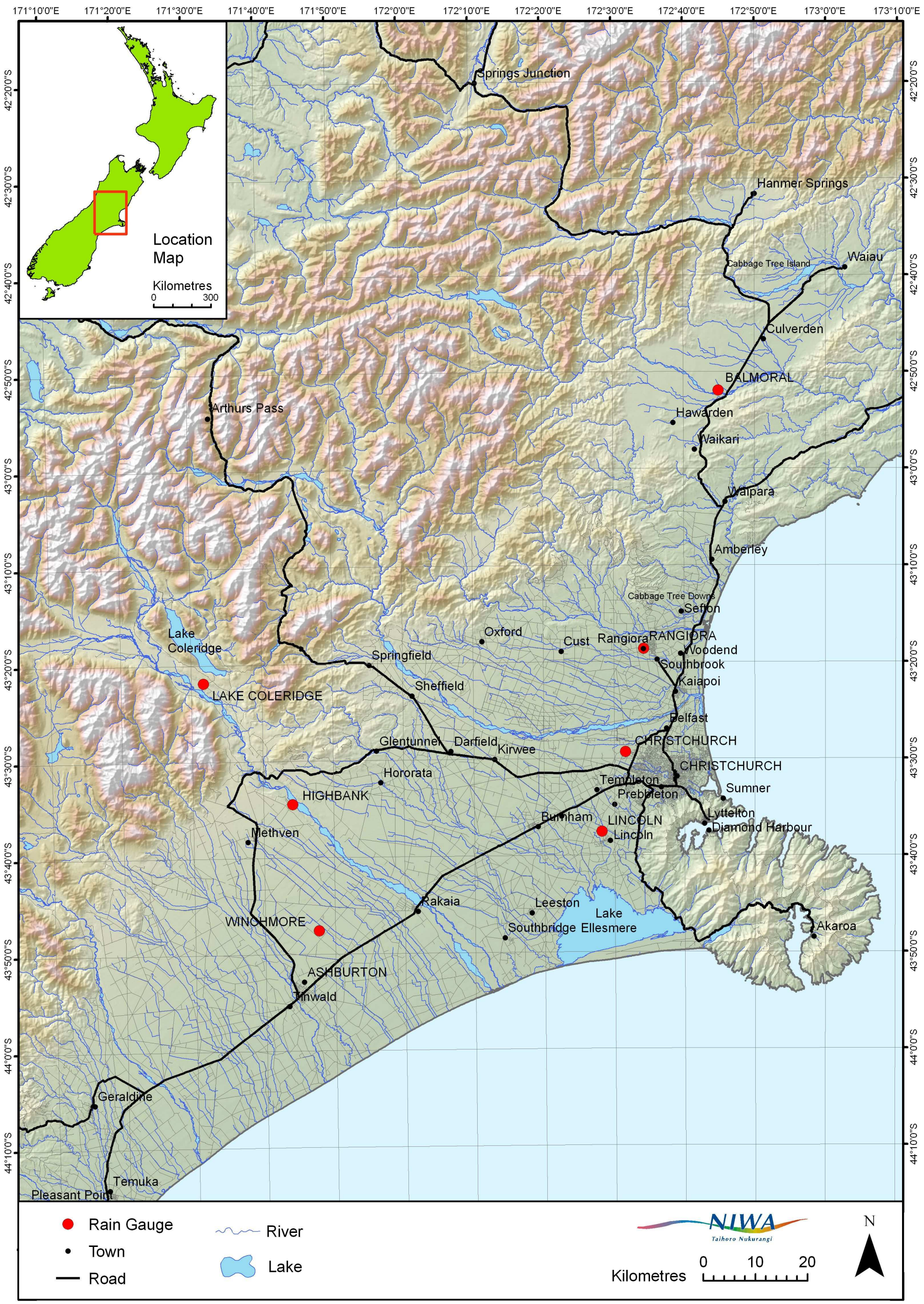
$$P(R_{T+t}(k) > r|\mathbf{R})$$

  where $T + t$ denotes some future time point.

## 4. Results

Simulation studies show that

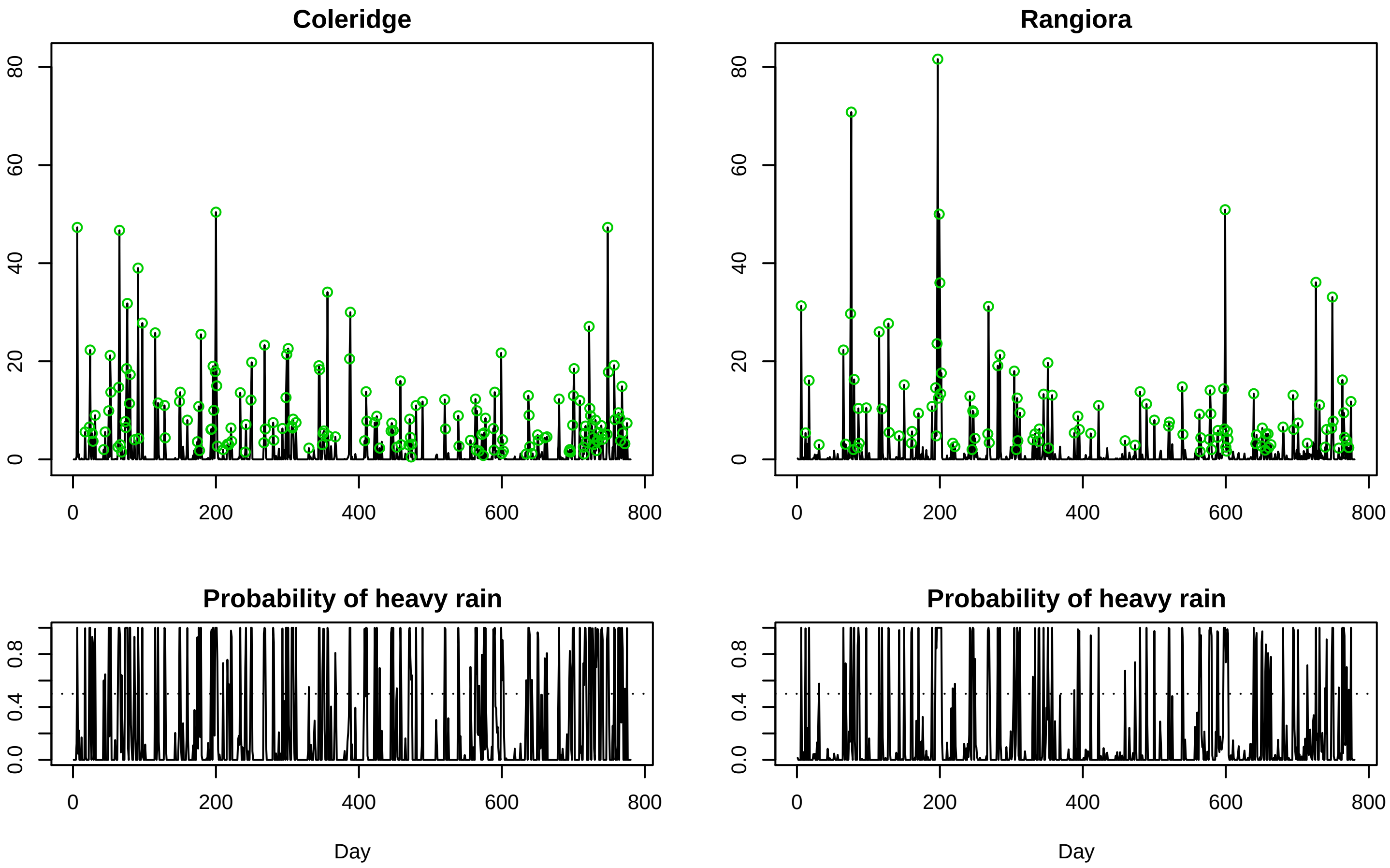
- full maximum likelihood (ML) performs best;
- marginal ML performs almost as well;
- both are significantly better than method of moments calibration in terms of accuracy and computational cost;
- sampling properties of ML estimators well approximated by asymptotic theory.




New Zealand rainfall data drawn from a small Canterbury network of 7 rainfall stations.

Focus on Coleridge and Rangiora rainfall for January and April over the period 1972–1997.

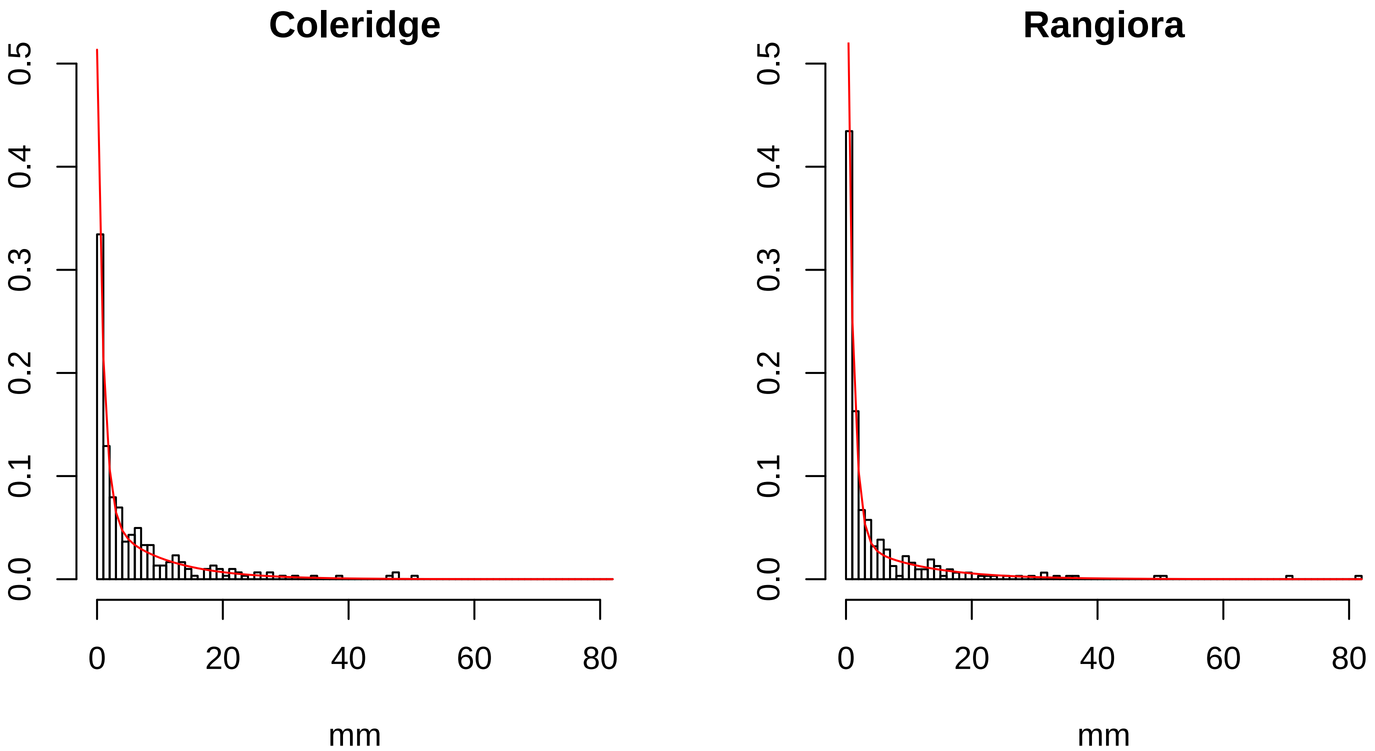
# Coleridge and Rangiora daily rainfall for April




Upper plots: daily rainfall with green indicating points classified as heavy rain.

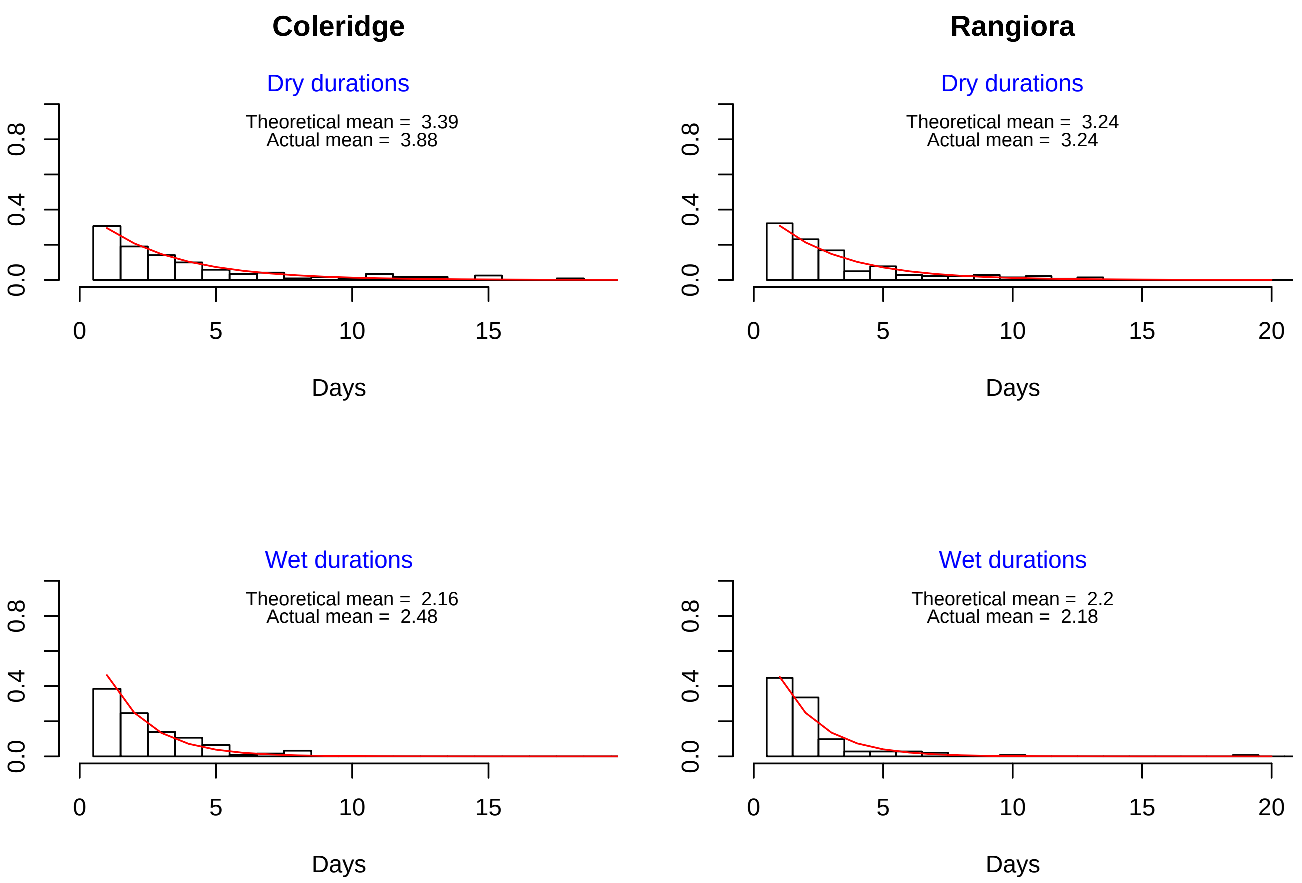
Lower plots: probability of heavy rain given the data.

# Coleridge and Rangiora daily rainfall for April



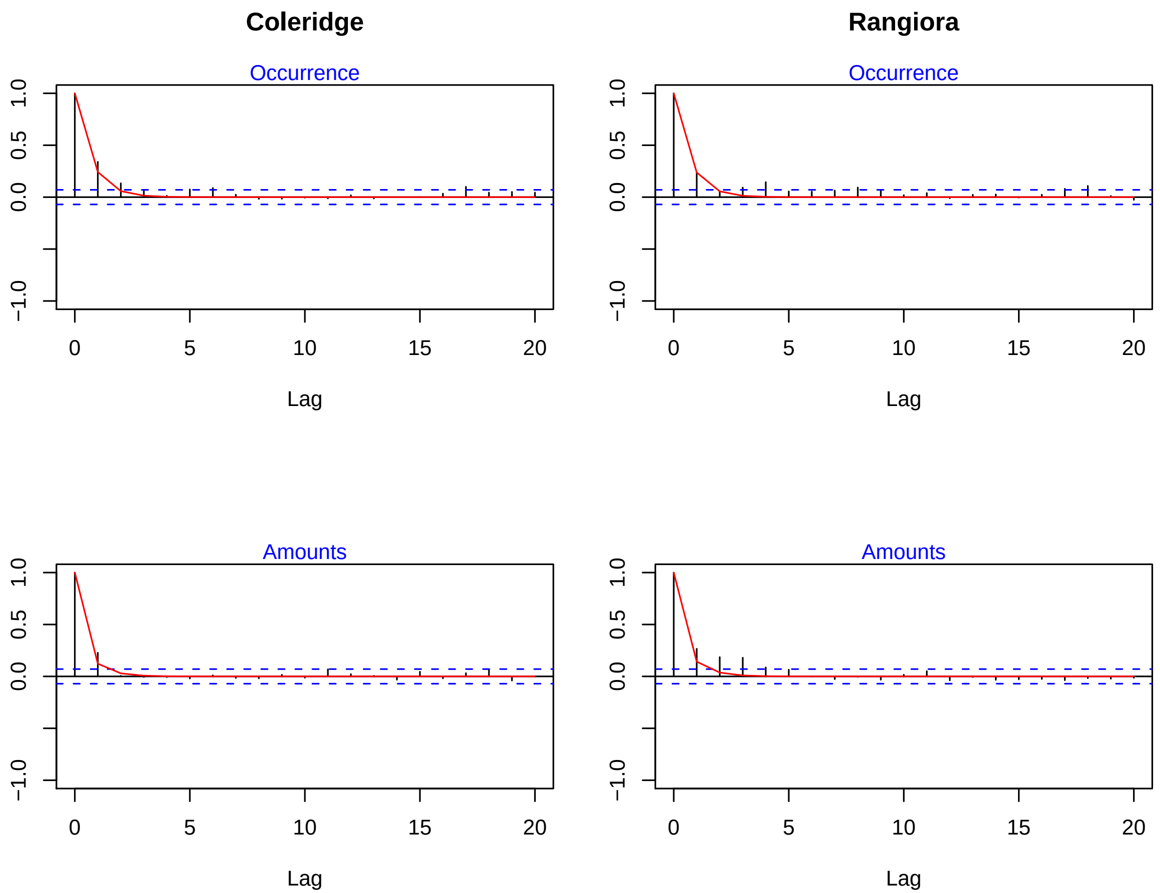

Histograms of daily rainfall on wet days with fitted exponential mixture distributions superimposed.

# Coleridge and Rangiora daily rainfall for April



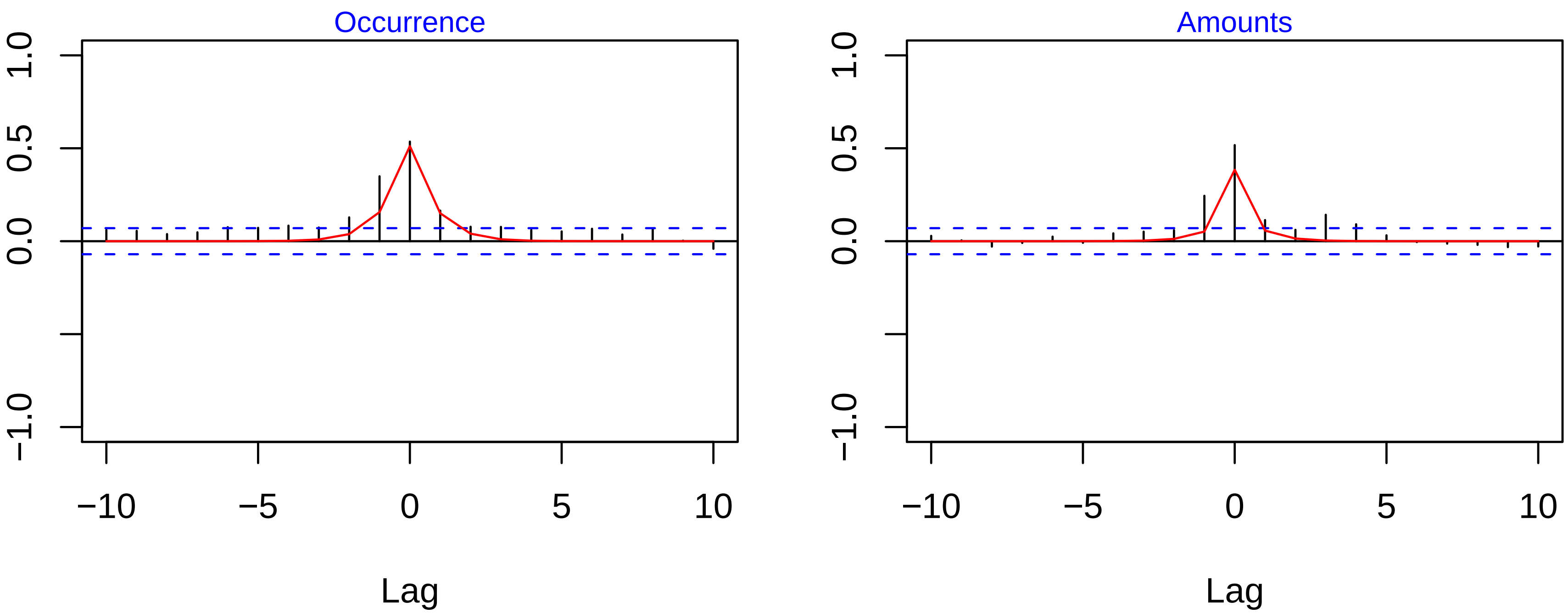

Histograms of dry and wet durations with fitted distributions superimposed.

# Coleridge and Rangiora daily rainfall for April




Autocorrelation functions of rainfall occurrence and amounts with fitted autocorrelation functions superimposed.

# Coleridge and Rangiora daily rainfall for April




Cross–correlation functions of rainfall occurrence and amounts with fitted cross–correlation functions superimposed.

## Data analysis summary

- AIC rarely supports Wilks model.
- Other reduced models explored, with spatially homogeneous parameters favoured in many cases.
- Rainfall distributions modelled reasonably well. (Statics)
- Dry durations and cross–correlations not always well–modelled. (Dynamics)

The above suggest that further structure (time and space) is needed to better explain the dynamics.

## 5. Conclusions

Wilks multisite weather generator model has been generalised to a local weather state HMM with

- copulas used to model spatial dependence;
- efficient statistical estimation procedures.

However need to

- augment the HMM model's dynamic structure in time and space so that it more closely reflects the data;
- incorporate stochastic seasonality;
- account for longer–term variation (ENSO, IPO etc).