# An application of Climate Data Science to New Zealand rainfall


John Sansom

*National Institute of Water and Atmospheric Research, New Zealand*

Jan Bulla

*University of Bergen, Norway*

Trevor Carey-Smith

*National Institute of Water and Atmospheric Research, New Zealand*

Peter Thomson

*Statistics Research Associates Ltd, New Zealand*

## Outline

1. Introduction
2. Breakpoint rainfall data and HSMM model
3. Precipitation states over aggregated time and space scales
4. Single-site analysis
5. Multi-site analysis
6. Conclusions

## 1. Introduction

Rainfall is a continuous-time phenomena typically recorded at a variety of spatial locations in the form of rainfall accumulations (daily, hourly etc).

High-resolution continuous-time rainfall data are not widely available.

Stochastic rainfall models are commonly fitted to daily rainfall aggregations, especially within stochastic weather generators. These models

- provide an imperfect description of the underlying dynamics and intensity of rainfall;
- suffer from well-known problems such as over-dispersion.

**Key question:**

What accumulation time scales are likely to result in more faithfull descriptions of the space-time dynamics of continuous-time rainfall?

First need to determine the space-time dynamics of continuous-time rainfall.

Here this is achieved by

- identifying suitable (synoptic) precipitation states (Dry, Showers, Rain);
- using a Hidden Semi-Markov Model (HSMM) of continuous-time rainfall;
- applied to (continuous-time) breakpoint rainfall data.

The HSMM model (Sansom and Thomson, 2001)

- uses a physically-meaningful hierarchy of precipitation states;
- accurately reflects rainfall dynamics;
- has proved useful in practice;
- but is currently univariate only.

What is breakpoint rainfall data and how does it differ from conventional rainfall accumulations?

## 2. Breakpoint rainfall data and HSMM model

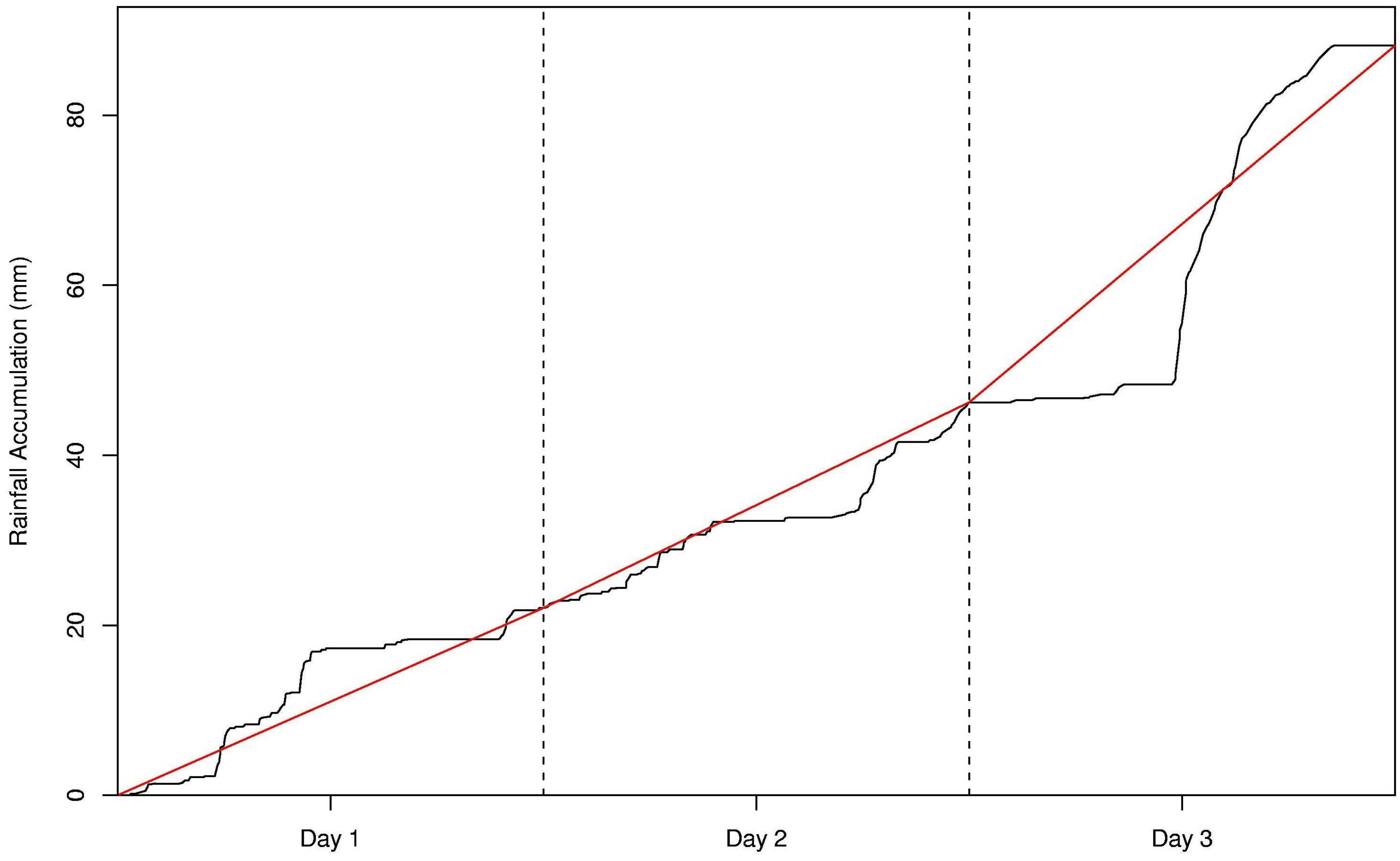
- Records times of rain-rate changes and steady rates between; i.e.

$$\mathbf{Y}_k = (R_k, D_k) \qquad (k = 1, \ldots, K)$$

where, for breakpoint $k$,

$$R_k = \text{rainfall rate}, \qquad D_k = \text{duration}.$$

- High resolution of around 6 seconds.
- Approximately 3000 breakpoints per annum, most of which give bivariate measurements of rate and duration.
- Efficient representation of continuous-time rainfall (especially dry periods) by comparison to conventional rainfall accumulations over hours (or less).
- Can be used to generate more conventional accumulations over any time interval.



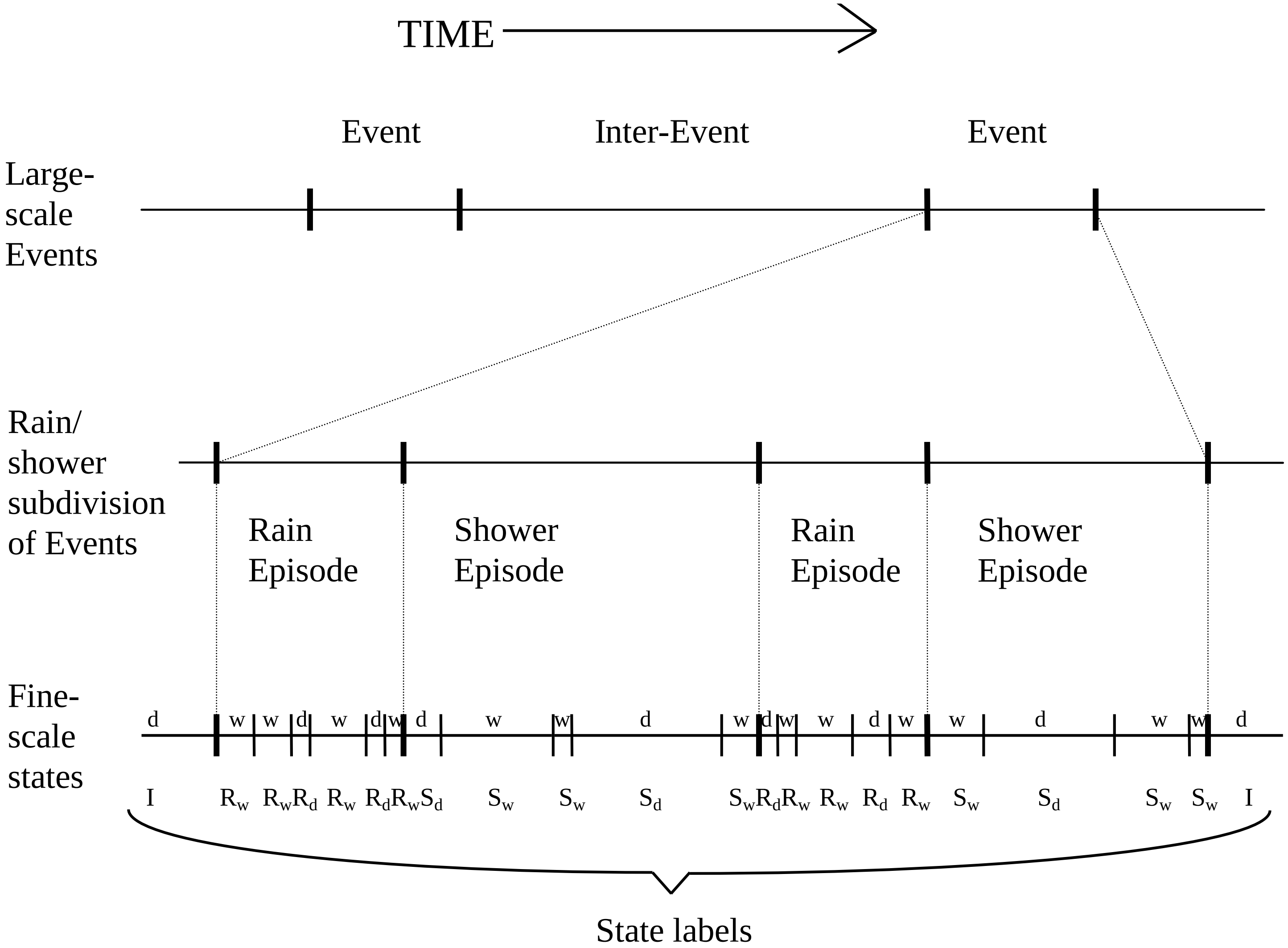

High resolution **breakpoint data** and daily accumulations.

## Breakpoint HSMM

Augments the data $\mathbf{Y}_k$ by precipitation states $S_k$ with the following hierarchy.

- Large-scale precipitation events and inter-event dry periods.
- Medium-scale rain and shower episodes within precipitation events.
- Fine-scale periods of steady rain or dry periods within episodes.

These states are generally hidden and must be inferred from the observed breakpoint data.



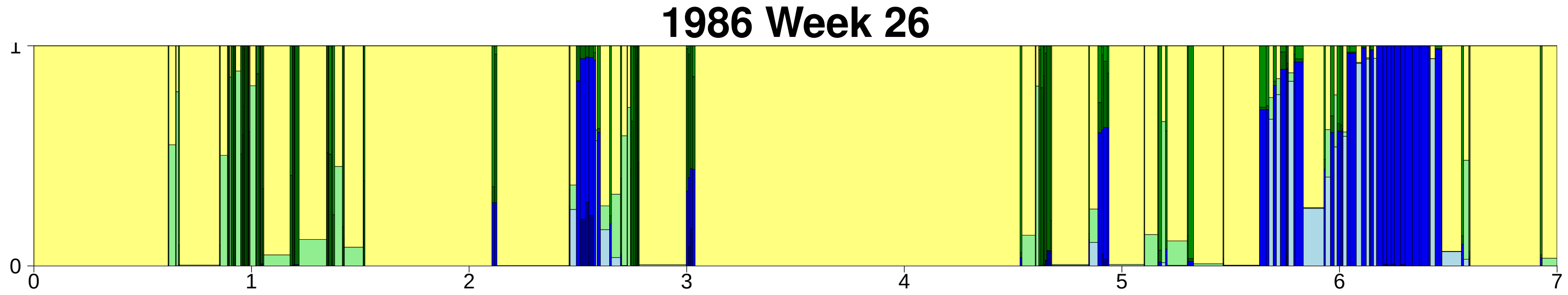

Hierarchical specification of (hidden) breakpoint rainfall states and their time scales.

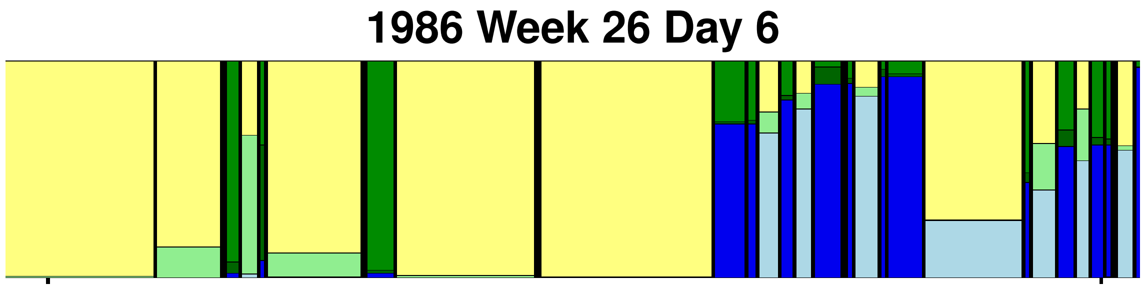
## For the breakpoint HSMM

- All rainfall dynamics are modelled by the states $S_k$ which are semi-Markov.
- Rainfall amounts are conditionally independent given the states.
- If all state sojourns are geometric the HSMM becomes a conventional HMM (hidden Markov model).
- Like the HMM, the HSMM is fitted using maximum likelihood and the EM algorithm.
- Key quantities are the data-determined classification probabilities

  $$\gamma_k(j) = P(S_k = j|\mathbf{Y}) \quad (k = 1, \ldots, K)$$

  where $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_K)$.
- For NZ data the HSMM has 7 states: 1 large-scale inter-event Dry state, 2 medium-scale states (Shower and Rain) each with 3 sub-states (light, heavy, intra-event dry).




Classification probabilities of the 7 states for one week of breakpoint rainfall data at Kelburn (Wellington, NZ) with time measured in days. Bar heights show the classification probabilities of the inter-event **Dry state**, **Shower state** and **Rain state**. The **increasing hues** of the Rain and Shower states correspond to **dry, light and heavy** precipitation respectively.

## NZ breakpoint data considered:

- 5 years from 1986 - 1990;
- 13 sites;
- 3 broad spatial scales.

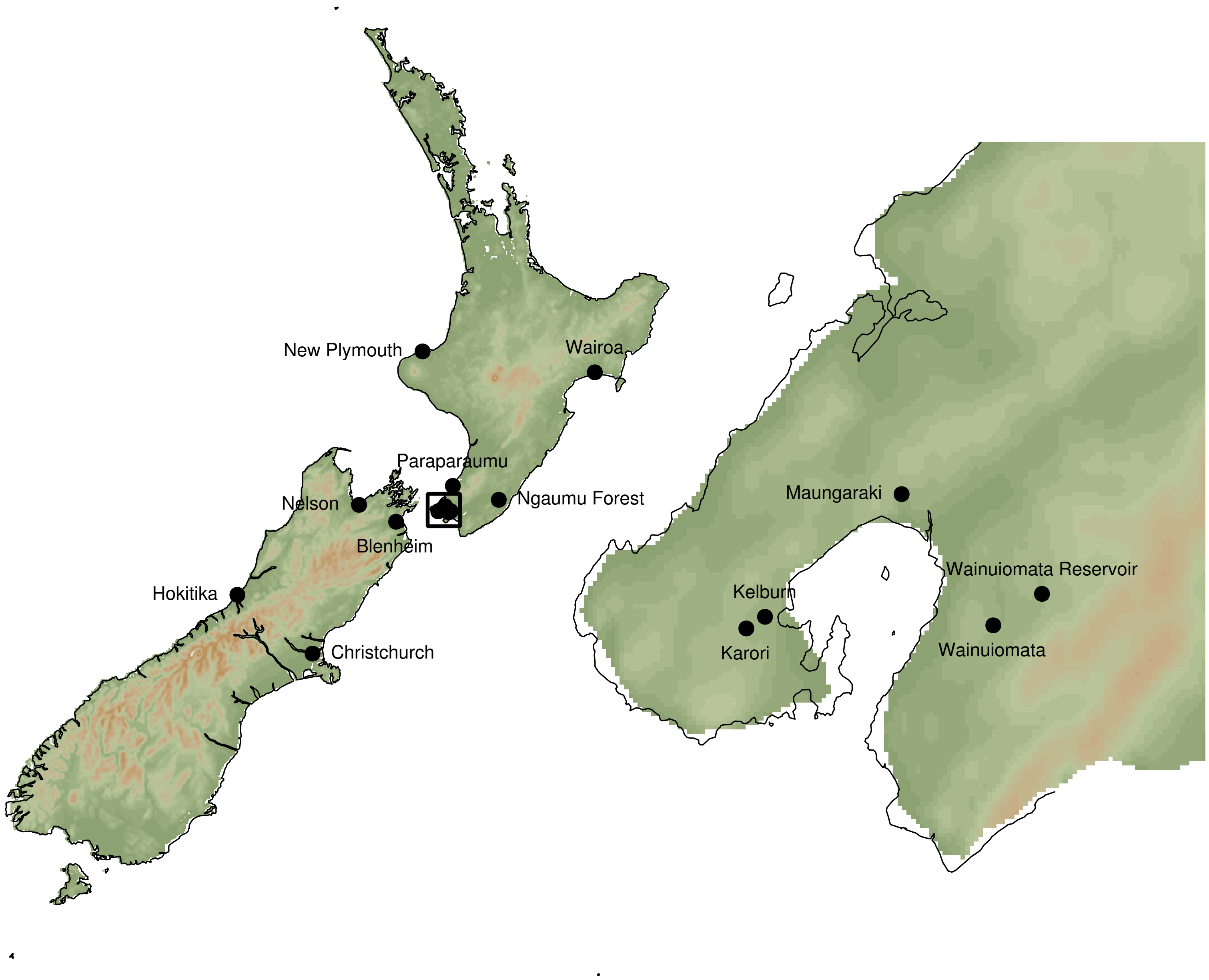
## Spatial scales/regions

**Local:** Kelburn, Karori, Maungaraki, Wainouiomata Reservoir, Wainuiomata.

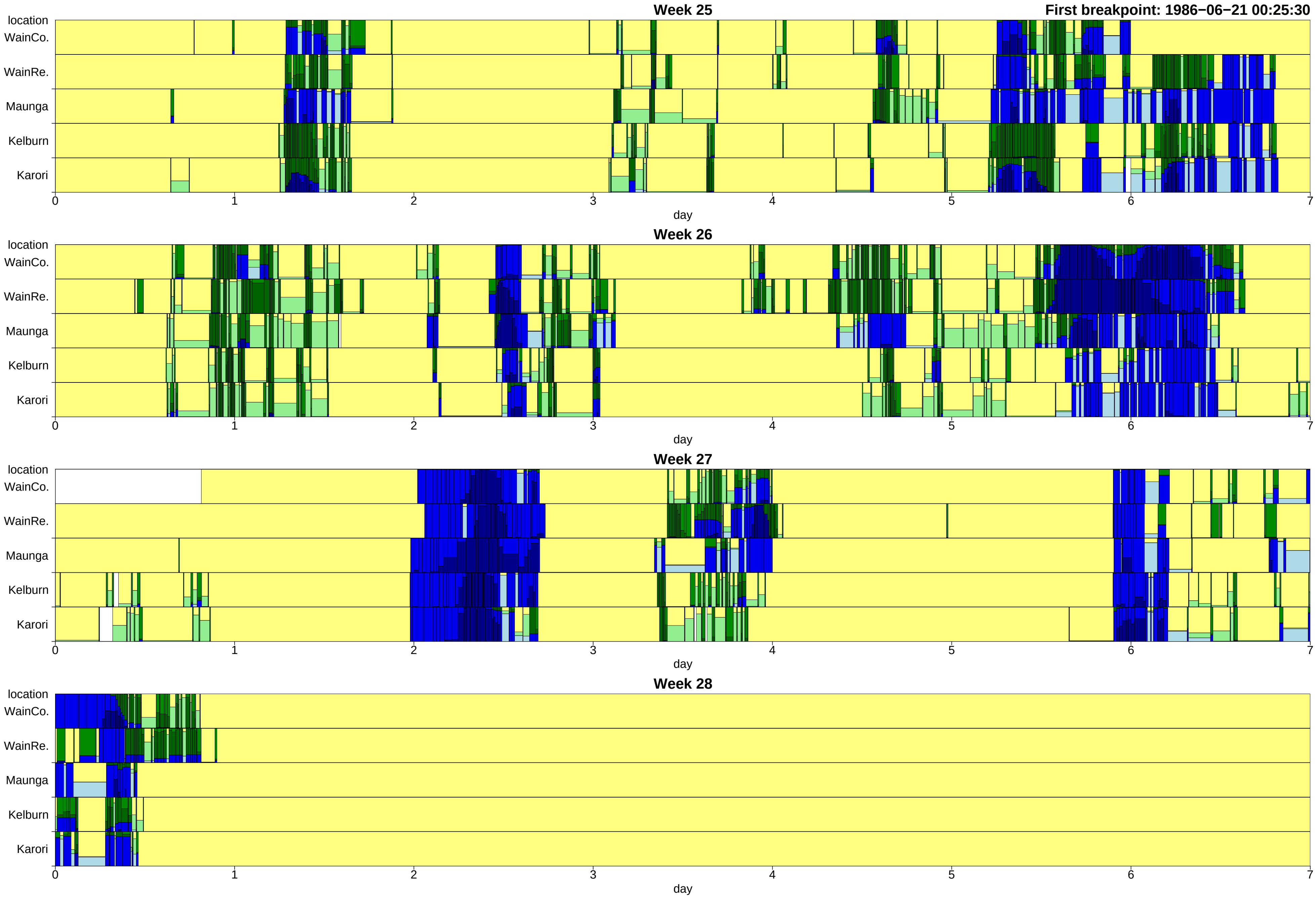
**Meso:** Kelburn, Ngaumu Forest, Paraparaumu, Blenheim, Nelson.

**Macro:** Kelburn, New Plymouth, Wairoa, Hokitika, Christchurch.

Kelburn is the common reference site with 5 sites per scale/region.

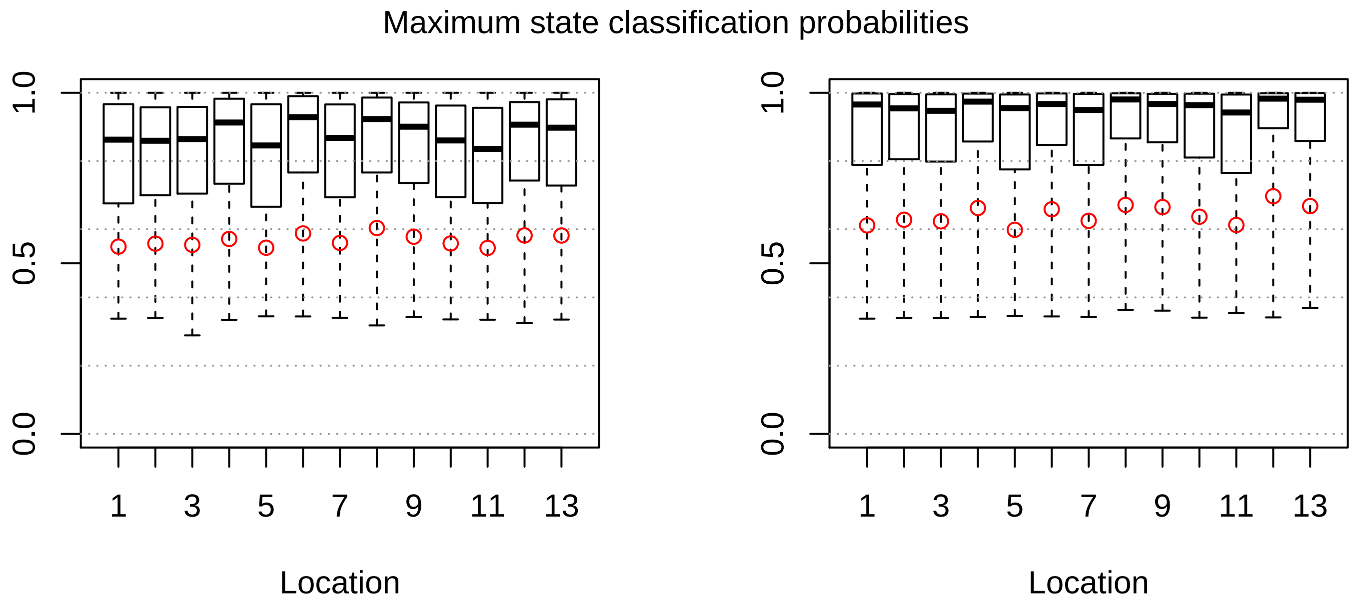



13 sites grouped into three spatial scales: **Local** (within 20km of Kelburn), **Meso** (around 100km from Kelburn and each other), **Macro** (around 200km from Kelburn and each other).




Classification probabilities of the 7 states for 4 weeks of breakpoint rainfall data at 5 sites in the **Local** region.

The HSMM provides an excellent and informative fit. The classification probabilities are generally unambiguous (definite) leading to secure state classifications that are physically meaningful.




Maximum state classification probability by location over 7 fine-scale states (**left**) and 3 synoptic-scale states Dry, Showers, Rain (**right**) with o the 10th percentile.

Classification probability plots show clear evidence of spatial coherence. However, comparisons across sites are difficult since breakpoints are site specific.

This leads us to determine classification probabilities over a **common uniform time scale** (minute, 10 minutes, hour, 6 hours, day etc).

## 3. Precipitation states over aggregated time and space scales

For any given site, divide time into equisaped intervals

$$E_n = [n\Delta, (n+1)\Delta)$$

of length $\Delta$ (minute, hour etc) and define the classification proportions

$$\widehat{\lambda}(j) = \frac{1}{\Delta} \int_{E_n} \bar{\gamma}_t(j) dt = \sum_{k=1}^{K} w_n(k) \gamma_k(j)$$

where

$\bar{\gamma}_t(j)$ = continuous-time classification probability for state $j$

$w_n(k)$ = proportion of $E_n$ falling in breakpoint interval $[T_k, T_{k+1})$.

The classification proportions $\widehat{\lambda}(j)$ are:

- weighted averages of the breakpoint state classification probabilities;
- discrete-time approximations of the continuous-time $\bar{\gamma}_t(j)$;
- readily calculated from the breakpoint data.

The accuracy of the approximation can be measured by

$$MSE_n(j) = \frac{1}{\Delta}\int_{E_n} (\bar{\gamma}_t(j) - \hat{\lambda}_n(j))^2 dt = \sum_{k=1}^{K} w_n(k)(\gamma_k(j) - \hat{\lambda}_n(j))^2$$

or its square-root (root-mean-square).

The definition of $\hat{\lambda}_n(j)$ can be extended to a spatial region (Local, Meso or Macro) to give

$$\hat{\lambda}_n(j) = \frac{1}{S}\sum_{s=1}^{S} \hat{\lambda}_n^{(s)}(j)$$

where $s$ indexes the sites in the region and the $\hat{\lambda}_n^{(s)}(j)$ are the single-site classification proportions given before.

In general, the $\hat{\lambda}_n(j)$ are:

- estimates of the proportion of time spent in state $j$ in time interval $E_n$ over all sites in the given region;
- useful for exploring the space-time dynamics of rainfall.

## 4. Single-site analysis

The HSMM vests all rainfall dynamics in the 7 precipitation states. So we focus exclusively on the stochastic properties of these states (**not** rainfall amounts).

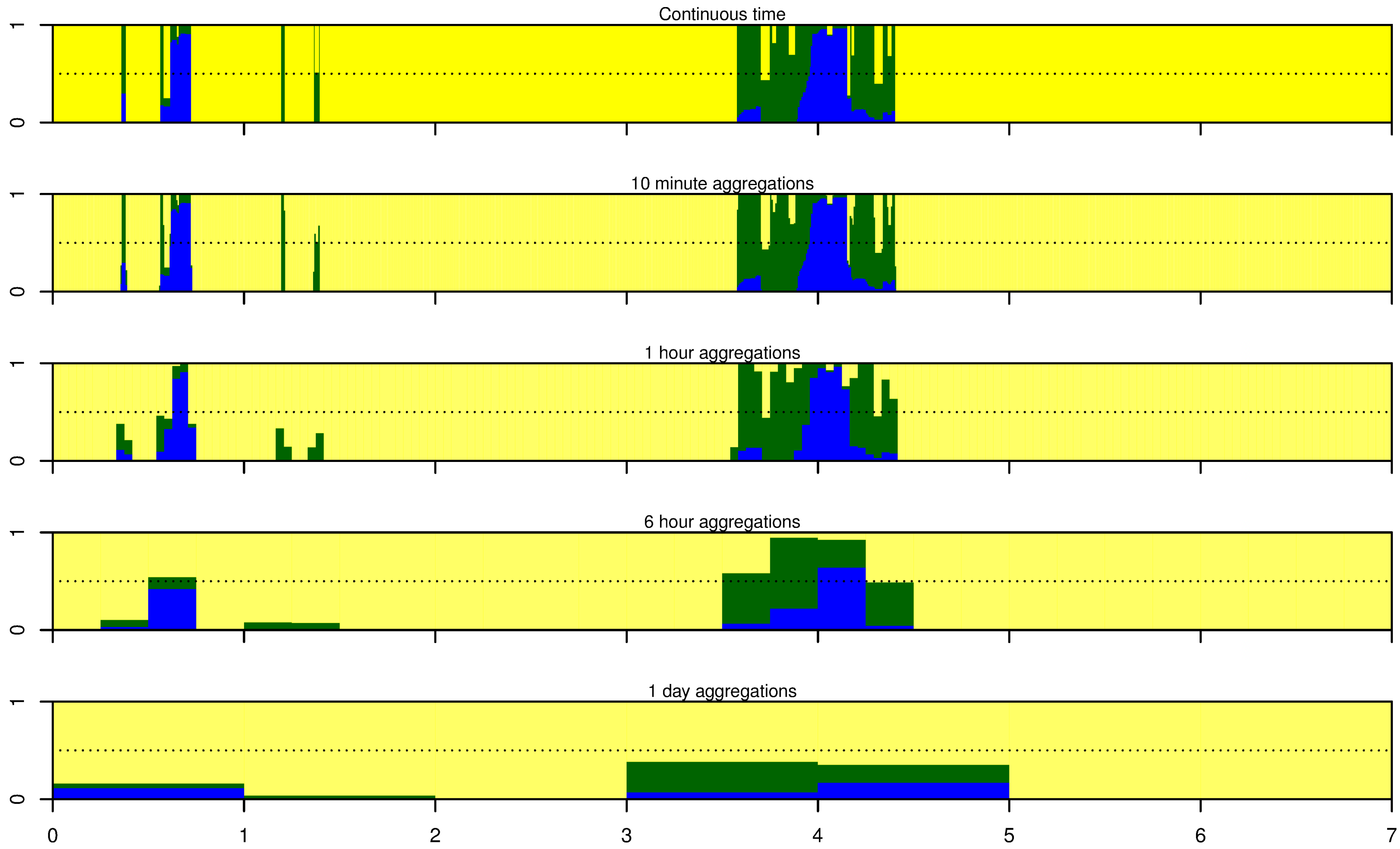
Furthermore, we restrict attention to the synoptic-scale states

**Dry**, **Shower**, **Rain**.

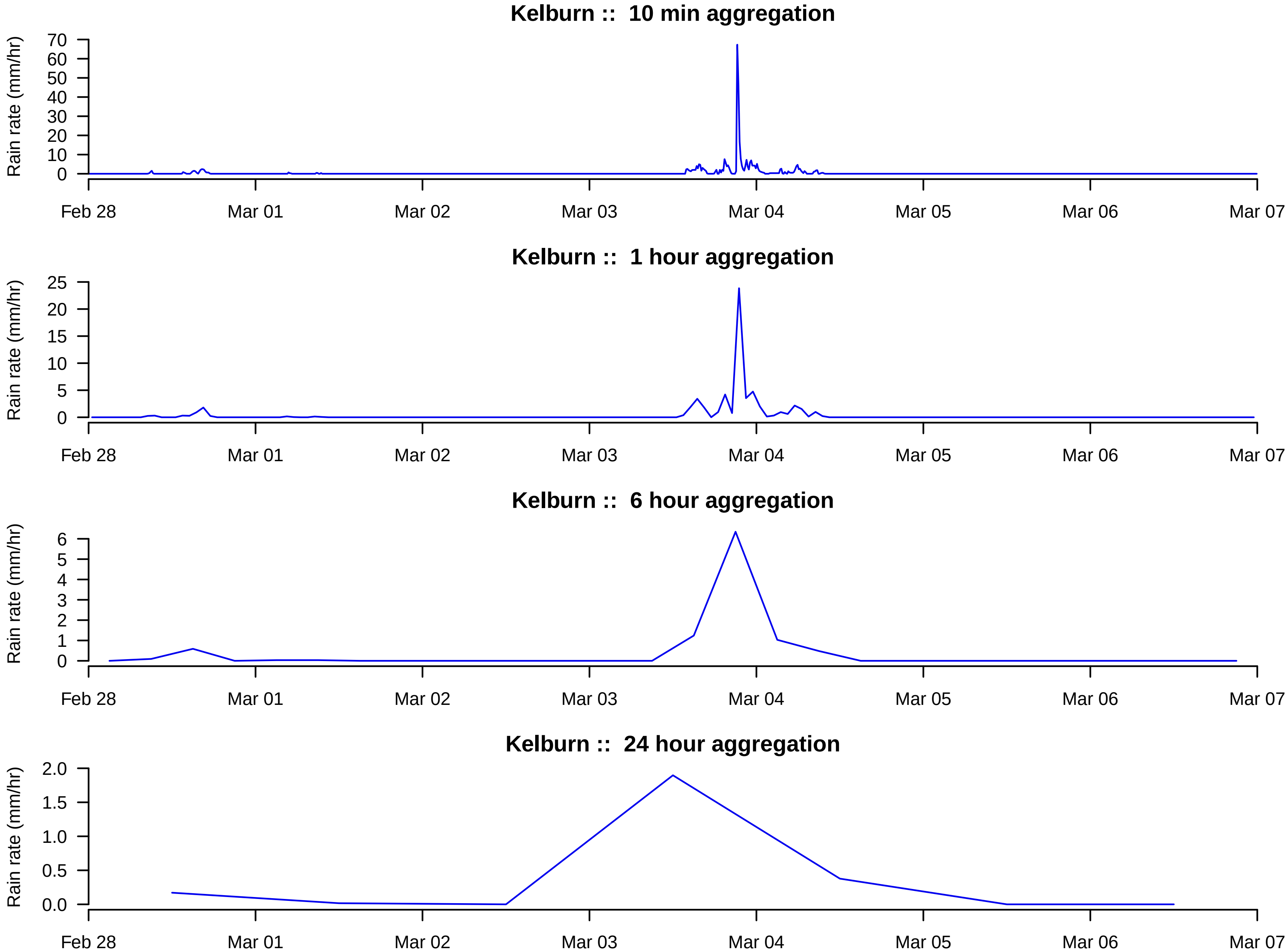
Although fine-scale states are also of interest, their properties are typically:

- location dependent since they control the shape of the rainfall mixture distribution at each site;
- difficult to recover from time series aggregations over coarser time scales.

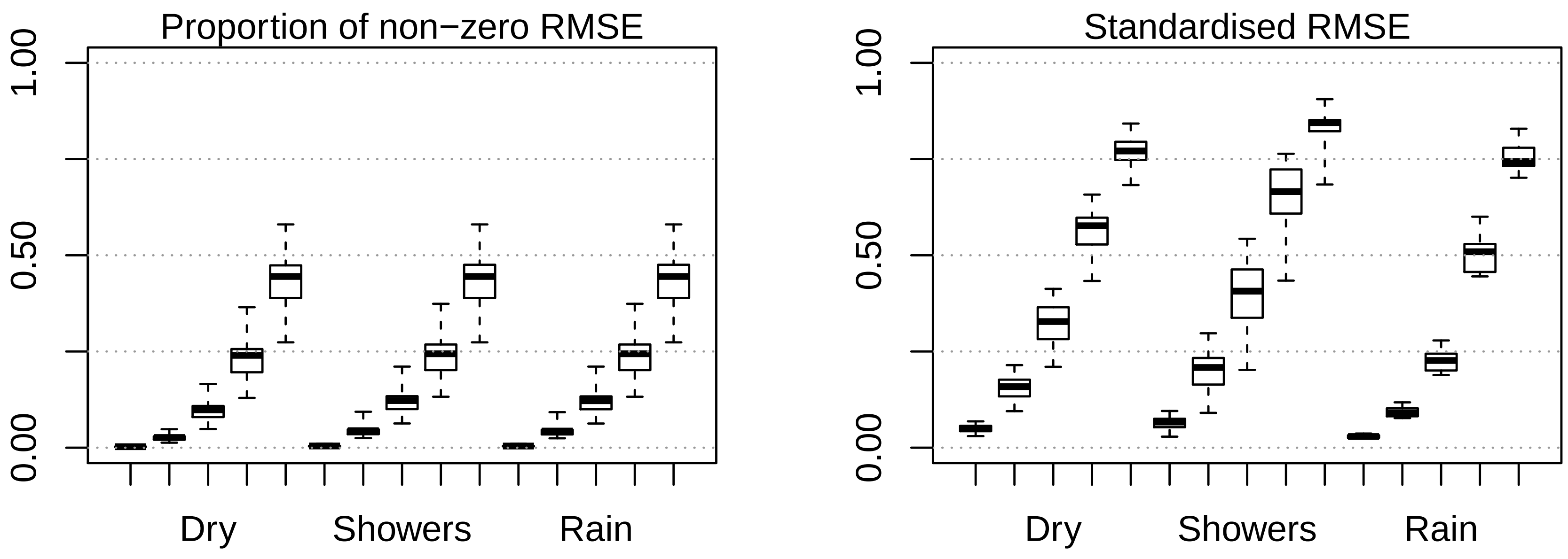
How well do the classification proportions approximate the underlying continuous-time classification probabilities?




**Top plot** shows the breakpoint classification probabilities for Kelburn over the week commencing 28 Feb 1986. **Remaining plots** show the classification proportions for 10 minute, hour, six hour and day time intervals.




**By comparison:** Kelburn rainfall accumulations for week commencing 28 Feb 1986. Note the impact of aggregation.




Boxplots of the proportions of non-zero RMSE values and standardised RMSE values for the approximation errors at each location.

- The approximation error is the difference between the breakpoint classification probability and the corresponding classification proportion;
- Within each state the increasing boxplots correspond to time intervals of 1 minute, 10 minute, 1 hour, 6 hour and 1 day.
- The standardised RMSE is the RMSE divided by the continuous-time standard deviation (worst case RMSE).

Now use the classification proportions to identify states and estimate state sojourns.

## State classification rules

For any given time interval

- Classify the state as Dry if

$$\textbf{Dry state classification proportion } > 0.5$$

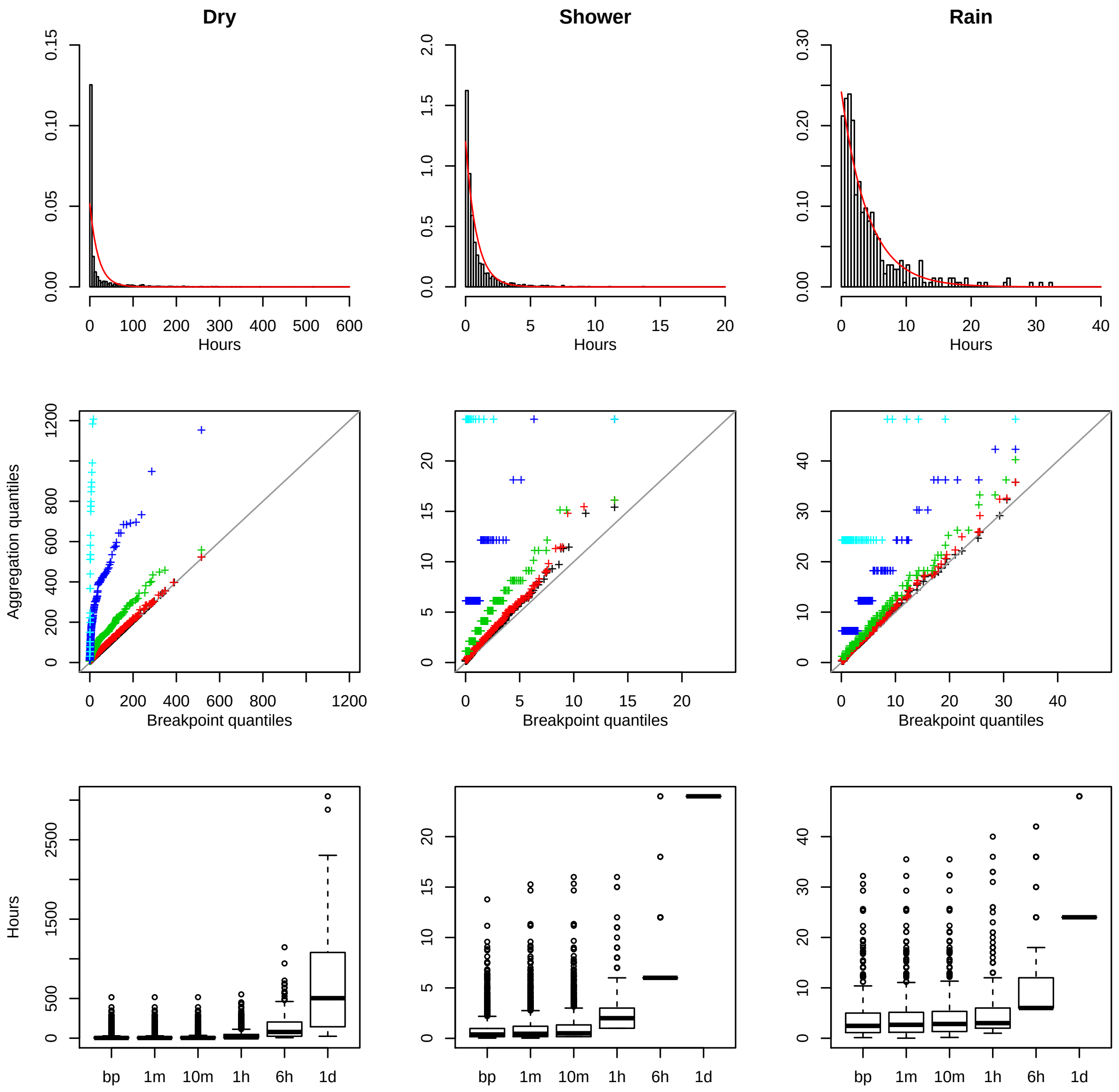
  and as a precipitation state (shower or rain) otherwise.

- Classify a precipitation state as a Shower state if

$$\textbf{Shower classification proportion} > \textbf{Rain classification proportion}$$

  and as a Rain state otherwise.

These simple rules respect the time-scale hierarchy of the states.




## Kelburn: estimated state sojourn distributions.

- **Top:** histograms of breakpoint sojourns with fitted exponentials for reference.
- **Middle:** Q-Q plots for 1 minute, 10 minute, 1 hour, 6 hour and 1 day sojourns versus breakpoint sojourns.
- **Bottom:** boxplots of breakpoint, 1 minute, 10 minute, 1 hour, 6 hour and 1 day sojourns.

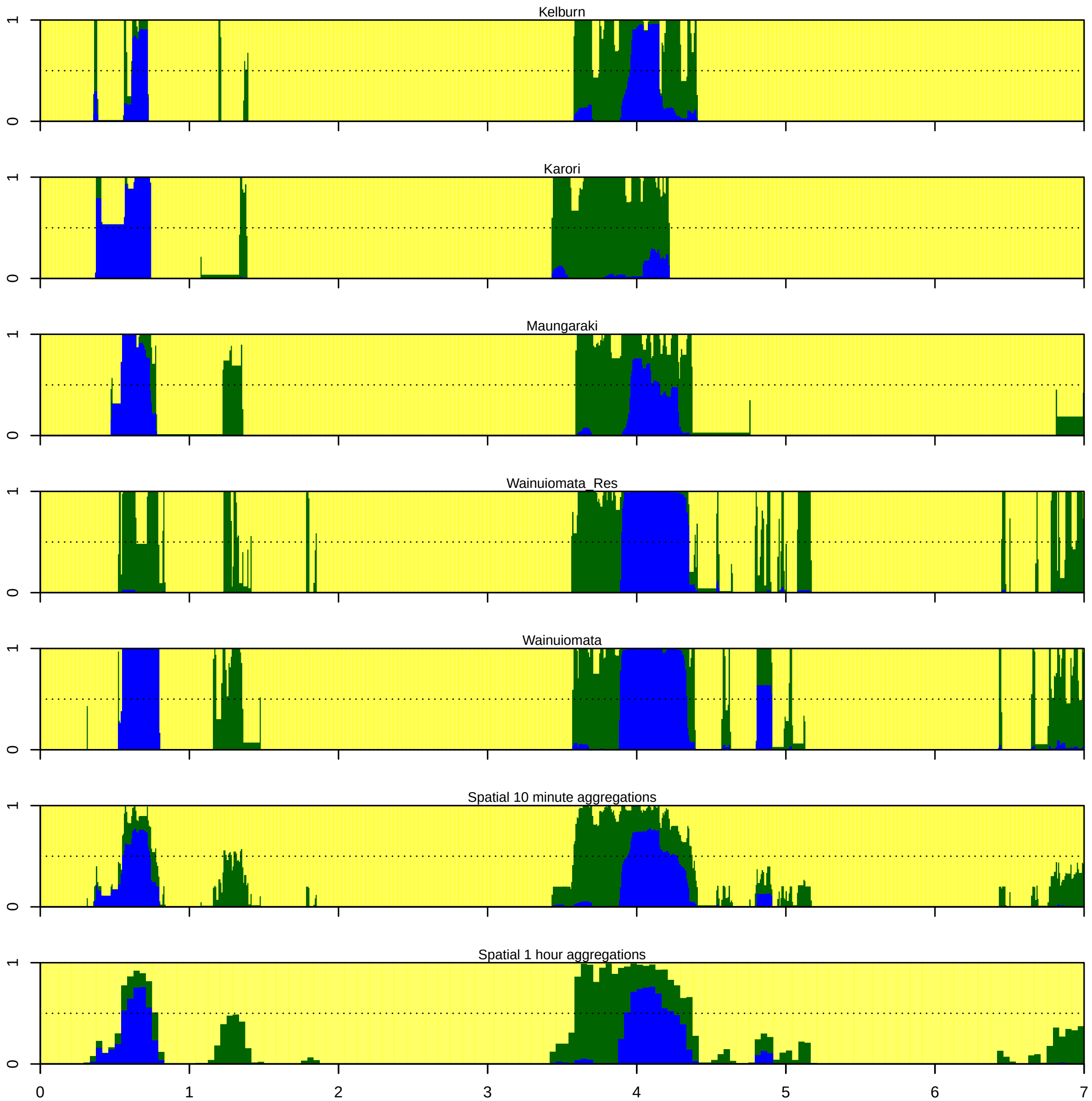
## 5. Multi-site analysis

Single-site analysis suggests time scales of at most 1 hour are required to approximate continuous-time rainfall dynamics. Consider 10 minute time-scale (1 hour similar) and spatial dynamics of regional rainfall states.

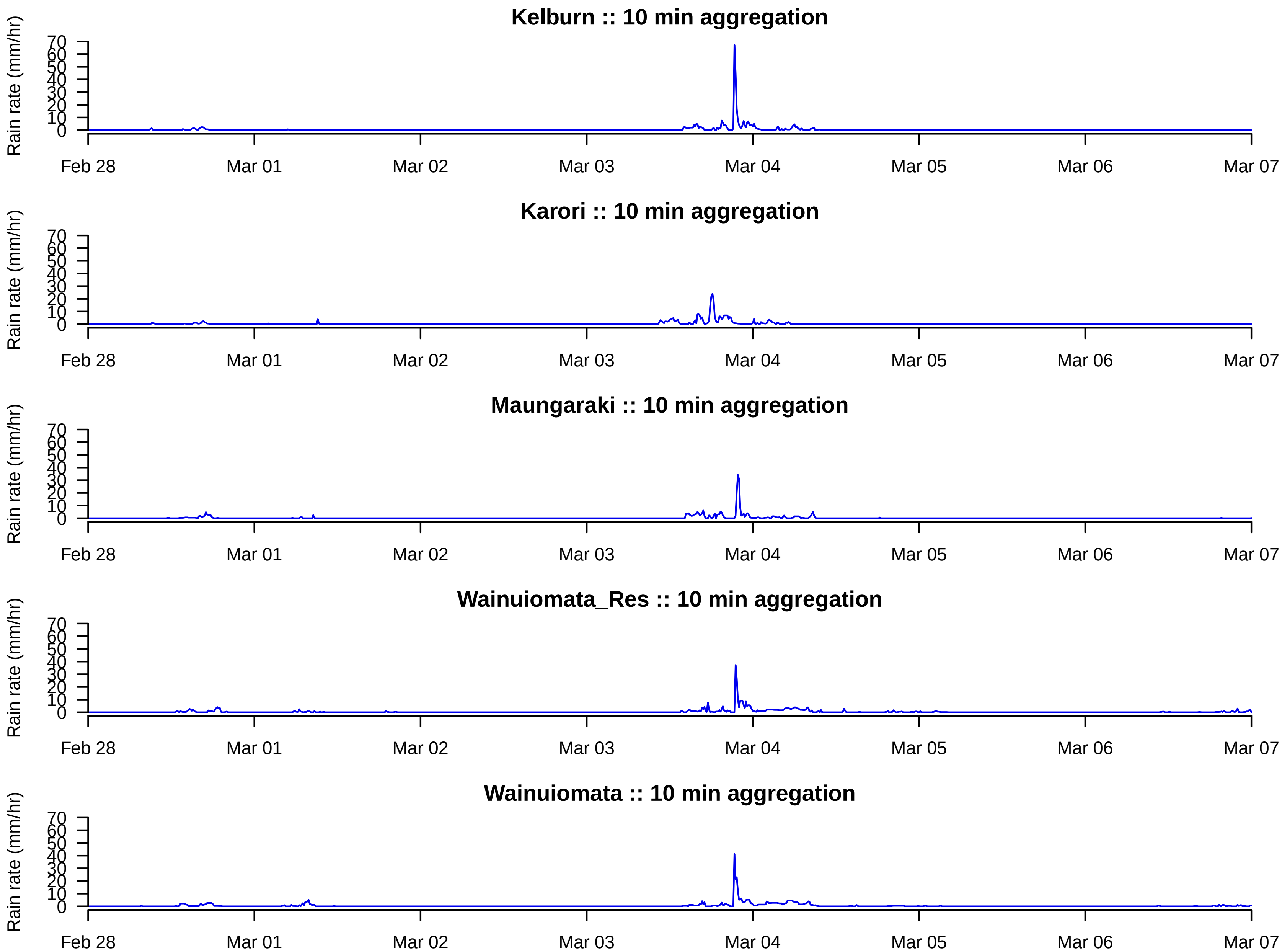
The regional classification proportions $\widehat{\lambda}_n(j)$ are space-time regional averages of state classification probabilities. To have meaning region should have:

- relatively homogeneous rainfall climatology;
- largely contemporaneous rainfall events.

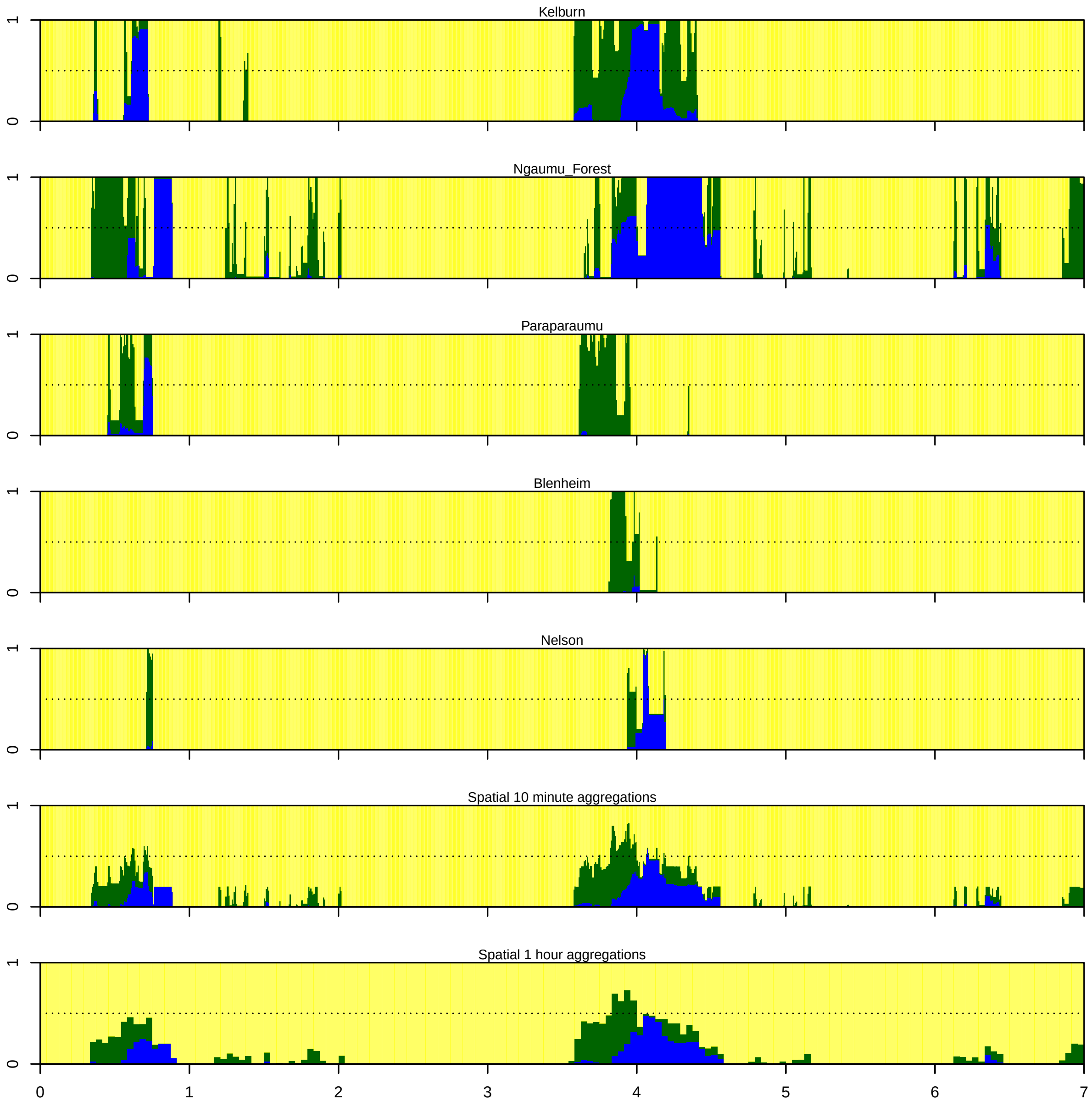
Spatial coherence is expected (especially for the Local region), but will become progressively weaker over coarser spatial scales. How much weaker?



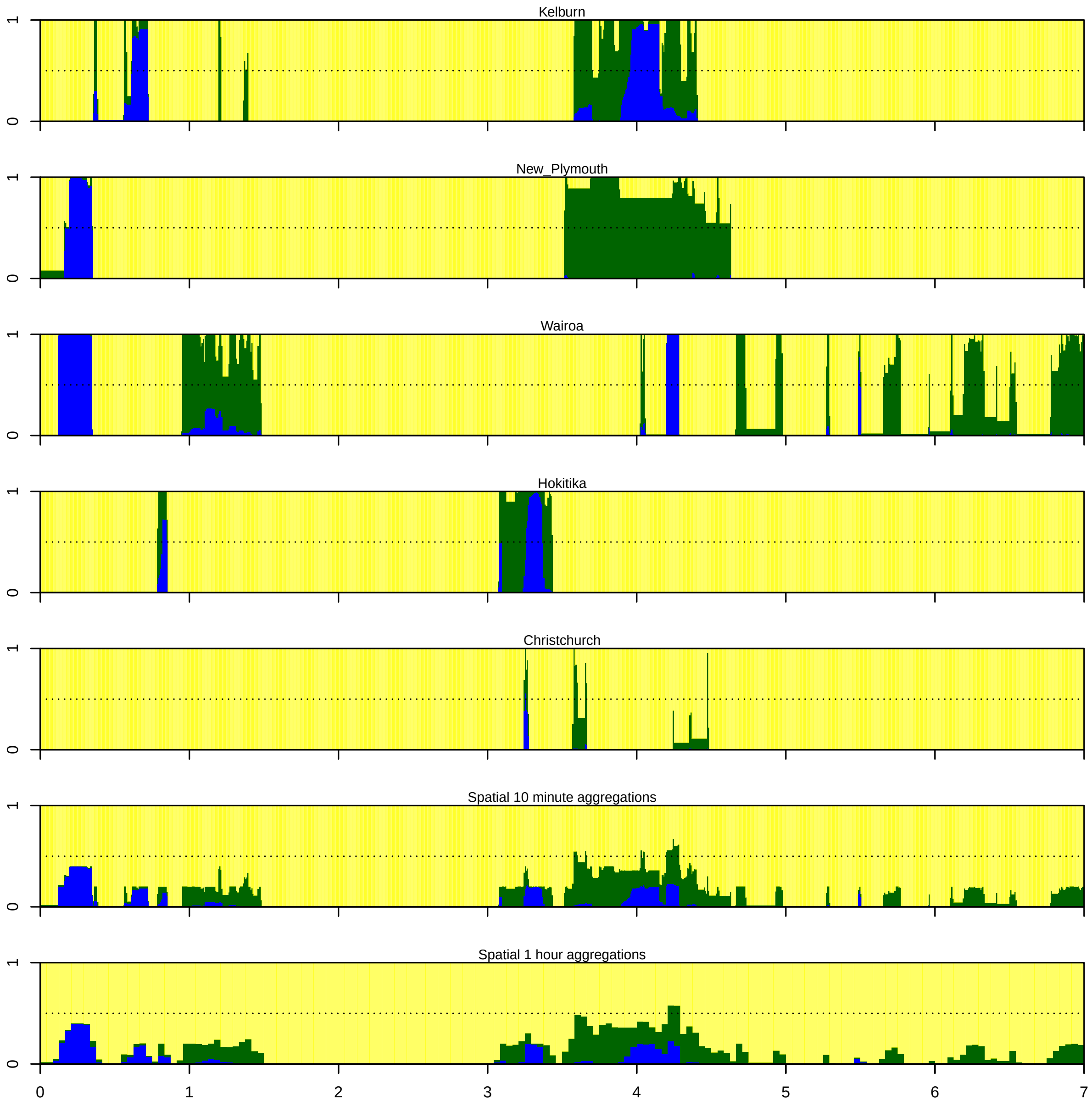

**Top 5 plots** show the classification proportions by site in the **Local region** at the 10 minute time scale over the week commencing 28 Feb 1986. **Bottom 2 plots** show the corresponding regional classification proportions for the 10 minute and 1 hour time scales.




**By comparison:** 10 minute rainfall accumulations by site in the **Local region** over the week commencing 28 Feb 1986.




**Top 5 plots** show the classification proportions by site in the **Meso region** at the 10 minute time scale over the week commencing 28 Feb 1986. **Bottom 2 plots** show the corresponding regional classification proportions for the 10 minute and 1 hour time scales.




**Top 5 plots** show the classification proportions by site in the **Macro region** at the 10 minute time scale over the week commencing 28 Feb 1986. **Bottom 2 plots** show the corresponding regional classification proportions for the 10 minute and 1 hour time scales.

Need to measure the spatial coherence between individual site classification proportions and a spatial average.

Let $Y_n$ be a series of site classification proportions and $X_n$ a spatial average. A measure of spatial coherence between $Y_n$ and $X_n$ is $D_{XY}$ where

$$D_{XY}^2 = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{Y_n - X_n}{s_{XY}}\right)^2 = \left(\frac{\bar{X}-\bar{Y}}{s_{XY}}\right)^2 + \left(\frac{s_X - s_Y}{s_{XY}}\right)^2 + 2(1-r_{XY})$$

and $s_{XY}^2$ is a pooled variance given by

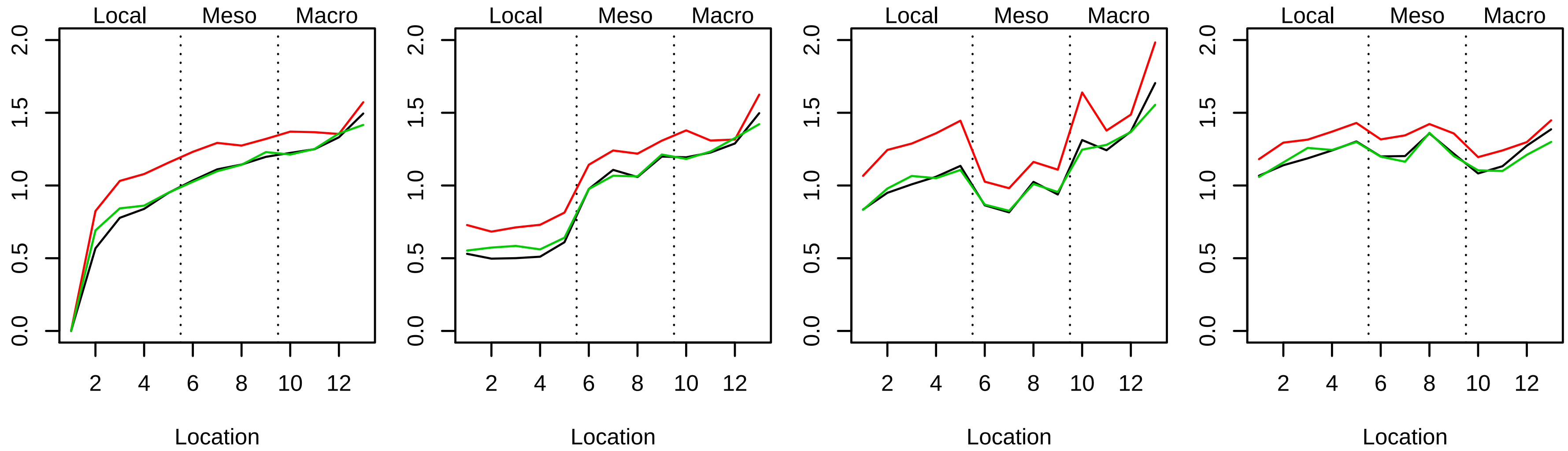
$$s_{XY}^2 = s_X s_Y = \text{geometric mean of } s_X^2 \text{ and } s_Y^2.$$

In our case $D_{XY}$ is largely influenced by the sample correlation $r_{XY}$.

Since $X_n$ and $Y_n$ are proportions

$$s_X^2 \quad = \quad \bar{X}(1-\bar{X}) - \frac{1}{N}\sum_{n=1}^{N} X_n(1-X_n) \quad \leq \quad \bar{X}(1-\bar{X})$$

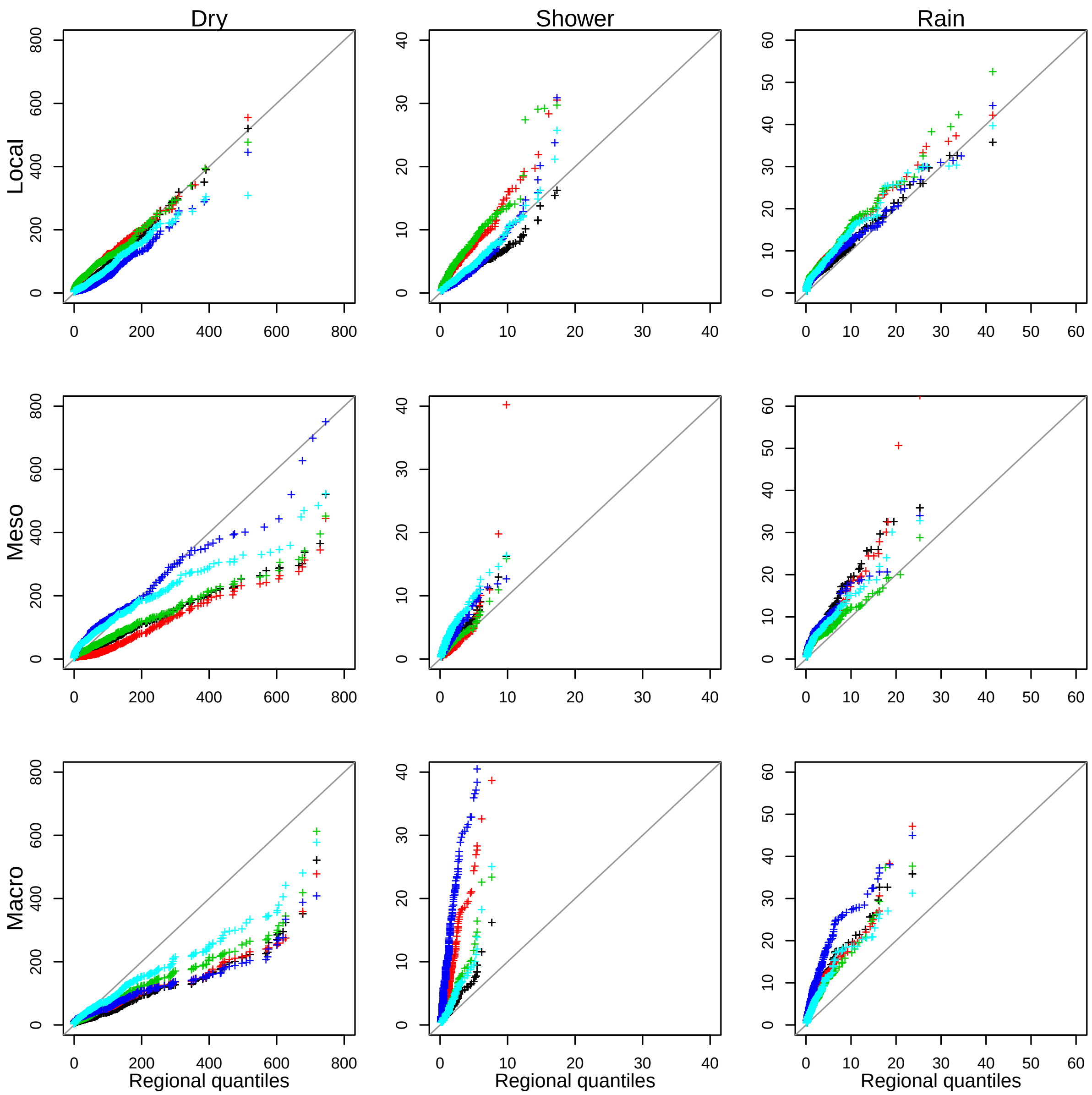
with a similar result for $s_Y^2$. Variances are greatest when proportions are 0 or 1 (Bernoulli) and are otherwise moderated by a measure of their uncertainty.




Plots of $D_{XY}$ comparing the state classification proportions by site at the 10 minute time scale with the corresponding (regional) state classification proportions for Kelburn reference site (**Left**), Local region (**Left centre**), Meso region (**Right centre**), Macro region (**Right**). Each plot shows the values of $D_{XY}$ for the **dry states**, shower states and rain states.

Note that:

- $D_{XY}$ for Kelburn reference site increases with distance from Kelburn;
- Local region highly coherent compared to Meso or Macro regions;
- Shower states always less coherent than either Dry or Rain states.




Q-Q plots of the state sojourns (Dry, Shower, Rain) at the 10 minute time scale for sites within a region (Local, Meso, Macro) against the regional state sojourns estimated from the regional state classification proportions. In each case the first site is the **Kelburn reference site**.

## 6. Conclusions

Analysis of continuous-time rainfall state classification probabilities and their space-time averages indicates that:

- aggregation intervals (time scales) of at most 1 hour are needed to approximate the continuous-time dynamics of rainfall;
- nearby sites (Local spatial scale) are strongly spatially coherent;
- spatial coherence is much weaker for Meso and Macro spatial scales.

Furthermore:

- the synoptic-state classification probabilities and proportions yield highly informative plots useful for rainfall modelling;
- if accurate and reliable modelling of continuous-time rainfall dynamics is the objective, then these results suggest that rainfall accumulations of 1 hour or better are necessary (not 6 hourly or daily).