Climate inference on Australian daily rainfall using distributed MCMC

Michael Bertolacci

PhD student, The University of Western Australian

Collaborators

Ed Cripps, UWA

John Lau, UWA

Sally Cripps, USyd

Ori Rosen, UTEP

Australian daily rainfall

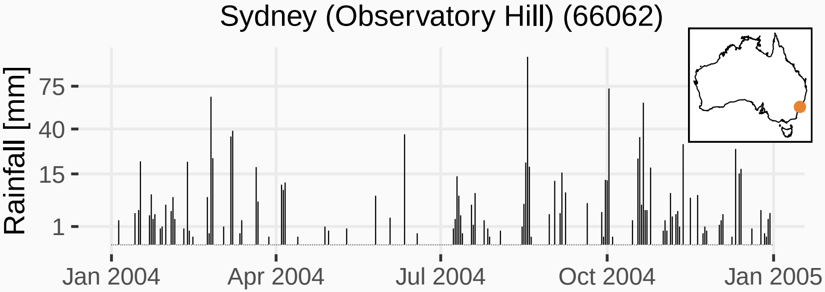
When you think of daily rainfall data, what do you think of?

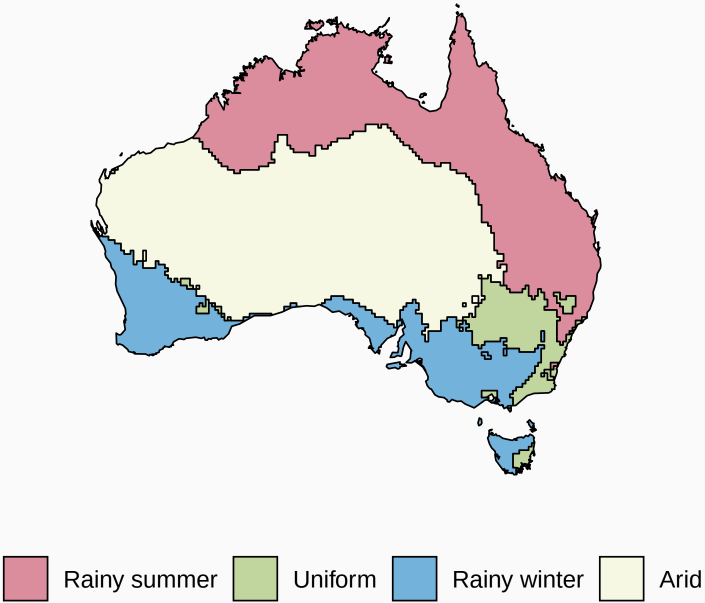
When you think of daily rainfall data, what do you think of?

Rain!

No rain

Rain

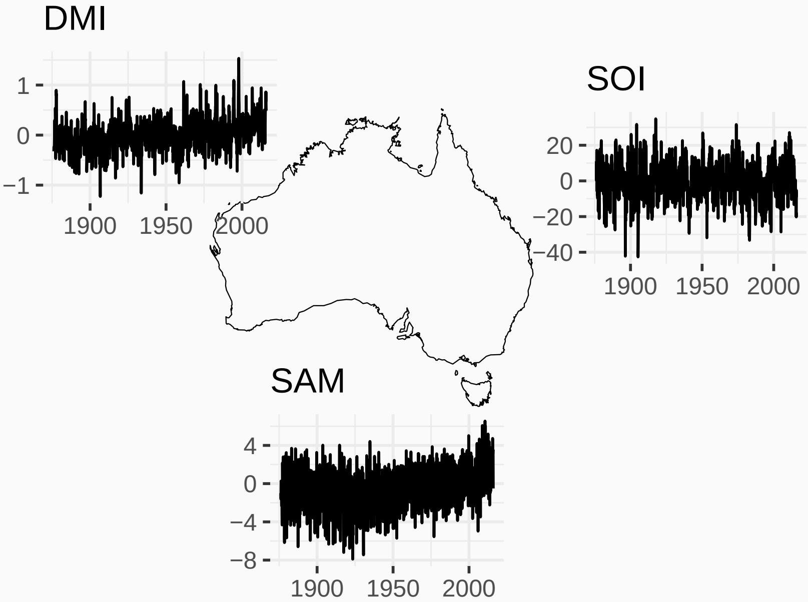

Sydney (Observatory Hill) (66062)
Rainfall [mm]
75
40
15
1
Jan 2004
Apr 2004
Jul 2004
Oct 2004
Jan 2005

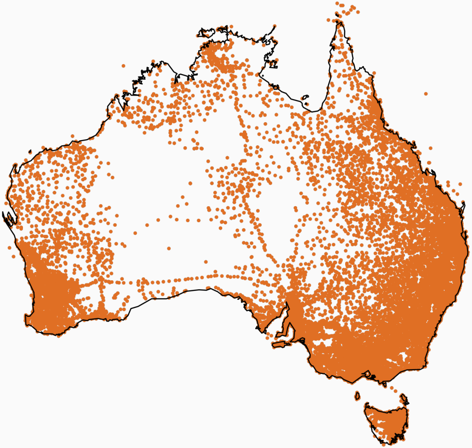

Rainy summer
Uniform
Rainy winter
Arid

Indian Ocean
Dipole

El Niño
Southern
Oscillation

Southern
Annular Mode


DMI
1
0
−1
1900
1950
2000
SOI
20
0
−20
−40
1900
1950
2000
SAM
4
0
−4
−8
1900
1950
2000



17,606 sites. Between 89 and 51,102 observations per site. Total 294 million observations.

Model

The Bayesian framework:

- Likelihood
- Prior

$$\boxed{p(\theta \mid y)} \propto \boxed{p(y \mid \theta)}\,\boxed{p(\theta)}$$

- Posterior

Spatial (between-site)

Latitude, longitude

$$\boldsymbol{w}_s$$

$$\boldsymbol{\sigma} \qquad \boldsymbol{\mu}_{s,k,\cdot} \sim GP(m(\boldsymbol{w}), k(\boldsymbol{w}, \boldsymbol{w}')) \qquad \boldsymbol{a}, \boldsymbol{b}$$

Temporal (within-site)

Climate drivers, seasonality

$$\boldsymbol{x}_{t,s} \qquad \boldsymbol{\beta}_{s,k,\cdot} \sim N(\boldsymbol{\mu}_{s,k,\cdot}, \sigma_k^2)$$

$$\operatorname{logit} \pi_{t,s,k} = \frac{\boldsymbol{x}'_{t,s}\boldsymbol{\beta}_{s,k,\cdot}}{\sum_{k'=0}^{K} \boldsymbol{x}'_{t,s}\boldsymbol{\beta}_{s,k',\cdot}}$$

$$y_{t,s} \sim \pi_{t,s,0}\delta(0) + \sum_{k=1}^{K} \pi_{t,s,k}\mathrm{Ga}(a_k, b_k)$$

Spatial (between-site)

Latitude, longitude

$\boldsymbol{w}_s$

$\boldsymbol{\sigma}$ $\quad \boldsymbol{\mu}_{s,k,\cdot} \sim GP(m(\boldsymbol{w}), k(\boldsymbol{w}, \boldsymbol{w}'))$ $\quad \boldsymbol{a}, \boldsymbol{b}$

Temporal (within-site)

Climate drivers, seasonality

$\boldsymbol{x}_{t,s}$

$$\boldsymbol{\beta}_{s,k,\cdot} \sim N(\boldsymbol{\mu}_{s,k,\cdot}, \sigma_k^2)$$

$$\operatorname{logit} \pi_{t,s,k} = \frac{\boldsymbol{x}_{t,s}'\boldsymbol{\beta}_{s,k,\cdot}}{\sum_{k'=0}^{K} \boldsymbol{x}_{t,s}'\boldsymbol{\beta}_{s,k',\cdot}}$$

$$z_{t,s}$$

$$(y_{t,s} \mid z_{t,s} = k) \sim \begin{cases} \delta(0) & \text{if } k = 0 \\ \mathrm{Ga}(a_k, b_k) & \text{otherwise} \end{cases}$$

Estimation

MCMC for $i = 1, \dots, N$ iterations:

1. Sample $(a_k, b_k \mid \boldsymbol{\beta}_{\cdot,k,\cdot}, \boldsymbol{z}, \boldsymbol{y})$ for $k = 1, \dots, K$
2. Sample $(\sigma_k^2 \mid \boldsymbol{\beta}_{\cdot,k,\cdot}, \boldsymbol{\mu}_{\cdot,k,\cdot}, \boldsymbol{y})$ for $k = 1, \dots, K$
3. Sample $(\boldsymbol{\mu}_{\cdot,k,\cdot} \mid \boldsymbol{\beta}_{\cdot,k,\cdot}, \sigma_k^2)$ for $k = 1, \dots, K$
4. Sample $(\boldsymbol{\beta}_{s,k,\cdot} \mid \boldsymbol{\mu}_{s,k}, \boldsymbol{z}_s, \sigma_k^2)$ for $s = 1, \dots, S$ and $k = 1, \dots, K$
5. Sample $(z_{t,s} \mid \boldsymbol{\beta}_{s,\cdot,\cdot}, y_{t,s})$ for $t = 1, \dots, T$ and $s = 1, \dots, S$

MCMC for $i = 1, \ldots, N$ iterations:

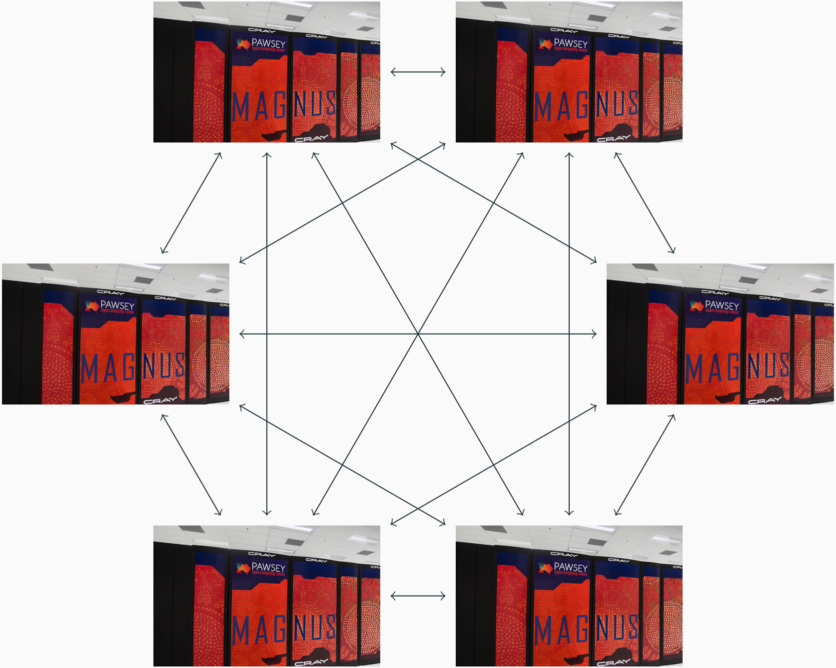
1. Sample $(a_k, b_k \mid \boldsymbol{\beta}_{\cdot,k,\cdot}, \boldsymbol{z}, \boldsymbol{y})$ for $k = 1, \ldots, K$
2. Sample $(\sigma_k^2 \mid \boldsymbol{\beta}_{\cdot,k,\cdot}, \boldsymbol{\mu}_{\cdot,k,\cdot}, \boldsymbol{y})$ for $k = 1, \ldots, K$
3. Sample $(\boldsymbol{\mu}_{\cdot,k,\cdot} \mid \boldsymbol{\beta}_{\cdot,k,\cdot}, \sigma_k^2)$ for $k = 1, \ldots, K$
4. Sample $(\boldsymbol{\beta}_{s,k,\cdot} \mid \boldsymbol{\mu}_{s,k}, \boldsymbol{z}_s, \sigma_k^2)$ for $s = 1, \ldots, S$ and $k = 1, \ldots, K$
5. Sample $(z_{t,s} \mid \boldsymbol{\beta}_{s,\cdot,\cdot}, y_{t,s})$ for $t = 1, \ldots, T$ and $s = 1, \ldots, S$

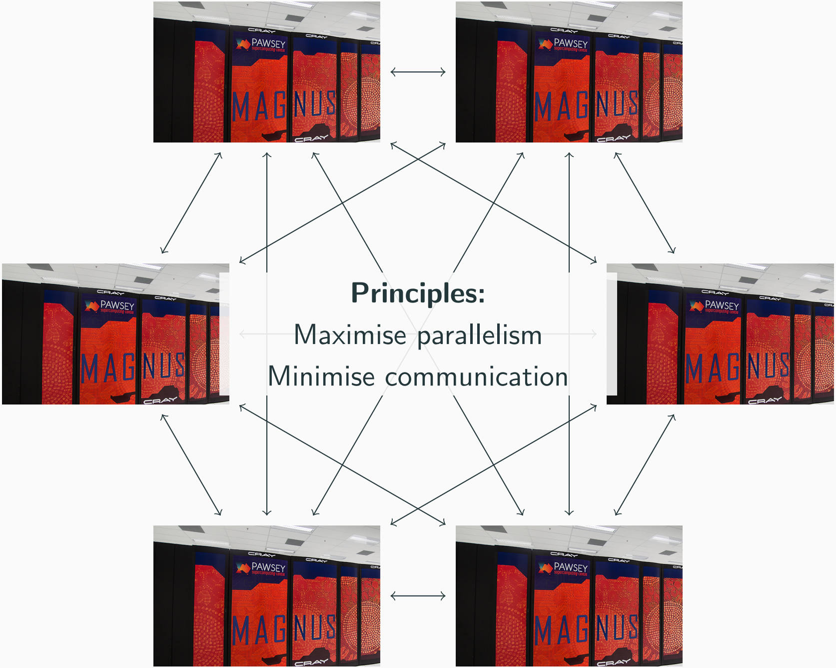
$$
\begin{aligned}
N &= 40,000 \text{ iterations} \\
S &= 17,606 \text{ sites} \\
T &\approx 17,000 \text{ per site} \\
ST &\approx 300,000,000 \text{ observations / latent variables} \\
NST &\approx 12,000,000,000,000 \text{ variables to sample}
\end{aligned}
$$

☹


WE'RE GONNA NEED
A BIGGER BOAT.


CRAY
PAWSEY
supercomputing centre
MAGNUS
CRAY
CRAY
CRAY
CRAY


CRAY
PAWSEY
MAGNUS
CRAY



Principles:

Maximise parallelism

Minimise communication

Step:

Initialise

Master node

a_k, b_k

σ_k

$\boldsymbol{\mu}_{\cdot,k,\cdot}$

Node 1		Node 2		Node 3	
$s = 1$	$\boldsymbol{\beta}_{s,k,\cdot}$	$s = 4$	$\boldsymbol{\beta}_{s,k,\cdot}$	$s = 7$	$\boldsymbol{\beta}_{s,k,\cdot}$
2	$z_{t,s}$	5	$z_{t,s}$	8	$z_{t,s}$
3	$y_{t,s}$	6	$y_{t,s}$	9	$y_{t,s}$

Step:

Sample

Master node

a_k, b_k

σ_k

$\boldsymbol{\mu}_{\cdot,k,\cdot}$

$s = 1$	$\boldsymbol{\beta}_{s,k,\cdot}$	$s = 4$	$\boldsymbol{\beta}_{s,k,\cdot}$	$s = 7$	$\boldsymbol{\beta}_{s,k,\cdot}$
2	$z_{t,s}$	5	$z_{t,s}$	8	$z_{t,s}$
3	$y_{t,s}$	6	$y_{t,s}$	9	$y_{t,s}$
Node 1		Node 2		Node 3	

Step:

Sample

Master node

a_k, b_k

$\boxed{\sigma_k}$

$\boldsymbol{\mu}_{\cdot,k,\cdot}$

$s = 1$	$\boldsymbol{\beta}_{s,k,\cdot}$
2	$z_{t,s}$
3	$y_{t,s}$

Node 1

$s = 4$	$\boldsymbol{\beta}_{s,k,\cdot}$
5	$z_{t,s}$
6	$y_{t,s}$

Node 2

$s = 7$	$\boldsymbol{\beta}_{s,k,\cdot}$
8	$z_{t,s}$
9	$y_{t,s}$

Node 3

Step:

Sample

Master node

a_k, b_k

σ_k

$\boldsymbol{\mu}_{\cdot,k,\cdot}$

$s = 1$	$\boldsymbol{\beta}_{s,k,\cdot}$
2	$z_{t,s}$
3	$y_{t,s}$

Node 1

$s = 4$	$\boldsymbol{\beta}_{s,k,\cdot}$
5	$z_{t,s}$
6	$y_{t,s}$

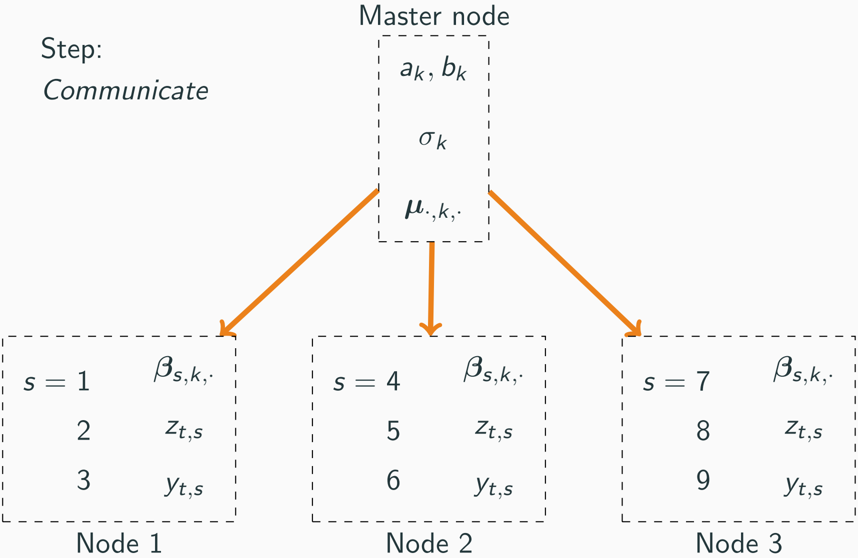
Node 2

$s = 7$	$\boldsymbol{\beta}_{s,k,\cdot}$
8	$z_{t,s}$
9	$y_{t,s}$

Node 3

Step:

Communicate



Step:

Sample

Master node

a_k, b_k

σ_k

$\boldsymbol{\mu}_{\cdot,k,\cdot}$

$s = 1$	$\boldsymbol{\beta}_{s,k,\cdot}$
2	$z_{t,s}$
3	$y_{t,s}$

Node 1

$s = 4$	$\boldsymbol{\beta}_{s,k,\cdot}$
5	$z_{t,s}$
6	$y_{t,s}$

Node 2

$s = 7$	$\boldsymbol{\beta}_{s,k,\cdot}$
8	$z_{t,s}$
9	$y_{t,s}$

Node 3

Step:

Sample

Master node

a_k, b_k

σ_k

$\boldsymbol{\mu}_{\cdot,k,\cdot}$

$s = 1$	$\boldsymbol{\beta}_{s,k,\cdot}$
2	$z_{t,s}$
3	$y_{t,s}$

Node 1

$s = 4$	$\boldsymbol{\beta}_{s,k,\cdot}$
5	$z_{t,s}$
6	$y_{t,s}$

Node 2

$s = 7$	$\boldsymbol{\beta}_{s,k,\cdot}$
8	$z_{t,s}$
9	$y_{t,s}$

Node 3

Step:

Sample

Master node

a_k, b_k

σ_k

$\boldsymbol{\mu}_{\cdot,k,\cdot}$

$s = 1$	$\boldsymbol{\beta}_{s,k,\cdot}$	$s = 4$	$\boldsymbol{\beta}_{s,k,\cdot}$	$s = 7$	$\boldsymbol{\beta}_{s,k,\cdot}$
2	$z_{t,s}$	5	$z_{t,s}$	8	$z_{t,s}$
3	$y_{t,s}$	6	$y_{t,s}$	9	$y_{t,s}$
Node 1		Node 2		Node 3	

Step:

Sample

Master node

a_k, b_k

σ_k

$\boldsymbol{\mu}_{\cdot,k,\cdot}$

$s = 1$	$\boldsymbol{\beta}_{s,k,\cdot}$
2	$\boxed{z_{t,s}}$
3	$y_{t,s}$

Node 1

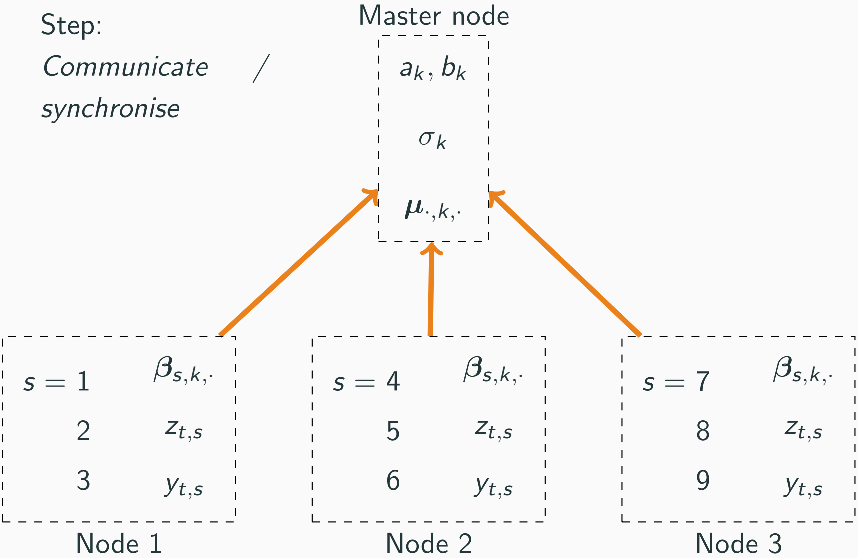
$s = 4$	$\boldsymbol{\beta}_{s,k,\cdot}$
5	$\boxed{z_{t,s}}$
6	$y_{t,s}$

Node 2

$s = 7$	$\boldsymbol{\beta}_{s,k,\cdot}$
8	$\boxed{z_{t,s}}$
9	$y_{t,s}$

Node 3

Step:
Communicate / synchronise



Master node

Step:

Sample

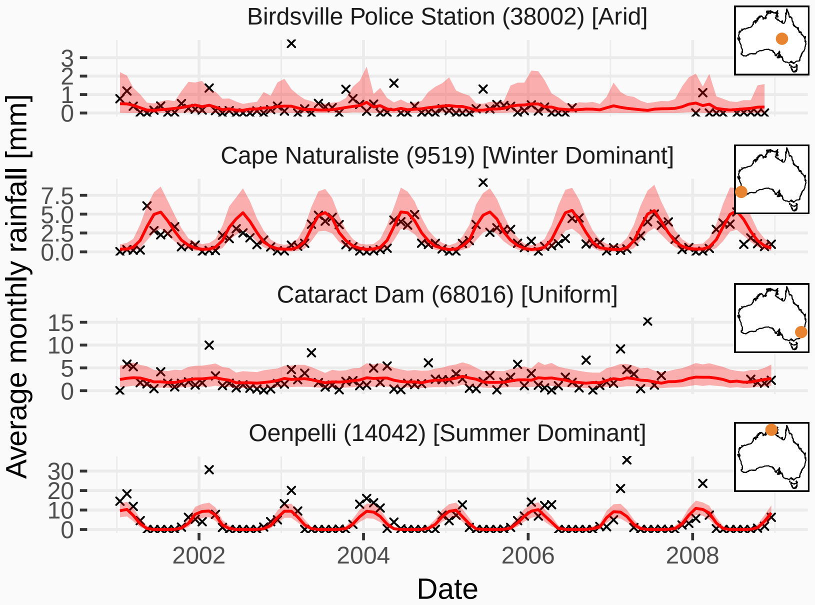
a_k, b_k

σ_k

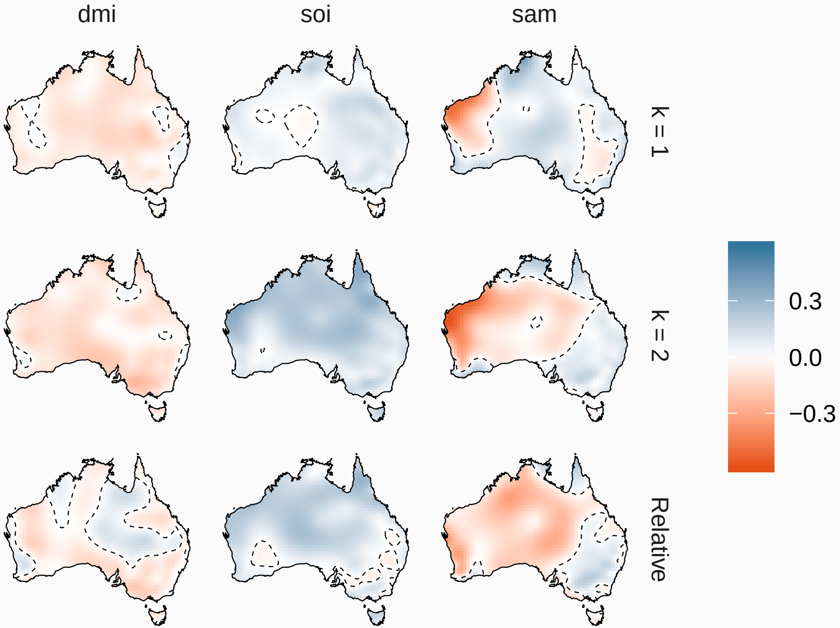
$\boldsymbol{\mu}_{\cdot,k,\cdot}$

Node 1		Node 2		Node 3	
$s = 1$	$\boldsymbol{\beta}_{s,k,\cdot}$	$s = 4$	$\boldsymbol{\beta}_{s,k,\cdot}$	$s = 7$	$\boldsymbol{\beta}_{s,k,\cdot}$
2	$z_{t,s}$	5	$z_{t,s}$	8	$z_{t,s}$
3	$y_{t,s}$	6	$y_{t,s}$	9	$y_{t,s}$

Results



Birdsville Police Station (38002) [Arid]
Cape Naturaliste (9519) [Winter Dominant]
Cataract Dam (68016) [Uniform]
Oenpelli (14042) [Summer Dominant]
Average monthly rainfall [mm]
Date
0
1
2
3
0.0
2.5
5.0
7.5
0
5
10
15
0
10
20
30
2002
2004
2006
2008



dmi
soi
sam
k = 1
k = 2
Relative
0.3
0.0
−0.3

Software:

- bomdata: `https://github.com/mbertolacci/bomdata/`
- climatedata: `https://github.com/mbertolacci/climatedata/`
- storm: `https://github.com/mbertolacci/storm/`

Paper:

Bertolacci, M., Cripps, E., Rosen, O., Lau, J. W., and Cripps, S. (2019), "Climate inference on daily rainfall across the Australian continent, 1876–2015," Annals of Applied Statistics, 13, 683–712.

Supported by resources provided by the Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia.