A spatial Markov model for climate extremes

Brian Reich (NCSU) and Ben Shaby (PSU)

October 25, 2016

Spatial extremes

- EVA can benefit greatly from spatial methods
- Spatial methods can map risk and borrow strength over space to estimate rare-event probabilities
- Accounting for spatial dependence is necessary for valid inference
- Methods and software in this area are developing rapidly to meet a growing demand

Gaussian data: Geostatistical v areal models

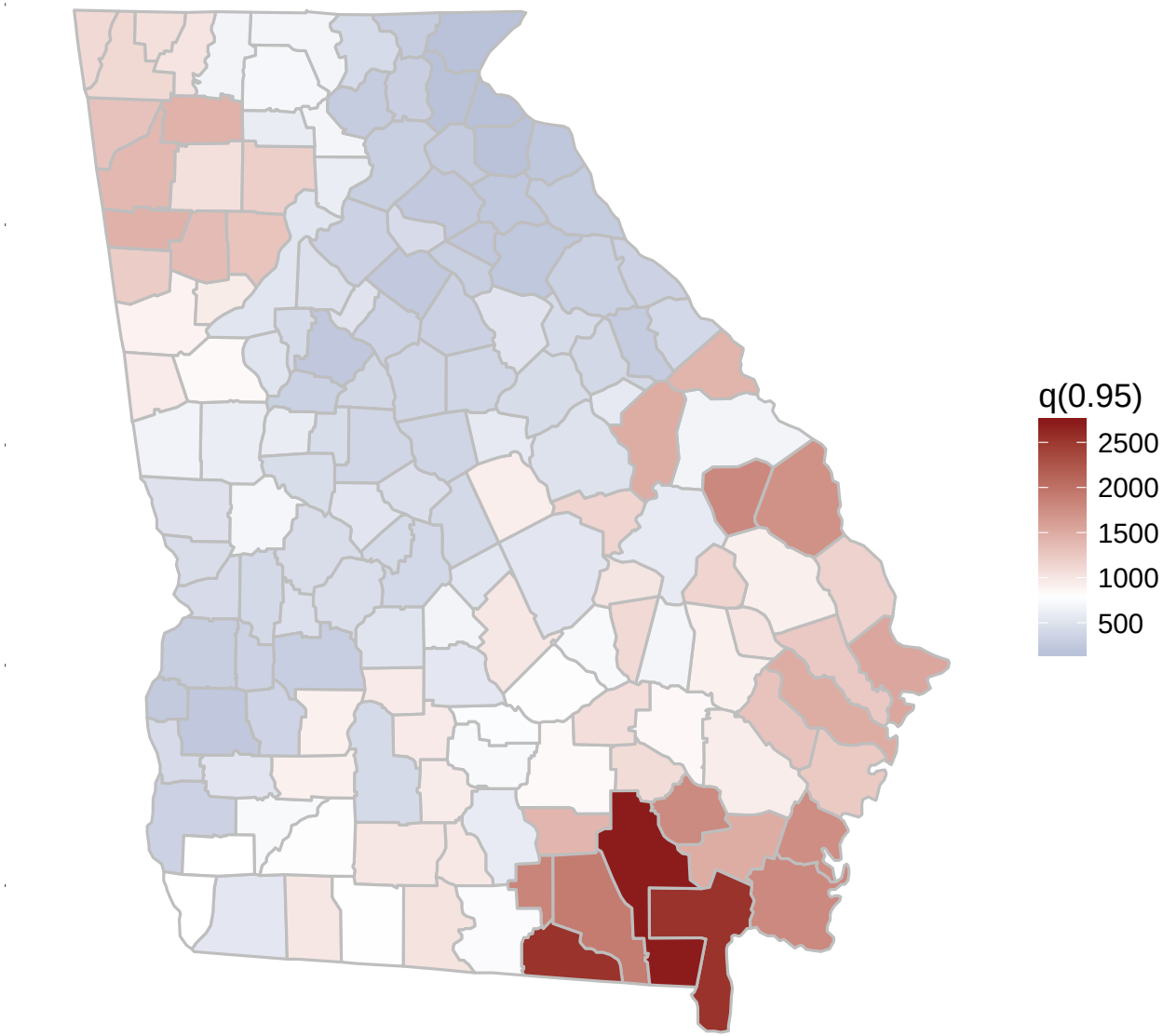
- **Geostats**: we sample n of an uncountable number of potential spatial locations
- Example: precipitation monitors
- Common methods: Matern correlation, Kriging, etc.

- **Areal**: the entire domain is partitioned into n regions
- Example: county-level disease rates
- Common methods: Conditionally autoregressive (CAR) model

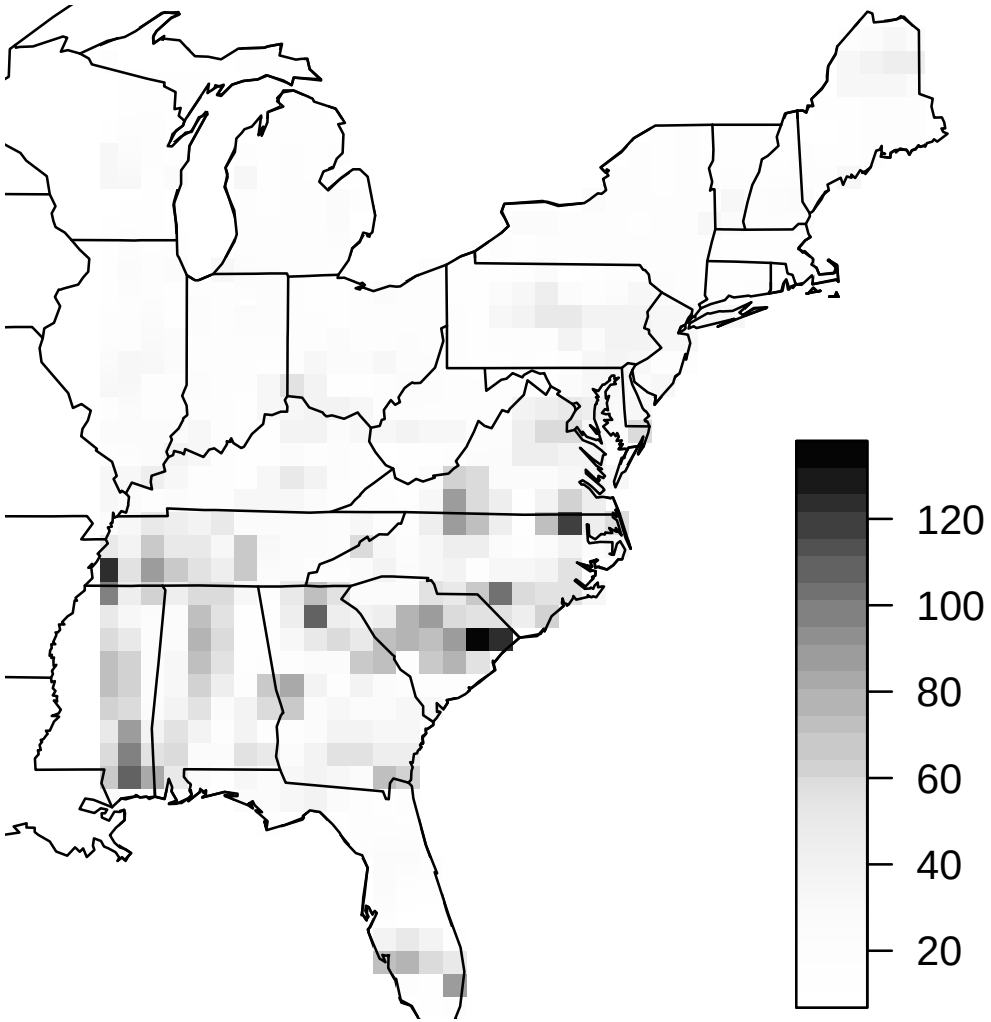
Gaussian data: Geostatistical v areal models

- Compared to geostat models, CAR models have computational advantages because they are defined locally
- In extremes, max-stable processes are the analogy of geostat models
- Max-stable processes are far more complex (conceptually and computationally) than Geostat models
- Climate data are often areal
- To our knowledge there is currently no analogy of CAR models for extremes

Example 1 – Forest fires in GA



Example 2 – Climate model precip output



Areal data model - Marginal distributions

- Y_{it} is the annual maximum in region i and year t
- The marginals are GEV with spatiotemporal parameters
- Location: $\mu_{it} = \sum_{l=1}^{L} X_{lt}\beta_{il}$, where
  - X_{lt} are known B-spline basis functions of time
  - β_{il} are unknown basis coefficients
- Scale: $\log(\sigma_{it}) = \beta_{i0}$
- Shape: $\xi_{it} = \xi$

Areal data model - Marginal distributions

- GEV parameters $\boldsymbol{\beta}_i = (\beta_{i0}, \beta_{i1}, ..., \beta_{iL})^T$ have MCAR prior

- $\boldsymbol{\beta}_i | \boldsymbol{\beta}_j, j \neq i \sim \text{Normal}\left(\boldsymbol{\gamma} + \rho(\bar{\boldsymbol{\beta}}_i - \boldsymbol{\gamma}), \frac{1}{m_i}\Sigma\right)$
  - $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, ..., \gamma_L)^T$ is the mean vector
  - $\bar{\boldsymbol{\beta}}_i$ is the mean of region i's neighbors
  - $\rho \in (0, 1)$ controls the strength of spatial dependence
  - Σ is an $L+1 \times L+1$ covariance matrix

Residual dependence

- To produce reliable Bayesian inference we must account for residual dependence
- Let $Z_{it} = [1 + \xi(Y_{it} - \mu_{it})/\sigma_i]^{1/\xi}$
- The residuals have unit Frechet distribution

$$Z_{it} \sim GEV(1, 1, 1)$$

- We will specify a spatial Markov model for Z_{it} as a function of neighboring Z_{jt}

Residual dependence

- For notational simplicity, we temporarily omit the temporal subscript, $Z_{it} \rightarrow Z_i$.
- Spatial dependence is introduced via a random partition of the n regions
- The adjacency structure of the regions determines the partition probabilities
- The random partition model captures spatial clusters of extreme events
- Within these clusters, we assume a multivariate GEV (MGEV) distribution to model extremal dependence

Residual dependence

- Assume the regions are partitioned into K clusters
- $g_i = k$ indicates that region i is allocated to cluster k
- Observations in different clusters are independent
- Observations within a cluster follow an exchangeable MGEV distribution
- We use the symmetric logistic extremal measure with dependence parameter $\alpha \in (0, 1)$
- Small α gives dependence, $\alpha = 1$ gives independence

Residual dependence

- The cluster labels follow the spatial Potts model

$$p(g_1, ..., g_n|\phi) \propto \exp\left(\sum_{i\sim j} \phi I(g_i = g_j)\right)$$

- $\phi > 0$ determines the strength of spatial dependence
- This leads to a Markov model for the labels

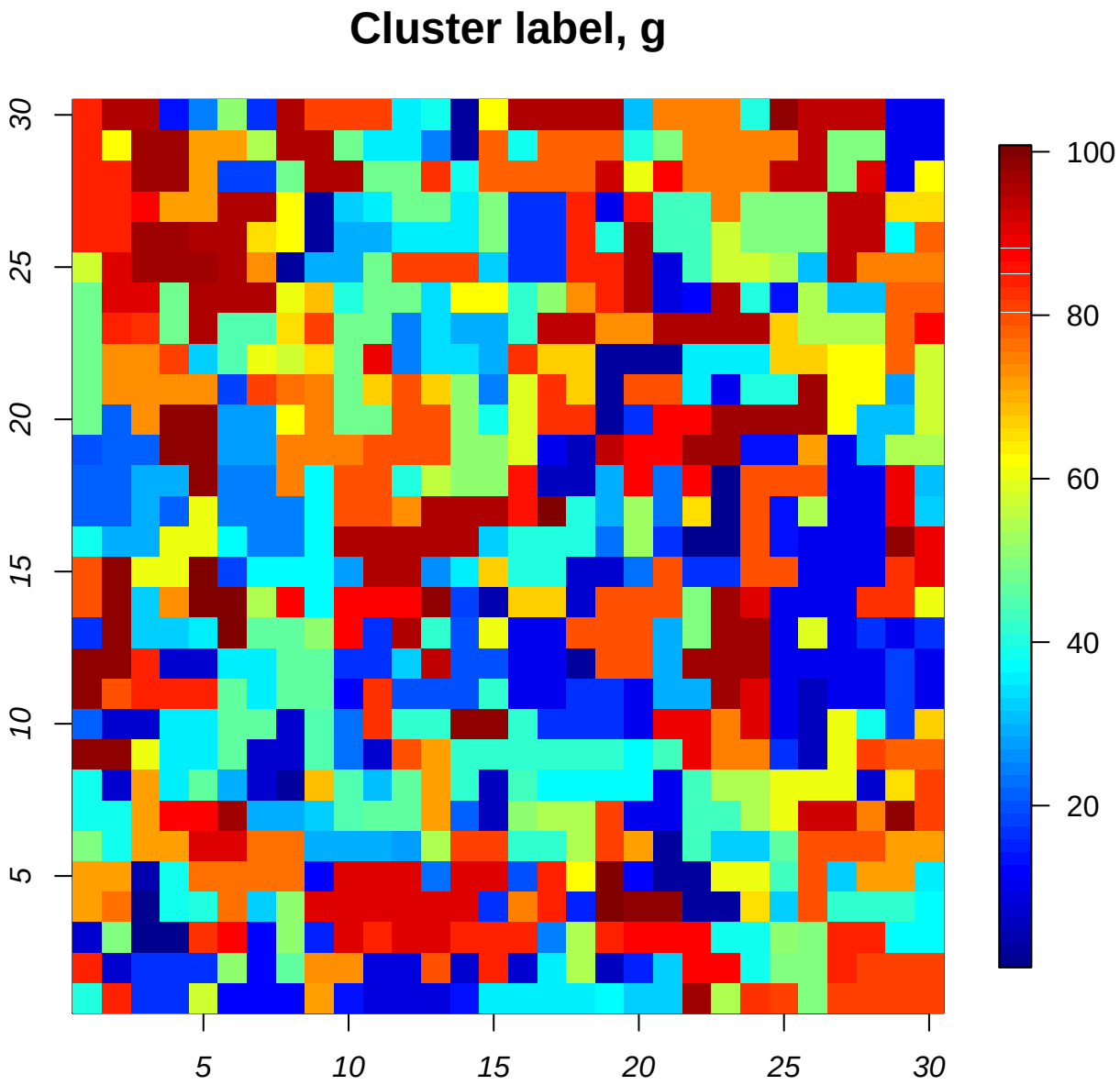
$$\text{Prob}(g_i = k|g_j, j \neq i) \propto \exp\left[\phi \sum_{j\sim i} I(g_j = k)\right]$$

- The log odds of label k increases by ϕ for each neighbor in cluster k

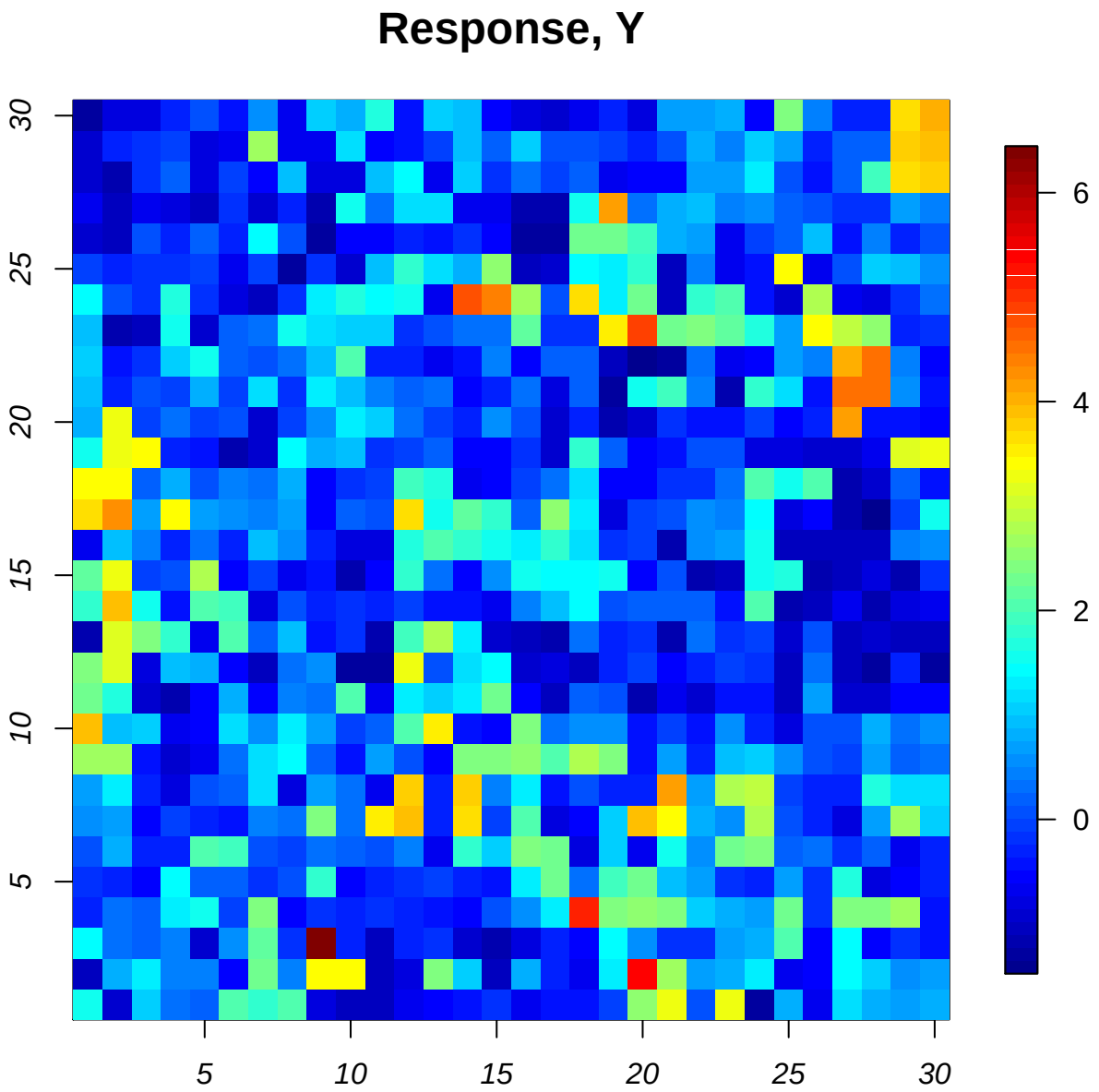
Realizations

- The following slides show realizations of the process with $\mu_i = 0$, $\sigma_i = 1$ and $\xi_i = 0.1$
- We plot both the labels, g_i, and the responses, Y_i
- The spatial dependence parameter ϕ ranges from small to large
- The MGEV parameter α ranges from small (dependence) to large (independence)

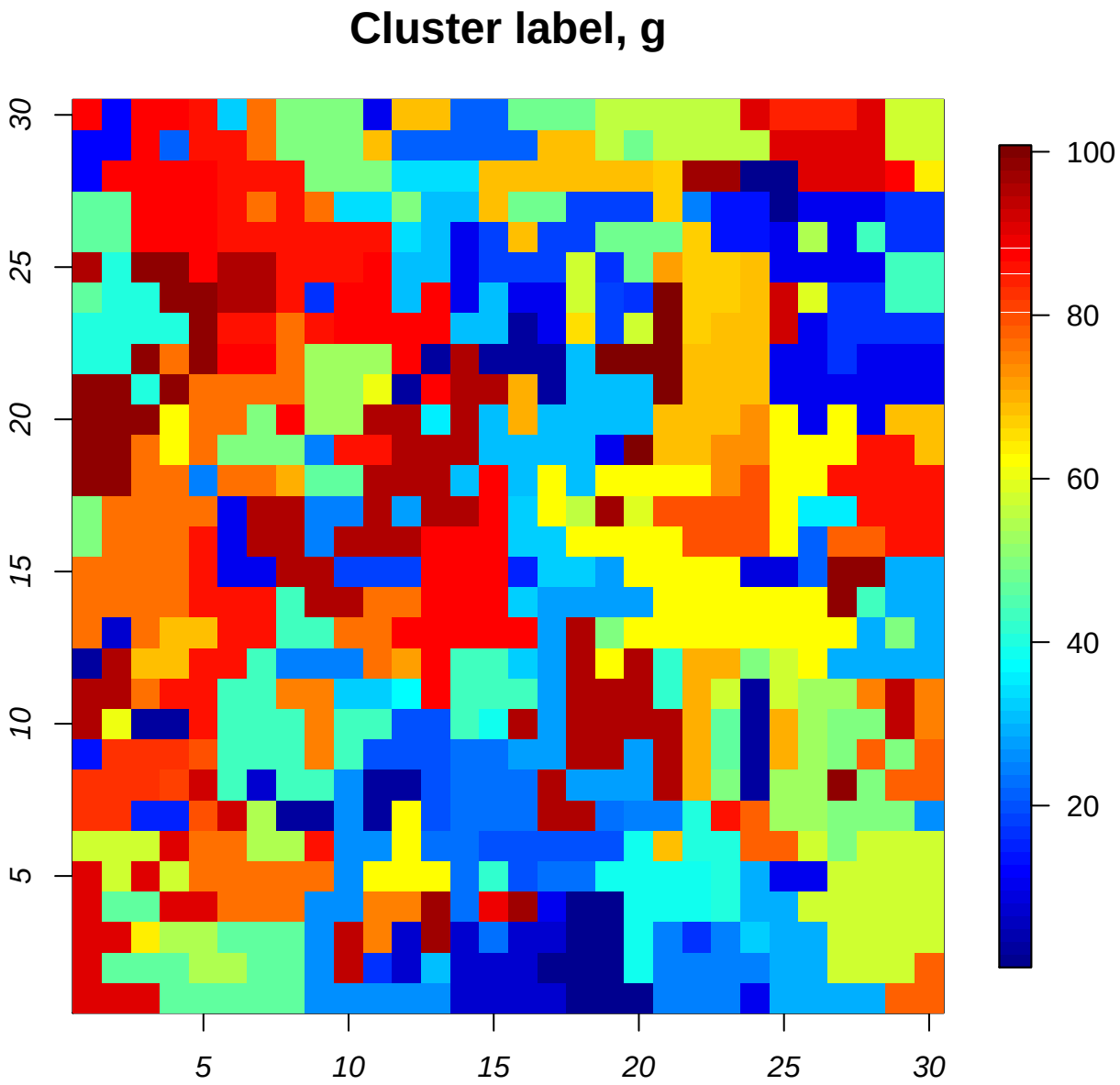
Small ϕ, small α



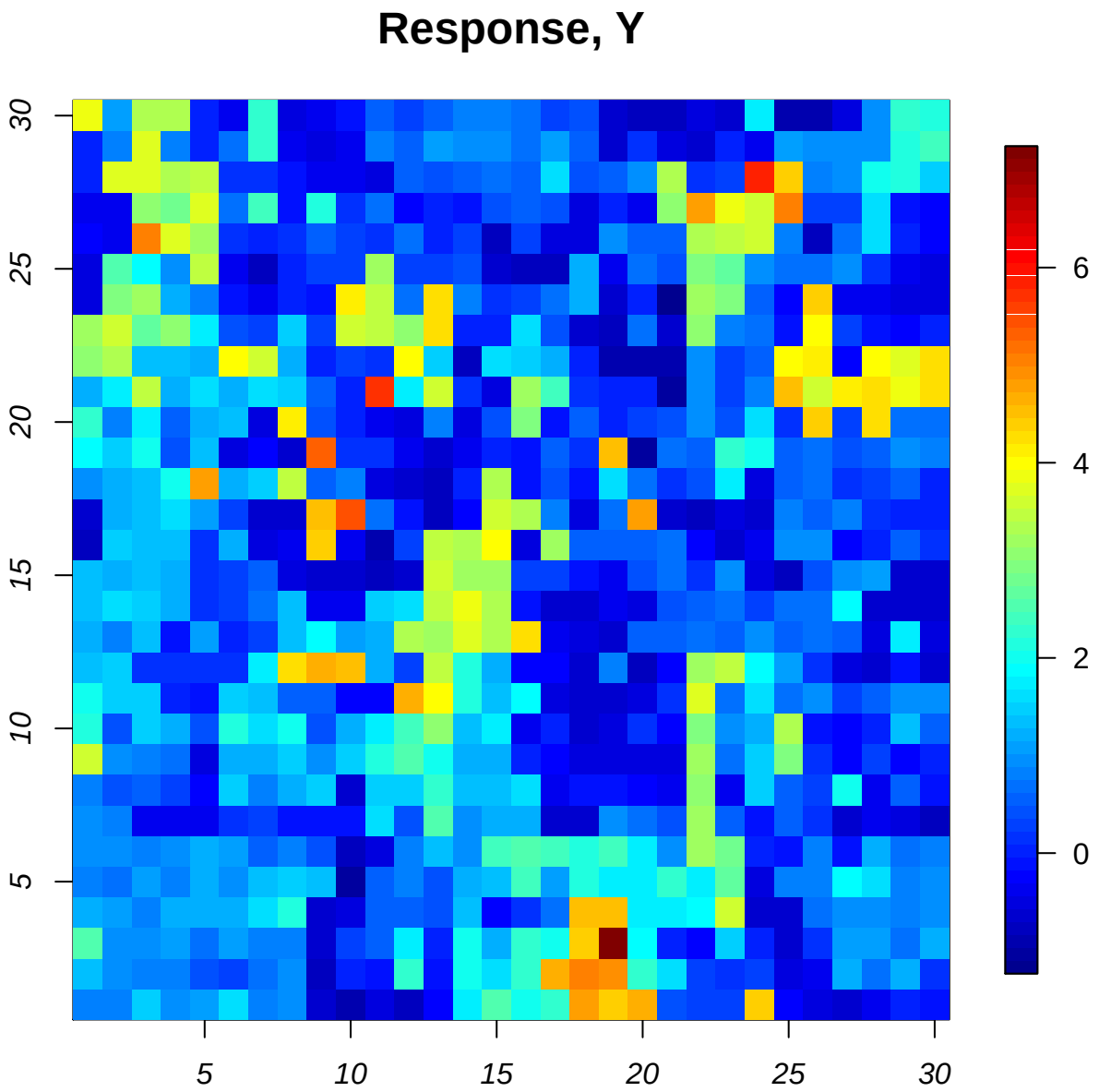
Small ϕ, small α



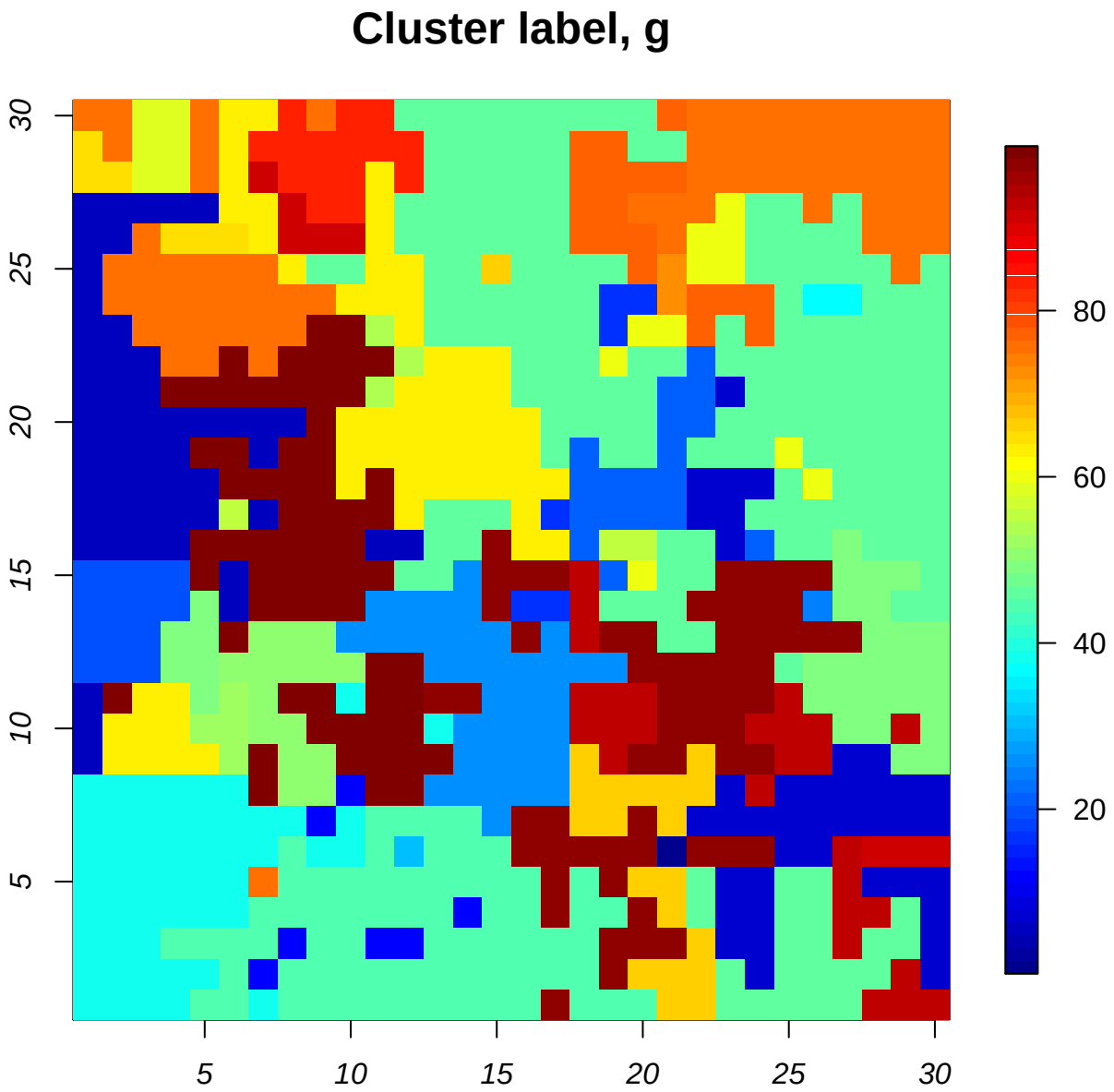
Medium ϕ, small α



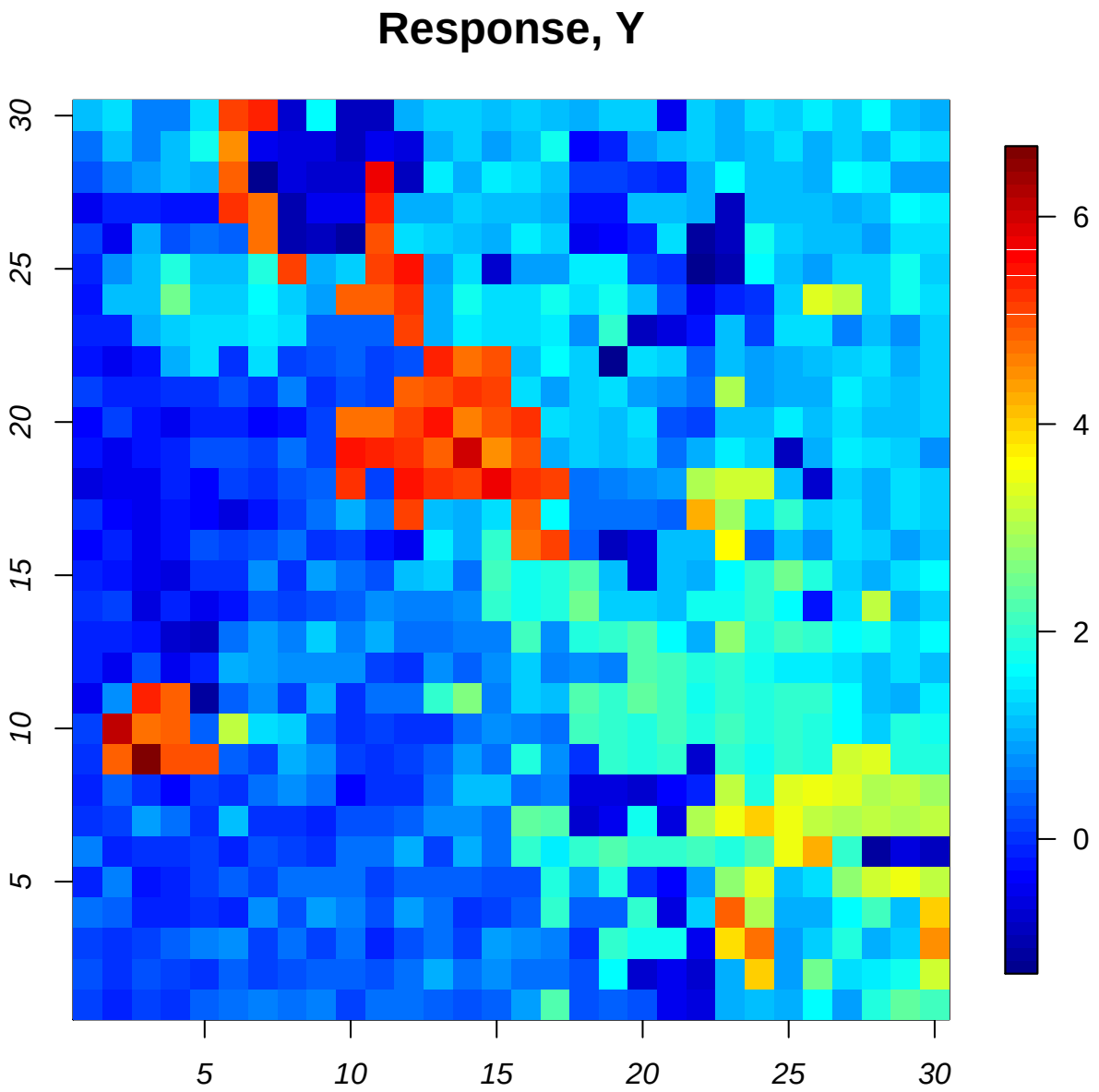
Medium ϕ, small α



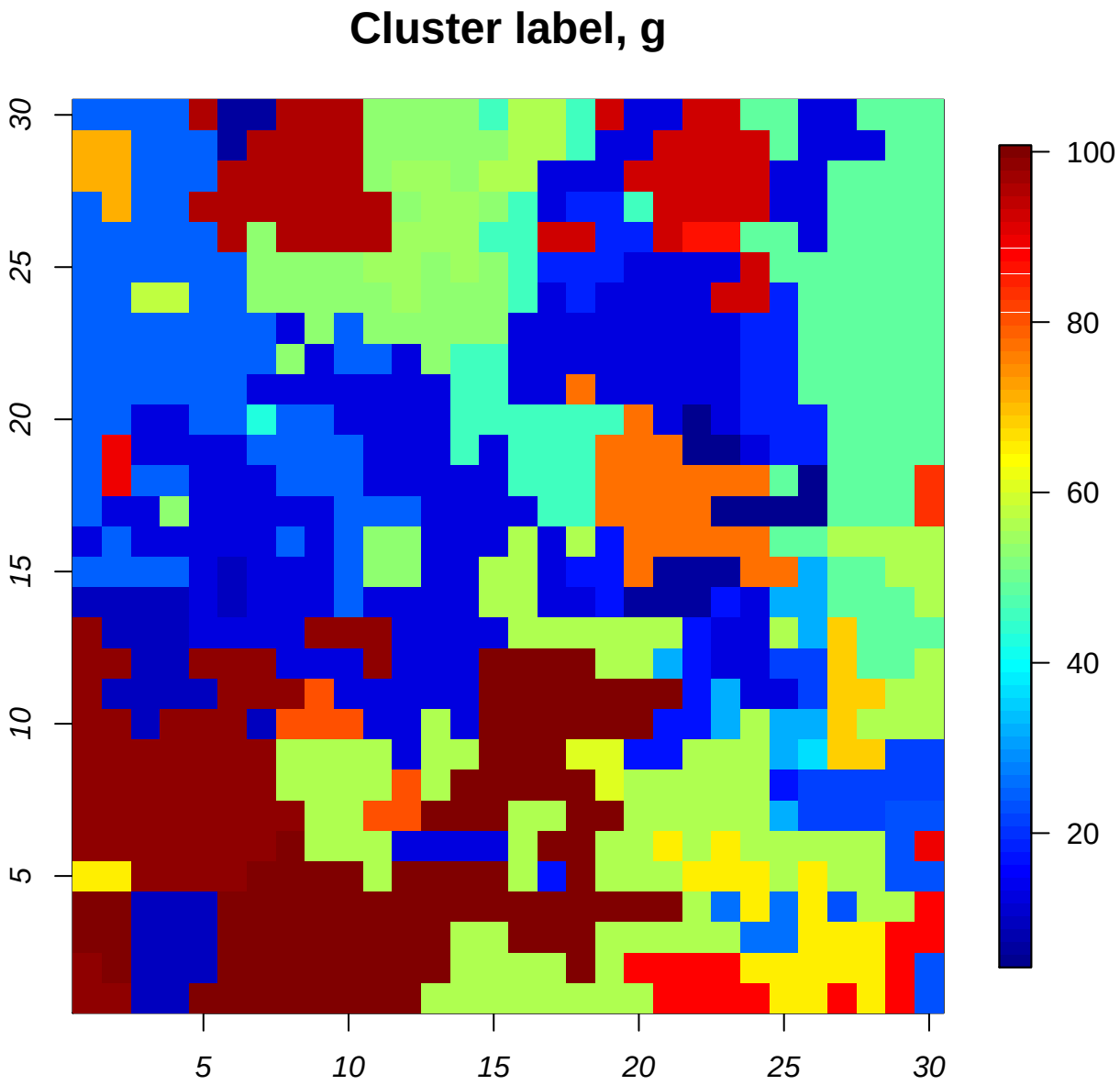
Large ϕ, small α



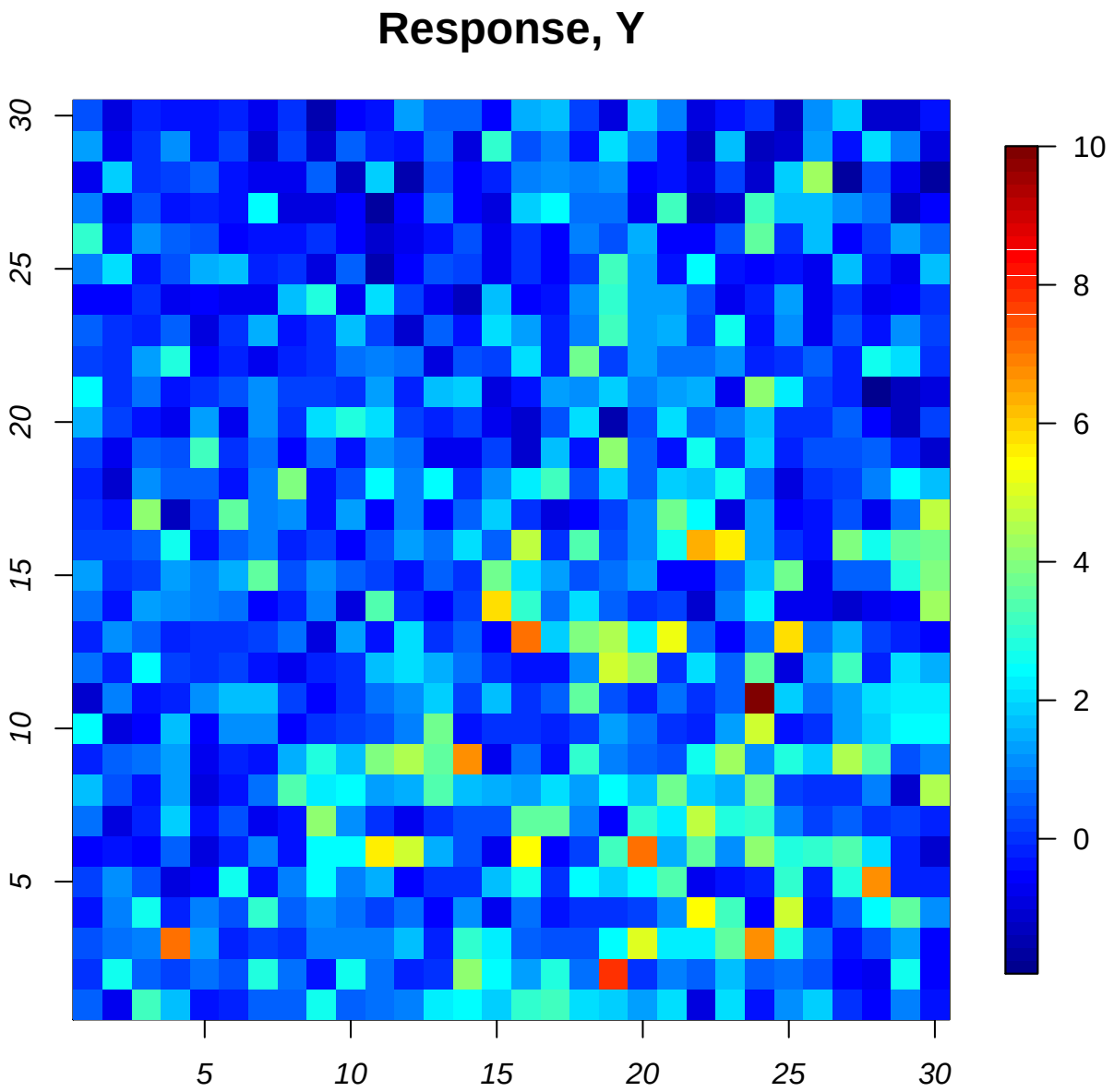
Large ϕ, small α



Large ϕ, large α



Large ϕ, large α



Asymptotic properties

- Asymptotic dependence is often measured by

$$\chi_{ij} = \lim_{z \to \infty} \text{Prob}\left(Z_i > z | Z_j > z\right)$$

- Z_i and Z_j are asymptotically independent if $\chi_{ij} = 0$

- For the Potts/MGEV model,

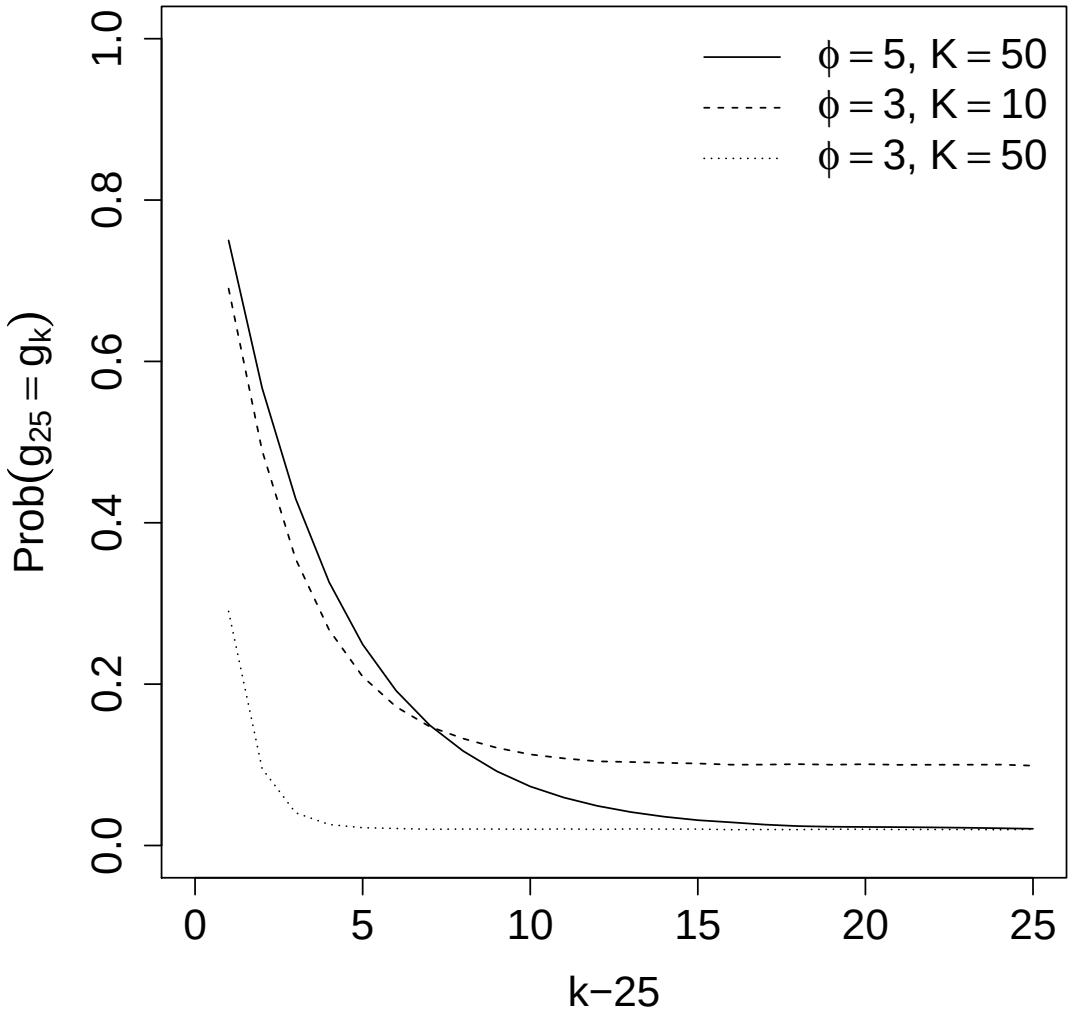
$$\chi_{ij} = \pi_{ij}(\phi)(2 - 2^{\alpha})$$

- $\pi_{ij}(\phi) = \text{Prob}(g_i = g_j) \in (1/K, 1)$ from the Potts model

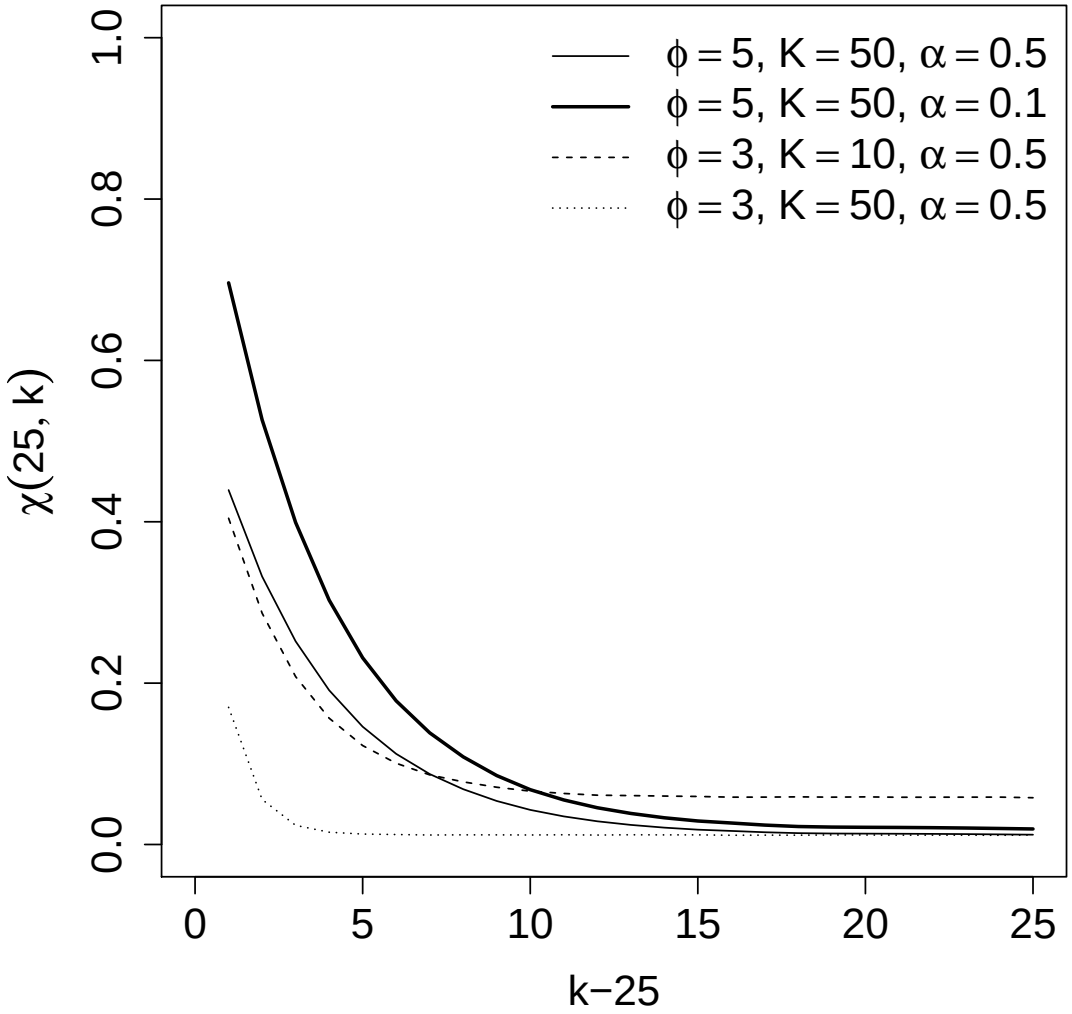
Asymptotic properties

- The following slides plot χ_{ij} for a linear one-dimensional grid of $n = 50$ locations with first-order neighbors
- The plots give $\pi_{25,k}(\phi) = \text{Prob}(g_{25} = g_k)$ and extremal dependence measure $\chi_{25,k}$
- The Potts probabilities do not have a closed form
- We use Monte Carlo simulation

Asymptotic properties - $\pi_{25,k}$



Asymptotic properties - $\chi_{25,k}$



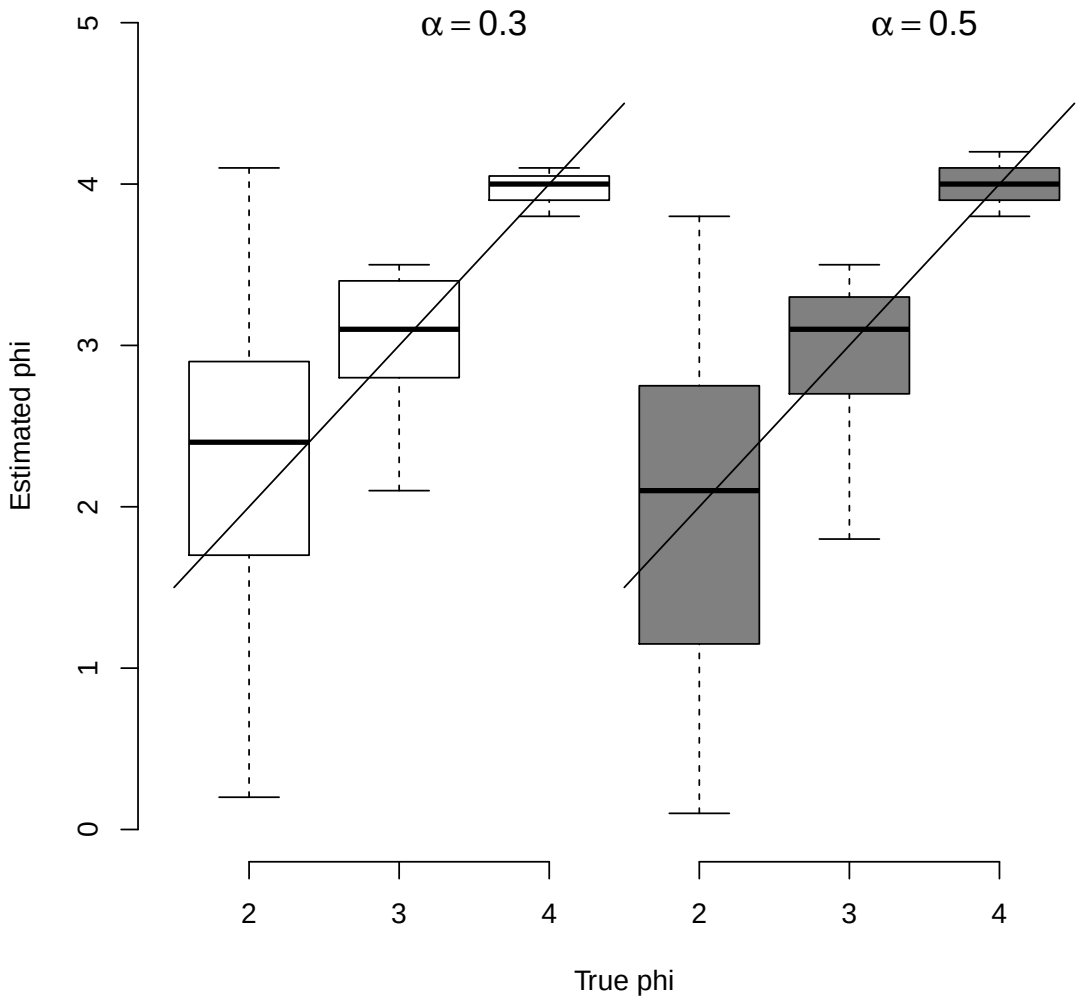
Computation - overview

- We primarily use MCMC
- The Potts parameter is hard to estimate
- After a reparameterization, the likelihood factors across sites
- All MCMC updates are local and fast

Computation - Estimating the Potts parameter

- Updating ϕ is difficult because the Potts distribution has an intractible normalizing constant
- We use a plug-in estimator
- Recall $\chi_{ij} = \pi_{ij}(\phi)[2 - 2^{\alpha}]$
- We fix ϕ at the value that maximizes correlation between $\pi_{ij}(\phi)$ and empirical estimates of χ_{ij}

Simulation study



Computation - random effects

- The MGEV distribution with symmetric logistic dependence has a mixture representation
- Let $A_k \sim$ generalized inverse gamma be a random effect for cluster k
- Then for sites in cluster k,

$$Z_i | g_i = k \sim \text{GEV}(A_k^\alpha, \alpha A_k^\alpha, \alpha)$$

  independent over i

- The likelihood now factors across all sites

Computation - sketch of MCMC

- β_j: independence sampler from prior
- g_{kt}: Gibbs
- A_{kt}: Gibbs/Metropolis
- The rest is straightforward Gibbs

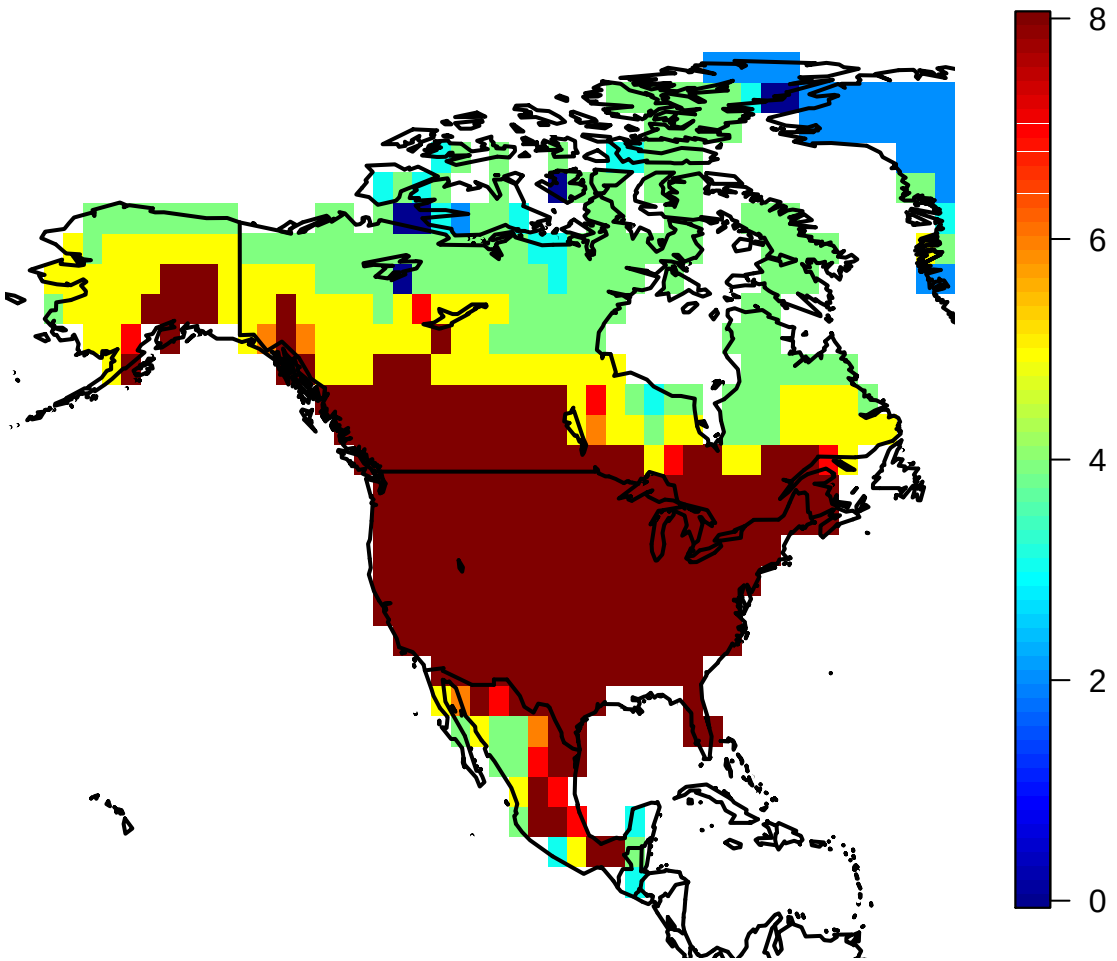
CLIMDEX data

- The CLIMDEX/GHCNDEX data repository contains a suite of gridded climate indices
- Each index is calculated annually over the period from 1950–2015
- Data are provided on the 2.5 x 2.5 degree grid of $n = 509$ locations
- We study eight indices

CLIMDEX data

Abbreviation	Description
TXx	Annual maximum of daily max temp (C)
TNx	Annual maximum of daily min temp (C)
TXn	Annual minimum of daily max temp (C)
TNn	Annual minimum of daily min temp (C)
Rx1day	Annual maximum of daily precip (mm)
Rx5day	Annual maximum of 5-day average precip
CDD	Maximum length of dry spell
CWD	Maximum length of wet spell

CLIMDEX data - number of indicies per site



CLIMDEX data - Settings

- We take $K = n$ possible clusters
- All other priors are uninformative
- We compare $L = 4, 6, 8$ temporal basis functions using CV

Cross validation MAD (of standardized data)

L	TXx	TNx	TXn	TN
4	0.081	0.083	0.030	0.033
6	0.080	0.081	0.029	0.033
8	0.079	0.081	0.028	0.032

L	Rx1day	Rx5day	CDD	CWD
4	0.259	0.281	0.272	0.414
6	0.257	0.281	0.271	0.413
8	0.259	0.280	0.270	0.408

Coverage of 80% intervals

L	TXx	TNx	TXn	TNn
4	82.7	82.0	87.3	86.1
6	81.6	81.4	87.1	87.1
8	81.6	81.3	86.6	86.7

L	Rx1day	Rx5day	CDD	CWD
4	81.3	81.1	79.9	79.8
6	80.9	81.2	80.0	79.6
8	81.0	81.1	79.5	79.6

Coverage of 90% intervals

L	TXx	TNx	TXn	TNn
4	91.7	91.1	95.1	94.3
6	91.3	91.5	94.9	94.4
8	91.4	91.0	94.5	94.7

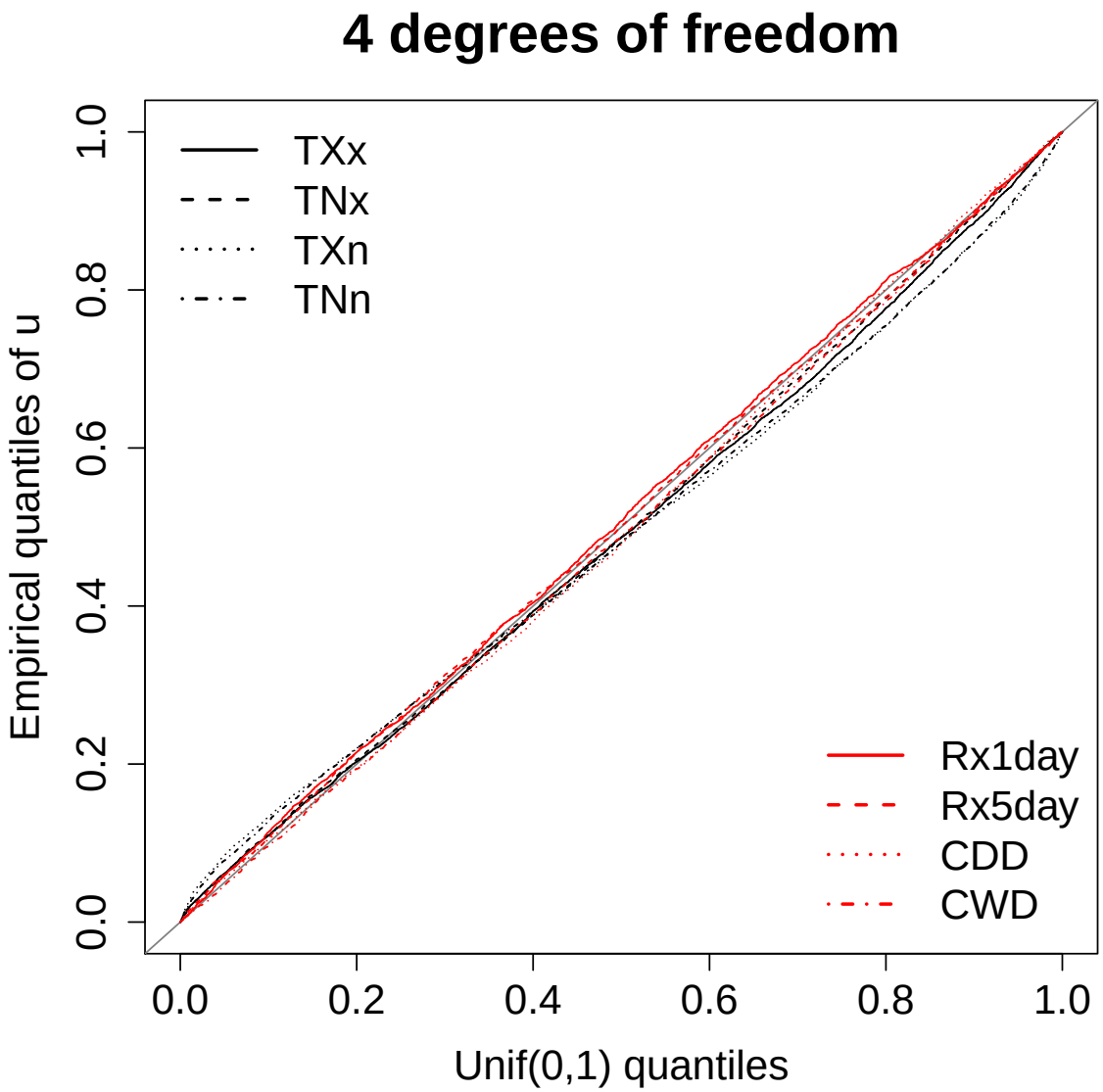
L	Rx1day	Rx5day	CDD	CWD
4	91.0	91.4	90.2	89.7
6	90.4	91.4	90.2	89.3
8	90.8	91.2	90.3	89.3

Coverage of 95% intervals

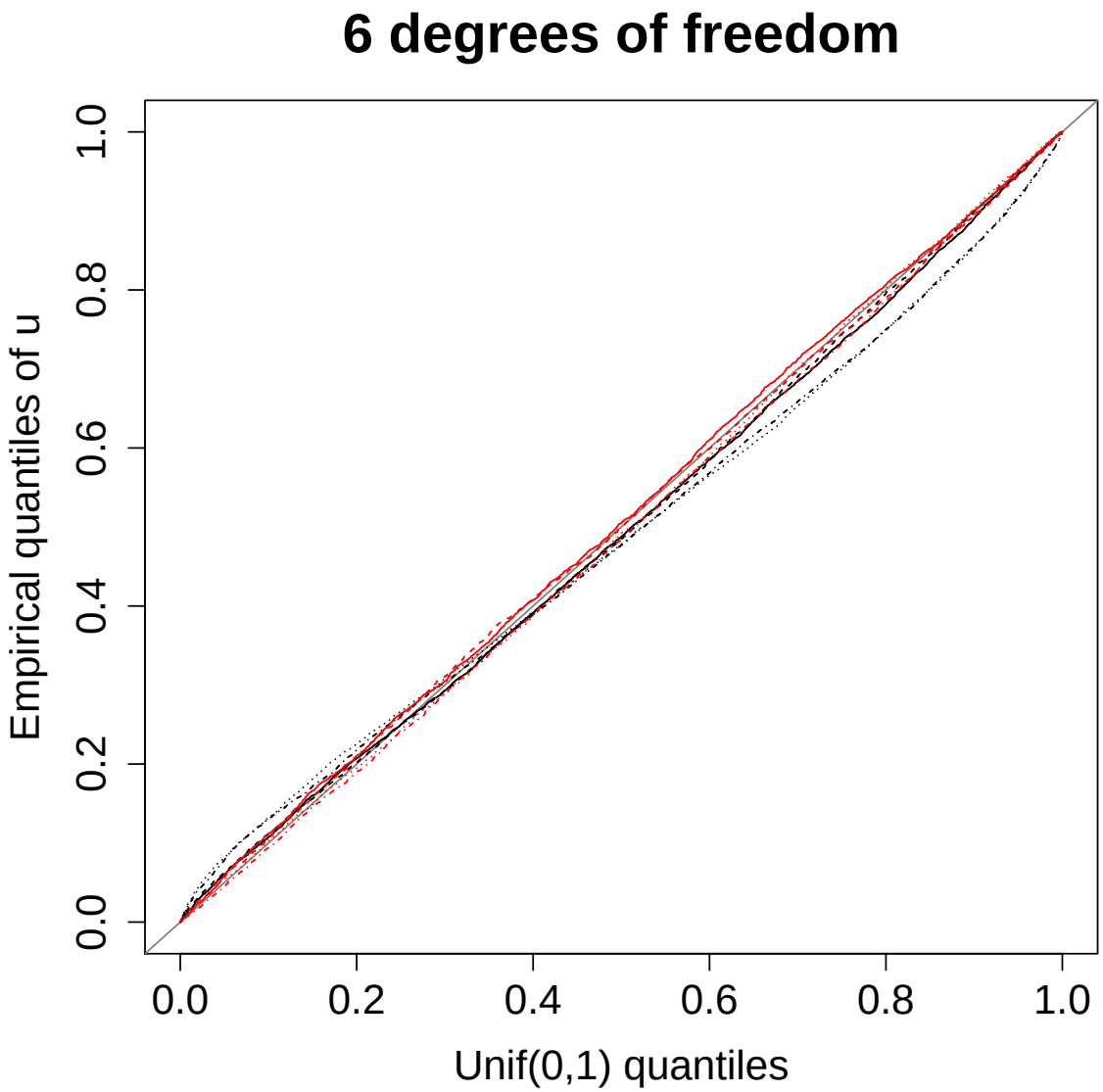
L	TXx	TNx	TXn	TNn
4	96.0	96.0	97.8	97.5
6	95.8	95.9	97.7	97.7
8	95.8	95.7	97.7	97.8

L	Rx1day	Rx5day	CDD	CWD
4	95.2	95.7	95.1	94.5
6	95.1	95.9	94.9	94.6
8	95.2	95.7	95.2	94.5

CLIMDEX data - Reliability plot with $L = 4$



CLIMDEX data - Reliability plot with $L = 6$

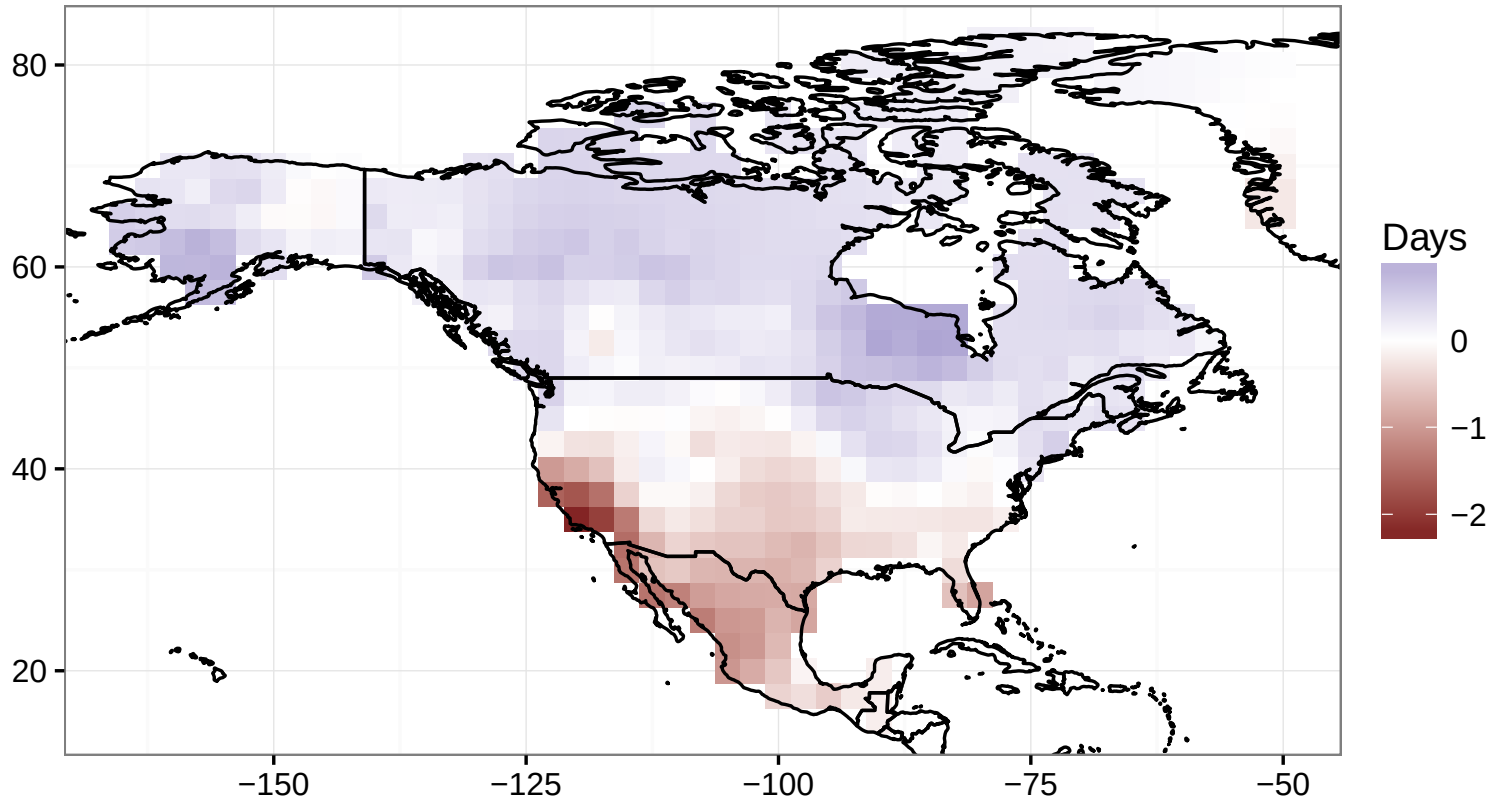


CLIMDEX data

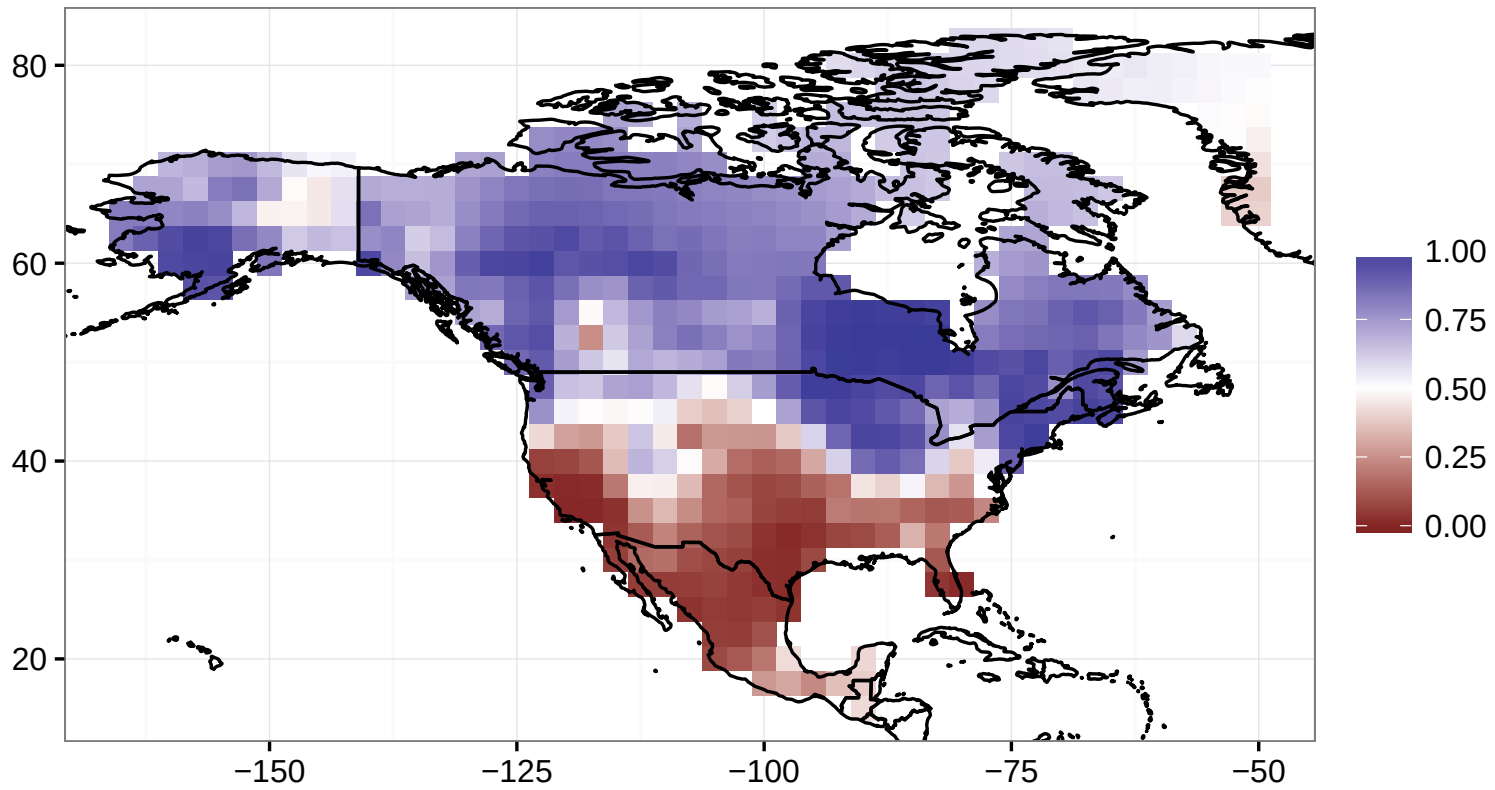
- We summarize the results using the posterior distribution of the decadal average change
- We map posterior means and posterior probability the change is positive
- We also plot the data versus fitted values for several pixels of interest
- These plots illustrate the non-linear fit of the GEV location

CDD - mean change

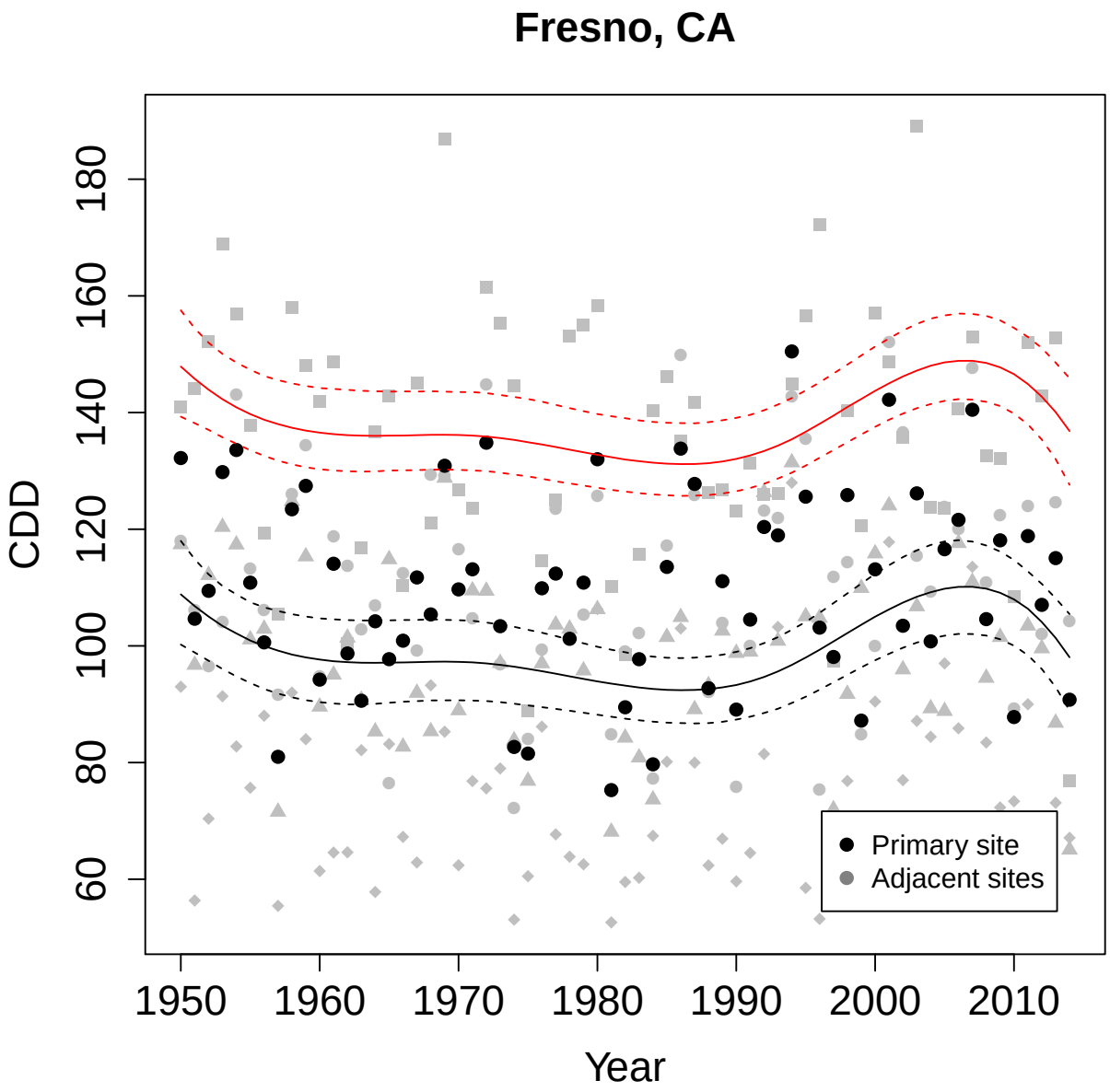
(e) Change in GEV location per decade



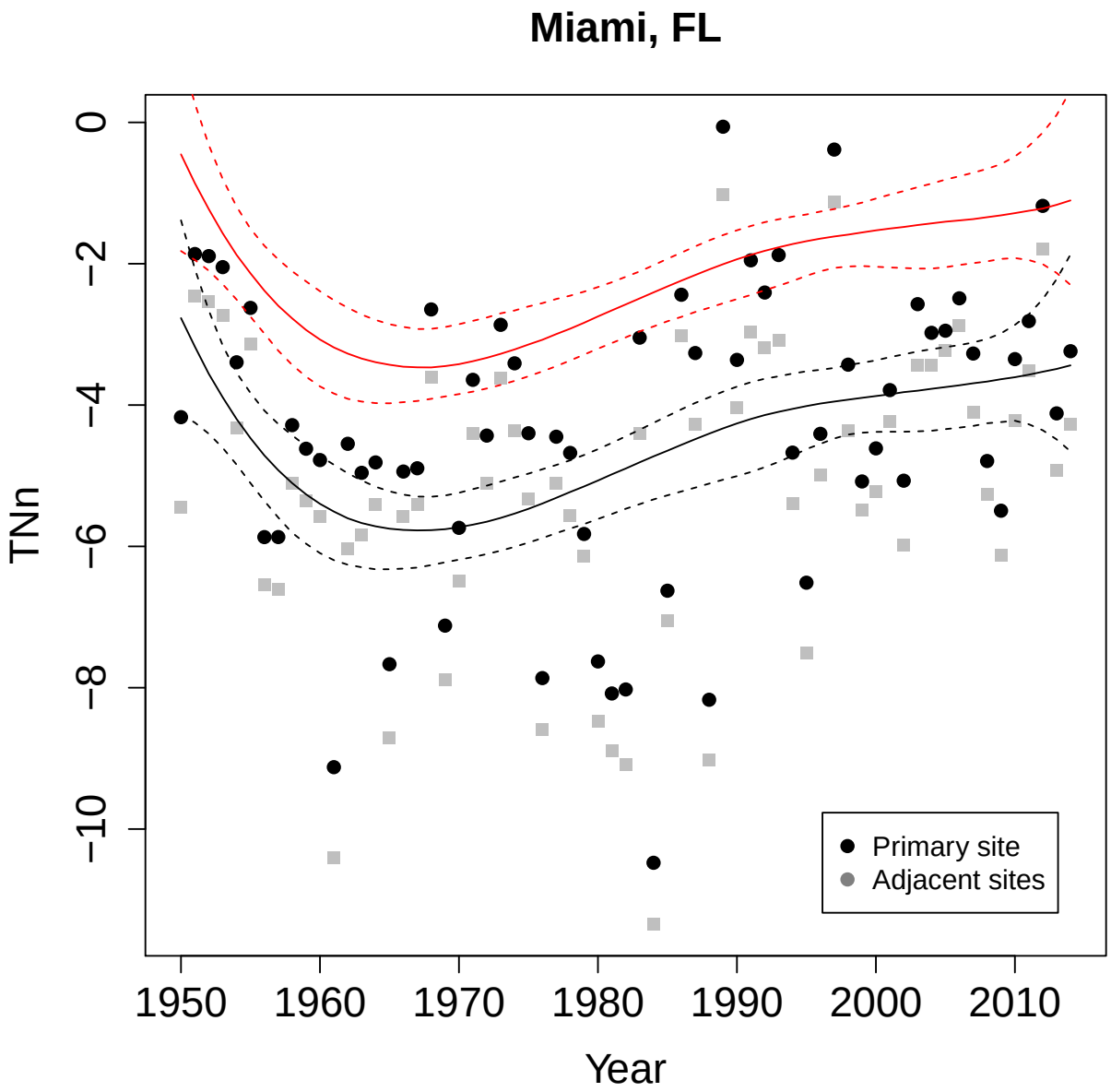
CDD - prob change>0



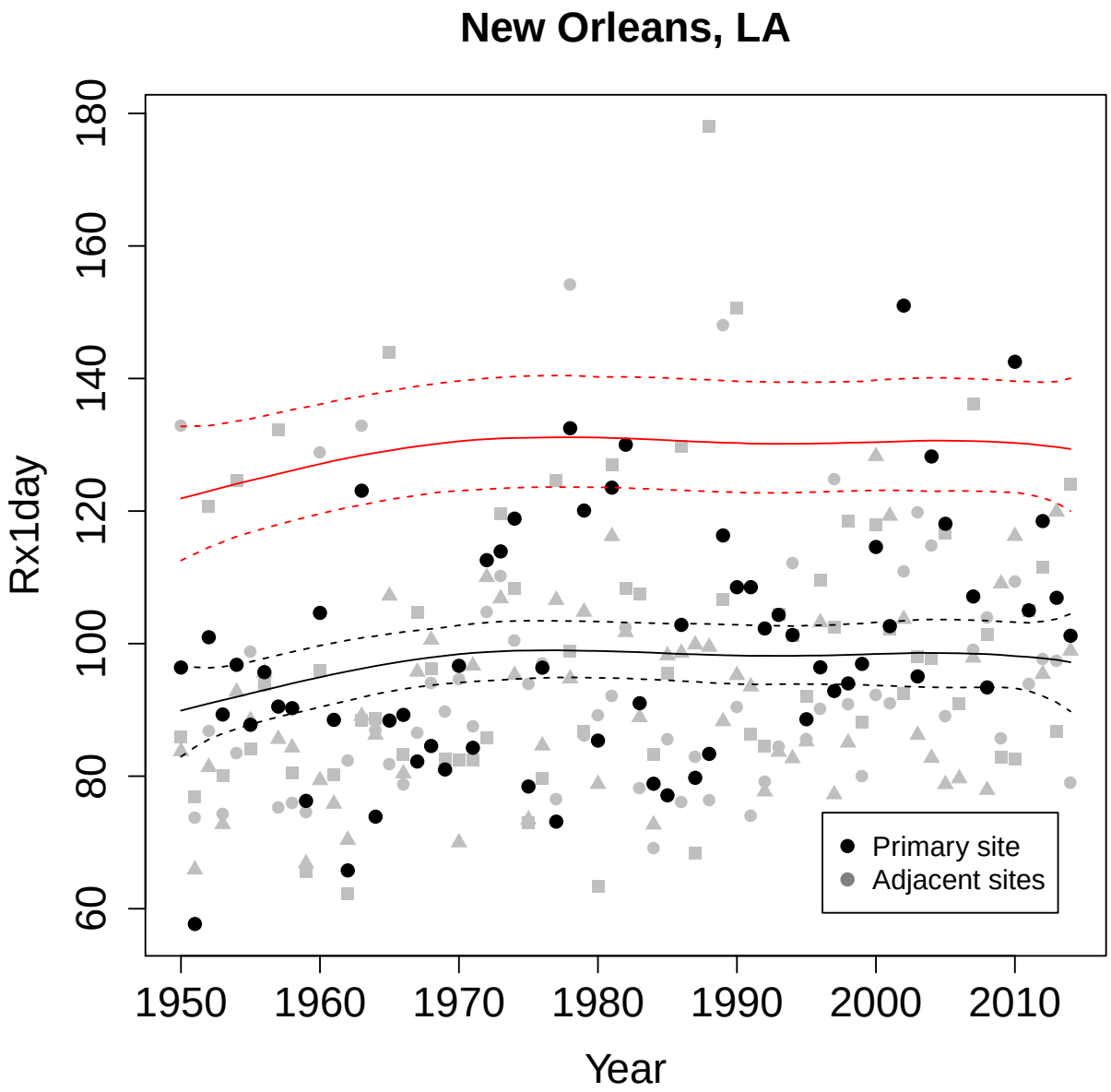
Time series plots



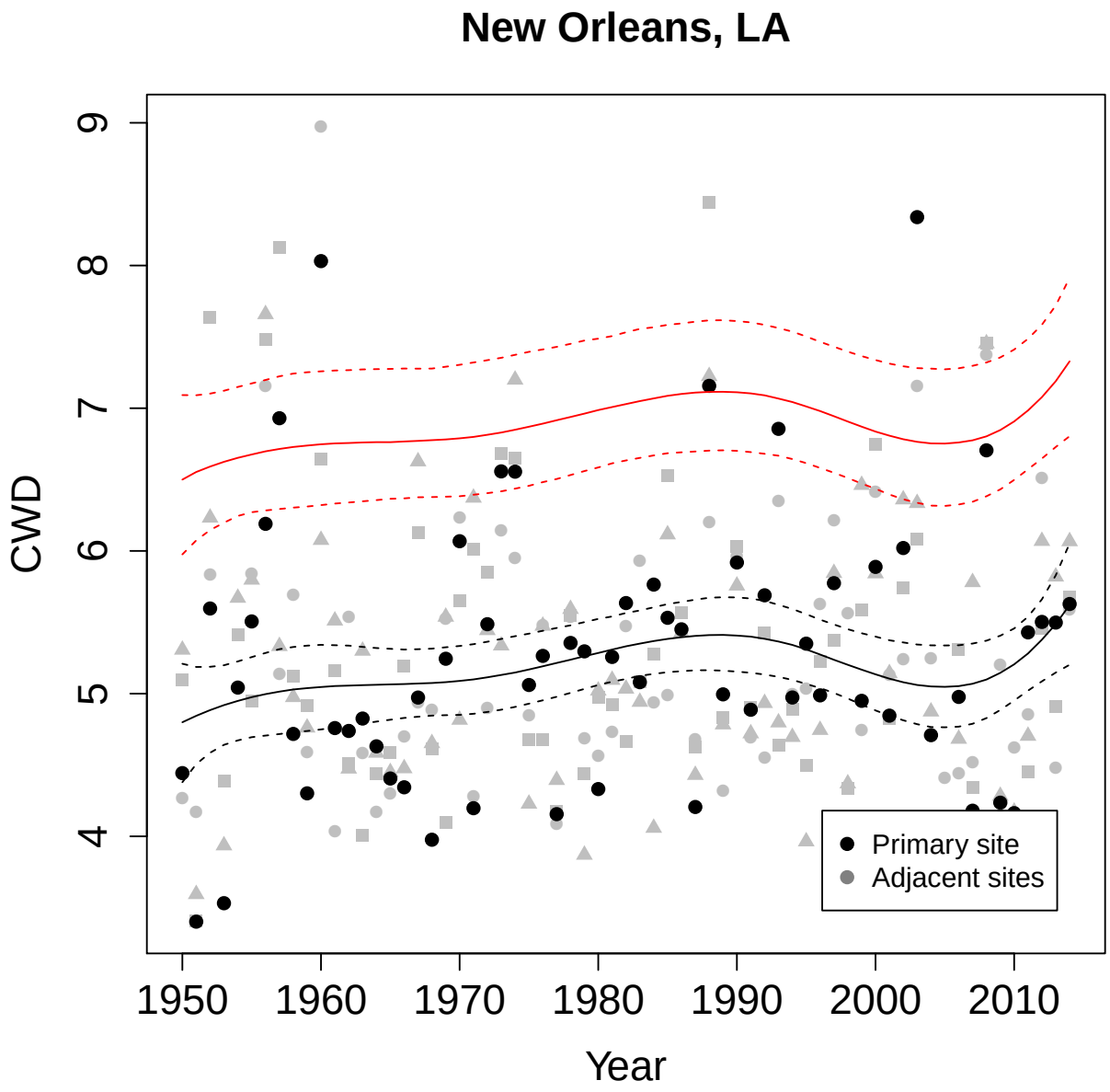
Time series plots



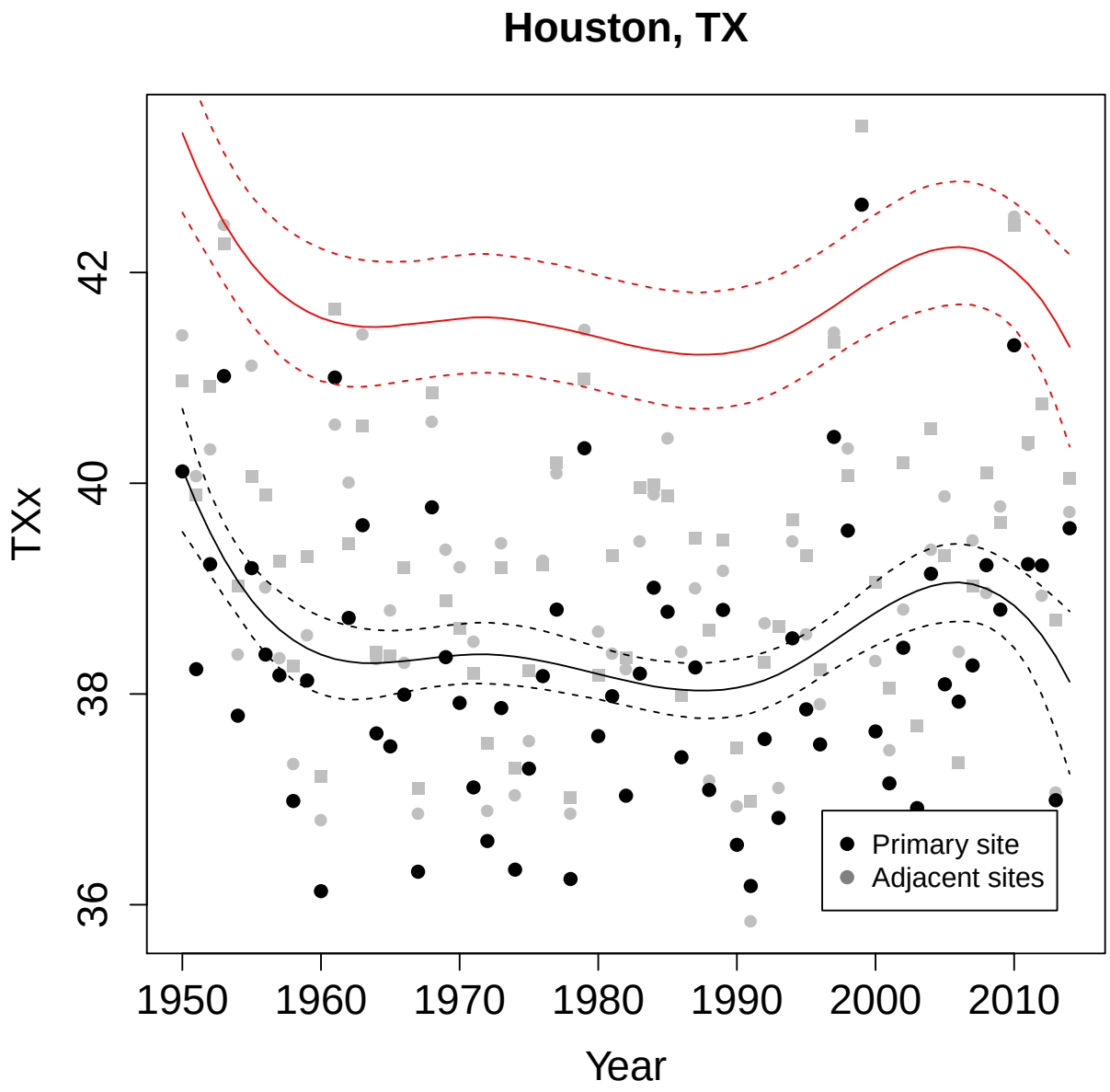
Time series plots



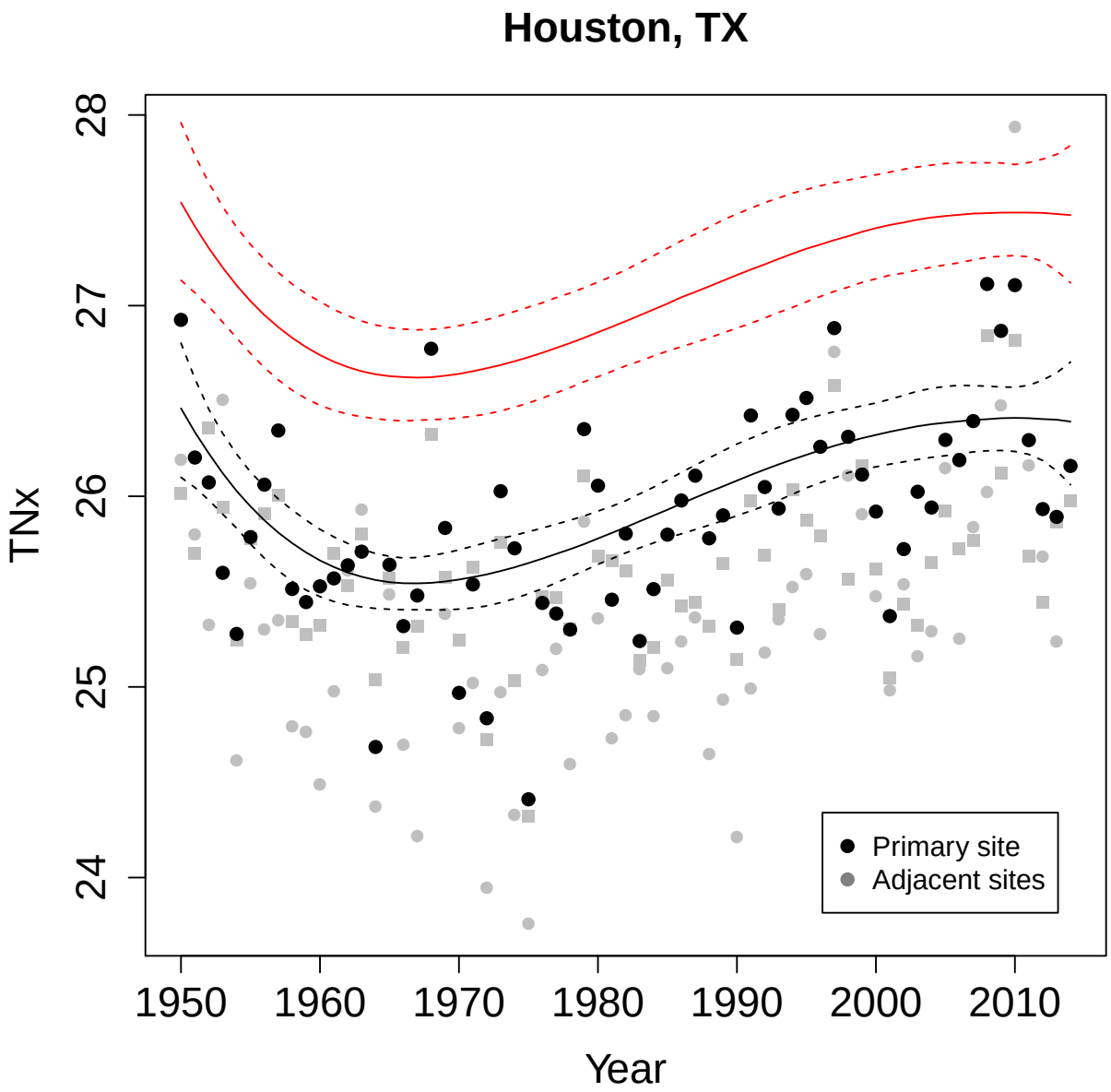
Time series plots



Time series plots



Time series plots



Summary

- We have proposed a Markov model for extremes
- Can we find a max-stable Markov model?
- We'd like to be able to compute the full posterior of ϕ
- Points over a threshold extension should be easy
- Support: NSF, DOI, EPA
- Thanks!