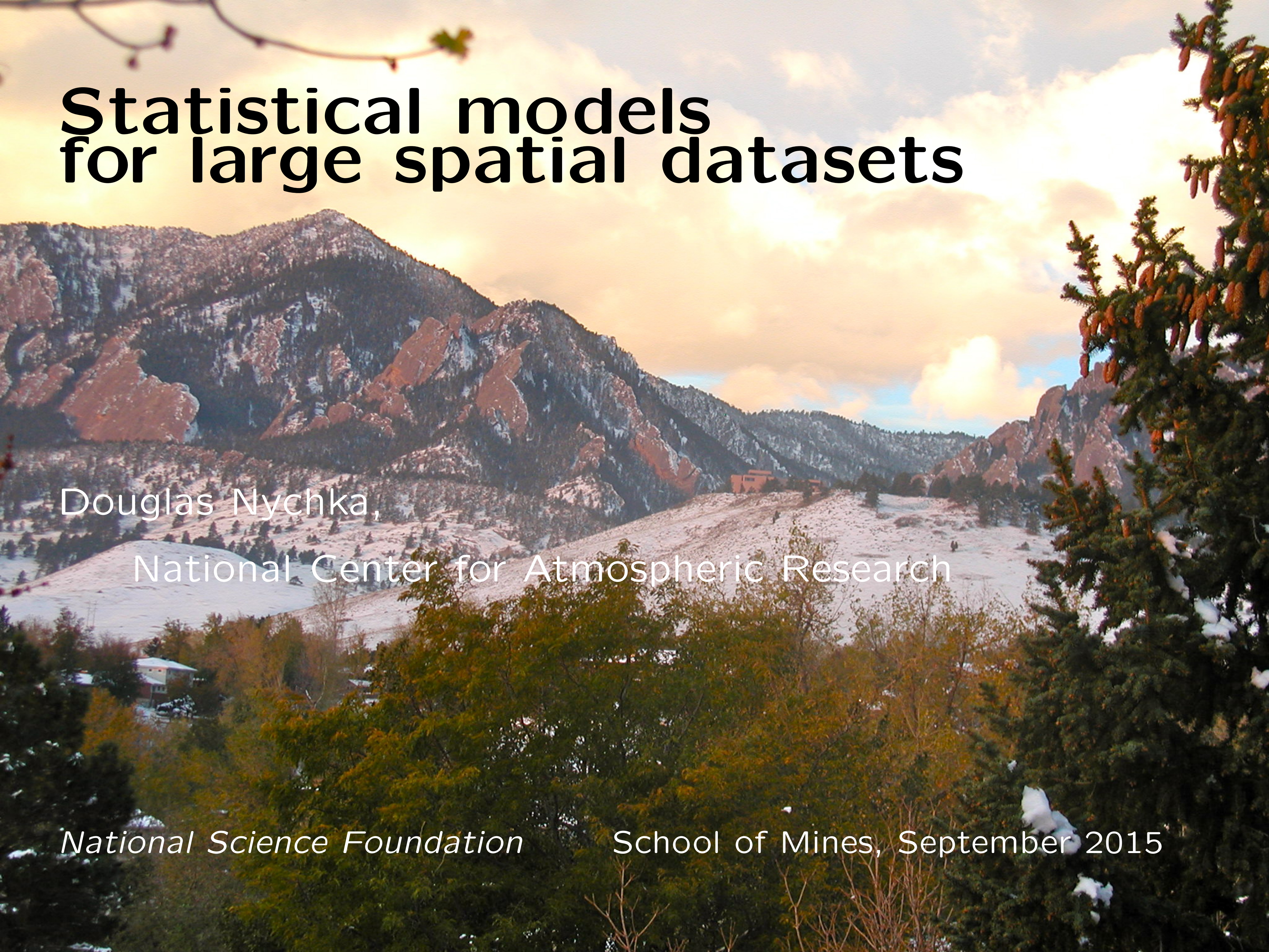


Statistical models for large spatial datasets

Douglas Nychka,

National Center for Atmospheric Research

National Science Foundation School of Mines, September 2015

Introduction

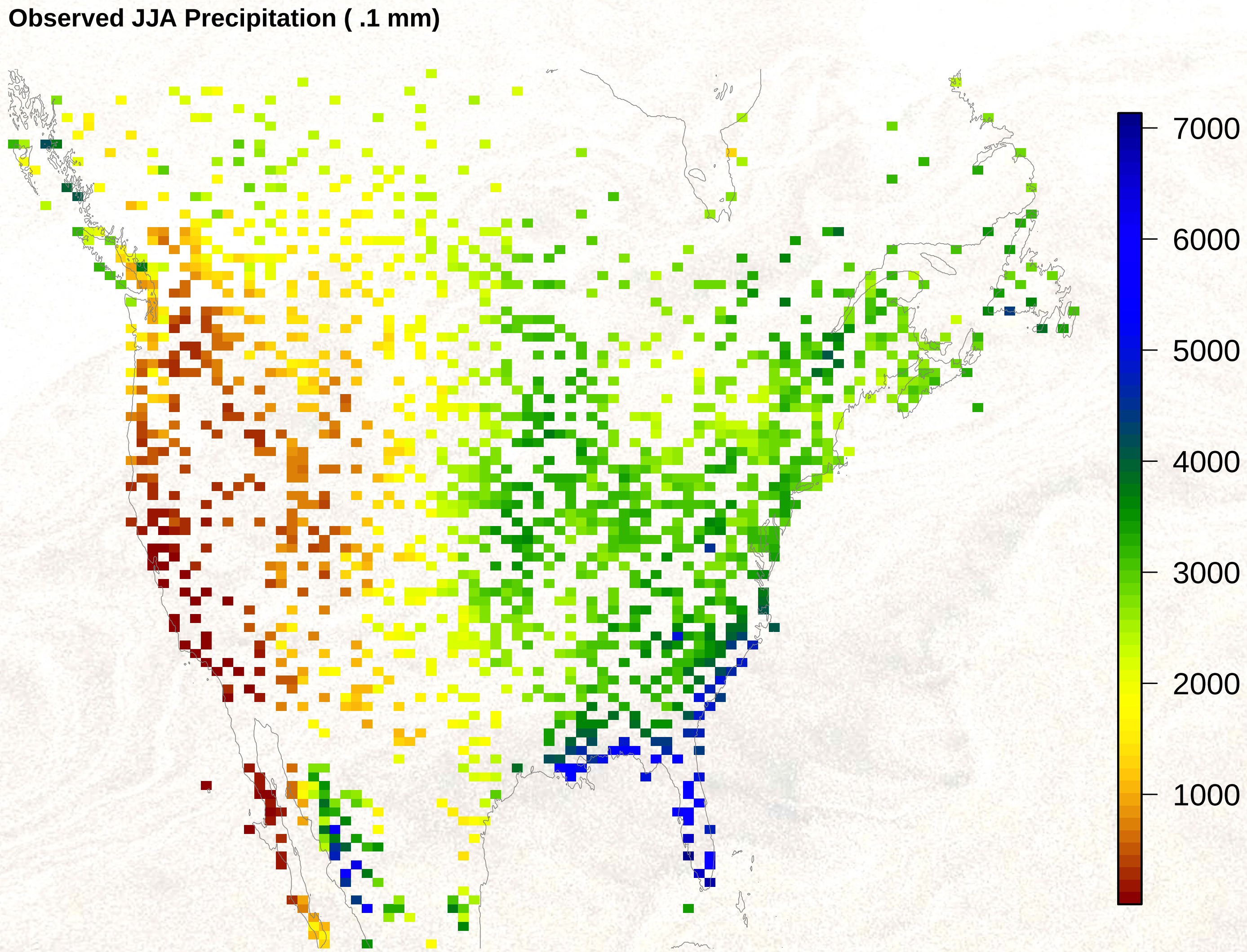
- Rainfall and Regional climate NARCCAP
- An additive model and Hilbert spaces.
- Some cartoons and a spatial model
- LatticeKrig – properties
- Future changes in the seasonality
- Tomography of the solar corona

Credits:

- Dorit Hammerling, SAMSI/STATMOS/NCAR
- Soutir Bandyopadhyay, Lehigh U
- Nathan Lenssen, Columbia
- Tamra Greasby, U Denver
- Finn Lindgren, U Bath
- Jim Gattiker, LANL
- John Paige, NCAR, U Washington
- Luke Burnett, Saint Olaf (Kevin Delmasse, Sarah Gibson)

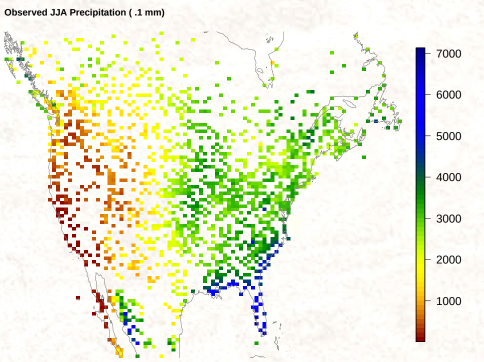
Observed mean summer precipitation

1720 stations reporting, "mean" for 1950-2010



The statistical problem

What is the summer rainfall at places where there is no data?



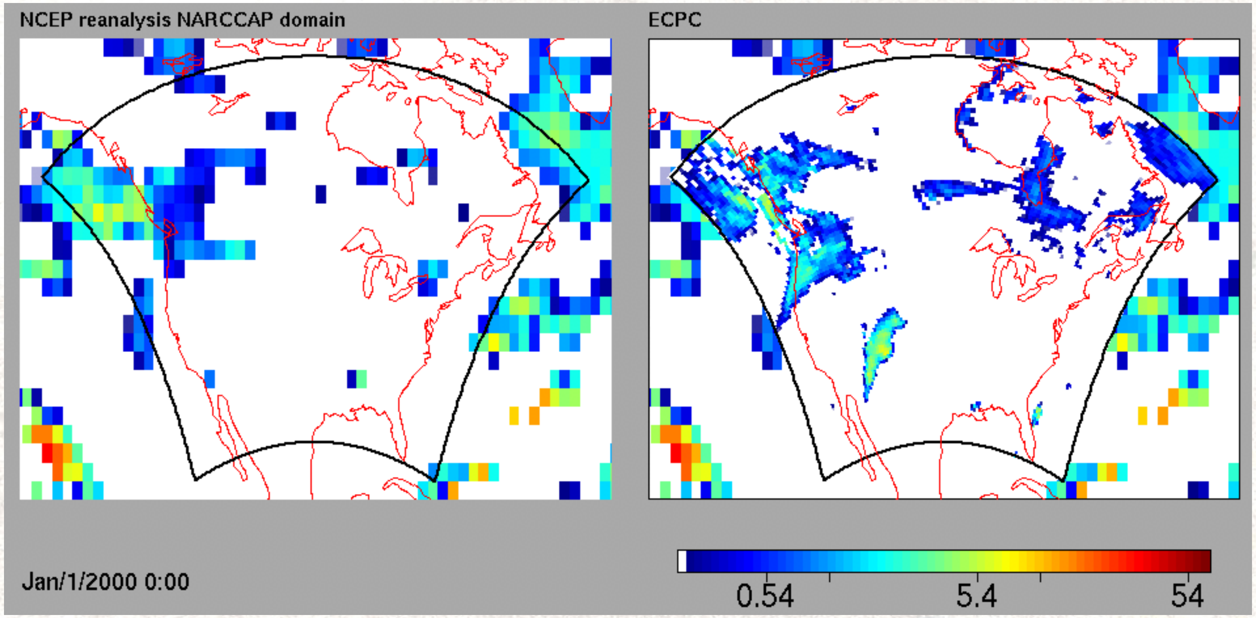
What is the uncertainty in the estimates?

A climate model grid box (?)

An approach to Regional Climate

- Nest a fine-scale weather model in part of a global model's domain.

Regional model simulates higher resolution weather based on the global model for boundary values and fluxes.

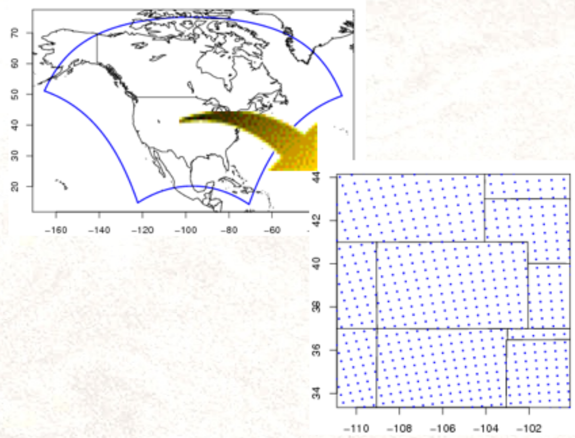


A snapshot from the 3-dimensional RSM3 model (right) forced by global observations (left)

- Consider different combinations of global and regional models to characterize model uncertainty.

NARCCAP – the design

4GCMS × 6RCMs:



12 runs – balanced half fraction design

- ● Driven by observations
- ■ 2× 2 subset

GLOBAL MODEL	REGIONAL MODELS					
	MM5I	WRF	HADRM	REGCM	RSM	CRCM
GFDL			●	●	○	
HADCM3	○		●		●	
CCSM	●	■				■
CGCM3		■		●		■
Reanalysis	●	●	●	●	●	●

A designed experiment is amenable to a statistical analysis and can contain more information.
But just 2-d temperatures fields are 72Gb of data.

Climate change

How will the seasonal cycle for temperature change in the future?

Additive model for curve fitting

Connection with data:

$$\boldsymbol{y}_i = g(x_i) + e_i$$

or

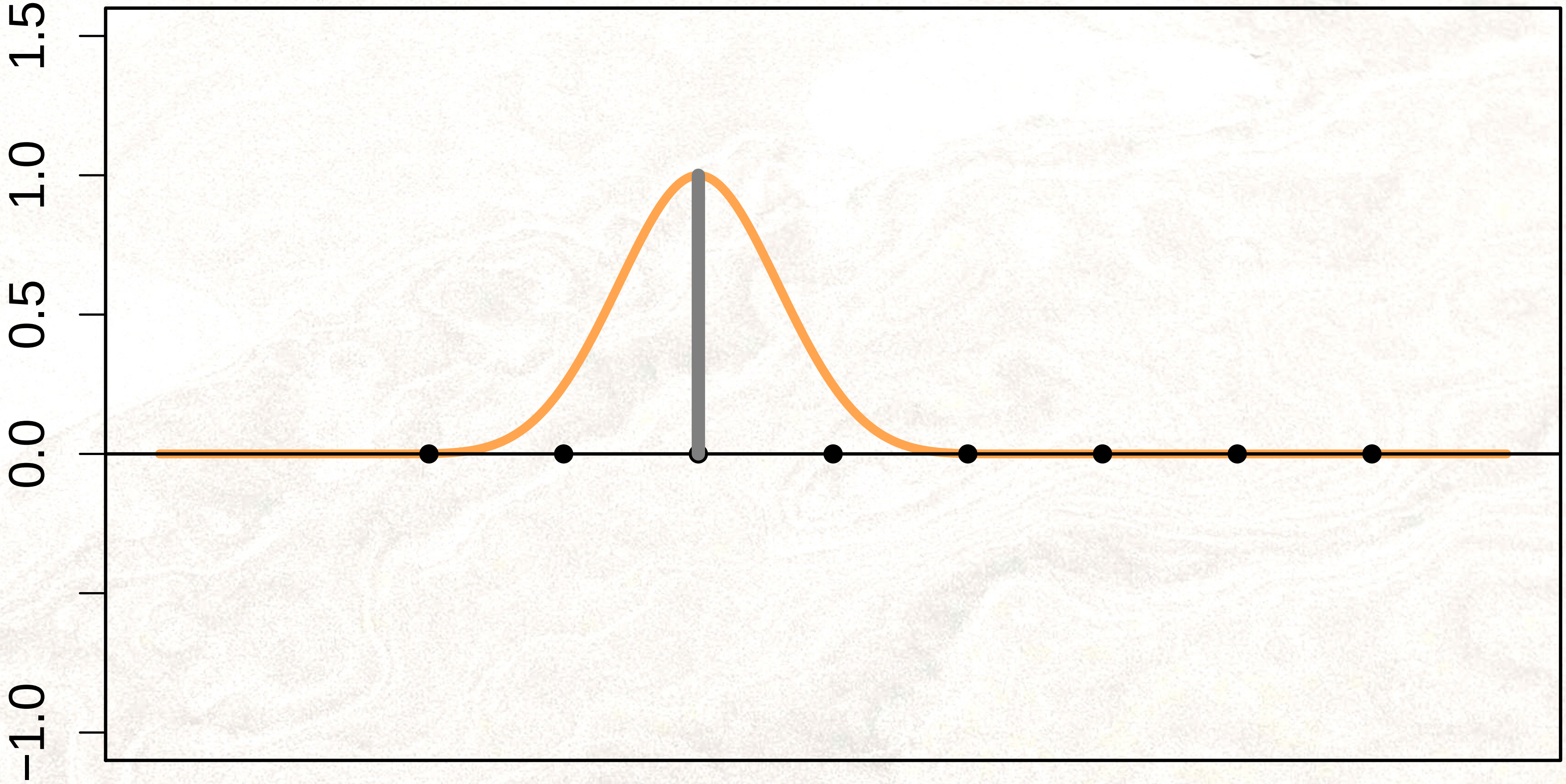
$$\boldsymbol{y}_i = L_i(g) + e_i$$

- Observations made at irregular locations
  – or as a linear functional

- ... and some random error added.

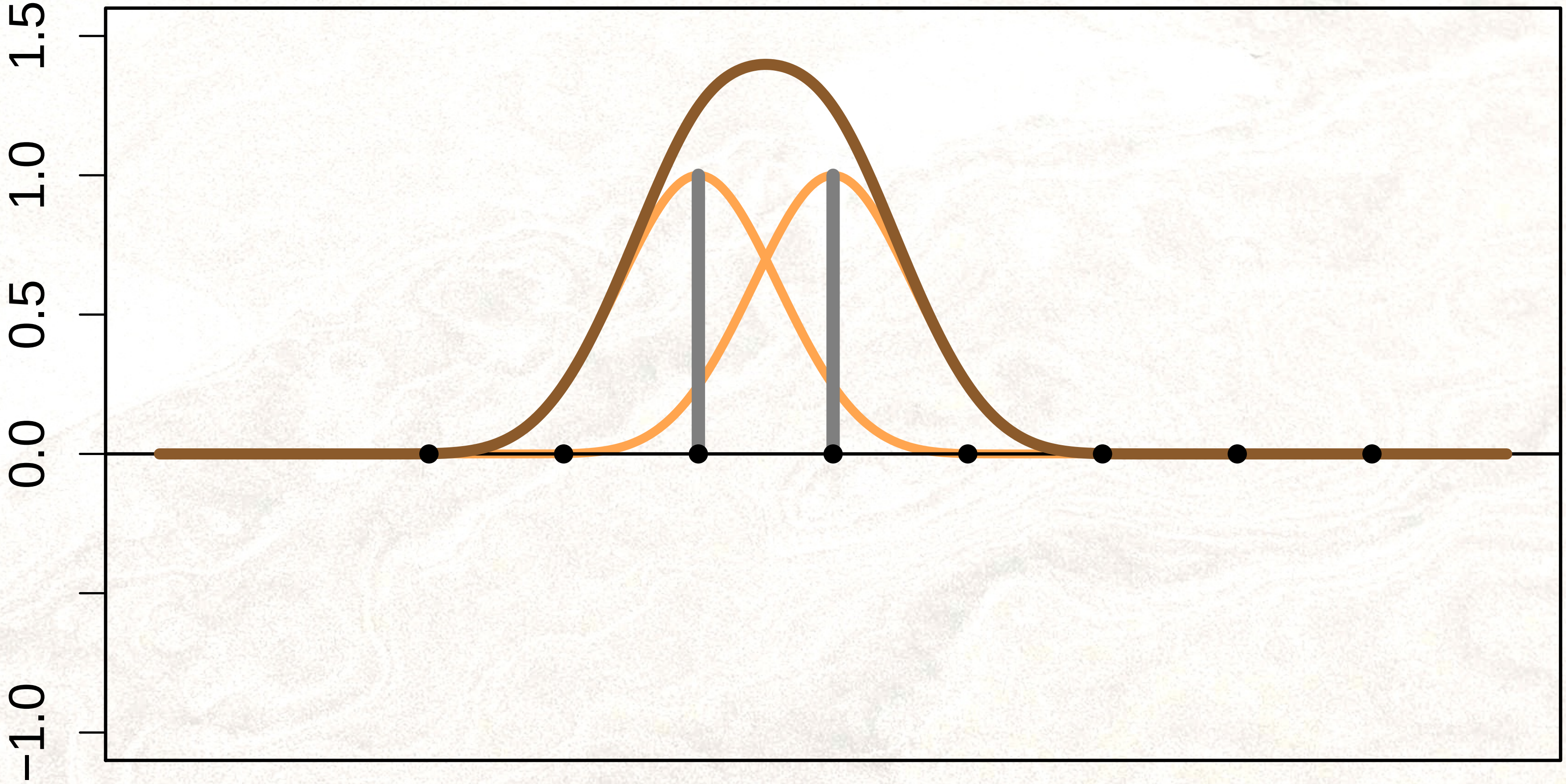
Representing the surface: $g(x) = \sum_j \phi_j(x) c_j$

Building a curve from bumps



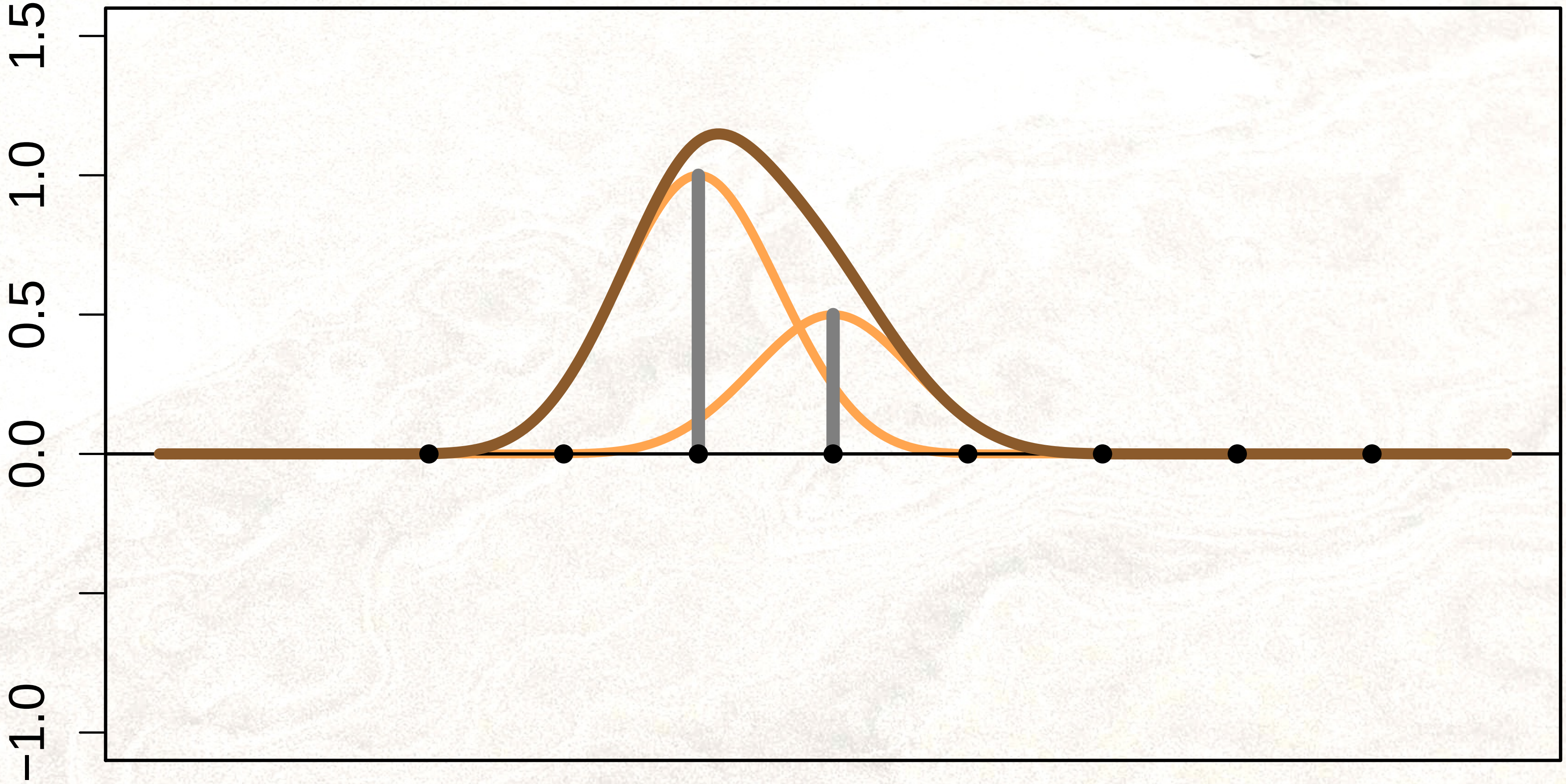
Single bump

Building a curve from bumps



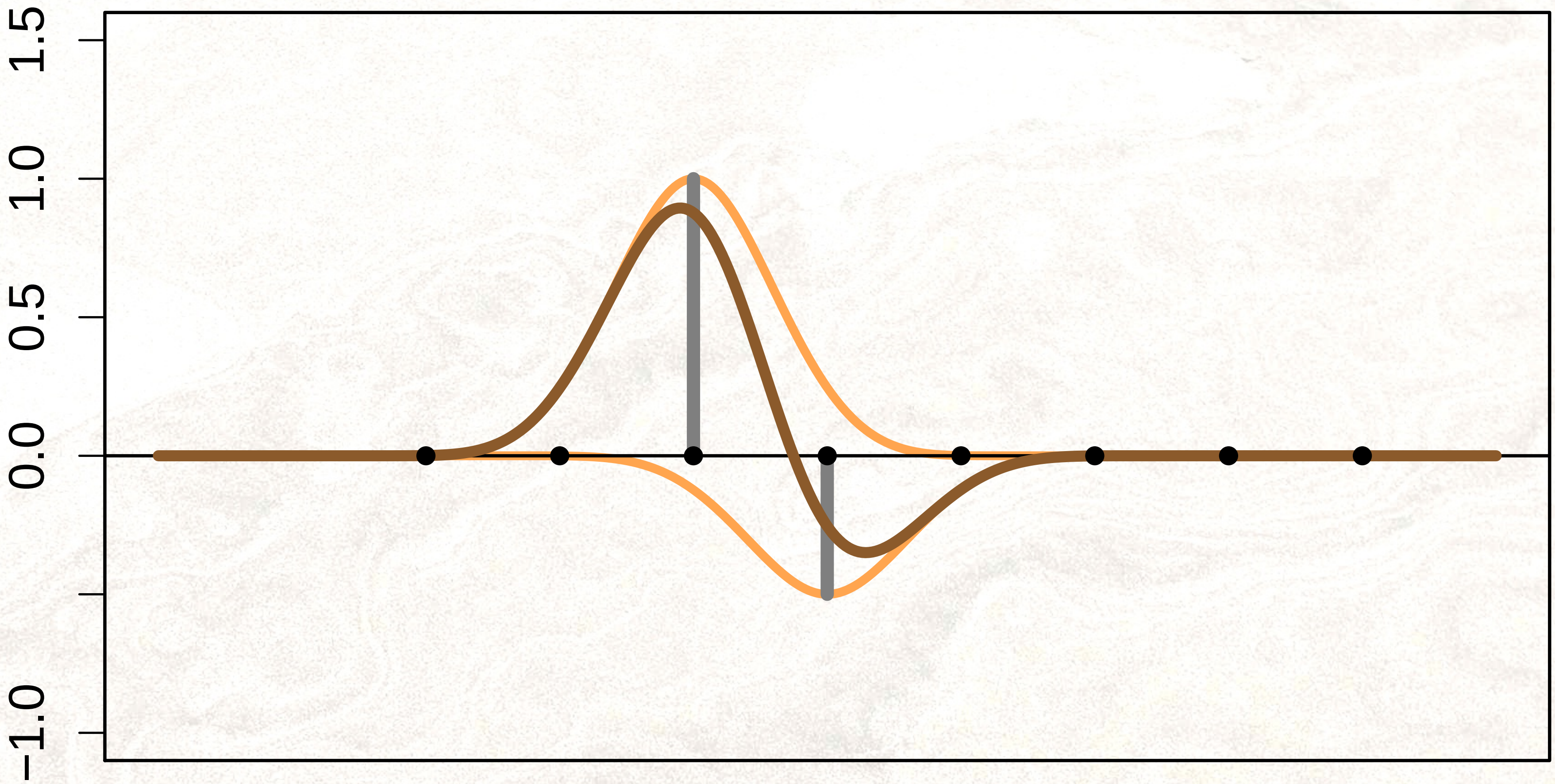
Two bumps same height

Building a curve from bumps



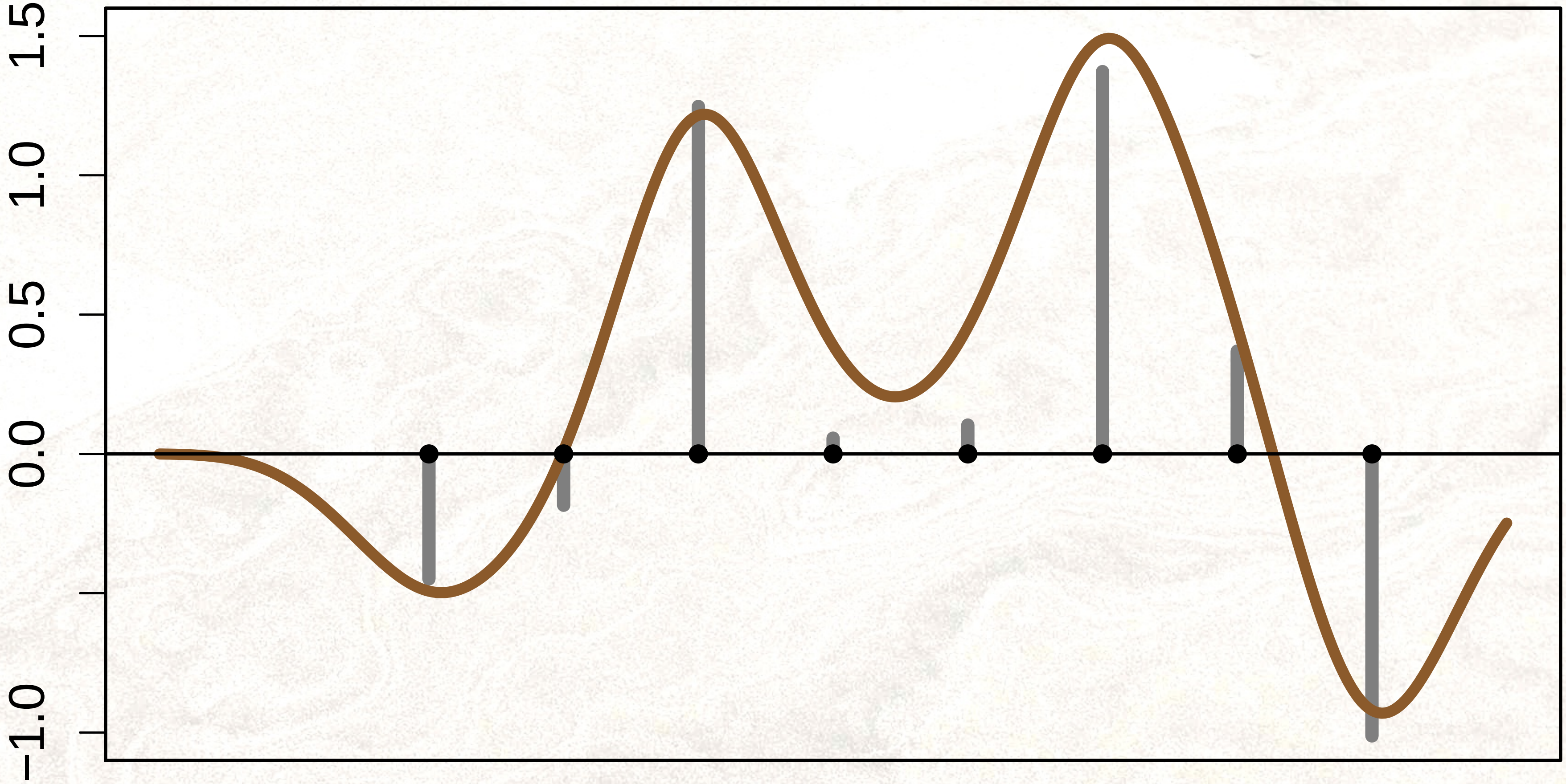
Two bumps different heights

Building a curve from bumps



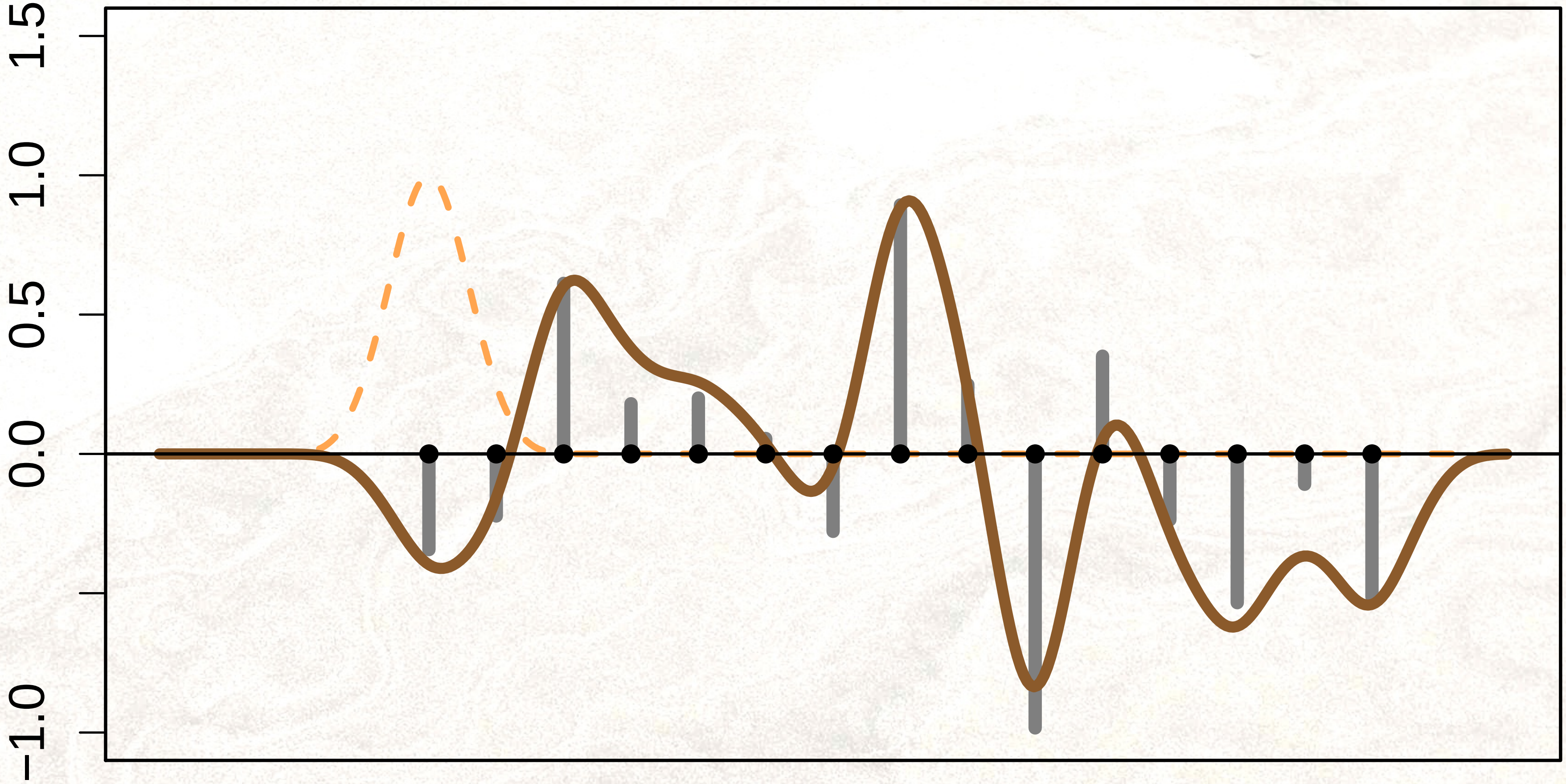
Two bumps different heights

Building a curve from bumps



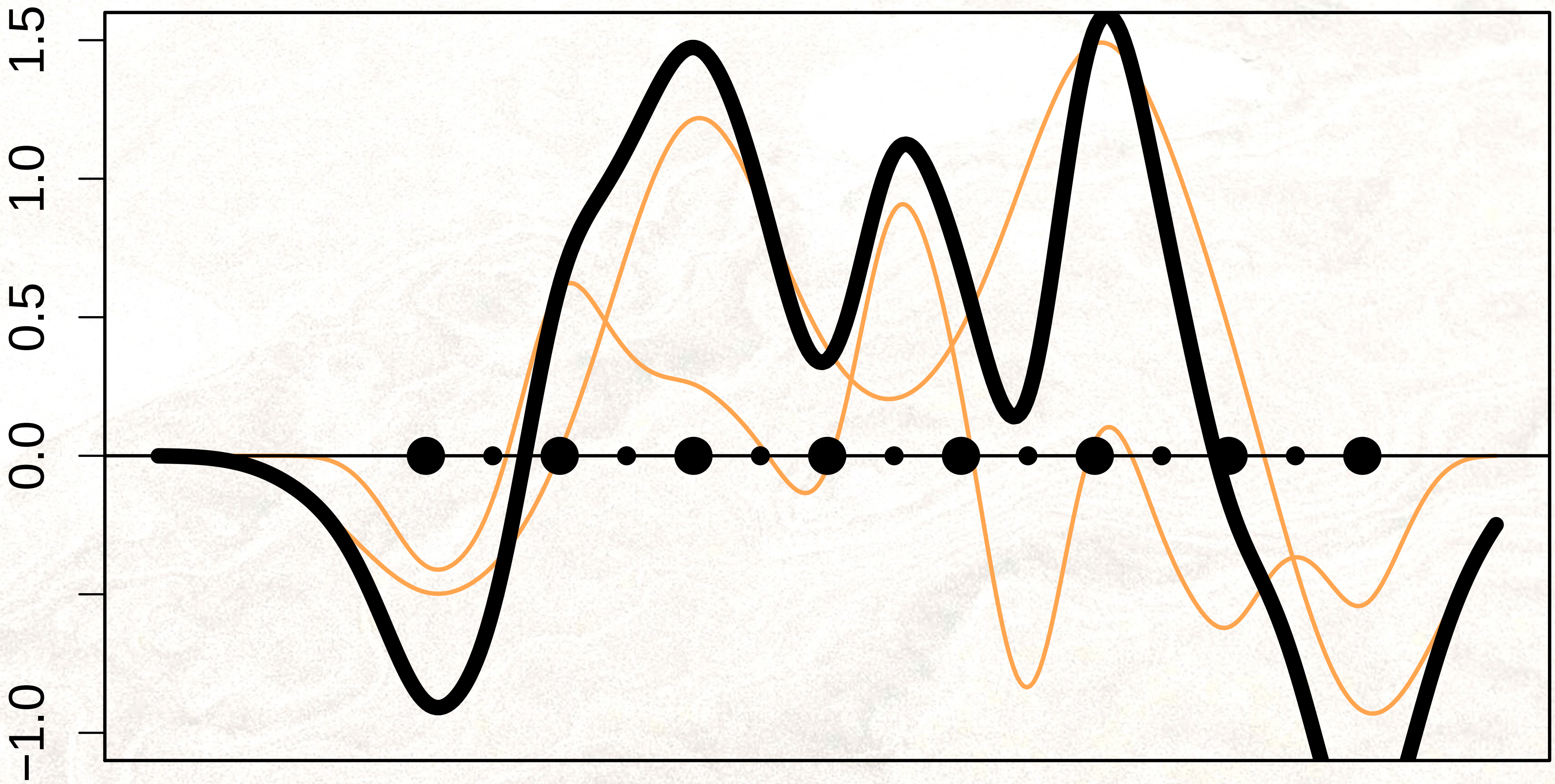
Eight bumps – all different heights

Building a curve from bumps



16 bumps – all different heights

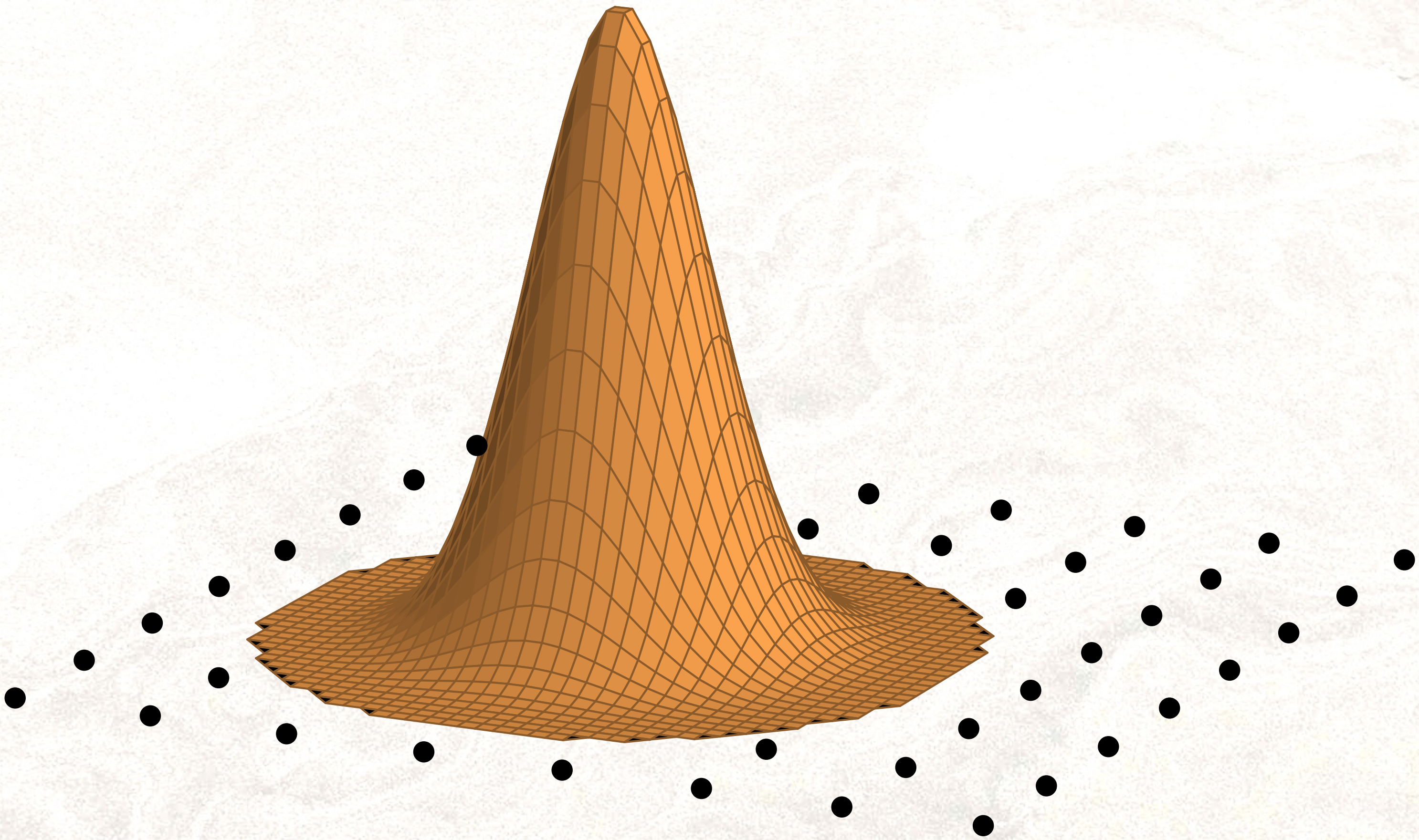
Building a curve from bumps



Adding them together

bumps = basis functions, bump heights = coefficients

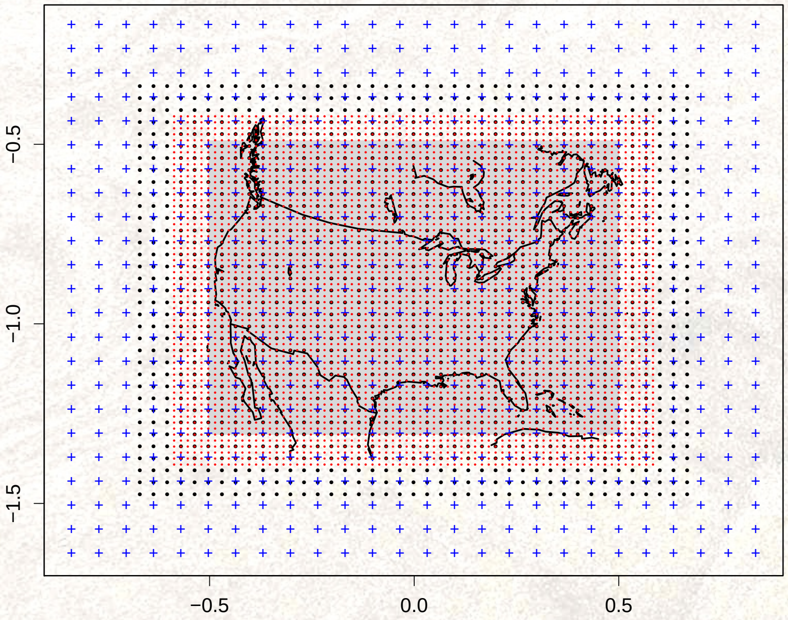
Going to two dimensions

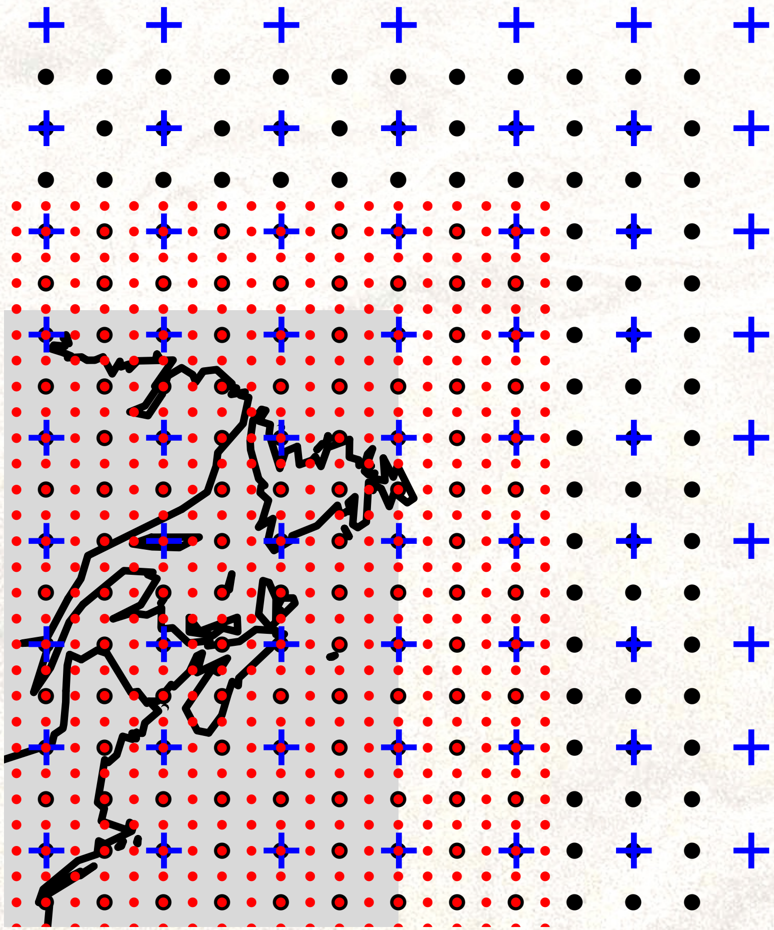


Example of a 2-d bump

A lattice example

- Three levels
- Extra points on margins to minimize edges effects
- About 4000 total lattice points





A statistical model for y and g

- X a regression matrix with $X_{i,j} = \phi_j(\boldsymbol{x}_i)$

–or some other linear operator applied to g, $X_{i,j} = L_i(\phi_j)$

Observations:

$$\boldsymbol{y} = X\boldsymbol{c} + \boldsymbol{e} \quad \boldsymbol{e} \sim MN(0, \sigma^2 I)$$

Process:

$$g(x) = \sum_j \phi_j(x) c_j, \quad \boldsymbol{c} \sim MN(0, \rho Q^{-1})$$

Potential Priors:

$$[\rho, \sigma^2, Q]$$

Part of a Gibbs sampler

"Full conditional for coefficients": $[\boldsymbol{c}|\boldsymbol{y}, \rho, \sigma^2, Q]$

Multivariate normal with mean:

$$\hat{c} = (\boldsymbol{X}^T \boldsymbol{X} + (\sigma^2/\rho) Q)^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

Precision:

$$(1/\sigma^2) \boldsymbol{X}^T \boldsymbol{X} + (1/\rho) Q$$

- Create a model where all matrices are sparse and finding $\hat{c}$ is fast
- Sampling from full conditional is also fast.
- Likelihood/posterior computation for ρ, σ^2, Q dominated by

$$det((1/\sigma^2) \boldsymbol{X}^T \boldsymbol{X} + (1/\rho) Q_a))$$

More about Q

Some coefficients:

$$\begin{array}{ccccc} . & . & . & . & . \\ . & . & c_1 & . & . \\ . & c_2 & c_* & c_3 & . \\ . & . & c_4 & . & . \\ . & . & . & . & . \end{array}$$

Some weights:

$$\begin{array}{ccccc} . & . & . & . & . \\ . & . & -1 & . & . \\ . & -1 & a & -1 & . \\ . & . & -1 & . & . \\ . & . & . & . & . \end{array}$$

A spatial autoregression:

$\boldsymbol{B}$ a matrix where each row has 4 nonzero weights corresponding to the first order neighbors and diagonal element, a

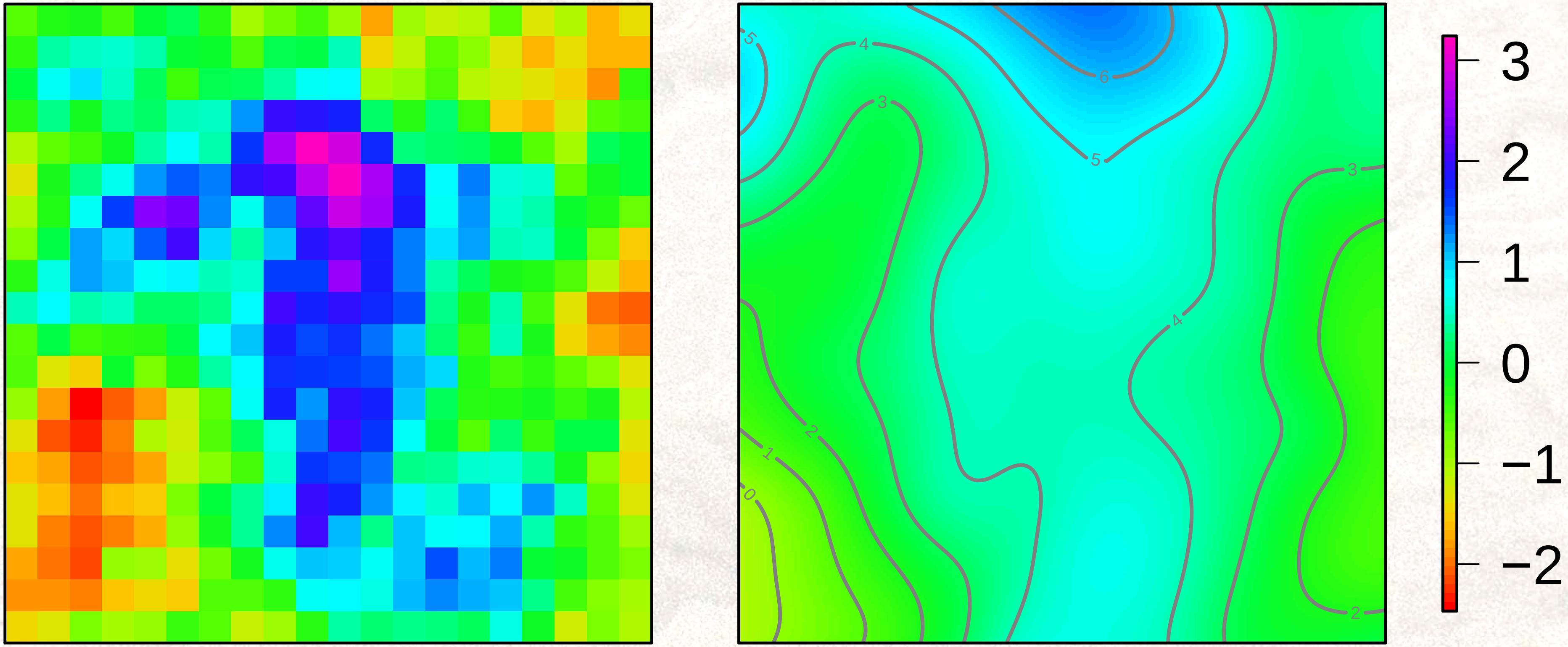
$$\boldsymbol{Bc} = \text{iid } N(0,1)$$

- a needs to be greater than 4, related to a range parameter.
- Precision matrix $\boldsymbol{Q} = \boldsymbol{B}^T\boldsymbol{B}$ Covariance matrix $= \boldsymbol{Q}^{-1} = \boldsymbol{B}^{-1}\boldsymbol{B}^{-T}$

Applying the basis functions

16×16 example with $a = 4.01$

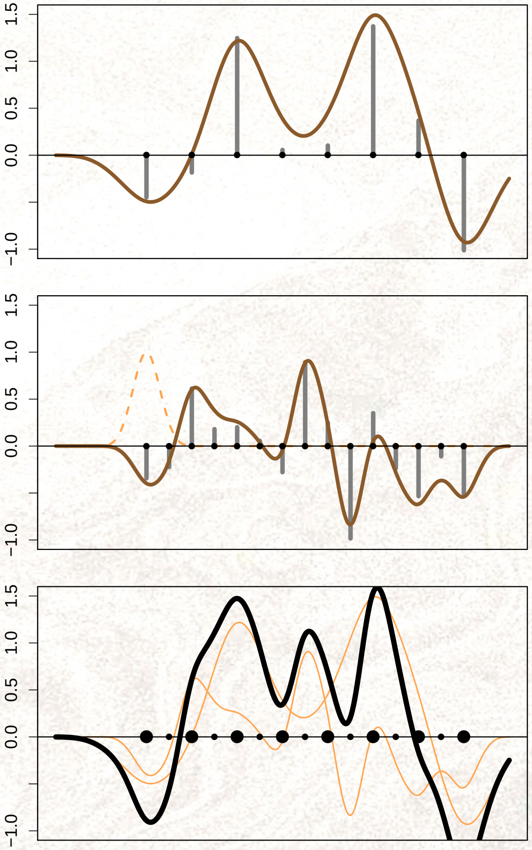
Coefficients on the lattice *Expanding with basis functions*



$$c_k \rightarrow \sum \phi_k(x) c_k$$

More than one level:

Adding different resolutions together:



$$g(\boldsymbol{x}) = \rho(\alpha_1 g_1(\boldsymbol{x}) + \alpha_2 g_2(\boldsymbol{x}) + \alpha_3 g_3(\boldsymbol{x}) + \ldots)$$

$$Q = (1/\rho) \begin{bmatrix} \alpha_1 B_1^T B_1 & 0 & 0 \\ 0 & \alpha_2 B_2^T B_2 & 0 \\ 0 & 0 & \alpha_3 B_3^T B_3 \end{bmatrix}$$

- ρ marginal variance of the process
- $\alpha_1, \alpha_2, \alpha_3$ relative weight for each level.

Kriging

Danie G. Krige

South African Mining Engineer who pioneered the field of geostatistics.

Kriging= Krig[e] + ing

Methodology for estimating a surface based on irregular observations.

A view of Kriging as a minimization problem

Kimeldorf and Wahba (1970)

(fit of the surface to the data) + (roughness of the surface)

- Want a surface that tracks the observations but is not overly rough and irregular.

The equivalent variational problem:

$$\min_{\boldsymbol{c}} (\boldsymbol{y} - X\boldsymbol{c})^T (\boldsymbol{y} - X\boldsymbol{c}) + \lambda \boldsymbol{c}^T \boldsymbol{Q}\boldsymbol{c}$$

- $\boldsymbol{y}$ the data, X matrix of basis functions, $\boldsymbol{c}$ coefficients, Q roughness matrix.

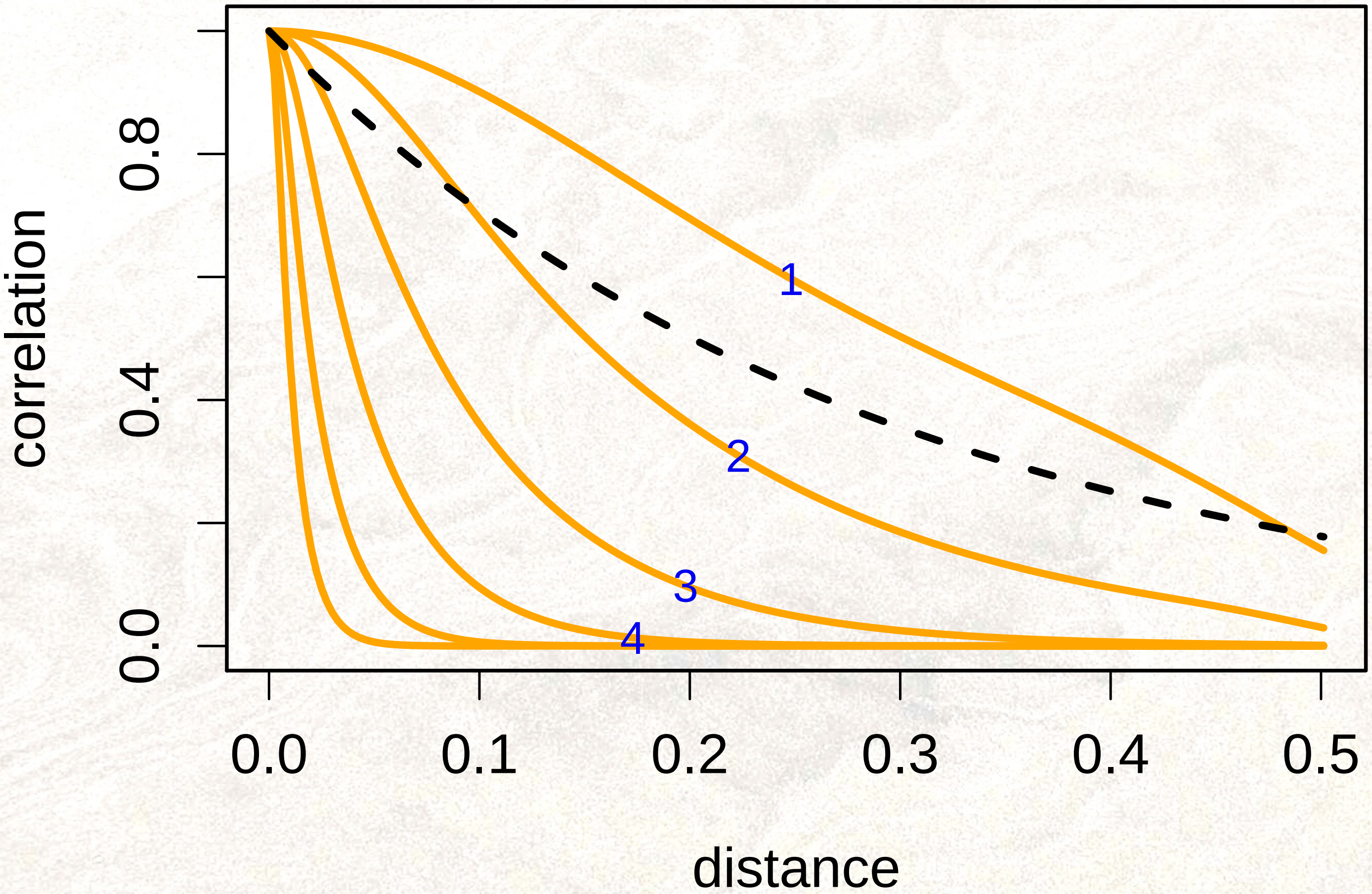
- Q is a penalty matrix for $\boldsymbol{c}$
  Minimizer: $\hat{c} = (\boldsymbol{X}^T\boldsymbol{X} + (\sigma^2/\rho)Q)^{-1}\boldsymbol{X}^T\boldsymbol{y}$

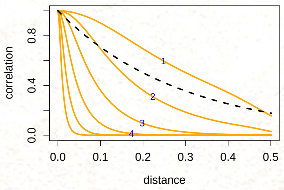
X is really any matrix that connects the data to the coefficients. (E.g. $L_i(g)$)

Benefits of a multi-resolution

Approximating an exponential covariance

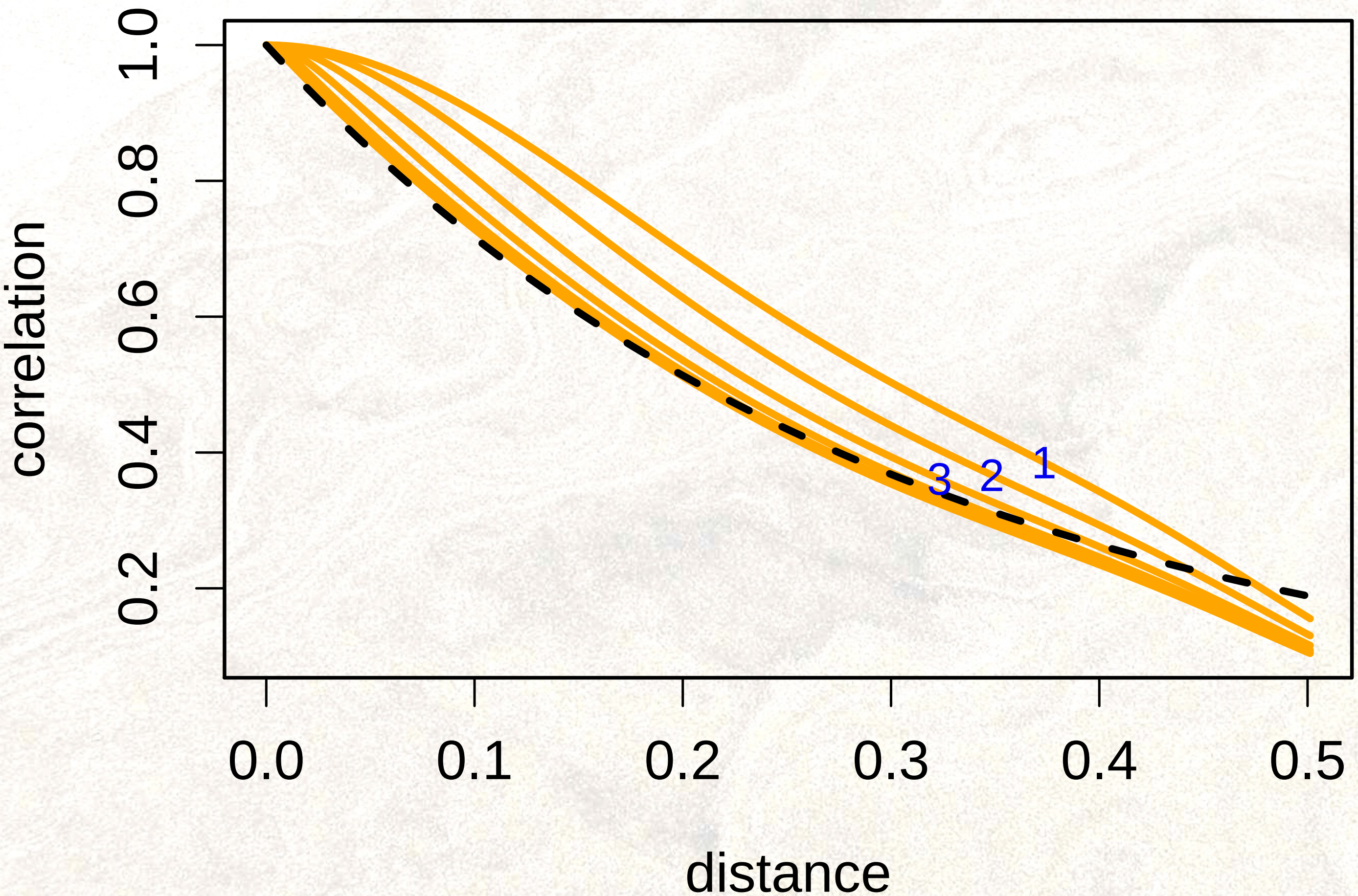
Correlation functions for 6 levels and a target exponential





Weighting by $2^{-level/2}$

Correlation functions adding levels and the target exponential



Timing

On my mac laptop and in R

— i.e. a single core and `LatticeKrig`

- Computation may be dominated by : matrix setup

normalization to stationarity

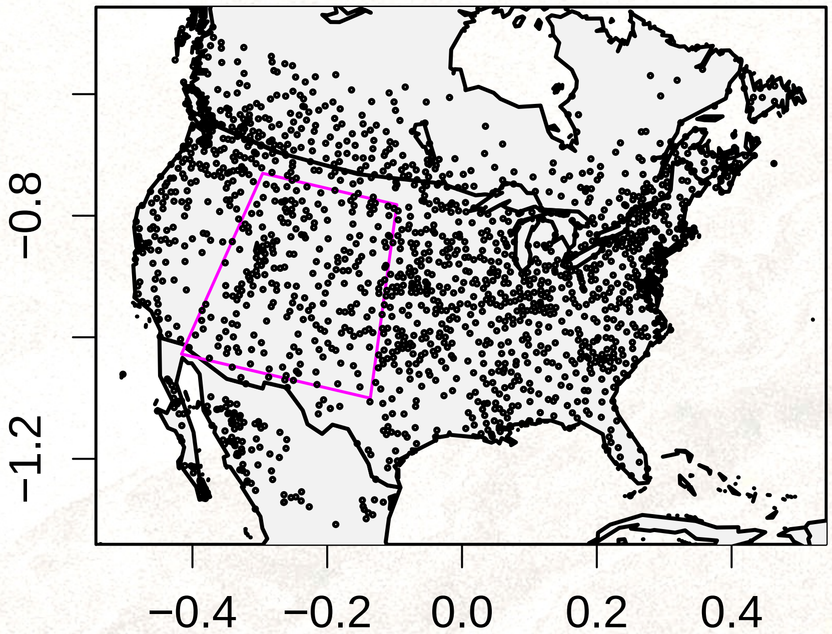
Cholesky decomposition

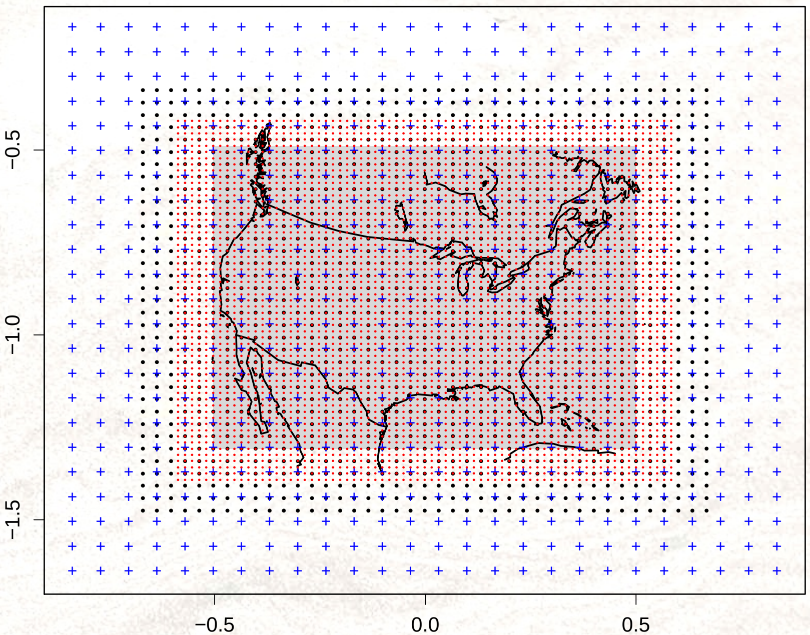
- For 20,000 observations:

the standard Kriging (dense Cholesky) is $\approx$ 20 minutes

`LatticeKrig` (sparse Cholesky) is $\approx$ 10 seconds.

NA Summer rainfall



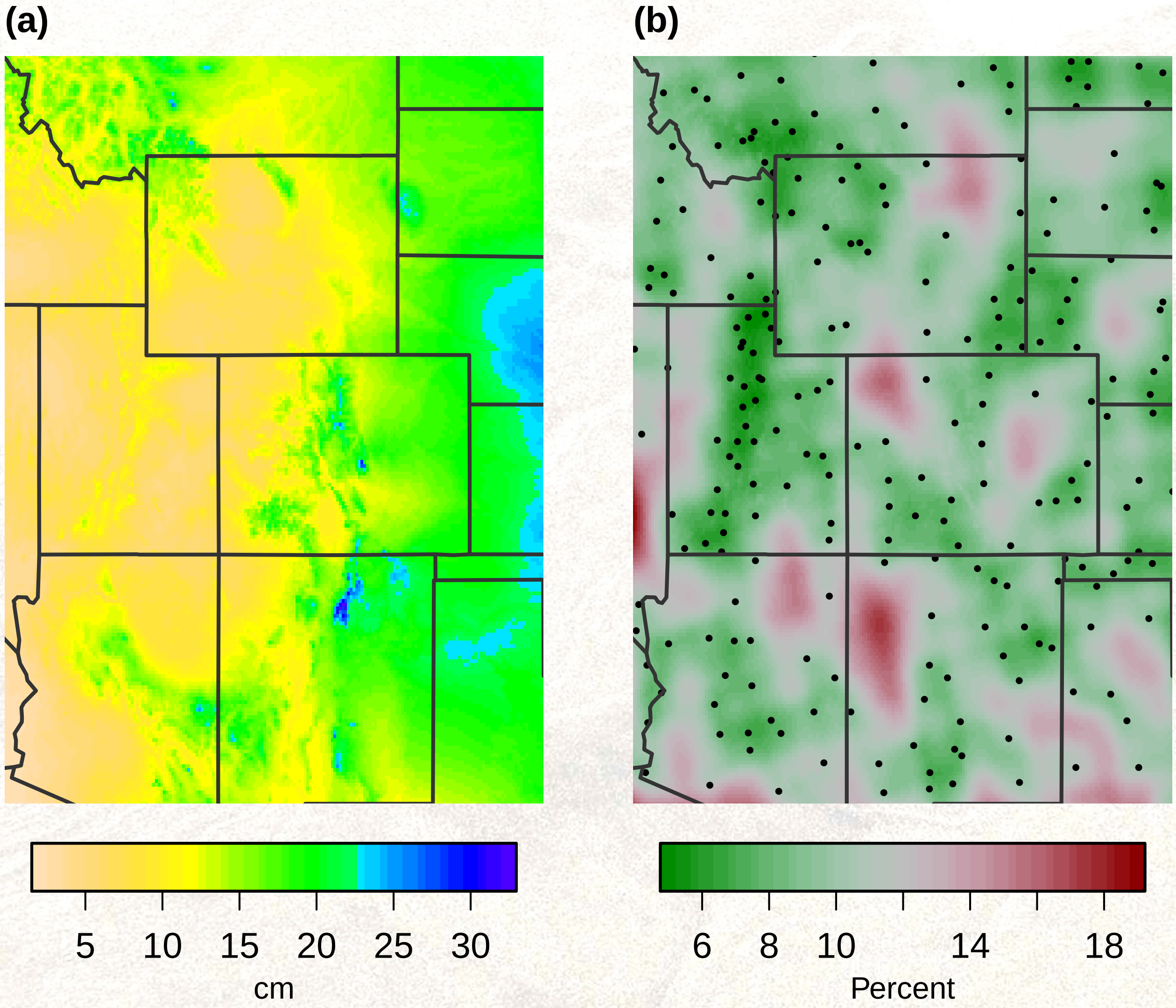


Three levels of resolution

- ≈ 4000 basis functions total.
- statistical parameters found by maximum likelihood
- coefficients found by "kriging"
- uncertainty found by Monte Carlo ensemble
- includes linear adjustment for elevation

Estimated summer rainfall

Predicted JJA rainfall (cm) Pointwise standard errors (percent)



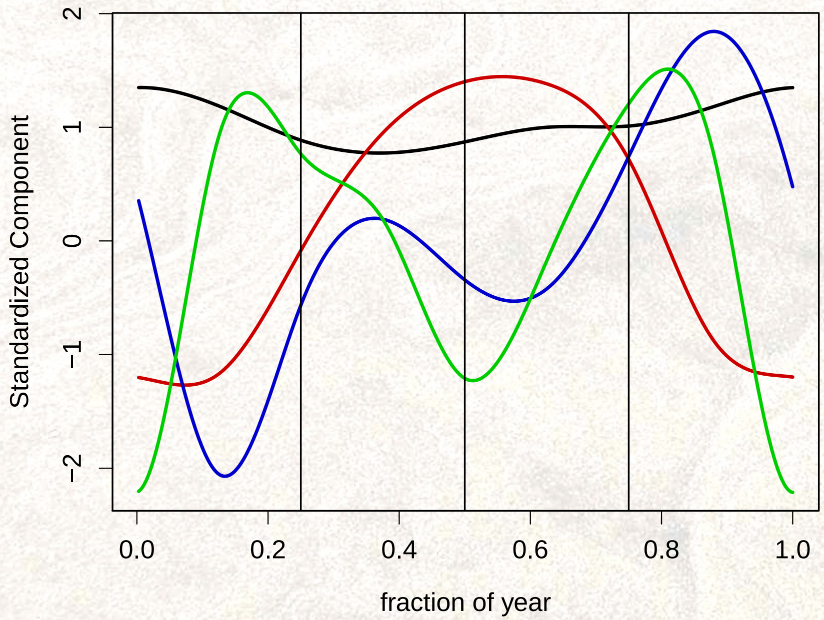
Climate change

How will the seasonal cycle for temperature change in the future?

Back to NARCCAP

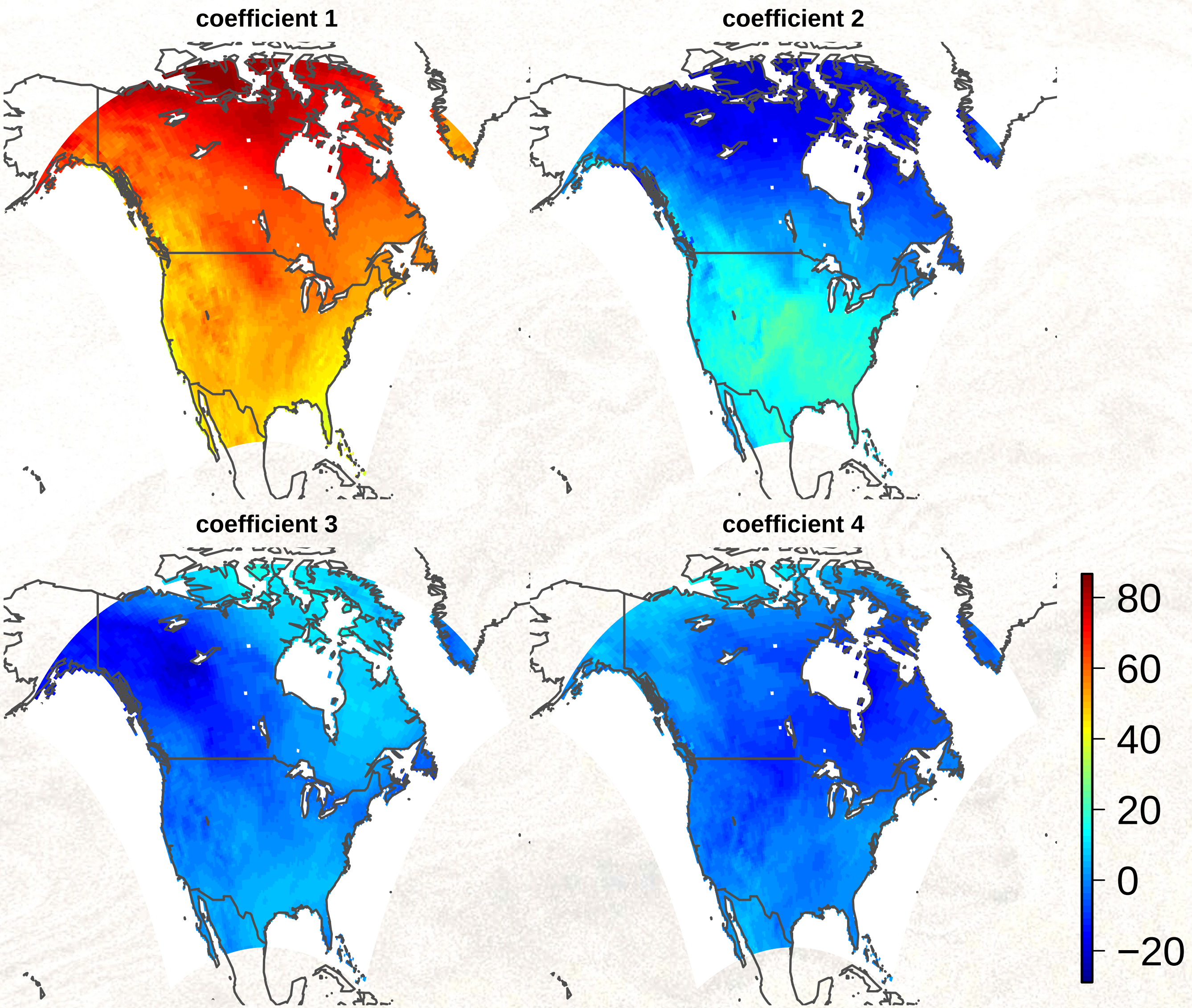
- A 2×2 subset of NARCCAP (4 global/regional combinations)
- (Future - Present) seasonal cycle expand in 4 principle components ... gives 4 "amplitude" spatial fields for each model.
- Approximately 9000 spatial locations

Seasonal PCs
(future - present)



NARCCAP domain

Coefficient fields – CRCMccsm

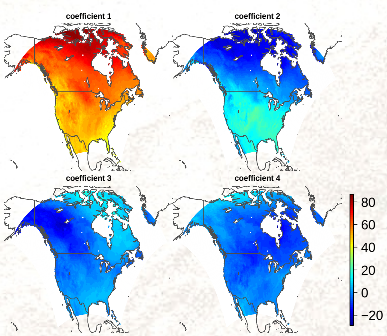


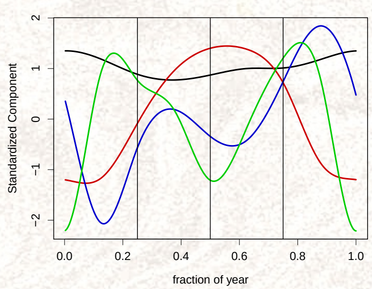
There are four of these!

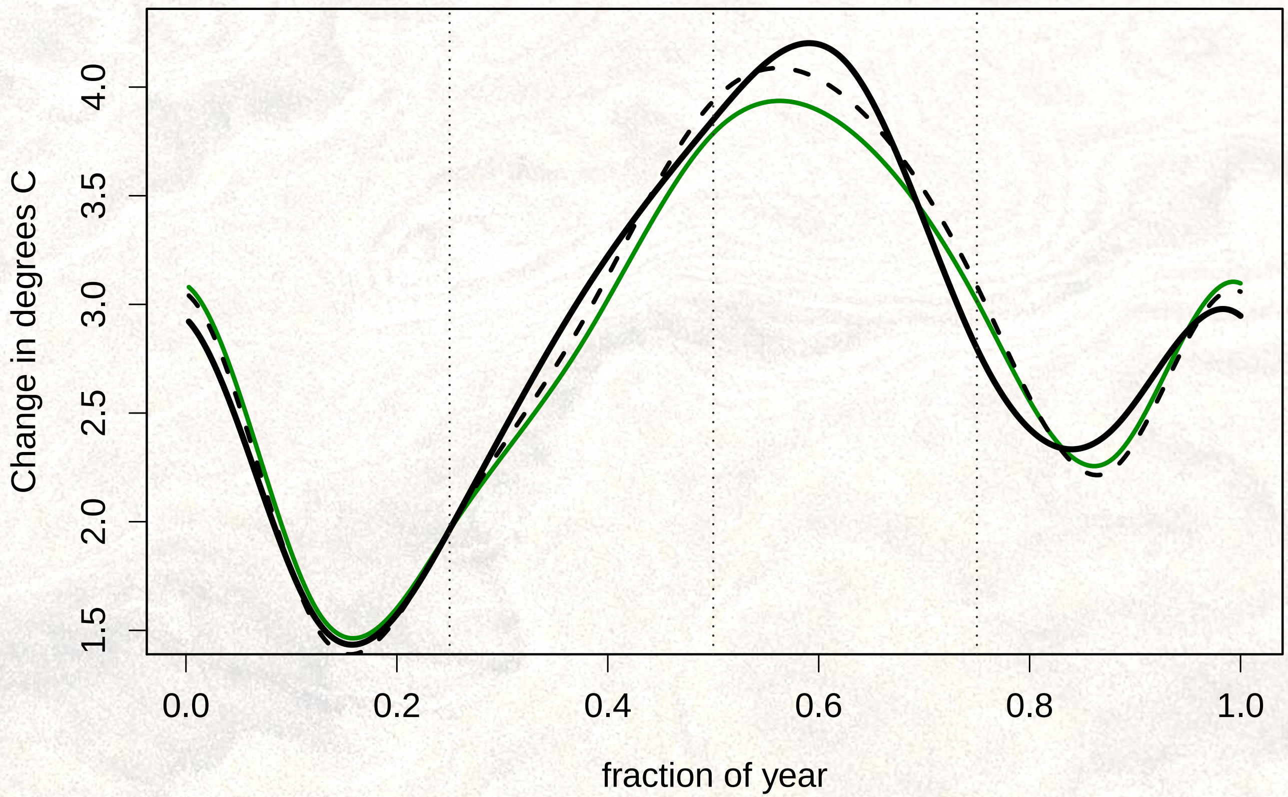
Example for Boulder grid box

change in season $= \alpha_1 PC_1 + \cdots + \alpha_4 PC_4$

Results for one regional model (CRCM/ccsm)







Solid - Raw, Dashed - projection to 4 EOFS/PCs,

With spatial smoothing

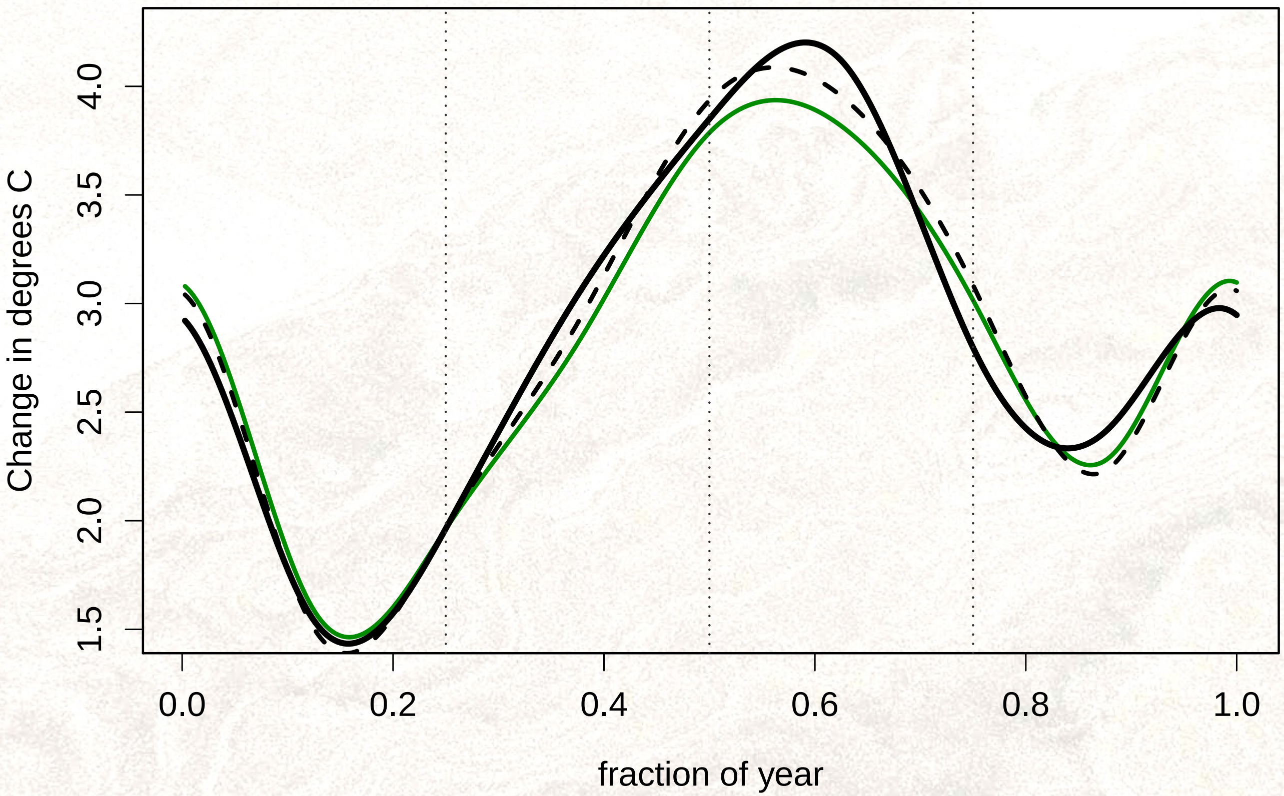
Spatial model

- *Four coefficients of seasonal profile for the four model combinations – and at each grid box*

$4\times\ 4 = 16$ fields total each with 9K locations.

- Smooth the 16 fields with LatticeKrig model using covariance close to a thin plate spline.

Results for Boulder grid box and CRCM/ccsm

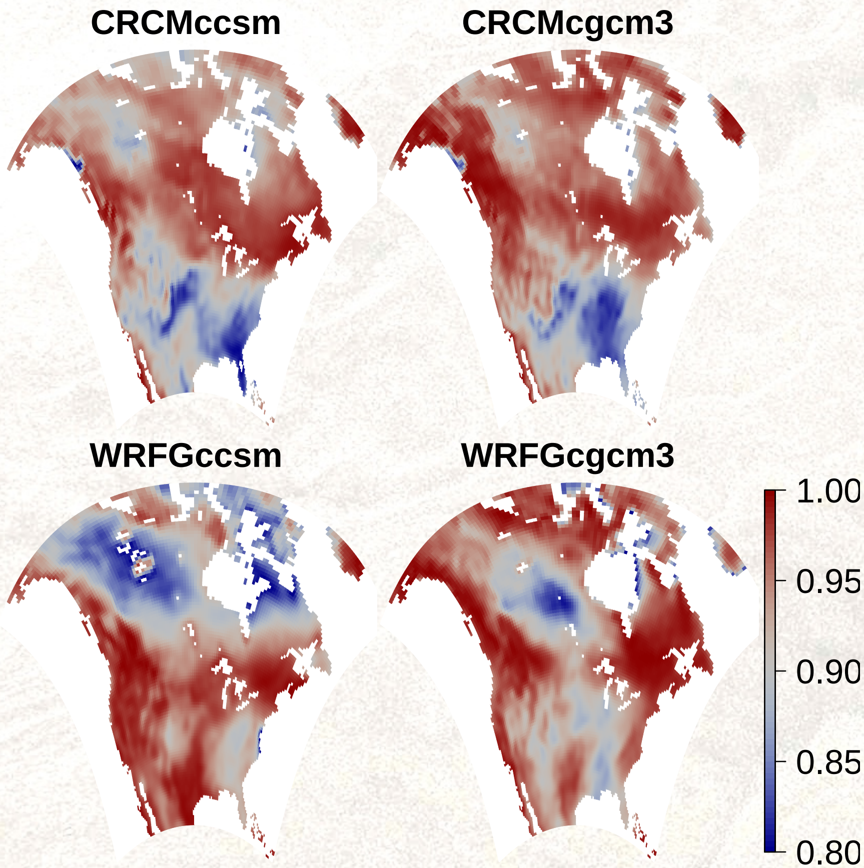


Solid - Raw, Dashed - projection to4 EOFS/PCs,
With spatial smoothing

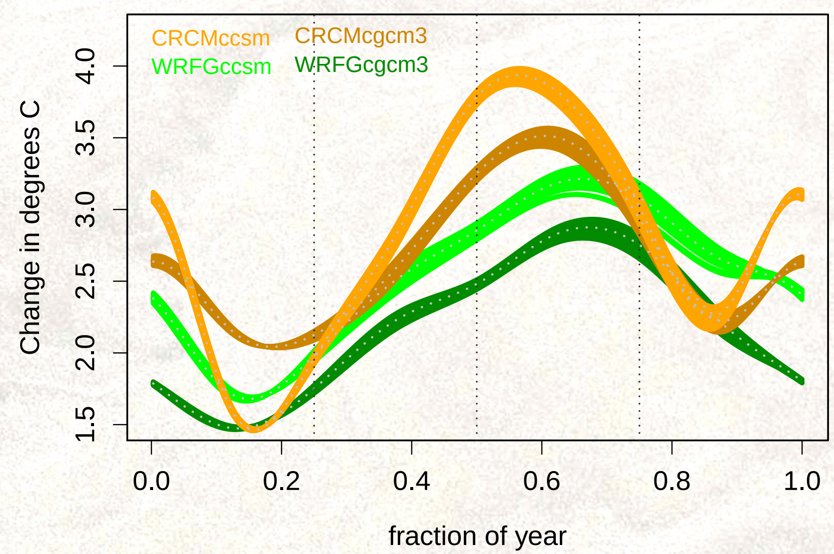
Results

- Thin plate spline-like model (1 level $120 \times 55 \approx 6000$ basis functions)
- λ found by MLE (equivalent to sill and nugget)
- Conditional simulation of fields (facilitates nonlinear statistics)

R^2 for first PC



Inference for Boulder grid box

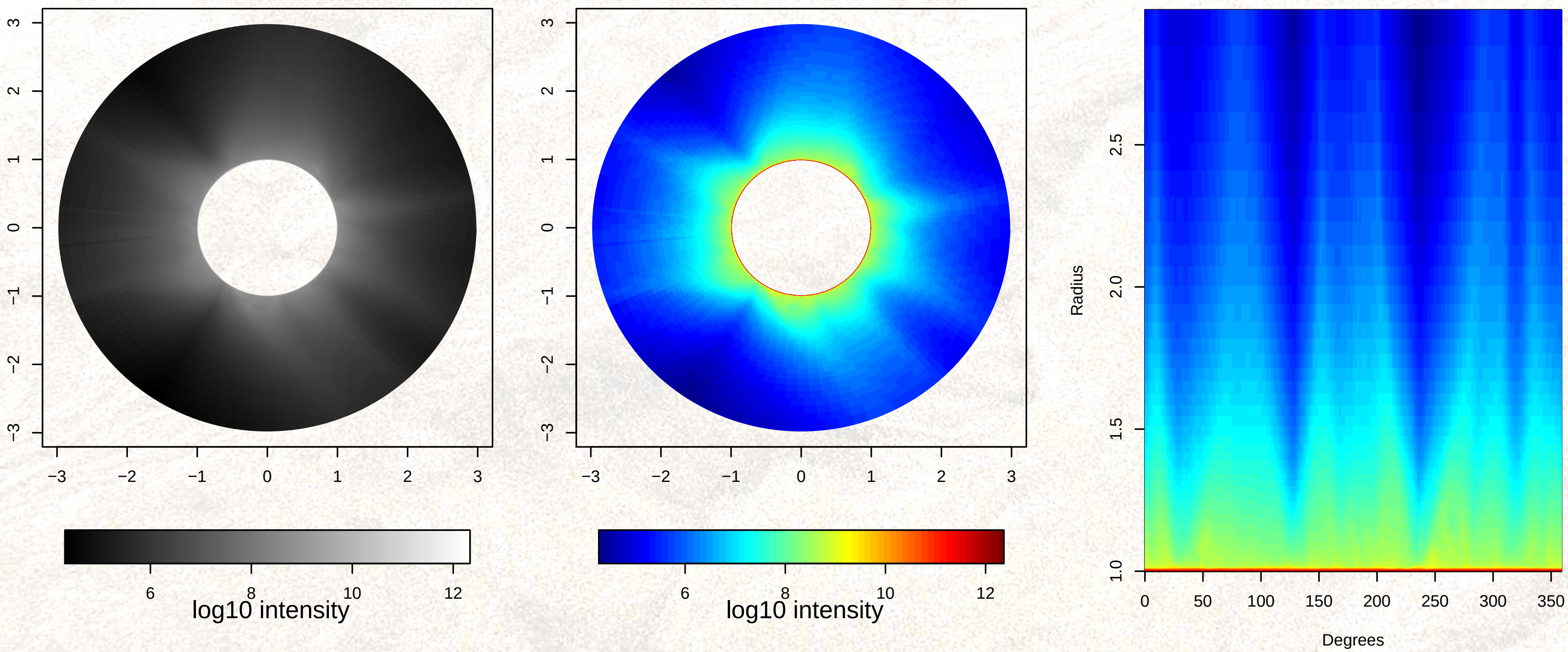


Electron density in the corona

(Luke Burnett, Kevin Delmasse, Sarah Gibson)

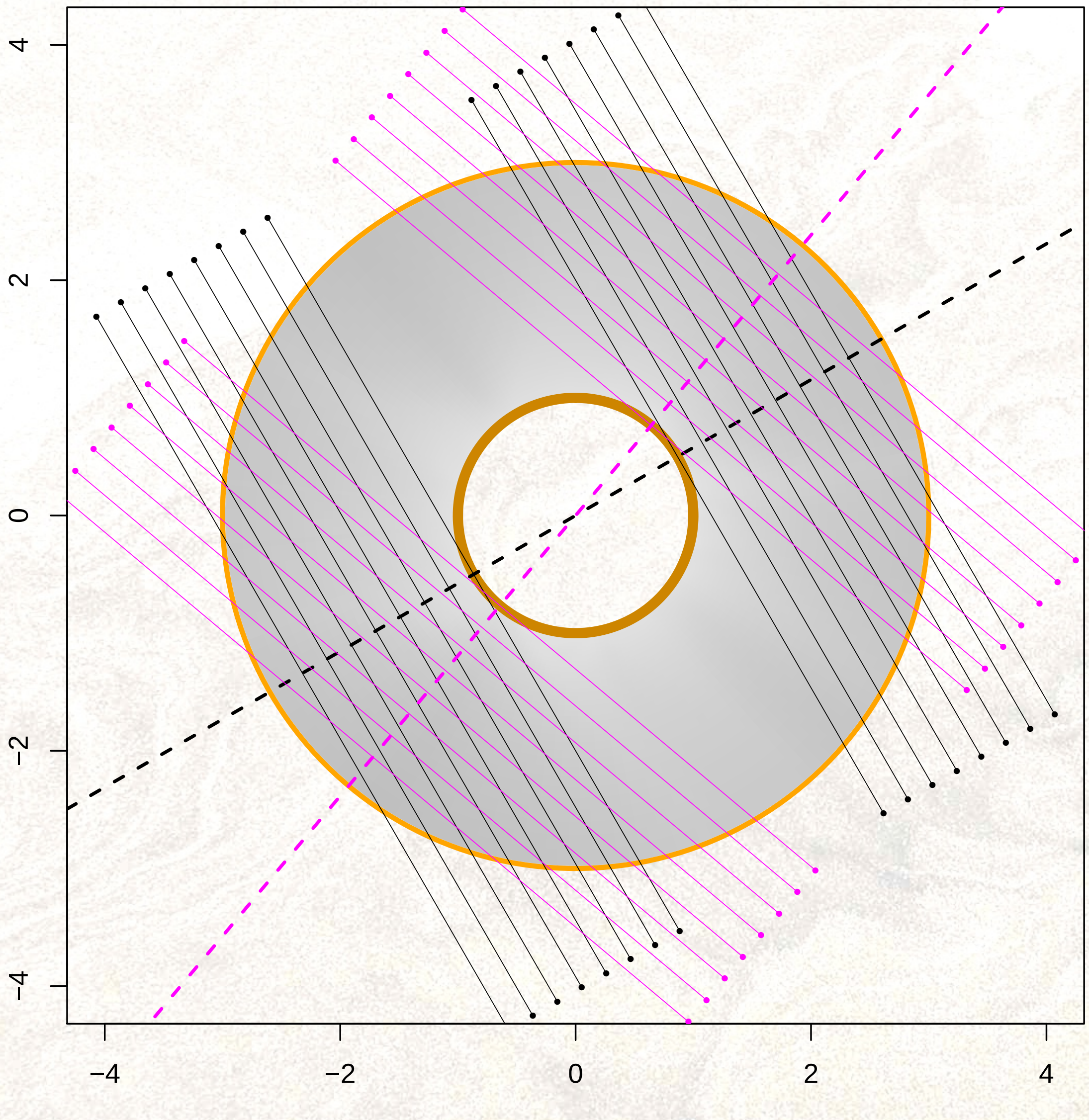
- Observations are integrals through corona.
- Goal is reconstruction of the density based on different viewing angles.

Equitorial slice for electron density
(*Predictive Science product* time = 2144^{th} Carrington rotation)



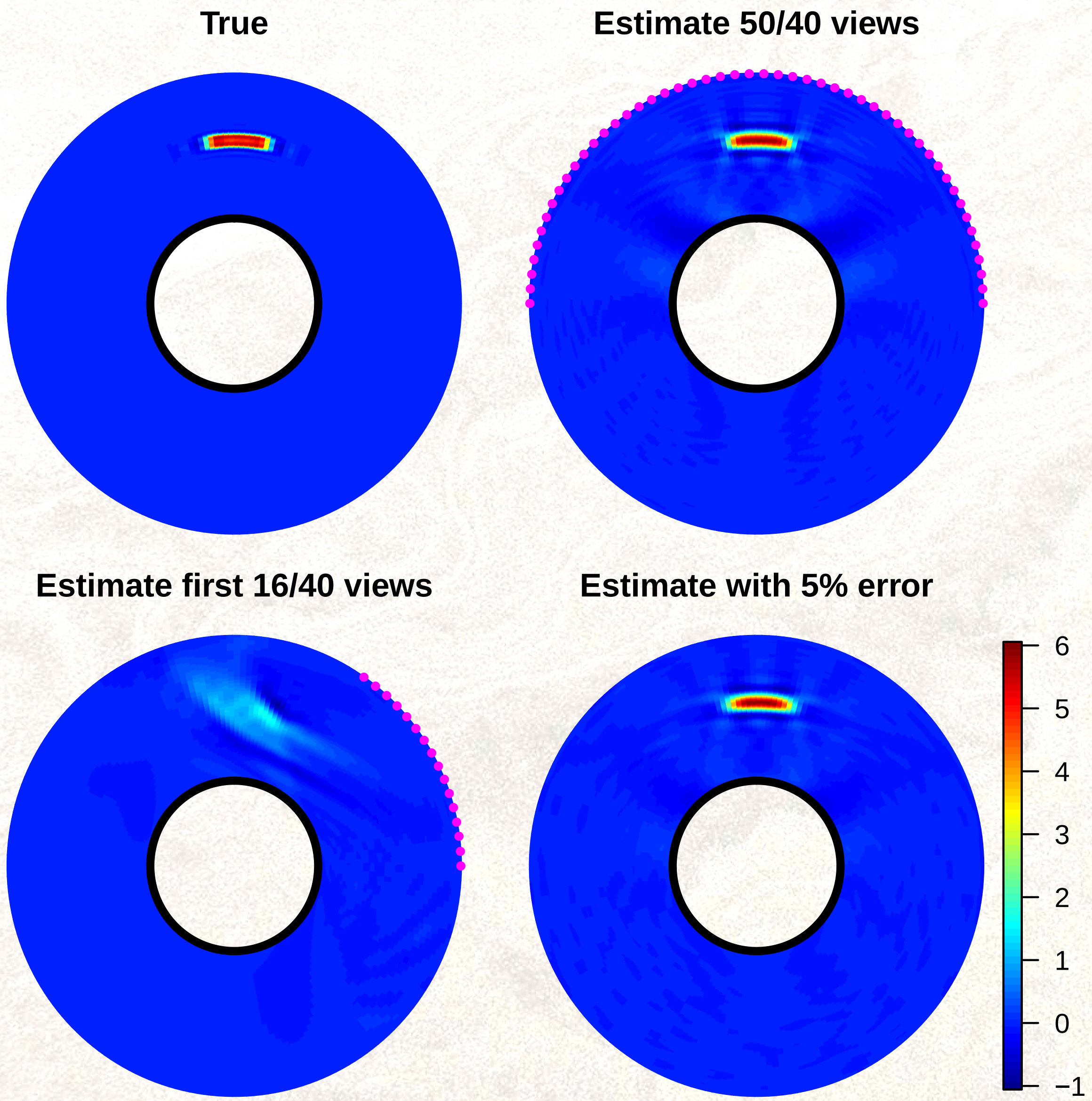
Observations of Corona

Two viewing angles each with 16 lines of sight: (2/16)



Reconstructions of simple phantom

LatticeKrig with $\sim$ 5000 basis functions,
50 angles with 40 lines of sight each.



Summary

- **Computational efficiency gained by compact basis functions and sparse roughness (precision) matrix.**

- **Multi-resolution can approximate standard covariance families (e.g. Matern)**

- **Easy to generate uncertainty measures.**

Exploit parallel strategies for larger problems

See `LatticeKrig` contributed package in R

Thank you!