Multi-resolution models for large data sets

Douglas Nychka,

National Center for Atmospheric Research

National Science Foundation

SIAM Austin May, 2013

Credits

- Steve Sain, Tamra Greasby, NCAR
- Dorit Hammerling, SAMSI
- Soutir Bandyopadhyay, Lehigh
- Finn Lindgren, U Bath, UK
- James Gattiker, LANL

Outline

- Surface observations of rainfall
- Regional Climate simulation and NARCCAP
- Compact basis functions (Φ),

  Markov Random fields (H)
- The multi-resolution model
- Covariance for summer precipitation.
- Changes in the seasonality for future climate

Key idea: Introduce a sparse basis and precision matrices without compromising the spatial model.

Estimating a curve or surface.

An additive statistical model:

Given n pairs of observations (x_i, y_i), $i = 1, \ldots, n$

$$y_i = g(x_i) + \epsilon_i$$

ϵ_i's are random errors and g is an unknown, smooth realization of a Gaussian process.

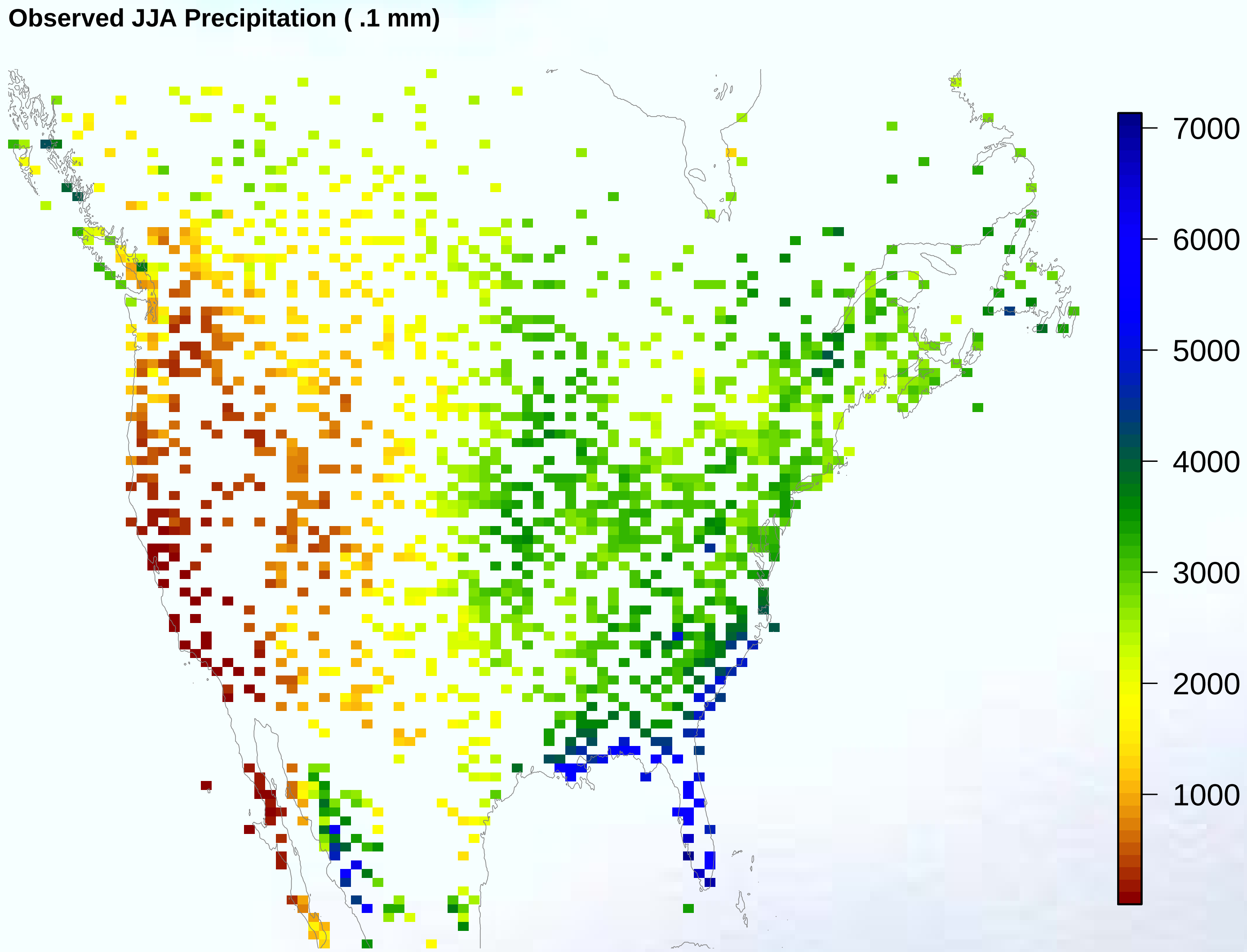
Estimate $g(\boldsymbol{x})$

Quantify the uncertainty of the estimate ...

Statistical perspective: You need a model

Observed mean summer precipitation

1720 stations reporting, "mean" for 1950-2010



Current Climate

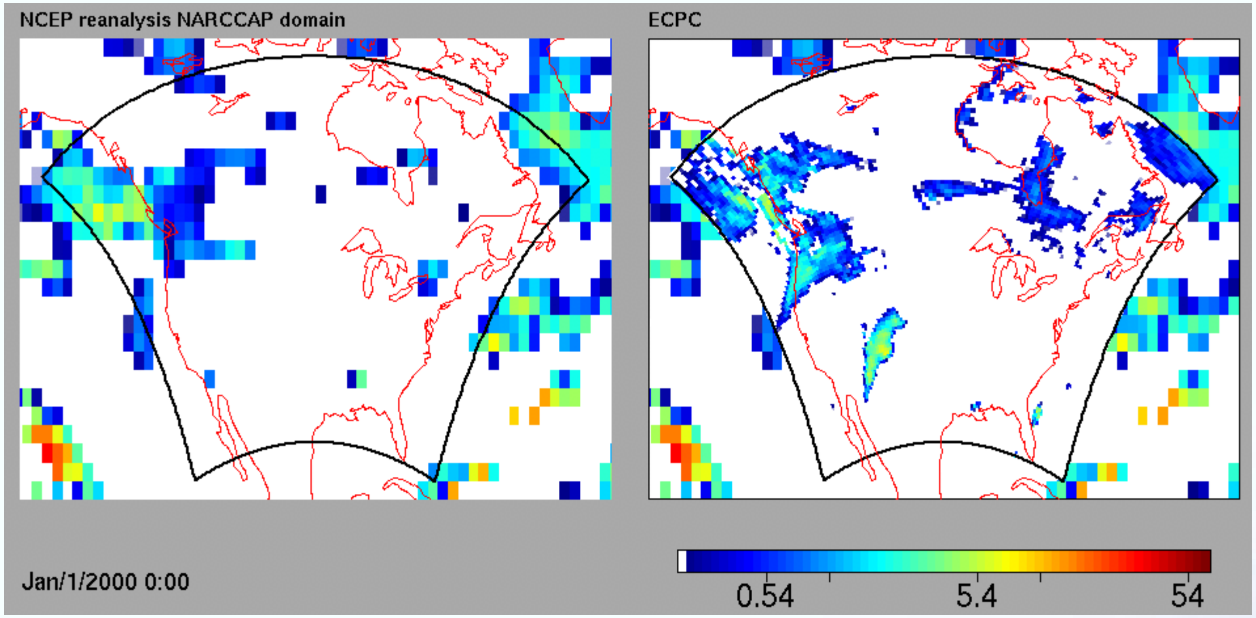
What is the spatial pattern for expected rainfall?

A climate model grid box (?)

An approach to Regional Climate

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

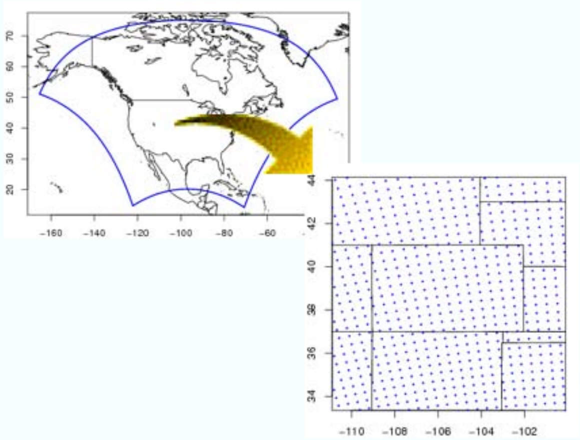
Regional model simulates higher resolution weather based on the global model for boundary values and fluxes.



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

Global observations × 6RCMs

X High resolution global atmosphere

GLOBAL FORCING	REGIONAL MODELS						
	MM5I	WRF	HADRM	REGCM	RSM	CRCM	time slice
GFDL			●	●	○		X
HADCM3	○		●		●		
CCSM	●	■				■	X
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?

The goals:

- *Estimate $g(x)$ based on the observations*
- *Quantify the uncertainty in the estimate.*
- *Handle larger spatial data sets in a interactive mode*

The goals:

- *Estimate $g(x)$ based on the observations*
- *Quantify the uncertainty in the estimate.*
- *Handle larger spatial data sets in a interactive mode*

I am not interested in spatial data

$$y_i = g(x_i) + \epsilon_i$$

Nonlinear autoregressions

Z_t a time series

$Y_i \equiv Z_t$, $\boldsymbol{x}_i \equiv Z_{t-1}, Z_{t-2}, \ldots$

Nonparametric regression

$\boldsymbol{y}_i$ a response and $\boldsymbol{x}_i$ covariates

Basic least squares setup is a first step in algorithms for nongaussian and quantile regression.

As a spline (or flexible form)

$$\min_{\boldsymbol{c}} \sum_i (\boldsymbol{y}_i - g_{\boldsymbol{c}}(\boldsymbol{x}_i))^2 + \lambda \boldsymbol{c}^T Q \boldsymbol{c}$$

How this is done ...

Michael Grab, Gravity Artist

gravityglue.com

Random Effects/Linear model for g

$\{\Phi_j\}$: m basis functions

$$g(x) = \sum_j \Phi_j(\boldsymbol{x})\boldsymbol{c}_j$$

A linear model:

$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{c} + \boldsymbol{\epsilon}$

Random effects:

$\boldsymbol{c} \sim MN(0, \rho\boldsymbol{P})$ and $\boldsymbol{\epsilon} \sim MN(0, \sigma^2\boldsymbol{I})$

Implied Covariance:

$E[g(\boldsymbol{x})g(\boldsymbol{x}')] = \sum_{j,k} \Phi_j(\boldsymbol{x})\rho\boldsymbol{P}_{j,k}\Phi_k(\boldsymbol{x}')$

Also $\boldsymbol{P} = (\boldsymbol{H}^T\boldsymbol{H})^{-1}$

$\lambda = \sigma^2/\rho$ plays an important role as a parameter.

Key ideas for large data sets

- Inverse of $\boldsymbol{P}$ chosen to be sparse.
- Basis functions have compact support.
- Still have a useful spatial model!

The estimate

Find $\boldsymbol{c}$ *by:*

Ridge regression/ conditional expectation/BLUE/ Posterior mean

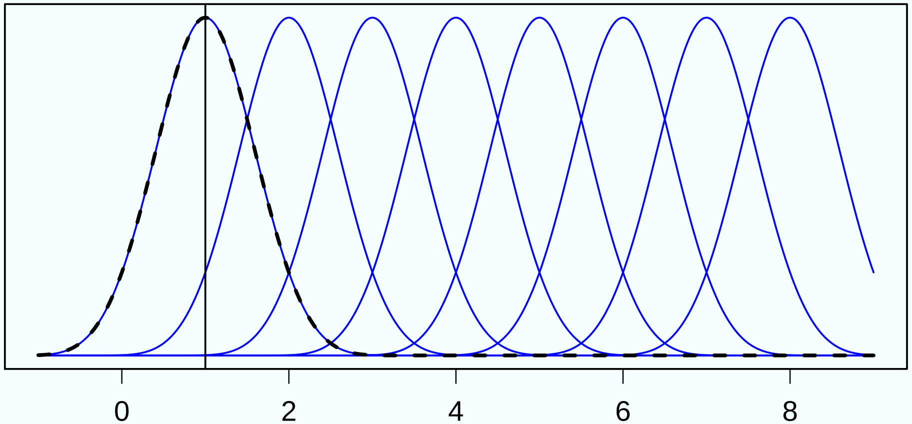
$$\hat{g}(x) = E[g(x)|\boldsymbol{y}, \boldsymbol{P}] = \sum_{k=1}^{n} \hat{c}_k \Phi_k(\boldsymbol{x})$$

$$\hat{\boldsymbol{c}} = \left(\Phi^T \Phi + \lambda \boldsymbol{P}^{-1}\right)^{-1} \Phi^T \boldsymbol{y}, \quad \lambda = \sigma^2/\rho$$

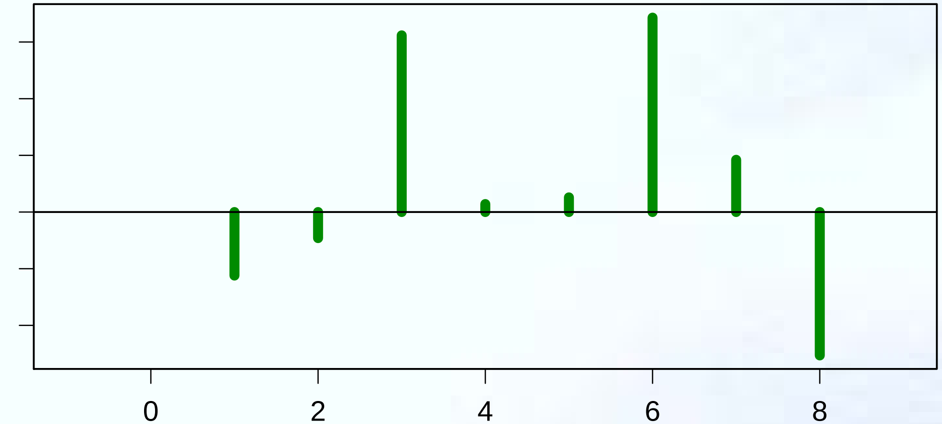
Φ^T, $\Phi^T\Phi$, $\boldsymbol{P}^{-1}$ are sparse.

A 1-d cartoon ...

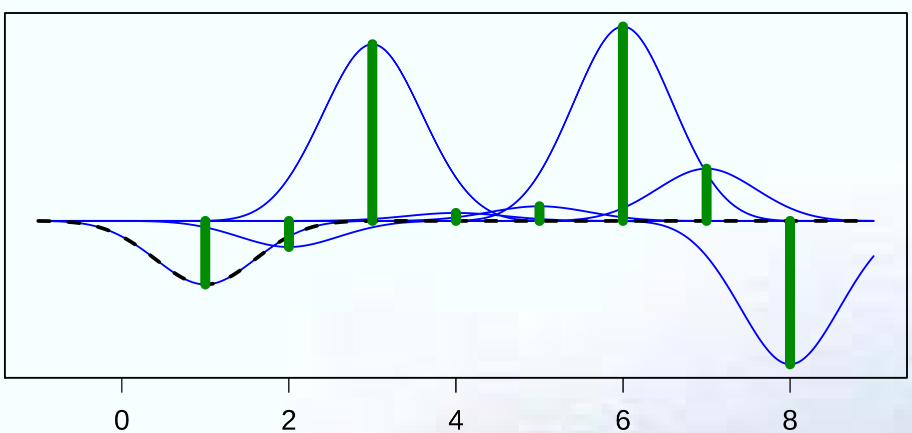
8 basis functions



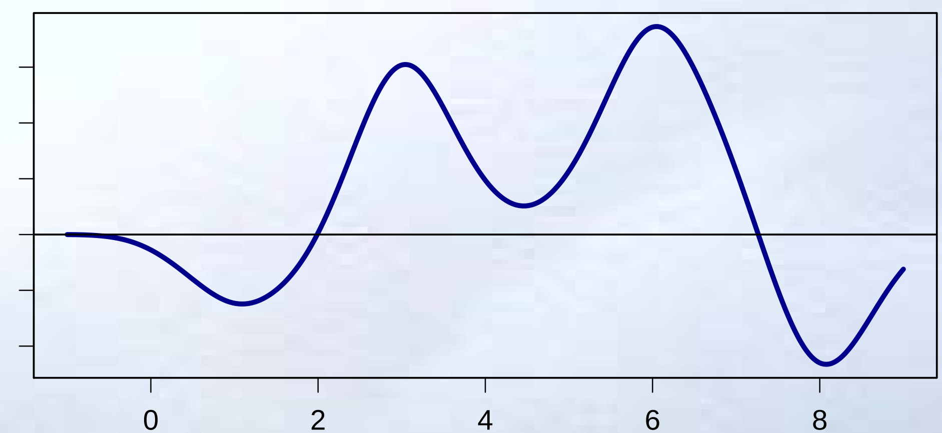
8 (random) weights



weighted basis

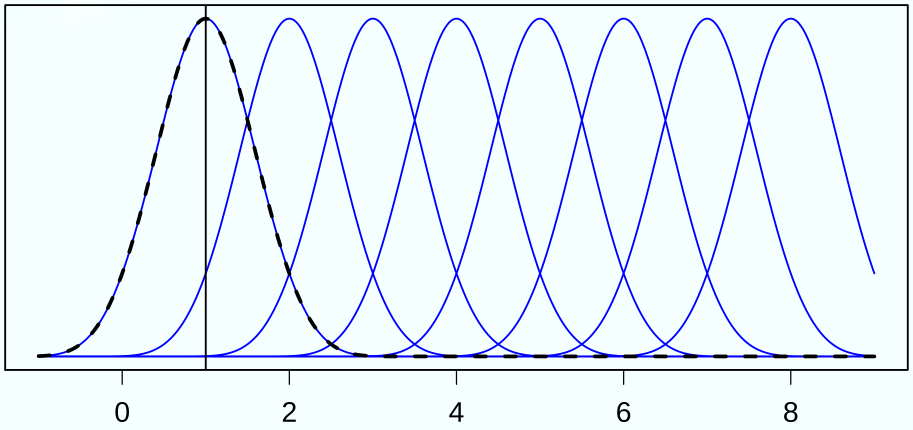


Random curve

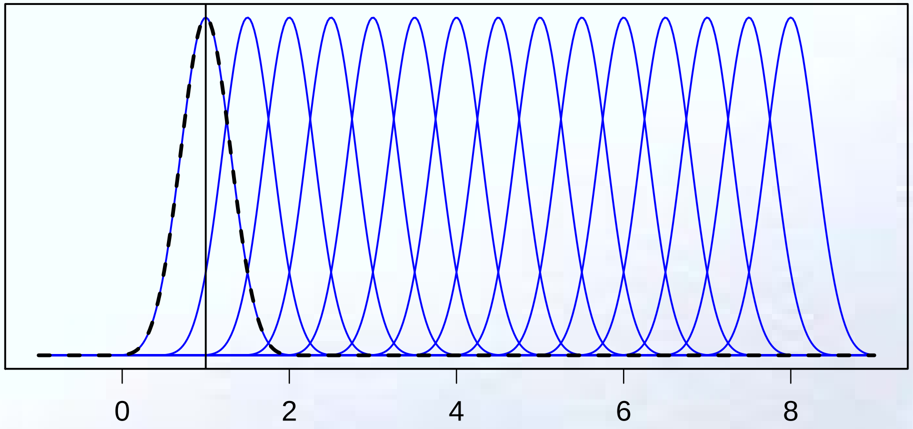


A Multiresolution

8 basis functions

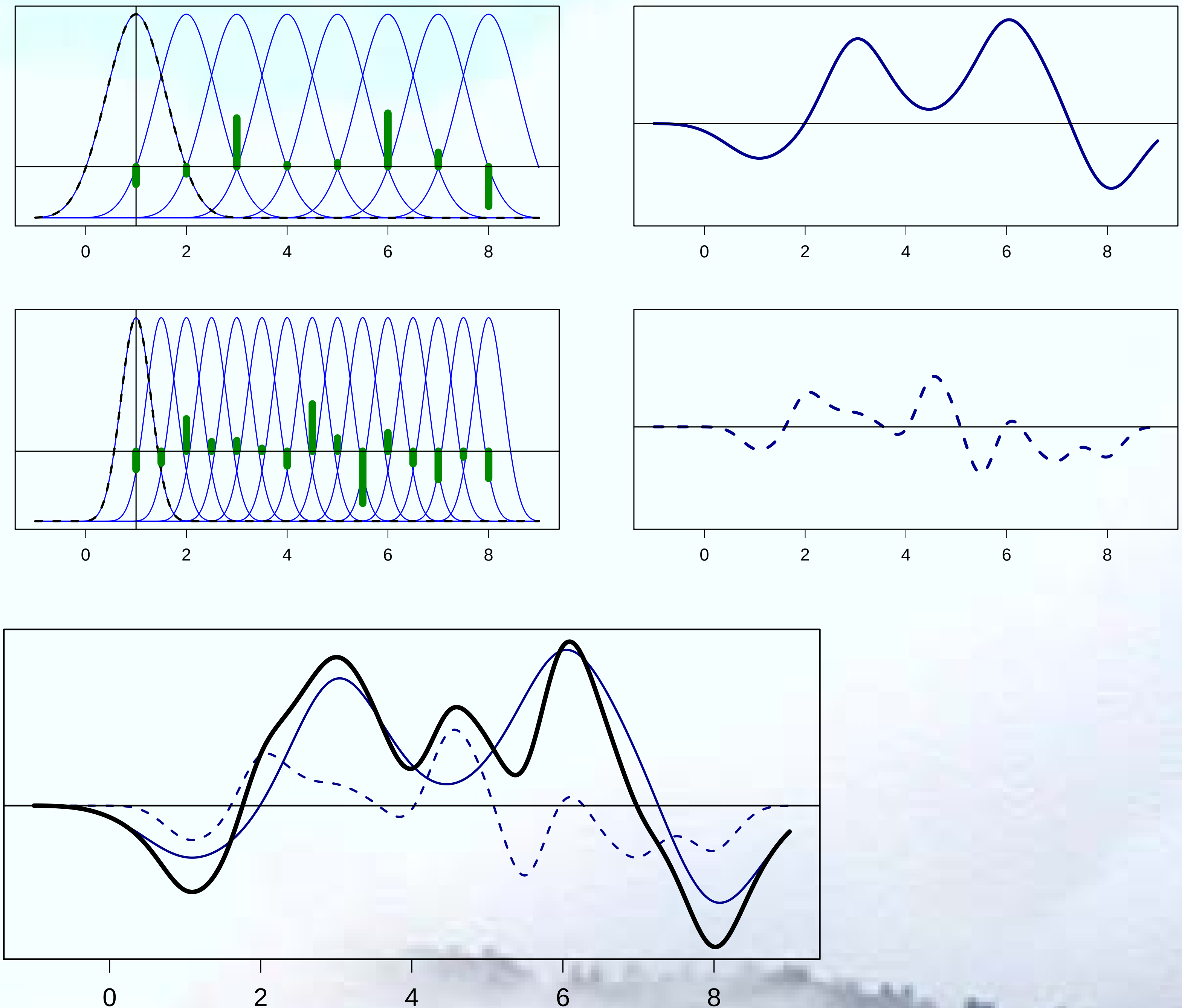


16 basis functions



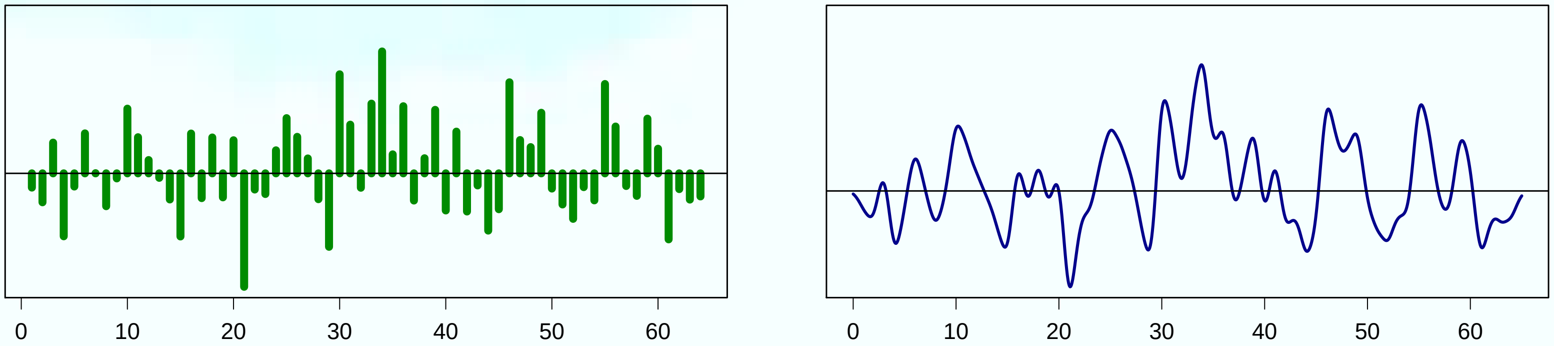
⋮

Adding them up

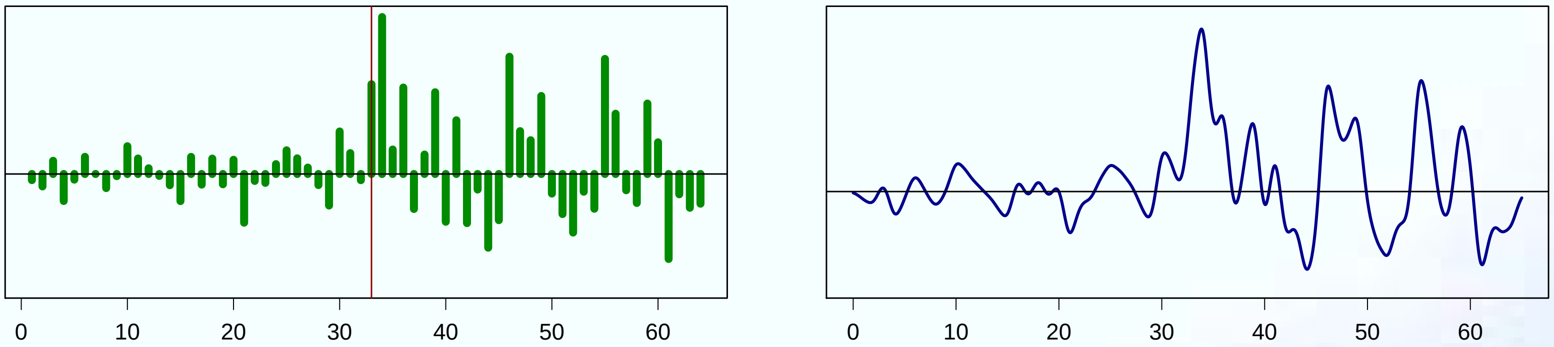


Distributions of coefficients

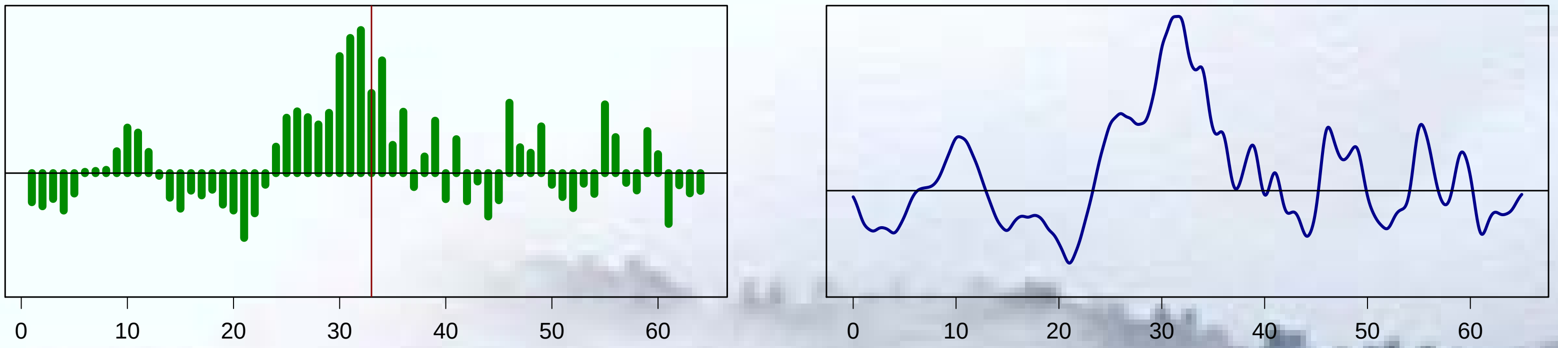
Uncorrelated (stationary)



Different variability

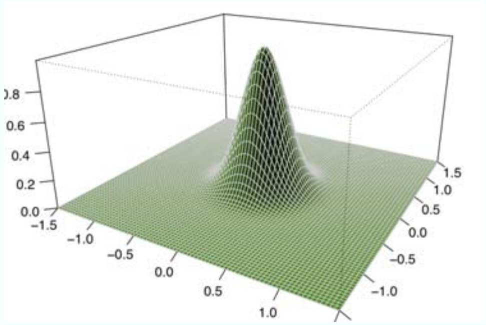


Different Correlation



A recipe for 2-d RBFs

Basis function$_j(\boldsymbol{x}) = \varphi(||\boldsymbol{x} - \boldsymbol{u}_j||/\theta)$



2-d Wendland

- φ is a positive definite, compactly supported function – a nice bump.
- $\{\boldsymbol{u}_j\}$ basis centers on a regular grid
- θ scale set to provide some overlap

Four level multi-resolution starting with 11×11 *grid has 8804 basis functions.*

A recipe for P^{-1}

Recall: $g(x) = \sum_j \Phi_j(\boldsymbol{x})\boldsymbol{c}_j$

$\boldsymbol{c}$ at each resolution level is a Markov random field:

$$(4+\kappa^2)\boldsymbol{c}_j - \sum_{l\in\mathcal{N}} \boldsymbol{c}_l = e_j, \qquad H\boldsymbol{c} = \boldsymbol{e}$$

$\{e_j\}$ are uncorrelated N(0,1) and $\mathcal{N}$ is 4 nearest neighbors.

Weights in lattice format:

.	.	.	.	.
.	.	-1	.	.
.	-1	$(4+\kappa^2)$	-1	.
.	.	-1	.	.
.	.	.	.	.

Precision matrix for c is sparse: $P = (H^T H)^{-1}$

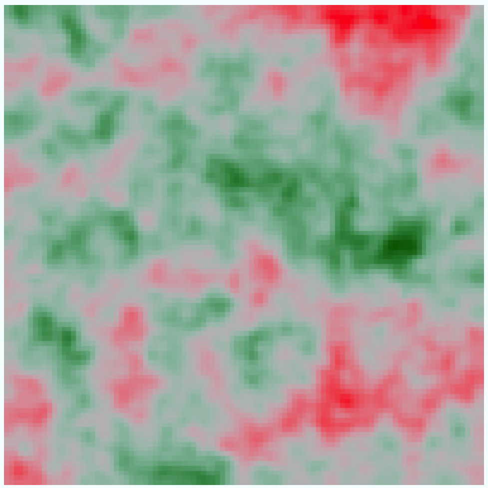
Two dimensions

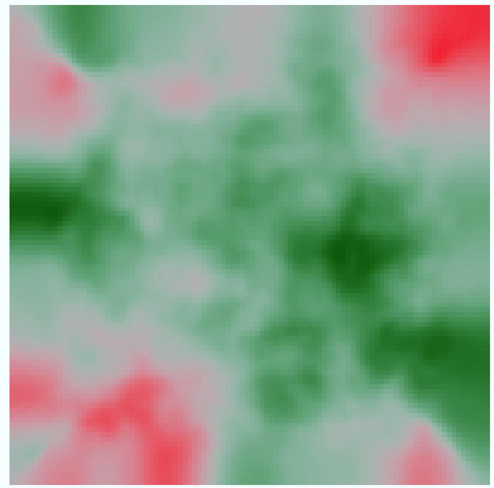
Combination of 4 levels starting with an 8×8 grid

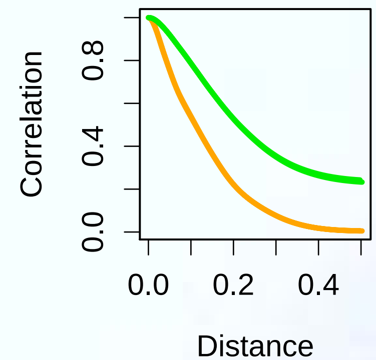
Uncorrelated weights | Correlated weights

Rougher fields

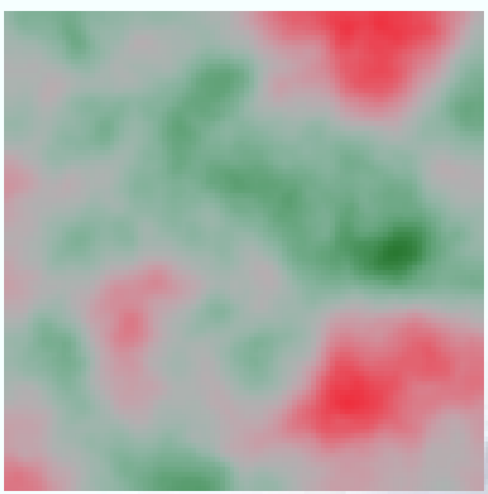
Correlation function2

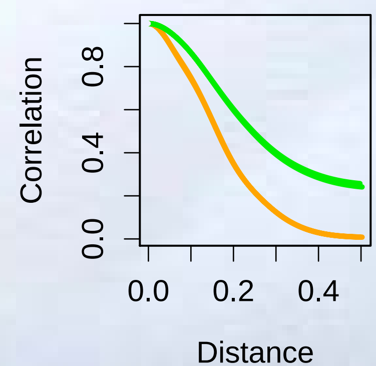






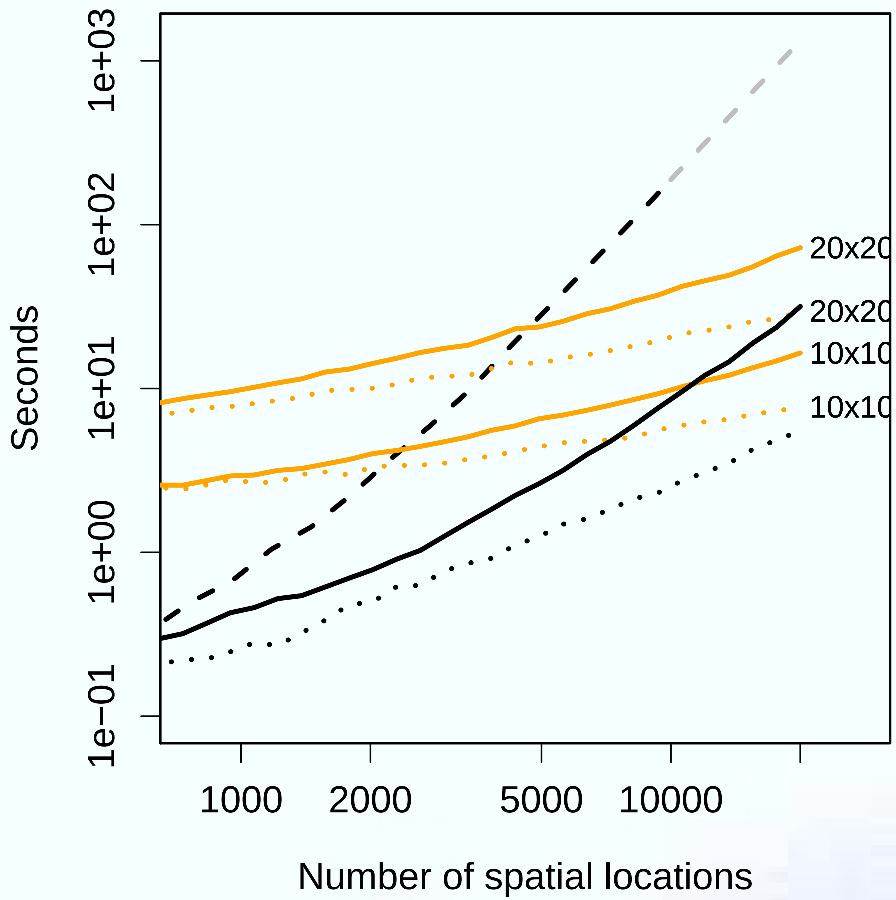
Smoother fields





Timing

An evaluation of the likelihood using the standard dense matrix Kriging and LatticeKrig



Standard Model:

dashed – exponential covariance

Lattice Krig model:

Solid - with normalization,

short dashed - without

grid = number of locations

four levels 10×10 **M** $\approx$ **8000**

four levels 20×20 **M** $\approx$ **30000**

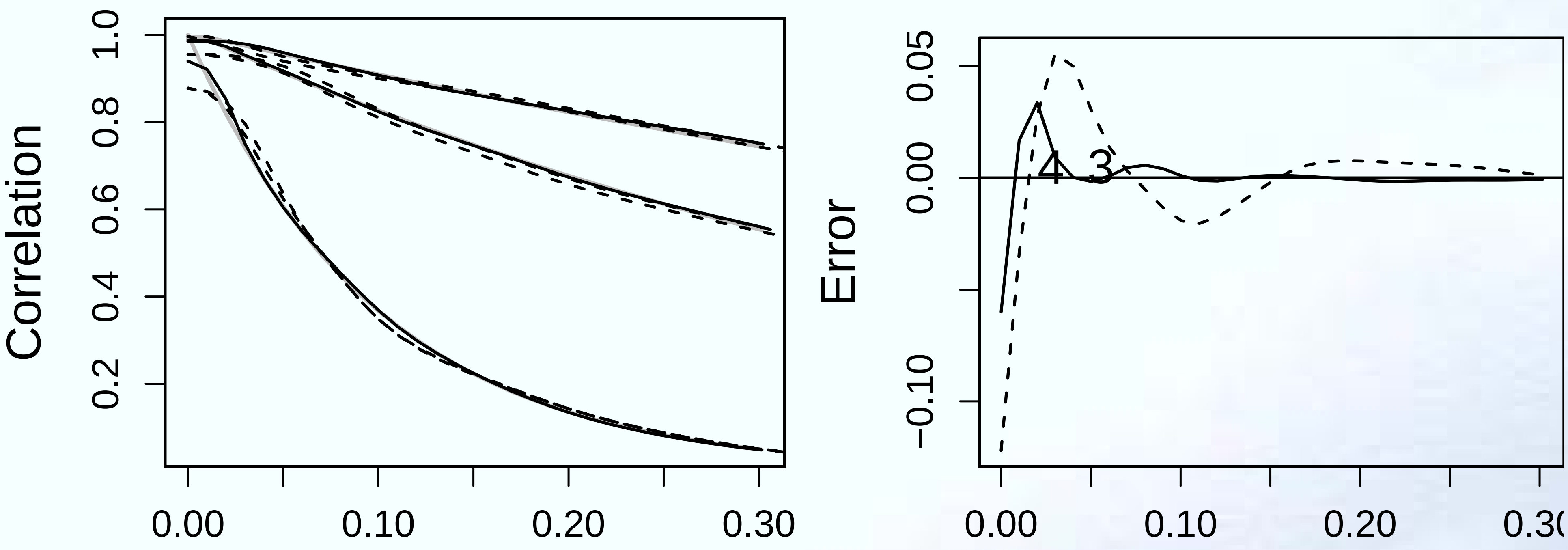
At 20,000 observations:

standard Kriging about 21 minutes , LatticeKrig is 5-10 seconds.

Flexibility of LatticeKrig model

Fitting an exponential (minimizing mean squared error)

- First level resolution of 10×10
- 3 levels, 4 levels, target exponential

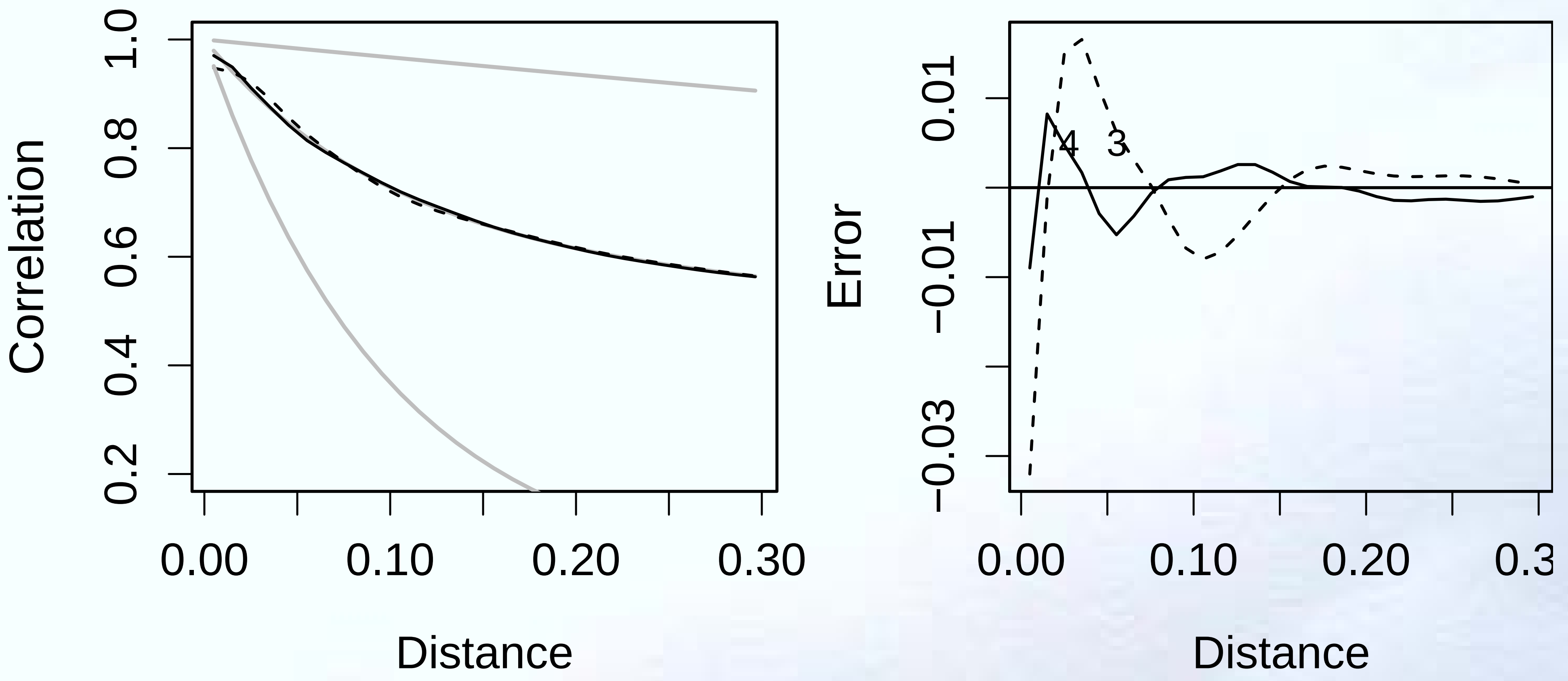


Also works well for approximating smoother covariances.

More Flexibility of LatticeKrig model

Fitting a mixture of exponentials

- First level resolution of 10×10
- 3 levels, 4 levels, target: .4Exp(.1) + .6Exp(3)



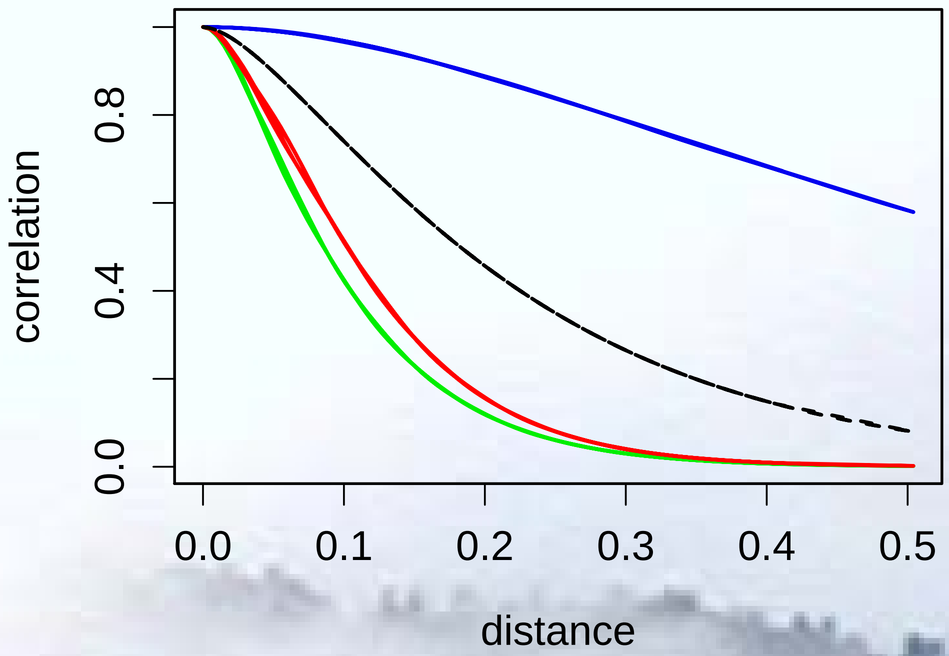
Back to climate data

Some details for observed data:

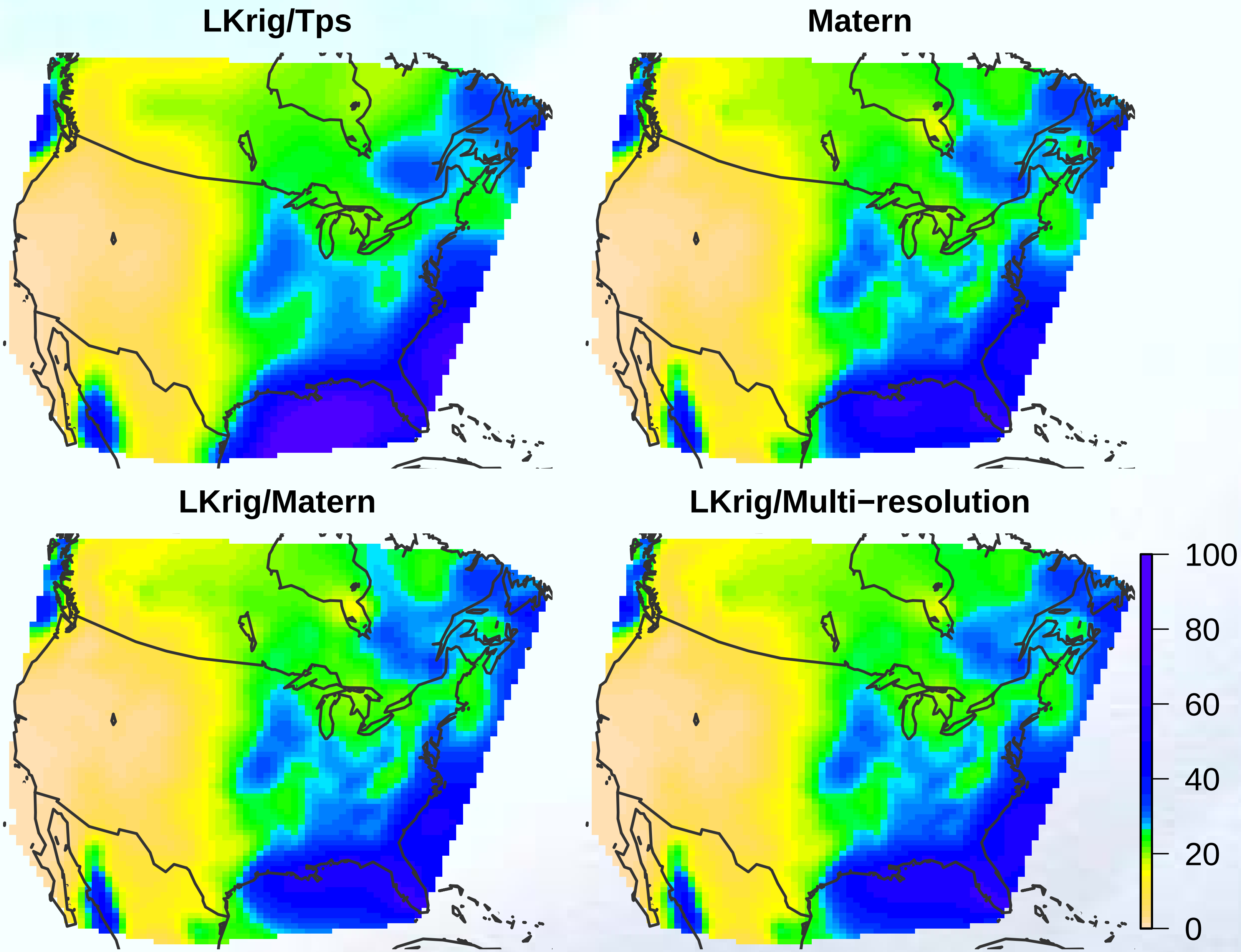
- Used log transformation and stereographic projection for locations
- Elevation included as linear fixed effect.
- Covariance parameters found by maximum likelihood

Estimated covariance functions

Matern, thin plate like , Matern-like, Multiresolution (3 levels)



Predicted surface



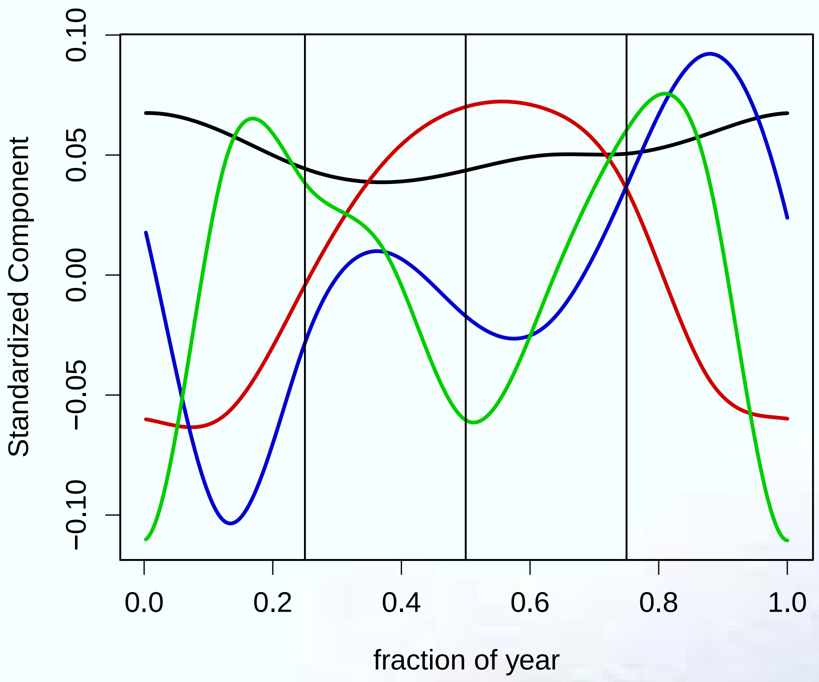
Climate change

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 coefficient spatial fields for each model.
- Approximately 8000 spatial locations

Seasonal PCs
(future - present)

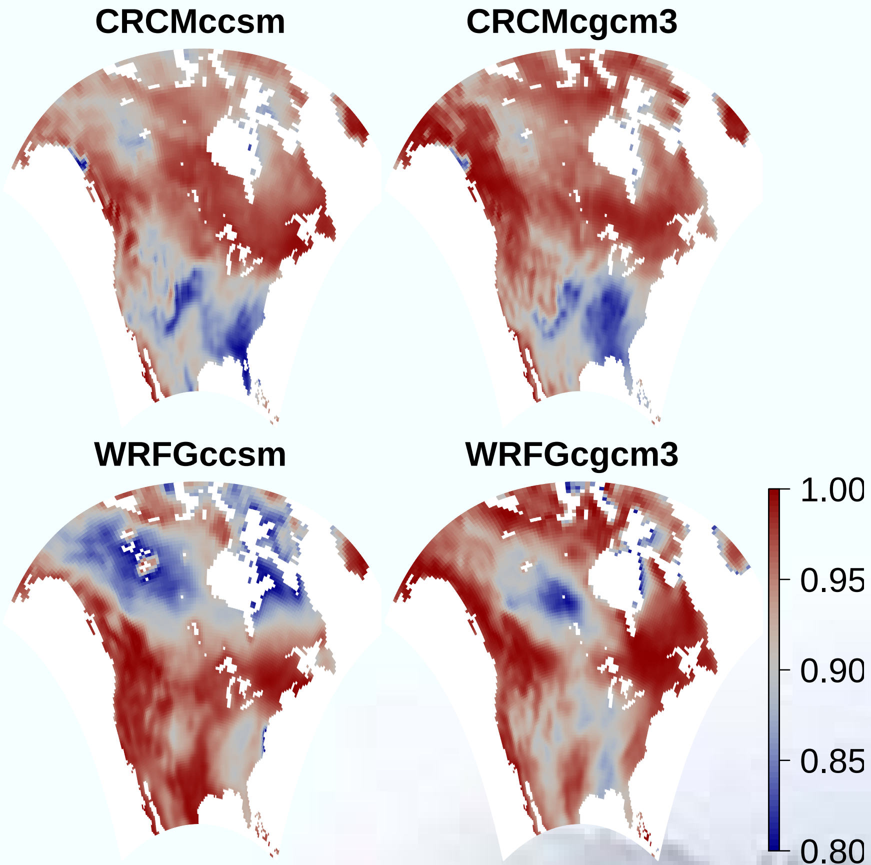


NARCCAP domain

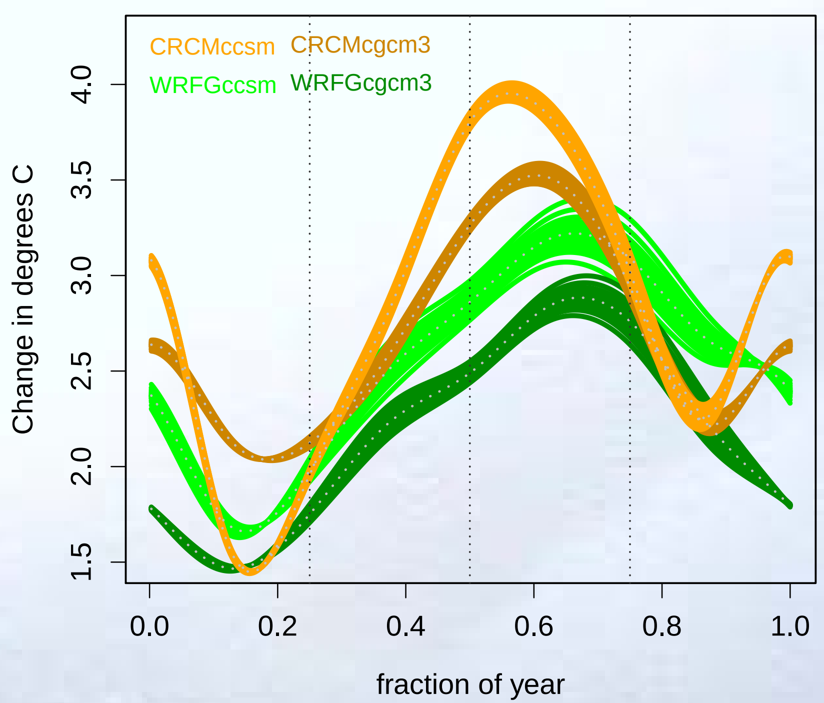
Results

- Thin plate spline 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)
- Works in **United Econoplus** !

R^2 for first PC



Inference for Boulder grid box



Summary

- Computational efficiency gained by compact basis functions and sparse precision matrix.

- Flexibility in model to account for nonstationary spatial dependence.

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

See `LatticeKrig` *package in R*

Thank you!