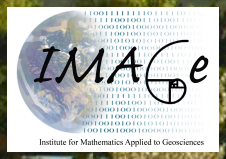
Spatial data analysis: a brief introduction

Douglas Nychka,
National Center for Atmospheric Research



Supported by the National Science Foundation

Outline

- Colorado springtime temperatures and a spatial analysis
- Easy to use functions from the `fields` R package

Analysis Work Flow

- EDA and plots — Yogi Bera

- MODEL STEP for spatial data
  consider climate covariates and the covariance (correlation)
- COMPUTE STEP estimate parameters
  and find conditional distribution (ensembles)

Yogi Bera

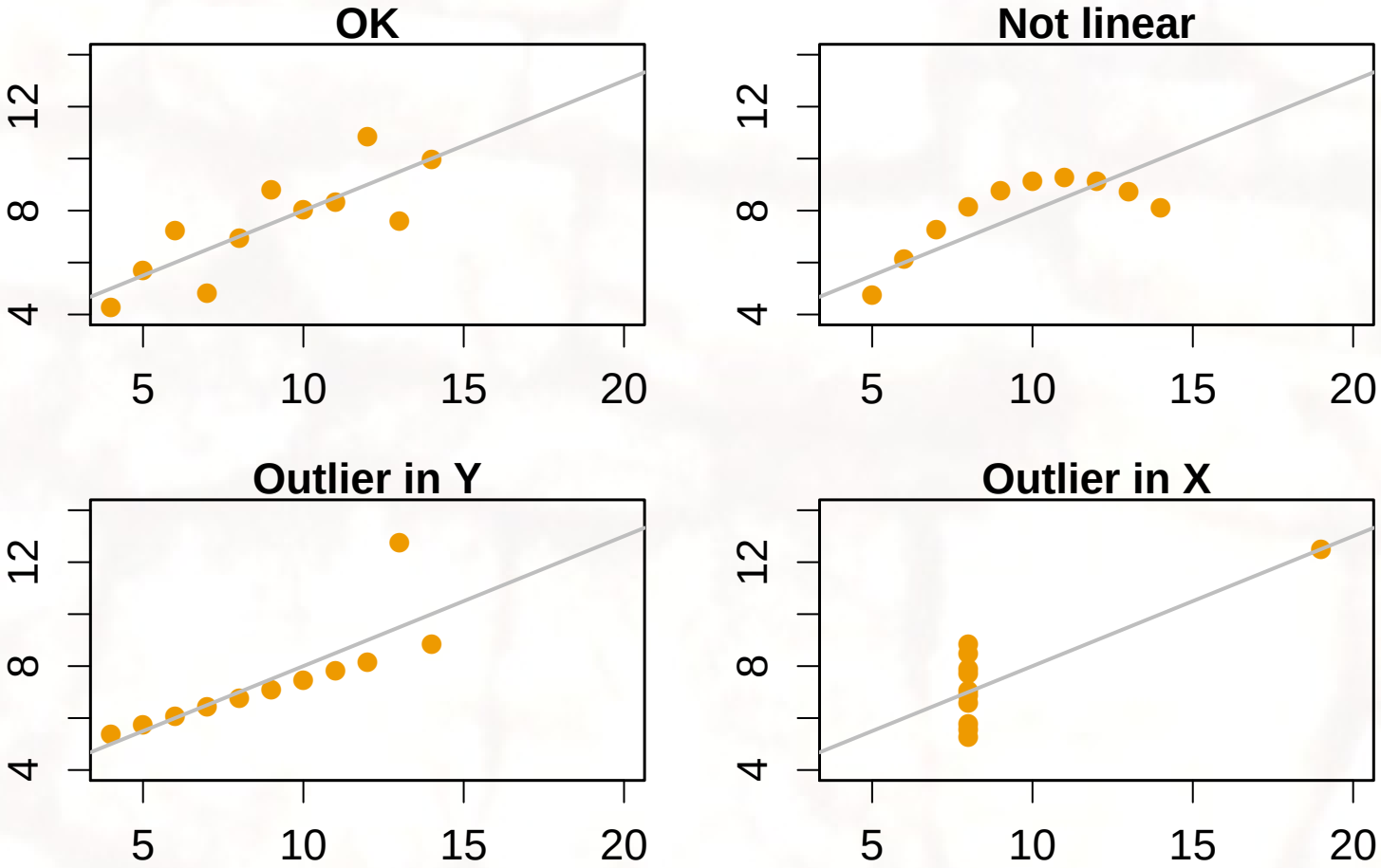
"You can observe a lot just by looking."

This is why statisticians like high level languages like R.

e.g Fitting a line to data by least squares.

Anscombe's Quartet

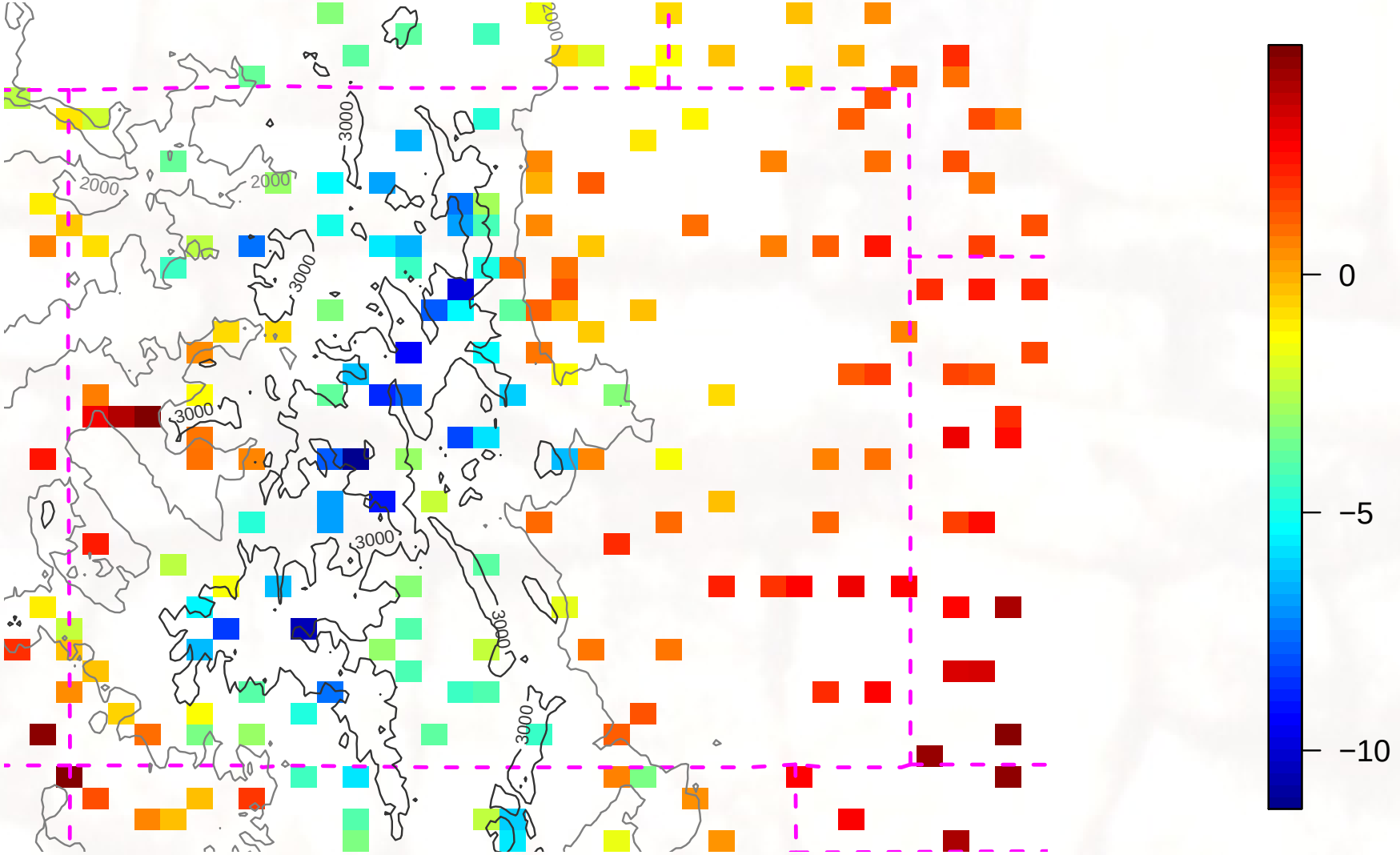
Things that can go wrong.



All these data sets have the same least squares line and the same correlation ($\sim$ .82).

Estimating a curve or surface.

Colorado MAM average tmin



Goals:

1. *Prediction:* Determine the climate at locations where there are not stations.

2. *Uncertainty:* Quantify the error in the predictions.

3. *Summarize the spatial/temporal structure:* One tool for comparing observations to models and models to models for physical insight

The additive statistical model:

Observation = climate covariates

+ Smooth function (location) + error

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

$$y_i = z_i\boldsymbol{\beta} + g(x_i) + \epsilon_i$$

ϵ_i's are random errors
z_i climate covariates (e.g. elevation) , parameters ($\boldsymbol{\beta}$)
and g is an unknown smooth function.

Main Ideas

MODEL STEP

Use observed data to tease out a statistical model for g

COMPUTING STEP

Find the distribution of g and β *given* the observations

e.g. This is Kriging, or Bayesian statistics or splines

Start with g being a Gaussian process described by a mean function and a covariance function.

Statistical models for a surface

Covariance functions

$g(\boldsymbol{x})$ is a random surface with $E[g(\boldsymbol{x})] = \mu(\boldsymbol{x})$

Covariance function k has two spatial arguments

$$k(\boldsymbol{x}_1, \boldsymbol{x}_2) = COV(g(\boldsymbol{x}_1), g(\boldsymbol{x}_2))$$

- Usually $\mu(\boldsymbol{x})$ is constant or even zero.
- Often k only depends on distance of separation between locations.

Exponential

$$k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \rho e^{-\text{ distance}(\boldsymbol{x}_1, \boldsymbol{x}_2)/\theta}$$

Isotropic

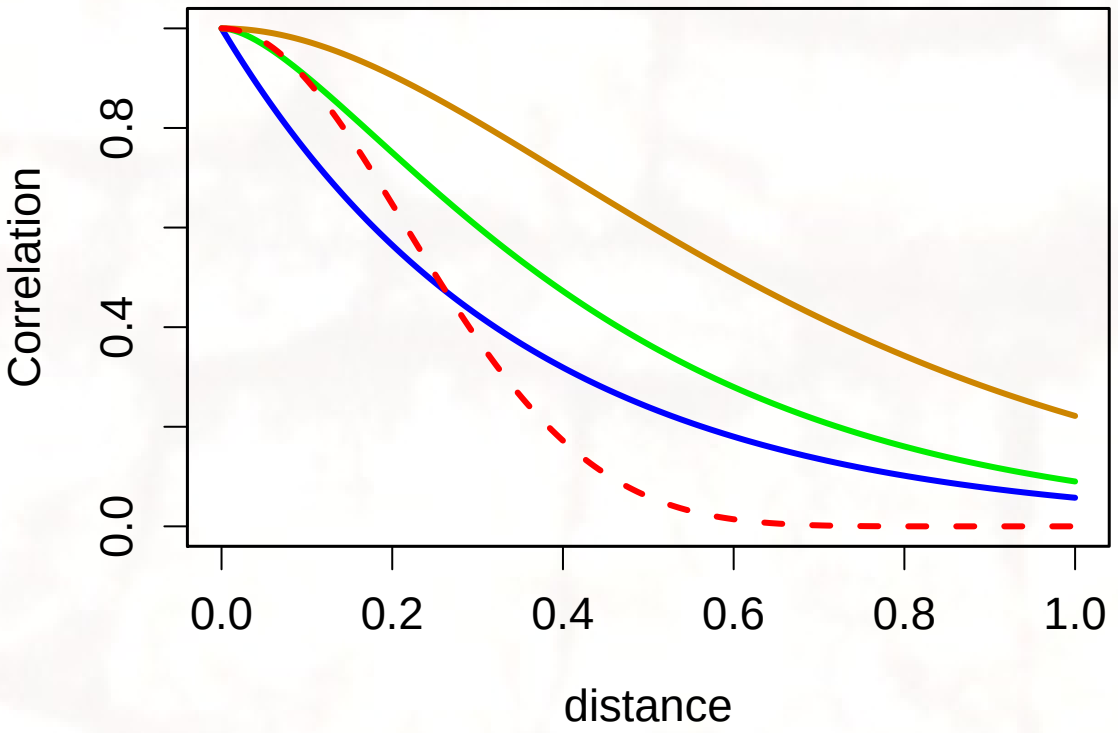
$$k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \rho\Phi(\text{distance}(\boldsymbol{x}_1, \boldsymbol{x}_2)/\theta)$$

Families of correlation functions

Matern:

$\phi(d) = \rho\psi_\nu(d/\theta))$ with ψ_ν a Bessel function.

$\nu =$ .5, 1.0 , 2.0



- θ a range parameter
- ν smoothness at 0.
- ψ_ν is an exponential for $\nu = 1/2$ as $\nu \to \infty$ Gaussian.
- As ν increases the process is smoother.

Wendland:

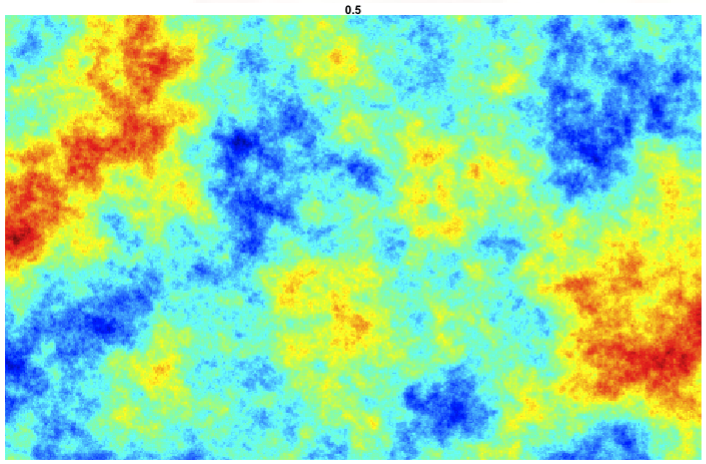
Polynomial that is exactly zero outside given range.

Compactly supported Wendland covariance (d=2, k=3)

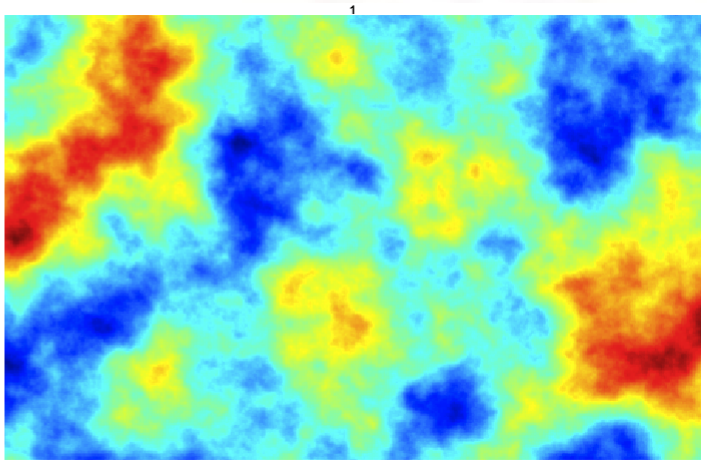
What do these processes look like?

Varying the smoothness:

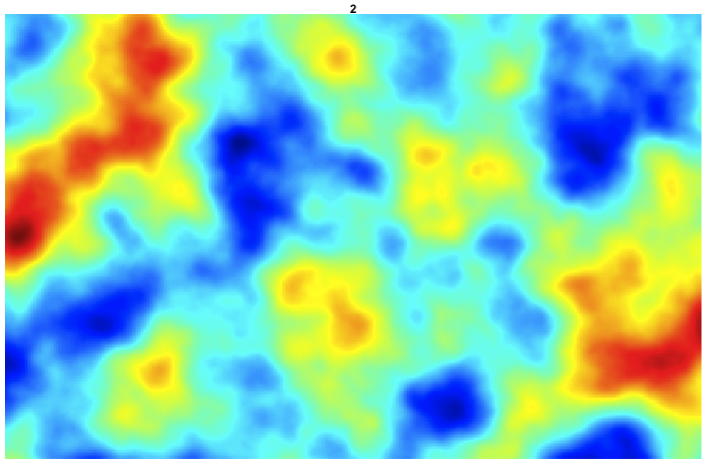
Matern (.5)



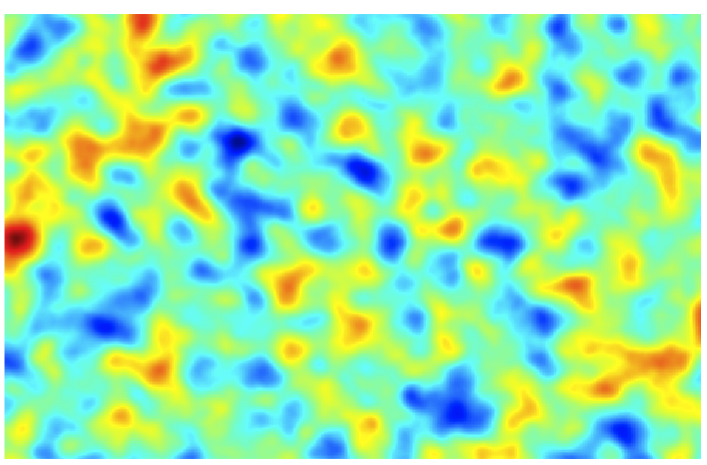
Matern(1.0)



Matern (2.0)



Wendland (2.0)



Variogram as an EDA tool

How to estimate the spatial from a single field?

- Plot for a spatial data set or spatial field plot $\frac{(\boldsymbol{y}_i - \boldsymbol{y}_j)^2}{2}$ against the distance of separation. "On the average" this should be follow the theoretical curve that is the variance of the data *minus* the covariance function .

covariance $\rightarrow$ variogram or variogram $\rightarrow$ covariance

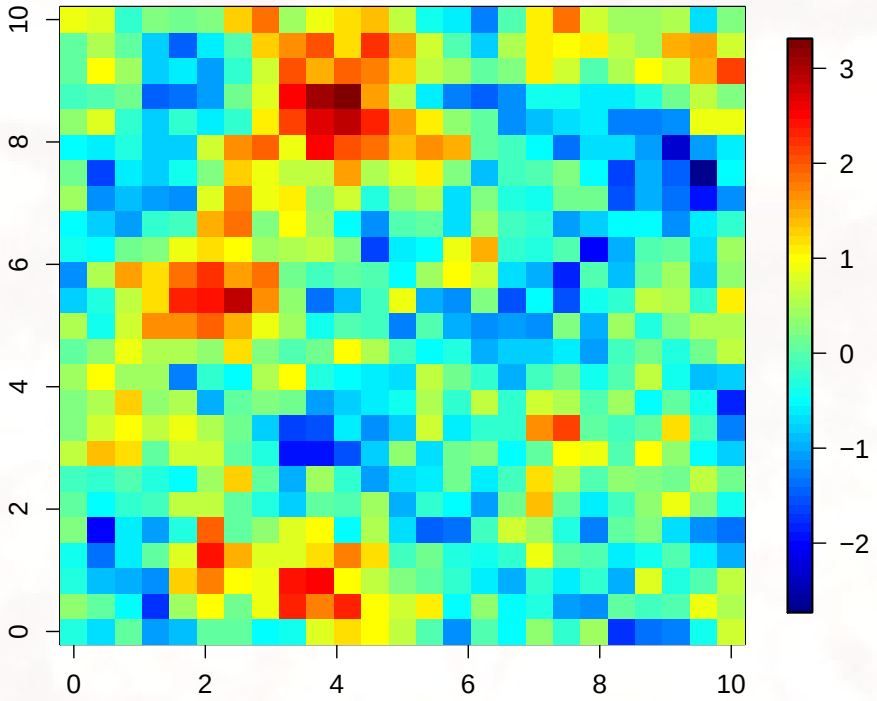
- If one can find find the variogram then one can transform back to the covariance function.

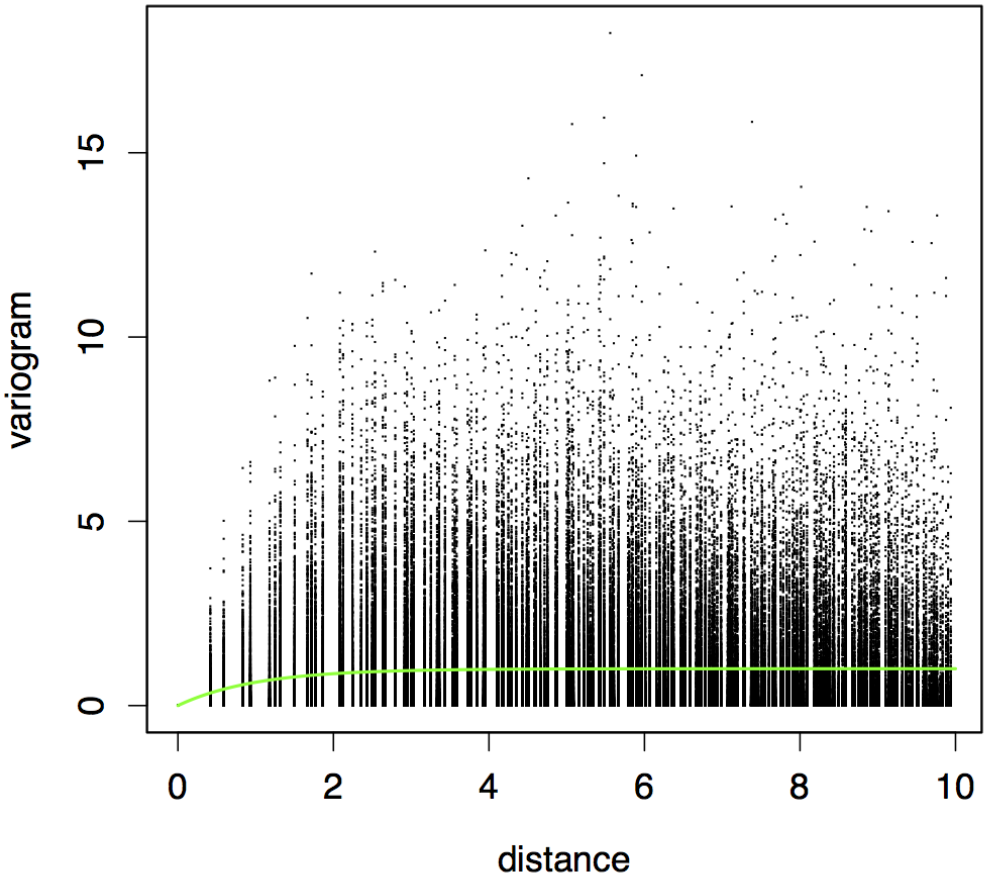
- Need to be careful about how the variogram behaves close to zero distance. The variogram estimates $\rho + \sigma^2$ right at zero – not just ρ

- Great EDA tool – terrible for actually estimating parameters!

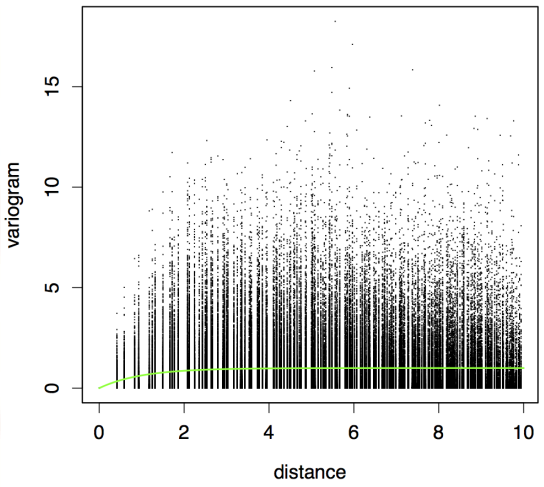
A variogram example

Sample field with true exponential variogram on a 25×25 grid.





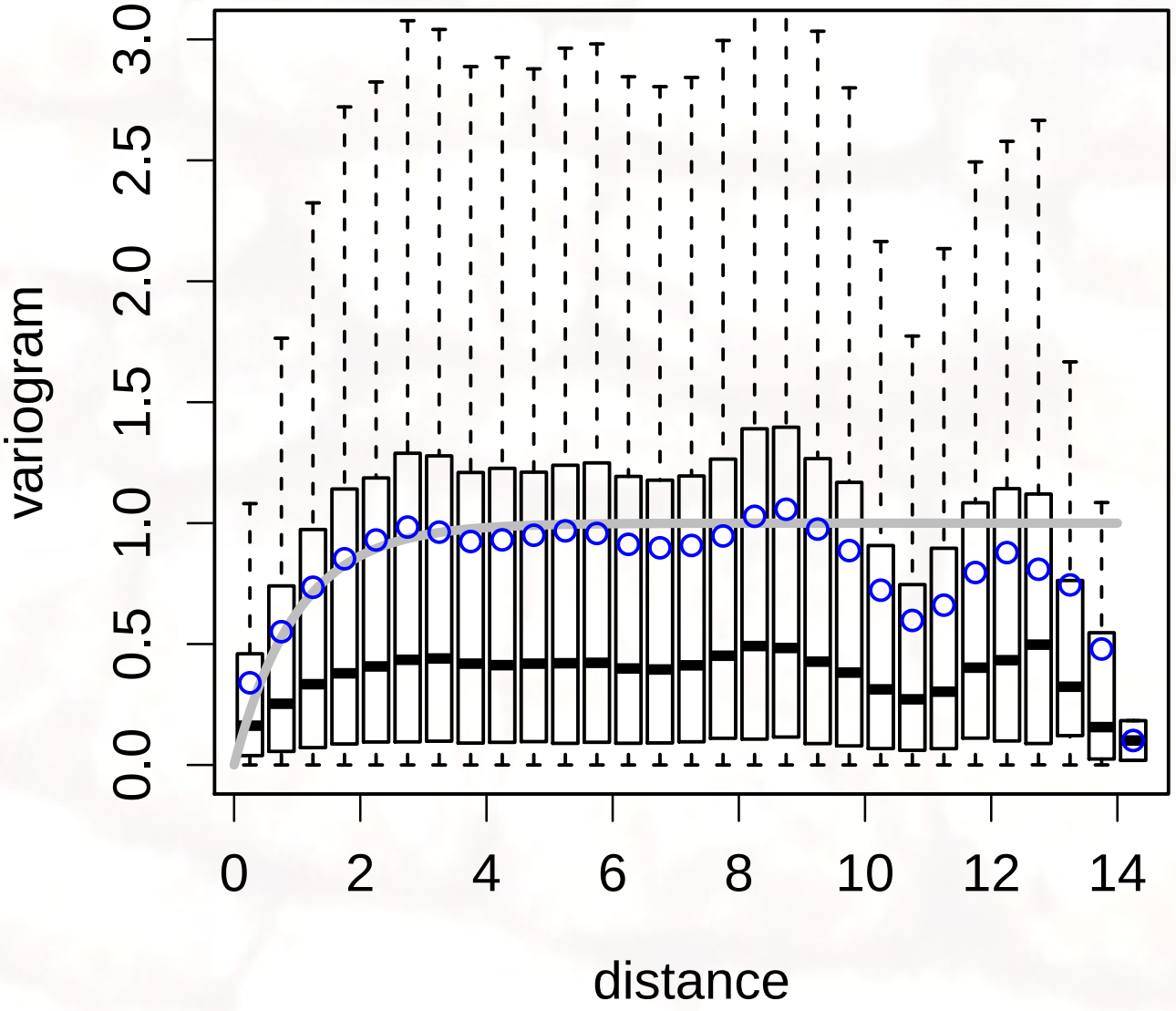
Based on (625 *choose* 2) total pairs
$\approx$ 200K differences: $(y_i - y_j)^2/2$



Houston, we have a problem.

Improving the variogram

Binning observations by distance ranges, finding box plots and adding the mean for each bin.

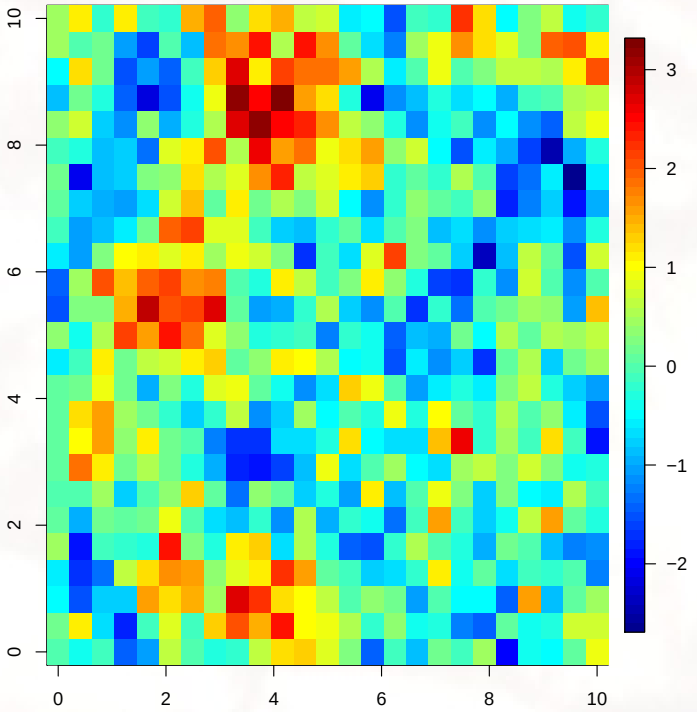


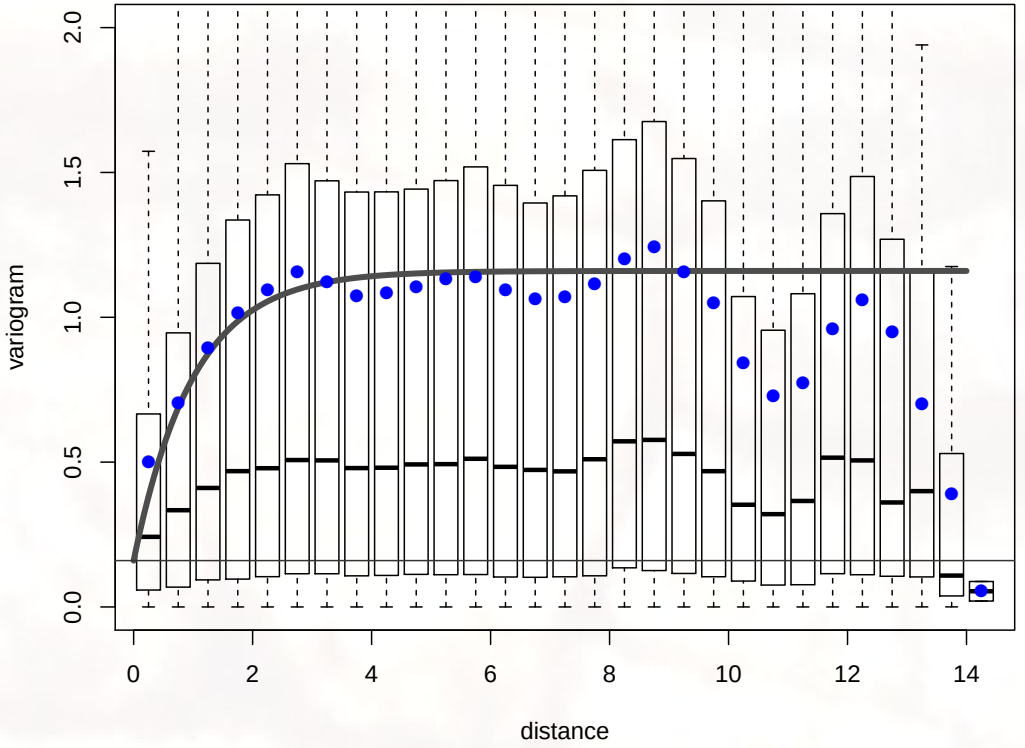
Identifying the nugget σ^2

Recall the additive model:

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

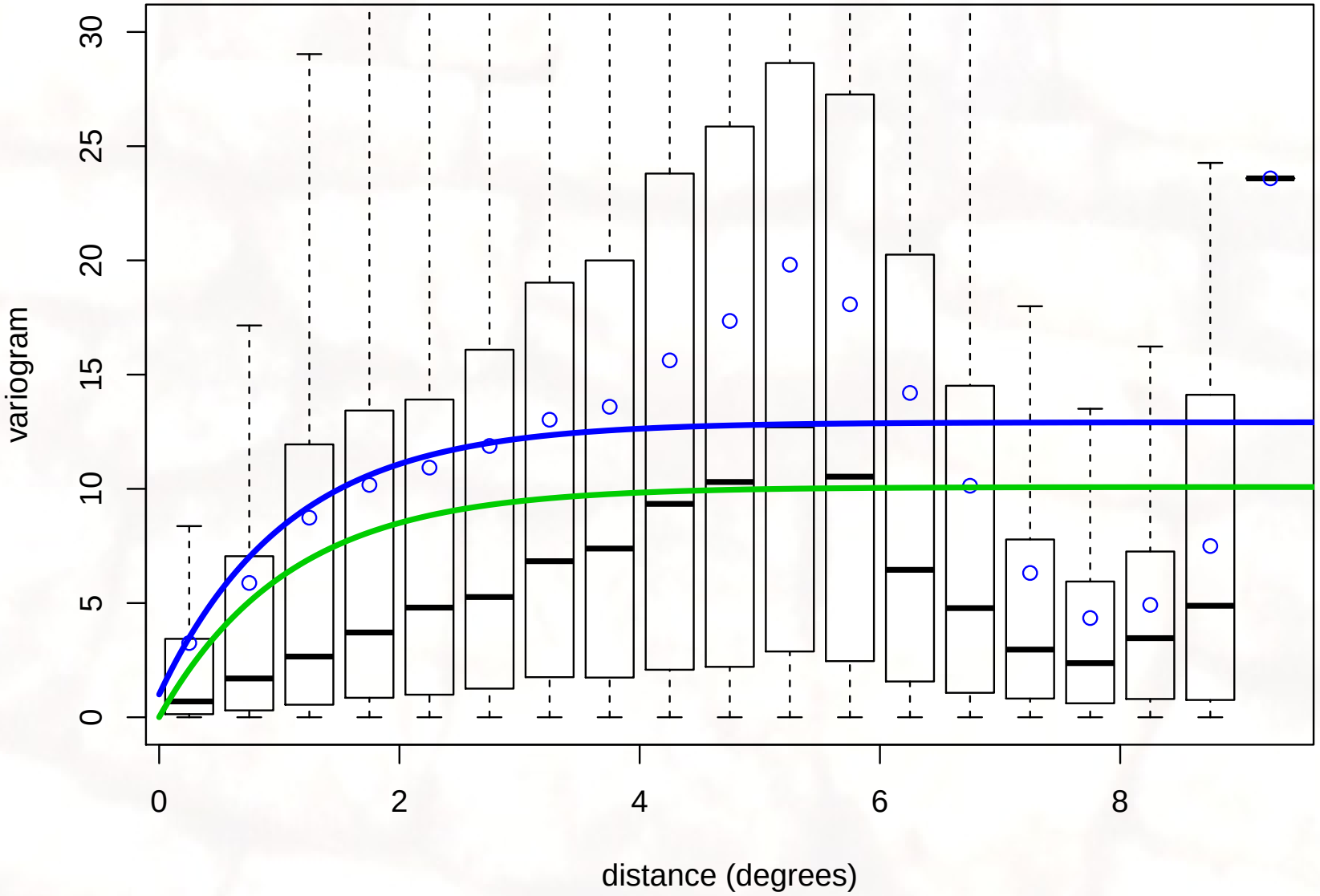
Correlations among the observations due to the smooth field but the measurement error is uncorrelated. Adding measurement error to the example ($\sigma = .4, \sigma^2 = .16$)



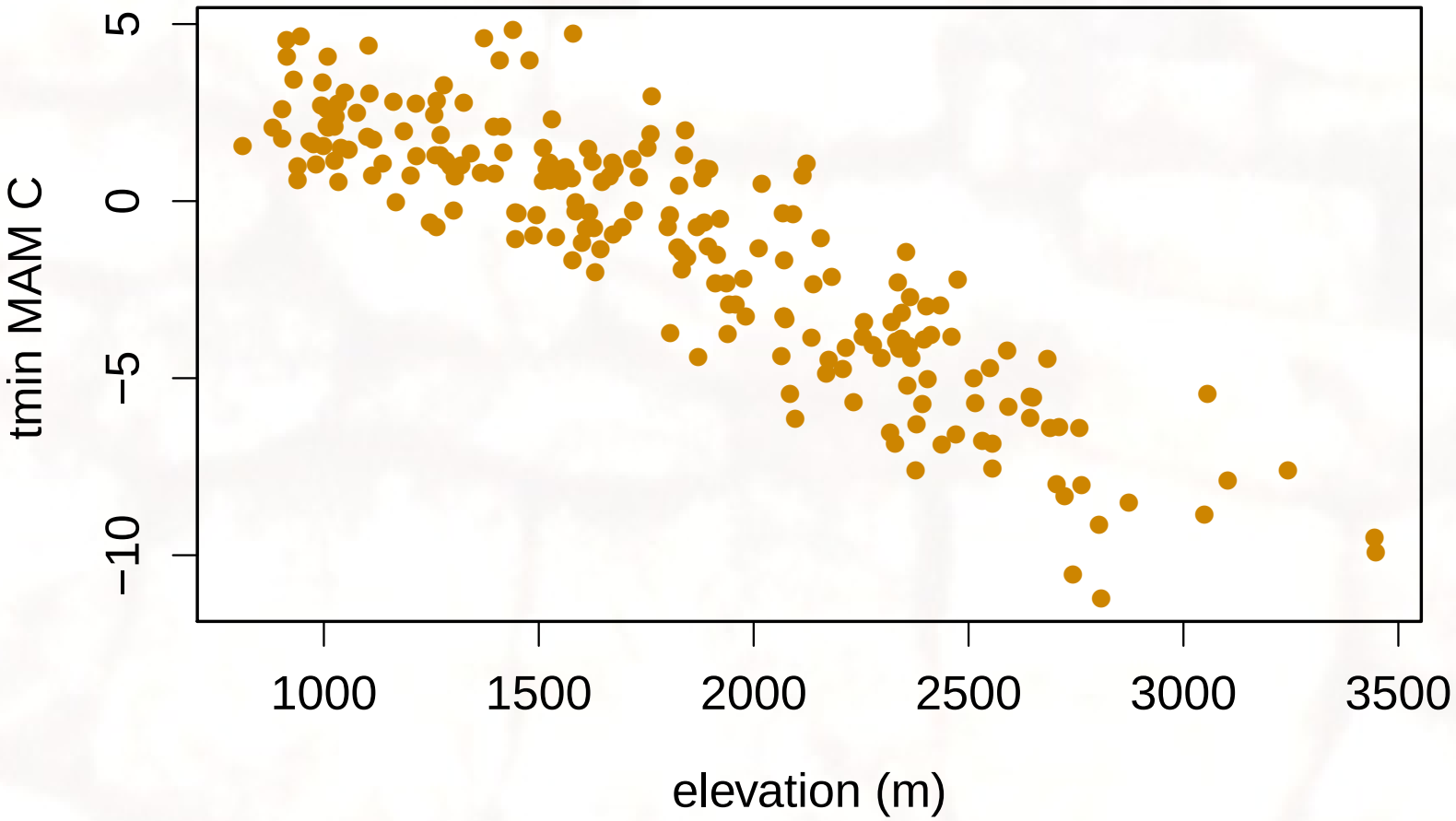


Back to the CO climate data

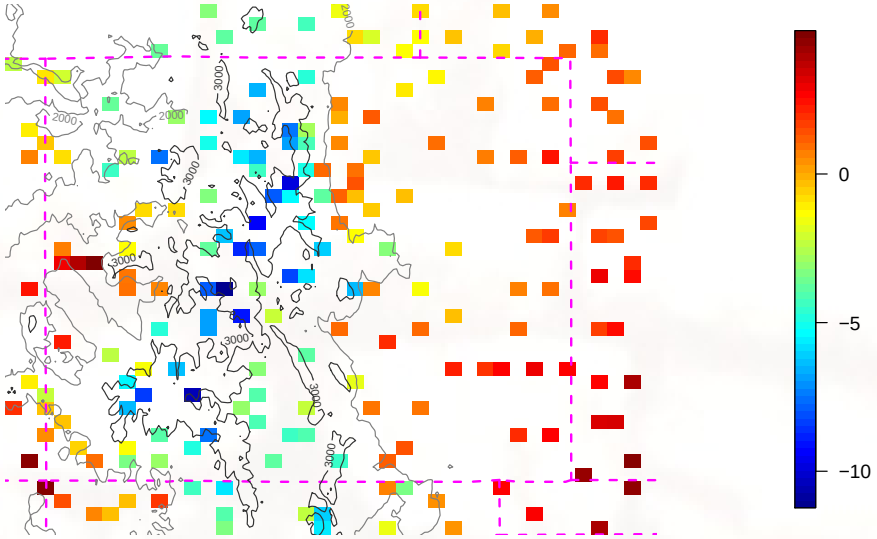
Variogram for the MAM station averages.



What about elevation?



Fitting the MAM Colorado temps



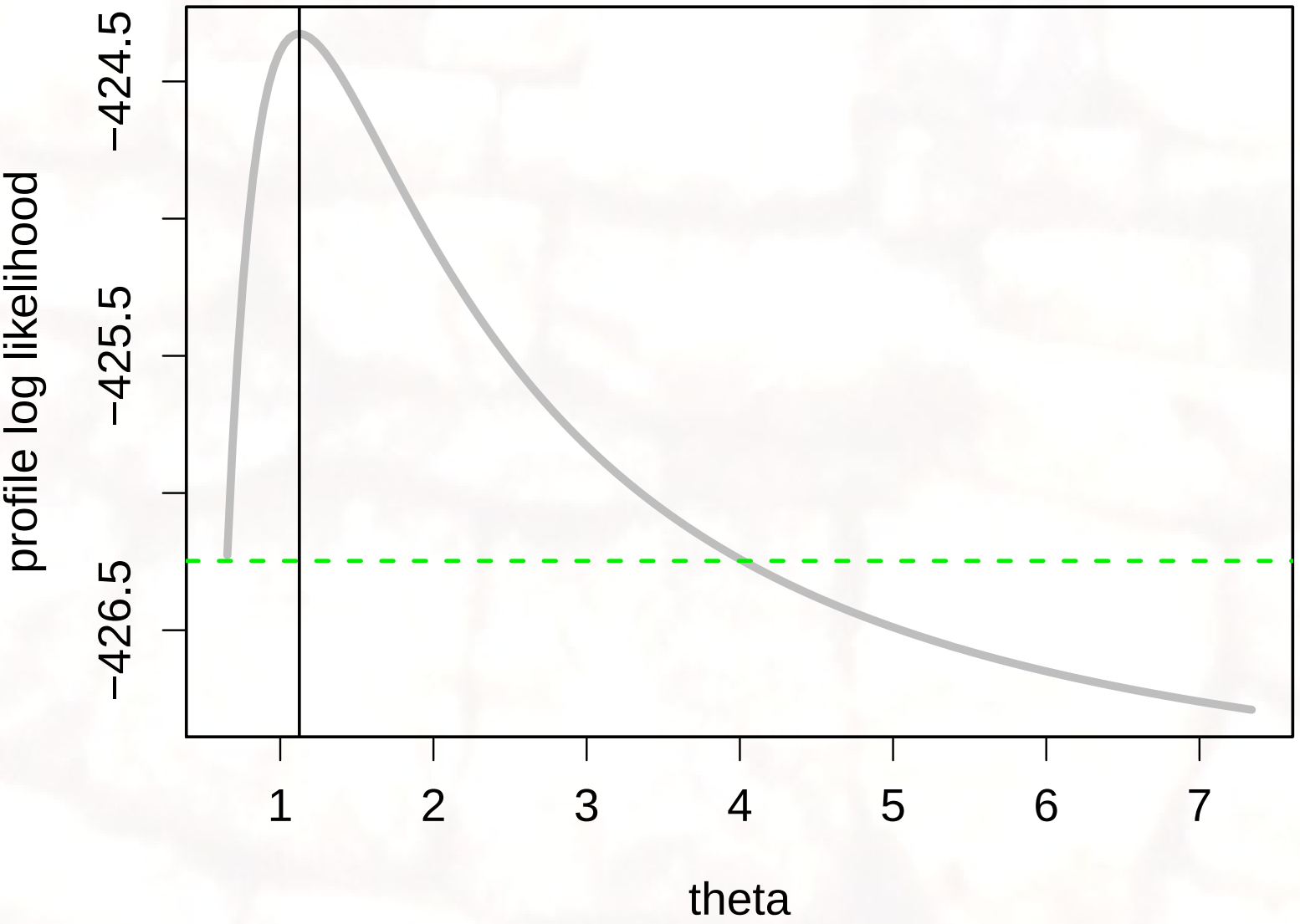
$$\text{Observed}_i = \beta_1 + \beta_2\ \text{lon} + \beta_3\ \text{lat} + \beta_4\text{elevation} + \text{g}(\boldsymbol{x}_i) + \text{error}$$

Use maximum likelihood to find:

- $\boldsymbol{\beta}$,
- VAR(g), VAR(error)
- range and smoothness of Matern.

Profile likelihood for θ

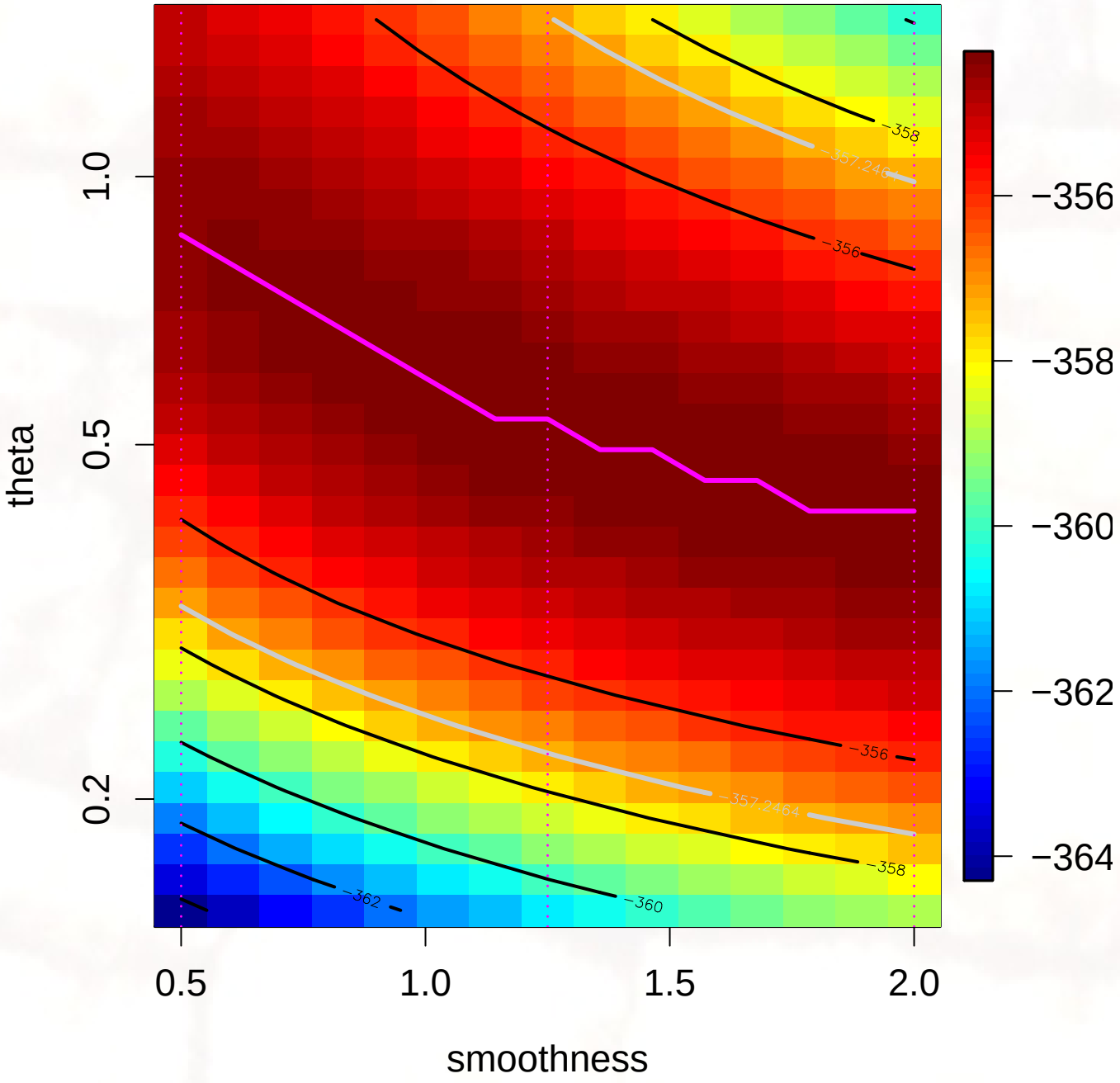
No elevation term and smoothness of 1.0



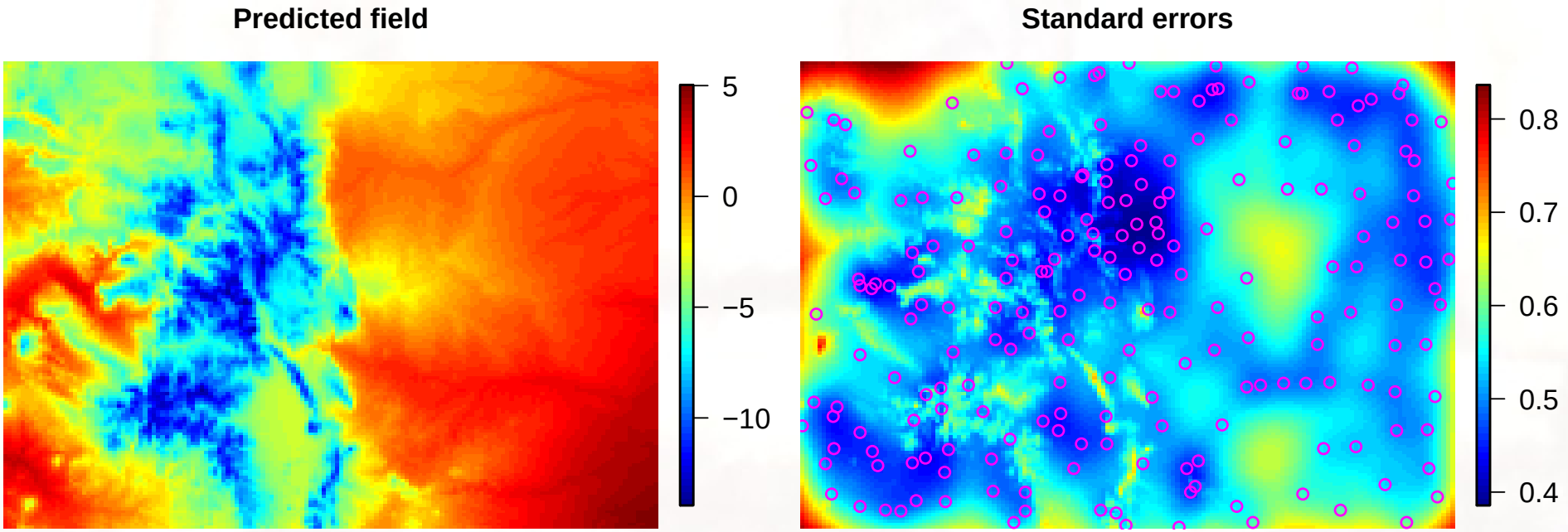
Line at approximate 95% confidence bound.

Likelihood θ and smoothness

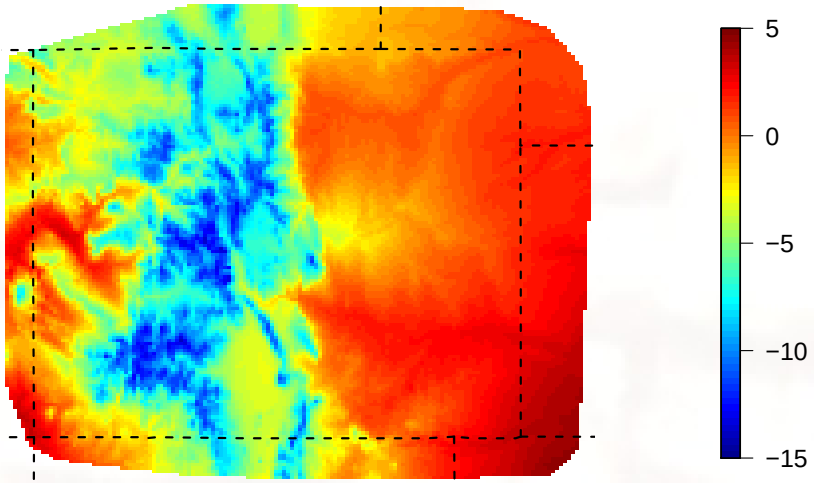
Marginal maxima, 95% confidence bounds



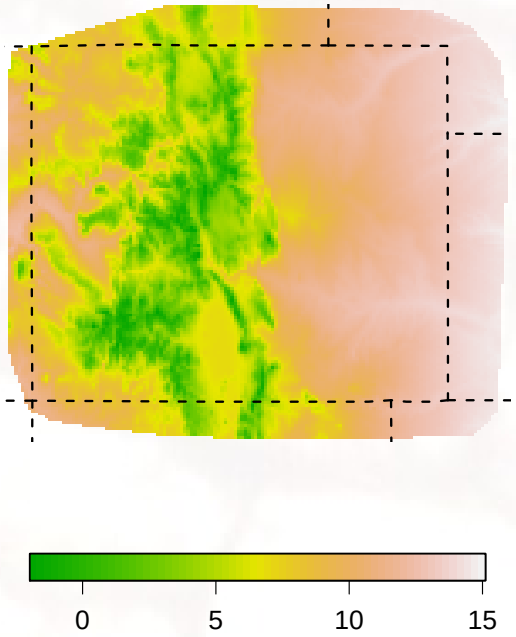
Predicted temperature field



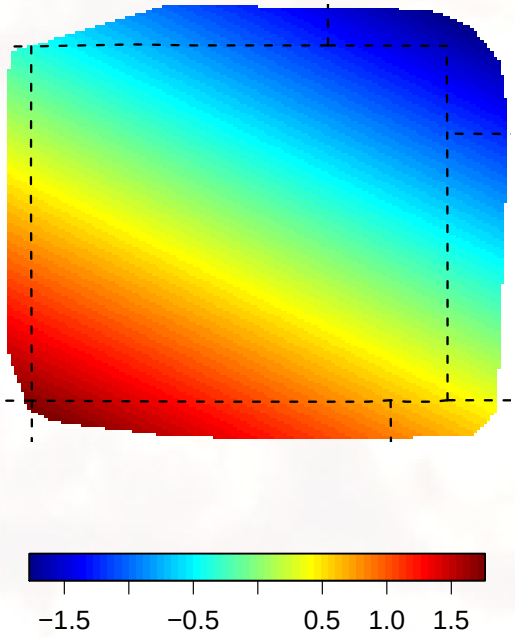
Decomposition of the fit



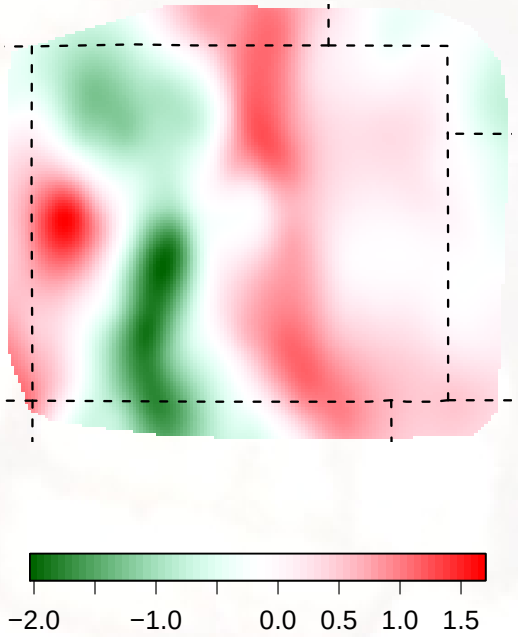
= elevation



+ longitude and latitude



+spatial process



Inference beyond the mean

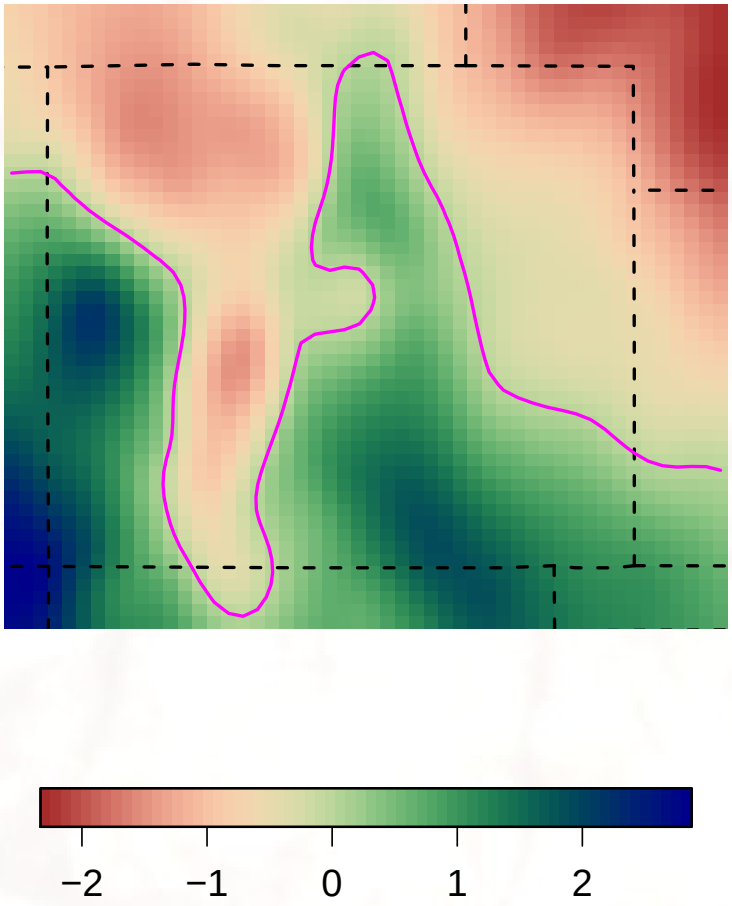
100 Weddings and the average

$\overline{wedding}$

J. Salavon *Cabinet* **15**

The mean surface w/o elevation

Components based on lon/lat and smooth surface.

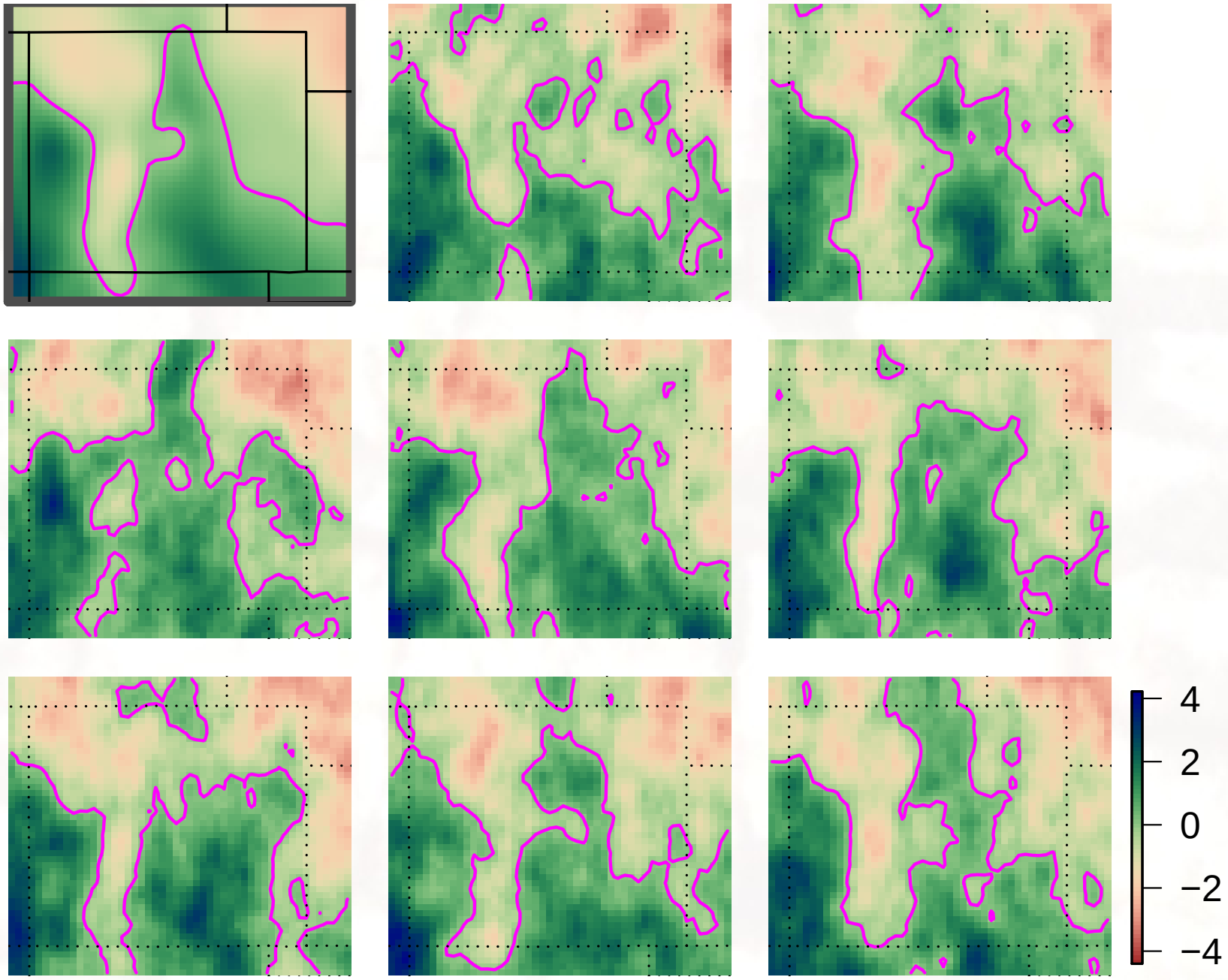


What is uncertainty of the zero contour?

Conditional simulation

For fixed covariance parameters sample conditional distribution of field given station data.

Ensemble for Colorado MAM temps



Analysis Work Flow

- EDA and plots
- Model for climate covariates
- Model for covariance
- Estimate parameters
- Compute conditional distribution (ensembles)

Thank you