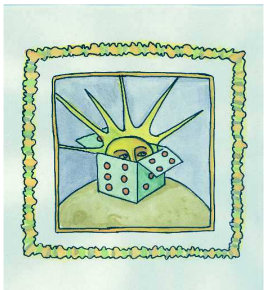
Infilling Sparse Records of Precipitation Fields

Douglas Nychka and Craig Johns
Chris Daly and Tim Kittel

Geophysical Statistics Project,
National Center for Atmospheric Research
www.cgd.ucar.edu/stats

- The problem
- Spatial models for precip
- Ensembles for uncertainty
- Computing
- Validation



National Center for Atmospheric Research

$\approx$ 1000 people total, several hundred PH D (physical) scientists,
half the budget ($\approx$ 60M) is a single grant from NSF-ATM

Research most aspects related to the atmosphere
Climate, Weather, the Sun, Ocean/atmosphere, Ecosystems, Economic impacts, Air quality, Instrumentation, Scientific computing and ...

Statistical methods for the geosciences

Complete historical climate “data sets”

The historical monthly precipitation record for the US consists of $\approx$ 12K stations reporting from 1985- present. But there are many missing station records.

Goal Create a complete set of observations:

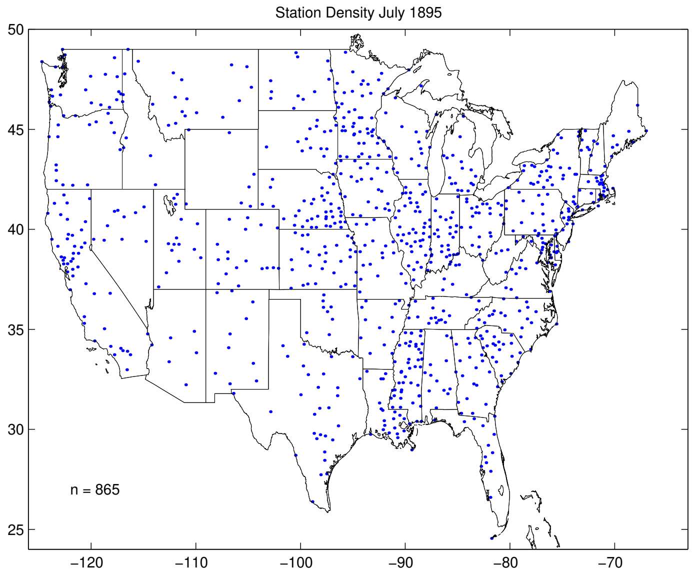
- For exercising ecological models.
- As a first step in interpolating to a regular, fine grid e.g. PRISM (Parameter-elevation Regressions on Independent Slopes Model)

Method should generate ensembles: Precipitation fields that are consistent with the data and equally likely. Variation among ensembles is a measure of uncertainty.

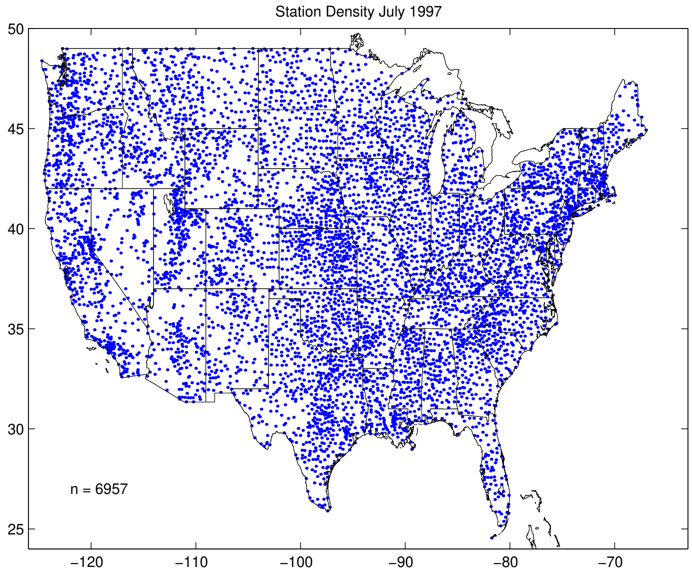
Challenge Balance adequate statistical models with efficient computational methods.

Solution Local Bayesian spatial models applied to scaled measurements and using covariance information from other time periods.
Nearest neighbor models make the “ production runs” possible in MATLAB.

An example of the changing station density.

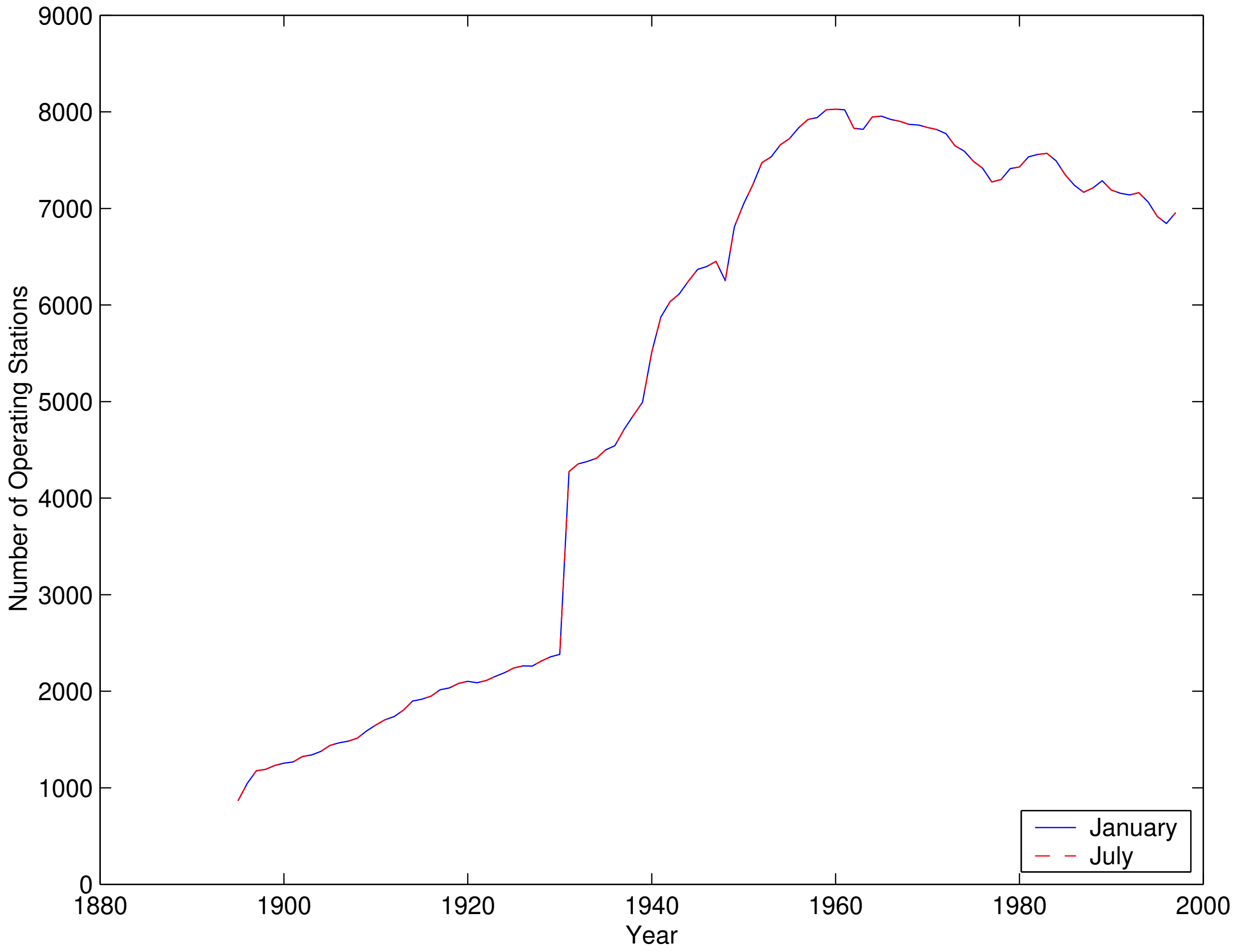


1895



1997

Number of stations operating over time.



Other problems with precip

- Non-Gaussian distribution
- Means vary over space
- Nonstationary covariance
- Using temporal structure

Infill estimates: Regression vs Kriging

z: station to infill at t

Z_t: neighboring station values

Two approaches to infilling:

NN Regression: For other time periods with data, regress station measurements on past nearest neighbors. Use regression relationship to predict at time when station is missing.

$$\mathbf{z} = Z\beta + error$$

$$\hat{z}_t = \mathbf{z}_t^T \hat{\beta}$$

Spatial prediction: Given a covariance function for the precipitation field predict the missing station value using geostatistics (Kriging).

$$\hat{z}_t = COV(z_t, Z_t) \left(COV(Z_t, Z_t)^{-1} \right) Z_t$$

Some Disadvantages

NN Regression Not useful if there are few close NNs with data.

Spatial prediction: Covariance function needs to be estimated. May not adapt to local correlations among stations.

Neither basic approach accounts for uncertainty in the model.

Bayes model

With a suitable transformation of raw precipitation ...

$[\ \mathbf{Z}_t \mid \Omega\ ] \sim N(0, \Omega)$ are independent multivariate normal.

Ω has Inverse-Wishart prior with centering K and degrees of freedom ν.

Observed $\mathbf{Z}_t^{(1)}$ and the unobserved $\mathbf{Z}_t^{(0)}$

The mean prediction for infilling:

$$\hat{\mathbf{Z}}_t^{(0)} = Q_{0\,1}\, Q_{1\,1}^{-1}\, \mathbf{Z}_t^{(1)} \tag{1}$$

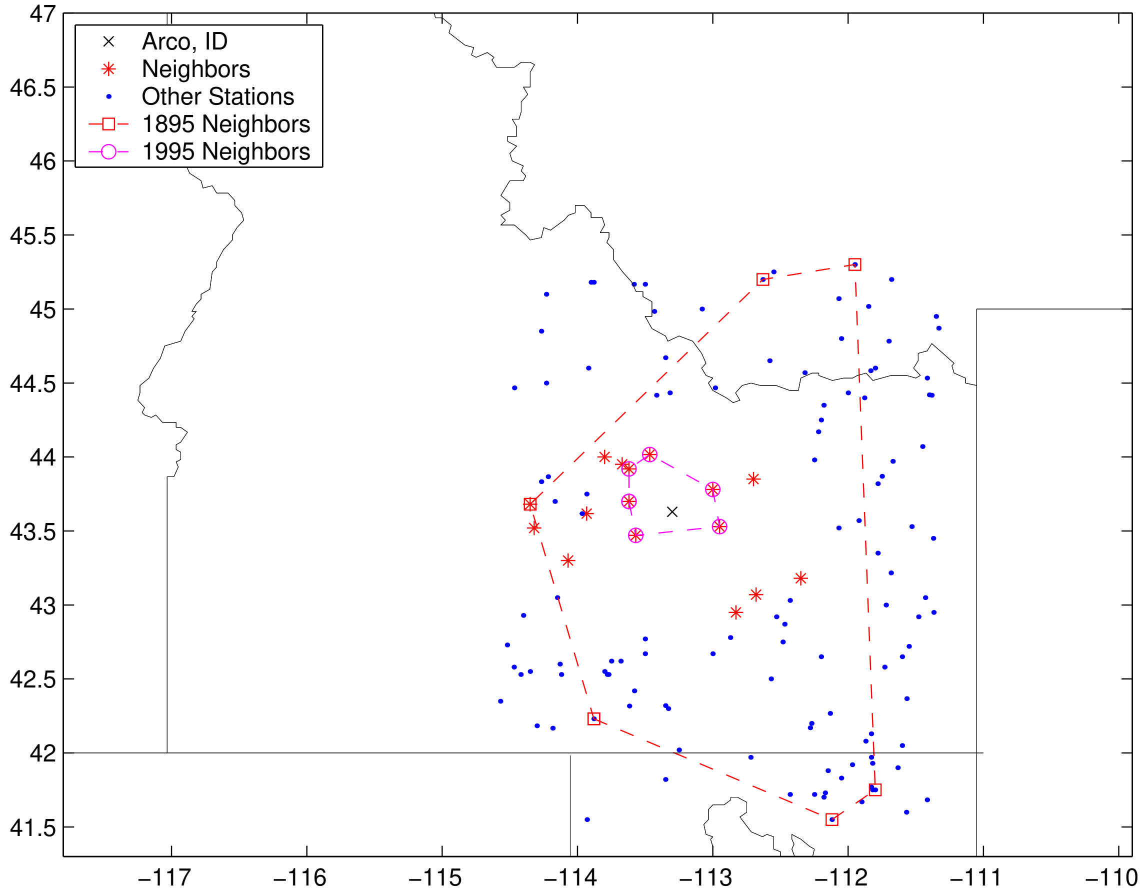
where Q is of the form

$$Q = \sum_{i \neq t} \mathbf{Z}_i \mathbf{Z}_i^T + \nu K$$

The key is that the covariance matrix Q is a sum of sample covariances (Regression $\nu = 0$)
and
a prior covariance (KrigingO $\nu \rightarrow \infty$)

Example of nearest neighbors

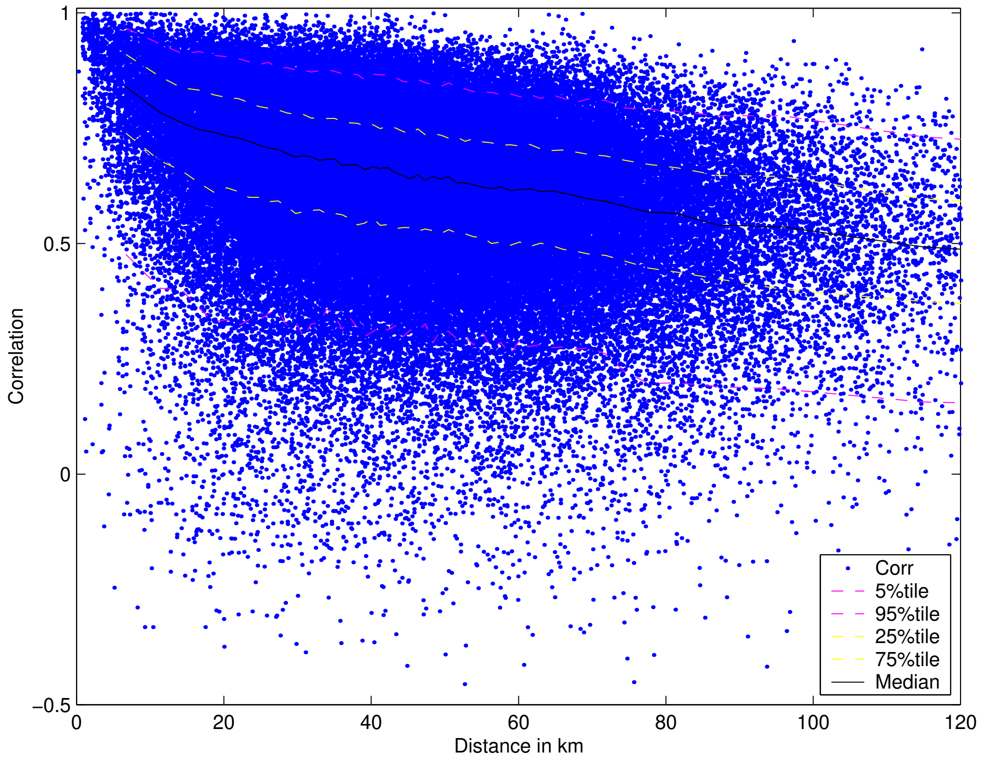
Neighborhoods for Arco, Idaho



Prior for Ω

ν controls the weight given to the prior covariance.
Some practical approaches

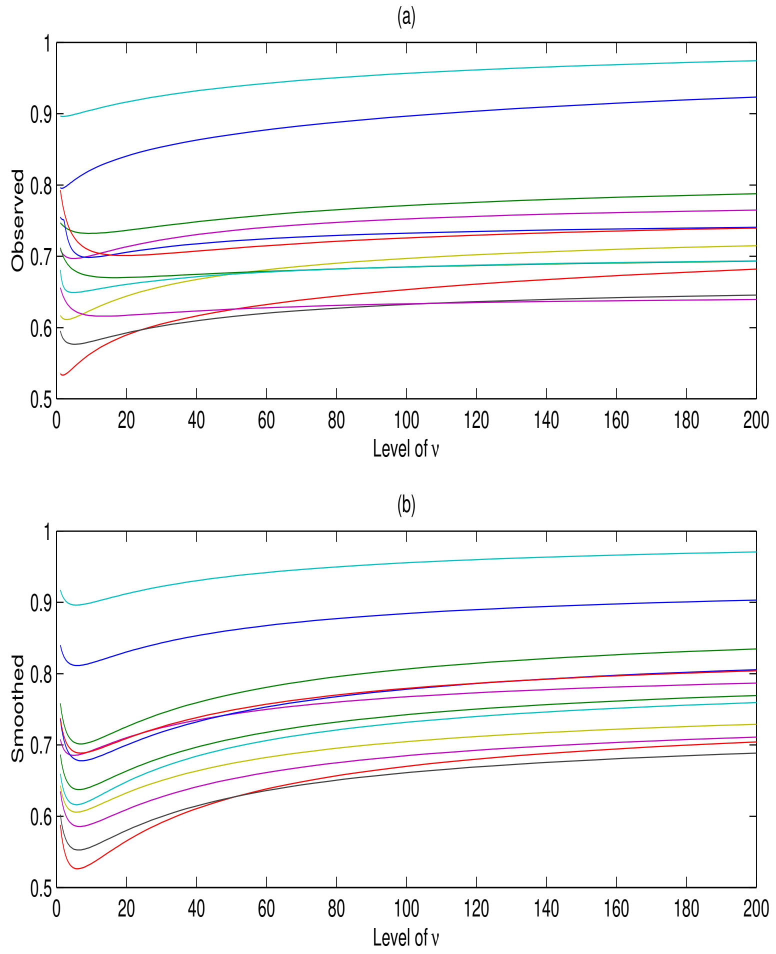
- Estimate K as a stationary covariance model from the data.
- Estimate spatially dependent values for ν using leave-one-out cross-validation.



Smoothing CV curves for representative stations

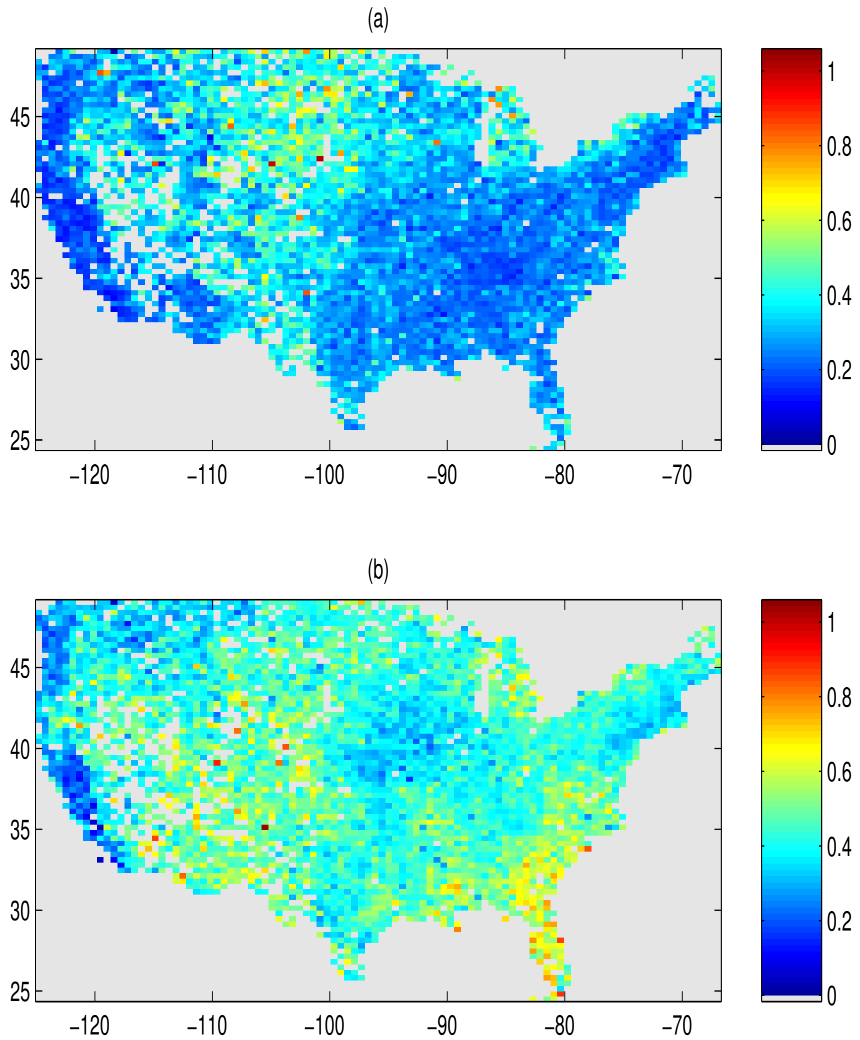
Vary ν over a grid and find the RMSE for the infill estimate when values are available.

Smooth across space and grid using EOFS.



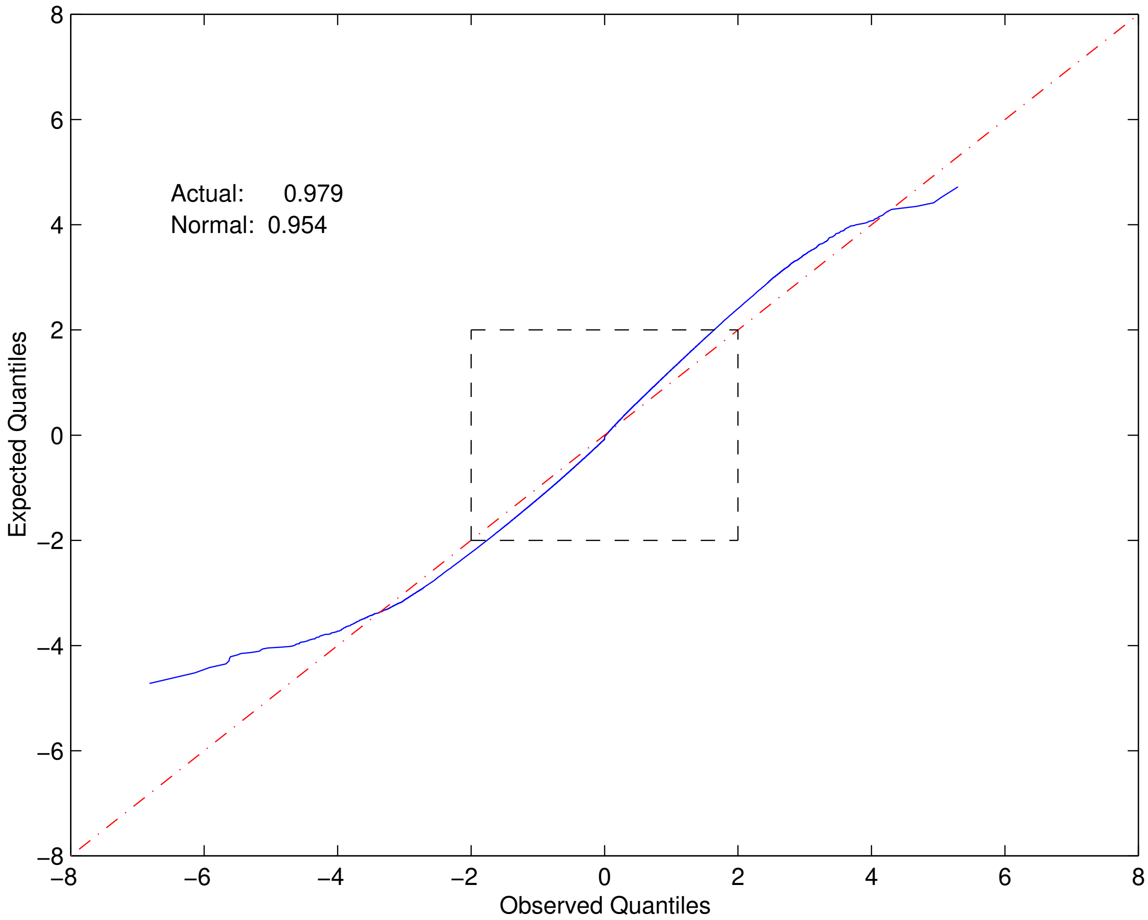
Results

CV Standardized Prediction Error (Jan, July)



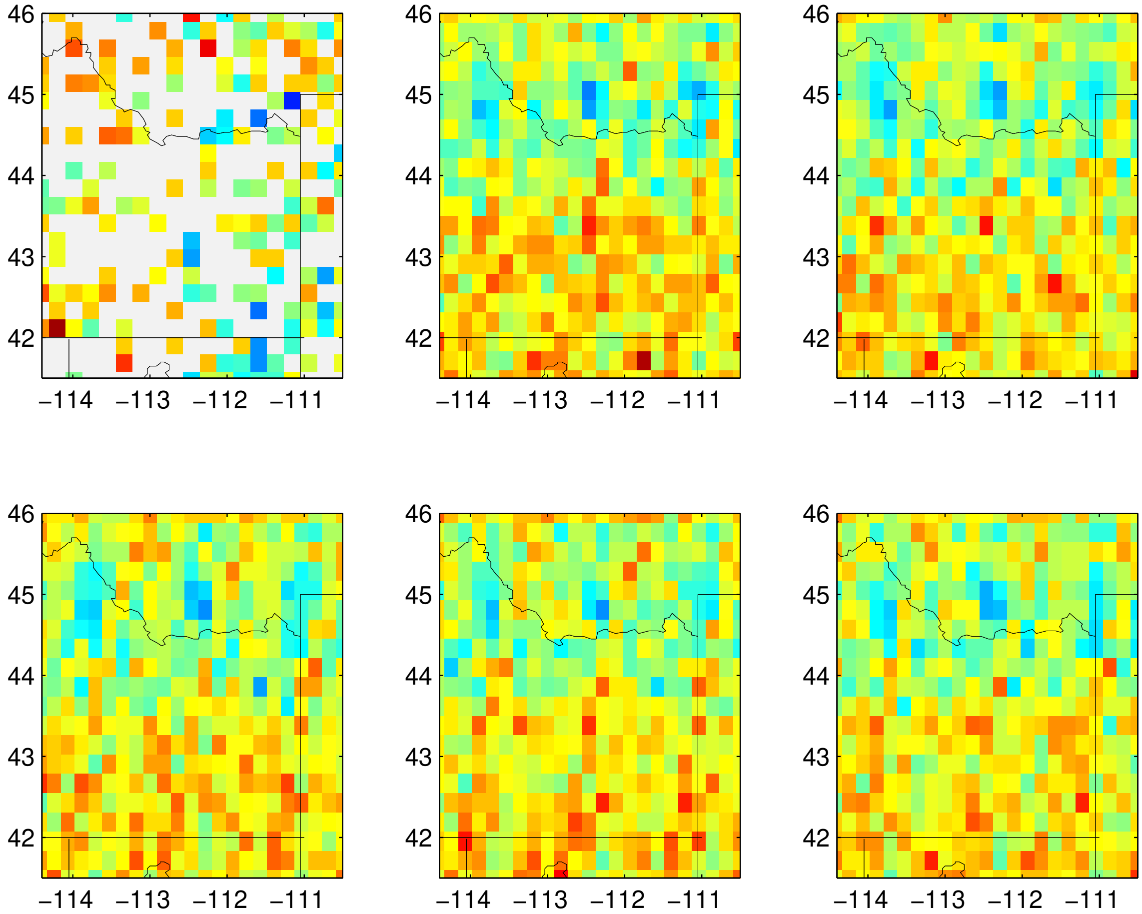
Validity of standard errors

Q-Q Plot of Standardized Prediction Errors



An example of ensembles

Southern Idaho, single time point.



Conclusions

- Regression and Kriging approaches can be combined using a Bayes model.
- Prior degrees of freedom (ν) can be estimated and has spatial structure.
- Ensembles can be generated that include nonstationary empirical covariance relationships.
- Nearest neighbor approximations make it computations feasible in MATLAB.