A Transformed, Thresholded Gaussian Model for Precipitation Extremes
or
What I Have Been Thinking About While Constantly Flying Between North Carolina and Colorado

Richard L. Smith
Department of Statistics and Operations Research
University of North Carolina, Chapel Hill
and
Geophysical Statistics Project
IMAGe
NCAR

Extreme Values Reading Group
March 30, 2006

I. INTRODUCTION AND MOTIVATION

II. STATISTICAL MODEL

III. ESTIMATING A THRESHOLDED GAUSSIAN PROCESS

IV. APPLICATION TO THE RAINFALL DATA

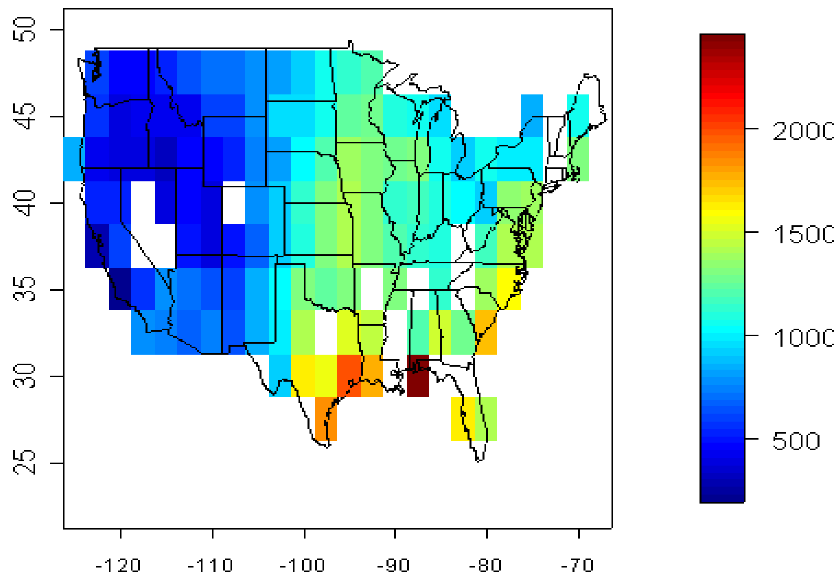
V. RESULTS

VI. SUMMARY AND CONCLUSIONS

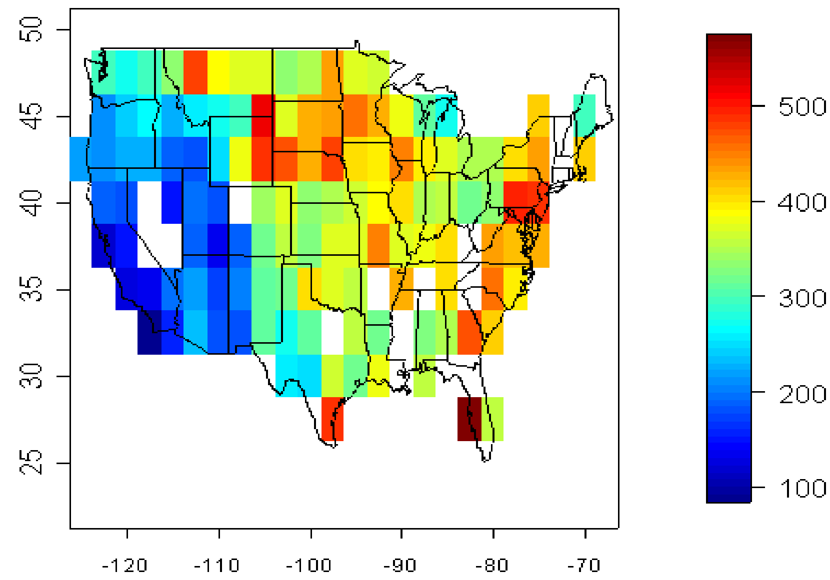
VII. REFERENCES

- Data from 5873 observational rainfall stations and 288 NCEP gridcells
- Restrict to 1970–1999 summer data and stations with no more than 10% missing data
- Threshold set at a given percentile of all available observations (including 0s)
- No declustering — assume each day is independent of every other
- Fit "point process" form of POT model and calculate 50-year return values
- For each grid cell containing at least 10 observational stations, compute RV50 from pooled data and compare with gridcell RV50
- Ratios of point to grid RV50 range from 1.34 to 8.85

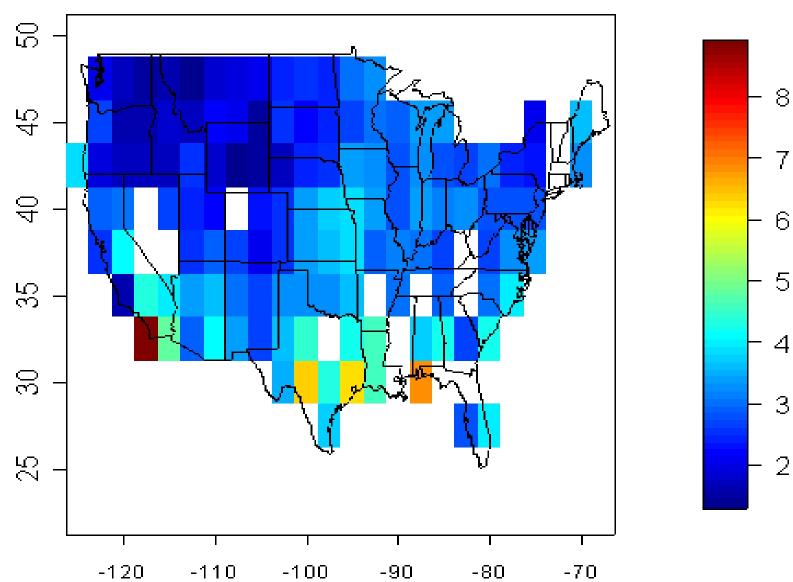
OBSV 50-Year Return Values



NCEP 50-Year Return Values



Ratio OBSV:NCEP



Objectives of the study:

1. Use spatial statistics to interpolate daily data and thereby estimate a grid-cell average for each day

2. Estimate 50-year return values based on estimated daily grid-cell averages

3. Provide an independent evaluation of how well NCEP is doing

The objectives are different from Elizabeth Shamseldin's project presented last week because I don't directly address the down-scaling issue. I view it as a complementary project aimed at evaluating the quality of NCEP or any other reanalysis or RCM we may choose to examine.

II. STATISTICAL MODEL

- Fit GEV to tails of distribution — equivalent to

$$1 - F(y) = \Pr\{Y > y\} \approx \frac{1}{T}\left(1 + \xi\frac{y-\mu}{\psi}\right)_+^{-1/\xi}, \; y \geq u,$$

u is threshold, $x_+ = \max(x, 0)$ and T is number of relevant days per year (here 92)

- Estimate $F(0)$ by sample proportion of zeros

- For $0 < y < u$, divide range into 20 equiprobable intervals and assume $F(y)$ is piecewise linear within each interval. Thus we have an estimate of $F(y)$ for the entire range of y

- Define $Z = \Phi^{-1}(F(Y))$ so that Z has marginal $N[0,1]$ distribution. Values $Y = 0$ are transformed into $u^* = \Phi^{-1}(F(0))$ which becomes the *natural threshold* — Z values censored at u^*.

- We assume the underlying Z process is a Gaussian spatial process.

Related ideas in the literature:

Coles and Tawn (1996) developed a similar approach but assuming Z is max-stable. They argued that this is more suitable for studying extremal properties but it is unclear how realistic this model is. Also, their paper is based on a particular representation of max-stable processes and the estimation methods are computationally intensive for the kind of applications we have in mind.

Sansó and Guenni (2000, 2004) proposed an embedded Gaussian model similar in spirit to what I propose here. They fitted the model in a fully Bayesian approach with MCMC. However their data examples are much more limited than the data considered here and I question whether a computationally intensive MCMC approach is feasible for this problem.

III. ESTIMATING A THRESHOLDED GAUSSIAN PROCESS

Basic model: $Y = (Y_1, ..., Y_n)^T$ has a multivariate normal distribution, $N[0, \Sigma(\theta)]$ where $\Sigma(\theta)$ is a known function of finite-dimensional parameter θ. However we observe only those values Y_i for which $Y_i > u$ — rest are censored.

This model is replicated independently for each of N days.

How should we estimate θ?

On any given day, let $A = \{i : \ Y_i > u\}$ and $B = \{i : \ Y_i \le u\}$. Suppose $|B| = m$. Also let Y_A, Y_B be the corresponding subvectors.

If we condition on Y_A, then we can write down the (MVN) conditional distribution of Y_B. Therefore, the contribution from this day to the exact likelihood is given by

$$f(Y_A) \times \text{Pr}\{Y_i \le u, \ i \in B \mid Y_A\}$$

and could be evaluated exactly if we had an efficient algorithm to evaluate the m-dimensional multivariate normal distribution.

Option 1: Use exact or simulated MVN distribution function.

Exact algorithms due to, e.g. Schervish (1984), but does not work well in high dimensions. Usual recommendation in high dimensions is simulation, but this takes us back to MCMC evaluation.

Option 2: Use EVT approximation to MVN distribution.

Possible use of Stein-Chen method, e.g. Roos (1994), Raab (1998)

It's not clear to me that the error in such approximations is small (in this sort of context), or exactly how one would determine that issue.

Option 3: Break up into a series of bivariate normal approximations, for which there are well-established algorithms, e.g. Owen (1956), Donnelly (1973), Young and Minder (1974).

So if we reorder the data so that $B = \{1, ..., m\}$, $A = \{m+1, ..., n\}$, the idea is to approximate $\Pr\{Y_i \le u,\ i = 1, ..., m \mid Y_A\}$ by

$$\begin{aligned}
\Pr\{Y_1 \le u \mid Y_A\} \quad & \times \quad \Pr\{Y_2 \le u \mid Y_1 \le u, Y_A\} \\
& \times \quad \Pr\{Y_3 \le u \mid Y_2 \le u, Y_A\} \\
& \vdots \\
& \times \quad \Pr\{Y_m \le u \mid Y_{m-1} \le u, Y_A\}.
\end{aligned}$$

Problem with this: does not lead to consistent estimators.

Consider $n = 3$. Break up likelihood into 8 components corresponding to $(Y_1 > u,\ Y_2 > u,\ Y_3 > u)$, $(Y_1 > u,\ Y_2 > u,\ Y_3 \le u)$,..., $(Y_1 \le u,\ Y_2 \le u,\ Y_3 \le u)$. Only the last of these is different from exact MLE so let's concentrate on that case.

Exact likelihood would be

$$\mathrm{Pr}\{Y_1 \le u\} \cdot \mathrm{Pr}\{Y_2 \le u \mid Y_1 \le u\} \cdot \mathrm{Pr}\{Y_3 \le u \mid Y_1 \le u, Y_2 \le u\}.$$

Approximate this by

$$\mathrm{Pr}\{Y_1 \le u\} \cdot \mathrm{Pr}\{Y_2 \le u \mid Y_1 \le u\} \cdot \mathrm{Pr}\{Y_3 \le u \mid Y_2 \le u\}.$$

The difference in log likelihoods is

$$\ell_{\text{approx}} - \ell_{\text{exact}} = \log \frac{\mathrm{Pr}\{Y_3 \le u \mid Y_2 \le u\}}{\mathrm{Pr}\{Y_3 \le u \mid Y_1 \le u, Y_2 \le u\}}.$$

When this is differentiated with respect to θ we have

$$E\left\{\frac{\partial \ell_{\text{exact}}}{\partial \theta}\right\} = 0$$

and so

$$\begin{aligned} E\left\{\frac{\partial \ell_{\text{approx}}}{\partial \theta}\right\} &= \frac{\partial}{\partial \theta}\left\{\log \frac{\Pr\{Y_3 \le u \mid Y_2 \le u\}}{\Pr\{Y_3 \le u \mid Y_1 \le u, Y_2 \le u\}}\right\} \\ &\quad \cdot \Pr\{Y_1 \le u, Y_2 \le u, Y_3 \le u\}. \end{aligned}$$

Unless Y_1 and Y_3 are conditionally independent given Y_2, this expression will not be 0.

In other words, the estimating equations are not unbiased. Typically this is a necessary condition for consistency.

However, this suggests another approach.

Option 4: Apply the pairwise principle to the whole of the likelihood, not just part of it.

In other words, for a *fixed* ordering of indices, replace exact LH

$$L(Y_1) \cdot L(Y_2 \mid Y_1) \cdot L(Y_3 \mid Y_2, Y_1)...$$

by a pairwise approximation

$$L(Y_1) \cdot L(Y_2 \mid Y_1) \cdot L(Y_3 \mid Y_2)...$$

Here $L(Y_{i+1} \mid Y_i)$ is the conditional likelihood of Y_{i+1} given Y_i allowing for censoring, i.e.

1. If $Y_i > u,\ Y_{i+1} > u$, $L = f(Y_{i+1} \mid Y_i)$,
2. If $Y_i > u,\ Y_{i+1} \le u$, $L = \mathrm{Pr}\{Y_{i+1} \le u \mid Y_i\}$,
3. If $Y_i \le u,\ Y_{i+1} > u$, $L = \frac{\mathrm{Pr}\{Y_i \le u \mid Y_{i+1}\} f(Y_{i+1})}{\mathrm{Pr}\{Y_i \le u\}}$,
4. If $Y_i \le u,\ Y_{i+1} \le u$, $L = \frac{\mathrm{Pr}\{Y_i \le u, Y_{i+1} \le u\}}{\mathrm{Pr}\{Y_i \le u\}}$.

This is similar to proposals for approximate likelihood for spatial processes by Vecchia (1988) and Stein, Chi and Welty (2004).

The estimating equations are unbiased, essentially because each component of the approximate log likelihood is.

Therefore, estimates are consistent and nearly unbiased.

Not asymptotically efficient, but we can estimate approximate variances through information sandwich approximation.

Simulation Study

Use Matérn covariance function:

$$C_0(t) = \frac{\theta_1}{2^{\theta_3 - 1}\Gamma(\theta_3)} \left(\frac{2\sqrt{\theta_3} t}{\theta_2} \right)^{\theta_3} \mathcal{K}_{\theta_3} \left(\frac{2\sqrt{\theta_3} t}{\theta_2} \right).$$

Here $\theta_1, \theta_2, \theta_3$ are respectively scale, range and shape parameters.

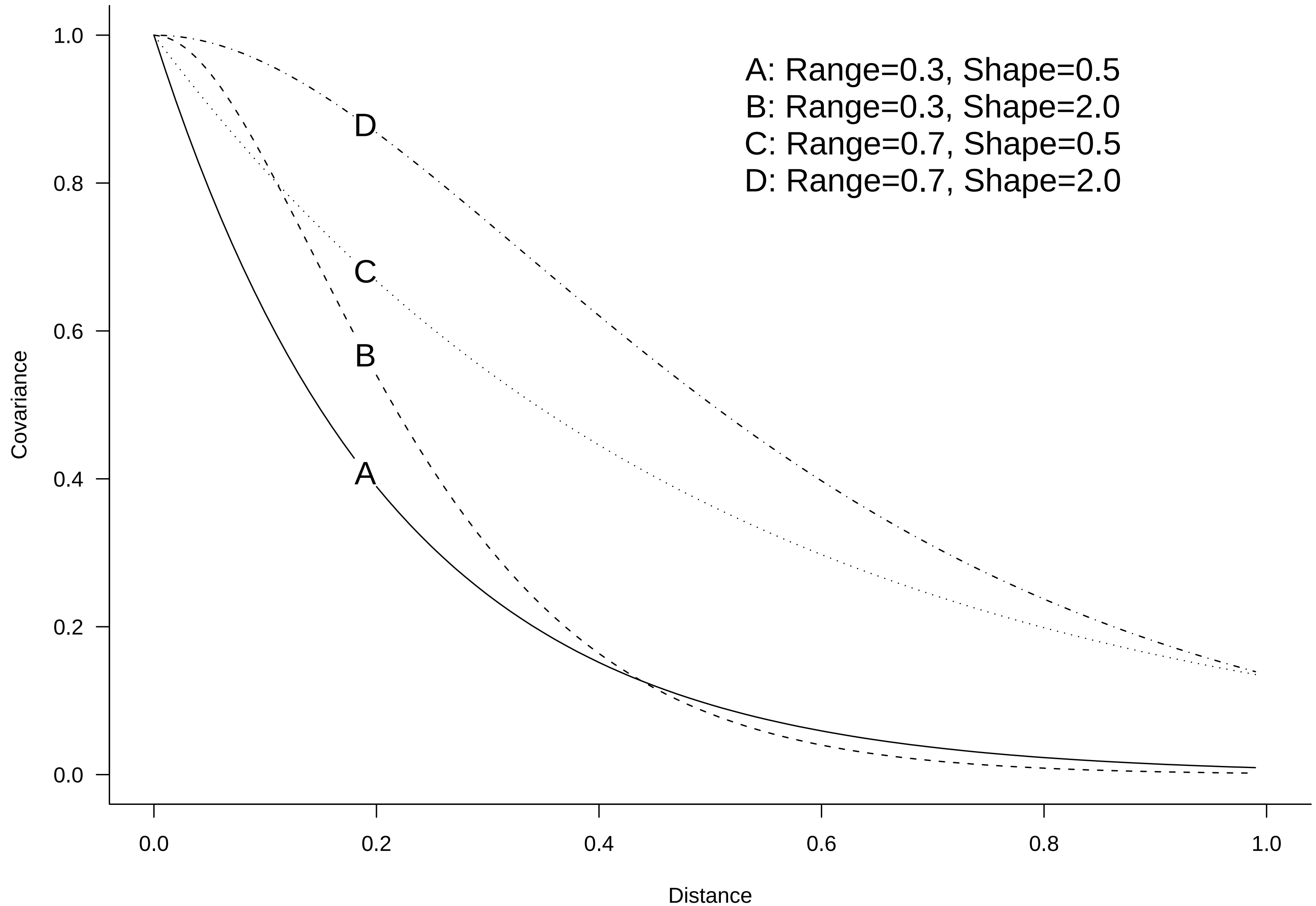
Simulate station locations in a 1×1 unit region — each sample consists of 100 days' independent data at 50 locations.

Employ two values of each of θ_2 and θ_3 (see Fig.).

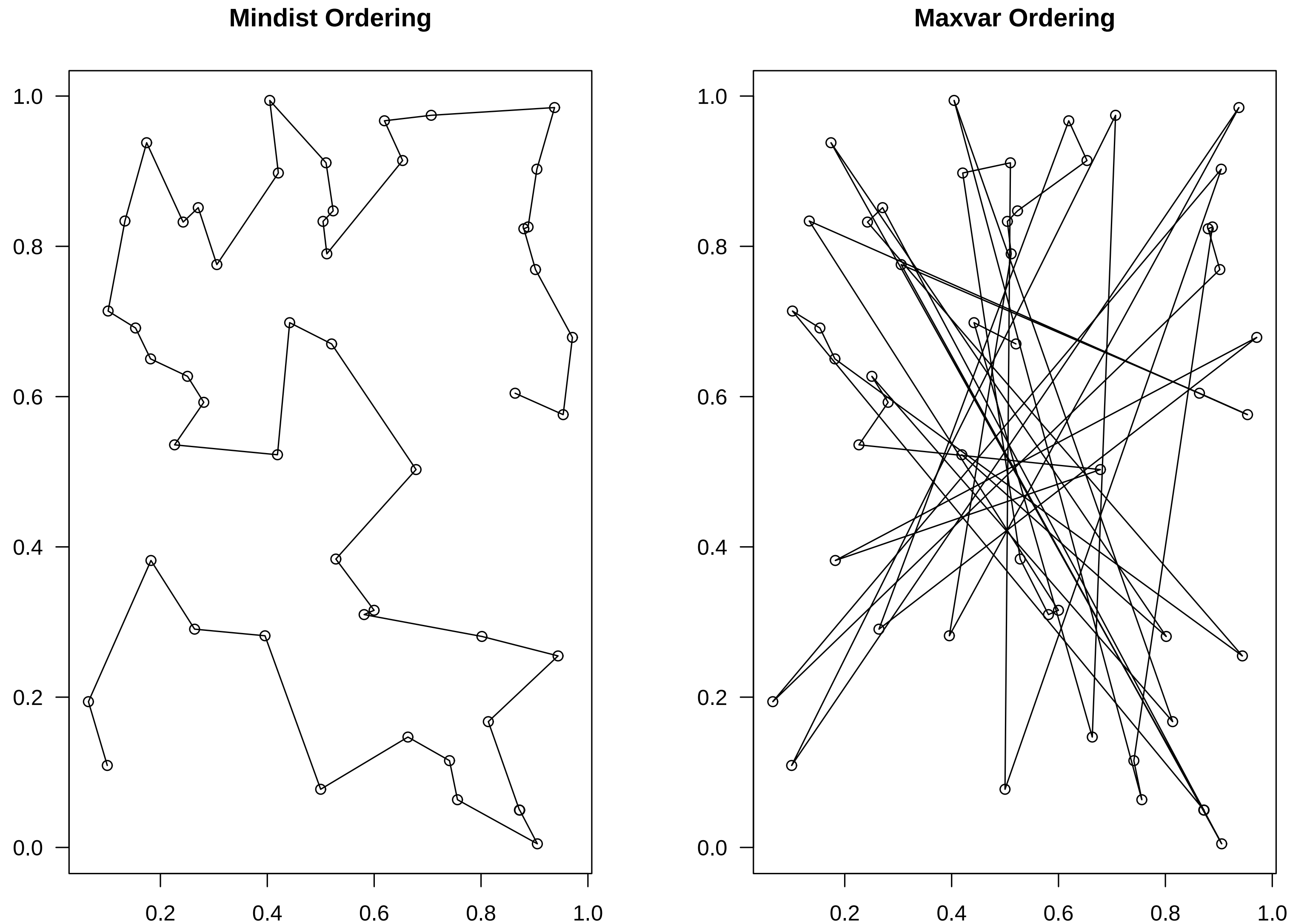
Use two methods of ordering stations: *mindist* and *maxvar*.

PL1, PL2 are pairwise likelihoods using maxvar, mindist orderings respectively

Options for threshold: none, $u = 1$ or $u = 2$. Exact MLE available only in no-threshold case.



Four Matern covariances used for simulation study



Two methods of ordering the spatial locations

The full simulation consists of 1000 replications used to compute:

1. Mean and RMSE of estimator
2. Coverage probability of nominal 90% and 95% confidence intervals *without* information sandwich correction (CP1, CP2)
3. Coverage probability of nominal 90% and 95% confidence intervals *with* information sandwich correction (CP3, CP4)

True scale=1.0, range=0.3, shape=0.5

Parameter	Estimator	u	Mean	RMSE	CP1	CP2	CP3	CP4
Scale	MLE	—	.998	.030	90	95	—	—
Scale	PL1	—	.998	.031	70	78	88	94
Scale	PL2	—	.998	.031	81	88	88	93
Scale	PL1	1	.999	.054	64	73	90	94
Scale	PL2	1	.999	.053	77	85	88	94
Scale	PL1	2	.999	.050	76	84	91	95
Scale	PL2	2	.999	.050	82	89	90	95
Range	MLE	—	.299	.012	89	95	—	—
Range	PL1	—	.299	.019	82	89	89	94
Range	PL2	—	.299	.014	83	89	89	94
Range	PL1	1	.298	.039	85	90	87	91
Range	PL2	1	.297	.027	84	91	87	92
Range	PL1	2	.300	.104	81	85	77	82
Range	PL2	2	.292	.066	84	89	82	87
Shape	MLE	—	.501	.011	89	93	—	—
Shape	PL1	—	.501	.015	86	92	88	93
Shape	PL2	—	.501	.013	87	93	87	92
Shape	PL1	1	.506	.037	86	92	86	91
Shape	PL2	1	.505	.031	88	93	86	91
Shape	PL1	2	.539	.119	87	93	74	82
Shape	PL2	2	.531	.098	90	95	78	85

True scale=1.0, range=0.7, shape=0.5

Parameter	Estimator	u	Mean	RMSE	CP1	CP2	CP3	CP4
Scale	MLE	—	1.000	.047	90	95	—	—
Scale	PL1	—	1.000	.052	49	58	89	94
Scale	PL2	—	1.000	.053	71	79	88	94
Scale	PL1	1	1.000	.084	48	56	88	93
Scale	PL2	1	1.000	.084	71	79	89	94
Scale	PL1	2	0.989	.087	52	60	84	90
Scale	PL2	2	.991	.088	66	75	85	91
Range	MLE	—	.702	.043	89	94	—	—
Range	PL1	—	.704	.074	48	56	89	93
Range	PL2	—	.701	.049	77	84	88	92
Range	PL1	1	.697	.118	58	67	85	90
Range	PL2	1	.699	.087	81	89	88	92
Range	PL1	2	.635	.226	71	80	69	74
Range	PL2	2	.674	.190	81	86	78	83
Shape	MLE	—	.500	.009	90	94	—	—
Shape	PL1	—	.500	.016	69	78	89	94
Shape	PL2	—	.501	.011	89	95	88	94
Shape	PL1	1	.504	.031	77	85	88	93
Shape	PL2	1	.503	.027	90	95	86	93
Shape	PL1	2	.536	.087	85	92	76	82
Shape	PL2	2	.521	.078	92	96	80	87

True scale=1.0, range=0.3, shape=2.0

Parameter	Estimator	u	Mean	RMSE	CP1	CP2	CP3	CP4
Scale	MLE	—	.999	.033	89	94	—	—
Scale	PL1	—	.999	.036	64	73	90	95
Scale	PL2	—	.998	.036	86	92	89	94
Scale	PL1	1	.999	.060	60	70	91	96
Scale	PL2	1	.999	.058	86	92	90	95
Scale	PL1	2	.998	.063	68	76	89	95
Scale	PL2	2	.999	.063	82	90	89	95
Range	MLE	—	.300	.007	90	95	—	—
Range	PL1	—	.299	.027	78	84	87	91
Range	PL2	—	.299	012	90	95	89	94
Range	PL1	1	.295	.047	69	73	75	79
Range	PL2	1	.298	.026	89	92	86	91
Range	PL1	2	.297	.074	73	81	72	79
Range	PL2	2	.292	.052	75	80	73	79
Shape	MLE	—	2.005	.062	89	95	—	—
Shape	PL1	—	2.115	.48	78	81	83	87
Shape	PL2	—	2.034	.20	94	96	92	95
Shape	PL1	1	3.43	3.21	62	64	64	67
Shape	PL2	1	2.18	.61	84	88	82	86
Shape	PL1	2	5.46	5.36	44	47	42	44
Shape	PL2	2	4.39	4.34	55	58	53	55

True scale=1.0, range=0.7, shape=2.0

Parameter	Estimator	u	Mean	RMSE	CP1	CP2	CP3	CP4
Scale	MLE	—	1.001	.051	79	84	—	—
Scale	PL1	—	1.002	.067	44	49	90	95
Scale	PL2	—	1.002	.069	86	93	90	94
Scale	PL1	1	1.003	.104	45	50	89	93
Scale	PL2	1	1.002	.106	85	89	90	94
Scale	PL1	2	.995	.109	47	55	87	92
Scale	PL2	2	1.010	.115	78	86	90	93
Range	MLE	—	.701	.106	81	85	—	—
Range	PL1	—	.703	.079	35	41	85	91
Range	PL2	—	.700	.050	86	90	87	90
Range	PL1	1	.694	.128	42	47	78	83
Range	PL2	1	.691	.098	74	78	75	80
Range	PL1	2	.638	.189	45	53	58	64
Range	PL2	2	.711	.159	75	80	73	80
Shape	MLE	—	1.999	.028	84	88	—	—
Shape	PL1	—	2.098	.471	48	60	81	85
Shape	PL2	—	2.045	.298	83	87	82	85
Shape	PL1	1	2.99	2.43	54	57	70	73
Shape	PL2	1	2.95	2.57	65	69	63	66
Shape	PL1	2	6.36	5.88	36	37	38	40
Shape	PL2	2	4.57	4.63	51	55	45	48

Conclusions from simulation study

1. Estimates generally unbiased except for shape parameter in threshold situation when true shape parameter is 2

2. RMSE is smallest for exact MLE, increases for approximate MLE as threshold rises

3. PL2 generally better than PL1 (i.e. prefer minimum-distance ordering of stations)

4. There is an overall problem of undercoverage of the confidence intervals. In many cases the information sandwich correction helps the situation, but not all.

IV. APPLICATION TO THE RAINFALL DATA

Step 1: Estimating the GEV and spatial parameters

- Common distribution for all stations within a grid cell
- Fit GEV to exceedances over a high threshold based on pooled data from stations. In most cases I used the 0.975 empirical quantile but in a few cases a higher threshold.
- Use piecewise linear approximation for CDF below threshold.
- Transform to normality, fit Gaussian model. In many cases, the Matérn covariance did not result in a satisfactory model fit so I switched to the exponential model with nugget

  $$C_0(t) = \theta_1 I(t=0) + \theta_2 \exp\left(-\frac{t}{e^{\theta_3}}\right) I(t>0).$$

  Estimation via pairwise likelihood based on ordering of stations that minimizes total distance (simulated annealing).
- Used natural threshold u^*, but also considered alternatives $u > u^*$ (a separate issue from the choice of threshold for initial GEV fit).

Example: Station 185, natural threshold $u^* = .5306$ (parameter estimates with SE by information sandwich in parentheses)

Threshold	$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\theta}_3$
.5306	1.037 (.015)	.665 (.015)	1.37 (.04)
1	1.007 (.013)	.630 (.017)	.84 (.04)
2	1.008 (.018)	.724 (.062)	.15 (.37)

- Range θ_3 decreases with increasing threshold, implying same Gaussian process does not apply to all levels.
- Generalization: replace Z by a mixture of Gaussian processes?
- Despite the clearly significant difference in the models fitted to different thresholds, it does not seem to make much difference to the interpolations (specific example later)

Step 2: Interpolating the missing and censored observations in the Gaussian process

Use MCMC fixing spatial covariance parameters. 1000 warm-up iterations, followed by either 500 or 1000 iterations.

Step 3: Interpolating Gaussian process to a 30×30 *array of locations within the grid cell*

Use second phase of MCMC in Step 2, every 10th iteration generate full sample of values conditional on observed and imputed data at observation locations.

Suppose $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ where index 1 represents observation stations and index 2 represents the 900 interpolation points. Calculate Cholesky $\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = U^T U$. Then simulate $Z_2 = \Sigma_{21}\Sigma_{11}^{-1}Z_1 + U^T Z_0$ where Z_0 is white noise.

A trick: All values of U less than .01 were set to 0. This speeds up computation time by a factor typically between 4 and 10.

Step 4: Transform each predicted value back to original marginal distribution by inverting initial transformation step.

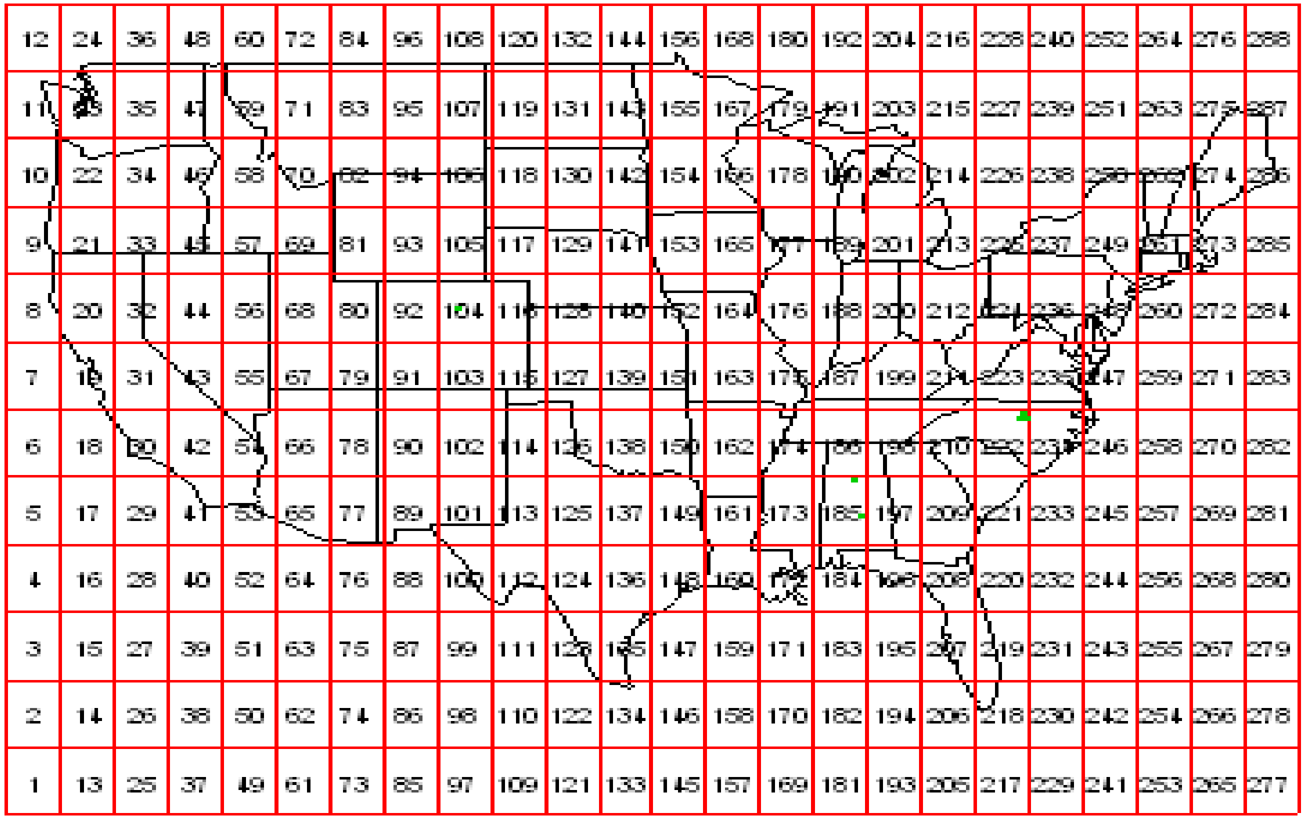
Step 5: Compute summary statistics

For each of the 900 interpolation points, average over iterations to obtain a single "predicted value" for that location for each day.

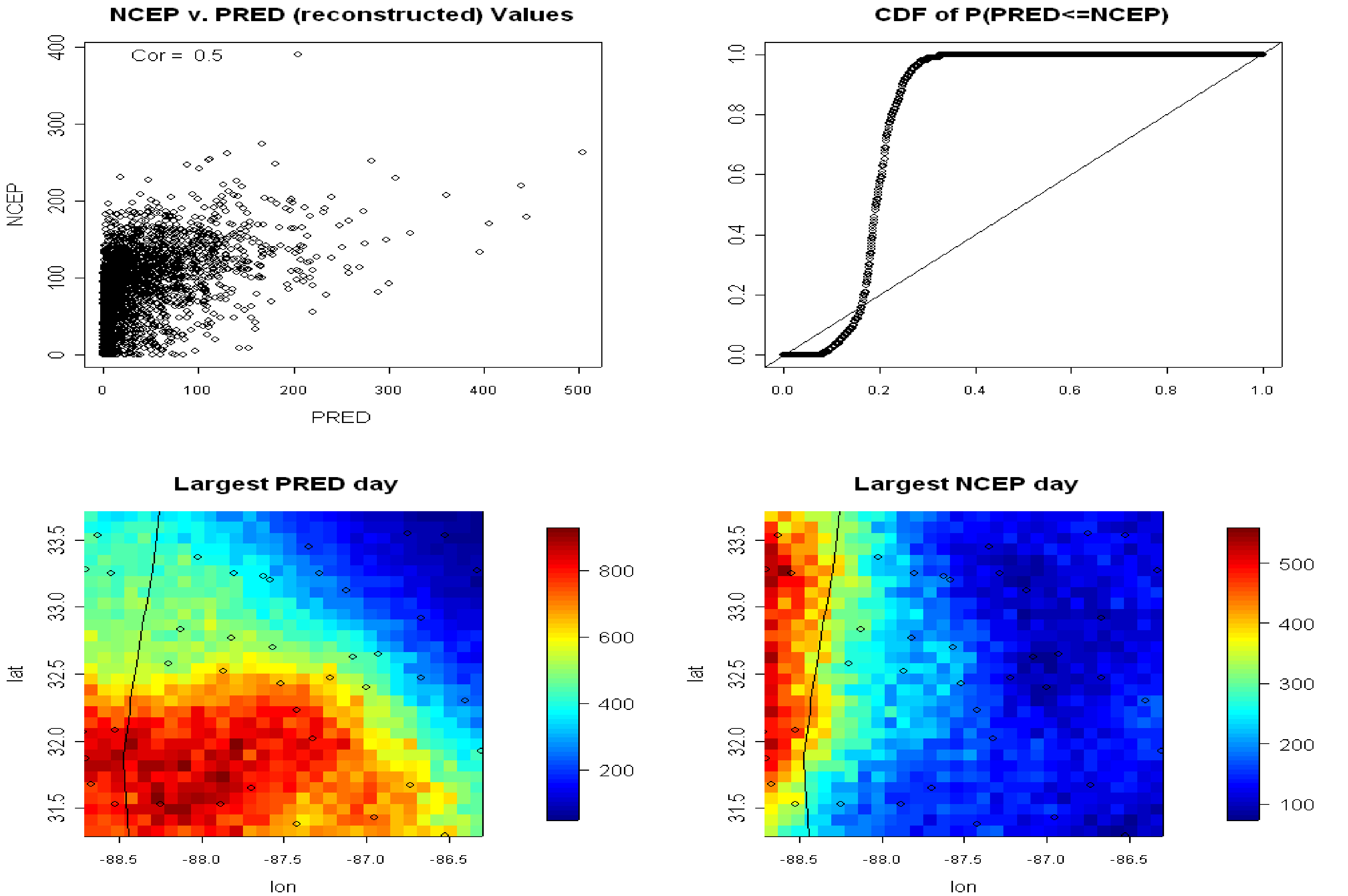
Also average over the interpolation points to produce a predicted value for grid-cell average, henceforth denoted PRED.

However we also utilize the information from the individual iterations to calculate a predictive distribution for the grid-cell average for each day.

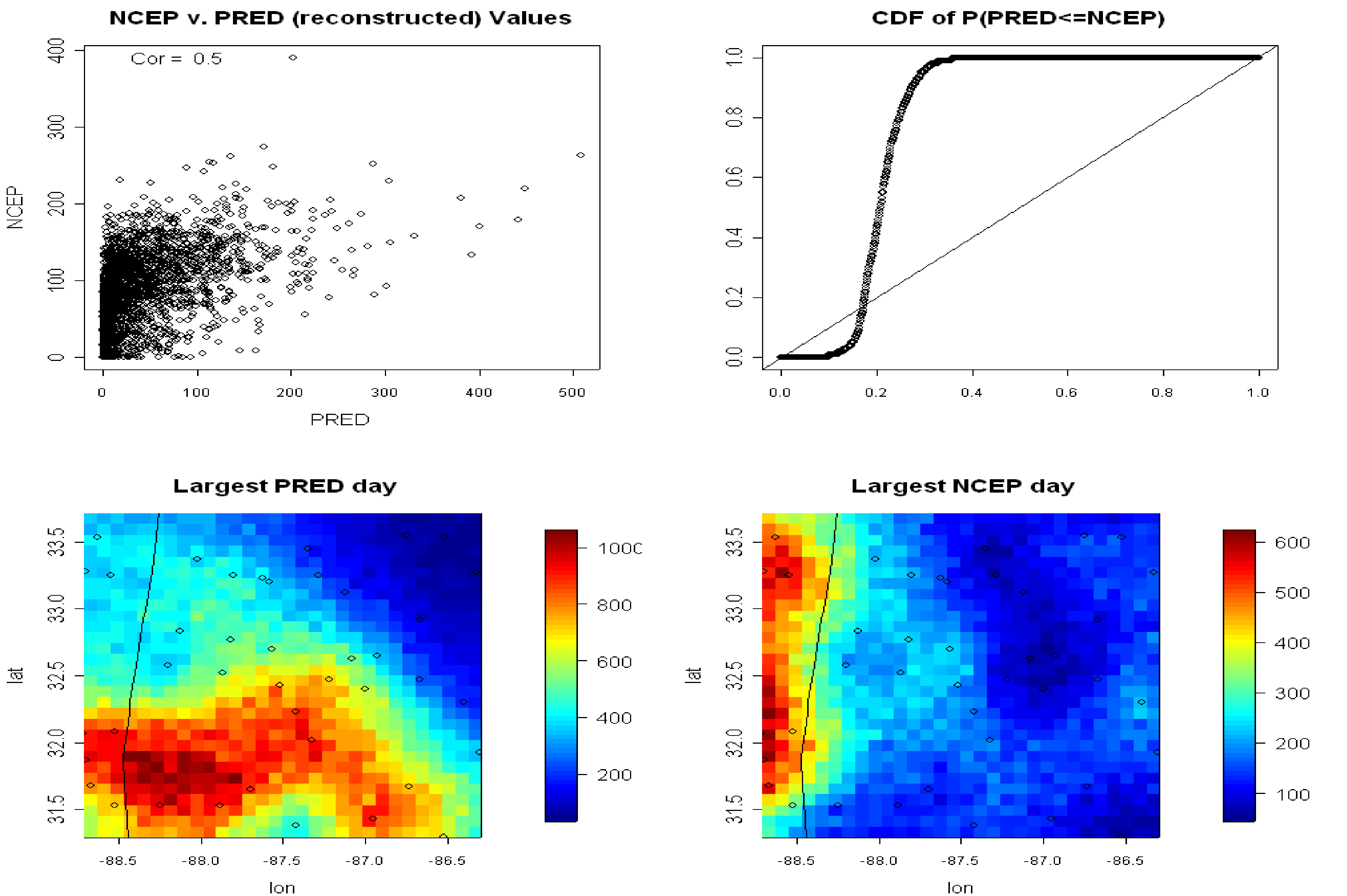
V. RESULTS



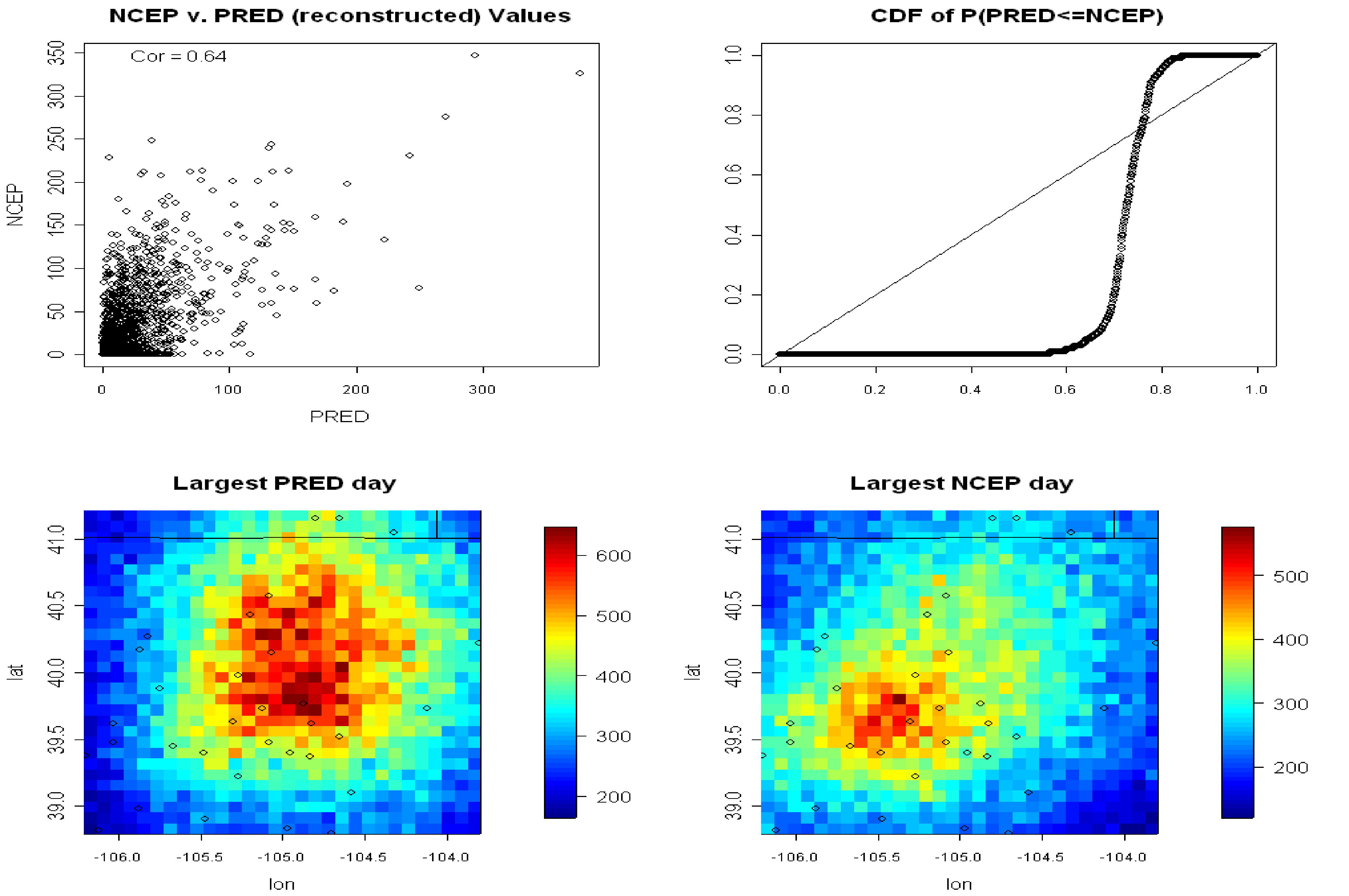
Grid Cell Representation



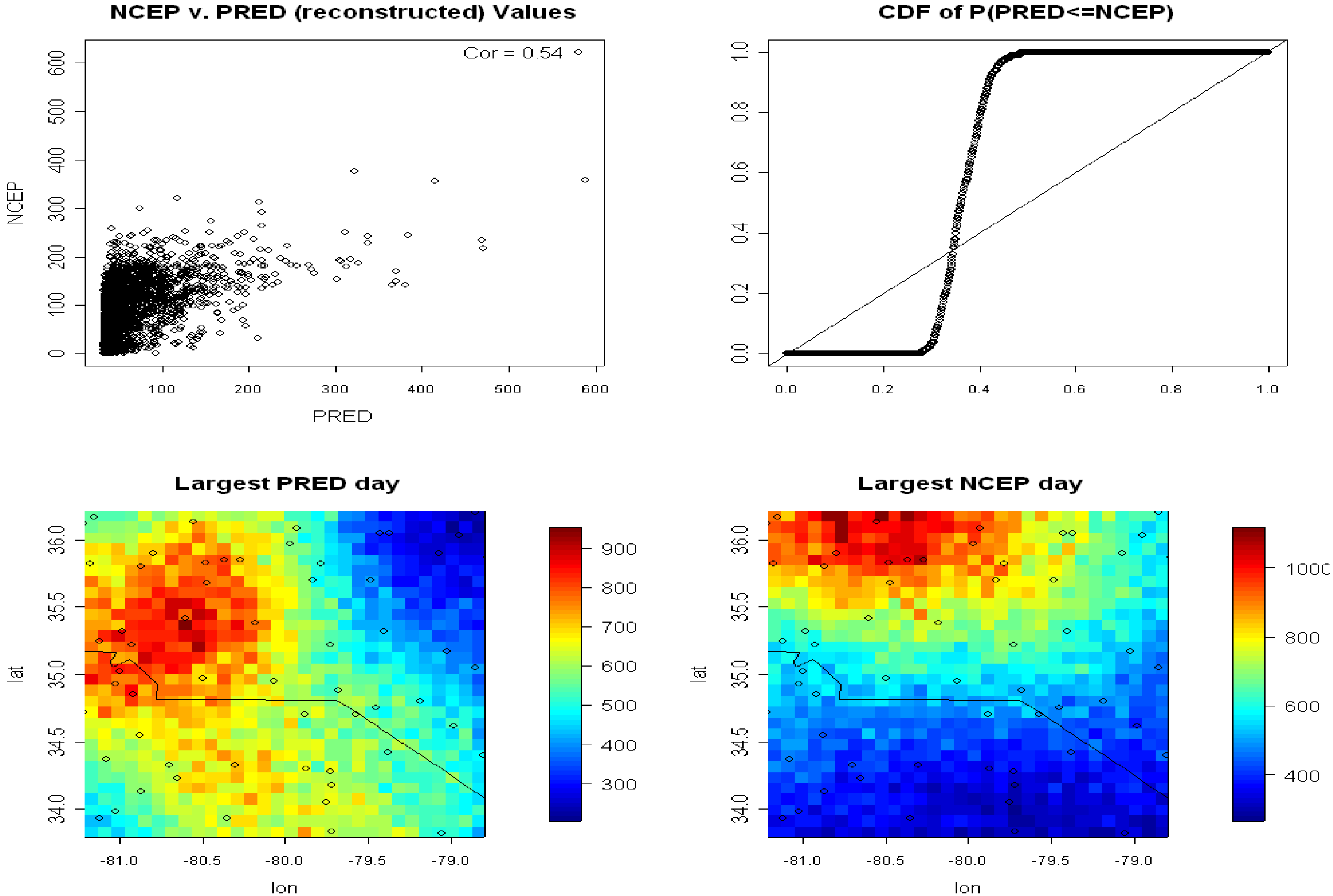
Results for Grid Cell 185 (Threshold 0.5306)



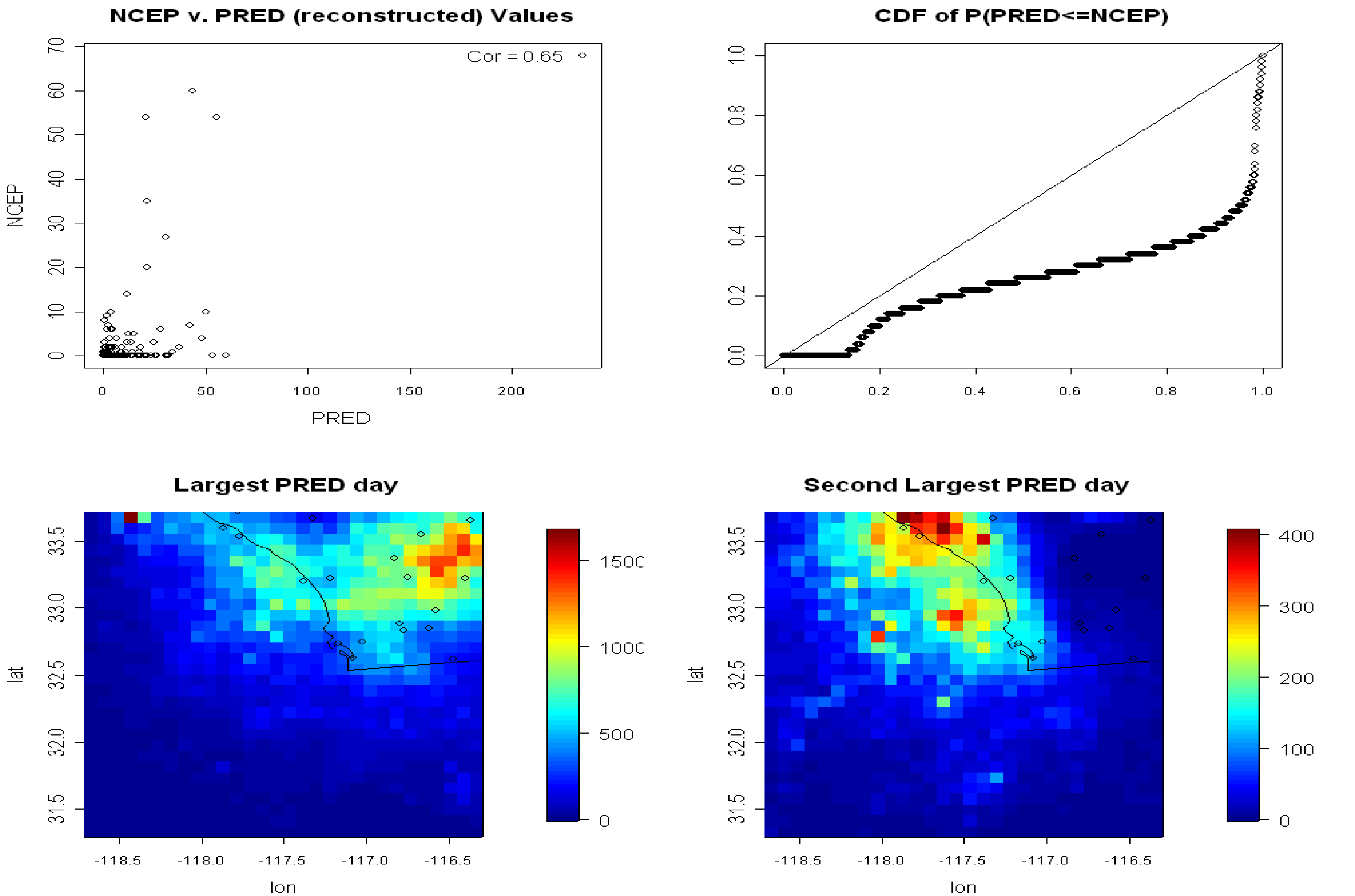
Results for Grid Cell 185 (Threshold 1.645)



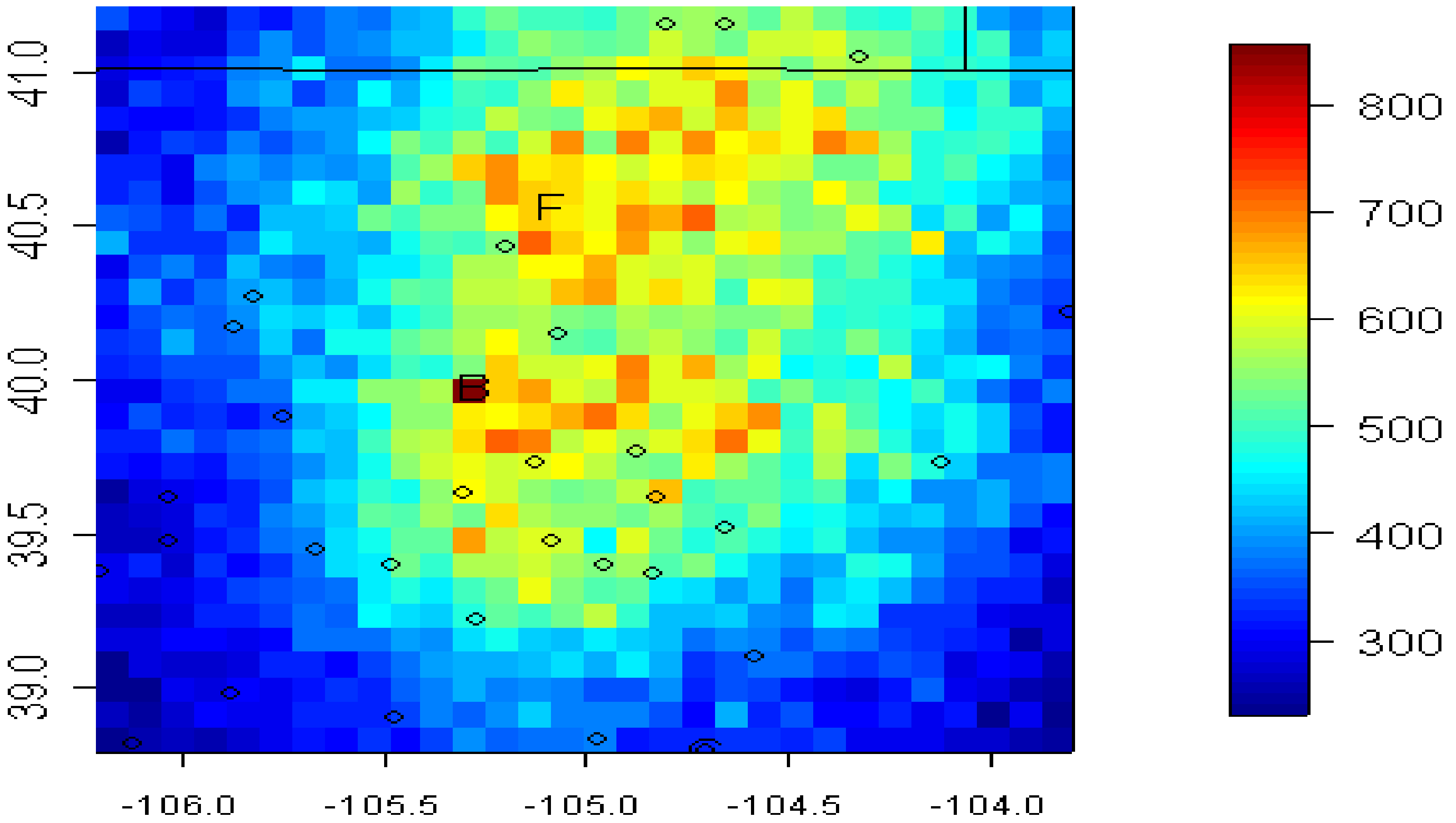
Results for Grid Cell 104



Results for Grid Cell 222



Results for Grid Cell 41



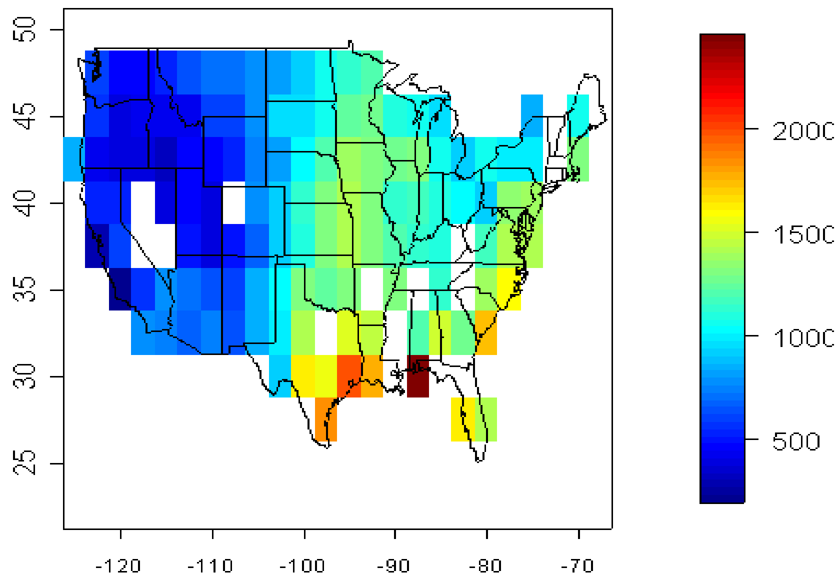
50-Year Return Values for Cell 104
(B=Boulder; F=Fort Collins)

Computing grid cell return values

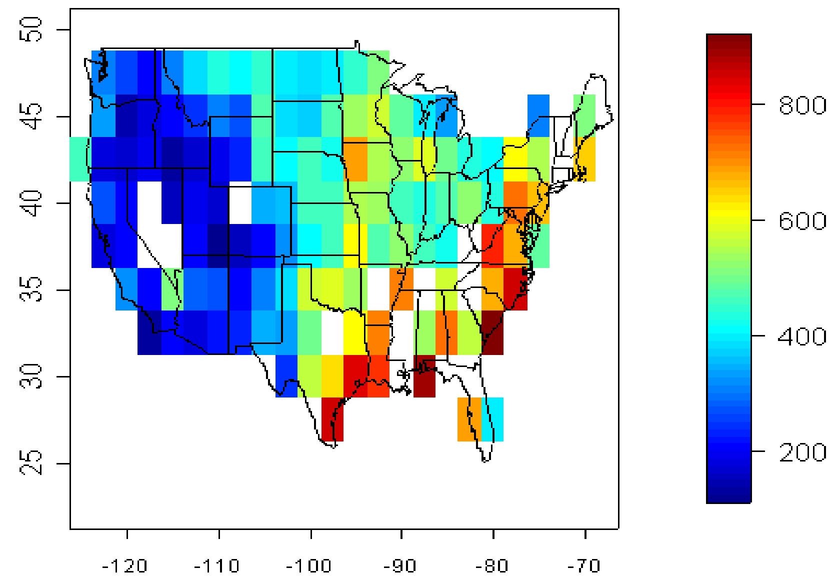
For OBSV: use same threshold as used in constructing the spatial analysis (in most cases, this was set at 97.5th percentile)

For PRED and NCEP: calculate RV50 (with delta method SE) using 95th, 96th, 97th, 98th, 99th percentiles as thresholds. Use the estimate for the lowest threshold that's consistent with every estimate above it, as judged by the SE.

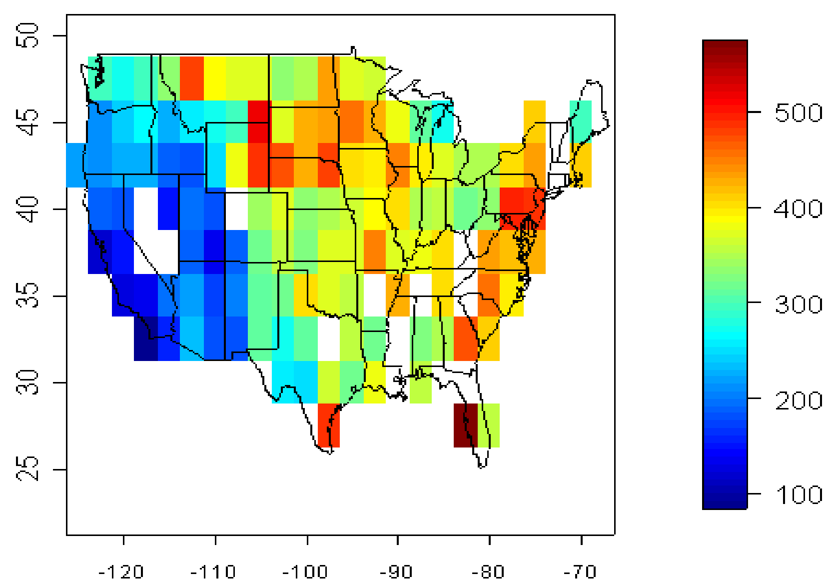
OBSV 50-Year Return Values

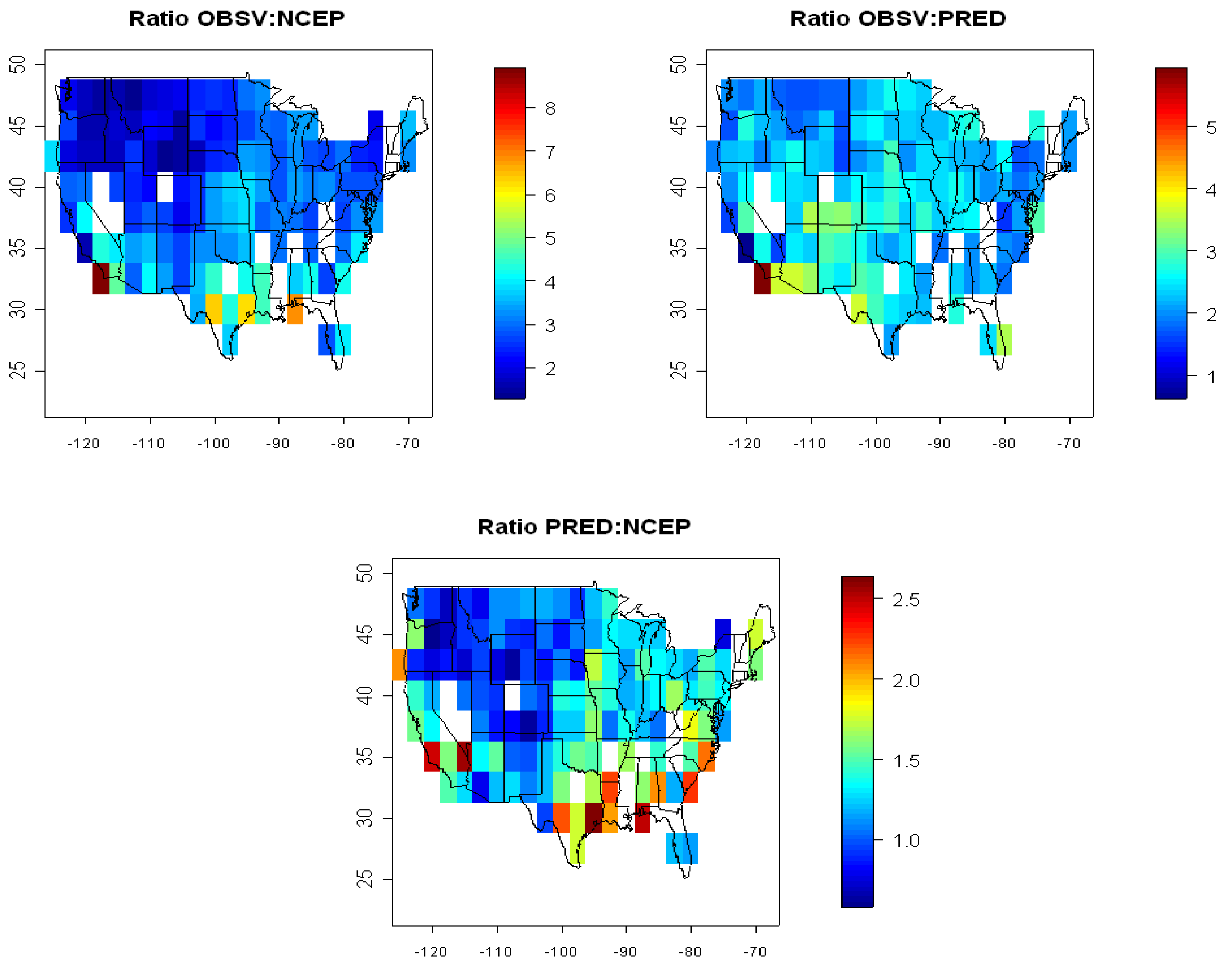


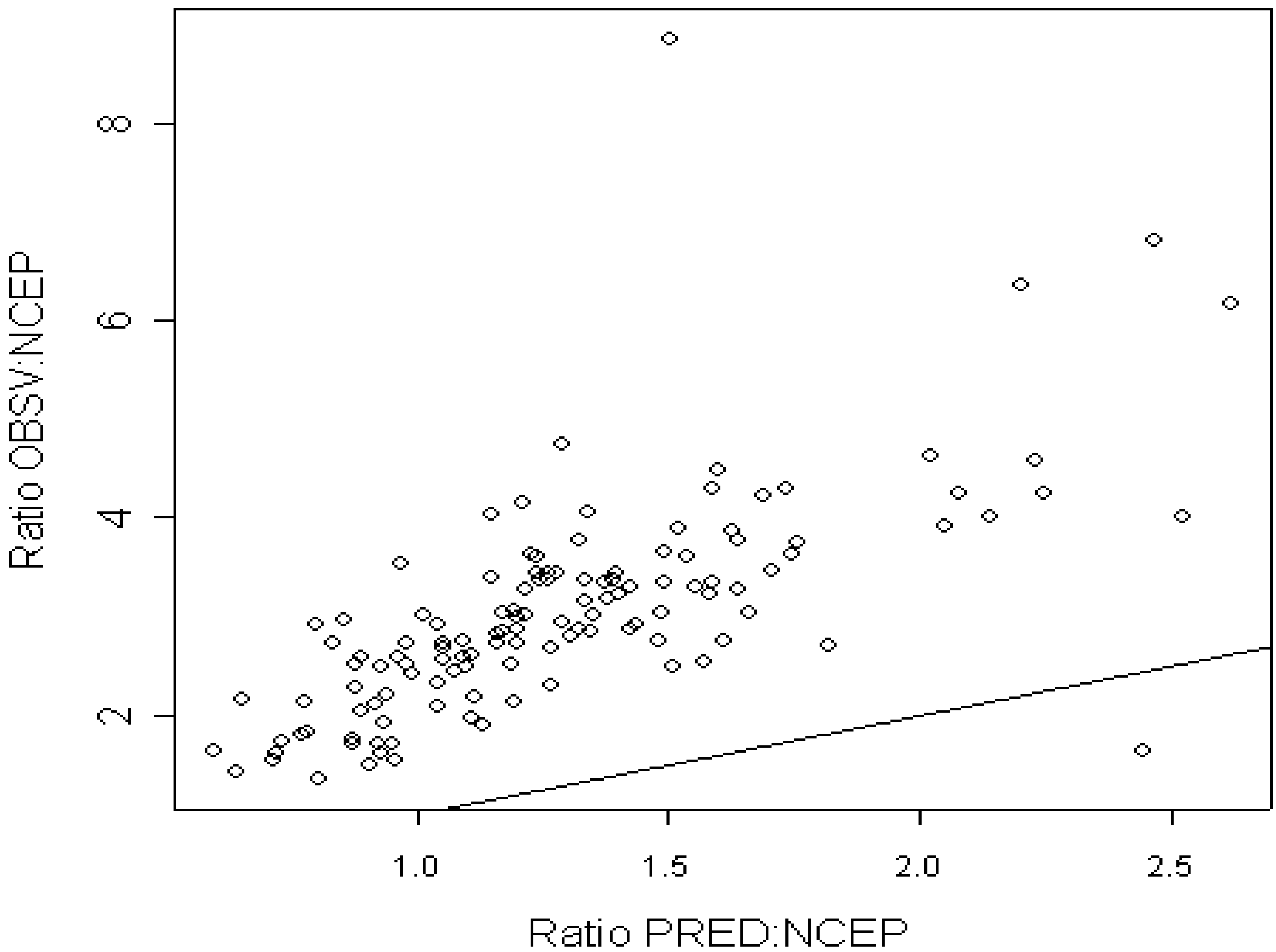
PRED 50-Year Return Values



NCEP 50-Year Return Values




Ratio OBSV:NCEP
Ratio OBSV:PRED
Ratio PRED:NCEP
50
45
40
35
30
25
-120
-110
-100
-90
-80
-70
8
7
6
5
4
3
2
1
2.5
2.0
1.5
1.0


8
6
4
2
1.0
1.5
2.0
2.5
Ratio OBSV:NCEP
Ratio PRED:NCEP

VI. SUMMARY AND CONCLUSIONS

- Return values computed from PRED are much closer to those computed from NCEP than the original estimates computed directly from the observational data

- However, clear discrepancies remain. In most cases, the return value computed from NCEP still underestimates that from PRED.

Future Work: Statistics

- Incorporate covariates and/or spatial dependence into parameters of marginal distributions

- Consider alternative spatial dependence models or extend to spatial-temporal processes

- Alternatives to Gaussian processes, such as mixtures?

Future Work: Climatology

The results show clearly that there is a smoothing effect — return values based on grid-cell averages are smaller than those based on individual observation stations.

However, NCEP seems to be taking this smoothing effect too far.

Could a more realistic spatial statistics representation of sub-grid-cell processes lead to improved parametrizations in climate and weather-forecasting models?

VII. REFERENCES

Coles, S.G. and Tawn, J.A. (1996), Modelling extremes of the areal rainfall process. *J.R. Statist. Soc. B* **58**, 329–347.

Donnelly, T.G. (1973), Algorithm 462. Bivariate normal distribution. *Commun. Assoc. Comput. Mach.* **16**, 638.

Owen, D.B. (1956), Tables for computing bivariate normal probabilities. *Ann. Math. Statist.* **27**, 1075–1090.

Raab, M. (1998), Compound Poisson approximation of the number of exceedances in Gaussian sequences. *Extremes* **1**, 295–321.

Roos, M. (1994), Stein's method for compound Poisson approximation: The local approach. *Annals of Applied Probability* **4**, 1177–1187.

Sansó, B. and Guenni, L. (2004), A Bayesian approach to compare observed rainfall data to deterministic simulations. *Environmetrics* **15**, 597–612.

Sansó, B. and Guenni, L. (2000), A non-stationary multisite model for rainfall. *J. Am. Statist. Assoc* **95**, 1064–1089.

Schervish, M.J. (1984), Multivariate normal probabilities with error bound. *Applied Statistics* **33**, 81–94.

Stein, M.L., Chi, Z. and Welty, L.J. (2004), Approximating likelihoods for large spatial data sets. *J.R. Statist.Soc. B* **66**, 275–296.

Vecchia, A. V. (1988), Estimation and identification for continuous spatial processes. *J. Roy. Statist B* **50** 297–312.

Young, J.C. and Minder, C.E. (1974), Algorithm AS76: An integral useful in calculating non-central t and bivariate normal probabilities. *Applied Statistics* **23**, 455–457.