A stochastic space-time model for annual precipitation extremes

Audrey Fu, Nhu D Le, Jim Zidek

U Washington, BC Cancer Agency, U British Columbia

Acknowledgements

- Francis Zwiers, Environment Canada

Outline

- **General perspectives**
- **Coupled Global Climate Model (CGCM)**
- **Precipitation extremes**
- **Review** extreme value theory-**shortcomings**
- **Alternate approach:** basic theory
- **Application:** Coupled Global Climate Model
- **Conclusions**

What's "extreme"?

For dams, hydro electric, water storage or flood control:
1000 year return period

NOTE: Meaning

$$P[\text{Dam failure in a given year}] = 1/1000$$

What's “extreme”?

Highway bridges:

100 year return period

NOTE: Not too extreme - 99th percentile

What's "extreme"?

EPA regulations for particulate pollution ($PM_{2.5}$):

- At each monitoring site, compute daily concentration averages
- Compute 98th percentile of these
- Compute T = 3 year average of these
- Requirement: $T \leq 65$ μg m^{-3} at each site

Physical vs Statistical Modelling

THEME 1: Statistics can help assess physical (simulation) models (if you must)

- The US EPA says you must!!
- Fuentes, Guttorp, Challenor (2003). NRCSE TR # 076.

Physical vs Statistical Modelling

- **THEME 2**: Physical and statistical models can produce synergistic benefits by ”melding” them.
  - Wikle, Milliff, Nychka, Berliner (2001). JASA.
  - Example: how can simulated (modelled) and real rainfall data be usefully combined?

Physical vs Statistical Modelling

THEME 3: Statistics can help interpret, analyze, understand, exploit outputs of complex physical models.

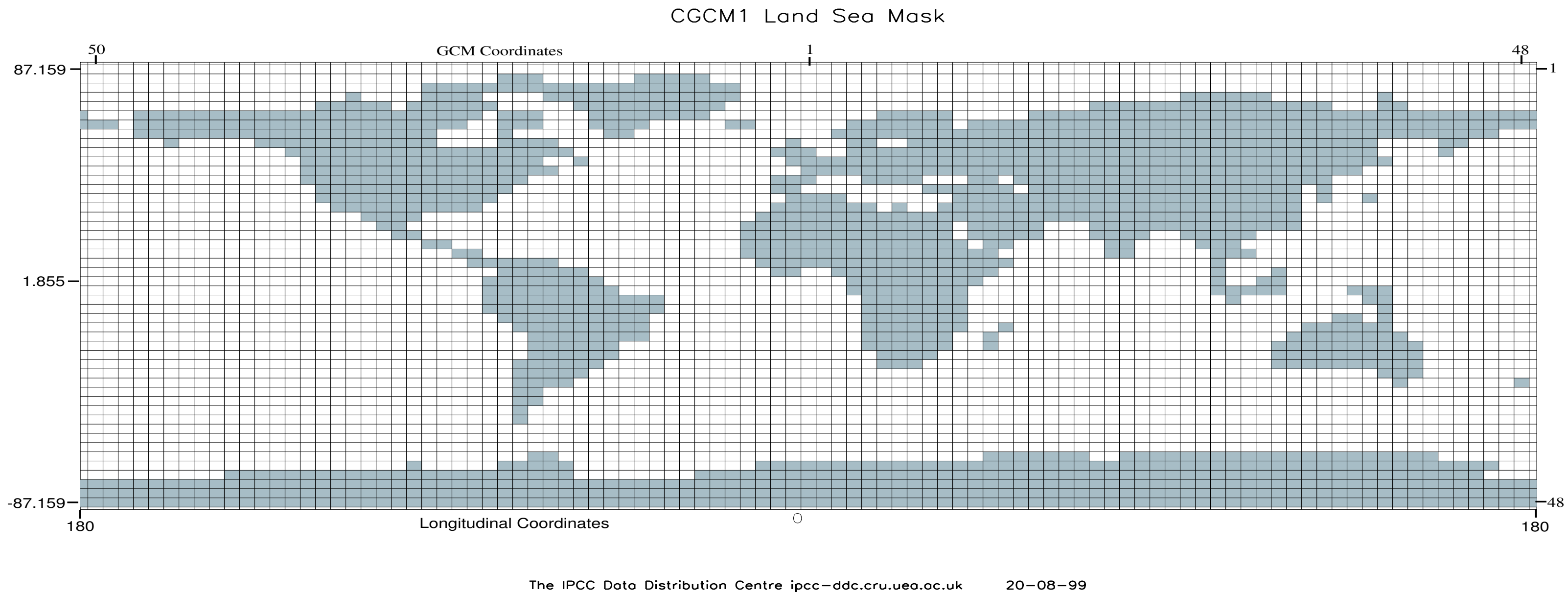
- Nychka (2003). Workshop presentation
- Example: statistical analysis of CGCM precipitation (precip) extremes gives coherent return values over space for design

Coupled Global Coupled Model

- ocean and atmosphere models run separately
  - over centuries
  - then coupled thru 14 yr “integration” periods
- output forced by input of greenhouse gas scenarios
  - eg as observed up to 1990 and 1% per yr increase in CO_2 to 2100

Coupled Global Coupled Model

- ocean and atmosphere models run separately
  - over centuries
  - then coupled thru 14 yr “integration” periods
- output forced by input of greenhouse gas scenarios
  - eg as observed up to 1990 and 1% per yr increase in CO_2 to 2100

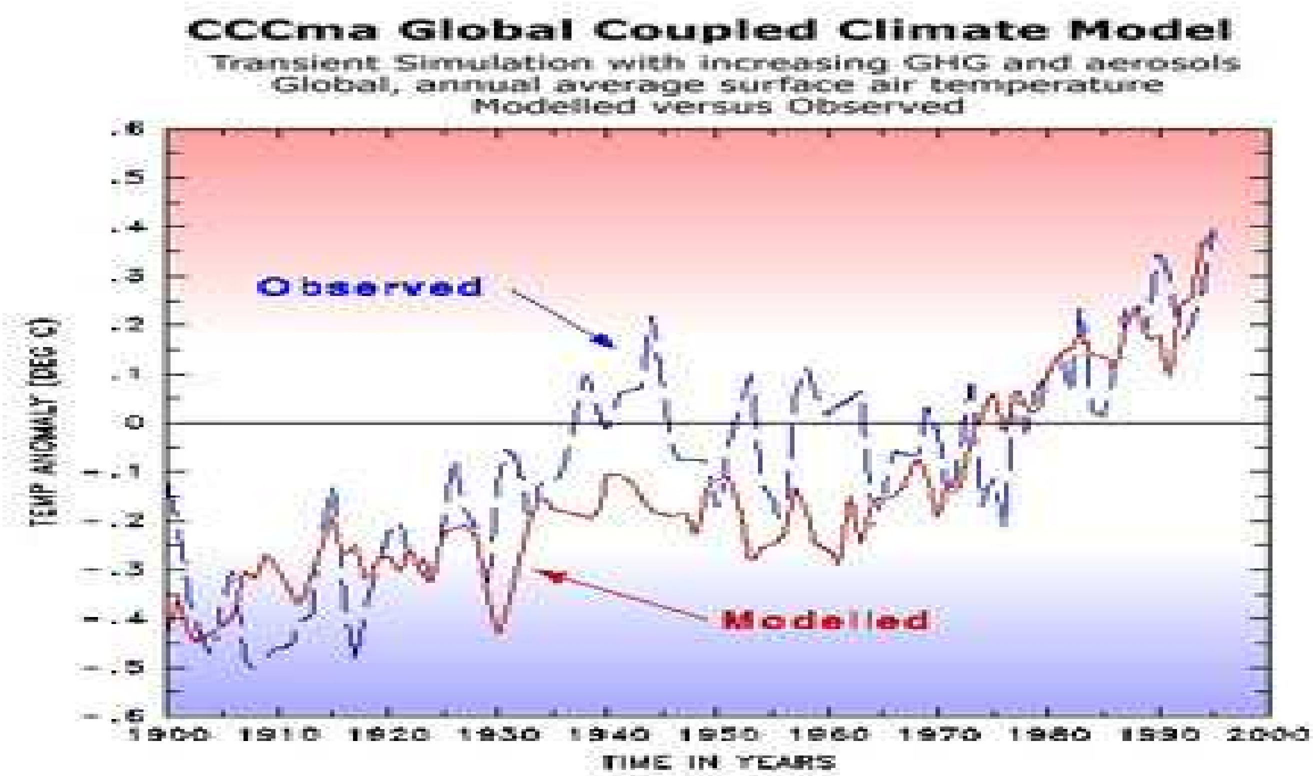


Coupled Global Coupled Model

- ocean and atmosphere models run separately
  - over centuries
  - then coupled thru 14 yr “integration” periods
- output forced by input of greenhouse gas scenarios
  - eg as observed up to 1990 and 1% per yr increase in CO_2 to 2100
- precipitation & latent heat released when local rel humidity hi enough
  - liquid water falls to the surface as precipitation

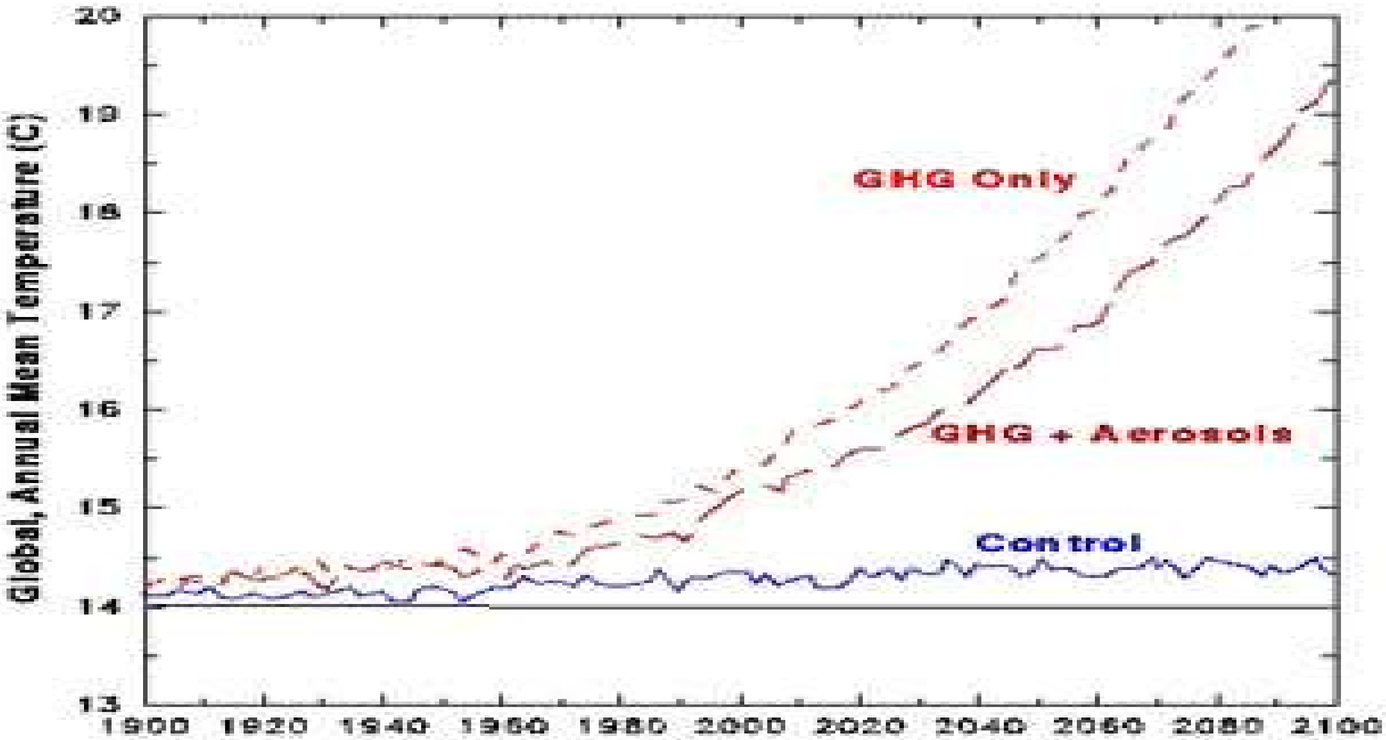
Coupled Global Coupled Model

“Confirmation” Run: modelled & observed global annual average surface temperature, 1900 - 1990. Scenario: like that above.



Coupled Global Coupled Model

Looking ahead under various scenarios



Precipitation extremes

EG: The 100 year rain!

- **return values** for annual max precipitation levels important - but Canada little monitored
- **solution**: simulate precipitation extreme fields using CGCM: 312 Canadian grid cells.
- **Required:**
  - spatially coherent cell return values!
  - joint 312 dimensional distribution to
    - enable prediction of T = number of 312 return value exceedances with E(T), SD(T), etc

CGMC Data

- 3 independent simulation runs of hourly precipitation (mm/day)
  - in 21-year windows (to look for trends)
  - 1975-1995 2040-2060 2080-2100
- 26×12 grid covers Canada, cell size = $(3.75^\circ)^2$
- gives $21 \times 3 = 63$ annual precipitation maxima per cell $\times$ time window

Modelling extreme fields

E.G.: annual precip maxima. Assume no autocorrelation

- Approach 1: multivariate extreme value theory
- Approach 2: use hierarchical Bayes

Single cell (site)

Assume $X_1, X_2, \cdots, X_n$ iid. Let $M_n = max\{X_1, X_2, \cdots, X_n\}$.
Fisher-Tippett (1928)showed:

$$P(\frac{M_n - b_n}{a_n} \leq x) \quad \rightarrow \quad H(x), \quad \text{as n} \rightarrow \infty$$

where H has **GEV** distribution

$$H(x) = \begin{cases} \exp\left[-\left\{1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right\}^{-1/\xi}\right], & 1 + \xi\left(\frac{x-\mu}{\sigma}\right) > 0, \xi \neq 0 \\ \exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right)\right] & \xi = 0 \end{cases}.$$

Single cell (site)

Alternatives: Generalized Pareto (GPD) model:
Cumulative distribution's right hand tail approximated by:

$$H(x) = 1 - \lambda \left[1 + \xi \left(\frac{x-u}{\sigma} \right) \right]_{+}^{-1/\xi}, \quad x > u$$

for parameters, $\lambda > 0,\ \sigma > 0,\ \text{and}\ \xi \in (-\infty, \infty)$.

NOTE: Assuming number of exceedances over time Poisson yields the Poisson–GPM method

Single cell (site)

Alternatives:
Peak over Threshold (POT) model:
Only values above a “threshold” are modelled:

$$P[X \leq x + u \mid X > u]$$

approximated the GPD model.

- Good idea since only extremes of interest
- However:
  - for large threshold “u” little data
  - for small “u” poor tail approximation
- tail model parameters depend on “u”
- results sensitive to the choice of “u”

Single cell (site)

Alternatives:

Probability Weighted Moment (PWM) model: unduly complex for certain purposes

Multiple cells

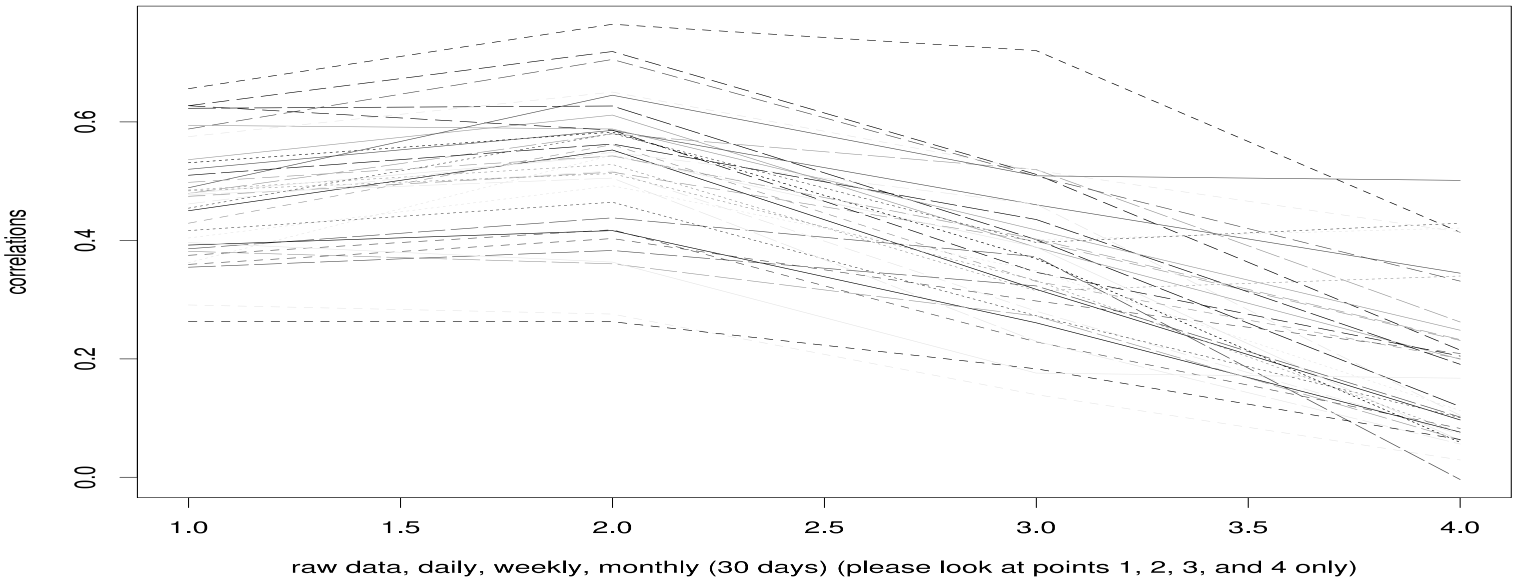
Approaches:

Extend Fisher-Tippett

Problems: leads to a big class of possible limit distributions. Moreover, extremes must be asymptotically dependent for large return periods.

Multiple cells

An example: Small inter - site correlations *Inter- site dependence declines with increasing extreme's"range" for many (not all!) site pairs [London and Vancouver analyses]*



Inter-site correlations for Vancouver's PM_{10} decline with max range.

Multiple cells

Approaches:

Specify individual cell (parametric) distributions. Put a joint distribution over their parameters

Problems: *ah hoc* and complex. No compelling dependence structure

Multiple cells

Point process (PP) approach:
Space–time points where threshold exceedance occurs is non-homogeneous Poisson process, intensity function:

$$\Lambda(A) = (t_2 - t_1)\Psi(y; \mu, \psi, \xi)$$

where

$$A = (t_1, t_2) \times (y, \infty)$$

$$\Psi(y; \mu, \psi, \xi) = \left[1 + \xi\left(\frac{y-\mu}{\psi}\right)\right]^{-1/\xi}, \; 1 + \xi\left(\frac{y-\mu}{\psi}\right) > 0$$

Problems: Complicated distribution; unclear how to extend to multivariate responses; what about fixed site monitors?

Multiple cells

Approaches:

Our hierarchical Bayesian method:

- Approximate transformed cell max precip data by joint multivariate t distribution
- Very flexible & lots of available theory

Problems:

- may not work for "extreme extremes".
- asymptotically independent sites

Our method: Assumptions

$\mathbf{Y}_j : p \times 1$ annual precipitation maxima; p cells, years $j = 1, \ldots, n$.

Sampling Distribution:

Conditional on $(\boldsymbol{\mu}, \Sigma)$ $\mathbf{X}_j \doteq \log \mathbf{Y}_j \overset{iid}{\sim} MVN_p(\boldsymbol{\mu}, \Sigma)$

Our method: Assumptions

$\mathbf{Y}_j : p \times 1$ annual precipitation maxima; p cells, years $j = 1, \ldots, n$.

Prior distribution:

- $\boldsymbol{\mu}|\Sigma \sim MVN(\boldsymbol{\nu}, F^{-1}\Sigma)$
- $\Sigma \sim W^{-1}(\Sigma|\Psi, m)$
- Ψ and m called *hyperparameters*
- F^{-1} re-scales Σ to mean's level of uncertainty

Implications

Posterior distribution (of precip max field):

- $\mathbf{X}_{p\times 1}|D \sim t\left(\tilde{\mathbf{x}}, \tilde{\Sigma}, l\right)$ with
- $\tilde{\mathbf{x}} = \boldsymbol{\nu} + (\bar{\mathbf{x}} - \boldsymbol{\nu})\hat{E}$
- $\tilde{\Sigma} = \frac{1+nF^{-1}-n\hat{E}F^{-1}}{l}\hat{\Psi}$ and
- $l = m + n - p + 1$,
  $\hat{\Psi} = \Psi + (n-1)S + (\bar{\mathbf{x}} - \boldsymbol{\nu})(n^{-1} + F^{-1})^{-1}(\bar{\mathbf{x}} - \boldsymbol{\nu})'$

The Hyperparameters

- ν: estimated by **smoothing spline** over all cells
- $\Psi = c \times \Phi$, Φ a **covariance matrix** estimated by *semivariogram*
- $\Phi_{ij} = Cov(X_i, X_j) = \sigma^2 - \gamma(h_{ij})$
  - σ^2 = common sample variance
  - h_{ij} = Euclidean distance between sites i, j
  - $\gamma(h)$ = isotropic semivariogram model fitted to data
- **EM algorithm** estimates c & degrees of freedom, m
- F^{-1} estimated by method of moments

The Hyperparameters

Justifying empirical Bayes:

- Posterior must be well-calibrated w.r.t. real max precip fields. Hence prior must be *fitted* to enable good match
- simplicity
- equates with using diffuse prior

Diagnostic tool

Validating **joint normality** assumption

Method
For any given year:

- delete data from selected sites
- predict them by $\hat{\mathbf{x}}_{\mathbf{u}} = \boldsymbol{\nu}_u + (\hat{\mathbf{x}}_g - \boldsymbol{\nu}_g)' \Psi_{gg}^{-1} \Psi_{gu}$ where
  - u means *ungauged* (missing) and g, *gauged*
  - ($\boldsymbol{\nu}_u$,$\boldsymbol{\nu}_g$) partitioned *prior mean* conformably
  - likewise for Ψ_{gg} and Ψ_{gu}

Diagnostic tool

Validating **joint normality** assumption

Method

Now see if vector of missing values falls into 95% credibility interval:

- $\{\mathbf{X_u} : (\mathbf{X_u} - \hat{\mathbf{x}}_\mathbf{u})' \Psi_{u|g}^{-1} (\mathbf{X_u} - \hat{\mathbf{x}}_\mathbf{u}) < b\}$ where
  - $b = (u \times P_{u|g} \times F_{1-\alpha,u,m-u+1}) \times (m - u + 1)^{-1}$
  - $\Psi_{u|g} = \Psi_{uu} - \Psi_{ug}\Psi_{gg}^{-1}\Psi_{gu}$ and
  - $P_{u|g} = 1 + F^{-1} + (\mathbf{x_g} - \boldsymbol{\nu}_g)'\Psi_{gg}^{-1}(\mathbf{x_g} - \boldsymbol{\nu}_g)$

Diagnostic tool

Validating **joint normality** assumption

Method

- do this repeatedly. Randomly remove subsets of sites of fixed size
- compute the relative coverage frequency - it should be around 95%!! Offers check on the validity of the model.

Estimating Return Values

- Definition: T year return value, X_T

  $P(X > X_T) = \frac{1}{T}$
- can be estimated for each cell from the joint posterior t distribution
- approximation: use log normal instead of log t distribution, $x_{1-T,i} = \tilde{x}_i + \Phi(1-T) \times \tilde{\sigma}_i$

CGCM Analysis

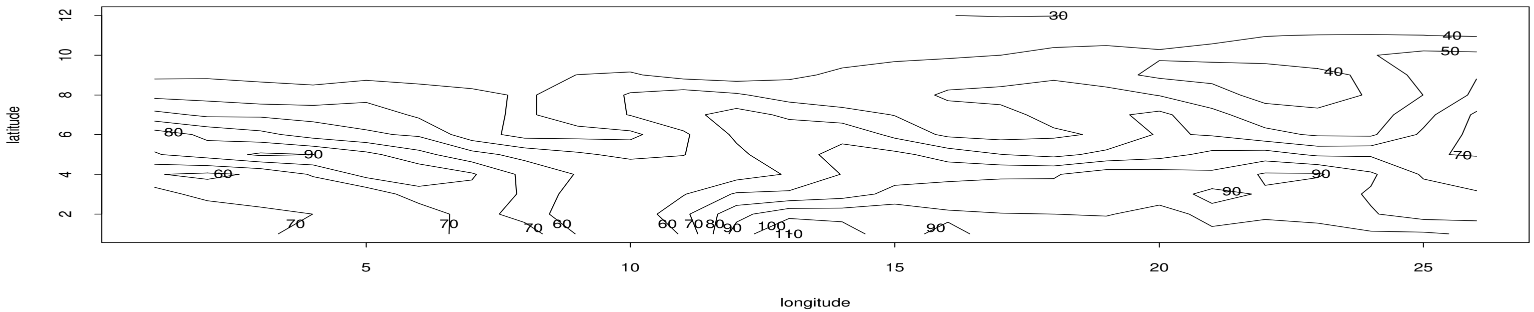
- BEFORE ANALYSIS: log - transform, de-trend
- RESIDUALS:
  - symmetric empirical marginal distribution
  - slightly heavier than normal tails
  - no significant autocorrelation
- AFTER ANALYSIS: re-trend, antilog-transform

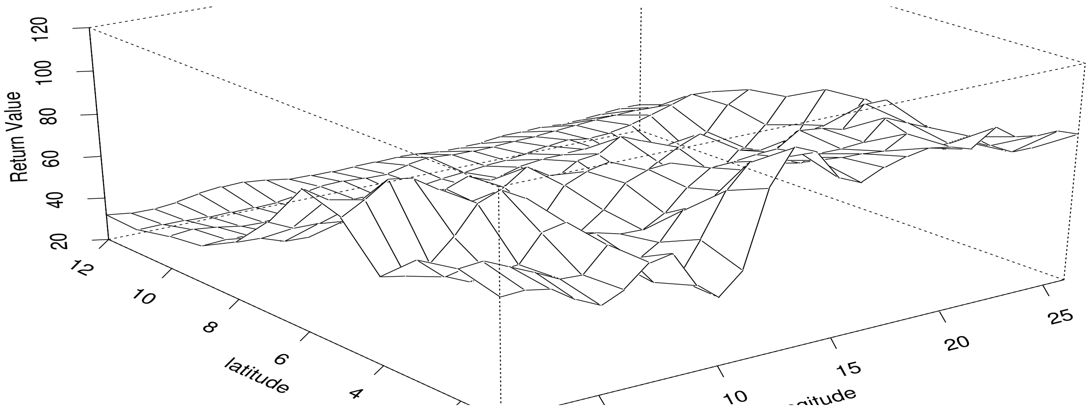
CGCM Statistical Model

- **HIERARCHICAL BAYES**: Normal - Inverted Wishart model for residuals
  - estimated variogram for 312*312 dimensional hypercovariance
- **RESULTING POSTERIOR**: 312 dimens'l, multivariate t-distribution, $m = 355$, $c = 49$ (for Ψ estimate)
  Posterior:
  - yields estimates of 312 marginal return values
  - enables simulation of 312 dim'l annual max precip field and
    - distribution of *stat's* computed from it
      - EG: T = # of (312) cells above return values, $E(T)$, predictive interval for T

Results

Contour, perspective plots of estimated 10-year return values (mm/day).





CGGM Stats Model Assessment

CROSS VALIDATION:

- randomly omit 30 of 312 cells repeatedly
- predict their values from rest from the joint t distribution.
- CONCLUSION: The joint t distribution fits the simulated data quite well

Credibility Level	Mean	Median
30%	35	35
95%	96	97
99.9%	99.9	99.9

Table 1: SUMMARY: cred'y ellips'd coverage probs

Discussion

- Joint distribution fits well. Also worked in another study on real air pollution data.
- Allows answers to complex question like chances of say 10 simultaneous exceedances of cell return values
- Suggests model could be used to design extreme precip monitoring networks
- the return value index reveals reasonable joint fit but needs further study

Concluding Remarks

- multivariate t distribution promising model for extreme space-time fields. Data needs to be transformable. But wealth of existing theory for multinormal makes pursuit worth it!
- empirical checking/diagnostics vital before using the method
- general theory allows extension to multivariate responses in each site & covariates too!
- Can posterior be trusted for extreme-extremes? Can any distribution?

References & Contact Info

- email: jim@stat.ubc.ca
- internet: http://www.stat.ubc.ca/<LINK Faculty Members>
- tech reports: http://www.stat.ubc.ca/<LINK Research Activities>
- **R-based software:** http://enviro.stat.ubc.ca
- Companion to Nhu D Le & James V Zidek (2006) *Statistics analysis of environmental space–time processes*. Springer. To Appear May 12.

NOTE: Chap 14 gives a tutorial on its use.