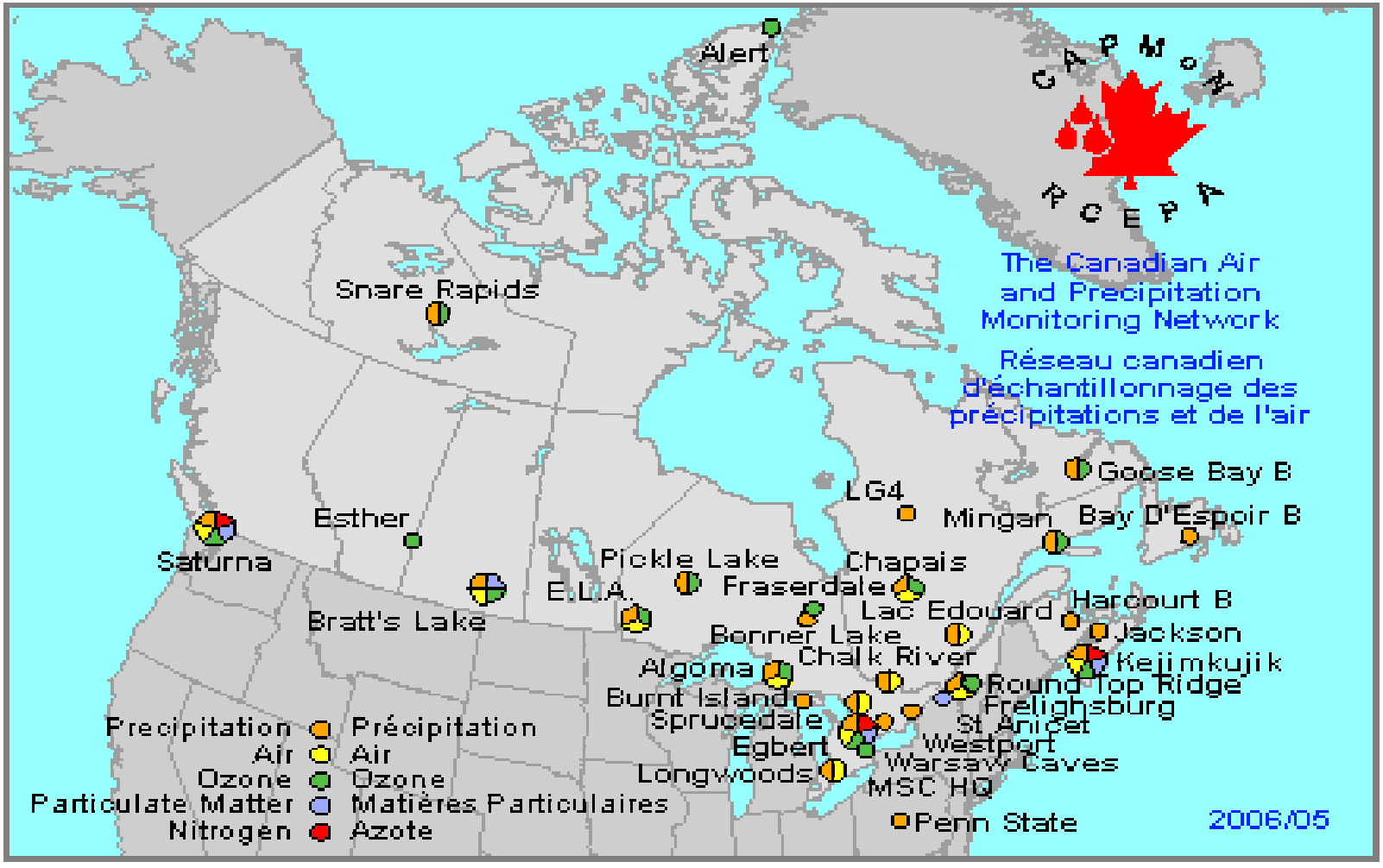
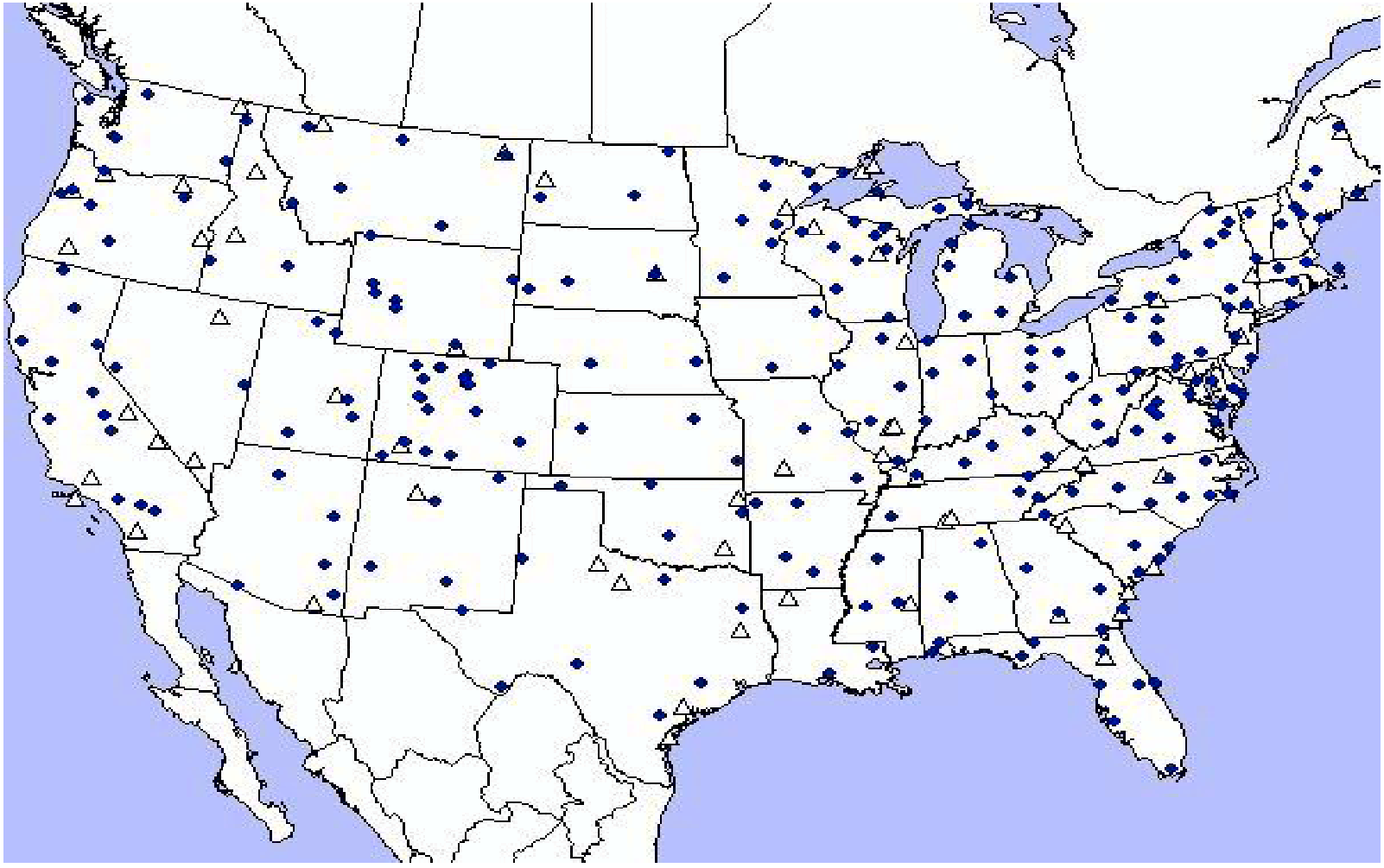
5.1 Designing networks

Example: the Capmon network. Combines various networks created for various purposes



The NADP/NTN network

Monitors multivariate responses related to “acid precipitation”



Why monitor?

General objectives: To:

- measure process responses at critical points:
  - Near a new smelter using arsenic
- enable predictions of unmeasured response
- enable forecasts of future responses
- provide process parameter estimates
  - physical model parameters
  - stochastic model parameters eg. covariance parameters
- address societal concerns

Why monitor? (cont'd)

Specific objectives: To:

- detect non-compliance with regulatory standards
- enable health effect assessments to be made
  - & provide good estimates of relative risk
  - determine how well sensitive sub-populations are protected
  - can include all life, not just human
- to assess temporal trends
  - are things getting worse?
  - is climate changing?

Overall

- to reduce uncertainty about some aspect of the world
  - one form of uncertainty (***aleatory***) cannot be reduced (outcome of fair die toss
  - the other (***epistemic***) (whether the die is fair) will be reduced, implying that the optimum design must be regularly revisited

What's uncertainty

- Laplace: "Probability is the language of uncertainty"
- DeFinetti: "In life uncertainty is everything"
- Statisticians: "variance" or "standard error"
- Kolmorov & Renyi: "Entropy"

Question

Study question 5.1 Suppose $X \sim N(0,1)$. Prove that uncertainty about X, i.e. Var(X||X|<C) is increasing as a function of C.

Research question 5.1 Suppose $X \sim N(\eta,1)$. Prove that uncertainty about X, i.e. Var(X||X|<C) is increasing as a function of C. Warning: quite a well-known, unsolved problem!! :-)

Possible design criteria

"Gauge" (add monitors to) sites that

- that maximally reduce uncertainty at their space-time points
  - measuring their responses eliminates their uncertainty
- best minimize uncertainty about their cousin's responses
- best inform about process parameters
- that best catch the non-compliers

Special problem: extreme values

Regulatory criteria metrics (risk) usually involves extremes.
Example: EPA'S PM_{10} criterion:
For particles of diameters of 10 micrometers or less:

Annual Arithmetic Mean: 50 μg m^{-3}

24 - hour Average: 150^{FN} μg m^{-3}

The three year average of 98-th annual percentiles of 24 hour averages must be $\leq$ 150 ($\mu g\ m^{-3}$) at all sites in an urban area. Complex metric $\Rightarrow$ need predictive distribute to simulate its distribution!

The bad news

1. **Insufficient data, spatial and temporal.**
2. **Extremes have small inter - site dependence**
   - **between some site pairs, not others**
3. **Conventional approaches fail**
4. **Multivariate extreme value distributions - not tractable**
   - **conditional computation (e.g. entropy) difficult**
   - **simulating extreme fields hard**
5. **Elusive design objective**

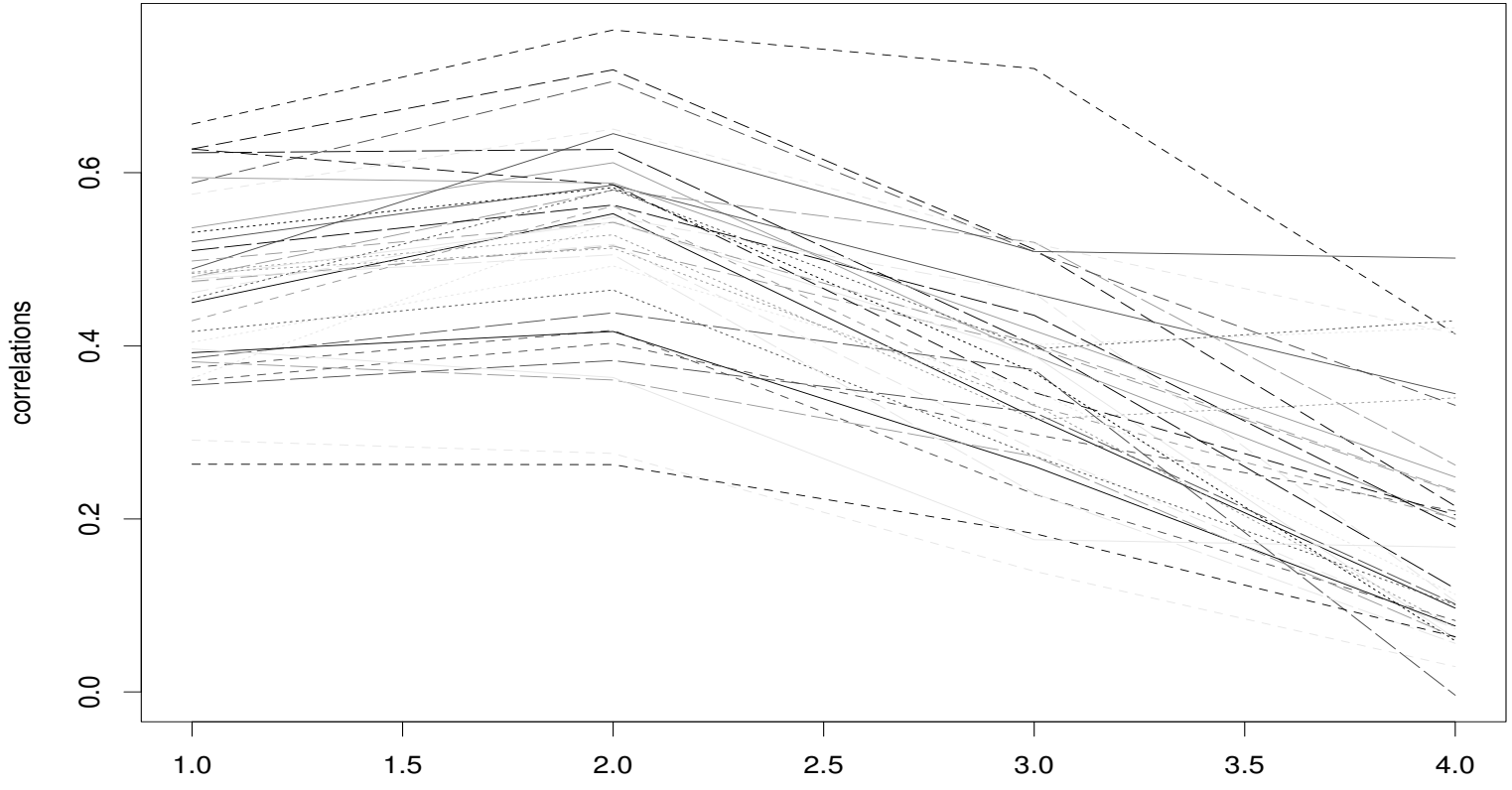
The good news

Joint distribution of extremes approximately a log multivariate t distribution. Hence can:

1. **have convenient conditional, marginal distributions**
2. **accommodate existing sites and historical data**
3. **permit simulation of complex metric distributions**
4. **have explicitly computable entropy's, regression models, etc**
5. **can enable "elusive objectives issue" to be bypassed**

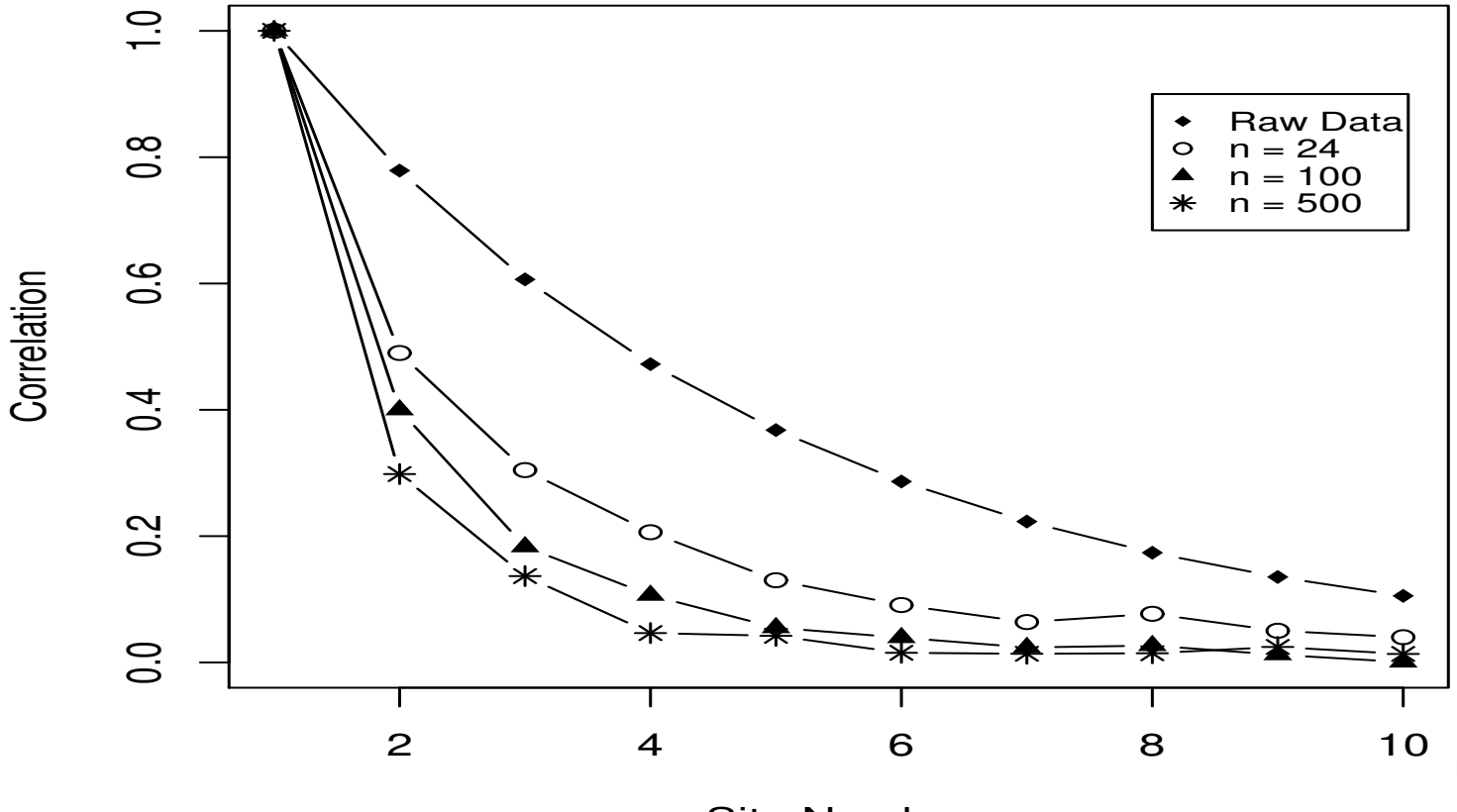
Small inter-site correlations

Inter- site dependence declines with increases in extreme's"range" for many, not all site pairs [London and Vancouver analyses]. Figure shows Vancouver's for PM_{10} decline with max range.



Inter-sites correlations (cont'd)

Simulation study: multinormal responses; maxima with varying ranges at 10 sites. Multivariate t results show smaller loss of dependence.
Inter-site correlations for maxima for simulated fields of extremes. Big n = light tails.



Designer challenges

- multiplicity of valid objectives
- unforeseen & changing objectives
- multiple responses at each site: which to monitor?
- must include prior knowledge & prior uncertainty
- should use realistic process models. (How?)
- must integrate with existing networks
- must deal with reality!!!

Questions

Study Question 5.2 Try the simulation experiment yourself and confirm that for heavier tails lead to increased intersite correlations

Discussion question 5.1 How might design criteria be arrived at in practice? Who is responsible for setting them.

Research question 5.2 Monitor placement should recognize such things as the geographical distribution of impacted populations (eg trees or fish). How can an optimal design be determined in such a context.

Research question 5.3 Develop a design theory in a non-Gaussian context.

Approaches to design

Space-filling designs

Probability based designs

- simple random sampling
- stratified, multistage designs
- e.g. (1) EPA's survey of lakes; (2) the EMAP project

Model Based

- Regression model approach
  - eg to estimate the slope put 1/2 the data at each end of the data range
- Random fields (prediction, e.g. entropy) approach

Other. In particular Zhengyuan Zhu (UNC) incorporates both of the latter, prediction and parameter estimation.

Entropy based approach

"Gauges" sites with greatest "uncertainty"

- **uncertainty = entropy**
- **maximally reduces uncertainty about "ungauged" sites**
- **best estimates predictive posterior distribution under entropy utility**

By - passes specification of objectives

Long history, currently popular

- **General: Good 1952, Lindley 1956, Shewry&Wynn 1987**
- **Network design: Caselton and Zidek 1984, Sebastiani&Wynn 2000, Zidek&Sun&Le 2000**

What's entropy?

Let $p = P(E)$= probability an uncertain event E occurs (heads on possibly bent coin). That uncertainty is reduced to none when outcome known, a reduction of say (for some function ϕ)
$\phi(p)$ if E occurs
$\phi(1-p)$ if not.

The expected reduction in uncertainty is
$p\phi(p) + (1-p)\phi(1-p)$
Simple assumptions imply:

$$\boxed{\phi(p) = log(p)}$$

Thus entropy reduction due to knowledge of E's occurrence (= "**uncertainty**" about E) is the **entropy** for the two point distribution $(p, 1-p)$:

$$p \log(p) + (1-p) \log(1-p)$$

Relative entropy

How much is that entropy?
Needs a reference level. Complete uncertainty about the coin, how its to be tossed and so on would point to a two point distribution $(q, 1-q)$ with $q = 1/2$. Thus the relative entropy would be
$I(p, q) = p \log (p/q) + (1-p) \log \{(1-p)/(1-q)\}$
Kullback-Leibler's measure of deviation of $(p, 1-p)$ from its reference level (that corresponds to a "state of equilibrium" in physics (thermodynamics).

Multiple events

$$I(p,q) = \sum_i p_i \log \{p_i/q_i\}$$

Continuous variables

Start with $p_i \sim f(x_i)dx_i$ & $q_i \sim g(x_i)dx_i$ as approximations. Then as $dx_i \to 0$, this entropy converges to

$$I(f,g) =) \int f \log f/g$$

.

Commonly $g \equiv 1$. However the Jacobean cancels under a "change of variables" so entropy is "intrinsic" measure of uncertainty.

Using entropy

X = vector of all-site responses at time n+1, monitored and unmonitored.

$$H(X, \theta) = H(X \mid \theta) + H(\theta)$$

where

$$H(X \mid \theta) = E[-\log(f(\tilde{X} \mid \tilde{\theta}, D)/h_1(\tilde{X})) \mid D]$$

$$H(\theta) = E[-\log(f(\tilde{\theta} \mid D)/h_2(\tilde{\theta})) \mid D].$$

Design goal

Add new sites to an existing network

- $X = (X^{(1)}, X^{(2)})$:all site responses, time n+1
- $X^{(2)}$: gauged now, time n
- $X^{(1)}$: ungauged sites, time n
- DESIGN GOAL: Partition $X^{(1)} = (X^{(rem)}, X^{(add)})$ NOW!
- $X^{(rem)}$: future ungauged sites
- $X^{(add)}$:future new network stations.

Entropy decomposition thm

U=$X^{(rem)}$;$G = (X^{(add)}, X^{(2)})$; X=[U,G]

Fundamental identity

TOT = PRED + MODEL + MEAS

where

$$\begin{aligned} PRED &= E[-\log(f(\tilde{U} \mid \tilde{G}, \tilde{\theta}, D)/h_{11}(\tilde{U})) \mid D], \\ MODEL &= E[-\log(f(\tilde{\theta} \mid \tilde{G}, D)/h_2(\tilde{\theta})) \mid D], \end{aligned}$$

and

$$MEAS = E[-\log(f(\tilde{G} \mid D)/h_{12}(\tilde{G})) \mid D].$$

Max'ing MEAS=Min'ing MODEL + PRED

Response distribution

Filter/transform responses.

$$X \mid \boldsymbol{\beta}, \Sigma \sim N(Z\boldsymbol{\beta}, I_n \otimes \Sigma)$$

$$\boldsymbol{\beta} \mid \Sigma, \boldsymbol{\beta}_0, F \sim N(\boldsymbol{\beta}_0, F^{-1} \otimes \Sigma)$$

$$\Sigma \sim GIW(\Psi, \delta)$$

Generalized IW

$$\Sigma^{[g]} \sim GIW(\Psi^{[g]}, \delta^{[g]})$$

$$\Gamma^{[u]} \sim IW(\Lambda_0 \otimes \Omega, \delta_0)$$

$$\tau^{[u]} \mid \Gamma^{[u]} \sim N\left(\tau_0 u, H_0 \otimes \Gamma^{[u]}\right)$$

$$\Gamma^{[u]} = \Sigma^{[u|g]} = \Sigma^{[u]} - \Sigma^{[ug]}(\Sigma^{[g]})^{-1}\Sigma^{[gu]};\ \tau^{[u]} = (\Sigma^{[g]})^{-1}\Sigma^{[gu]}$$

Predictive distribution

$$
\begin{aligned}
(X_{unob} \mid D, \mathcal{H}) \quad \sim \quad & \left(X^{[u]} \mid X^{[g_1^m, \ldots, g_k^m]}, D, \mathcal{H} \right) \times \\
& \prod_{j=1}^{k-1} \left(X^{[g_j^m]} \mid X^{[g_{j+1}^m, \ldots, g_k^m]}, D, \mathcal{H} \right) \\
& \times \left(X^{[g_k^m]} \mid D, \mathcal{H} \right).
\end{aligned}
$$

Predictive distribution

$$
\begin{aligned}
(X_{unob} \mid D, \mathcal{H}) \quad \sim \quad & \left(X^{[u]} \mid X^{[g_1^m, \ldots, g_k^m]}, D, \mathcal{H} \right) \times \\
& \prod_{j=1}^{k-1} \left(X^{[g_j^m]} \mid X^{[g_{j+1}^m, \ldots, g_k^m]}, D, \mathcal{H} \right) \\
& \times \left(X^{[g_k^m]} \mid D, \mathcal{H} \right).
\end{aligned}
$$

$$
\left(X^{[u]} \mid X^{[g_1^m, \ldots, g_k^m]}, D, \mathcal{H} \right) \sim
$$

$$
t_{n \times up} \left(\mu^{[u|g]}, \text{Dispersion}, \delta_0 - up + 1 \right).
$$

$$
\text{Dispersion} = (\delta_0 - up + 1)^{-1} (\Lambda_0 \otimes \Omega)
$$

‘Add’ entropy

$$H\left[X_{unob} \mid D\right] \;=\; H\left[X^{[u]} \mid X^{[g_1^m,\ldots,g_k^m]}, D\right] + \ldots$$

and

$$H\left[X^{[u]} \mid X^{[g_1^m,\ldots,g_k^m]}, D\right] \;=\; \frac{p}{2}\log|\Lambda_0| + \ldots$$

Optimal ‘add’ sites: max’ize $|\Lambda_0[add, add]|$

‘Add’ computation

- **NP-Hard:** No exact algorithms for big networks
- Inexact Methods:
  - Greedy
  - Greedy + Swap
- Exact Methods:
  - Complete enumeration
  - Branch and bound

How many sites

Compute entropy/number of sites as the number of sites varies. Eventually this reaches a max (bang for the buck)and then declines. Indicates when to stop on redesign.

Example

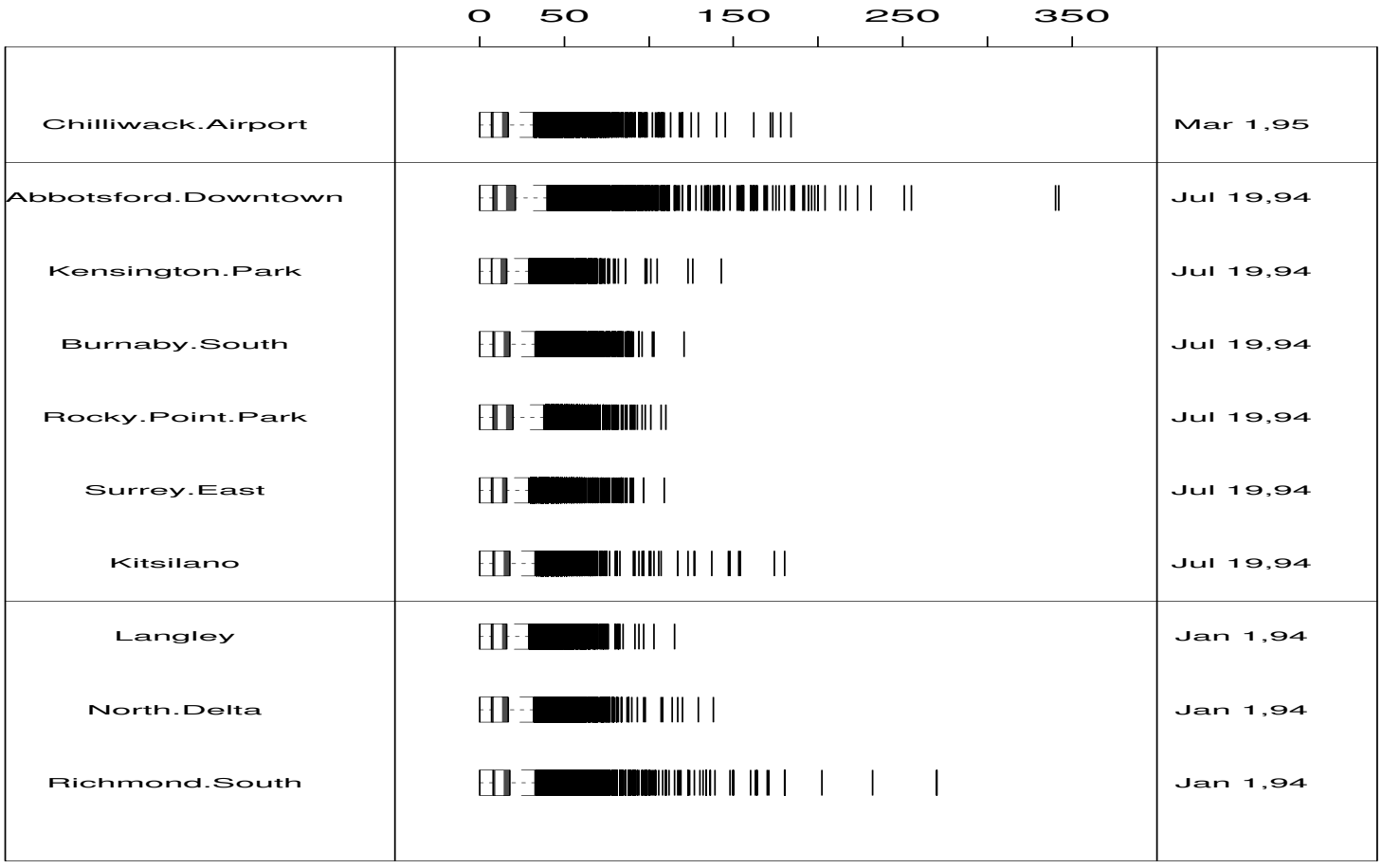
Hypothetical redesign of Vancouver's hourly PM_{10} field. Existing 10 stations $\longrightarrow$ to 16. 6 new stations from among 20 sites. Use ENTROPY

normal- generalized inverted Wishart predictive distribution

- **needs "whitened" residuals; space - time interaction $\longrightarrow$ use of 24 (hour) dimensional multivariate AR(1) model**
- **different 10 - station startups $\longrightarrow$ monotone ("staircase") data structure $\longrightarrow$ generalized inverted Wishart distribution $\longrightarrow$ different d.f. for each staircase step**
- **select the 6 new stations with jointly maximum conditional entropy**

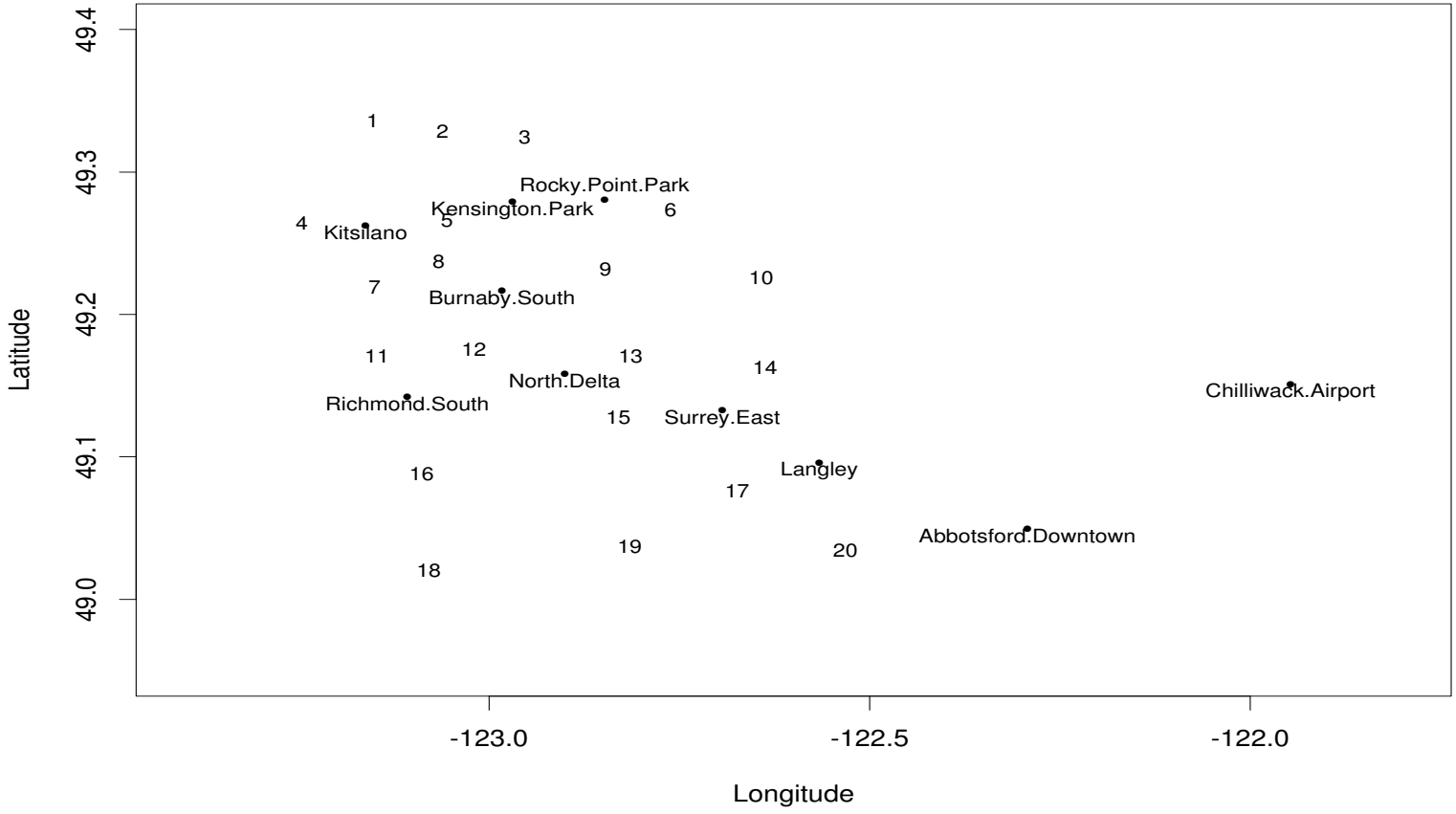
Results

Here is a look at PM_{10}'s levels at the 10 existing stations. Note differing startup times.



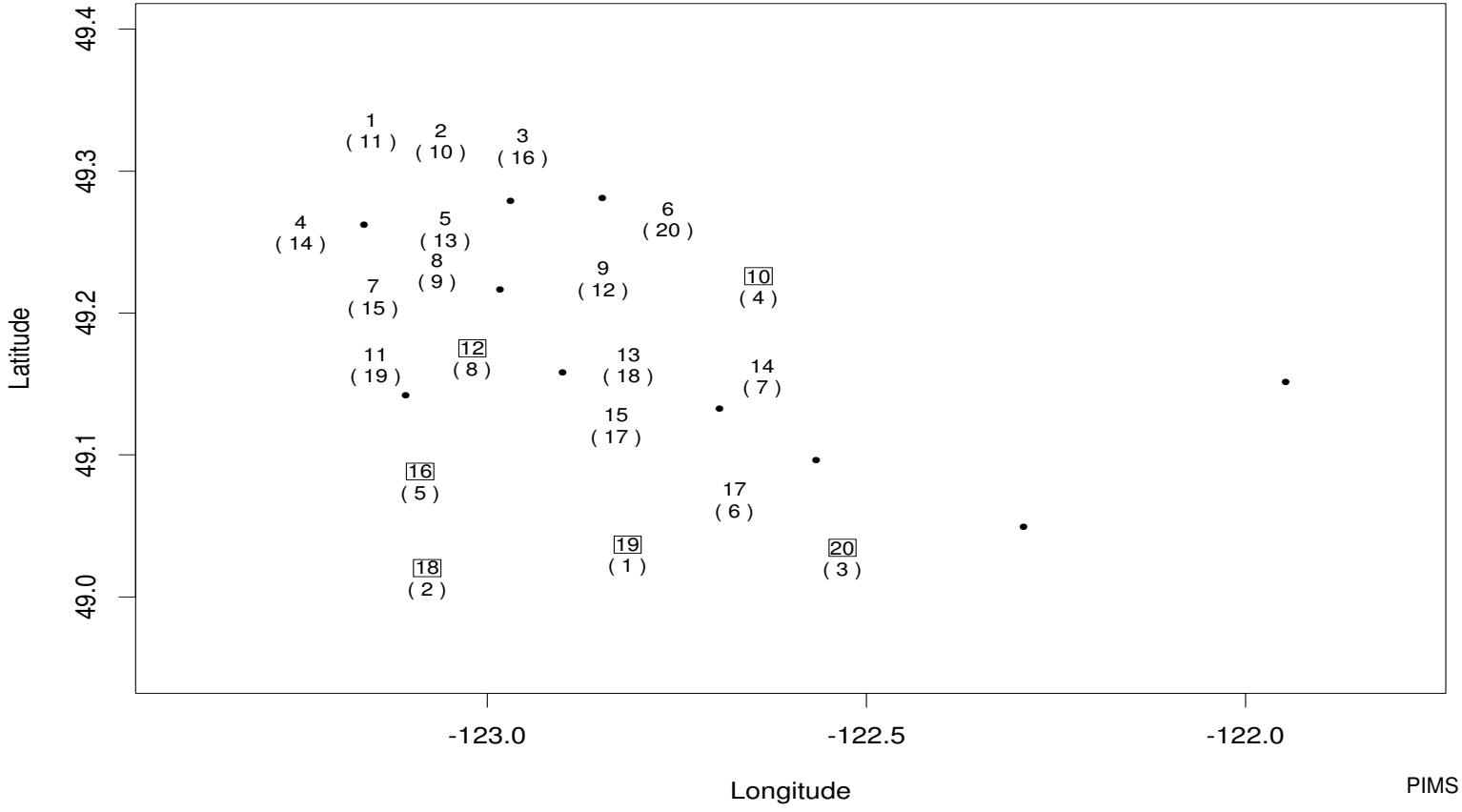
Results (cont'd)

The 10 PM_{10} monitoring sites and prospective new locations.



Results (cont'd)

Locations of the old and newly selected sites (square brackets). Rank by estimated variance is in curved brackets.



Noncompliance, not entropy!

Example: how well would Vancouver's 6 site, entropy-based, addition compare to an optimal noncompliance based addition?

Use 10 station hourly data for PM_{10}, February 28, 1999 & hierarchical Bayes predictive distribution

CRITERION:

$$\textbf{argmax PR}\{\textbf{daily max } PM_{10} \textbf{ Y}^{6added} \geq \textbf{50 } (\mu g\, m^{-3})\}$$

NOTES: Entropy does not work nearly as well on August 1, 1998!

About non-compliance detection designs

Probability, hence best design, day - dependent!

- which day?
- a simulated future day? Average day? Bad day?

How implemented?

- monitor sites most likely to comply?
- do not monitor sites least likely?
- what about existing sites?

Conclusions: **Selecting best non compliance designs presents challenging non-statistical issues.**

What about extremes?

Entropy theory can be made to work. **Approximate joint distribution of extremes field by log multivariate - t distribution**

Empirical results:

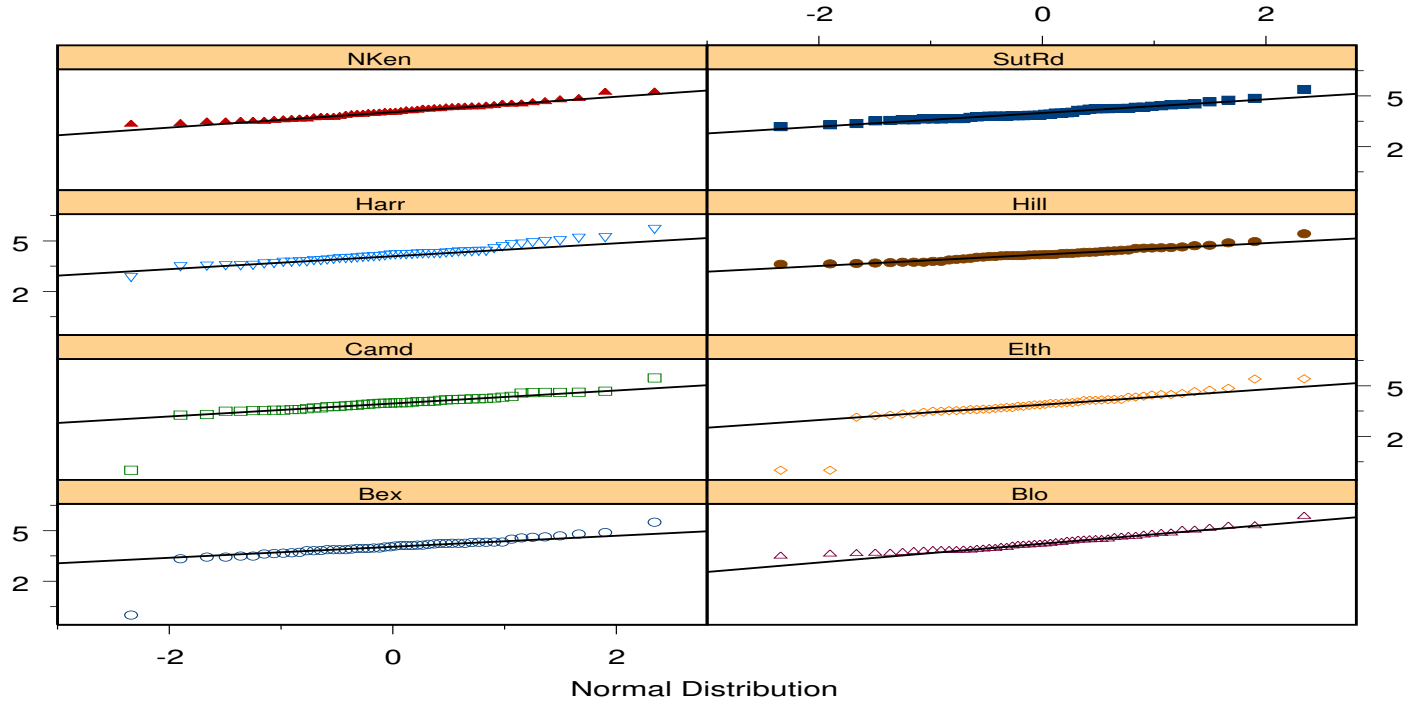


Figure 1: QQplots for weekly maxima of hourly $\log PM_{10}$

T approx approach cont'd

Empirical results $\longrightarrow$ well-calibrated 95% (etc) prediction intervals. Supports use of multivariate approximation even for extremes.

Credibility Level	Mean	Median
30%	35	35
95%	96	97
99.9%	99.9	1

Table 1: **Summary of coverage probabilities at different credibility levels for the simulated precipitation data over 319 grid cells, Canadian Climate Model**

Results for Vancouver redesign

The selected new sites are now determined pretty much by their posterior estimated variances. That is because the spatial correlation is now quite weak.

The good news

Use of log multivariate t distribution for extreme fields promising. But:

- **how far can approximation go $\longrightarrow$ need for theory**
- **test approximation case-by-case**
- **no substitute for knowledge of latent processes**
- **need to compare regular - and extreme-entropy designs.**

Recommendations to regulators

Spend some data dollars assessing current designs

Select simpler regulatory metrics.

Think about/articulate design purposes

Conclusions

Current urban networks may not be dense enough for adequate surveillance

Conventional designs inadequate but MaxEnt may be adapted/used

More knowledge of latent processes needed

More attention to design criteria for extremes needed

Designing for extremes $\longrightarrow$ significant challenges

References

Chang, H, Fu, H, Le, ND, Zidek, JZ. Perspectives on Designing Environmental Monitoring Networks for Measuring Extremes. EES. To appear