Reconciling Physical & Statistical Approaches to Modeling.

Jim Zidek

U British Columbia

Acknowledgements

- Prasad Kasibhatla, Duke
- Douw Steyn, UBC
- Co-authors:
  - Nhu Le, BC Cancer Agency
  - Zhong Liu, UBC

Outline

- **Origins:** of the talk
- **Physical modelling perspective**
- **Physical vs statistical modeling themes**
  - **Bayesian melding**
  - **Alternatives**
- **Applications**
- **Conclusions**

Origins

- Need to model environmental space -time fields over large space - time domains that challenge physical and statistical modelers
- Space time research study group: Statistical and Applied Mathematical Sciences Institute, Jan - May, 2003.

What's a Model?

"an abstract, analogue representation of the prototype whose behavior is being studied" (Steyn & Galmarini 2003)

5

Physical Modeler's Perspectives

Phys model classification

- **Analytic Models:**
  - variables in tractable math equations represent measurable attributes of the real thing
- **Physical Scale Models**
  - physical behavior of their measurable properties analogous to that of the real thing
- **Numerical Models**
  - variables obtained by numerical solution thought to be analogous to measurable attributes of the real thing
  - **Example:** Climate Models: IMS Talk 2

Controversy! The Oreskes Paper

The paper (OSB): *Oreskes, Schrader-Frechette & Belitz (1994) Science, 263, 641-646*

- highly influential
  - says physical models cannot be shown to represent reality - validation meaningless/pointless
  - still cited over 40 times per yr
  - used to justify not validating!

7

Controversy! The Oreskes Paper

The paper (OSB): *Oreskes, Schrader-Frechette & Belitz (1994) Science, 263, 641-646*

- dismisses common assessment practices
  - verification
  - validation
  - verifying numerical solutions
  - calibration
  - confirmation

Oreskes

On "confirmation" for example.

- **Confirmation:** concluding that simulated - real data match $\Rightarrow$ truth is **logical fallacy**: *"affirming the consequence"*

  EXAMPLE: **Hypothesis H**: "It is raining." **Model:** "If H, I will stay home and revise the paper." You find me at home and conclude H valid since data matches prediction under model hypothesis!
- poor predictions $\Rightarrow$ bad model!
- good predictions $\not\Rightarrow$ good model!
  - many good models possible
  - bad hypotheses could cancel each other

Oreskes

Summary:

“The primary purpose of models in heuristic...useful for guiding further study but not susceptible to proof... [Any model is] a work of fiction. ... A model, like a novel may resonate with nature, but is not the ‘real thing’.”

11

Steyn & Galmarini Counterattack!

- reject alternative: pure empiricism
- go for an compromise between pure empiricism models and “true” models:
  - models have predictive & heuristic value
  - but define “success” before assessment to avoid “gradualism”
  - they provide evidence of predictive value of models
- current hot topic in phys modelling & other communities

Phys - Stat Modelling Themes

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

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

Phys - Stat Modelling Themes

THEME 2: Statistics can help interpret, analyze, understand, exploit outputs of complex phys models [Nychka 2003].

- Example: statistics on modeled precipitation (precip) extremes gives coherent return values over space for design

15

Phys - Stat Modelling Themes

- **THEME 3**: Physical (phys) and statistical (stat) models can produce synergistic benefits by ”melding” them.
  - Wikle, Milliff, Nychka, Berliner (2001). JASA.
  - Example: how can simulated (modelled) and real ozone data be usefully combined?

Theme 3: Simulated + Real Data

Does this make sense?

- Example:

$$(2 \qquad + \ 1 \qquad)/2 = 1.5$$

Seems correct. But its actually nonsensical.

Theme 3: Simulated + Real Data

Does this make sense?

- Example:

$$(2 \qquad + \ 1 \qquad)/2 = 1.5$$

  Seems correct. But its actually nonsensical.

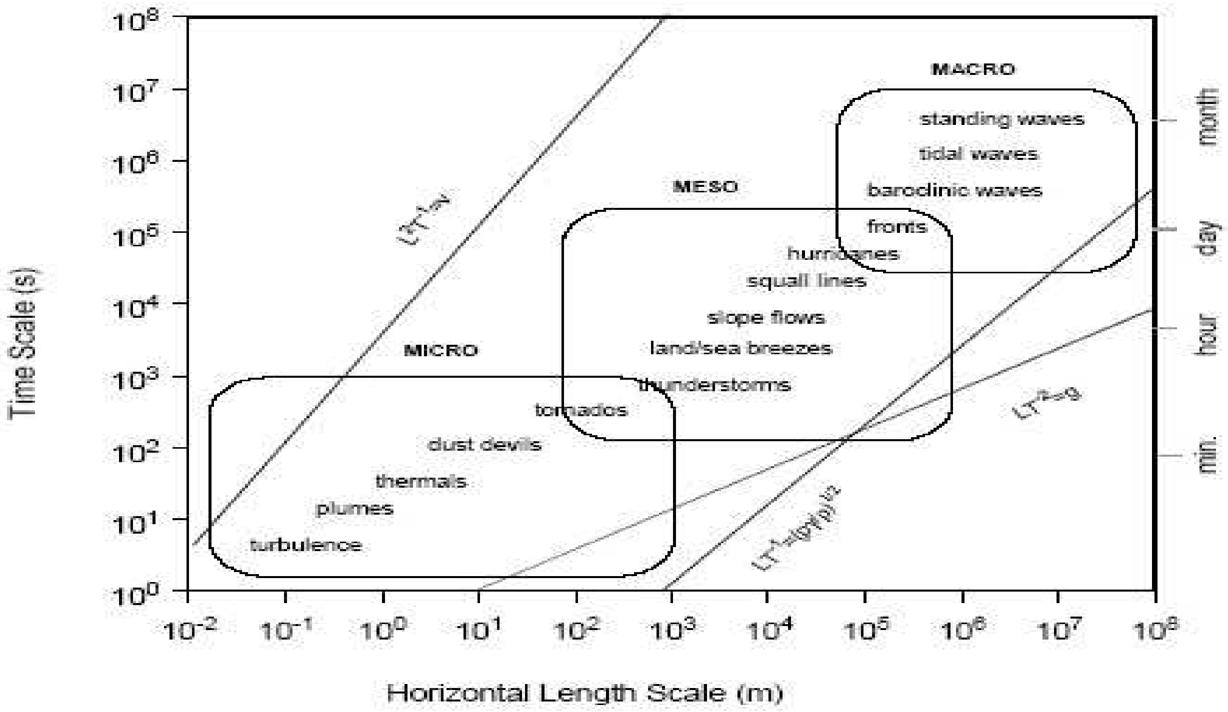
-

$$(2 \textbf{ cm} + 1 \textbf{ apple})/2 = 1.5$$

  Phys model data scales differ from real data

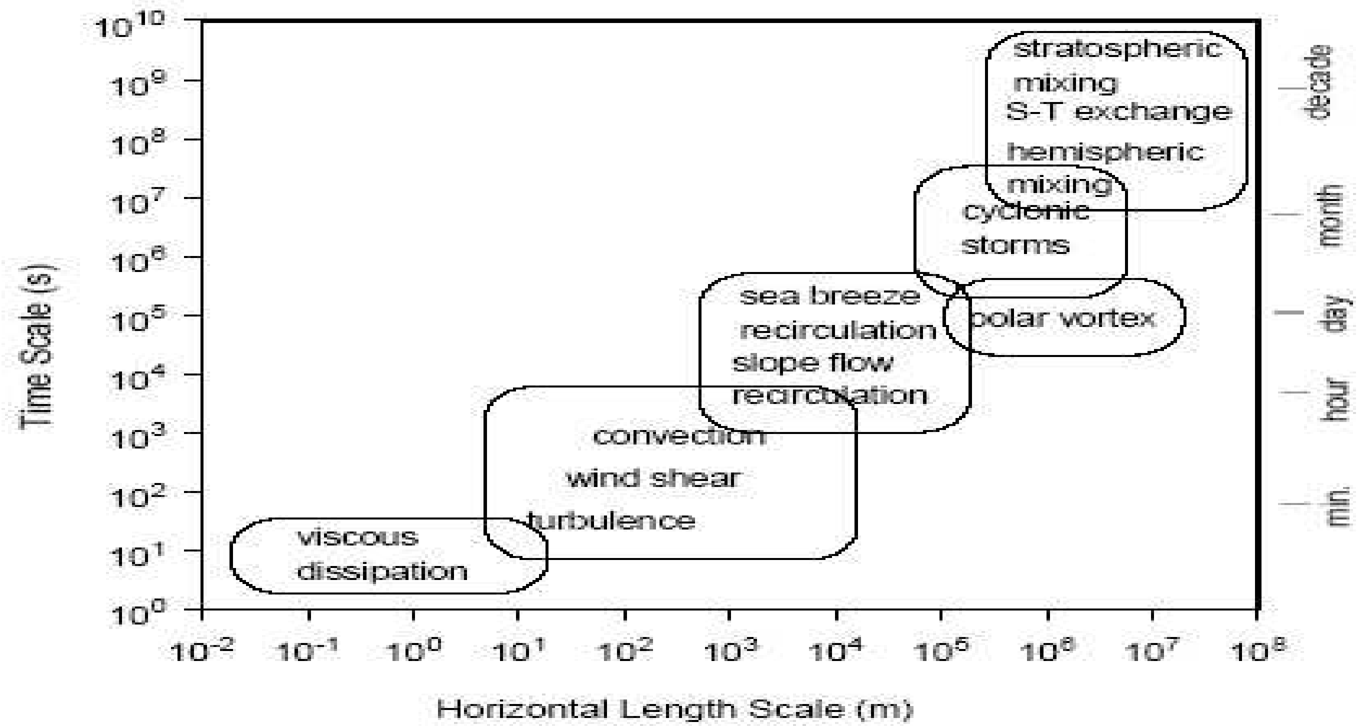
Model Dynamic Scales

The problem (Steyn & Galmarini 2003):



Continuous real data monitors: scale just 1 $m^2 \times$ few minutes - lower left hand corner!!

Model Dispersion Scales



Model Chemical Scales

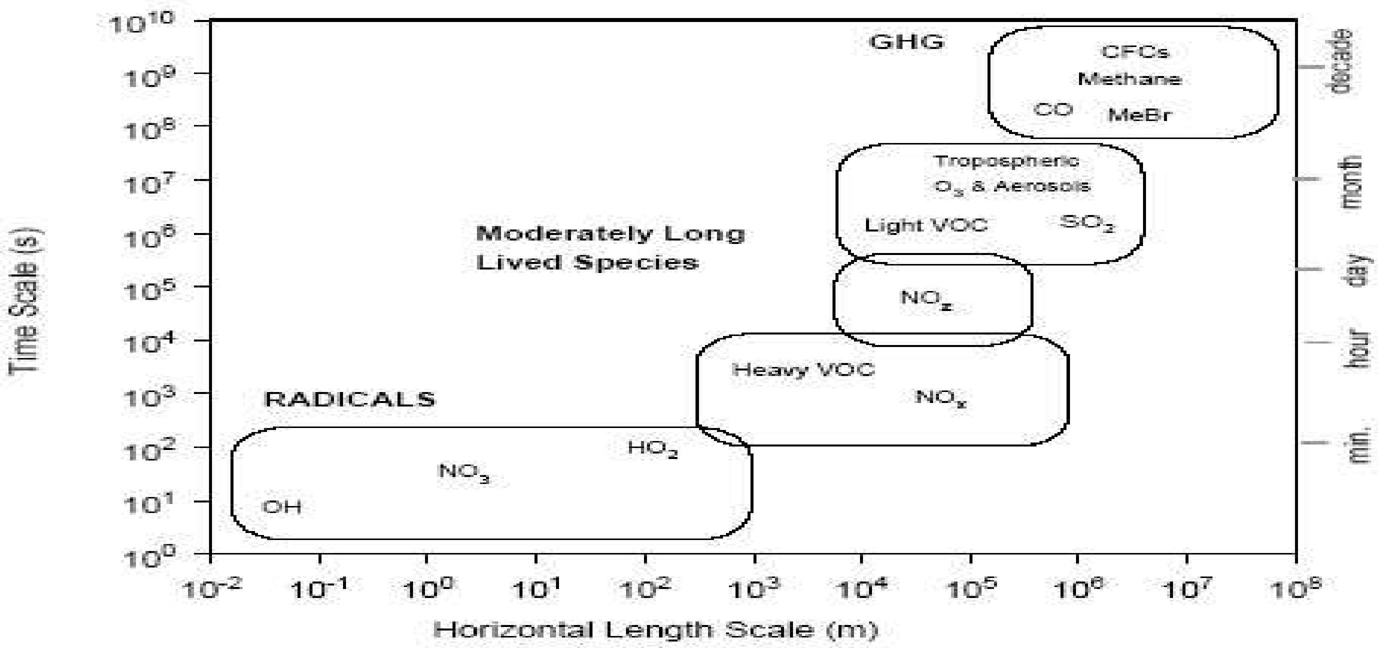
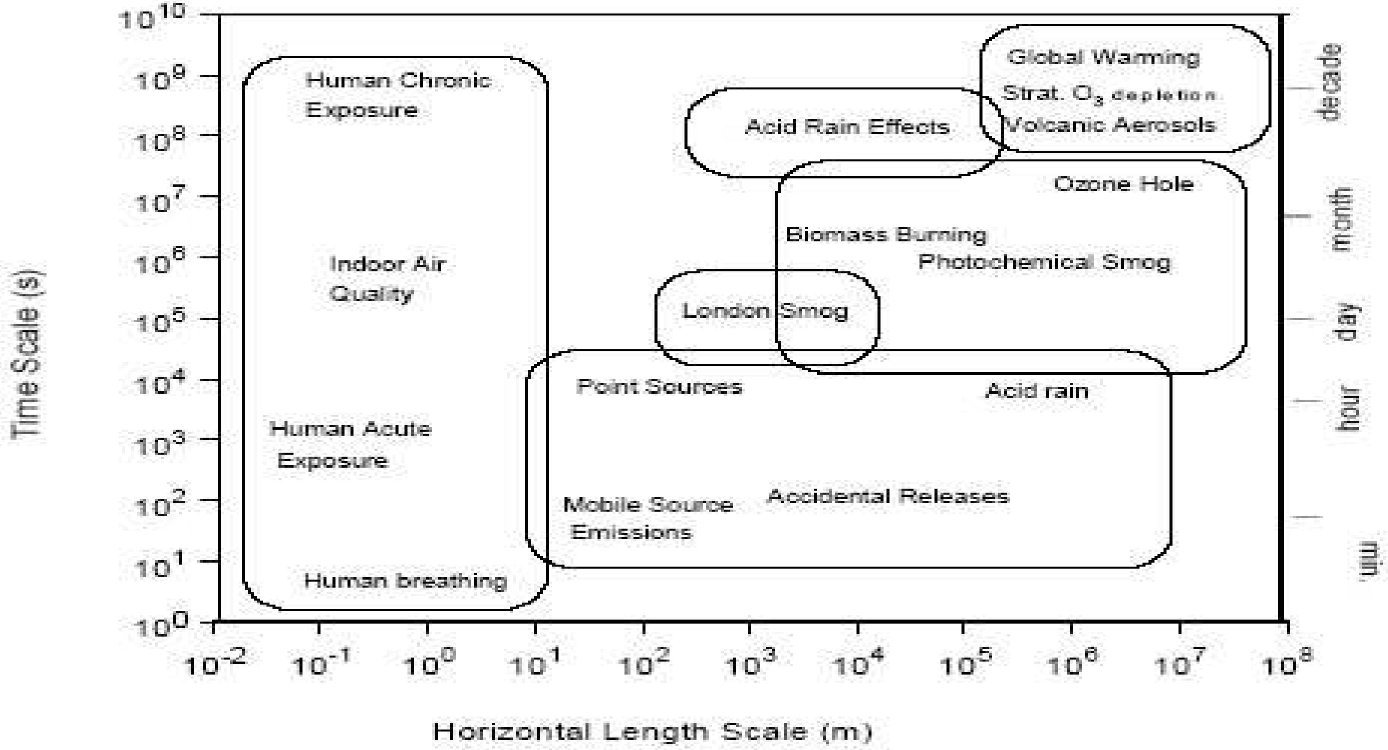


Figure 3. Time and space scales of atmospheric chemical processes

Model Human Scales



Phys vs Stat Models

Physical models:

- prior knowledge expressed by math equations (de's)
- can lead to big computer models
- yield deterministic response predictions
- can fail because of:
  - butterfly effect
  - nonlinear dynamics
  - lack of background knowledge
  - lack of computing power

25

Phys vs Stat Models

Statistical models:

- prior knowledge expressed thru stat models & priors
- can lead to big computer models
- yield predictive distributions
- can fail because of:
  - too simplistic off-the-shelf models
  - lack of background knowledge
  - lack of computing power

Phys vs Stat Models

May be strength in unity but:

- approaches differ
- big gulf between two cultural “attitudes”
- communication between camps strained
- route to reconciliation unclear

29

Phys vs Stat Models

- **General framework** [Berliner (2003)]:
  - measurement model
  - process model
  - parameter model
- fits nicely with hierarchical Bayesian modeling

Phys vs Stat Models

Strategies for combining depend on:

- purpose
- context
- # of mathematical equations involved

Phys vs Stat Models

Case 1: Only a few de's:

Example: $dX(t)/dt = \lambda X(t)$.

- solve it and make constants random:

  $$X(t) = \beta_1 \exp \lambda t + \beta_0$$

  (Wikle et al 2001)
- discretize the de and add noise to get a state space model: $X(t+1) = (1+\lambda)X(t) + \epsilon_t$ (Wikle et al 2001)
- use functional data analytic approach - incorporate de's via penalty term (as in splines; Ramsay & Silverman 1998?)
  $\sum_t (Y_t - X_t)^2 + \int (DX(t) - \lambda X(t))^2 dt$

Phys vs Stat Models

Case 2: Many mathematical (differential) equations, e.g. 100:

- construct better predictive density:
  - $f(real|simulated)$ eg input simulated value as prior mean
  - Mayer Alvo (1990??)
- view simulated data as real - build joint density ("**Bayesian melding**"):
  - $f(real, simulated) = \int f(real|truth) f(simulated|truth) \times \pi(truth) d(truth)$
  - Fuentes & Raftery (2005). Biometrics.

Bayesian melding

Combining measurement & and model data with different scales.

- Instrument ensemble $i = 1, \ldots, D$
- Model ensemble $i = 1, \ldots, M$

- s = spatial location
- B = grid cell
- $\hat{Z}_i(s)$ = measurement i at site s
- $\tilde{Z}_j(B)$ = model j output for cell B
- $Z(s)$ = truth links measurement and model

Extends Montserrat Fuentes and Adrian E. Raftery 2005 Biometrics

Bayesian melding

Combining measurement & and model data with different scales.

Measurement equations:

For all instruments, i:

$$\hat{Z}_i(s) \quad = \quad a_i^I(s) + b_i^I(s) Z(s) + e_i(s)$$

Here

$$\begin{aligned} Z(s) &= \text{``truth''} \\ e_i(s) &\sim N(0, \sigma_{ei}^2) \perp Z(s) = \text{measurement error} \\ a_i^I(s) &= \text{additive bias instrument i} \\ b_i^I(s) &= \text{multiplicative bias instrument i} \end{aligned}$$

Bayesian melding

Combining measurement & and model data with different scales.

Truth:

$$
\begin{aligned}
Z(s) &= \mu(s) + \epsilon(s) \text{ where} \\
\mu(s) &= \text{spatial trend} = \text{linear polynomial in lat, long} \\
\epsilon(s) &= \text{trend mispecification error} \\
&\sim N(0, \Sigma) \\
\Sigma &= \textit{spatial covariance matrix}
\end{aligned}
$$

Bayesian melding

Combining measurement & and model data with different scales.

Simulated data equations:

For model j:

$$\begin{aligned}
\tilde{Z}_j(s) &= a_j^M(s) + b_j^M(s)Z(s) + \delta_j(s) \\
a_j^M(s) &= \text{additive mispecification error, model j} \\
b_j^M(s) &= \text{multiplicative mispecification error, model j} \\
\delta_j(s) &= \text{model mispecification error} \\
&\sim N(0, \sigma^2_{\delta_j}) \perp Z(s), e(s)
\end{aligned}$$

Implementation

- Truth assumed as locally stationary ("**convolution approach**" of Fuentes - variation of Higdon):

$$Z(x) = \int_D K(x-s) Z_{\theta(s)}(x) ds$$

$Z_{\theta(s)}(x)$ = stationary spatial process over x for each fixed s. But Z not stationary! In fact:

- $C(s_1, s_2; \theta) = \int_D K(s_1 - s) K(s_2 - s) C_{\theta(s)}(s_1 - s_2) ds$ with
  - D = whole region

- For simplicity:

$$
\begin{aligned}
K(u) &= h^{-2}K_0(h^{-1}u)\text{!only choice of h being critical} \\
K_0(u) &= \frac{3}{4}(1-u_1^2)_+\frac{3}{4}(1-u_2^2)_+
\end{aligned}
$$

and

- $C_{\theta(s)}$=Matern covariance kernel with
  - $\theta(s) = (\nu_s, \sigma_s, \rho_s)$

50

Matern covariance model

Let $d = s_1 - s_2$. Then

$$\begin{aligned} Cov[Z(s_1), Z(s_2)] &= C_\theta(d) \\ &= \frac{\sigma}{2^{\nu-1}\Gamma(\nu)}(2\nu^{1/2}|d|/\rho)^\nu K_\nu(2\nu^{1/2}|d|/\rho). \end{aligned}$$

Here:

- σ = *sill* parameter
- ρ = *range* parameter (rate of decay of spatial correlation)
- ν = *smoothness* parameter ($\nu = 1/2$ yields exponential decay model)

Integrals to sums

Draw systematic location sample $s_m \in D, \; m = 1, 2, ..., M$. Then

$$C(s_1, s_2; \theta) \approx M^{-1} \sum_{m=1}^{M} K(s_1 - s_m) K(s_2 - s_m) C_{\theta(s_m)}(s_1 - s_2)$$

54

Prior - posterior modelling

- All variances assumed inverted gamma priors
- All multiplicative biases $b(s) = b$'s assumed spatially invariant (based on expert input) with joint Gaussian prior. Additive biases vary over space and get a joint Gaussian prior.
- The sill and range parameters for Matern get ANOVA forms over a regular grid indexed by (i, j):

$$\begin{aligned} \sigma_{i,j} &= \alpha + r_i + c_j + \epsilon_{i,j} \\ \rho_{i,j} &= \alpha' + r'_i + c'_j + \epsilon'_{i,j} \end{aligned}$$

  the $\{r_{i,j}\}$, etc, getting a joint Gaussian prior distribution
- All linear mean parameters get joint Gaussian parameters.

Prior - posterior modelling

- n monitoring sites & measurements $\hat{Z}$
- sample L grid points in each of m grid cells, B_i: $\{s_{1,B_i}, ..., s_{L,B_i}\}$

$$Z(B_i) \approx \frac{1}{L} \sum_{j=1}^{L} Z(s_{j,B_i}) + \delta(s_{j,B_i})$$

- split truth vector Z for $n + mL$ sites into Z_1 (n monitoring sites) Z_2 (mL sampling sites).

Prior - posterior modelling

Joint posterior distribution:

$$
\begin{aligned}
& p(\hat{\mathbf{Z}}, \tilde{\mathbf{Z}}, \mathbf{Z}, \beta, \theta, \mathbf{a}, \mathbf{b}, \sigma^{\mathbf{2}}_{\mathbf{e}}, \sigma^{\mathbf{2}}_{\delta}) \\
= \; & p(\hat{\mathbf{Z}}|\mathbf{Z}, \sigma^{\mathbf{2}}_{\mathbf{e}})\mathbf{p}(\tilde{\mathbf{Z}}|\mathbf{Z}, \mathbf{a}, \mathbf{b}, \sigma^{\mathbf{2}}_{\delta})\mathbf{p}(\mathbf{Z}|\beta, \theta)\pi(\sigma^{\mathbf{2}}_{\mathbf{e}}, \sigma^{\mathbf{2}}_{\delta}, \beta, \theta) \\
= \; & \Phi_{\sigma^2_e \mathbf{I}}(\hat{\mathbf{Z}} - \mathbf{Z_2})\boldsymbol{\Phi}_{\sigma^{\mathbf{2}}_{\delta}\mathbf{I}}(\tilde{\mathbf{Z}} - \mathbf{Z_1}) \\
& \Phi_{\Sigma(\theta)}(\mathbf{Z} - \mathbf{X}\beta)\pi(\sigma^{\mathbf{2}}_{\mathbf{e}})\pi(\sigma^{\mathbf{2}}_{\delta})\pi(\beta)\pi(\theta).
\end{aligned}
$$

Now go to MCMC

60

Spatial prediction

- Kriging: an alternative using data or simulated data
- unmonitored sites simple to predict with melding
  - part of MCMC
  - yields predictive distribution with 95% credibility (predictive) intervals

The Ozone Project

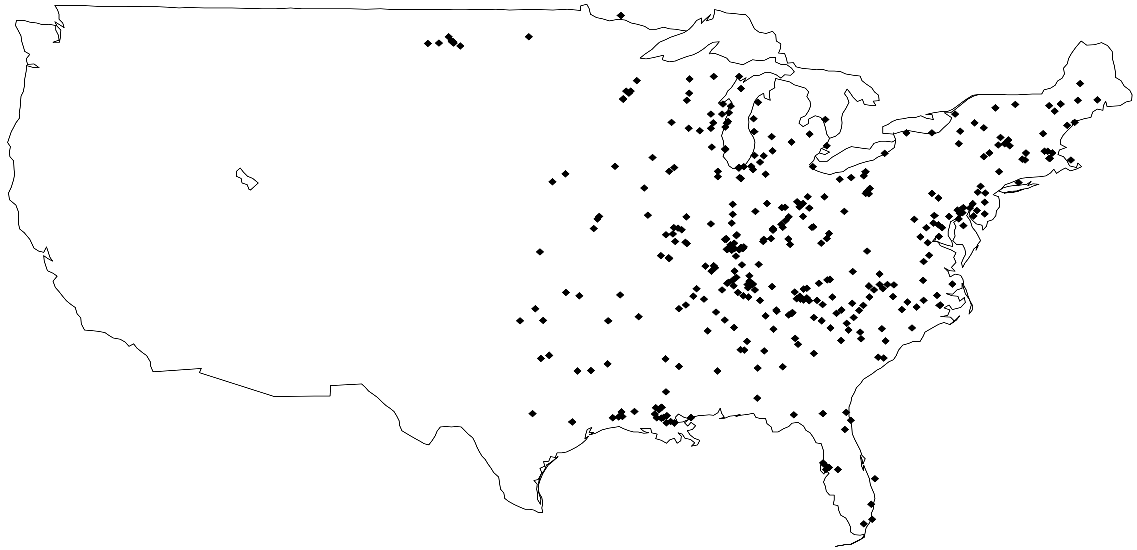
Air pollution transportation models:

- mathematical - computer models:
  - capture nonlinear photochemical interactions
  - predict/simulate air pollution
  - URM 1994, UAM-V 1995, CAMx 1997, SAQM 1997, MAQSIP 1996, MODELS-3 1998
- developed for variety of purposes:
  - assessing success of abatement strategies
  - regulation & control
- yield “simulated data”

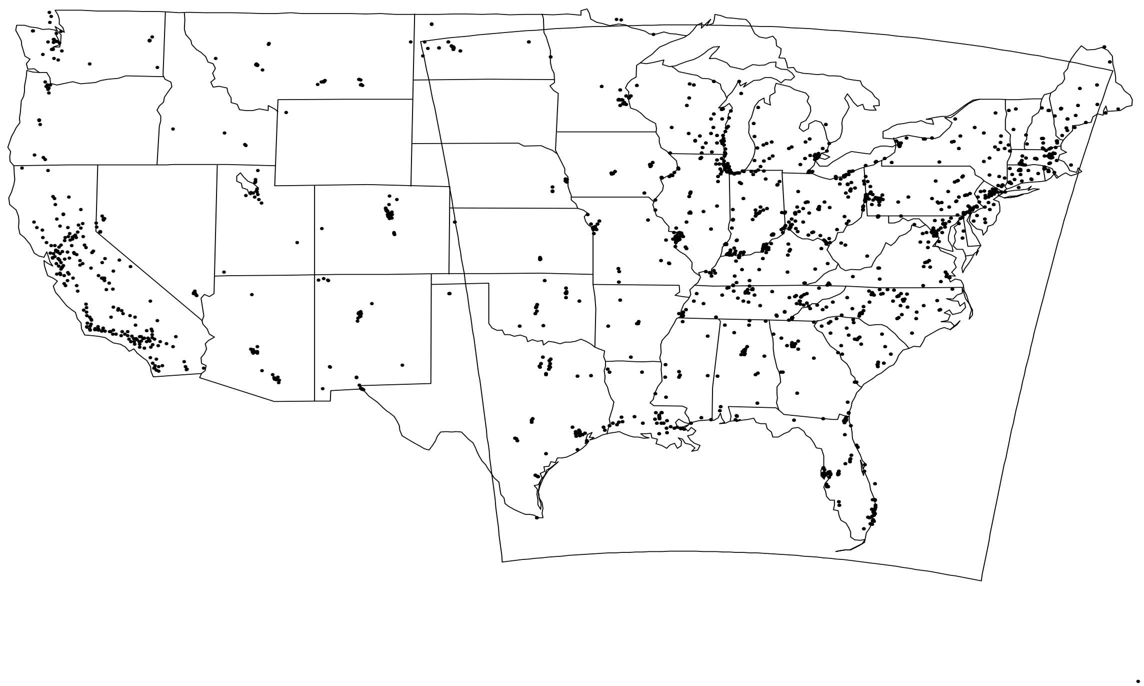
Ozone Chem Trans Mods

- hourly concentrations of ground level O_3 (ozone) over eastern US for 120 days from May 15 - Sept 11 1995.
- simulated concentrations from the **MULTISCALE AIR QUALITY SIMULATION PLATFORM** (MAQSIP) model
- measured concentrations from > 200 monitoring sites from US EPA's AIRS database.
- **BASIC QUESTIONS:**
  - Can simulated data help in spatial prediction of data?
  - Can data help recalibrate the simulated data?
  - Can simulated data be substituted for data?
  - How might they be combined?

Simulated Data (MAQSIP) cells

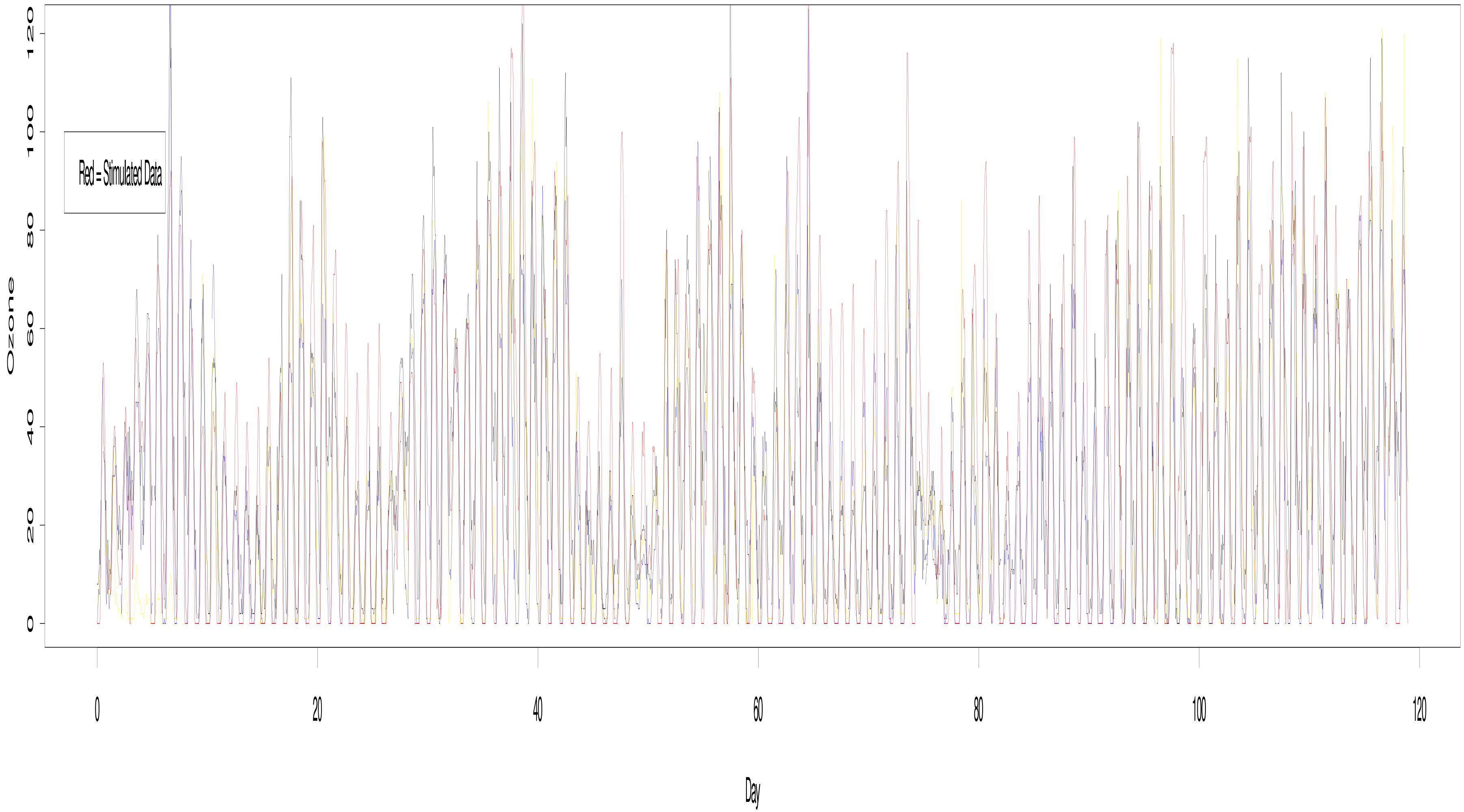


Real Data Ozone Sites



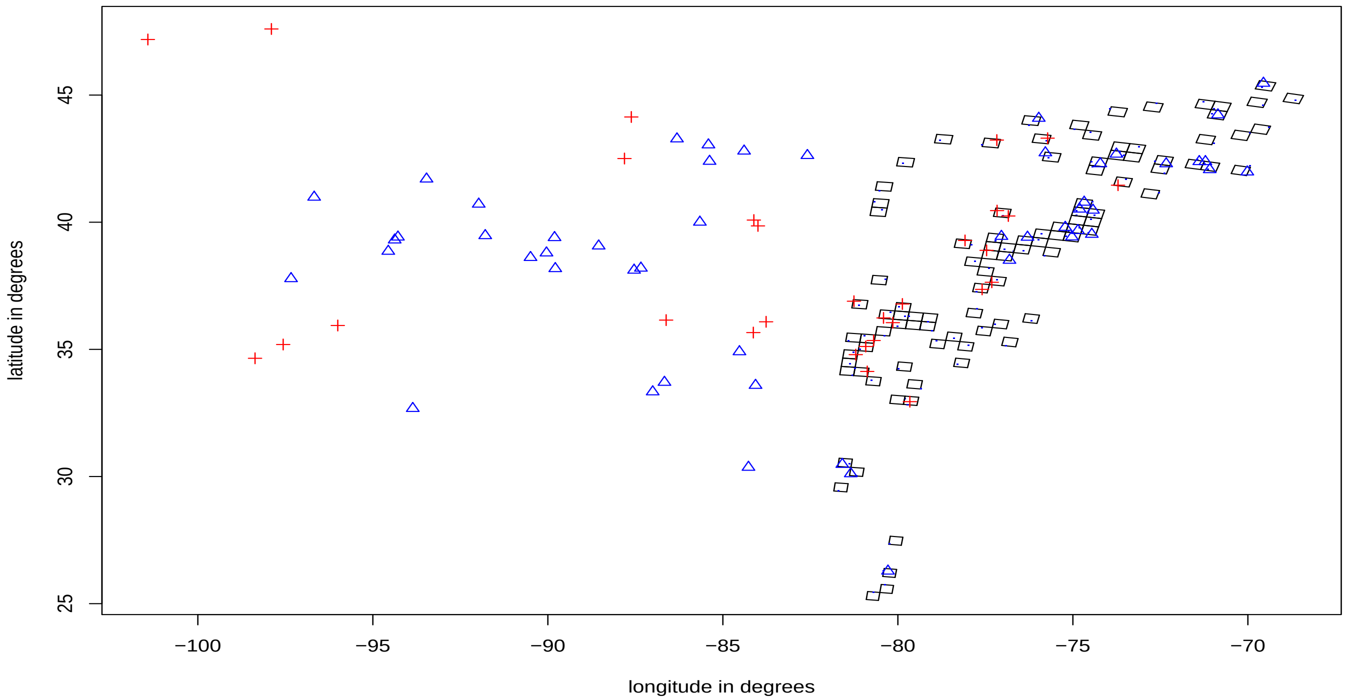
Comparisons

TS plots of 3 real & 1 simulated data series. Simulated more variable than real!



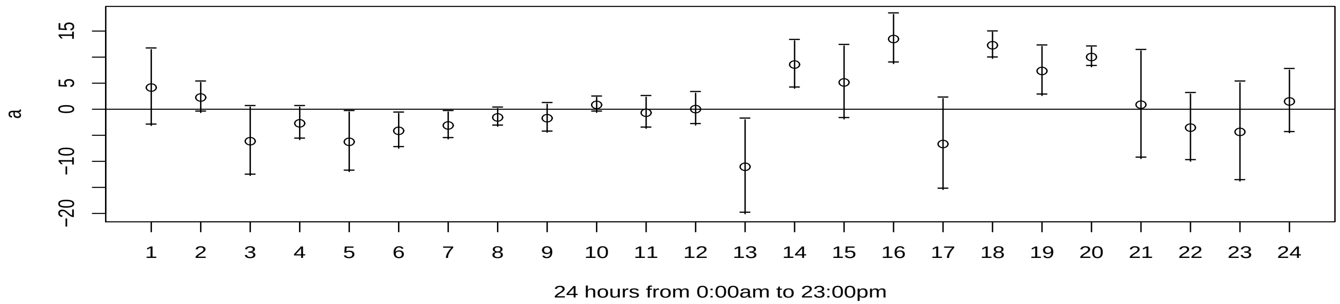
Ozone application results

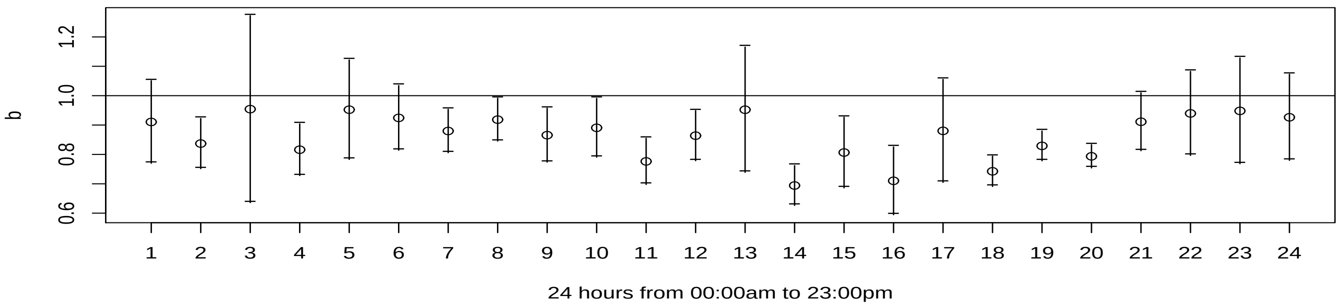
Site and grid cell locations. **TRIANGLES**- **51** monitored sites, **RECTANGLES** - **100** grid cells, **PLUSES**- **30** unmonitored sites



Ozone application results

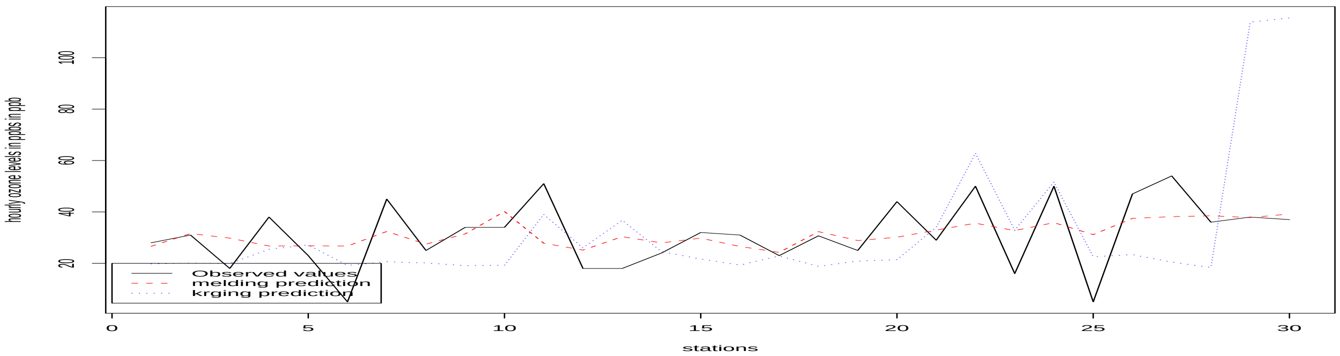
Substantial model bias at certain hours a (additive bias - upper panel) b (multiplicative bias - lower panel)

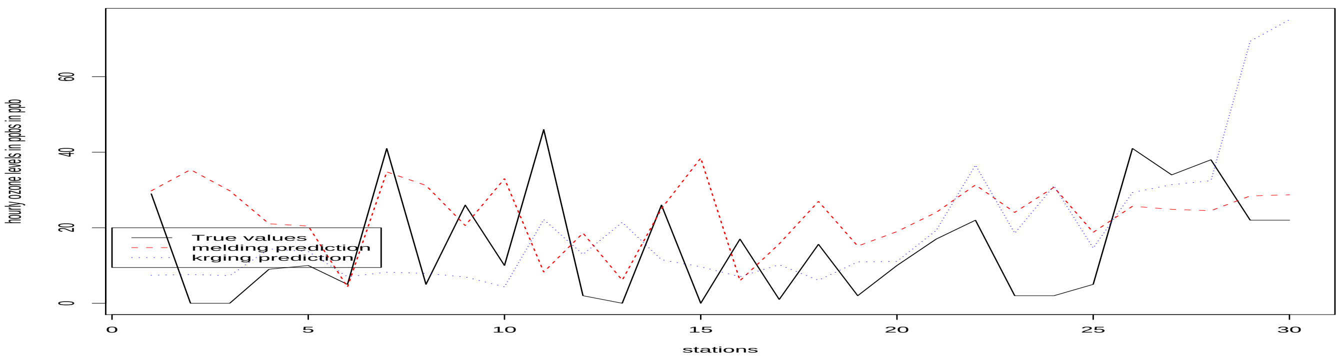




Ozone application results

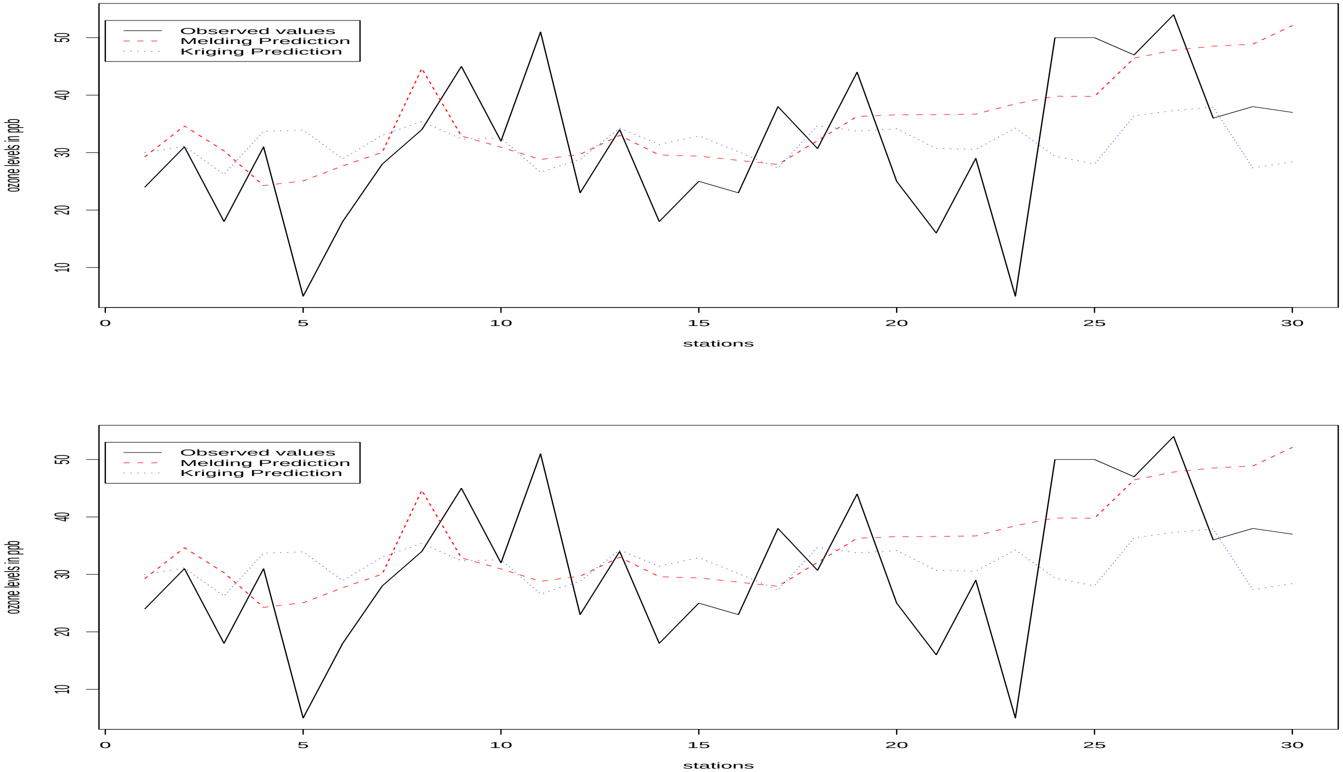
Comparison: melding (Red)and kriging (**Blue**) - 2 different hours 30 sites





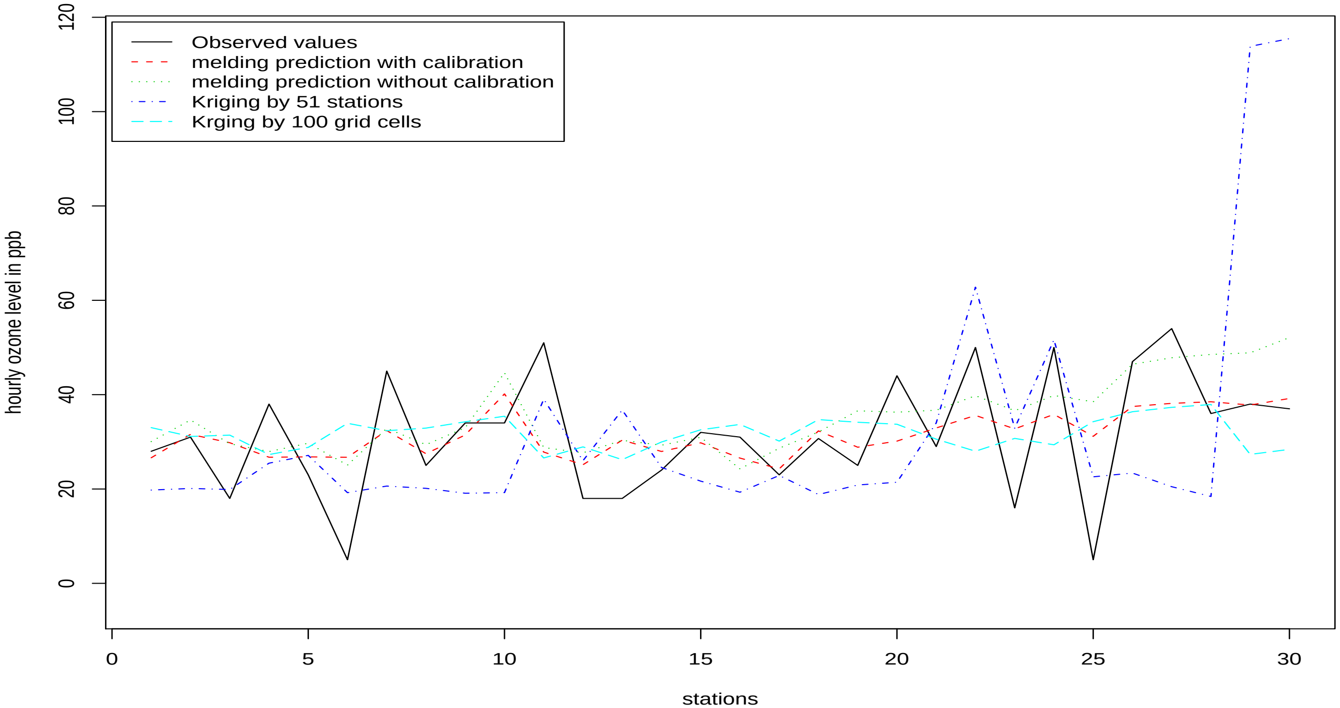
Ozone application results

Comparison: melding (Red) simulated data only & kriging (Blue) - 2 different hours 30 sites



Ozone application results

Comparison: melding & kriging - 1 hour 30 sites. **Red plot** - bias adjusted melding improves prediction. **dotted green plot** - no model bias adjustment. **dashed turquoise plot**-kriging the grid cell data



Melding report card

Pros:

- permits meso - micro scales scale integration
- makes good use of model outputs
- includes model - only ensembles (probabilistic weather forecasting)
- enables lots of model diagnostics - not merely prediction!
- generally better than Kriging (smaller mean squared prediction errors)
- implementation (R) software now online

80

Melding report card

Cons:

- computationally intensive
- large nos of grid cells & monitoring sites means big, numerically unstable covariance matrices
- approach to misaligned data not supported by physics
- Kriging a lot simpler for spatial prediction

Melding report card

Alternatives:

Two step regression:

$$\begin{aligned} O_{st} &= a_s + c_s M_{st} + N_{st} \\ N_{st} &= \rho N_{s,t-1} + \gamma_{1t} Z_{1t} + \cdots + \gamma_{24t} Z_{24t} + \epsilon_{st} \end{aligned}$$

where the $\{Z_{jt}\}$ are $0 - 1$ hour indicators.
This model can be fitted into a Bayesian framework and works better for spatial prediction than melding since it includes all the data over time.

Other applications

- Probabilistic weather forecasting
- Setting error bars on climate model predictions
- Calibrating deterministic model outputs for point inference

Concluding remarks

- Results for melding encouraging but work remains.
- Simulated data improves spatial predictive performance, melding beats Kriging and can produce better calibrated predictive intervals
- Physical statistical modelling part of a larger trend from “normal science” to “post - normal science”

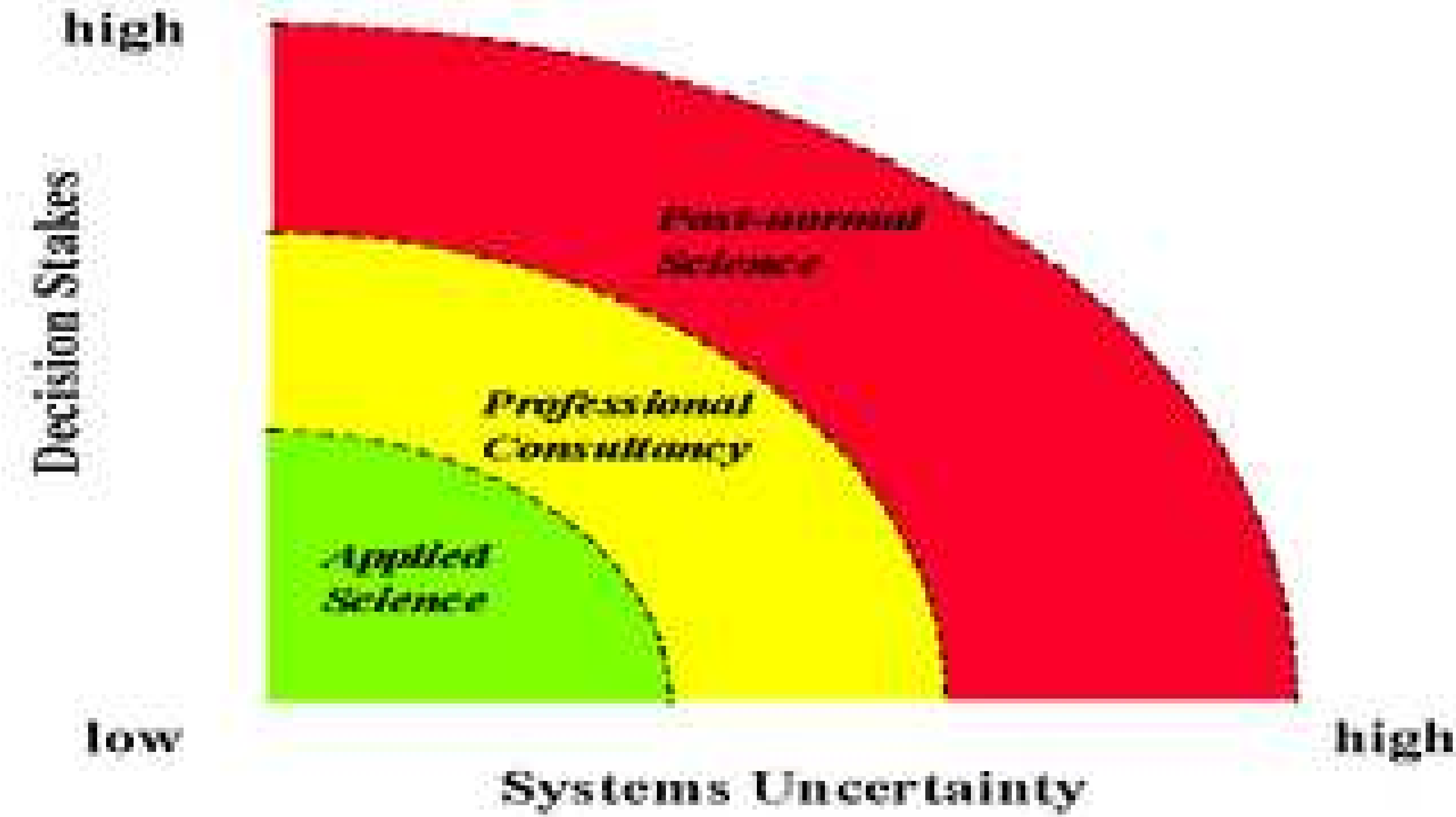
Concluding remarks

Funtowicz, Ispra Ravetz (2004?) Nusap.net:

> "...key properties of complex systems, radical uncertainty and plurality of legitimate perspectives....When facts are uncertain, values in dispute, stakes high, and decisions urgent the ...guiding principle of research science, the goal of achievement of truth,...must be modified. In post-normal conditions, such products may be ...an irrelevance."

Concluding remarks

From Funtowicz et al:



Extended version of this talk to be posted. Follow links from http://www.stat.ubc.ca/<faculty members LINK>

Concluding remarks

Other PIMS CRG events in 2008

- **Banff International Research Station for Mathematical Innovation and Discovery (BIRS): The Climate Change Impacts on Ecology and the Environment** May 4 -9, 2008.
- The 2008 annual **International Environmetrics Society Conference:**
  Quantitative Methods for Environmental Sustainability
  June 8-13, 2008, Kelowna, Canada.
  http://people.ok.ubc.ca/zhrdlick/ties08/call.htm

Concluding remarks

Other PIMS CRG events in 2008

- The **PIMS International Graduate Institute's Summer School**
  Computation in environmental statistics
  Tentatively: Jul 28-Aug 1, 2008, National Center for Atmospheric Research (NCAR), Boulder Colorado
- **Workshop on extreme climate events**
  Winter, 2008, Lund University

Concluding remarks

Email: jim@stat.ubc.ca

Homepage: http://www.stat.ubc.ca/ jim

Tech reports: http://www.stat.ubc.ca/Research/TechReports/

Melding software: http://enviro.stat.ubc.ca