Statistical Issues in the Mathematical Geosciences

Christopher K. Wikle
Department of Statistics
University of Missouri-Columbia

Outline

- General Framework
- Model Data Fusion
- Statistical Issues

NISS CMG Workshop, October 6, 2005

What is the Statistician's Role?

Statisticians are concerned with uncertainty!

Some general areas (all related) in the geosciences for which statisticians can contribute include:

- Model Data Fusion (Data Assimilation)
- Dimension reduction
- Multiscale modeling
- Multiprocess modeling
- Computation
- Spatio-Temporal field comparison
- Extremes

General Framework: Hierarchical Bayesian

It is increasingly common for statisticians to represent problems in geosciences in three stages (Berliner, 1996):

Basic Hierarchical Model

1. [data|process, parameters]
2. [process|parameters]
3. [parameters]

Bayes:

$$[\text{process,parameters}|\text{data}] \propto [\text{data}|\text{process, parameters}] \times [\text{process}|\text{parameters}][\text{parameters}]$$

Bayesian Hierarchical Modeling: Think Conditionally!!

Data $\rightarrow Y$ Process $\rightarrow X$ Parameters $\rightarrow \theta$

- Data Model(s): $[Y \mid X,\ \theta\ ]$
  - Simpler dependence structures through conditioning. e.g.,
    $$[Y_a, Y_b|X, \theta] = [Y_a|X, \theta][Y_b|X, \theta]$$
  - Change of resolution; misalignment
- Process Model(s): $[X \mid \theta\ ]$
  - Can build-up complicated dependence by conditional models. e.g.,
    $$[X_2, X_1|\theta] = [X_2|X_1, \theta][X_1|\theta]$$
  - Incorporate science!! (e.g., PDEs)
- Parameter Model(s): $[\ \theta\ ]$
  - Further conditioning, e.g., $[\theta_1|\theta_2][\theta_2]$; Incorporate science

Bayes:

$$[X, \theta|Y] \propto [Y|X, \theta][X|\theta][\theta]$$

Hierarchical Models (cont.)

Specifically, consider a *hierarchical state-space* framework: Data ($\mathbf{y}_t$), Process ($\mathbf{x}_t$), Parameters $\boldsymbol{\theta}$

Data Stage:

$$\mathbf{y}_t = \mathcal{H}(\mathbf{x}_t, \boldsymbol{\theta}_h)$$

Process Stage:

$$\mathbf{x}_t = \mathcal{M}(\mathbf{x}_{t-1}, \boldsymbol{\theta}_m)$$

Parameter Stage:

$$[\boldsymbol{\theta}_h, \boldsymbol{\theta}_m]$$

Bayesian framework provides a solution:

$$p(\mathbf{x}_{0:t}, \boldsymbol{\theta}|\mathbf{y}_{1:t}) \propto p(\mathbf{y}_{1:t}|\mathbf{x}_{0:t}, \boldsymbol{\theta}_h)p(\mathbf{x}_{0:t}|\boldsymbol{\theta}_m)p(\boldsymbol{\theta})$$

where $\mathbf{x}_{0:t} \equiv \{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_t\}$, $\mathbf{y}_{1:t} \equiv \{\mathbf{y}_1, \ldots, \mathbf{y}_t\}$, $\boldsymbol{\theta} \equiv \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_m\}$.

Hierarchical Models (cont.)

Many examples over the last few years of statisticians and geoscientists using this framework for "real" (although not necessarily "operational") problems: e.g.,

- winds (tropical Pacific, subpolar, coastal, sea breeze)
- Pacific sea surface temperature forecasts
- hurricane/tornado climatology
- air pollution
- glacial dynamics
- radar nowcasting

Most (if not all) of these examples considered various degrees of pre-existing scientific knowledge to inform the models (at the data, process, and/or parameter stages) and MCMC implementation.

Such "subjective" prior information is critical.

Hierarchical Models (cont.)

In general, a powerful paradigm, but practical limitations!

Fundamental Questions:

- To what extent do we know $\mathcal{H}$, $\mathcal{M}$, $\boldsymbol{\theta}_h$, $\boldsymbol{\theta}_m$?
  - $\mathcal{M}$ "known" - Deterministic model/data fusion
  - $\mathcal{M}$ "unknown" - e.g., short time-scale or long-lead forecasts in atmospheric science; one must estimate the evolution operator from the data and prior knowledge
  - $\mathcal{H}$ (measurement function) usually assumed known
  - $\boldsymbol{\theta}$ typically, only variance/covariance parameters are "unknown" (if any)
- What is the dimensionality of $\mathbf{y}_t$, $\mathbf{x}_t$, $\boldsymbol{\theta}$?
  - Makes a huge difference in how we implement a solution; approximations and "adhockery" are rampant!!

Model/Data Fusion or Data Assimilation (DA)

The "traditional" DA problem:

- $\mathcal{H}$ and $\mathcal{M}$ nonlinear and known
- state process covers multiple scales of variability
- many observations, yet sparse relative to state dimension
- model error and biases

NOTE: as a special case of the hierarchical Bayesian paradigm (with known parameters), the solution can be found from Bayes rule: $p(\mathbf{X}|\mathbf{Y})$.

In case of linear model operators $\mathcal{H}$, $\mathcal{M}$, and Gaussian error distributions, this can be derived analytically and gives the traditional optimal interpolation (i.e., kriging) formulas. Variational solution just the posterior mode (but no measure of uncertainty).

Sequential DA

Forecast Distribution:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})d\mathbf{x}_{t-1}$$

Analysis Distribution:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \propto p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$$

- Linear, Gaussian case: traditional Kalman Filter (KF).
- Linearized operators, Gaussian errors: Extended KF
- Monte Carlo sampling form evolution distribution, with added assumptions of linear measurement operator, Gaussian error distributions, and sample-based covariance estimates: Ensemble Kalman filter (EnKF).
- Monte Carlo sampling from the model evolution distribution, reweighted according to the data model (likelihood): sequential importance sampling (particle filtering)

Sequential DA (cont.)

Some Statistical Issues with EnKF

- When is it sufficient to just consider the first two moments?
- Can modifications to sequential importance sampling contribute? e.g., controlling degeneracy through data distribution factorization (e.g., Mark Berliner's talk at SAMSI DA workshop yesterday)
- How large does ensemble have to be?
- Parameter estimation? (can be done with sequential importance sampling; what about EnKF?). NOTE: state augmentation not always appropriate (high dimensional parameter space; significant prior information for parameters). Hybrid MC and MCMC methods?
- In general: add statistical rigor to implementations that seem to work well but have numerous approximations and ad hoc modifications.

General DA Statistical Issues

- **Parameter estimation** in large numerical models.

  [Work in progress: formulated stochastic convective initiation scheme in a regional climate model (MM5); used radar reflectivities as "data" and Bayes to obtain posterior probability distributions for the parameters]

- **Model uncertainty:** fundamental problem that model and "real world" do not live on the same attractor (see Z. Toth and L. Smith's talks at Oct. 5 SAMSI DA meeting); can statistical methods help???

- **Non-additive errors**

  [Work in other disciplines (e.g., Ecology) consider observational processes that have explicitly non-Gaussian error structures; e.g., species abundance with sampling uncertainty, and spatio-temporal dynamics from reaction diffusion equations with stochastic parameters; fully Bayesian implementation via MCMC.]

Dominant Statistical Issues

There are many statistical issues in both the general (hierarchical) modeling framework, and specific (data assimilation) problems.

Dimension Reduction

A fundamental problem in many atmospheric/ocean applications is the *curse of dimensionality* (i.e., sample size needed to estimate a function of several variables to a specified degree of accuracy grows exponentially as the number of variables increases).

This affects model/data fusion (DA) but also modeling processes for which explicit dynamics are less well-known. That is, when one must estimate the evolution operator (e.g., short term nowcasting, seasonal prediction)

Dimension Reduction (cont.)

Consider the data stage:

$$\mathbf{y}_t \longrightarrow \mathbf{x}_t$$

where $Dim(\mathbf{y}_t) >> Dim(\mathbf{x}_t)$.

Some Approaches:

- Non-probabilistic methods
  - Principal component analysis (EOFs), Wavelet's with thresholding, Projection pursuit, Principal curves, Vector quantization (self-organizing maps, elastic nets), Multidimensional scaling, Locally linear embedding
- Probabilistic methods
  - Factor analysis, Independent component analysis, Independent factor analysis, Generative topographic mapping

Dimension Reduction (cont.)

Probabilistic methods can all be characterized as some form of **latent variable model**. This framework (aka hidden process models, mixture models, hidden-Markov models, etc.) is equivalent to the hierarchical model we have already considered:

Let $\mathbf{y}_t$ be the high-dimensional observation vector, and $\mathbf{x}_t$ the underlying latent (unobservable) process of interest. As before, we specify a distribution for the data given this process $p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}_h)$: e.g.,

$$\mathbf{y}_t = \mathcal{H}(\mathbf{x}_t; \boldsymbol{\theta}_h)$$

where $\mathcal{H}$ is the map that controls the dimension reduction.

Most of the methods listed previously have not been used in a modeling sense for geophysical data. Need for estimation approaches for these in the context of state-space models - rich collaborative possibilities.

Multiscale Processes

Multiscale processes are ubiquitous in the natural world (e.g, the transfer of energy across scales in the atmosphere.)

- Data models: Data at different scales; the so-called "change of support" problem in spatial statistics. Exploration of these ideas for data sets of varying resolutions. (e.g., Wikle and Berliner 2005)

$$[Y_a, Y_c | X_b, \theta_{a,c|b}]$$

- Process models: link processes at various scales; e.g.,

$$[X_a | X_b, \theta_a][X_b | X_c, \theta_b][X_c, \theta_c]$$

  – graphical models, multigrid methods

- Parameter models: include scaling relationships (e.g., turbulent scaling relationships in variances)

Example: Multiresolution Model with Latent Process Dynamics

$$\mathbf{y}_t = \mathrm{W}\boldsymbol{\alpha}_t + \boldsymbol{\xi}_t$$

where W is a multiresolution operator and $\boldsymbol{\alpha}_t$ are the time-varying multiresolution coefficients.
Each resolution of $\boldsymbol{\alpha}_t$ is conditioned on an underlying dynamical process, $\mathbf{b}_t$:

$$\boldsymbol{\alpha}_t^j = \phi^j(\mathbf{b}_t; \boldsymbol{\theta}_\phi^j)$$

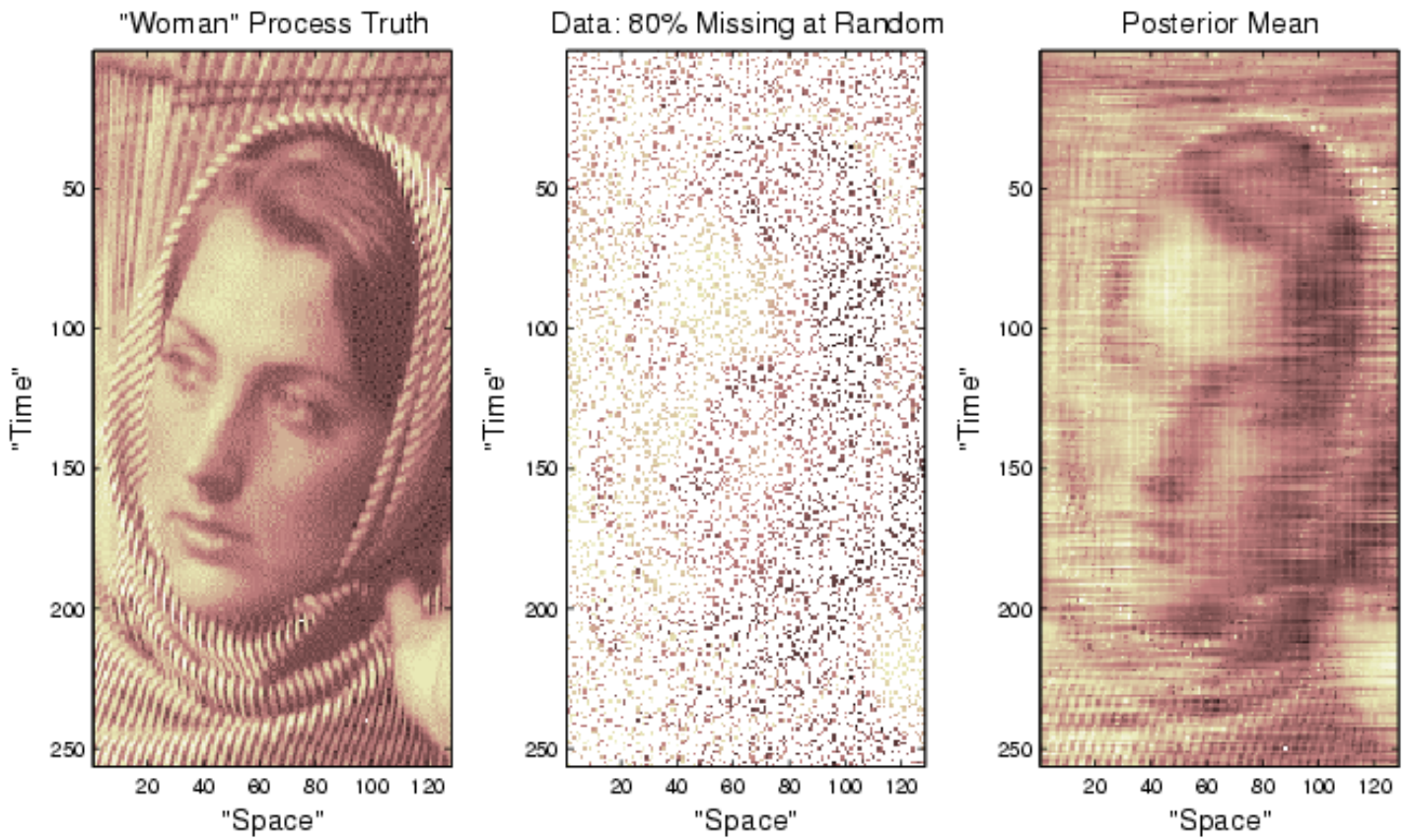
for each multiresolution scale, $j = 0, \ldots, J$.

$$\mathbf{b}_t = \mathcal{M}(\mathbf{b}_{t-1}; \boldsymbol{\theta}_m)$$

- $\phi(\ )$ maps the n_j-dimensional subprocess $\boldsymbol{\alpha}_t^j$ to the p-dim process $\mathbf{b}_t$ ($p \ll \Sigma_j\, n_j$) **[KEY]**

Multiresolution Hidden Process Model: Example

- Model with linear ϕ, $\mathcal{M}$, and Gaussian errors



Multiple Process Modeling

It is critical to link (or couple) multiple processes in geophysical systems (e.g., atmosphere/ocean)

This can be done hierarchically (e.g., Berliner, Milliff, Wikle, *Journal of Geophysical Research: Oceans*, 2003)

- Couple models of interacting spatio-temporal processes (atmosphere and ocean)
  - Hierarchical coupling of complicated systems; each of which is also modeled hierarchically
  - Use approximate dynamics; physical-statistical models
- Incorporate diverse datasets
- Include stochastic elements to adjust for model uncertainty, unmodeled components, etc.
- Quantify uncertainty in each phase

Coupled Atmosphere-Ocean Model

Data

* D_a Atmospheric data (scatterometer)
* D_o Ocean data (altimeter)

HBM Skeleton

1. $[D_a, D_o|\text{Atm}, \text{Ocean}, \theta_a, \theta_o]$
2. $[\text{Atm}, \text{Ocean}|\eta_a, \eta_{o|a}]$
3. $[\theta_a, \theta_o, \eta_a, \eta_{o|a}]$

Parameters

$\theta_a, \theta_o, \eta_a, \eta_{o|a}$

Multiprocess Model Example (cont.)

BHM Keys!

1. $[D_a, D_o|\text{Atm}, \text{Ocean}, \theta_a, \theta_o] = [D_a|\text{Atm}, \theta_a][D_o|\text{Ocean}, \theta_o]$

2. $[\text{Atm}, \text{Ocean}|\eta_a, \eta_{o|a}] = [\text{Ocean}|\text{Atm}, \eta_{o|a}][\text{Atm}|\eta_a]$

1. Atm & Ocean data are conditionally independent

2. Parameterized air-sea model is stochastic atmospheric model coupled to stochastic Ocean-given-Atmosphere model

* Posterior: $[\text{Atm}, \text{Ocean}|D_a, D_o]$

Implementation: MCMC for atmospheric model; Importance Sampling MC to link atmosphere and ocean (stochastic coupling)

Other Statistical Issues

- **Field Comparison:**

  Atmospheric scientists have a long history of at least considering comparisons (verifications) with recognition of the process behavior.

  Statisticians have very few good approaches for comparing fields in space and time, especially in a distributional (ensemble) context.

- **Extremes:**

  Of increasing interest due to the concern that anthropogenic changes in the environment will be manifest as changes in extremes.

  Extreme value theory is just now starting to be applied to spatial and spatio-temporal problems in statistics. Very little has been done in complicated non-linear dynamical systems. This is an area that would benefit from strong collaboration with applied mathematicians and dynamicists.

Computation

- Model applicability critically tied to computational practicality
- Related to issues of model/data fusion, sequential estimation, dimensionality, multiscaling, multi-process modeling
- Statisticians, as a group, are behind: we don't use parallel codes, and we are not very good at handling huge data sets!
- Subject matter scientists are wedded to certain numerical approaches and are often not willing to "retool".
- It is time to start considering computational approaches that may benefit both deterministic "rules" and statistical parameterizations in massively parallel environments!

Conclusion

- There are numerous sources of uncertainty in problems in the geosciences.
- It is essential that statisticians play a role. But, they must be willing to learn the science and consider new approaches!
- Similarly, geoscientists must be willing to consider new approaches to handling uncertainty.
- We are in a second "golden age" of statistics in which scientific problems are providing the motivation for new methodologies!
- **True collaboration is essential!**