Statistical Approaches to Combining Models and Observations

Mark Berliner
Department of Statistics
The Ohio State University
mb@stat.osu.edu

SIAM UQ12, April 5, 2012

Outline

I. Bayesian Hierarchical Modeling
II. Examples: Glacial dynamics; Paleoclimate
III. Using Large-scale Models
IV. Example: Ocean Forecasting
I. Bayesian Hierarchical Modeling (BHM)

- Bayesian Analysis:
  1. UQ via prob. modeling;
  2. update via Bayes’ Theorem;
  3. infer, predict & make decisions via prob. (“risk”) analysis

- Hier. Model: Sequence of conditional probability distributions (corresponding to a joint distribution)

- Framework for modeling:
  Observations $Y$  Processes (State Variables) $X$  Parameters $\theta$

  1. **Data Model**  $[Y \mid X, \theta]$
  2. **Prior Process Model**  $[X \mid \theta]$
  3. **Prior Parameter Model**  $[\theta]$

- Bayes’ Theorem:  $[X, \theta \mid Y]$
Strategies

1. Incorporating physical models
   Physical-statistical modeling (Berliner 2003 JGR)
   From “F=ma” to process model [ X | θ ]

2. Qualitative use of theory
   (e.g., Pacific SST model (Berliner et al. 2000 J. Climate)
   ******************************************

3. Incorporating large-scale computer model output

4. Combinations

5. Alternate Approaches
● Steady Flow of Glaciers and Ice Sheets
  – Flow: gravity moderated by drag (base & sides) & stuff
  – Simple models: flow from geometry

● Data
  – Program for Arctic Climate Regional Assessments
  – Radarsat Antarctic Mapping Project
    * surface topography (laser altimetry)
    * basal topography (radar altimetry)
    * velocity data (interferometry)
North East Ice Stream, Greenland
Physical Modeling: Surface: s, Thickness: H, Velocity: u

- Basal Stress: $\tau = -\rho g H \frac{ds}{dx}$ ( + ”stuff”)
- Velocities: $u = u_b + b_0 H \tau^n$ where $u_b = k \tau^p + (\rho g H)^{-q}$

Our Model

- Random Basal Stress: $\tau = -\rho g H \frac{ds}{dx} + \eta$
  where $\eta$ is a “corrector process” (“model error”)
- Random H: wavelet smoothing prior
- Random s: parameterized model from literature
- Random Velocities: $u = u_b + b H \tau^n + e$
  where $u_b = k \tau^p + (\rho g H)^{-q}$ or a constant (change-point)
  $b$ is unknown parameter, $e$ is a noise process
- Process Model: $[u \mid s, H, \eta \mid \eta \mid s, H]$

- Use of proxies:
  - Inverse problem:
    \[ \text{proxy} \approx f(\text{climate}) \rightarrow \text{climate} \approx g(\text{proxy}) \]
  - Inverse probability problem:
    \[ [\text{proxy data} \mid \text{climate}] \rightarrow [\text{climate} \mid \text{proxy data}] \]

- Boreholes: earth stores info about surface temp’s
  - Inverse: borehole data \( \approx f(\text{surface temp’s}) \)
  - Model:
    * Heat equation
    * Infer boundary condition
Data Model: $\vec{Y} | \vec{T}_r, q \sim N(\vec{T}_r + T_0 \vec{I} + q \vec{R}(k), \sigma^2_y I)$

- $\vec{T}_r$: reduced temperatures (literature: $\vec{T}_r = \vec{T} - T_0 \vec{I} - q \vec{R}(k)$)
- $T_0$: reference surface temperature
- $q$: surface heat flow
- $\vec{R}$: thermal resistances ($\propto k^{-1}$, thermal conductivities, adjusted for rock types, etc.)

Process Model:

- Heat equation applied to $\vec{T}_r$
- B.C.: Surface temp history $\vec{T}_h$ (assumed piecewise constant)
- Easy to solve the heat equation

$$\vec{T}_r | \vec{T}_h, q \sim N(A \vec{T}_h, \sigma^2 I)$$

$$\vec{T}_h | q \sim N(0, \sigma^2_h I)$$

Parameter Model: next
Spatial Hierarchy

- Nine boreholes: 5 in desert region, 4 in swell region
- Extend the hierarchy: for $j = 1, \ldots, 9$

**Data Model:**
\[
\tilde{Y}_j \mid \mathbf{T}_{rj}, q_j \sim N(\mathbf{T}_{rj} + T_{0j} \mathbf{1} + q_j \mathbf{R}_j(k_j), \sigma_{yj}^2 \mathbf{I})
\]

**Process Model:**
\[
\tilde{T}_{rj} \mid \tilde{T}_{hj}, q_j \sim N(A_j \tilde{T}_{hj}, \sigma_j^2 \mathbf{I})
\]
\[
\tilde{T}_{hj} \mid q_j \sim N(\bar{0}, \sigma_{hj}^2 \mathbf{I})
\]

**Parameter Model**
- $\tilde{T}_{h1}, \ldots, \tilde{T}_{h5}$ conditionally independent $N(\bar{\mu}_d, \gamma_d^2 \mathbf{I})$
- $\tilde{T}_{h6}, \ldots, \tilde{T}_{h9}$ conditionally independent $N(\bar{\mu}_s, \gamma_s^2 \mathbf{I})$
- $\bar{\mu}_d$ & $\bar{\mu}_s \sim N(\bar{\mu}_0, \gamma_0^2 \mathbf{I})$
- $q_1, \ldots, q_5$ conditionally independent $N(\nu_d, \eta_d^2)$
- $q_6, \ldots, q_9$ conditionally independent $N(\nu_s, \eta_s^2)$
- etc.
SRD-2, SR Desert

SRS-5, SR Swell
III. Incorporating Large-Scale Computer Models

Ensembles $\tilde{O} = O_1 = M(T_1), \ldots, O_n = M(T_n)$

(T’s include “controls, model names, etc.)

- **Data Model** Treat $\tilde{O}$ as ”observations”
  - $[Y, \tilde{O} | X, \theta]$ (include ”bias, offset, model error ..”)
  - Convenient for design of collection of $Y, \tilde{O}$

- **Process Model** Use $\tilde{O}$ to develop $[X | \theta]$
  - Kernel density estimate
    $$\Sigma_i \alpha_i k(x | O_i)$$
  - Gaussian process models; emulators; UQ

Model Output: $\tilde{O} = (O_1 = M(T_1), \ldots, O_n = M(T_n))$

$$[O | \theta, \tilde{O}] \rightarrow [X | \theta]$$

- **Parameter Model** from model output
  (eg: Berliner, Levine, & Shea 2003 *J. Climate*)

- **Combinations**
• Themes
  – “climate” = parameters of prob. dist. of “weather”
  – build or “parameterize” scales into dynamic model for X
• Future climate depends on future, but unknown, inputs.
• IPCC: construct plausible future inputs, ”SRES Scenarios”
  (CO$_2$ etc.)
• Assume a scenario and find corresponding projection
Hemispheric Monthly Surface Temperatures (X)

- Two models ($\mathbf{\bar{O}}$): PCM ($n=4$), CCSM ($n=1$) for 2002-2197
- 3 SRES scenarios (B1,A1B,A2).
Hierarchical Data Model for Model Output

- Scalar climate variable $X$; $m = 1, \ldots, M$ models (time fixed)
- $\tilde{O}_m$: ensemble of size $n_m$ of estimates of $X$ from model $m$.

1. Given means $\mu_m$ and variances $\sigma_{Ym}^2$, $m = 1, \ldots, M$;
   $\tilde{O}_m$ are independent and
   \[ \tilde{O}_m \mid \mu_m \sim \text{Gau}(\mu_m \mathbf{1}_{n_m}, \sigma_{Ym}^2 \mathbf{1}_{n_m}) \]

2. Given $\beta$, model biases $b_m$ and variances $\sigma_{\mu_m}^2$;
   $\mu_m$ are independent and
   \[ \mu_m \mid \beta, b_m \sim \text{Gau}(\beta + b_m, \sigma_{\mu_m}^2) \]

3. Given $X$,
   \[ \beta \mid X \sim \text{Gau}(X, \sigma_{\beta}^2) \text{ and } b_m \mid X \sim \text{Gau}(b_{0m}, \sigma_{b_m}^2) \]
Remarks

• Implied marginal dist.: “$\tilde{O}$ given $X$”: Integrating out $\beta$ induces dependence within & across ensembles

• Modify intuition about value of increasing ensemble size: “Infinite” ensembles do not give perfect forecasts: If all biases $= 0$, infinite ensembles give the value of $\beta$, not $X$

• Extensions to different model classes (more $\beta$’s) and richer models are feasible.
Model Overview

1. \([Y \mid X, \theta]\): measurement error model
   Gaussian with mean = true temp.
   & unknown variance (with a change-point)

2. \([X \mid \theta]\): Time series models with time varying parameters
   \[
   X_t = \mu_{i(t)} + \begin{pmatrix} \eta^n_j(t) \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ \eta^s_j(t) \end{pmatrix} (X_{t-1} - \mu_{i(t-1)}) + e_t
   \]

3. \([\theta]\) Climate = parameters of distribution of weather
   - Climate-weather: multiscale phenomena
   - Time evolution: \(\mu_{i(t)}\) slow; \(\eta_j(t)\) moderate;
   \(e_t\) fast, but variances of \(e_t\) are slow
   - \(\mu_i = A + B \ CO_2i + \text{noise}\)
   - Obs period: \(\eta_j = C + D \ |\text{SOI}|_j + \text{noise}\)
     Fore. period: AR model (i.e., SOI not observed)
   - Variances of \(e_t\): AR-like prior
Model Adapted to Decadal Forecasting
(Kim & Berliner 2012)
IV. Combining Approaches: Mediterranean Ocean Forecasting

1. Winds as a boundary condition for the ocean surface
   (Milliff et al. & Bonazzi et al. 2011 Quart. J. Roy. Met. Soc.)

2. Bayesian Multi-model Ensembling for Ocean Forecasting
   (Berliner et al. 2012)
Bayesian Hierarchical Models to Augment The Mediterranean Forecast System (MFS)

Nadia Pinardi  
INGV (I'istituto Nazionale di Geofisica e Vulcanologia)  
Univ. Bologna (MFS Director)

Alessandro Bonazzi, Srdjan Dobricic  
INGV, Univ. Bologna

Ralph Milliff  
CoRA

Chris Wikle  
Univ. Missouri

Mark Berliner  
Ohio State Univ.

Revolutionary Research . . . Relevant Results
Overview:

- MFS: deterministic operational forecasting system
- Boundary condition/forcing: Surface Winds
- Bayesian Hierarchical Model (BHM) to quantify Surface Vector Wind (SVW) distributions
- Ocean Ensemble Forecasting (OEF) using 10 member BHM-SVW ensembles

Key: Exploit abundant, “good” satellite wind data combined with physical modeling
Building wind dist. (BHM-SVW)

1. **Data Stage**
   - Satellite (QSCAT) and Numerical Weather Pred. Analyses (ECMWF)
2. **Process Model**: Rayleigh Friction Model  
(Linear Planetary Boundary Layer Equations)  

**Theory**  
\[
\begin{align*}
\frac{\partial u}{\partial t} - f v &= -\frac{1}{\rho} \frac{\partial p}{\partial x} - \gamma u \\
\frac{\partial v}{\partial t} + f v &= -\frac{1}{\rho} \frac{\partial p}{\partial y} - \gamma v
\end{align*}
\]  
(neglect second order time derivative)  

**discretize:**  
\[
V_t = \left[ 1 + \frac{2\gamma}{f^2\Delta} + \frac{\gamma^2}{f^2} \right]^{-1} 
\left[ \left( \frac{2\gamma}{f^2\Delta} \right) V_{t-1} + \left( \frac{1}{f^2} \right) D_x P_t + \left( -\frac{1}{f^2\Delta} - \frac{\gamma}{f^2} \right) D_y P_t + \left( \frac{1}{f^2\Delta} \right) D_y P_{t-1} \right]
\]
\[
U_t = \left[ 1 + \frac{2\gamma}{f^2\Delta} + \frac{\gamma^2}{f^2} \right]^{-1} 
\left[ \left( \frac{2\gamma}{f^2\Delta} \right) U_{t-1} + \left( -\frac{1}{f^2} \right) D_y P_t + \left( -\frac{1}{f^2\Delta} - \frac{\gamma}{f^2} \right) D_x P_t + \left( \frac{1}{f^2\Delta} \right) D_x P_{t-1} \right]
\]

**Our model**  
\[
V_t = -L_v|_{v(1)} V_{t-1} + c_v|_{p_x} D_x P_t + c_v|_{p_y} D_y P_t + c_v|_{p_y(1)} D_y P_{t-1} + \epsilon
\]
\[
U_t = -L_u|_{u(1)} U_{t-1} - c_u|_{p_y} D_y P_t + c_u|_{p_x} D_x P_t + c_u|_{p_x(1)} D_x P_{t-1} + \epsilon
\]
BHM Ensemble Winds

10 members selected from the Posterior Distribution (blue)

Ensemble mean wind (green); ECMWF Analysis wind (red)
BHM-SVW-OEF initial condition spread:

Sea Surface Temperature
Initial condition spread

Sea Surface Height
Initial condition spread

Uncertainty is concentrated at mesoscales
BHM-SVW-OEF 10 day forecast spread

Initial condition ensemble spread has amplified at the 10 day fcst in mesoscales
The forecast spread at 10 days

EEPS forced ensemble

BHM-SVW ensemble

ECMWF EPS forcing
Ineffective at producing flow field changes at mesoscales
Discussion:

• BHM methods produce realistic distributions of surface winds (SVW) (Milliff et al. 2011 J Roy Met Soc)
• BHM-SVW results used to in a new ocean ensemble forecasting method: BHM-SVW-OEF (Bonazzi et al 2011 J Roy Met Soc)
• The BHM-SVW-OEF produces 10-day-forecast spreads at mesoscales and in the upper thermocline
2. Bayesian Multi-model Ensembling for Ocean Forecasting (Berliner et al. 2012)

- profiles of wintertime (60 days) of temperature $T(z,t)$ (& salinity) (Rhodes Gyre region in Eastern Mediterranean Sea)
- 16 vertical levels $z = 0 \text{ m}$ to $z = 300 \text{ m}$
- BHM Winds produce ensembles from two models:
  1) Ocean PArallelise (OPA)
  2) Nucleus for European Modeling of the Ocean (NEMO)
- Model as in climate example: model-specific biases
- Prior: Analyzed fields