Is Data Assimilation Relevant to Climate Research?

Christopher Jones
Warwick Mathematics Institute,
University of Warwick, and
Department of Mathematics,
University of North Carolina at Chapel Hill

Newton Institute, Cambridge: August, 2010
OBS

DATA

PHYSICAL LAWS

MODEL

DA

OPTIMAL STATE ESTIMATE OR PREDICTION
Model:

\( x \in \mathbb{R}^N \) - state vector comprising all relevant dynamical variables (e.g. flow velocity, temperature, salinity, etc. at each grid point)

\[
dx^f = M(x^f, t)dt
\]

- prognostic model

\[
dx^i = M(x^i, t)dt + \sigma(t)dt
\]

- actual evolution

\[
E[\sigma(t)\sigma^T(t')] = \delta(t - t')Q(t)
\]

- covariance of the model residual

Observations:

\[
\eta^o_i = H_i[x^t_i] + \varepsilon_i
\]

\( H_i \) - observation operator

\( \varepsilon_i \) - observation error

\[
\eta^o_i \in \mathbb{R}^L, \text{ typically } L \ll N
\]

\[
E[\varepsilon_i\varepsilon^T_m] = \delta_{im}R_i
\]

- obs error covariance
State estimate:

\[ x_1^a(t_1), x_2^a(t_1), P^a(t_1) \]

Model forecast:

\[ x_1^f(t_1), x_2^f(t_1), P^f(t_1) \]

Initial conditions:

\[ x_1^f(t_0), x_2^f(t_0), P^f(t_0) \]

Measurement:

\[ \eta(t_1) = x_1^f(t_1) + \varepsilon \]

Gain Matrix

\[ x^a = x^f + K(\eta - H(x^f)) \]
Data Assimilation in Predictive Mode

Model + observations \[ \xrightarrow{\text{prediction}} \]

\[ \begin{array}{c}
t_0 \\
\text{model} \\
t_1 \\
\text{model} \\
t_2 \\
\text{update} \\
... \\
t_N \\
\end{array} \]

Interpolation at \( t = t_i \):

\[ x^a_i = x^f_i + K_i \left( \eta_i - H(x^f_i) \right) \]

Sequential DA/Forward Problem/Filtering
\[ x^a = x^f + K \left( \eta - H(x^f) \right) \]

Where the matrix \( K \) is determined by the requirement that \( x^a \) minimise the cost function:

\[
J(x) = \left\langle x - x^f, (P^f)^{-1} (x - x^f) \right\rangle + \left\langle \eta - H(x^f), R^{-1} (\eta - H(x^f)) \right\rangle
\]

\[
K = P^f H^T \left( R + H P^f H^T \right)^{-1}
\]

\( H \) (linearised) observation operator
\( R \) observation error covariance matrix
\( P^f \) determined by DA scheme being used!
\[ x^a = x^f + K \left( \eta - H(x^f) \right) \]

\[
K = P^f H^T \left( R + HP^f H^T \right)^{-1}
\]

Possibilities for calculating \( P^f \)

1. \( P^f \) is constant background error covariance matrix (Optimal Interpolation=OI)
2. If model is linear, \( P^f \) is evolved under model (Kalman Filter=KF)
3. If model is nonlinear, \( P^f \) is evolved under tangent linear model (Extended Kalman Filter=EKF)
4. \( P^f \) is built out of ensembles evolved under full nonlinear model (Ensemble Kalman Filter=EnKF)
Ensemble Kalman Filter (EnKF)

Error covariance is predicted via solution of full nonlinear system for a Monte-Carlo ensemble of states

Initial conditions:
\[ x^i_{1,j}(t_0), x^i_{2,j}(t_0), \quad j = 1, N_E \]

Model forecast:
\[
\begin{align*}
  x^f_{1,j}(t_1) &= \frac{1}{N_E} \sum_{j=1}^{N_E} x^i_{1,j}(t_1), \\
  x^f_{2,j}(t_1) &= \frac{1}{N_E} \sum_{j=1}^{N_E} x^i_{2,j}(t_1)
\end{align*}
\]

\[
P^f = \frac{1}{N_E - 1} \sum_{j=1}^{N_E} \left( x^f_j - x^f, x^f_j - x^f \right)
\]
DA in State Estimation Mode

Model run + observations \( t=t_1,\ldots,t_N \) \( \rightarrow \) state estimate

\[ x_0 \rightarrow t_0 \text{ model} \rightarrow t_1 \text{ model} \rightarrow t_2 \text{ model} \rightarrow t_3 \rightarrow \ldots \rightarrow t_N \]

\( \eta(t_0) \rightarrow \eta(t_1) \rightarrow \eta(t_2) \rightarrow \eta(t_3) \rightarrow \ldots \rightarrow \eta(t_N) \)

\( x(t_N) \)

4DVAR: Minimise the cost function:

\[
J(x) = \left\langle x_0 - x_0^*, B^{-1} \left( x_0 - x_0^* \right) \right\rangle + \sum_{j=1}^{N} \left\langle \eta_j - H(x(t_j)), R_j^{-1} \left( \eta_j - H(x(t_j)) \right) \right\rangle
\]

\( x_0^* = \text{estimate} \quad B = \text{background error covariance matrix} \)

Variational DA/Inverse Problem/Smoothing
Fishkill in Lake Kinneret

Vernieres et al. (2006)

Conjecture: due to “lifting” of lower layer of oxygen-free water

• Occasional “fishkill”
• Feeding of 5,000??
Model:

- Stably stratified during summer
- Strong westerly sea breeze

\[
\begin{align*}
\frac{du^{(1)}}{dt} - fv^{(1)} + g \frac{\partial}{\partial x} \left( h^{(1)} + h^{(2)} + D \right) - A_n \nabla^2 u^{(1)} - F_u &= 0 \\
\frac{dv^{(1)}}{dt} + fu^{(1)} + g \frac{\partial}{\partial y} \left( h^{(1)} + h^{(2)} + D \right) - A_n \nabla^2 v^{(1)} - F_v &= 0 \\
\frac{dh^{(1)}}{dt} &= 0 \\
\frac{du^{(2)}}{dt} - fv^{(2)} + g \frac{\partial}{\partial x} \left( h^{(1)} + h^{(2)} + D \right) + g \frac{\partial}{\partial y} \left( h^{(2)} + D \right) - A_n \nabla^2 u^{(1)} &= 0 \\
\frac{dv^{(2)}}{dt} + fu^{(2)} + g \frac{\partial}{\partial y} \left( h^{(1)} + h^{(2)} + D \right) + g \frac{\partial}{\partial y} \left( h^{(2)} + D \right) - A_n \nabla^2 v^{(1)} &= 0 \\
\frac{dh^{(2)}}{dt} &= 0
\end{align*}
\]

\[
\frac{\partial}{\partial x} (\rho u^{(1)} h^{(1)}) + \frac{\partial}{\partial y} (\rho v^{(1)} h^{(1)}) + \frac{\partial}{\partial z} (\rho \frac{\partial h^{(1)}}{\partial z}) + \mathbf{F}^{(1)} = 0 \\
\frac{\partial}{\partial x} (\rho u^{(2)} h^{(2)}) + \frac{\partial}{\partial y} (\rho v^{(2)} h^{(2)}) + \frac{\partial}{\partial z} (\rho \frac{\partial h^{(2)}}{\partial z}) + \mathbf{F}^{(2)} = 0
\]
Model run
Data

• Thermistor chains
With thermistor data assimilated...
A comparison

Neither model nor data on their own show “fishkill,”
But, together, they do!
Climate: is DA relevant?

Weather prediction: DA is strikingly successful

Climate prediction: neither approach is, in an obvious way, useful or even appropriate

Conventional wisdom: climate prediction is NOT an initial value problem, but a forcing and boundary value problem
Antarctic sea ice cover

Challenges

New system complexity
Ultra-high resolution
Data assimilation
Needs of user community
Where are we heading?
Beachhead for DA in Climate

**Paleoclimate:** state estimation DA is becoming recognised as an appropriate and useful tool

**Ocean:** IPCC AR5 focus on decadal predictions—need to get ocean initial condition right! DA can be very useful in this endeavour.

Key issues in both areas: *long* timescales and *sparse* data

Workshop: Data Assimilation and Climate Research, June 2010
Dalton Minimum  1790-1820

(Van der Schrier and Barkmeijer, Climate Dynamics 2005)

- Cold in Western Europe
- Decreased solar activity
- Volcanoes: 1809 and 1815

Bjerknes Hypothesis:

Anomalous ocean-atmosphere interaction

↓

Southerly flow of cold polar waters into NE Atlantic Ocean

↓

Western Europe is cold even into Summer months
Van der Schrier and Barkmeijer introduce a DA technique to assess relationship between anomalous SLP and cold period.
Atmospheric circulation:

QG streamfunction ($m^2 s^{-1}$)

Target used in DA

Difference (DJF) between DA and control runs
Fig. 13 Difference in 2 meter temperature, averaged over 25 years, for the winter (DJF) season (fig. a) and the summer (JJA) season (fig. b) between the data-assimilated and the control run. In both seasons, north-western Europe is under the data-assimilated run colder than in the control run. All temperature changes shown are statistically significant at the 95% level. Figures c and d show the difference in reconstructed surface air temperature for the period 1790-1820 AD with respect to 1971-2000, for the winter (DJF) season (fig. c) and the summer (JJA) season (fig. d). Reconstructed temperatures are from Luterbacher et al. (2004)
Conclusions:
1. Anomalous ocean/atmosphere interaction is directly related to Western European cooling,
2. Independent of possible primary forcing due to solar or volcanic activity
3. MOC weakening is present but not significant

Anomalous wind-stress curl (DJF)  
SST difference between DA and control (DJF)
\[ t = t_0 \]

Model forecast:
\[ x_1^f(t_1), x_2^f(t_1), P^f(t_1) \]

Initial conditions:
\[ x_1^f(t_0), x_2^f(t_0), P^f(t_0) \]

\[ t = t_1 \]

State estimate:
\[ x_1^a(t_1), x_2^a(t_1), P^a(t_1) \]

Measurement:
\[ \eta(t_1) = x_1^f(t_1) + \varepsilon \]

Bayes

Gain Matrix

\[ P_{\text{posterior}}(x|\eta) = P_{x|\eta}^{\text{obs}}(x|x) R_{\text{prior}}^{\text{post}}(\eta - \left( x|f \right)) \]
Forecast step:
\[ p(x, t_0) \rightarrow p(x, t_1) \]
\[ \frac{\partial p}{\partial t} + \frac{\partial (M_ip)}{\partial x_i} = \frac{1}{2} \frac{\partial^2 (Q_{ij}p)}{\partial x_i \partial x_j} \]

Bayes step (update/analysis):
\[ p(x, t_1) \rightarrow p(x, t_1 \mid y^o) \]
\[ p(x, t_1 \mid y^o) = \frac{p(y^o \mid x)p(x, t_1)}{\int p(y^o \mid z)p(z, t_1)dz} \]

But: computationally prohibitive, state \( \approx 10^6 \)
Posterior and Cost Function

Predictive:

\[ P(x | \eta) \propto \exp(-J(x)) = \exp \left( -\left< x - x^f, \left( P^f \right)^{-1} (x - x^f) \right> - \left< \eta - H(x^f), R^{-1} (\eta - H(x^f)) \right> \right) \]

State estimation:

\[ P(x | \eta) \propto \exp(-J(x)) = \exp \left( -\left< x_0 - x_0^*, B^{-1} (x_0 - x_0^*) \right> - \sum_{j=1}^{N} \left< \eta_j - H(x(t_j)), R_j^{-1} (\eta_j - H(x(t_j))) \right> \right) \]

The minimisation gives the posterior mode

log posterior of the Lorenz ’63 system (courtesy of Jochen Voss)

See: Data Assimilation: Mathematical and Statistical Perspectives, Apte, J, Stuart and Voss, IJNMF 2008
Perturbed Cellular Flow Field

\[
\frac{\partial u}{\partial t} = v - \frac{\partial h}{\partial x}, \\
\frac{\partial v}{\partial t} = -u - \frac{\partial h}{\partial y}, \\
\frac{\partial h}{\partial t} = -\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y},
\]

\[
u(x, y, t) = -2\pi l \sin(2\pi kx) \cos(2\pi ly) u_0 + \cos(2\pi my) u_1(t),
\]

\[
\dot{u}_0 = 0,
\]

\[
\dot{u}_1 = v_1,
\]

\[
\dot{v}_1 = -u_1 - 2\pi mh_1,
\]

\[
\dot{h}_1 = 2\pi mv_1,
\]

\[
x = u(x, y, t)
\]

\[
y = v(x, y, t)
\]

Apte, Stuart and J., Tellus A 2008
Assimilating from trajectory staying in one cell

Compare:
- EnKF
- Metropolis-Hastings
Problem with EnKF

After first observation and assimilation

Before second observation

After SECOND observation and assimilation
Climate Models

Primitive eqns of GFD solved on grid

Dynamical core of model

Much of climate science is devoted to setting the forcing terms. This includes:

1. Parametrising sub-gridscale processes
2. Parametrising missing effects
3. Information flow to other parts of model
4. Setting ambient conditions (radiative forcing)
5. Tuning the model
Tuning the Model

Model runs → Model output checked → Resetting of model parameters

Key signatures, e.g.
1. Does it give correct ice thickness?
2. Do currents lie in correct locations?
3. Is tropical precip right?

Accumulated experience ← Expert judgment

OBS → DATA
1. Uncertainty is an intrinsic property of “The” Model, or a collection of models, and it can be found by just looking at them hard enough.

Uncertainty of the model can have its origin in many places: data used in tuning, expert judgment etc. Perhaps we need to unpack the model?

2. Data assimilation’s primary use is to get the “best” estimate.

Data assimilation tells us how to put information together from models and data how the cumulative error is formed. The error may be arbitrarily complex and the issue is how to sample the PDF.
• DA could be used in tuning climate models
• Its use would be similar to the way the paleoclimate work by van der Schrier and Barkmeijer
• If done from a Bayesian perspective, the procedure would have implications for uncertainty due to model tuning.
Challenges

- DA is nowhere near ready to be applicable to climate
- Need to develop DA to deal with
  - High dimensions
  - Nonlinearity
  - Sparse data
  - Multiple timescales (and space)
- Shift perspective from getting “best estimate” of truth to finding PDF that represents all available information