Data Assimilation for Dynamic Systems Models

Discussion of the paper “A Statistical Overview and Perspectives on Data Assimilation for Marine Biogeochemical Models” by Dowd, Jones, Parslow

Computer Experiments Reading Group
November 3, 2015

The Plan

“Data assimilation refers to the quantitative methods by which the information in dynamic models and data are combined to provide estimates of the state and its key parameters”

This discussion will focus specifically on the use of data assimilation for inference and prediction for nonlinear dynamic systems models.

The Model

Interested in inference and prediction for large-scale spatio-temporal models defined implicitly as partial differential equations (PDE).

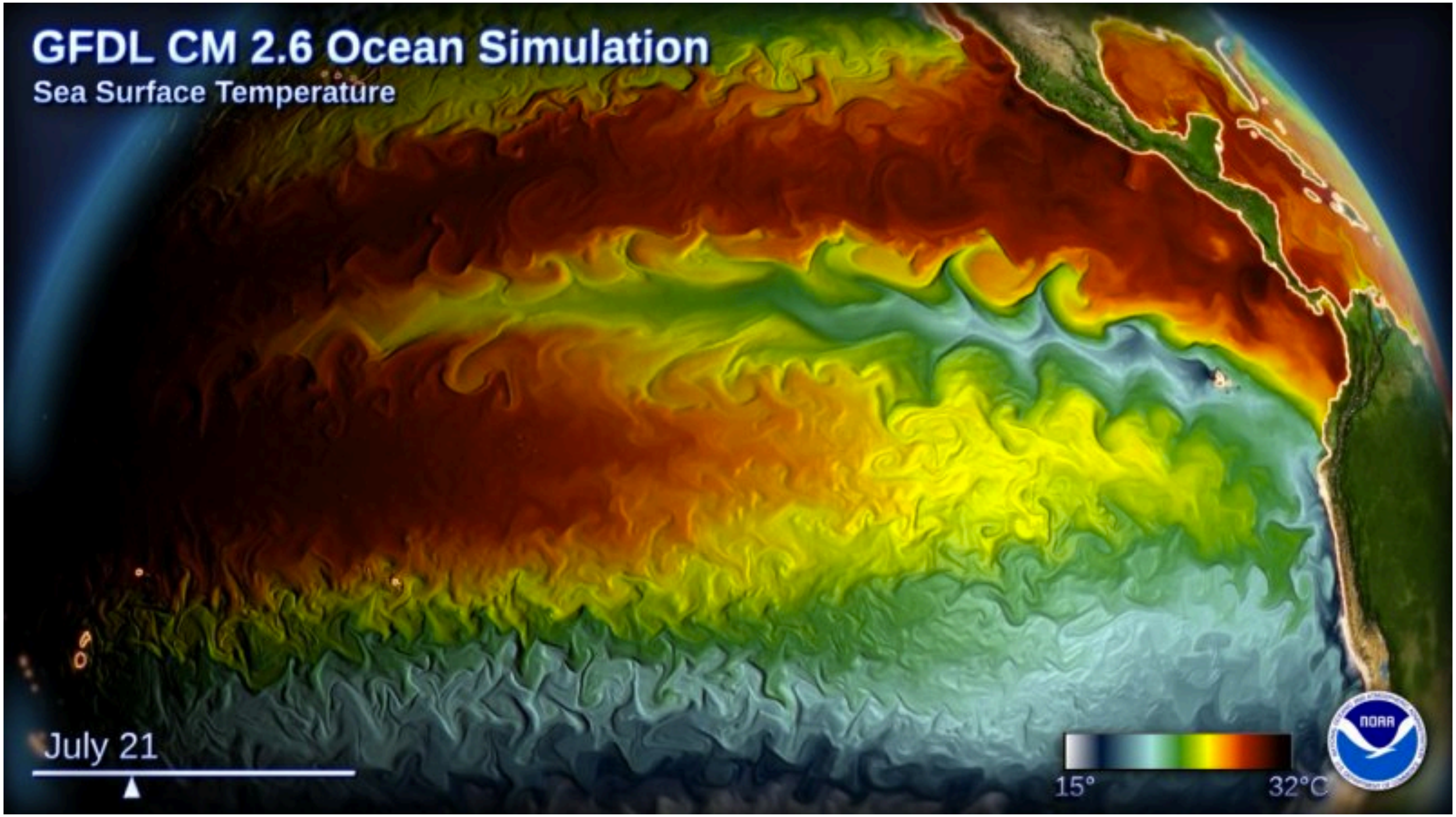
An example from biogeochemistry (BGC):

$$\frac{\partial x^{(i)}}{\partial t} + u \cdot \nabla x^{(i)} - \nabla \cdot \left(\kappa \nabla x^{(i)} \right) = f_i \left(x, \theta, \gamma \right), \quad i = 1, \ldots, m.$$

where

- $x = [x^{(1)}, \cdots, x^{(m)}]^\top$ – 3-d spatial field
- u – 3-d current field
- κ – matrix of diffusion coefficients
- f_i – governing equations (photosynthesis, predation, etc.)
- θ – biological parameters (may be high dimensional)
- γ – forcing fields (light, temperature, etc.)

Global Ocean Circulation Model: SST snapshot



Slide from "Miscellaneous Thoughts on Ocean Data Assimilation" by M. Dowd, presented at the 2015 STATMOS/NCAR Workshop on Data Assimilation

The Model

What we need to keep in mind about these models

Issues:

- The BGC parameters, θ, are uncertain and vary by season etc.
- The BGC governing equations (RHS) are uncertain (they are not perfect models of reality)
- Ocean circulation models (LHS) are uncertain (spatial resolution limitations, etc.)
- Forcing fields, γ, are uncertain
- No closed form expression for the states, x, means we make inference on a discrete approximation of the PDE and not the PDE itself

(Stochastic models?)

The Data

Data, $y_t, t = 1, \ldots, T$ for these applications can be quite complex

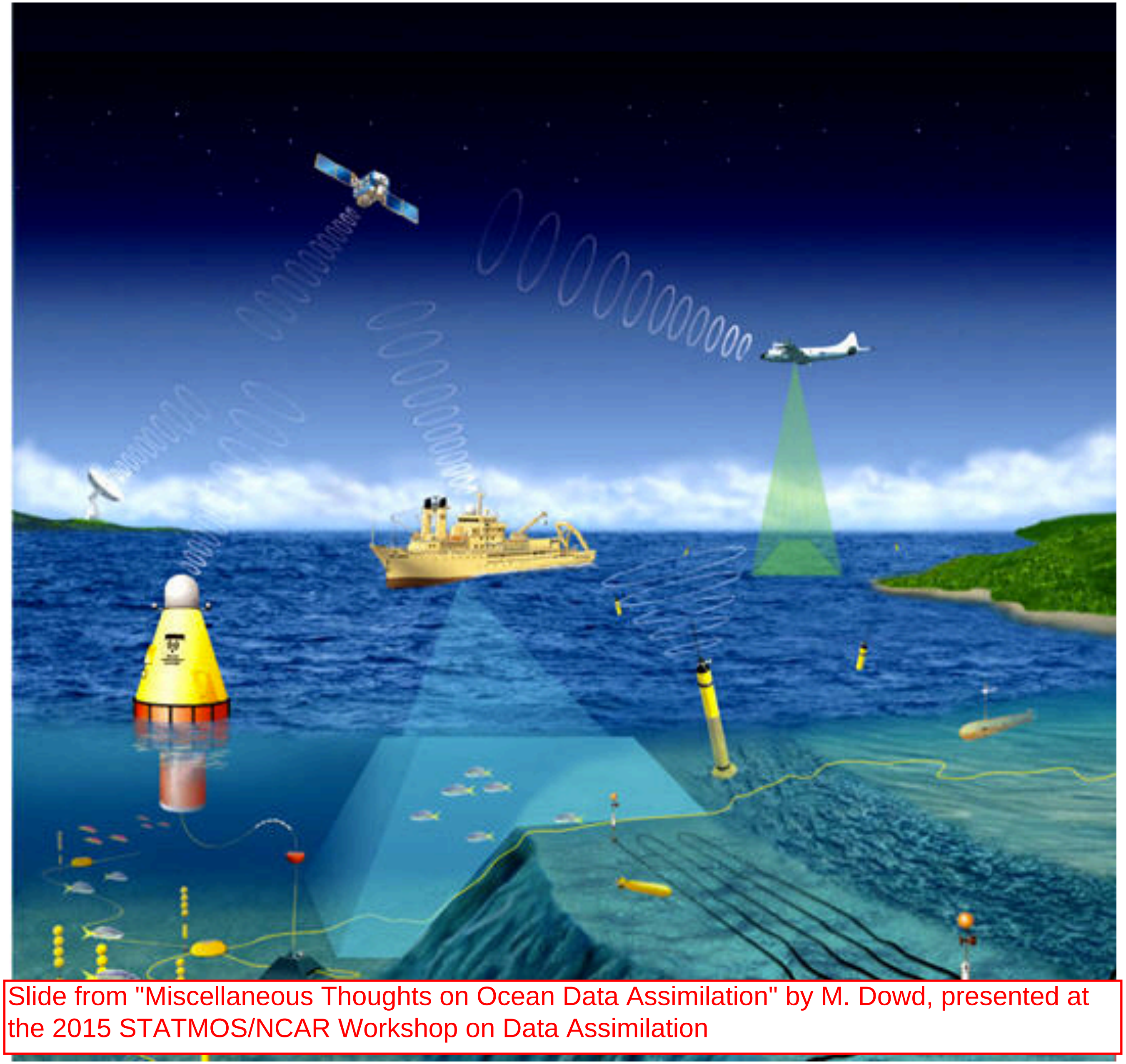
Examples:

- Indirect measurements of x at different locations and water depths (usually systematic sampling)
- Tracer data
- Remote sensing

Issues:

- Uncertain error models $y_t = h(x_t, \theta, \epsilon_t)$
- Indirect observations
- Combining different data types

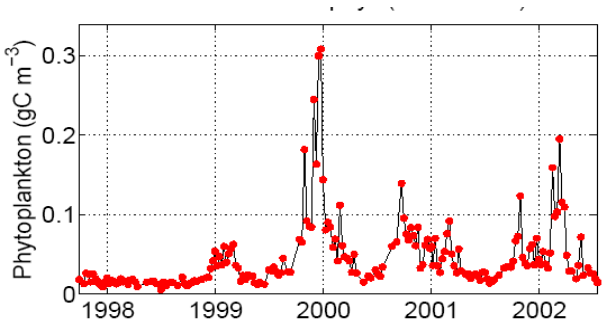
Measurements of the Physical Ocean



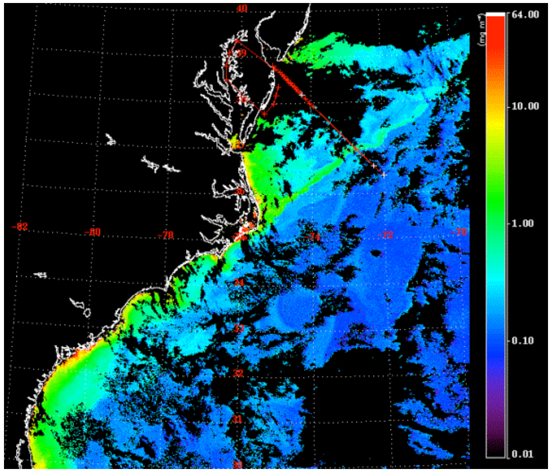
Slide from "Miscellaneous Thoughts on Ocean Data Assimilation" by M. Dowd, presented at the 2015 STATMOS/NCAR Workshop on Data Assimilation

Measurements of Ocean Biology

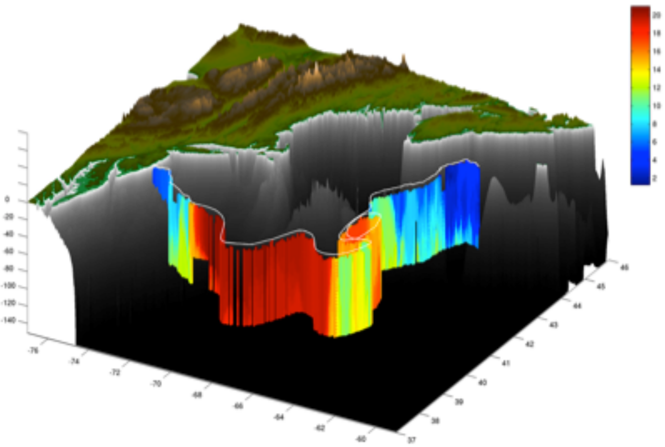
Slide from "Miscellaneous Thoughts on Ocean Data Assimilation" by M. Dowd, presented at the 2015 STATMOS/NCAR Workshop on Data Assimilation



time series
(moored instruments
- plankton, nutrients)



spatial data
(satellite chorophyll)



time-space series
(glider robots
- plankton)

- **OBSERVATION REVOLUTION IN OCEAN SCIENCES,**
- **STILL A PARTIALLY OBSERVED SYSTEM AND UNDERSAMPLED**
- **COMPLEX SPATIO-TEMPORAL SAMPLING PROTOCOLS**

Inference and Prediction

We are interested in the joint distribution of the state and the parameters conditional on the observations:

$$[x_{1:T}, \theta \mid y_{1:T}]$$

Hierarchical model:

$$[x_{1:T}, \theta \mid y_{1:T}] \propto \underbrace{[y_{1:T} \mid x_{1:T}, \theta]}_{\text{data model}} \quad \underbrace{[x_{1:T} \mid \theta]}_{\text{PDE forward model}} \quad \underbrace{[\theta]}_{\text{prior}}$$

Data Assimilation

Two main ways we could do (usually approximate) inference and prediction on $x_{1:T}$ and θ

1. Variational approaches (optimization based, “fast”, MAP estimation)
2. Sampling based methods (sampling from exact or approximate target posterior, “slow”, posterior uncertainty)

Variational Methods

Optimization problem of the following objective:

$$J = \sum_{t=1}^{T} ||y_t - h(x_t, \theta, \epsilon_t)||^2_{\Sigma^{-1}_{y_t - h(x_t, \theta, \epsilon_t)}}$$

with respect to θ and subject to the constraint

$$x_t = d(x_{t-1}, \theta, \gamma_t),$$

which is a discretized transition model for the state (approximate PDE solution obtained using numerical time-stepping from the previously estimated state x_{t-1}).

Variational Methods

Examples:

Algorithms include 3DVAR, 4DVAR

Issues:

Minimization of J is challenging and computationally expensive

- objective multimodality
- θ has high dimension
- discretization method used to obtain transitions of x from one time point to the other are computationally expensive

Sampling-based Estimation

The target density for the DA problem is $[x_{1:T}, \theta \mid y_{1:T}]$

Examples:

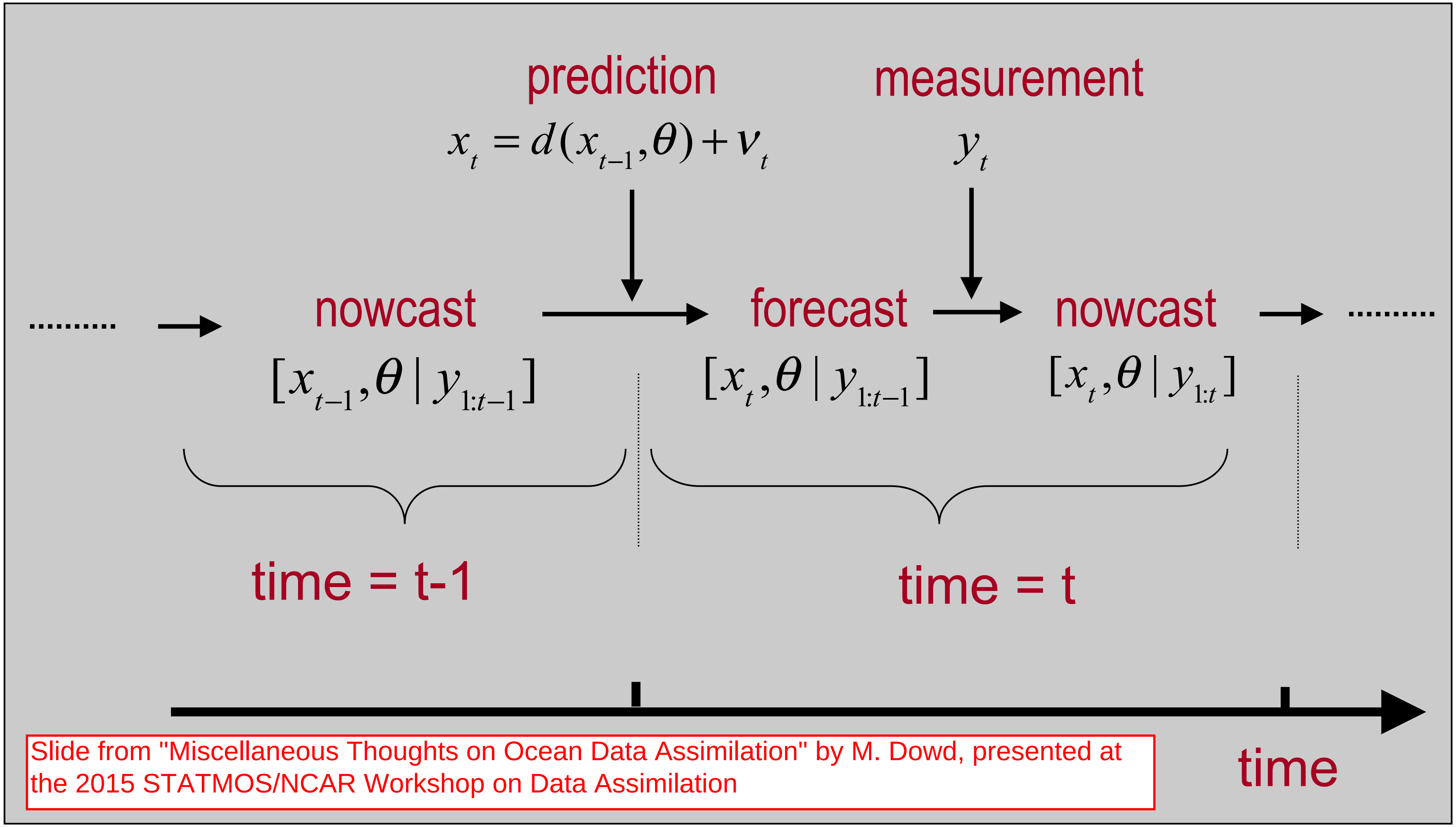
MCMC, particle filter, ensemble Kalman filter

Issues:

- MCMC targets the posterior directly, but it is slow
- Particle filtering can be faster. Iterates between two steps: prediction $x_t \mid y_{1:t-1}, \theta$ and measurement update $x_t \mid y_{1:t}, \theta$ but does not incorporate prior information on θ
- Kalman filter is exact for linear, additive Gaussian models, but targets a rough approximation of the posterior when the model is nonlinear

Sequential Methods for DA

Single stage transition of system from time *t-1* to time *t*



prediction

$x_t = d(x_{t-1}, \theta) + \nu_t$

measurement

y_t

nowcast

$[x_{t-1}, \theta \mid y_{1:t-1}]$

forecast

$[x_t, \theta \mid y_{1:t-1}]$

nowcast

$[x_t, \theta \mid y_{1:t}]$

time = t-1

time = t

time

Slide from "Miscellaneous Thoughts on Ocean Data Assimilation" by M. Dowd, presented at the 2015 STATMOS/NCAR Workshop on Data Assimilation

- *Recursive estimation of system state through time*
- *Forecast and Measurement steps*

Approaches for Parameter Estimation

1. *State Augmentation*: append parameters to the state

$$\tilde{x}_t = \begin{pmatrix} x_t \\ \theta_t \end{pmatrix}$$

Can use ‘standard’ sequential MC methods with iterative filtering

2. *Likelihood Methods*: Use sample based likelihoods

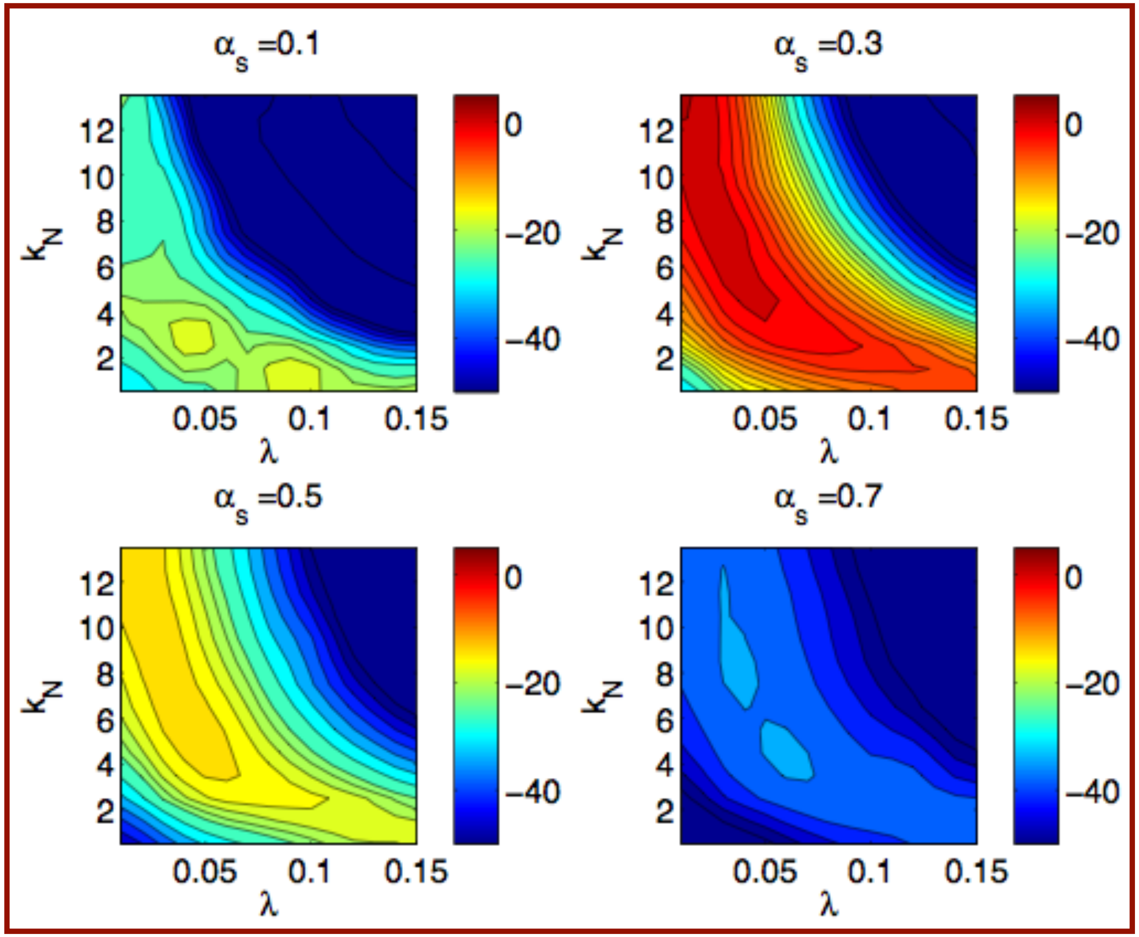
$$[y_{1:T} \mid \theta] = L(\theta \mid y_{1:T}) = \prod_{t=1}^{T} \int [y_t \mid x_t, \theta][x_t \mid y_{1:t-1}, \theta]\, dx_t$$

3. *Bayesian Hierarchical:* particle MCMC, SMC^2

Slide from "Miscellaneous Thoughts on Ocean Data Assimilation" by M. Dowd, presented at the 2015 STATMOS/NCAR Workshop on Data Assimilation

Sample Based Likelihood Surface

$$L(\theta \mid y_{1:T}) \approx \frac{1}{n}\prod_{t=1}^{T}\left(\sum\nolimits_{i=1}^{n}[y_t \mid x_{t|t-1}^{(i)},\theta]\right)$$



Slide from "Miscellaneous Thoughts on Ocean Data Assimilation" by M. Dowd, presented at the 2015 STATMOS/NCAR Workshop on Data Assimilation

Challenges

The paper identifies a number of open problems:

- High dimensional states: x_t (often $m > 10^6$); variational methods and Kalman filter are used, but rely on approximation
- Numerical resolution limits the number of ensemble sizes (usually < 1000); surrogate models and emulators are often used
- Most DA techniques do not incorporate prior information on θ, which can be very informative

Challenges

The paper identifies a number of open problems:

- Dynamic model complexity and model selection
- Can we obtain useful information from fitting a nonlinear model with so many sources of uncertainty?
- Specifying the data model: this modeling step is often neglected in BGC applications
- Sampling design