Statistical Data Assimilation for Biological Ocean Models

Mike Dowd[1]

Katja Fennel[2], Paul Mattern[1,2]

Jon Briggs[3], Renate Meyer[3]

[1]Dept of Mathematics & Statistics, Dalhousie University
[2]Dept of Oceanography, Dalhousie University
[3]Dept of Statistics, University of Auckland

Outline

- Application Area: Biological Ocean Models
- Statistical Data Assimilation
- Research Applications
  1. Particles Filters
  2. Location Particle Smoother
  3. Emulators for uncertainty analysis/parameter estimation
  4. Copulas for predictive distribution (model errors)
- Directions

Problem Statement

Target Application: Lower trophic level ocean biological processes (i.e., plankton dynamics and nutrient cycling)

Models: Ocean biology embedded within ocean circulation models

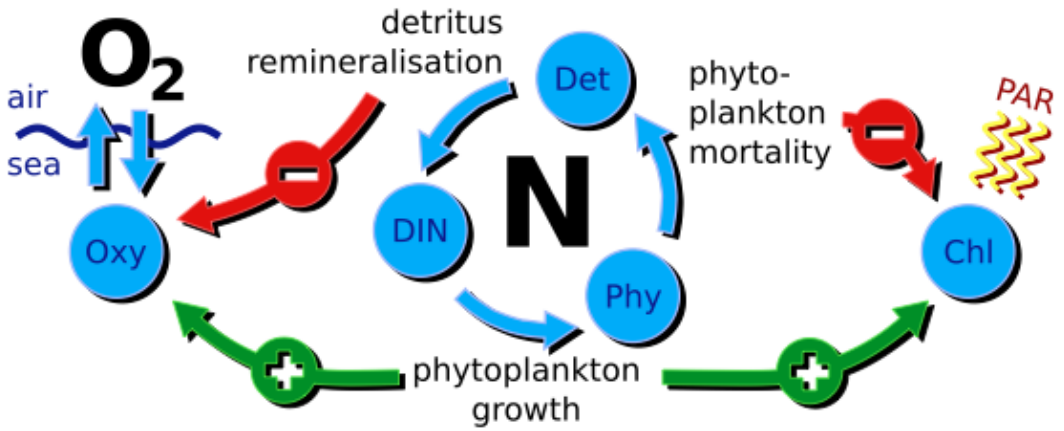
Data: Sparse, indirect and noisy. Spatio-temporal data. Emerging complex data types/sampling strategies (gliders)

Goals: Joint parameter and state estimation. Typically emphasize retrospective (hindcast) studies.

***Issues*:** Large-scale estimation problems, uncertain governing equations and parameters, complex observation errors.

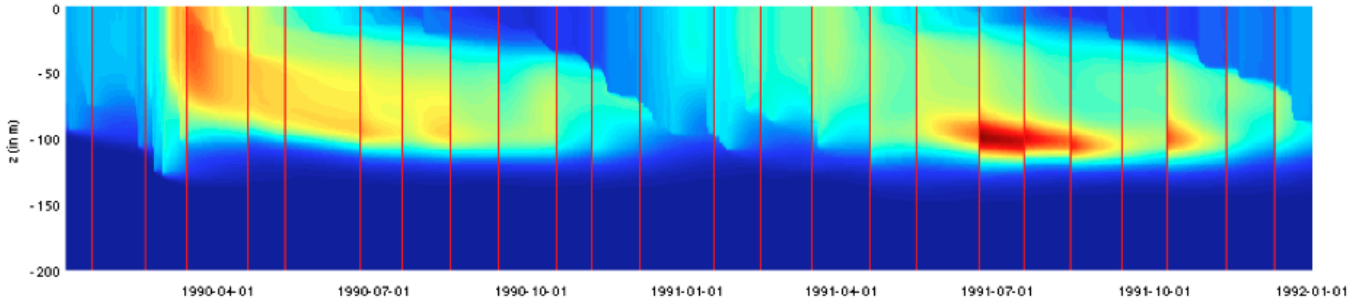
Biological Ocean Models

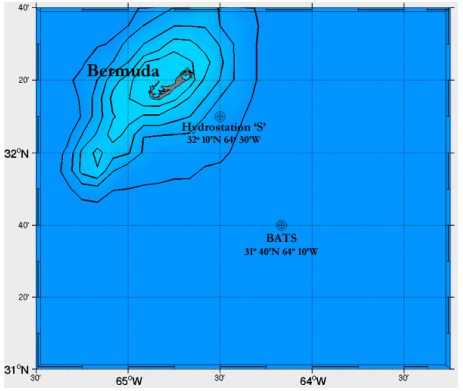
Research platforms (real applications) based on “PZND” or biogeochemical models:



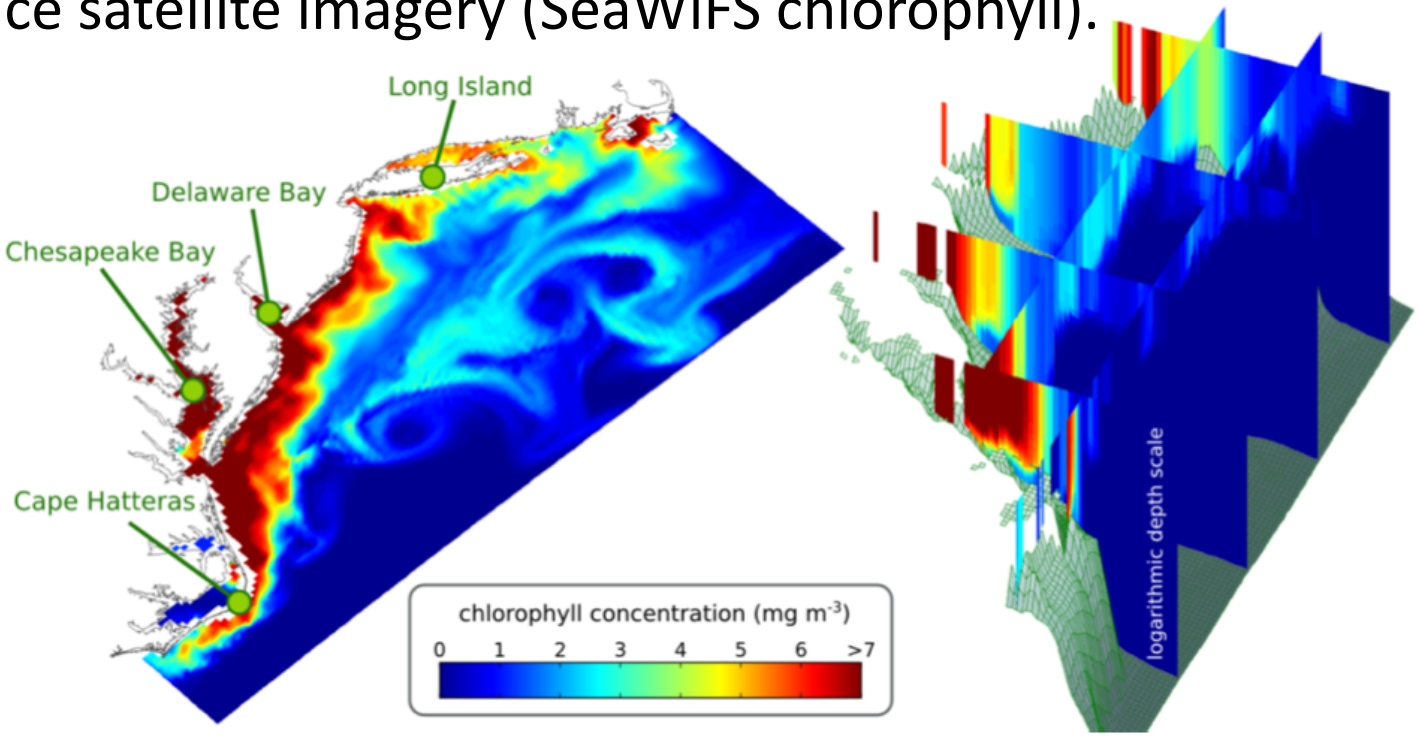
These are coupled to ocean circulation models as system of interacting non-linear tracer equations

- 1-D model of vertical ocean mixing with biology, and vertical profile measurements (Bermuda Atlantic Time Series)





- 3-D circulation (ROMS) model(s) with biology, and surface satellite imagery (SeaWiFS chlorophyll).



Statistical Data Assimilation

- Nonlinear Regression:

$$Y = D(\theta) + E$$

- State Space Model:

$$x_t = d(x_{t-1}, \theta) + e_t \qquad x_t \sim p(x_t, \theta \mid x_{t-1})$$

or

$$y_t = h(x_t) + v_t \qquad y_t \sim p(y_t \mid x_t)$$

- Hierarchical Bayesian Model

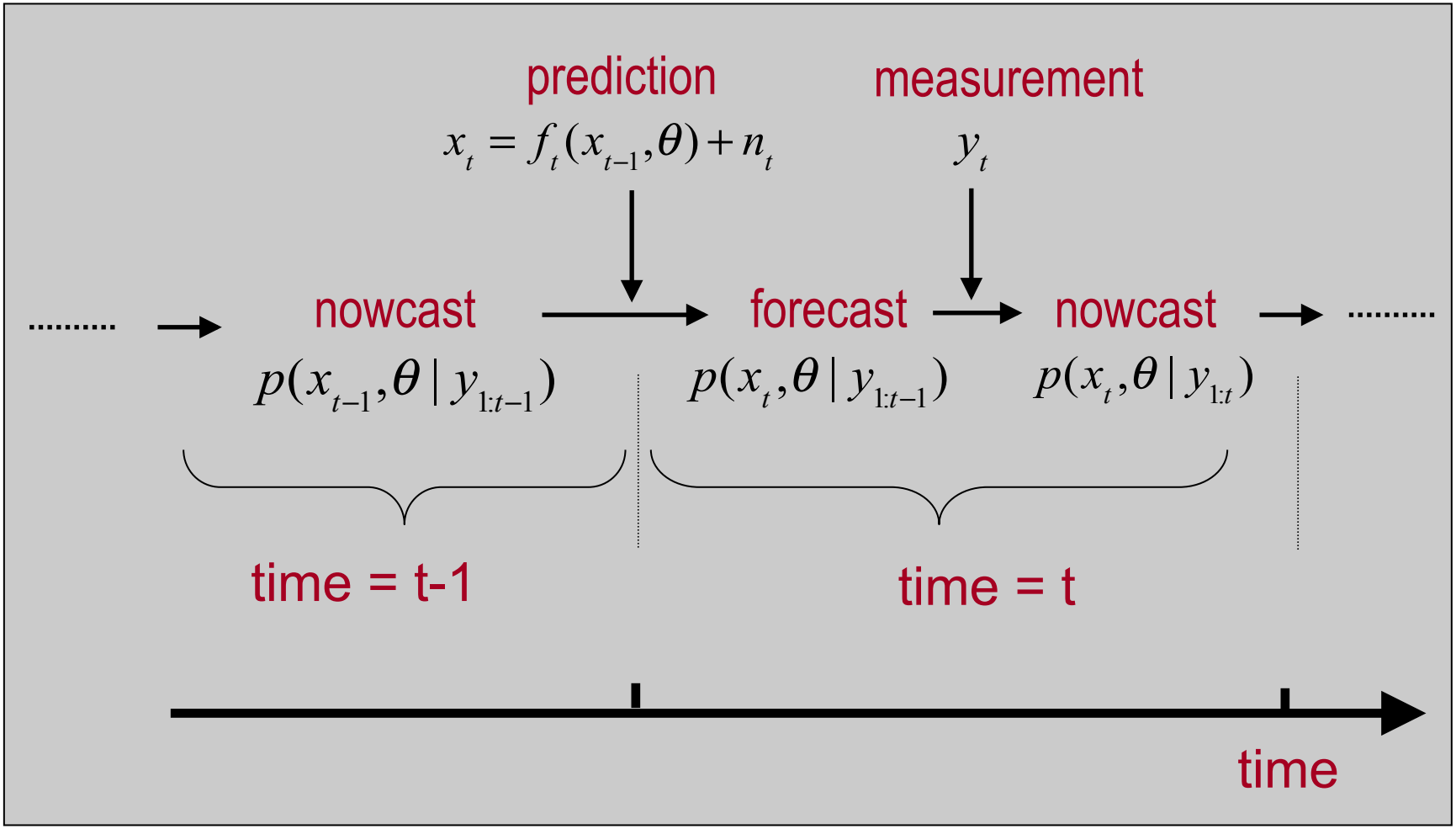
$$p(X, \theta \mid Y) \propto p(Y \mid X, \theta) \cdot p(X \mid \theta) \cdot p(\theta)$$

***Computational Approaches*:**

- Nonlinear time dependent least-squares (adjoint/cost function)
- Sequential Monte Carlo: particle Filters/Smoothers and EnKF
- MCMC, particle MCMC

Elements

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



- *Recursive estimation of system state through time (prediction/assimilation)*
- *Use samples (ensembles) to represent distributions*
- *Estimate parameters via sample based likelihood, or state augmentation*

1. A Particle Filter for a Large Scale System

Problem: Standard particle filtering (Sequential Importance Resampling) suffers from "weight collapse" or sample degeneration.

This results from prediction ensemble far away from observed state (due to small ensemble size, and high dimensional state space).
➔Likelihood poorly represented and estimation compromised.

Working Solutions: ensemble Kalman filter, look-ahead particle filters

What basic modifications do we have to make to implement a simple particle "filter" for our biological ocean model?

We made the following modifications to allow for sequential MC for our 3D ocean model:

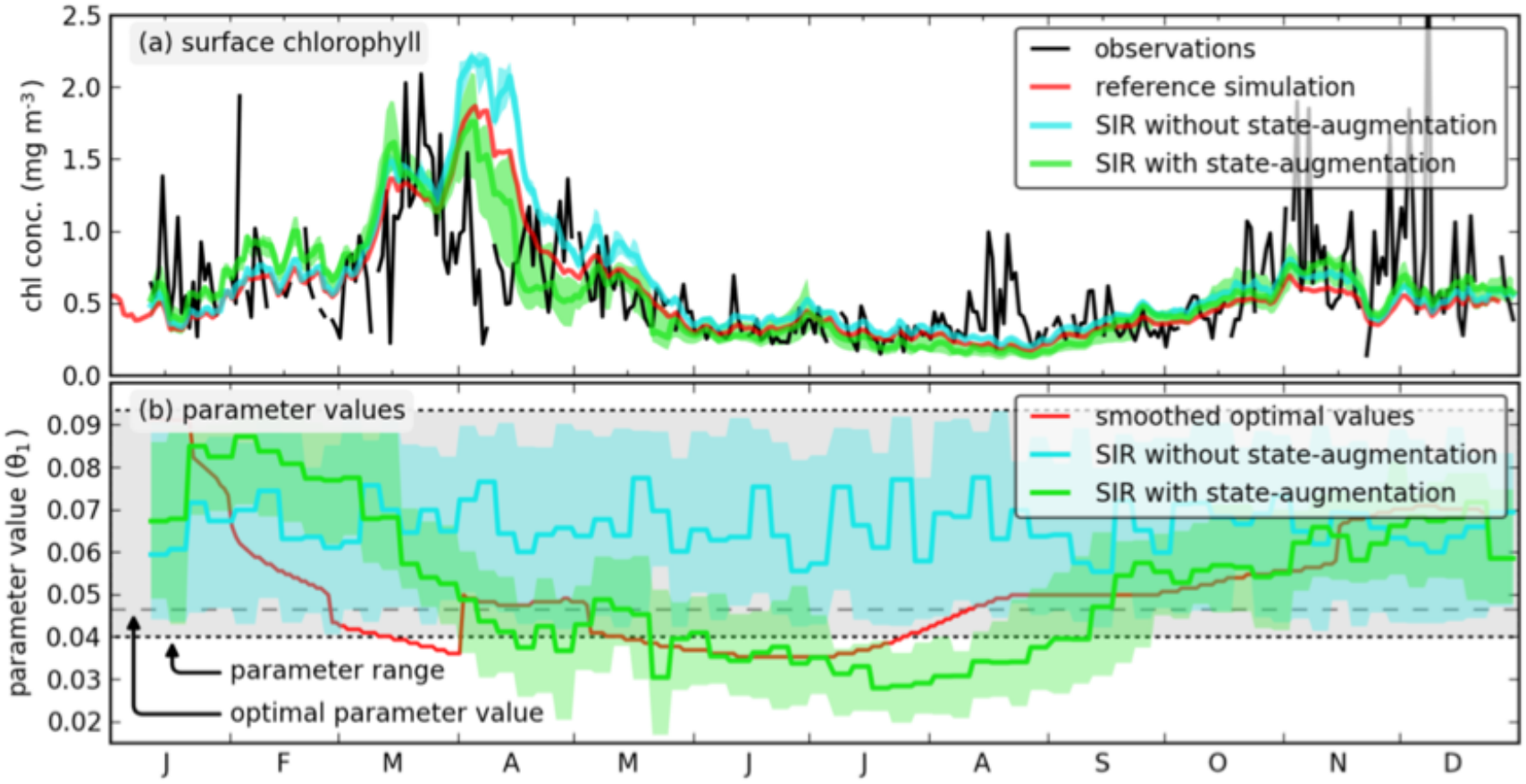
1. **Likelihood function**: based on spatial distance metrics, assign variance (=weights) via predictive skill. Treats weight collapse.

2. **Error subspace**, i.e. introduce stochasticity only through biological parameters. Treats high dimensionality.

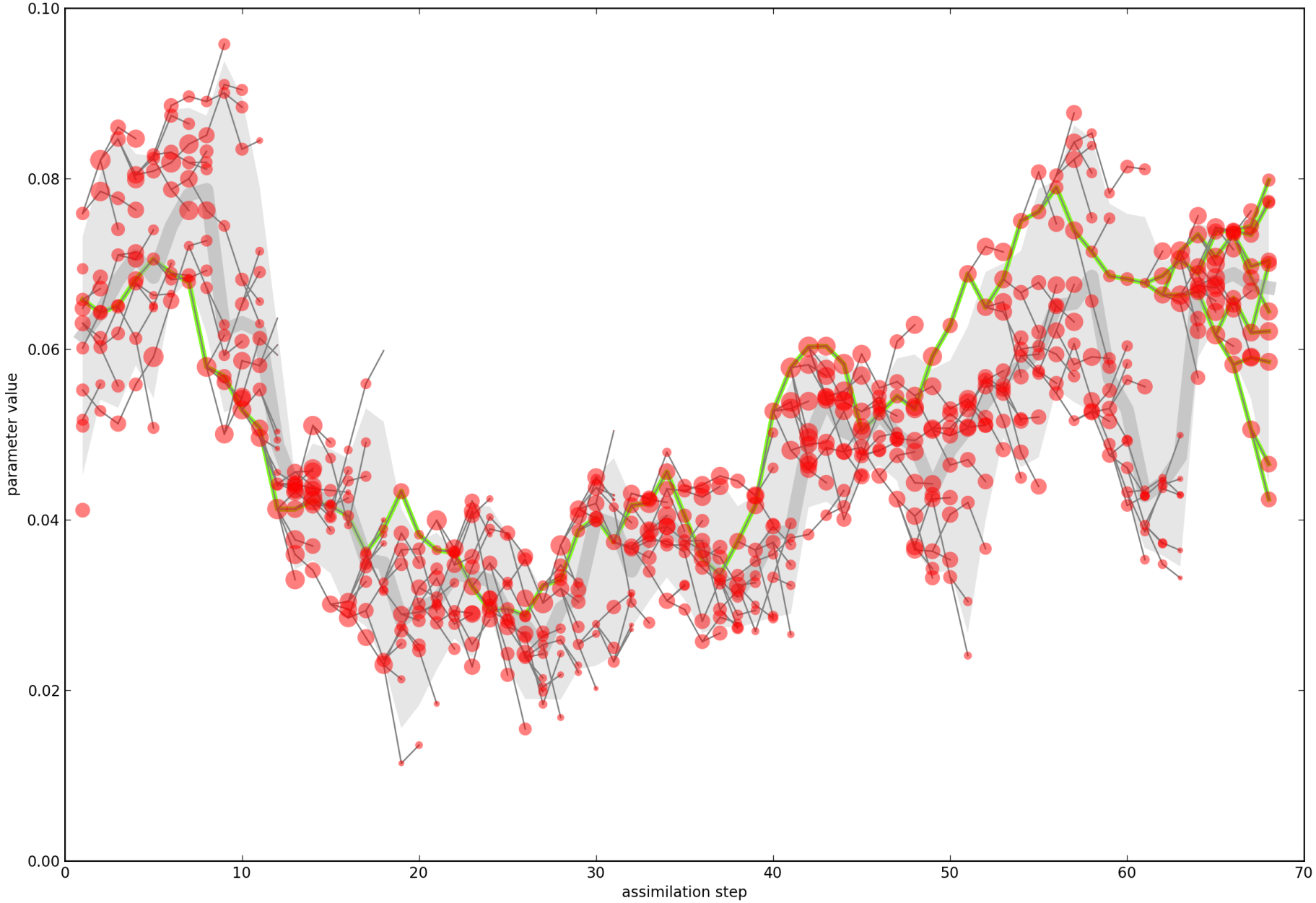
3. **Fixed lag smoother**, incorporate observations from multiple times into observation update. Treats Robustness.

4. **State Augmentation**: provides for adaptive parameter estimates. Treats bias.

Particle Filtering Results: Mid Atlantic Bight



Ensemble Member History from Parameters



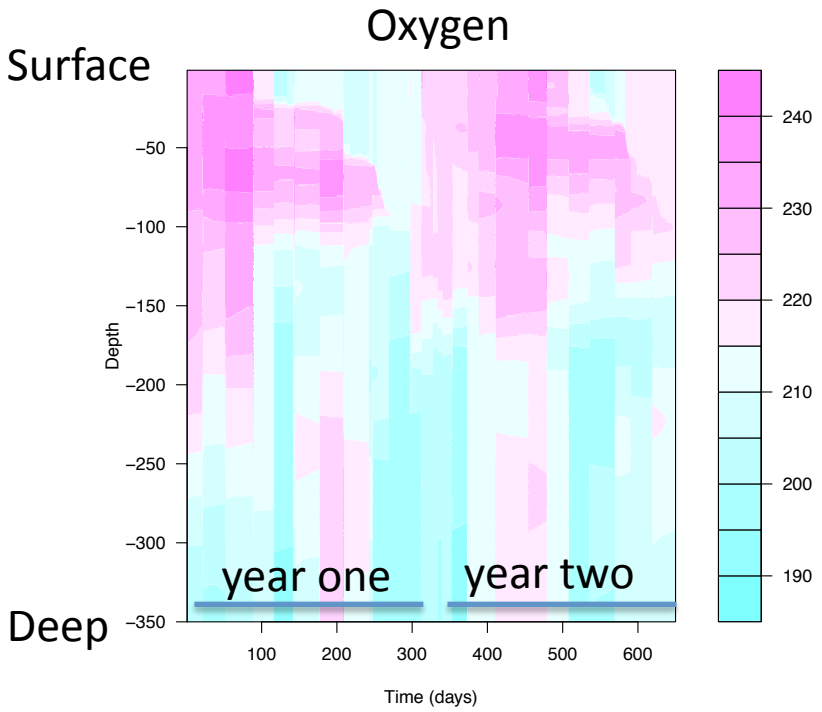
2. A Location Particle Smoother for Spatio-Temporal Systems

IDEA: run particle filter through time, but apply particle smoother ***in the spatial domain*** at every time step.

- Relies on use of time-domain predictive distribution as proposal distribution, and sequential importance sampling.

- Assumes conditionally independent observations errors.

Successful for multivariate state estimation for 1D biological model (Bermuda Atlantic Time Series site)



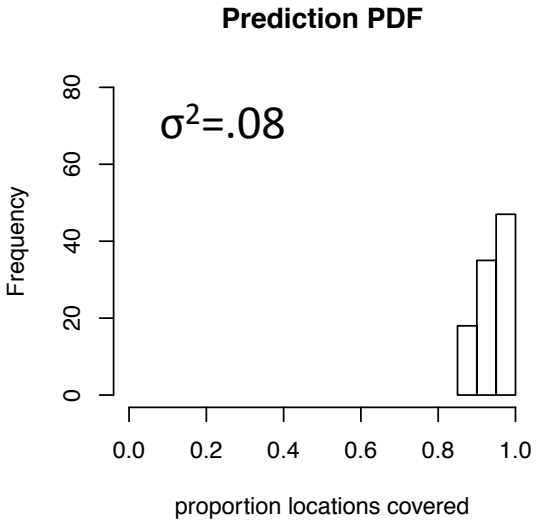
Performance of Location Particle Smoother

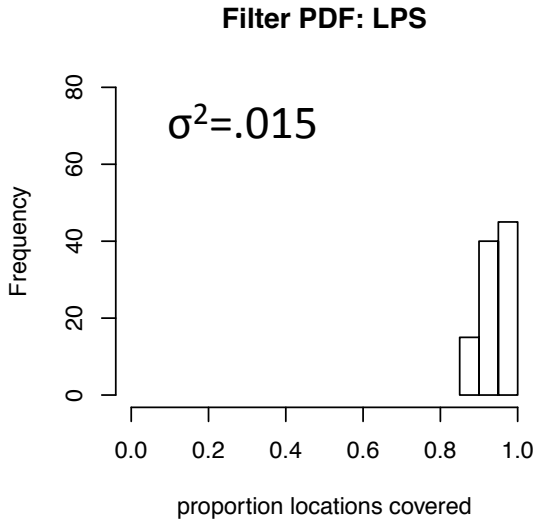
Simulation study:

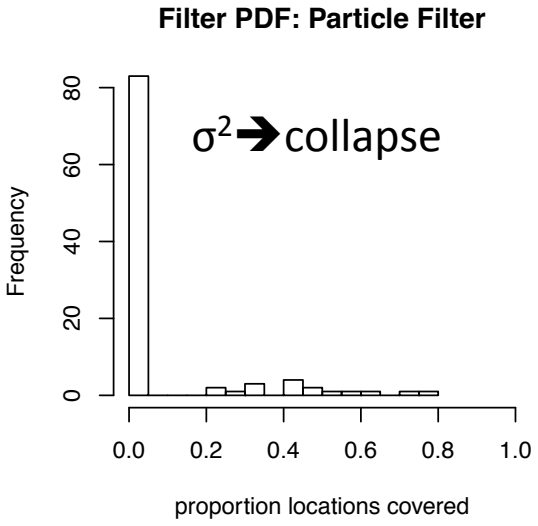
- Lorenz96 system
- 10 true states generated
- 10 replicates per state
- Cauchy observation error
- (plus different error types, mis-specifications wrt variance and tail density)

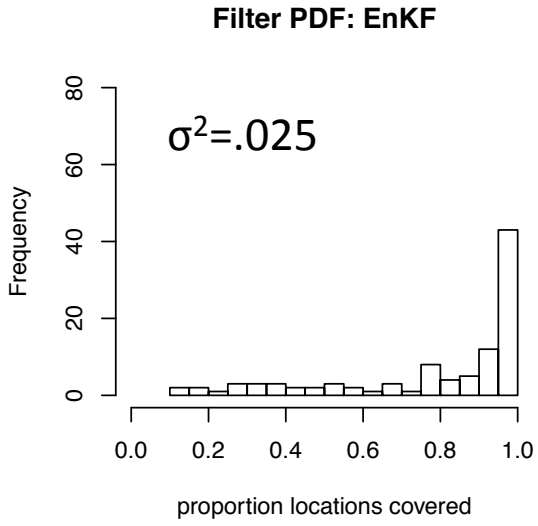
Performance Metrics:

- Percent coverage of true state by 95% credible intervals
- Average ensemble variance (σ^2)









PROS: outperforms EnKF for non-Gaussian likelihood

CONS: less computationally efficient, no better for Gaussian

3. Emulators

- Low dimensional (=computationally efficient) representation of complex computer code (numerical models).

- Research efforts in Statistics emphasize Gaussian process models. Used for calibration, uncertainty analysis, and experimental design

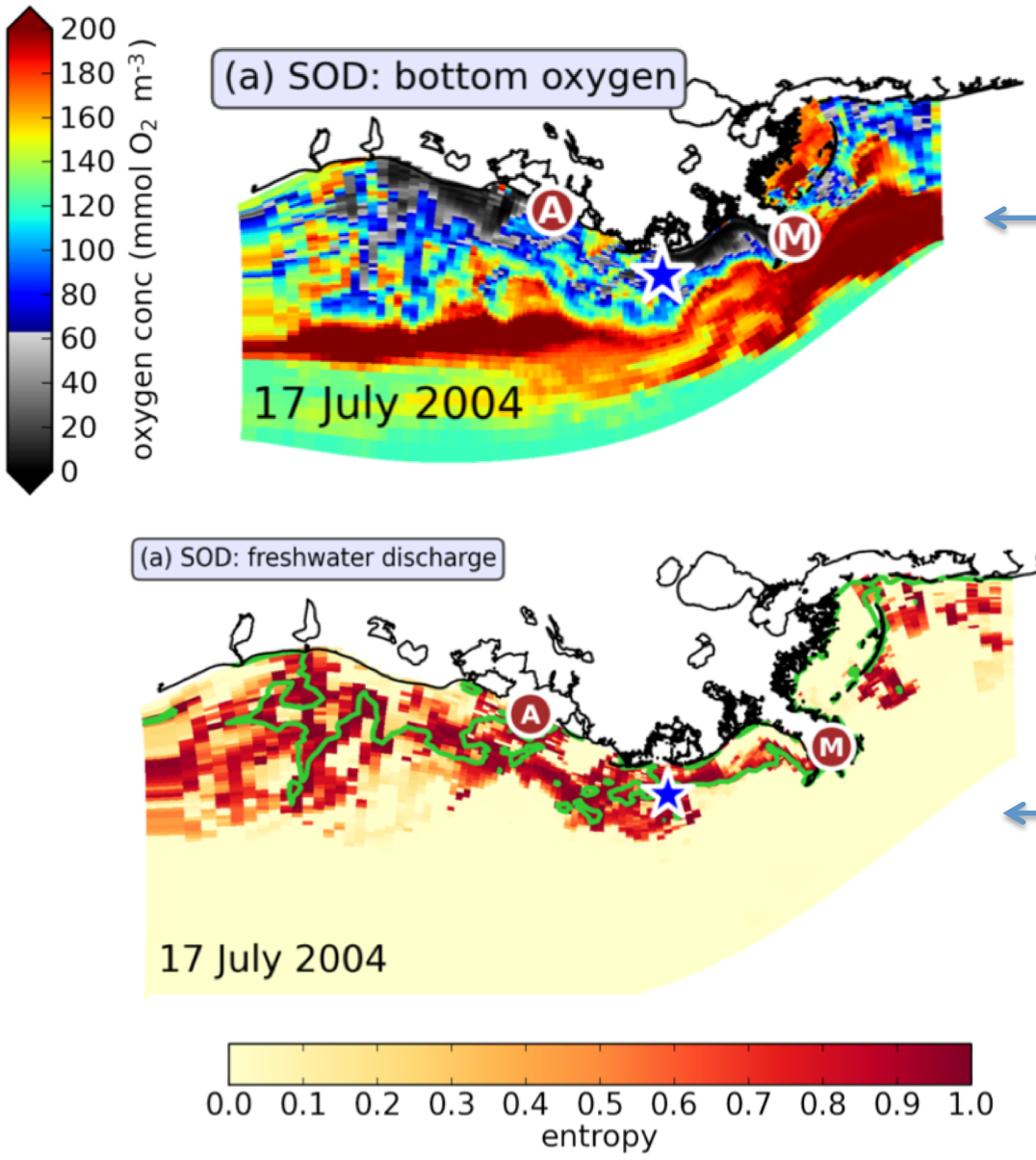
- We have used the "polynomial chaos" emulator in our work:

$$D(s,t,\theta) \approx \sum_k a_k(s,t)\,\phi_k(\theta)$$

where ϕ_k is the polynomial (type as dictated by pdf of θ), and a_k are coefficients (obtained via Gaussian quadrature).

➔ *Uncertainty Analysis* and *Parameter Estimation*

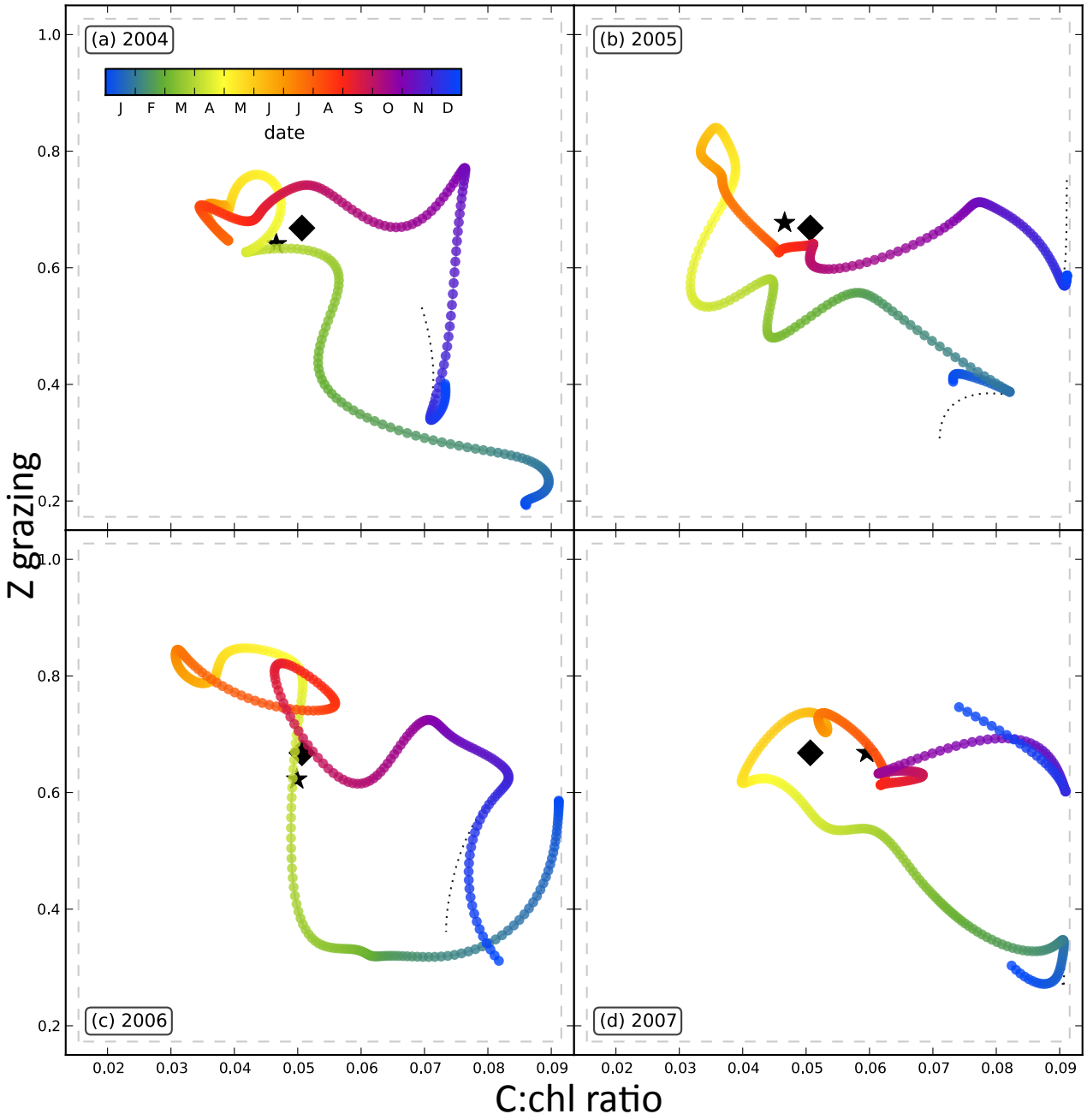
Uncertainty Analysis: Gulf of Mexico "Dead Zone"



Predicted Bottom Oxygen Concentration (defines oxic/anoxic zones)

Uncertainty (entropy) in the oxic state, due to freshwater discharge

Parameter Estimation



- Seasonal co-evolution of 2 model parameters (Zooplankton grazing, carbon:chl)
- Estimated with emulator using nonlinear least squares based on image distance metric (compares satellite to model surface field)
- Clear seasonal signals evident in these parameters; inclusion improves prediction skill of model.

4. Multivariate Error Distributions via Copulas

- We want: $x_t \sim p(x_t \mid x_{t-1})$ - predictive density/ model error dist

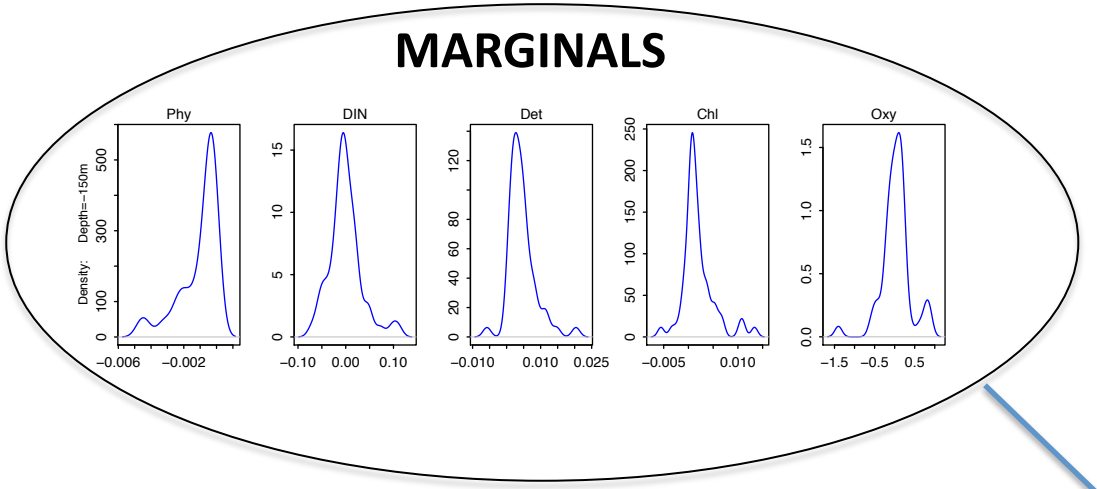
- We have: $x_t = d(x_{t-1}) + e_t$ - a numerical model to generate samples

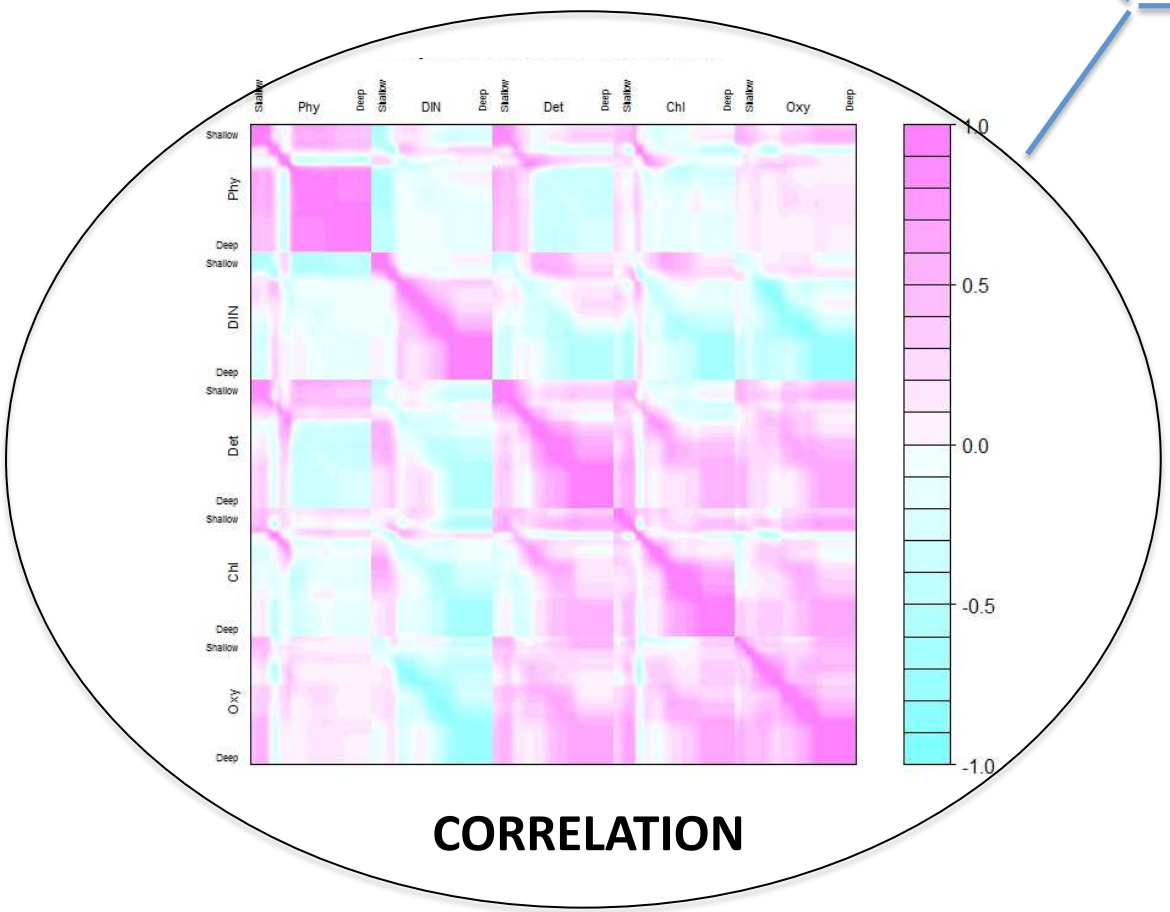
Rather than a sample-based representation of model errors, what about a (semi-)parametric representation?

Must be: accurate, flexible, easy to sample from, "high" dimensions…

Idea: create multivariate distributions using copulas …

- control the marginal distributions, and dependence separately
- construction and sampling from copulas is standard statistical task
- can be derived from "ensembles" using method of moments, or via parametric models (➔ anisotropic, non-homogenous ++)





$$P(x_1,\ldots,x_n) = C(P(x_1),\ldots,P(x_n))$$

Samples from predictive density, or model errors

DIRECTIONS

- Hierarchical Bayesian framework conceptually useful ➔ e.g. particle MCMC for dynamic models. Practically .. where / how to make approximations and their consequences.

- Sample based solutions. SIR is not the only particle filter – can use generic proposals (evenKF). Also smoothers, can include priors

- Model errors characterization important. How to do represent stochastic process (via samples, distributions).

- Role of Emulators? Need to incorporate emulator error in hierarchy, i.e. $p(x_e|x)$, and provide for efficient construction.

- Validation? Design for sample-based numerical experimental for assessing consistency, efficiency, asymptotics, robustness.

- Emerging approaches for dynamic systems from Statistics.