Probabilistic Approaches to Data Assimilation for Earth Systems

Mark Buehner (Environment Canada)
Pierre Gauthier (Université de Québec à Montréal)
Kayo Ide (University of Maryland)
Robert Miller (Oregon State University

February 17, 2013 – February 22, 2013

1 Overview of the Field

Estimation and prediction of the state of a system by combination of observations and model output is a task that is common to many scientific fields. In the earth science community, it is referred to as data assimilation. Much of the impetus for research in data assimilation in the earth sciences comes from numerical weather prediction, in which the initial conditions for a forecast must be derived from incomplete and noisy data, combined with the output of previous forecasts. Many important problems in analysis, prediction and monitoring of earth systems rely on data assimilation to provide a detailed description of the system along with reliable confidence intervals.

The first data assimilation systems implemented for earth systems used the machinery of inverse problems, dating back to Legendre and Gauss, as well as the mathematics of filtering developed in the engineering community following the development of the Kalman filter. By the mid 1990s most data assimilation systems for the ocean and atmosphere were based on least squares methods. The methods of choice for operational numerical weather prediction (NWP) were derived as methods for finding the minimum $\mathbf{x}$ of a quadratic cost function with the typical form:

$$J(\mathbf{x}) \quad = \quad (\mathbf{x}-\mathbf{x}^b)^T B^{-1}(\mathbf{x}-\mathbf{x}^b) + (\mathbf{y}^o - \mathbf{h}(\mathbf{x}))^T R^{-1}(\mathbf{y}^o - \mathbf{h}(\mathbf{x})) \tag{1}$$

where $\mathbf{x}$ is the state vector of the model and $\mathbf{y}$ is the observation, with superscripts b for the background for the optimization and o for observed data, and $\mathbf{y} = \mathbf{h}(\mathbf{x})$ is the functional relationship between the state and observation. In this form, quality of the analysis $\mathbf{x}^a$ given as the optimizer of the cost function depends on not only $\mathbf{x}^b$ and $\mathbf{y}^o$, but also their respective error covariance matrices $\mathbf{B}$ and $\mathbf{R}$, as well as $\mathbf{h}$. Quite often in the earth science applications, $\mathbf{R}$ and $\mathbf{h}$ are assumed to be (perfectly) known. When $\mathbf{B}$ is statistically estimated, the method for obtaining $\mathbf{x}^a$ is called 3DVAR, where 3D stands for the three-dimensional space. This estimation method is instaneous in time (lacking fourth dimension) and assumes synchronous observations at the time of the analysis. In the operational NWP, this typically repeats every 6hr which defines the assimilation window. By allowing the observations to be asynchronous and letting the state evolve over the assimilation windodw according to the numerical model equations the method becomes 4DVAR. A 4DVAR system, like the one described here, in which the model is assumed to be perfect, is sometimes known as a "strong constraint" system. The models, of course, are not perfect, but initial condition error dominates the forecast error in numerical weather prediction early in the forecast. If the errors in the background and the

observation are Gaussian random variables with covariance matrices $\mathbf{B}$ and $\mathbf{R}$ then the 3DVAR and 4DVAR solutions are the unbiased minimum variance estimates, as well as the maximum likelihood estimates, but the models are not unbiased, the errors can be far from Gaussian and the error covariances are, in any case, poorly known. Nonetheless, most operational weather centers used 3DVAR or 4DVAR with simplified error estimates since the year 2000.

Other data assimilation problems in earth systems were computed by methods based on the Kalman filter, in which model-data misfits at a given time are used to correct a forecast based on the best available initial condition. Following the notation of [2], given $\mathbf{x}_j^a$, our best guess at the state of the system at time t_j, a linear model specified by $\mathbf{M}$ and observations $\mathbf{y}_j^o$ the Kalman filter is specified by the following sequence of operations:

$$\begin{aligned}
\mathbf{x}_{j+1}^f &= \mathbf{m}(\mathbf{x}_j^a) && (2)\\
\mathbf{x}_{j+1}^a &= \mathbf{x}_{j+1}^f + \mathbf{K}(\mathbf{y}_{j+1}^o - \mathbf{h}(\mathbf{x}_{j+1}^f)) && (3)\\
\mathbf{K} &= \mathbf{P}_{j+1}^f \mathbf{H}^T (\mathbf{H}\mathbf{P}_{j+1}^f \mathbf{H}^T + \mathbf{R})^{-1} && (4)\\
\mathbf{P}_{j+1}^f &= \mathbf{M}\mathbf{P}_j^a \mathbf{M}^T + \mathbf{Q} && (5)\\
\mathbf{P}_{j+1}^a &= (I - \mathbf{K}\mathbf{H})\mathbf{P}_{j+1}^f && (6)
\end{aligned}$$

where $\mathbf{K}$ is the so-called the Kalman gain gives the weight of observations with respect to thebackground, $\mathbf{H}$ is the tangent linear model of $\mathbf{h}$, $P_j^{f,a}$ is the error covariance of $\mathbf{x}_j^{f,a}$ estimated by the tangent linear model $\mathbf{M}$ of the numerical model $\mathbf{m}$, and $\mathbf{Q}$ is the covariance of the model error, which is assumed to appear as additive white noise on the right side of (2). Most earth system models of interest involve nonlinear models and very high dimensional state spaces so direct application of (2) - (6) is not practical. One common method of applying Kalman filter techniques is the Ensemble Kalman filter (EnKF) in which an ensemble of state vectors $\mathbf{x_j}^{(f,a)(i)}, i = 1, \ldots, N$ are calculated, and the forecast/analysis error covariance matrices are taken to be the sample covariance of $\mathbf{x}_j^{(f,a)(i)}, i = 1, \ldots, N$.

Refinement of 4DVAR methods to incorporate flow-dependent errors by means of ensemble calculations, as well as refinements of the EnKF to allow for more efficient error estimates with necessarily small ensembles formed major themes of the presentations from groups involved in NWP. New techniques for error estimation were also presented, as well as techniques for state and parameter estimation in high dimensional highly nonlinear systems.

A more general way of looking at the data assimilation problem is as an exercise in the calculation of conditional probabilities. If we assume that the earth system under study is governed by a random process, then the solution of the data assimilation problem is the probability density function (pdf) of the system conditioned upon observations. If the system evolution is governed by linear dynamics and the noise sources are Gaussian, then the conditional probability will be Gaussian and the methods will be equivalent, but this is very rarely the case in analysis of earth systems. The system state vector $\mathbf{x}$ in most cases of interest will be of enormous dimension, and direct estimate of the conditional pdf must proceed by Monte-Carlo methods. Monte-Carlo methods for estimation of pdfs are often known as "particle methods," since each sample can be viewed as a particle moving through state space. The challenge in development of particle methods is the incorporation of observed data into the ensemble of particle trajectories. The EnKF can be viewed as a particle method in this light, but it is known to run into trouble in highly non-Gaussian settings; see, e.g., [3], [5].

Particle methods are the most general approach to the data assimilation problem in highly nonlinear and non-Gaussian settings in that they provide an explicit estimate of the relevant pdf. The main difficulty encountered in applications of particle methods to earth systems stems from the fact that the extremely high dimensional spaces in the problems of interest are necessarily sparsely sampled by ensembles of practical size. It is therefore a matter of extreme difficulty to construct particle trajectories that give rise to the observations as they actually occurred, since, in the high dimensional setting, any given value of an observed quantity is highly improbable. This so called "curse of dimensionality," a term coined by Bellman (e.g., [1]) can lead to extremely pessimistic estimates of the ensemble sizes needed to produce useful results; see, e.g., [4].

2 Recent Developments and Open Problems

3 Presentation Highlights

3.1 Current State-of-the-Art Short- to Medium-Range Operational Data Assimiltion for Numerical Weather Prediction

3.1.1 "4D-Ensemble-Var. A Development Path for Data Assimilation at UK Met Office" by Andrew Lorenc

- Why are we doing it? What is wrong with 4D-Var? Why do we need an ensemble?
- Addressed by:
  - Hybrid-4D-Var
  - Flow-dependent covariances from localised ensemble perturbations.
  - 4DEnVar. No need to integrate linear & adjoint models.
  - An Ensemble of 4DEnVar. Or alternatives.
- Preliminary results, and plans.
  - EnVar (i.e. both hybrid-4D-Var & 4DEnVar)
  - 4DEnVar
  - EDA (i.e. an ensemble of 4DEnVar assimilations)

3.1.2 "Recent Developments in Data Assimilation for Global Deterministic NWP: EnVar vs. 3D-Var and 4D-Var" by Mark Buehner

- The ensemble-variational (EnVar) data assimilation approach
- Recent results from using EnVar compared with standard 3D-Var and 4D-Var (but NO comparisons with 4D-Var- Bens or EnKF)
- Comparison of EnVar with 3D-Var and 4D-Var
  - EnVar produces similar quality forecasts as 4D-Var below 20hPa in extra-tropics, significantly improved in tropics
  - Above 20hPa, scores similar to 3D-Var, worse than 4D-Var; potential benefit from raising EnKF model top from 2hPa to 0.1hPa
- EnVar is an attractive alternative to 4D-Var:
  - Like EnKF, uses full nonlinear model dynamics/physics to evolve covariances; no need to maintain TL/AD version of model
  - Instead, makes use of already available 4D ensembles
  - More computationally efficient and easier to parallelize than 4DVar for high spatial resolution and large data volumes
  - Computational saving allows increase in analysis resolution and volume of assimilated observations; more computational resources for EnKF and forecasts
- Next step
  - Finalize testing EnVar with goal of replacing 4D-Var in operational global prediction system during 2013 in combination with other changes
  - Early results from using EnVar in regional prediction system as 4D-Var replacement look promising

3.1.3 "Hybrid Variational-Ensemble Data Assimilation at NCEP" by Daryl Kleist

- Hybrid Var/Ens at NCEP
- OSSE-based hybrid exeriments (Joint OSSE)
  - International, collaborative effort between ECMWF, NASA/GMAO, NOAA (NCEP/EMC, NESDIS, JCSDA), NOAA/ESRL, others
  - ECMWF-generated nature run (c31r1)
- Summary
  - 4DENSV seems to be a cost effective alternative to 4DVAR
  - Inclusion of time-invariant static B to 4DENSV solution is beneficial for dual-resolution paradigm
  - Extension to 4D seems to have larger impact in extratropics (whereas the original introduction of the ensemble covariances had largest impact in the tropics)
  - Increased convection in 4D extensions remains a mystery (it is not the weak constraint on unphysical moisture as I originally hypothesized)
  - Original tuning parameters for inflation were utilized. Follow-on experiments with tuned parameters (reduced inflation) and/or adapative inflation should yield even more impressive results.
- Future work
  - NCEP successfully implemented hybrid variational-ensemble algorithm into GDAS
  - NCEP aggressively pursuing application of hybrid to other systems
  - Have already run preliminary tests for 1981-1983 periods, attempting to capture QBO transitions (a difficult problem for reduced observing system periods)
  - Extensions to the GDAS hybrid are ongoing, including 4DEnsVar

3.2 Long-Range Data Assimilation and Climate Issues

3.2.1 "Diagnostics of Data Assimilation and Models for Environmental and Climate Prediction" by Pierre Gauthier

- Assessing the impact of observations and its applications using
  - Information content
  - Observing System Experiments
  - Observation impact on the quality of the forecasts
- Observability of precursors to instability
  - Observability of structure functions has been defined in observation space as a correlation between innovations and the structure function
  - Even though those structures do correspond to structure that will grow the most or grow to correct the forecast error at a given lead time
  - Reduced rank Kalman filters do not seem to be appropriate to represent the background error covariances in an assimilation system
  - Evolved covariances as estimated with an Ensemble Kalman filter would be more appropriate for an hybrid 4D-Var assimilation
- Diagnosing dynamical balance based on physical tendencies
  - Impact of using an analysis produced by a different model

  - Driving a limited-area model for regional climate applications with analyses produced by a different model

- Summary
  - Numerical simulations of the atmosphere are central to better understanding the complexities of the Earth system
  - Climate and weather forecasting systems now need to take into account interactions with the oceans, the land, ice, snow, atmospheric chemistry

3.2.2 "How Warm is it Getting? The Determination of a Trend in a Multi-Scale Problem" by Juan Restrepo

- Trend Problem: Define a set of simple universal rules with which to compute an underlying tendency, given a finite (non-stationary/multi-scale) data set
- Applications: Problems that critically depend on ar trend calculation
  - Global warming (sun radiation, CO2 averages, global temperature estimates).
  - Mean sea level (land ice melt and its effect on sea rise). Variability of
  - Glacial ice packing.
  - Long-term ocean sea surface temps (SST) data: PCA has an ENSO-line signal, not in ocean models.

3.3 Flow-Dependent Covariance: Ensemble Kalman Filter and Hybrid Approach

3.3.1 "Parallel Implementation of an Ensemble Kalman Filter" by Peter Houtekamer

- Parallel computing issues for EnKF
- EnKF algorithsm issues for parallel computing such as
  - Monte Carlo approach to data assimilation
  - Sequential algorithm
  - Time interpolation
  - Localization
- Scaling by
  - Computation of analyses
  - Strong scaling
  - Weak scaling
  - Application towards better results
- Discussion and future steps
  - Code complexity and evolution
  - Model error component, correlated observation errors, deduction of model spin-up and shorten the data assimilation window.

3.3.2 "Realtime Hurricane Prediction with EnKF Assimilation of Airborne Doppler Observations" by Fuqing Zhang

- PSU WRF-EnKF HFIP Stream1.5 Model and Filter Configurations for 2012
  - High resolution (4.5 to 3km),
  - More vertical levels (30 to 43),
  - Improved surface flux
- Comparison: EnKF, 3/4DVar, E3/4DVar
  - TC intensity forecasts can significantly improved with cloudpermitting EnKF assimilation of high-resolution inner-core radar observations
  - EnKF has great promise at meso- and convective scales but additional benefits may come from hybrid/coupling with 3D/4DVar
  - It might be premature to write off the adjoint-based 4DVar
  - We need to more effectively account for model error in DA and EF
- Summary on WRF-EnKF Hurricane Prediction
  - Hurricane intensity forecast can be greatly improved through using advanced DA techniques and a cloud resolving NWP model with assimilation of high-resolution inner-core observations
  - Average over 100+ NOAA P3 Doppler missions, the PSU WRF-EnKF forecasts with assimilation of Vr has the day 1 to 5 mean intensity forecast error 20-40% smaller than NHC official forecasts
  - The PSU WRF-EnKF experimental system performed well for all landfalling hurricanes during 2008-2012. It also shows great promise in predicting hurricane-induced rainfalls, as well as uncertainties
  - Future of hurricane prediction: better inner-core observations, better data assimilation, better forecast model, better computing resources

3.3.3 "Recent Advances in EnKF" by Eugenia Kalnay

- Review of a few recent advances in LETKF
  - Running in Place
  - Effective assimilation of precipitation
  - Ensemble Forecast Sensitivity to Observations (EFSO)
  - Parameter estimation and carbon cycle data assimilation
  - Estimation of surface heat and moisture fluxes
  - Sensible and latent heat fluxes (SHF, LHF)
  - Estimation of wind stress in addition to SHF and LHF
- Summary
  - We have shown the feasibility of simultaneous analysis of meteorological and carbon variables within LETKF framework through simulation experiments.
  - The system LETKF-C has been tested in a intermediate-complexity model SPEEDY-C with excellent results.
  - Multivariate data assimilation with localization of the variables (Kang et al. 2011)
  - Advanced data assimilation methods for CO2 flux estimation have been explored (Kang et al. 2012)
  - A short window is better than a long window.
  - We are implementing the LETKF-C to NCAR CAM 3.5 model and real observations

3.3.4 "Representing Model Uncertainty in Ensemble Data Assimilation" by Jeffery Whitaker

- Evaluating schemes for representing model uncertainity
  - Using EPS: Spread/error consistency, probabilisic scores; hard to know whether improvement comes simply from reducing spread deficiency
  - Using EnKF: Tougher test if multiplica1ve inflation used as baseline, since scheme must do more than increase variance.
  - Evolution of all errors in DA cycle (not just model error) must be represented. Model error may not be dominant.
- Methods for representing model error:
  - Multiplicative inflation
  - Additive inflation
- Lesson learned from simple model experiments
  - Improving background-error covariances in an EnKF is a tough test for a model error scheme.
  - Multiplicative inflation and stochastic physics/additive inflation sample different sources of error in the DA
  - Any of these methods can only do so much improving the forecast model will usually have a larger impact on the data assimilation.
- Results from NCEP Operationa 3D EnVar system
  - As in simple model, its hard to improve upon ad-hoc inflation.
  - We have progressed to the point where state-of-the-art stochastic schemes can perform as well as additive inflation in DA.
  - Hopefully, we can do better by treating model uncertainty the process level in each parameterizationscheme.

3.3.5 'Hybrid Data Assimilation without Ensemble Filtering" by Ricardo Todling

- Background
  - Hybrid DA includes: re-centering plus inflation
  - Evaluations in GEOS DAS suggest: 1) Hybrid approach provides noticeable improvements only when using additive inflation, i.e., EnKF alone doesn't do it; 2) Forecasts from EnKF analyses plus additive inflation result in mild spread within the background time window; 3) It seems that much of the initial (analysis) spread can be simulated with additive inflation; 4) Appreciable background spread is obtained in the latter case
- Question: how does hybrid-DA perform when the ensemble filter is dropped and an ensemble of analyses is created from simply additively inflating the central analysis?
- Summary:
  - Overall 3d-hybrid approach gives positive results in GEOS DAS with noticeable reduction of model biases and improved skill scores
  - Filter-free scheme works just as well as EnKF in sustaining ensemble
  - Would be nice to study skill of NMC-like perturbations in an EPS
  - Advantages of Filter-Free Hybrid
    - Really inexpensive way of generating ensemble
    - Avoids need to maintain two analysis systems
    - Avoids contradictions when calculating adjoint-based obs impact

3.3.6 "GMAO Hybrid Ensemble 3D-Var" by Amal El Akkaraoui

- Current status of GMAO Hybrid Ensemble 3DVa
  - Hybrid results are significantly positive: largely positive for the tropical winds around 500-200mb, and slightly positive to neutral elsewhere;
  - Still need to try to get more impact for temperature;
  - More tuning and testing with higher resolution, more members, different localization scales.
- Filter free GMAO Hybrid - New considertions
  - Start with the two decoupled schemes: the ensemble (S-EnKF) and the variational (3D-Var).
  - Both systems do not need to communicate with each other in order to perform, and then we ask them to work together to the fullest of their potential, hoping for the best outcome.
  - An important aspect of this exercise is how well each system performs when alone.
  - An equally important one is how do they work together as a team when in hybrid mode, and how do they both affect each others performance
- Summary
  - Additive inflation is essential to the performance of the ensemble $\rightarrow$ We need to try the hybrid with the 0.4 scaling factor.
  - The resolution difference between the ensemble and the central is critical for the re-centering step. $\rightarrow$ Re-evaluate the need for the re-centering in the "same-resolution" configuration.

3.3.7 "Hybrid 4DVAR and Nonlinear EnKS Methods without Tangents and Adjoints" by Elhoucine Bergou

- Problem statement: for a stochastic nonlinear system, find the best estimate for the trajectory evolution.
- Methods proposed
  - Globalization methods
  - LM-ENKF methods
- Summary
  - Solve the linear least-squares from 4DVar by EnKS, naturally parallel over the ensemble members.
  - Linear algebra glue is cheap
  - Finite dfferences $\rightarrow$ no tangent and adjoint operators needed.
  - Add Tikhonov regularization to the linear least-squares $\rightarrow$ Levelberg-Marquardt method, guaranteed convergence.
  - Cheap and simple implementation of Tikhonov regularization within EnKS as an additional observation.

3.4 Bayesian Approach

3.4.1 "New Developments in Ensemble Kalman Fltering and Smoothing Where a Bayesian Approach Gives an Advantage" by Marc Bocquet

- Motivation: Getting more from the ensemble
  - Idea: even under Gaussian assumption of the true distribution, the pdf extracts more information than just mean and covariance.

  - : Using Gaussian assumptions, and being only interested in the filtering problem, one can get more information on the pdf.

- Approach: The Primal EnKF-N, Dual EnKF-N, Iterative EnKF, Iterative EnKS
- Conclusions:
  - The finite-size scheme (EnKF-N) is a no-inflation scheme that accounts for sampling errors and is as efficient as optimally tuning inflation on geophysical toy models. It is actually equivalent to an adaptive inflation scheme.
  - It also helps understand whats wrong in the EnKF, and why an hyperprior underlies this DA method.
  - The iterative ensemble Kalman filter has been generalised to an iterative ensemble Kalman smoother (IEnKF). It is an En-Var method. It is tangent linear and adjoint free. It is, by construction, flow-dependent.
  - Though based on Gaussian assumptions, it can offer (much) better retrospective analysis than standard Kalman smoothers in mildly nonlinear conditions.
  - When affordable, it beats other Kalman filter/smoothers in strongly non-linear conditions.

3.4.2 "Bayesian Learning of Stochastic Dynamical Models: State, Parameters, also Model Formulations?" by Pierre Lermusiaux

- Motivation: KL-based uncertainty prediction schemes
- Holistic Challenge in Ocean (Bayesian) Estimation
  - Prognostic Equations for Stochastic Fields of Large-Dimension
  - Non-Gaussian Data Assimilation
  - Learn/Predict the Optimal Model and Optimal Data (Adaptive Modeling and Adaptive Sampling)
- Conclusion
  - Prognostic DO Equations for Stochastic Fields: Applications to several 2D NS/cases; Adapt the size of the subspace (continuous criterion) + DO numerics
  - GMM-DO Data Assimilation: Mixtures fit in stochastic subspace (using EM and BIC for now); Bayesian non-Gaussian update in the subspace; Much better than EnKF & ESSE when: sparse/noisy data, limited ensemble size.
  - Augment Modeling with Machine Intelligence; Learning Models: Bayesian inference of state, parameter and model eqs.; Path Planning, Adaptive Sampling

3.4.3 "On Non Gaussian Ensemble Data Assimilation" by Emmanel Cosme

- Motivation: improved data assimilation using ensemble
- Proposed methods
  - Univariate Rank Histogram Filter
  - Multivariate Rank Histogram Filter: Scheme 1 and Scheme 2
  - EnKF with Gaussian anamorphosis
- Conclusion on the mutivariate rank histogram filter
  - The analysis scheme is utterly deterministic;
  - Localization is natural;
  - Divergence is almost impossible for observed variables;
  - Flexible in terms of implementation
  - Balance between expense and performance (Scheme 1 and Scheme 2) to be acheieved

3.4.4 "Application of the Implicit Particle Filter to a Model of Nearshore Circulation" by Robert Miller

- Motivation: Application of Particle Filter to a nearshore circulation
- Model: a shallow-water model: highly nonlinear with large dimension
  - The linearized system is unstable, and the calculations blow up in a time comparable to the assimilation cycle
  - Interesting behavior of 4DVAR
  - Two distinct cases are considered: i) High drag case, regular wavelike flow; 2) Low drag case, aperiodic flow
  - Assimilation fails for low drag case with assumption of steady forcing
  - Must use incorrect assumption of unsteady forcing to get a solution to 4DVAR in the low drag case
- Data assimilation method: Implicit Particle Filter
  - Generate a random vector of state dimension;
  - Calculate the most probable state given the initial value and the minimizer
  - Choose the updated state so that each particle satisfy the propoer conditions
- Conclusions
  - The good news: The implicit particle filter, (in this case, the optimal importance filter) can be implemented efficeintly on models of geophysical interest; The resulting analysis looks good
  - The bad news: We are still cursed by dimensionality!
  - Next steps (no particular order): Sparse observations in time; Direct appeal to dynamical structure; Parameter estimation

3.5 Challenges

3.5.1 "Conditions for Successful Data Assimilation" by Mattias Morzfeld

- Motivation:
  - Use a mathematical model to make predictions about a physical process by assimilating noisy data to the model
  - Focus on the problem that small errors grow quickly and become large errors
- Objective: What are the conditions under which DA can be successful?
- Summary: Analysis for linear Gaussian case only. Nonlinear/non-Gaussian problems must be analyzed in each particular case.
  - Numerical data assimilation hopeless unless effective dimension is small (probability mass is concentrated on a low dimensional manifold)
  - Boundedness of effective dimension induces balance condition between errors in model and data
  - In practice, effective dimension often small because correlations in errors
  - Particle filters can work in high dimensions, provided their implementation is sound
  - Variational data assimilation requires well boundedness of covariance matrix, i.e. balance condition between errors in prior, model and data

3.5.2 "Implicit Parameter Estimation" by Brad Weir

- Motivation
  - Parameters control the growth/death rates of species and their interactions
  - Little to no a priori knowledge
  - Many are impossible to determine from in situ measurements alone
  - Models combine different species into functional groups: Parameters determine dominant species and their behavior; Fewer groups = stronger parameter dependence on specific ecosystem
  - Assimilate data to find appropriate estimates
- Approach: particle filter with implicit sampling
  - Nonparametric: strong theoretical basis for nonlinear/non-Gaussian problems
  - Generally applicable: smoother and filter forms; state and/or parameter estimator; applicable to deterministic and stochastic models
  - Optimized for observations: explores important regions in sample space; does not blindly explore space and eliminate improbable samples (like many particle filters and MCMC methods)
  - Many implementations: allows problem-specific tuning (hint ...)
- Conclusions
  - Implicit sampling is theoretically applicable to state and parameter estimation in a very general setting
  - In strongly non-Gaussian problems, can use a Robbins-Monro iteration to refine the Hessian and generate samples with acceptable weights
  - Refinement and sampling significantly improves the confidence limits from those given by local Gaussian assumption
  - If chlorophyll is the only information about parameters, can find more accurate estimates than quadratic/Gaussian interpretation suggests
  - This lets us define ecological regions with greater precision

3.5.3 "Statistical Data Assimilation for Biological Ocean Models" by Mike Dowd

- Motivation for biological ocean problems (ocean biology embeded with ocean circulation models) with biological processes sucah as plankton dynamics and nutrient cycling
  - Joint parameter and state estimation: Typically emphasize retrospective (hindcast) studies.
  - Issues: Large-scale estimation problems, uncertain governing equations and parameters, complex observation errors
- Approach
  - Particles Filters
  - Location particle smoother
  - Emulators for uncertainty analysis/parameter estimation
  - Copulas for predictive distribution (model errors)
- Summary
  - Hierarchical Bayesian framework conceptually useful $\rightarrow$ particle MCMC for dynamic models, Practically ... /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 characterizaCon important. How to do represent stochastic process (via samples, distribuCons).
- Role of Emulators? Need to incorporate emulator error in hierarchy and provide for efficient construruib
- Validation? Design for sample-based numerical experimental for assessing consistency, efficiency, asymptotic, robustness.
- Emerging approaches for dynamic systems from statistics.

3.5.4 "Data Analysis in Low Noise Regime" by Jonathan Weare

- Motivation: Regime transition happens
  - How does this event occur, i.e. what rearrangements have to happen to trigger the event?
  - How does the frequency or severity of the event depend on various environmental parameters?
  - Can we predict the event from data in real-time?
- Approach: Two-step recursive filtering procedure
  - Starting from an ensemble of copies evolve each copy forward to the next observation time (ti+1) and compute the contribution to the weight from the next observation
  - Resample the copies according to the weights, i.e. duplicate copies with large weights and eliminate copies with low weight.

  Repeat for next observation. Various approximations are needed to make this scheme practical on large problems (e.g. the ensemble Kalman filter).
- Theoretical points
  - Vanishing statistical error in contrast to the exponentially growing error before.
  - Errors introduced by numerical discretization can be addressed

3.5.5 "Multi-scale Data Assimilation" by Kayo Ide

- Motivation
  - As the model advances and becomes capable of complex physical processes, ocean model dynamics exhibits multi-scale phenomena.
  - Observing system network come in a variety of resolution and inhomogeneity.
- Multiscale (MS) data assimilation formulation: Handle scale by scale, sequentially from large to small scale.
  - MS Partition of the state and observations, and respective error covariance matrices
  - Handling of the representativeness error
- Demonstration
  - Idealized experiments: how MS handles scales better
  - Observing system experiments using Regional Ocean Modeling System (ROMS): 1) Observing System Simulation Experiments (OSSEs) using realistic observing system network; and 2) experiments with real data
- Discussion and challenges
  - MS 3D-Var works well

  - Two main elements: 1) Successive application of localization from Large-scale to smaller scales; 2) Separation of observing system network
  - Using real data and validation against independent observations show significant reduction of bias by the MS scheme.
  - On-going extensions: EnKF and Hybrid

3.5.6 "Incorporating Representativity Error in Data Assimilation" by Nancy Nichols

- Background: Errors of Representativity
  - Data assimilation combines observations with a model prediction.
  - Observations can contain information at smaller scales than the model can resolve.
  - Errors of representativity are the result of small scale information in observations being incorrectly represented in the model.
- Approach
  - Investigate the structure and properties of representativity errors;
  - Incorporate representativity errors in data assimilation.
- Structure of Representativity Error
  - Correlated and state and time dependent;
  - Reduced by increasing model resolution or increasing observation length scale;
  - Depend only on distance between observations and not the number.
  - In atmopshere: vary with height, more significant for humidity than temperature.
- Representativity Error in Data Assimilation. Methodology
  - Select initial observation error covariance matrix
  - Run data assimilation and gather samples of background and analysis innovation
  - Compute the correlation matrix using the samples
  - Symmetrize and localize to obtain new estimate for observation error covariance matrix
  - Repeat steps from rolling window of length
- Summary
  - Time-varying observation error covariance matrices can be estimated
  - Including the estimated observation error covariances in the data assimilation scheme can improve the analysis

4 Scientific Progress Made

The previous BIRS workshop on Mathematical Advancement in Geophysical Data Assimilation (08w5096) initiated a series of new research directions for data assimilation as interdisciplinary science. Significant progress has been made through interactions between mathematics and geosciences. Yet, there are many questions remain to be asked and challenges to overcome.

The objective of this intensive workshop was to advance data assimilation for earth systems by bringing together mathematicians, particularly those working in dynamical and stochastic systems, statisticians, and domain scientists who have a vested interest in fusing data with models. In the highly successful 2008 BIRS workshop, key mathematical questions in data assimilation were identified for atmosphere, ocean, and coupled atmosphere-ocean systems. The years following the workshop have seen a blossoming of new approaches to data assimilation, much of which has proceeded along the lines discussed in depth at BIRS in 2008. In this workshop, we focused on the following topics:

1. Current-State-of-the-Art for the short- and medium-range operational data assimilation;
2. Long-range and climate-scale data assimilation.
3. Flow-dependent covariance estimation
4. Bayesian Approach
5. Challenges

5 Outcome of the Meeting

The objective of this workshop was explore and discuss areas and specific problems in which collaborative efforts with the mathematical community could help to address fundamental and challenging issues in data assimilation in the interdiscplinary setting. The BIRS workshop brought together practioners of data assimilation with mathematicians and statisticians at a place where the intensive focus and energy could serve to define the way forward. It offered a unique and much needed opportunity to go beyond what we were able to achieve in the previous program. The outcome of the workshop is expected once again to lead to significant contributions to data assmilation of the earth system through the development of new statistical, dynamical and computational strategies.

During the workshop, several issues were raised and discussed. The interdisciplinary data assimilation community is now at a point to start working outside of the existing frameworks.

References

[1] R. E. Bellman, *Adaptive control processes: A guided tour*, Princeton University Press, Princeton, NJ, 1961.

[2] K. Ide, P. Courtier, M. Ghil and A. C. Lorenc, Unified notation for data assimilation: Operational, sequential and variational, *J. Meteorol. Soc. Jpn.* **75** (1997), 181 – 189.

[3] R. N. Miller, E. F. Carter and S. T. Blue, Data assimilation into nonlinear stochastic models, *Tellus*, **51A** (1999), 167 – 194.

[4] C. Snyder, T. Bengtsson, P. Bickel and J. Anderson, Obstacles to high-dimensional particle filtering, *Mon. Wea. Rev.* **136** (2008), 4629 – 4640.

[5] B. Weir, R. N. Miller and Y. H. Spitz, Implicit estimation of ecological model parameters, *Bull. Math. Biol.* **75** (2013), 223-257.