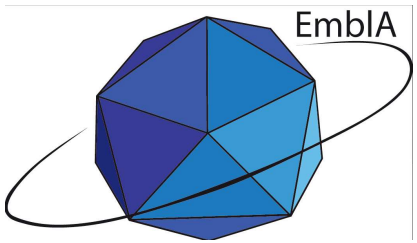
Data Assimilation in Geoscience: Atmosphere and Ocean

Alberto Carrassi

Nansen Environmental and Remote Sensing Center - Norway



Data Assimilation group: Laurent Bertino (Lead), Alberto Carrassi (Co-Lead), Colin Grudzien (PD), Patrick Raanes (PD), Matthias Rabatel (PD), Abhishek Shah (PhD) and Maxime Tondeur (MsC).

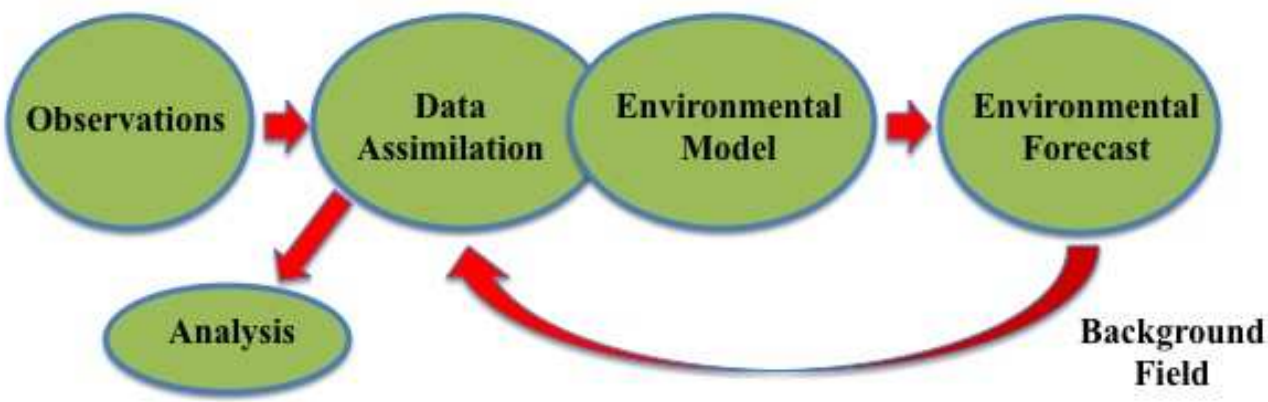
Indo-Norwegian Winter School on Operational Oceanography

INCOIS, Hyderabad - India 17^{th} October 2015

Data Assimilation - Overview

Data Assimilation *is the entire sequence of operations that, starting from the observations and possibly from a statistical/dynamical knowledge about a system, provides an estimate of its state*

- numerical weather & ocean prediction
- hydrology
- you name it...

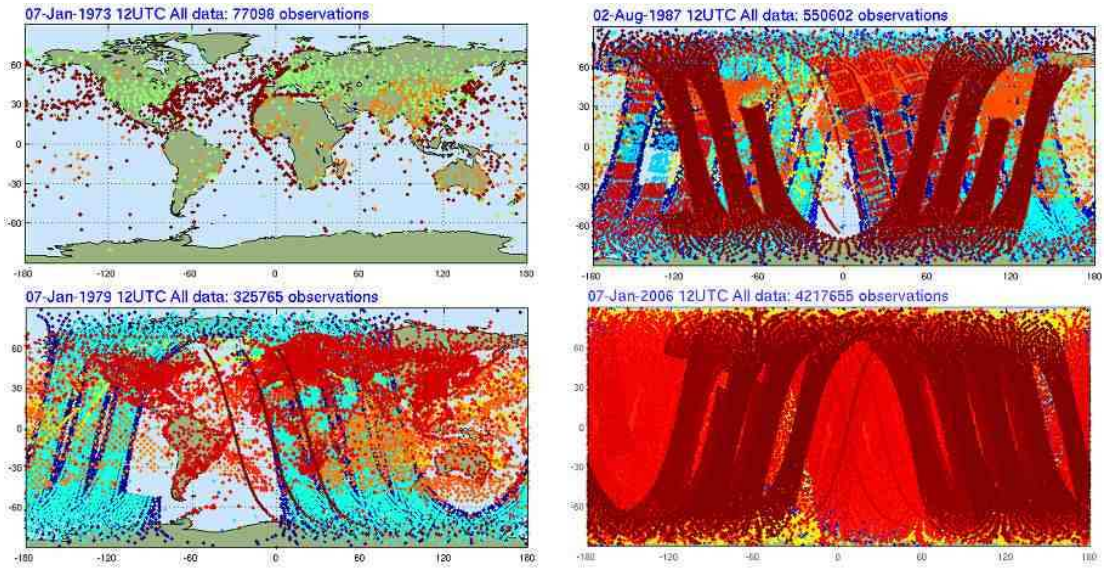


Data Assimilation - Overview

Typical sources of information are:

- observations (synoptic profiles, onboard measurements, remote sensing, etc...)
- background field (climatological, short range forecast)
- evolution dynamics (set of differential equations, numerical model ...)

All these information are combined in a statistical fashion to obtain the best-possible estimate the **analysis**



Data Assimilation Theory & Methods

1 - Data Assimilation Theory & Methods

Basic Definitions and Problem Statement

OBJECTIVE:

estimate the state of an unknown system based on an imperfect model and a limited set of noisy observations:

$$\mathbf{x}_k = \mathcal{M}_k(\mathbf{x}_{k-1}) + \mu_k \quad k = 1, 2, ...,$$

$$\mathbf{y}^o_k = \mathcal{H}(\mathbf{x}_k) + \varepsilon^o_k \quad k = 1, 2, ...,$$

- $\mathbf{y}^o \in \mathcal{R}^p$ and $\mathbf{x} \in \mathcal{R}^n$ - $p \ll n$ in realistic geophysical applications
- $\{\mu_k\}_{k=1,2...}$ and $\{\varepsilon^o_k\}_{k=1,2...}$ *assumed* to be random error sequences, white in time, and uncorrelated between them
- Collect state estimates and observations as: $\mathbf{X}_k = \{\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_k\}$ and $\mathbf{Y}^0_k = \{\mathbf{y}^0_0, \mathbf{y}^0_1, ..., \mathbf{y}^0_k\}$

Smoothing, Filtering or Prediction ?

1. Smoothing $\rightarrow$ Estimate the state at all times $\equiv \mathbf{X}_k$ based on $\mathbf{Y}^0_k$
2. Filtering $\rightarrow$ Sequential Estimate of the present state
3. Prediction $\rightarrow$ Estimate the state at future times $\equiv \mathbf{x}_{k>l}$ based on $\mathbf{Y}^0_l$

Probabilistic Approach

In the probabilistic framework, problems (1)-(2)-(3) are expressed as the estimation of the corresponding conditional probability density functions:

1. Smoothing $\rightarrow$ Estimate the state at all times $\equiv \mathbf{X}_k$ based on $\mathbf{Y}_k^0 \rightarrow \mathcal{P}(\mathbf{X}_k|\mathbf{Y}_k^0)$
2. Filtering $\rightarrow$ Sequential Estimate of the present state $\rightarrow \mathcal{P}(\mathbf{x}_k|\mathbf{Y}_k^o)$
3. Prediction $\rightarrow$ Estimate the state at future times $\equiv \mathbf{x}_{k>l}$ based on $\mathbf{Y}_l^0 \rightarrow \mathcal{P}(\mathbf{x}_{k>l}|\mathbf{Y}_l^0)$

The PDFs $\mathcal{P}$ fully characterise the estimation problem!

The error PDFs associated to all the information sources read:

- $\mathcal{P}(\mathbf{x}_0)$ PDF of the initial conditions - Prior/Background
- $\mathcal{P}(\mu_k) = \mathcal{P}(\mathbf{x}_k - \mathcal{M}_k(\mathbf{x}_{k-1})) = \mathcal{P}(\mathbf{x}_k|\mathbf{x}_{k-1})$ - Model Error PDF
- $\mathcal{P}(\varepsilon_k^o) = \mathcal{P}(\mathbf{y}_k^0 - \mathcal{H}(\mathbf{x}_k)) = \mathcal{P}(\mathbf{y}_k|\mathbf{x}_k)$ - Observational Error PDF

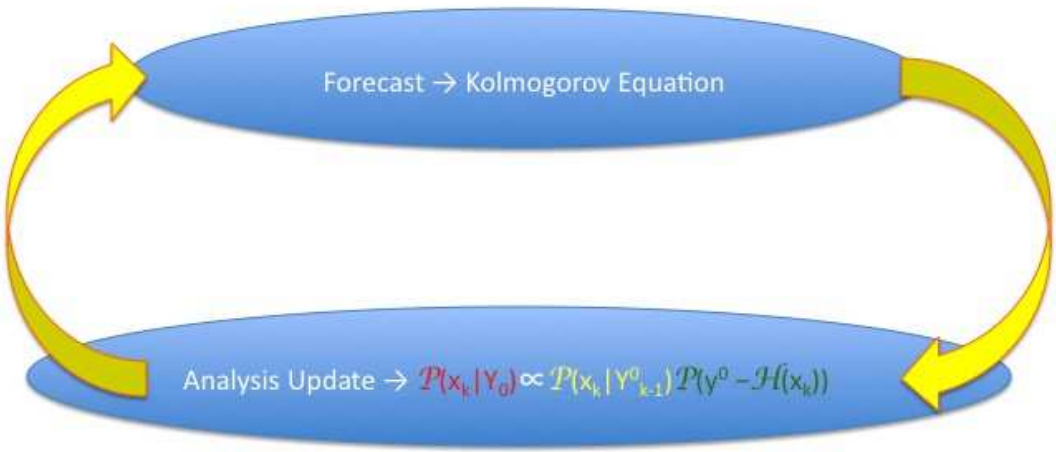
Probabilistic Approach

With Bayes's rules....

SMOOTHING

$$\mathcal{P}(\mathbf{X}_k|\mathbf{Y}_k^0) \propto \mathcal{P}(\mathbf{x}_0)\Pi_{i=1}^{k}\mathcal{P}(\mathbf{x}_i - \mathcal{M}_i(\mathbf{x}_{i-1}))\mathcal{P}(\mathbf{y}_i^0 - \mathcal{H}(\mathbf{x}_i))$$

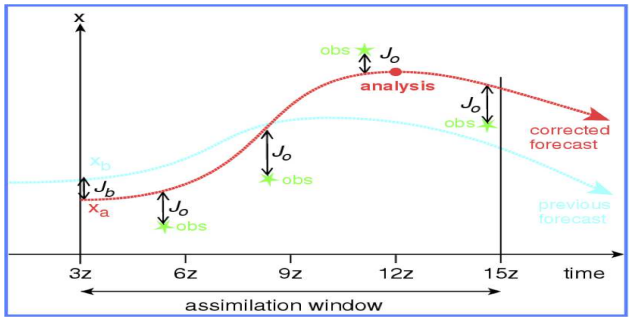
FILTERING



- The Bayesian framework ideally solves the inference problem BUT it is hardly affordable in geoscience
- The Particle Filter attempts to solve this problem and its potential application in geoscience has received much attention in recent years. See van Leuween, 2009 (MWR) for a review of PF in Geosciences. I will take this point later ...

A Gaussian World: 4D-Variational Assimilation

Under Gaussian assumptions things turn much easier $\Rightarrow$ PDFs are described with only MEAN and COVARIANCE



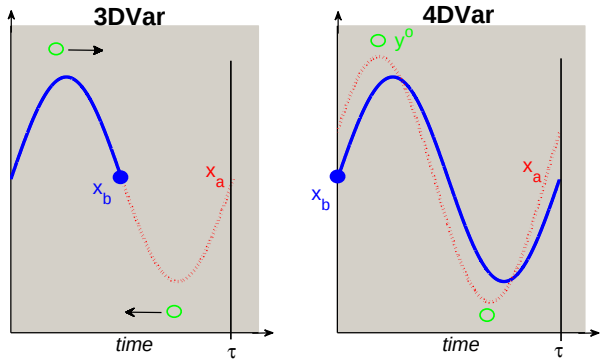
Initial condition, observational and model errors are all Gaussian and mutually uncorrelated $\Longrightarrow$ solving the SMOOTHING problem leads to the 4DVar formulation, *i.e.* minimise a penalty function as:

$$2J = \sum_{i=1}^{k} \mu_i^T (\mathbf{P}^m)_i^{-1} \mu_i + \sum_{i=1}^{k} [\mathbf{y}_i^0 - \mathcal{H}(\mathbf{x}_i)]^T \mathbf{R}_i^{-1} [\mathbf{y}_i^0 - \mathcal{H}(\mathbf{x}_i)] + (\mathbf{x_0} - \mathbf{x_b})^T (\mathbf{P}^f)^{-1} (\mathbf{x_0} - \mathbf{x_b})$$

- $\mathbf{P}^f$ - Background error covariance matrix
- $\mathbf{R}$ - Observational error covariance matrix
- $\mathbf{P}^m$ - Model error covariance matrix

4D-Variational Assimilation - Some Facts

- The sequence (trajectory) $\mathbf{X}_k$ which minimizes J is the maximum likelihood estimator of the PDF $\mathcal{P}(\mathbf{X}_k|\mathbf{Y}_k^0)$
- It provides the *"best"* possible fit to the observations, given the initial guess and the *imperfect* model
- The strong-constraint 4DVar makes the assumption of perfect model and the latter is appended as a strong-constraint when doing the minimization
- The minimization of J can be done in principle by solving the associated Euler-Lagrange (EL) equations (Le Dimet and Talagrand, 1986 Tellus)
- The *Method of Representer* is an efficient way to solve the EL eqs for linear dynamics (Bennett, 1982, chapter 5)
- *Descent Methods* are used in the case of large nonlinear systems (Talagrand and Courtier, 1987 QJRMS)
- The choice of the *Control Variable* defines the size of the problem to be solved and characterises different formulations of the 4DVar (see *e.g.* Trémolet, 2006; Bocquet, 2009)
- $\mathbf{P}^f$ is implicitly evolved within the assimilation window but it is not available for the next analysis cycle
- When observations are assimilated (as they were) at the same time the 3DVar is recovered
- 4DVar (under "strong" simplified assumptions) is operational in several weather services, among them MetOffice, ECMWF and Meteo France.



A Gaussian World: The Kalman Filter

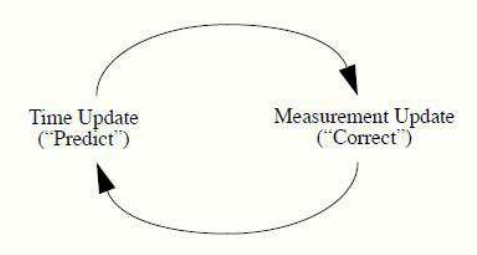
Under the same hypotheses of Gaussianity and mutual uncorrelation of errors the filtering problem reduces to the estimation of the mean and covariance.

Forecast:

$$\mathbf{x}_k^f = \mathbf{M}\mathbf{x}_{k-1}^f + \mu_k$$
$$\mathbf{P}_k^f = \mathbf{M}_k\mathbf{P}_{k-1}^a\mathbf{M}_k^T + \mathbf{P}^m{}_k$$

Analysis:

$$\mathbf{x}_k^a = \mathbf{x}_k^f + \mathbf{K}_k\left[\mathbf{y}_k^o - \mathcal{H}_k(\mathbf{x}_k^f)\right]$$
$$\mathbf{P}_k^a = [\mathbf{I} - \mathbf{K}_k\mathbf{H}_k]\mathbf{P}_k^f$$



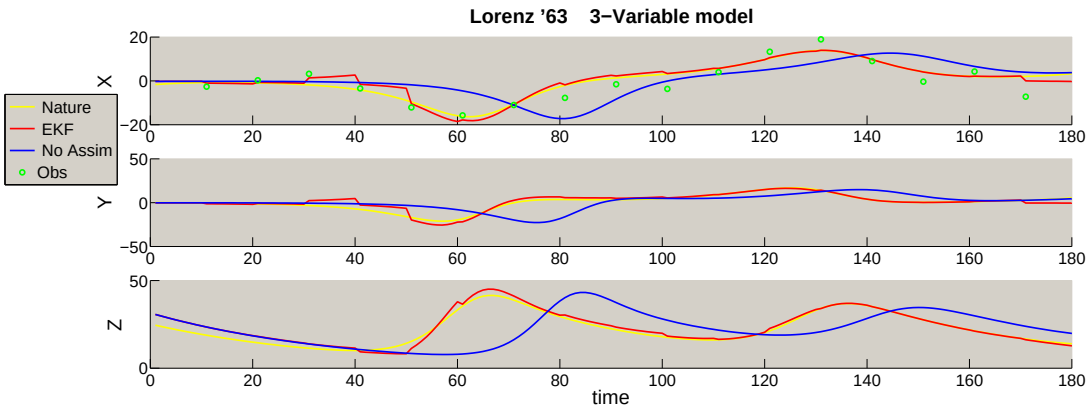
- $\mathbf{x}_k^a/\mathbf{P}_k^a$ - Analysis state/covariance at time t_k
- $\mathbf{x}_k^f/\mathbf{P}_k^f$ - Forecast state/covariance at time t_k
- **Kalman gain matrix** $K = \mathbf{P}^f\mathbf{H}^T[\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R}]^{-1}$
- The analysis $\mathbf{x}^a$ is optimal in the sense that it minimizes the analysis error variance
- When all errors are Gaussian the minimum variance estimate is also the maximum likelihood estimate (out of unimodality maximum likelihood estimators are of questionable relevance)

Kalman Filter (KF) and Extended KF (EKF)

For linear dynamics and observational operator the KF provides a closed set of estimation equations (Kalman, 1960).

Extension to nonlinear dynamics - Extended Kalman Filter

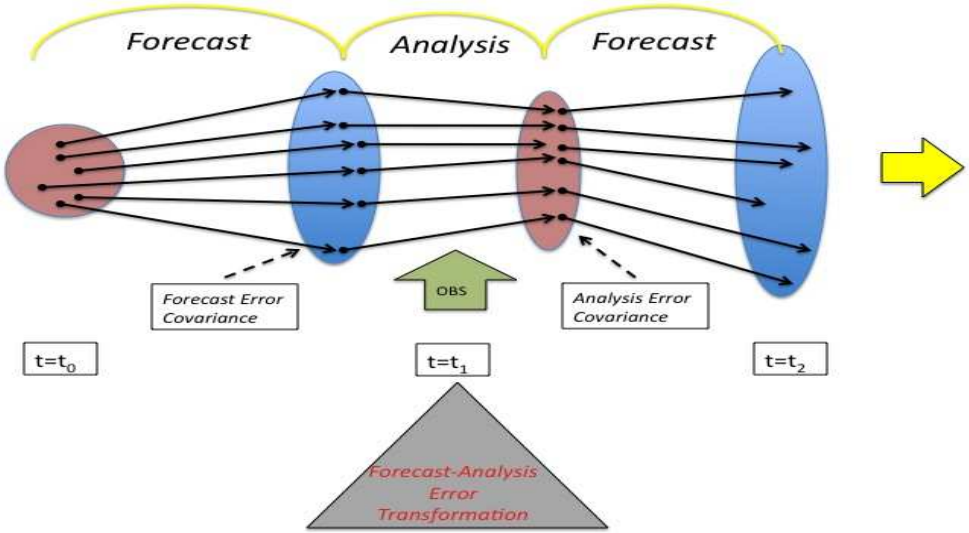
- The extended Kalman Filter (EKF) is a first order approximation of the KF
- The tangent linear model is used to forward propagate the forecast uncertainty (*i.e.* the error covariance)
- The full nonlinear model is used to evolve the state estimate
- The analysis update is the same as in the standard KF
- The introduction of the EKF in geoscience is due to Ghil and Malanotte-Rizzoli (1991) *AdvGeohys*
- The EKF response to different degree of nonlinearity has been studied in Miller, Ghil & Gauthiez (1994) *JAS*
- The EKF is almost-operational for ECMWF soil analysis (*e.g.* de Rosnay *et al.*, 2012 QJRMS)



Ensemble Based Data Assimilation Algorithms

In the ensemble-based DA the forecast/analysis error covariances are approximated using an ensemble of M model trajectories

see *e.g.* Evensen, 2009



- Ensemble based covariances $\mathbf{P}^{f,a} = \frac{1}{M-1}\sum_{i=1}^{M}(\mathbf{x}_i^{f,a} - \bar{\mathbf{x}}^{f,a})(\mathbf{x}_i^{f,a} - \bar{\mathbf{x}}^{f,a})^T$
- In Geoscience $M \ll n$
- Flow dependent description of the forecast error
- Provide automatically a set of initial conditions for ensemble prediction schemes.
- The forecast-analysis transformation characterises the ensemble-based algorithms (*Deterministic* vs *Stochastic*).

Stochastic or Deterministic ?

Ensemble data assimilation algorithms can be divided into Stochastic and Deterministic

Stochastic (Monte-Carlo approach)

- In this class of algorithms the observations are treated as a random ensemble by adding noise at each analysis update
- Each ensemble trajectory assimilates a different realization of the observation vector and undergoes an independent analysis update
- The standard Ensemble Kalman Filter (EnKF) belongs to this family (see *e.g.* Houtekamer and Mitchell, 1998 MWR)
- The EnKF has proved efficiency in a number of geophysical applications (see Evensen, 2003 Ocean Dyn for a review)

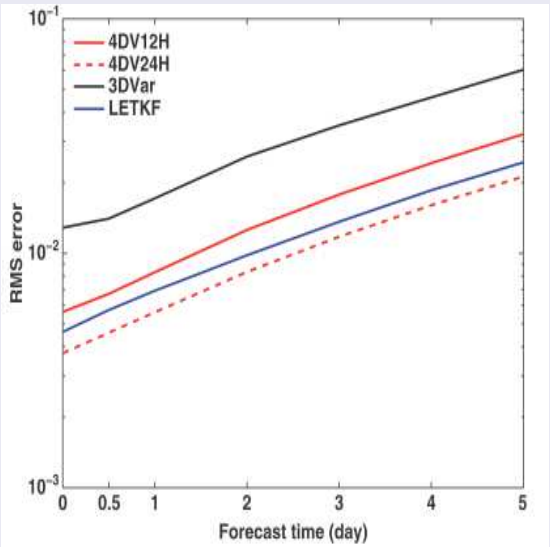
Deterministic (Square-Root approach)

- In this class of algorithms the step $\mathbf{P}^f \rightarrow \mathbf{P}^a$ is made through a linear transformation $\mathbf{T}$
- It avoids the introduction of extra noise at the analysis update
- $\mathbf{T}$ is usually defined under the constraint that $\mathbf{P}^a$ matches some desired value (*i.e.* the EKF one, the Hessian of a penalty function)
- The solution (a square-root matrix) is not unique and the particular choice characterises the algorithm (see Tippet *et al.*, 2003 MWR).
- Algorithms belonging to this family: ETKF, LETKF, EnSRF, MLEF (see Whitaker and Hamill, 2002 MWR; Bishop *et al.*, 2001 MWR; Hunt *et al.*, 2007 Physica D)

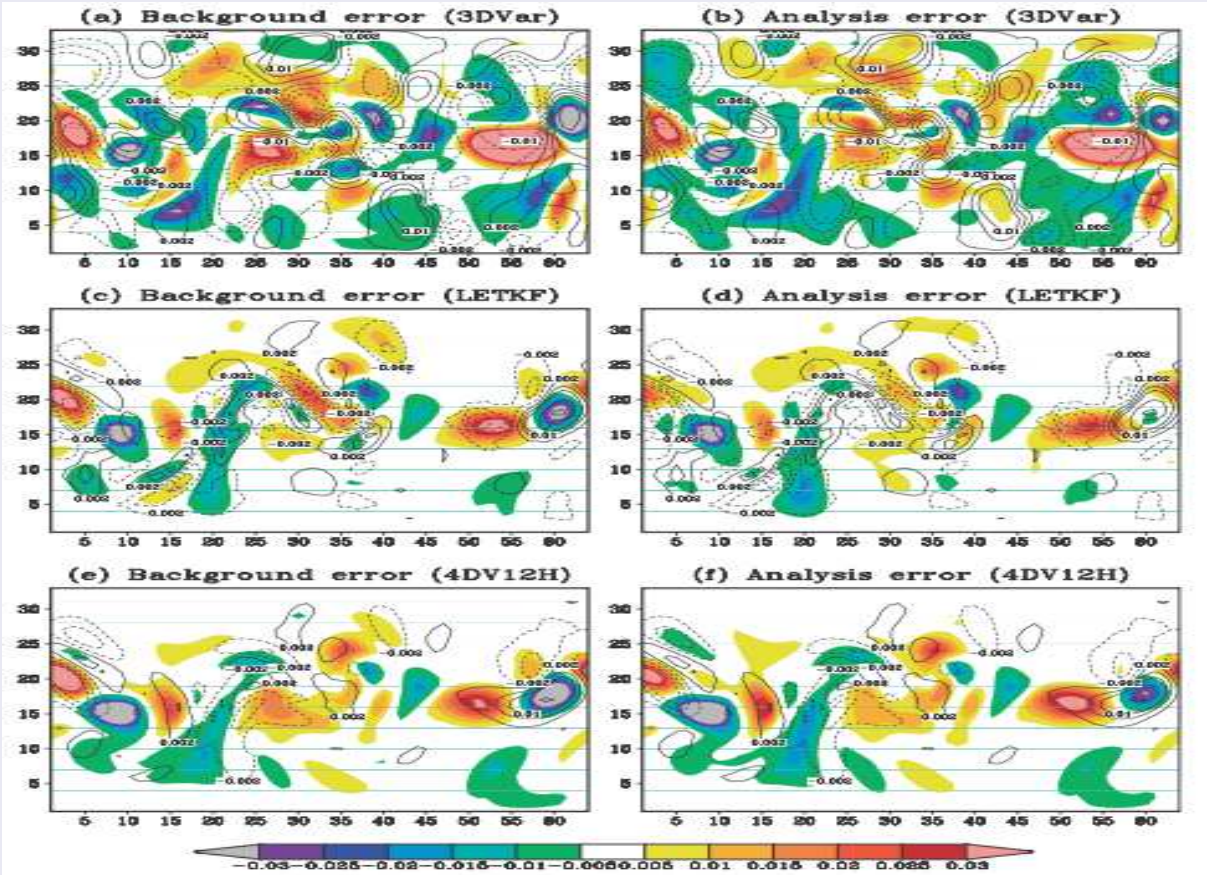
Ensemble-based or Variational: the comparison

- Results with a Quasi-Geostrophic model by Rotunno and Bao, 1996
- Ensemble-based scheme $\Longrightarrow$ Local Ensemble Transform Kalman Filter (Hunt *et al.*, 2007 Physica D)

Forecast Error



- Background (left) & Analysis (right) error in COLOR
- Analysis Increment in CONTOUR



From Yang, Corazza, Carrassi, Kalnay & Miyoshi and Kalnay, 2009

Dealing with Geophysical Systems

When dealing with realistic Atmosphere/Ocean dynamics DA faces a number of obstacles....

- The Atmosphere and the Ocean are example of nonlinear chaotic systems $\Rightarrow$ Flow-dependent description of the estimation error (EnKF, MLEF, AUS ...)
- Sources of nonlinearities: model $\mathcal{M}$, obs operator $\mathcal{H}$, first guess $\mathbf{B}$. Nonlinearities *"push out"* of Gaussianity $\Rightarrow$ Non-Gaussian analysis framework (see *e.g.* Fletcher and Zupanski, 2006 QJRMS and Bocquet et al., 2010 MWR)
- Models are not perfect - incorrect parametrizations of physical processes, numerical discretizations, unresolved scales, etc..
- Data Assimilation in geoscience is a Big Data problem $\Rightarrow$ Computationally suitable solutions (see Fisher talk at WMO-DA Symposium 2013)
- Some physical quantities are bounded distributed $\Rightarrow$ Adapt Gaussian methods to incorporate observations with limited range (Anamorphosis, Bertino XXX; See Abhiskek Shah poster !)

1. How Ensemble and Variational Methods have dealt with these issues ?
2. *Control of Chaos* $\Rightarrow$ DA methods for chaotic dynamics
3. *Model Error* $\Rightarrow$ DA methods accounting for model error

Dealing with Geophysical Systems

2 How data assimilation methods have dealt with the special problems encountered in Geosciences

2.1 - Variational Methods Adaptation
2.2 - Ensemble Methods Adaptation
2.3 - DA for chaotic systems
2.4 - DA with imperfect models - Treatment of model error in DA

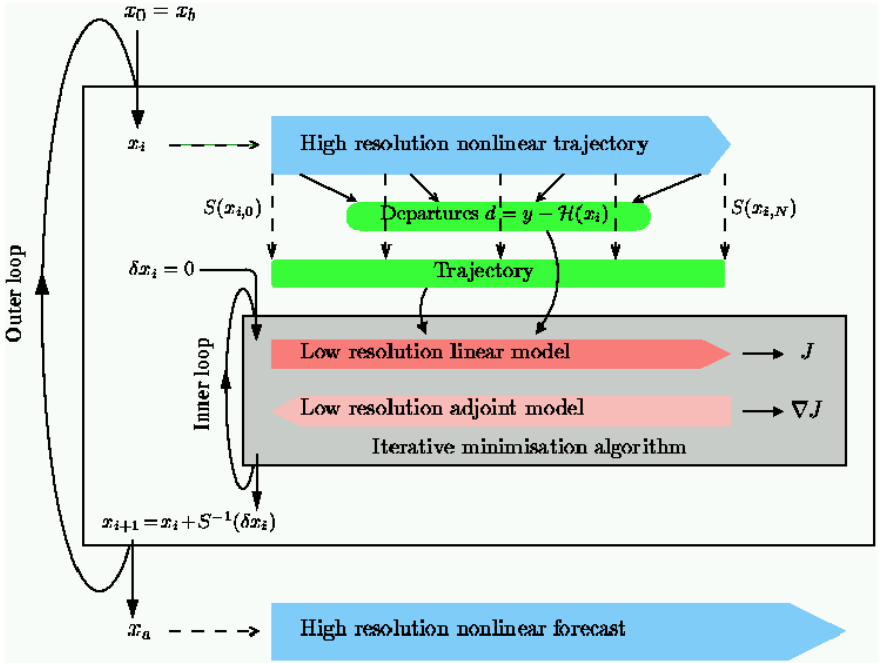
2.1 How to deal with geophysical systems: Variational

Main Drawbacks of Variational Approach:

1. Non-Quadratic cost-function in 4DVar
2. with possible Multiple Minima
3. maximum likelihood approach questionable
4. No flow-dependent error description

Proposed Solutions:

- Problem (1) and (2) are alleviated in the Incremental 4DVar (Courtier *et al.*, 1994 QJRMS).



From Andersson et al., 2005 ECMWF-Tech.Rep. 479

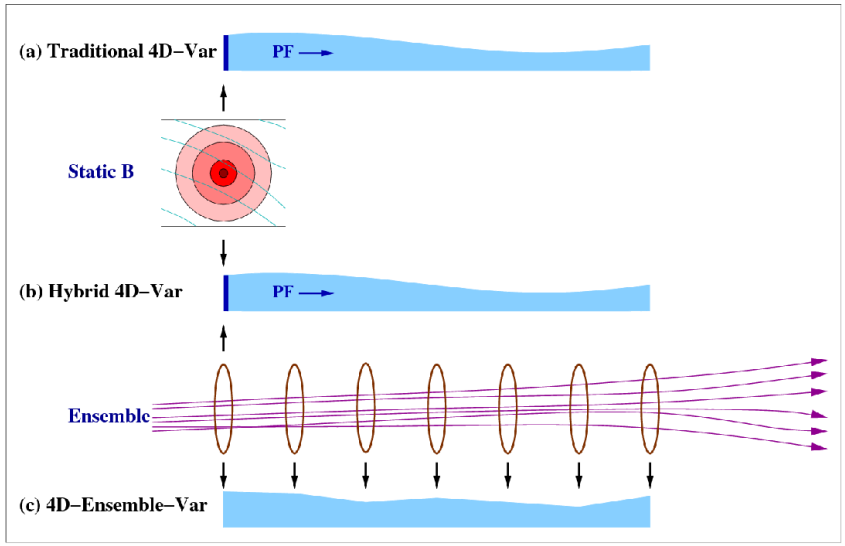
How to deal with geophysical systems: Variational

- Problem (4) is implicitly overcome with the Long Window 4DVar but ... problems (1)-(3) can be made worst
- Problems (1)-(3) are partly solved using the Weak-Constraint 4DVar but ... appropriate model error covariances need to be prescribed and the size of the control variable too big (see *e.g.* Trémolet, 2006; Carrassi and Vannitsem, 2010)
- Hybrid 3/4DVar-Ensemble algorithms attempt to tackle all problems at the same time (see Barker and Clayton, 2011 ECMWF Ann. Seminar for a review and for details on the operational implementation at MetOffice).

Example: ETKF $\leftrightarrow$ 4DVar at MetOffice (from Barker and Clayton, 2011 ECMWF Ann. Seminar)

Two hybrid strategies:

- Hybrid 4DVar operational at MetOffice (Use a combination of static and ensemble cov at the initial time)
- 4D-Ensemble-Var mid-long term development (Use ensemble cov within the entire assimilation window $\Rightarrow$ No need for Tangent/Adjoint model) See Buehner *et al.*, 2010 MWR



From Barker and Clayton, 2011 ECMWF Ann. Seminar

How to deal with geophysical systems: Ensemble Schemes

Main Drawbacks of Ensemble Based Approach:

1. Sampling Error (M $O(100)$)
2. Use only observations at analysis time
3. Only the Gaussian approximation of the flow-dependent $\mathbf{P}^f$ is accounted for at the analysis update

Proposed Solutions:

- Sampling errors (problem (1)) are mitigated using Covariance Localization $\Rightarrow$ Effective increase the rank of $\mathbf{P}^f$; but:
  - dynamical consistency is broken
  - the actual optimal size for the localization is time-dependent $\Rightarrow$ Flow-Dependent Covariance Localization (Bishop and Hodyss, 2011)
- Variance Underestimation (still problem (1)) $\Rightarrow$ Multiplicative or Additive Inflation
  Multiplicative Inflation (See *e.g.* Anderson and Anderson, 1999 MWR):
  - $\mathbf{P}^f \rightarrow (1+\alpha)\mathbf{P}^f$
  - keep the same rank/structure of $\mathbf{P}^f$, only the explained variance is modified
  - the inflation can be made adaptive $\Leftrightarrow$ more inflation where/when required: based on Kalman gain (Sacher and Bartello, 2008 MWR), on analysis error variance (Whitaker and Hamill, 2012 MWR)

  Additive Inflation:
  - add random noise to $\mathbf{P}^f$ or $\mathbf{P}^a$
  - the process introduce new structures in the error space spanned by the ensemble covariances
  - a combined additive/multiplicative scheme has been proposed by Zhang *et al.*, 2004 MWR
  - an ensemble based algorithm without the need of inflation has been proposed recently (Bocquet, 2011 NPG)
- An Hybrid approach is used to deal with problem (2) $\Rightarrow$ Several ensemble schemes introduce the time dimension to assimilate observations simultaneously over a given reference period (see *e.g.* Hunt *et al.*, 2004 Tellus; Sakov *et al.*, 2010 Tellus)
- Solution to problem (3) $\Rightarrow$ Particle Filters but ....

DA for chaotic dynamics

A data assimilation method explicitly designed for chaotic systems...

Assimilation in the Unstable Subspace $\Leftrightarrow$ Confine the analysis correction in the unstable subspace

- The growth of the initial uncertainty strongly projects on the unstable manifold of the forecast model.
- The AUS approach consists in confining the analysis update in the subspace spanned by the leading unstable directions $\mathbf{E}$; the analysis solution reads:

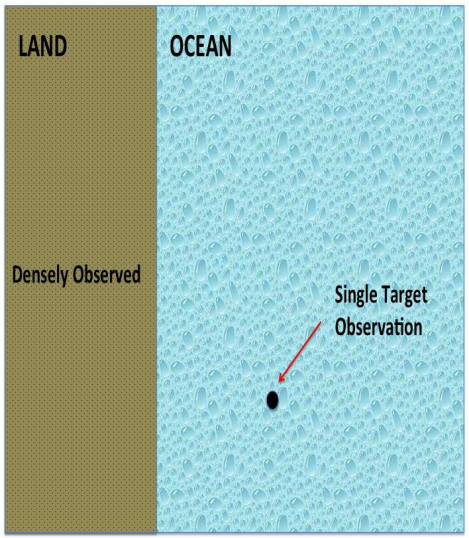
$$\mathbf{x}^{\mathbf{a}} = \mathbf{x}^{\mathbf{b}} + \mathbf{E}\Gamma\mathbf{E}^T\mathbf{H}^T(\mathbf{R} + \mathbf{H}\mathbf{E}\Gamma\mathbf{E}^T\mathbf{H}^T)^{-1}(\mathbf{y}^o - \mathbf{H}\mathbf{x}^{\mathbf{b}})$$

- While all assimilation methods, more or less implicitly, exert some control on the flow dependent instabilities, AUS exploits the unstable subspace, as key dynamical information in the assimilation process.
- Applications to atmospheric and oceanic models showed that even dealing with high-dimensional systems, an efficient error control can be obtained by monitoring only a limited number of unstable directions. (See Palatella, Carrassi, Trevisan, 2013 J. Phys. A)

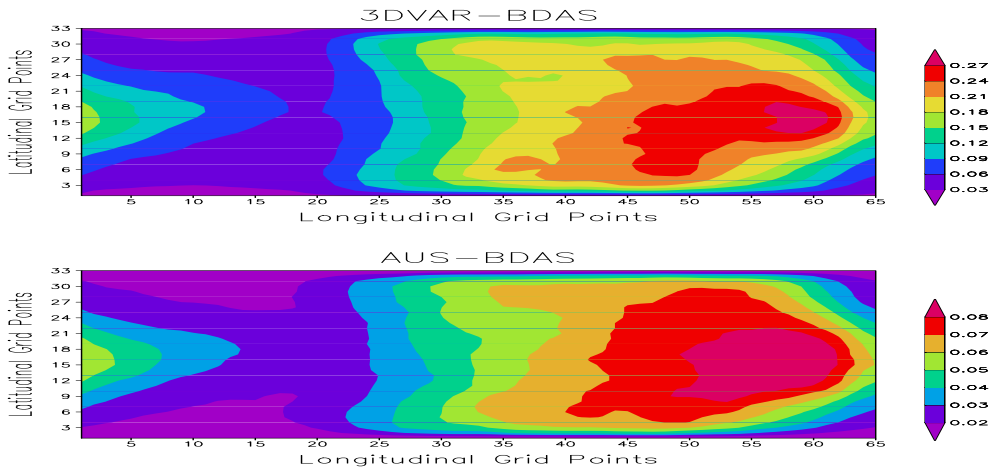
AUS and Target Observations

TARGET OBSERVATION STRATEGY: Breeding on the Data Assimilation System BDAS

- Quasi-geostrophic atmospheric model (Rotunno and Bao, 1996 MWR)
- Perfect model setup - Observation Dense area (1-20 Longitude) - Target Area, one obs between 21-64 Longitude



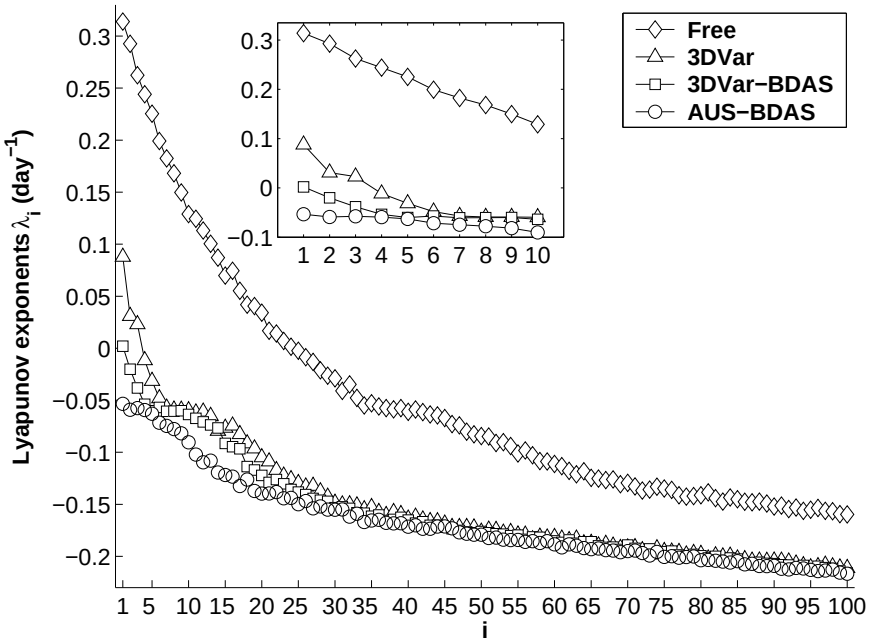
Experiment	Ocean Obs Type/Positioning/Assimilation	RMS Error
LO	-	0.462
FO	vert.Prof/fixed(in the max(err))/3DVar	0.338
RO	vert.Prof/random/3DVar	0.311
3DVar-BDAS	vert.Prof/BDAS/3DVar	0.184
AUS-BDAS	temp.1-Level/BDAS/AUS	0.060



Carrassi et al., 2007 Tellus

DA as a nonlinear stability problem

Can efficient DA methods be constructed to achieve the asymptotic stabilization of the system ?



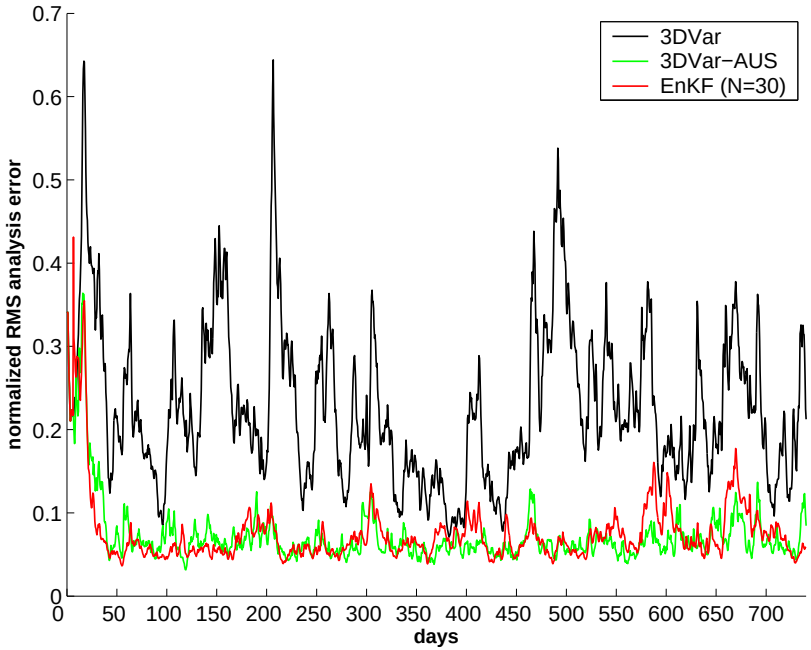
adapted from *Carrassi, Ghil, Trevisan & Uboldi, 2008 CHAOS*

- DA provides a stabilizing effect (compare 3DVar with free system Lyapunov spectrum) but ...
- if the DA is designed to kill the instabilities, the estimation error is efficiently reduced

Hybrid 3DVar - AUS

Enhancing the performance of a 3DVar by using AUS

Comparison with EnKF



adapted from *Carrassi, Trevisan, Descamps, Talagrand & Uboldi, 2008 NPG*

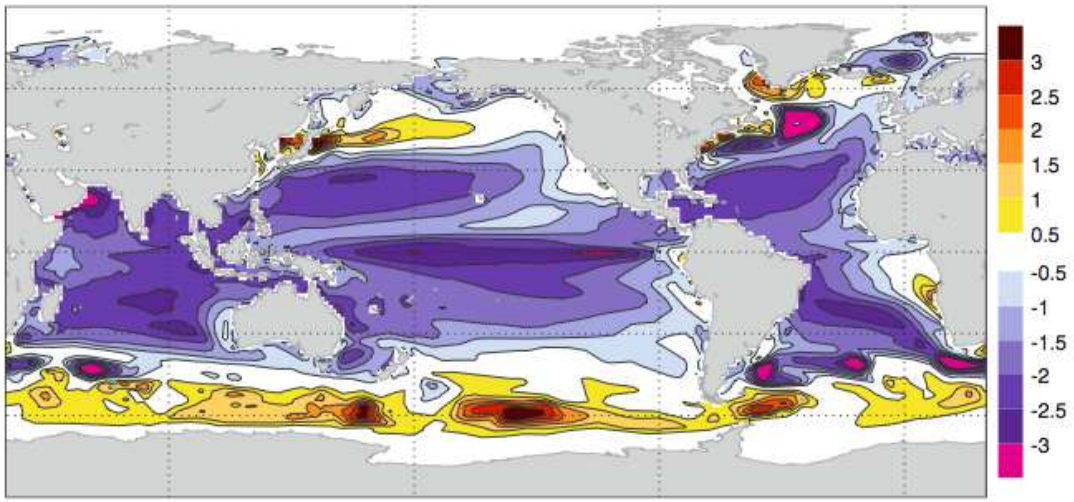
- A network of randomly distributed obs (vertical soundings)
- 3DVar-AUS: (1) AUS assimilate the obs able to control an unstable mode; (2) 3DVar process the remaining obs
- 3DVar-AUS comparable to EnKF with only one BDAS mode $\Rightarrow$ Reduced computational cost and implementation on a pre-existing 3DVar scheme

controlling errors: **what about model error ?**

In the past, model error has been considered small with respect to the (growth of) initial condition error, and thus often neglected

Nowadays model error is recognized as a main source of uncertainty in NWP, seasonal and climate prediction

- In DA for NWP the presence of model error may cause underestimation of the variance (inflation ...)
- On seasonal to climatic timescales model error becomes more evident, through the emergence of biases



SST bias - fcst year 14 - 23. ECMWF IFS model coupled with NEMO ocean model.

Adapted from Magnusson et al., 2012

controlling errors: **what about model error ?**

Fundamental problems making difficult an adequate treatment of model error in data assimilation:

- large variety of possible error sources (incorrect parametrizations of physical processes, numerical discretizations, unresolved scales, etc..)
- the amount of available data insufficient to realistically describe the model error statistics
- lack of a general framework for model error dynamics

OBJECTIVES

1. Identifying some general laws for the evolution of the model error dynamics (with suitable *application-oriented* approximations)
2. Use of these dynamical laws to prescribe the model error statistics required by DA algorithms

Formulation

Let assume to have the model:

$$\frac{d\mathbf{x}(t)}{dt} = f(\mathbf{x}, \lambda)$$

used to describe the true process:

$$\frac{d\hat{\mathbf{x}}(t)}{dt} = \hat{f}(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \lambda^{'}) + \epsilon \hat{g}(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \lambda^{'})$$
$$\frac{d\hat{\mathbf{y}}(t)}{dt} = \hat{h}(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \lambda^{'})$$

- $\hat{g}(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \lambda^{'})$ represents the dynamics associated to extra processes not accounted for by the model;
- $\hat{h}(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \lambda^{'})$ - unresolved scale

Mean Error & Covariance

- the model does not describe the scale given by $\hat{h}(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \lambda^{'})$
- the correlation between i.c. and model error neglected (standard hyp. in DA)

Mean Estimation error evolution in the resolved scale

$$< \delta\mathbf{x}(t) >=< \mathbf{x}(t) - \hat{\mathbf{x}}(t) >=< \delta\mathbf{x}_0 > + \int_{t_0}^{t} d\tau < f(\mathbf{x}, \lambda) - \hat{f}(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \lambda) >$$

Evolution of the estimation error covariance in the resolved scale

$$\mathbf{P}(t) =< \delta\mathbf{x}_0 \delta\mathbf{x}_0^T > + \int_{t_0}^{t} d\tau \int_{t_0}^{t} d\tau^{'} < [f(\mathbf{x}, \lambda) - \hat{f}(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \lambda)][f(\mathbf{x}, \lambda) - \hat{f}(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \lambda)]^T >$$

- the important factor controlling the evolution is the difference between the velocity fields $f(\mathbf{x}, \lambda) - \hat{f}(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \lambda)$

These covariance and correlations are exactly what we need in DA !

These equations are NOT suitable for realistic geophysical applications - Some approximation is required

Short Time Approximation

- the contribution $f(\mathbf{x},\lambda) - \hat{f}(\hat{\mathbf{x}},\hat{\mathbf{y}},\lambda)$ is treated as a deterministic process
- the short time evolution of $\delta\mathbf{x}(t)$ and $\mathbf{P}(t)$ read:

$$\delta\mathbf{x}(t) \approx < [f(\mathbf{x},\lambda) - \hat{f}(\hat{\mathbf{x}})] > t + O(2)$$

$$\mathbf{P}(t) \approx < \delta\mathbf{x}_0 \delta\mathbf{x}_0^T > + < [f(\mathbf{x},\lambda) - \hat{f}(\hat{\mathbf{x}},\hat{\mathbf{y}},\lambda)][f(\mathbf{x},\lambda) - \hat{f}(\hat{\mathbf{x}},\hat{\mathbf{y}},\lambda)]^T > t^2 + O(3)$$

- the important factor controlling the evolution is the difference between the velocity fields $f(\mathbf{x},\lambda) - \hat{f}(\hat{\mathbf{x}},\hat{\mathbf{y}},\lambda)$
- if the term $f(\mathbf{x},\lambda) - \hat{f}(\hat{\mathbf{x}},\hat{\mathbf{y}},\lambda)$ were delta-correlated the short-time evolution is bound to be linear

Carrassi et al., 2008 & Carrassi and Vannitsem, 2011

DA in the presence of model error

Error of Unresolved Scales $\Rightarrow \mathbf{P}^m \approx < (f - \hat{f})(f - \hat{f})^T > \tau^2$

...needs to estimate the statistics of the vel. fields discrepancy.

Solution proposed:

- Use of the analysis increments of a reanalysis data-set :

$$f - \hat{f} = \frac{d\mathbf{x}}{dt} - \frac{d\hat{\mathbf{x}}}{dt} \approx \frac{\mathbf{x}_r^f(t+\tau_r) - \mathbf{x}_r^a(t)}{\tau_r} - \frac{\mathbf{x}_r^a(t+\tau_r) - \mathbf{x}_r^a(t)}{\tau_r} = \frac{\delta \mathbf{x}_r^a}{\tau_r} \Rightarrow$$

$$< \delta\mathbf{x}(\tau) > \approx \bar{\mathbf{b}}_m = < \delta\mathbf{x}_r^a > \frac{\tau}{\tau_r} \qquad \mathbf{P}^m(t) \approx \bar{\mathbf{P}}^m(t) = < \delta\mathbf{x}_r^a \delta\mathbf{x}_r^{aT} > \frac{\tau^2}{\tau_r^2}$$

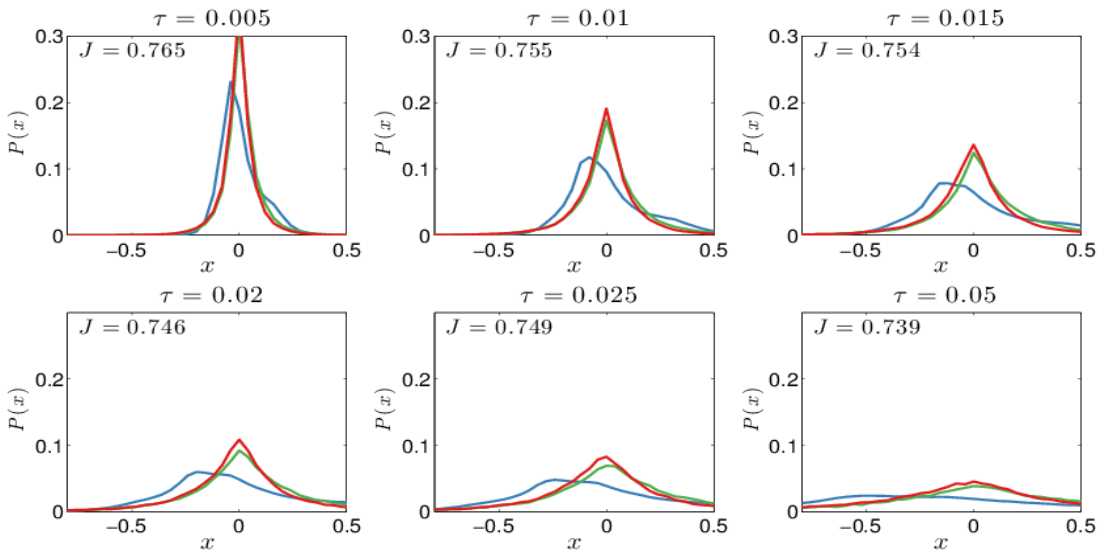
- τ_r reanalysis assimilation interval
- τ current assimilation interval

Testing the approximation

Lorenz (1996) with two scales (*large scale* - **x**; *small scale* - **y**)

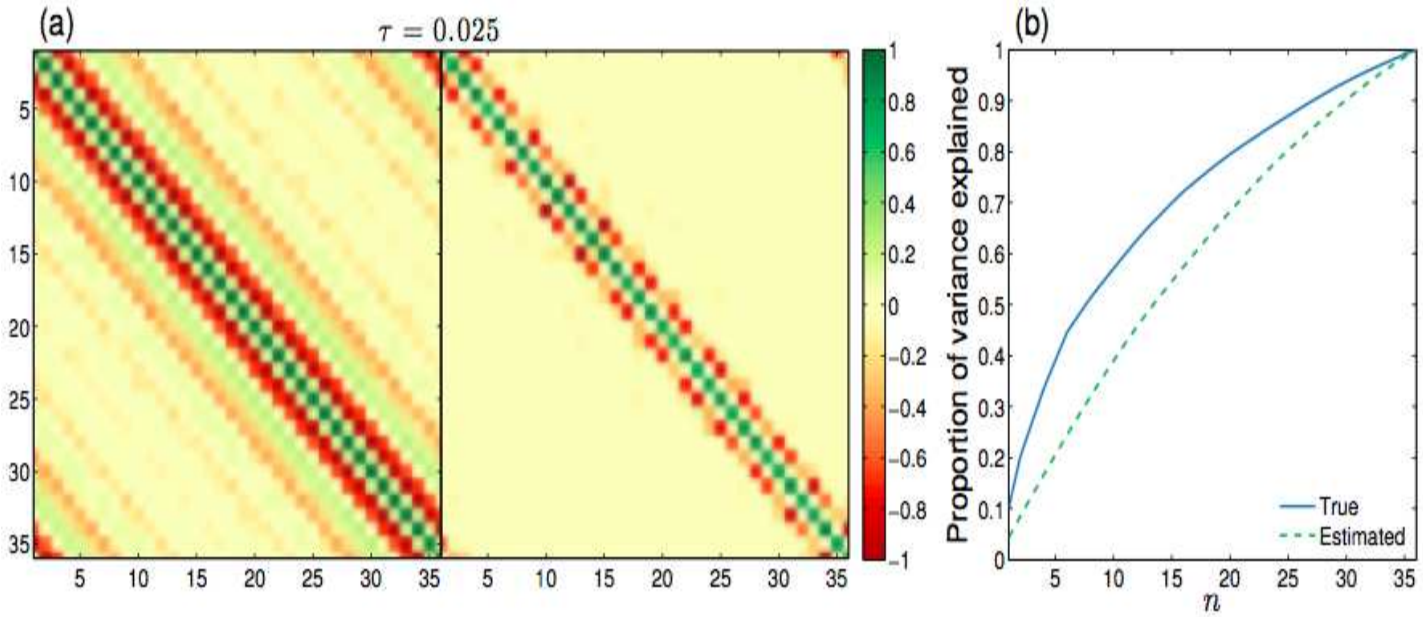
$$\frac{dx_i}{dt} = (x_{i+1} - x_{i-2})x_{i-1} - x_i + F - \frac{hc}{b}\sum_{j=1}^{10} y_{j,i}, \qquad i = \{1, ..., 36\}$$

$$\frac{dy_{j,i}}{dt} = -cby_{j+1,i}(y_{j+2,i} - y_{j-1,i}) - cy_{j,i} + \frac{hc}{b}x_i, \qquad j = \{1, ..., 10\}$$



PDF of $\delta\mathbf{x}_r^a$ for an observed (green) and unobserved (red) component, as well as the true model error PDF (blue), for different forecast lengths τ. J-values - joint probability between the true model error PDF and unobserved $\delta\mathbf{x}_r^a$.

Testing the approximation



(a): True (left) and estimated (right) $\mathbf{P}_m$.
(b): Proportion of the variance explained as a function of eigenvalue number for the true and estimated covariances.

Short-Time Ensemble Transform Kalman Filter – ST-ETKF

1. Standard ETKF with tuned localization & inflation:

$$\mathbf{P}^f \Longrightarrow (1+\delta)\mathbf{P}^f \circ \mathbf{\Omega}(r)$$

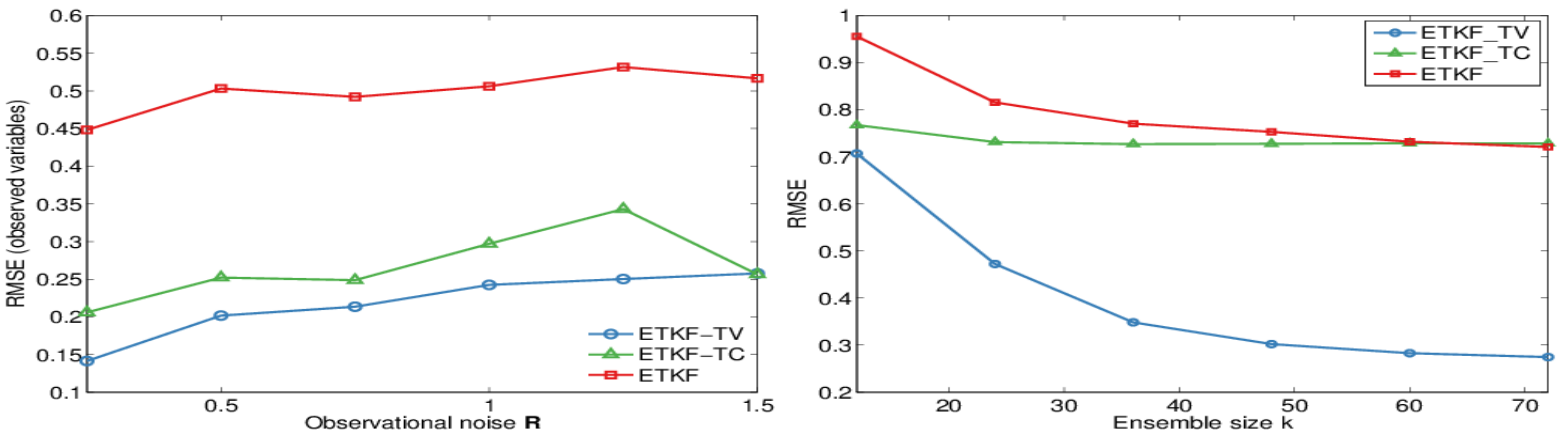
2. ETKF with Time Constant model error treatment (similar to ST-EKF) ETKF-TC:

$$\mathbf{x}^f \Longrightarrow \mathbf{x}^{\mathbf{f}} - \alpha\bar{\mathbf{b}}_m, \qquad \mathbf{P}^f \Longrightarrow (1+\delta)\mathbf{P}^f \circ \mathbf{\Omega}(r) + \alpha^2\overline{\mathbf{P}}_m$$

3. ETKF with Time Varying model error treatment ETKF-TV:

$$\mathbf{x}^f_{i,j} = \mathcal{M}(\mathbf{x}^a_{i,j}) - \alpha\eta_{i,j}\frac{\tau}{\tau^r} \qquad \eta_{i,j} \in \mathcal{N}(\overline{\mathbf{b}}_m, \overline{\mathbf{P}}_m) \qquad i = 1, ..., k$$

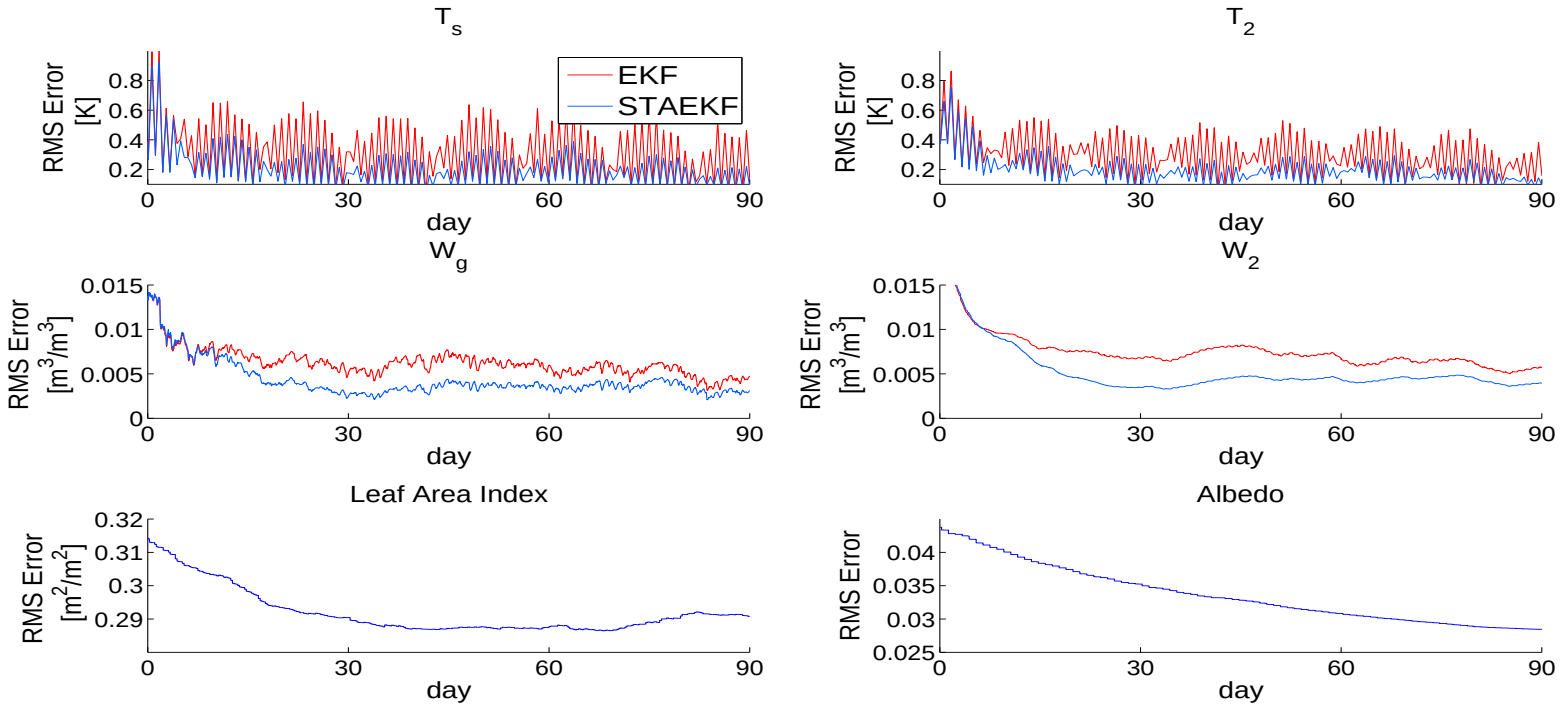
Skill as a function of Observational Error (left) & Ensemble Size (right)



From Mitchell & Carrassi 2015

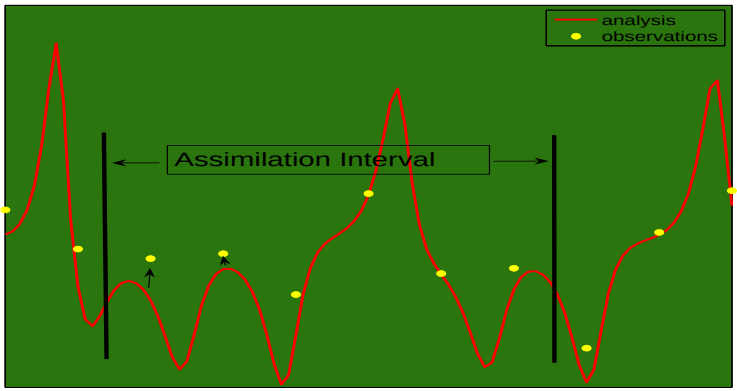
ST-AEKF - Parameter Estimation in Soil Model

- Land Surface model ISBA (Mahfouf and Noilhan, 1996)
- State Variables: soil temperature (T_s and T_2) and moisture content (w_g and w_2).
- Observations of screen-level variables (temperature and humidity at 2 meter)
- Parametric error in the *Leaf Area Index* (LAI) and *Albedo*
- Comparison between EKF and ST-AEKF



Carrassi et al., 2012

4DVar in the presence of model error - Short Time Weak Constraint 4DVar



- assimilate observations distributed over the time window τ
- analysis state as the minimum of a cost-function:

$$2J = \int_0^\tau \int_0^\tau (\delta \mathbf{x}_{t_1}^m)^T (\mathbf{P}^m)_{t_1 t_2}^{-1} (\delta \mathbf{x}_{t_2}^m) dt_1 dt_2 + \sum_{k=1}^{M} \epsilon_{\mathbf{k}}{}^T \mathbf{R}_k^{-1} \epsilon_{\mathbf{k}} + \epsilon_b^T \mathbf{B}^{-1} \epsilon_b$$

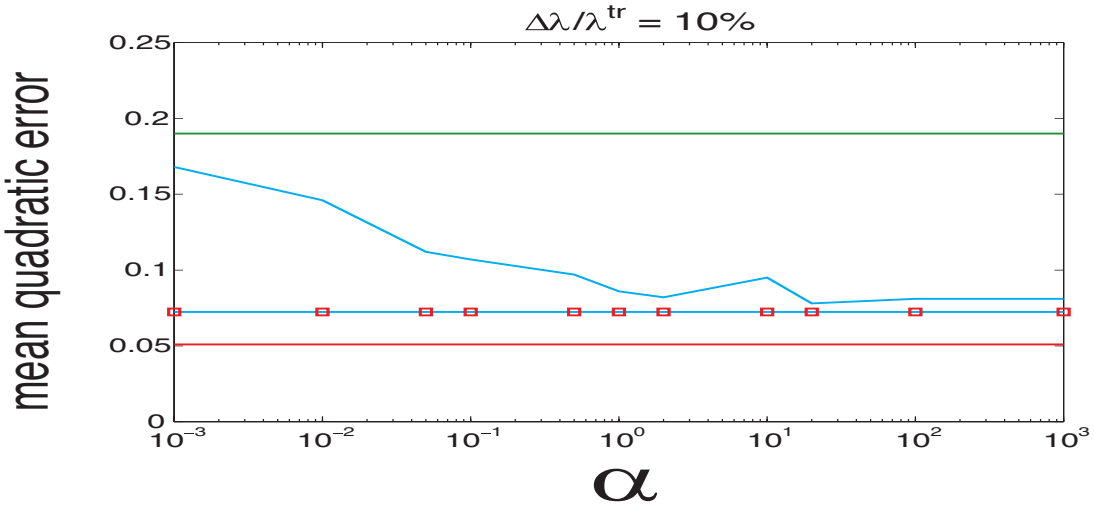
Estimate model error covariances/correlations using

$$\mathbf{P}(t_1, t_2) \approx \mathbf{Q}(t_1 - t_0)(t_2 - t_0)$$

DA and Model Error - ST-Weak Constraint 4DVar

Correlated vs Uncorellated Model Error

- Lorenz 3-variable (1963) system
- **Strong-constraint** $\Rightarrow$ Model Assumed Perfect !
- **Weak constraint 4DVar with uncorrelated model error** $P_t^m = \alpha \mathbf{B}$ (blue) $\Rightarrow$ Model Err Uncorrelated; Mod Err Covariance scaled as Background Cov
- **Weak constraint 4DVar with uncorrelated model error** $P_t^m = \mathbf{Q}(t - t_0)^2$ (red marks) $\Rightarrow$ Model Err Uncorrelated; Mod Err Covariance quadratic in time
- **Short-time weak constraint 4DVar** $\Rightarrow$ Model err as a fully-corr process



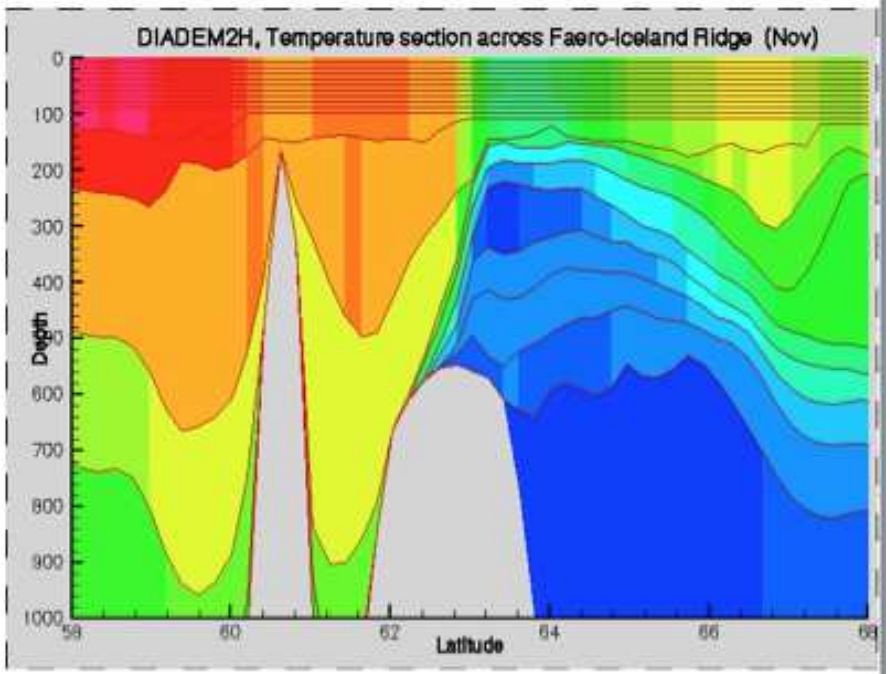
Carrassi & Vannitsem, 2010

Operational Data Assimilation

3 Operational Data Assimilation - The TOPAZ system at NERSC

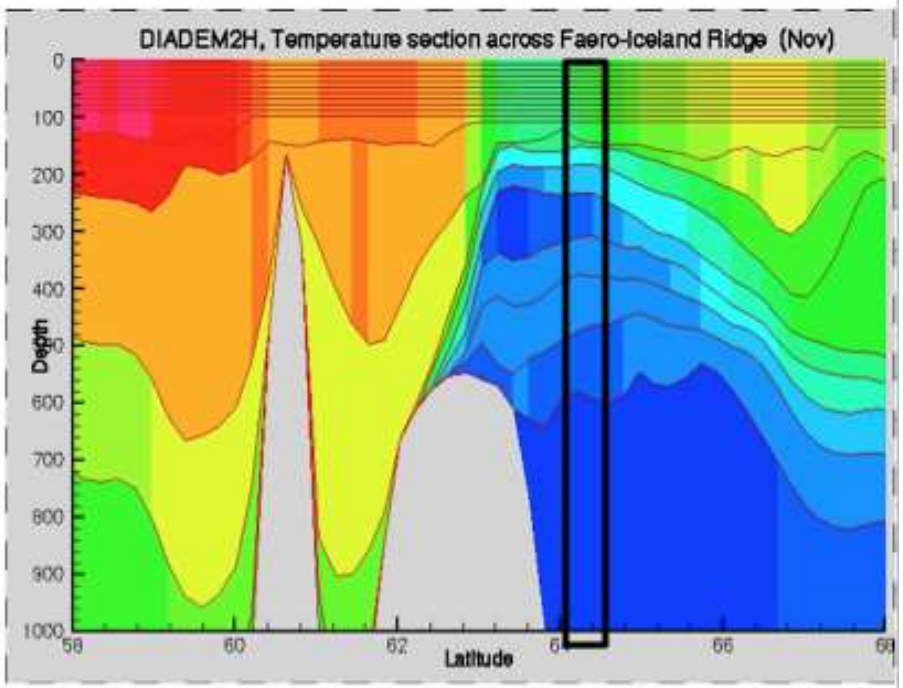
Operational Ocean Prediction with EnKF at NERSC

- 3D numerical ocean model
  - Hybrid Coordinate Ocean model, HYCOM (U. Miami)
- Hybrid vertical coordinate
  - Isopycnal in the interior
  - Z-coordinate at the surface
  - TOPAZ4 uses 28 layers
- Coupling to sea ice model
  - EVP dynamics ...
  - Semtner Thermodynamics
- Data assimilation:
  - EnKF (probabilistic) ...



Operational Ocean Prediction with EnKF at NERSC

- 3D variables
  - Temperature
  - Salinity
  - Layer thickness (can be zero)
  - X-current
  - Y-current
- 2D variables
  - Sea ice area
  - Sea ice thickness
  - Snow depths
  - Barotropic currents + pressure
- Typical grid size
  - Horizontal: 800x880
  - Vertical: 28
  - **Total unknowns: ~10^8**
    - **Need to perform *local* analyses**



Evensen 2002

Operational Ocean Prediction with EnKF at NERSC

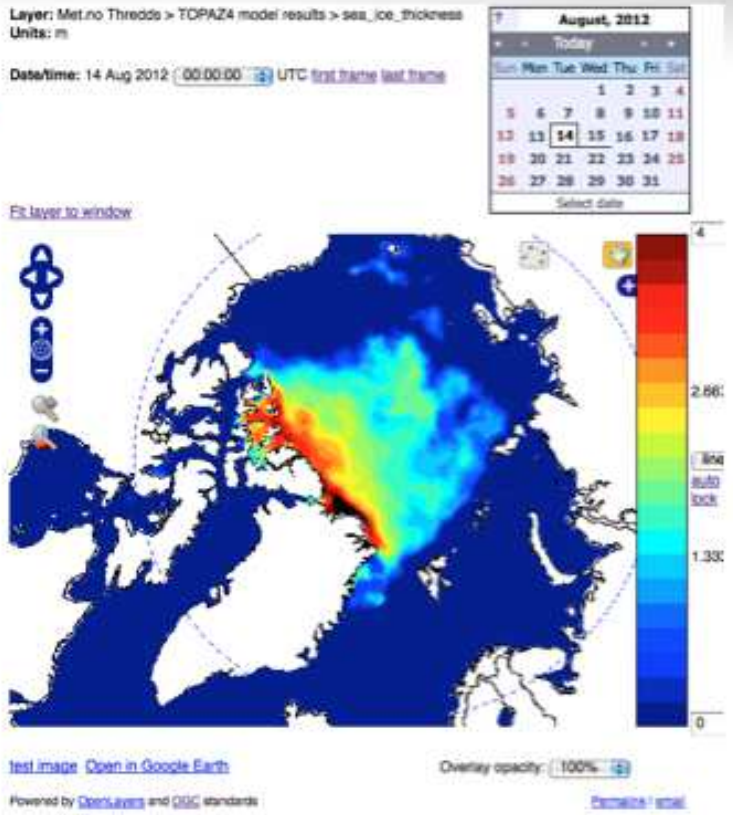
- **Ensemble Forecast**
- 2500 CPU hours / cycle
- Embarrassingly parallel
- 100x **133 CPU 11 min** jobs
- Each job requires **400 Mb**
  - MPI parallelization

- **Analysis**
- 20 CPU hours / update
- 6 datasets simultaneously
- One **20 CPU 1h job**
- Memory required **1 Gb**
  - MPI parallelization

- HPC Machine:
- Cray XE6m, updated 2012
  - 22272 cores, 205 Tflop/s
  - 676 nodes (32-cores)
  - 1-4 Gb per node

Operational Ocean Prediction with EnKF at NERSC - TOPAZ System

- Exploited operationally at met.no
  - Since 2008
  - Ecosystem added in Jan. 2012
- 20 years reanalysis at NERSC
  - Took 2 years to produce
  - 3-years ecosystem reanalysis
- MyOcean (Arctic MFC)
  - Free distribution of data
  - Dynamical viewing (Godiva2)
- Data used by ECMWF wave model (J. Bidlot)
  - Sea ice edge forecast
  - Surface currents



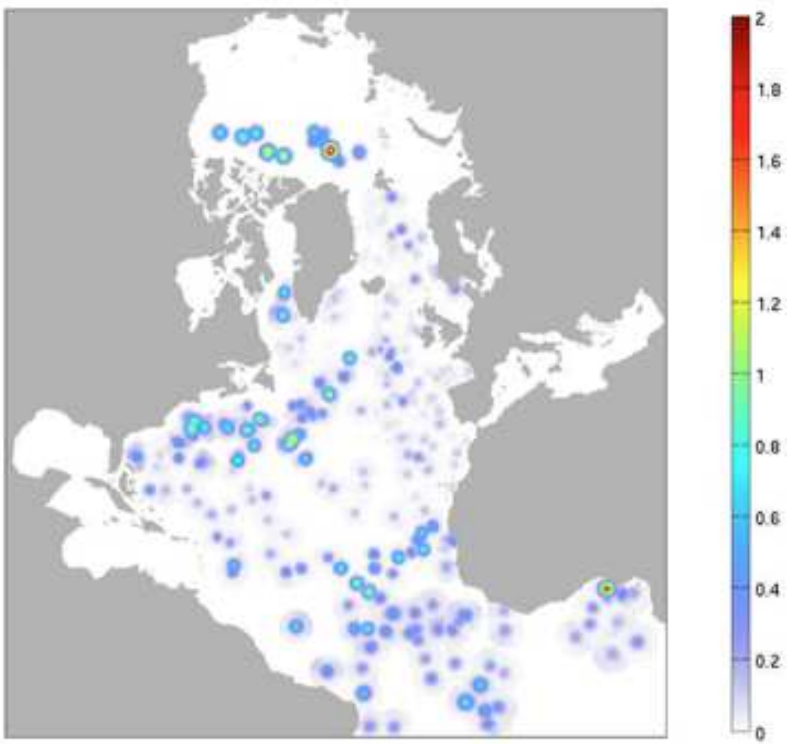
Ice thickness forecast for 14th Aug. 2012

Operational Ocean Prediction with EnKF at NERSC - TOPAZ System

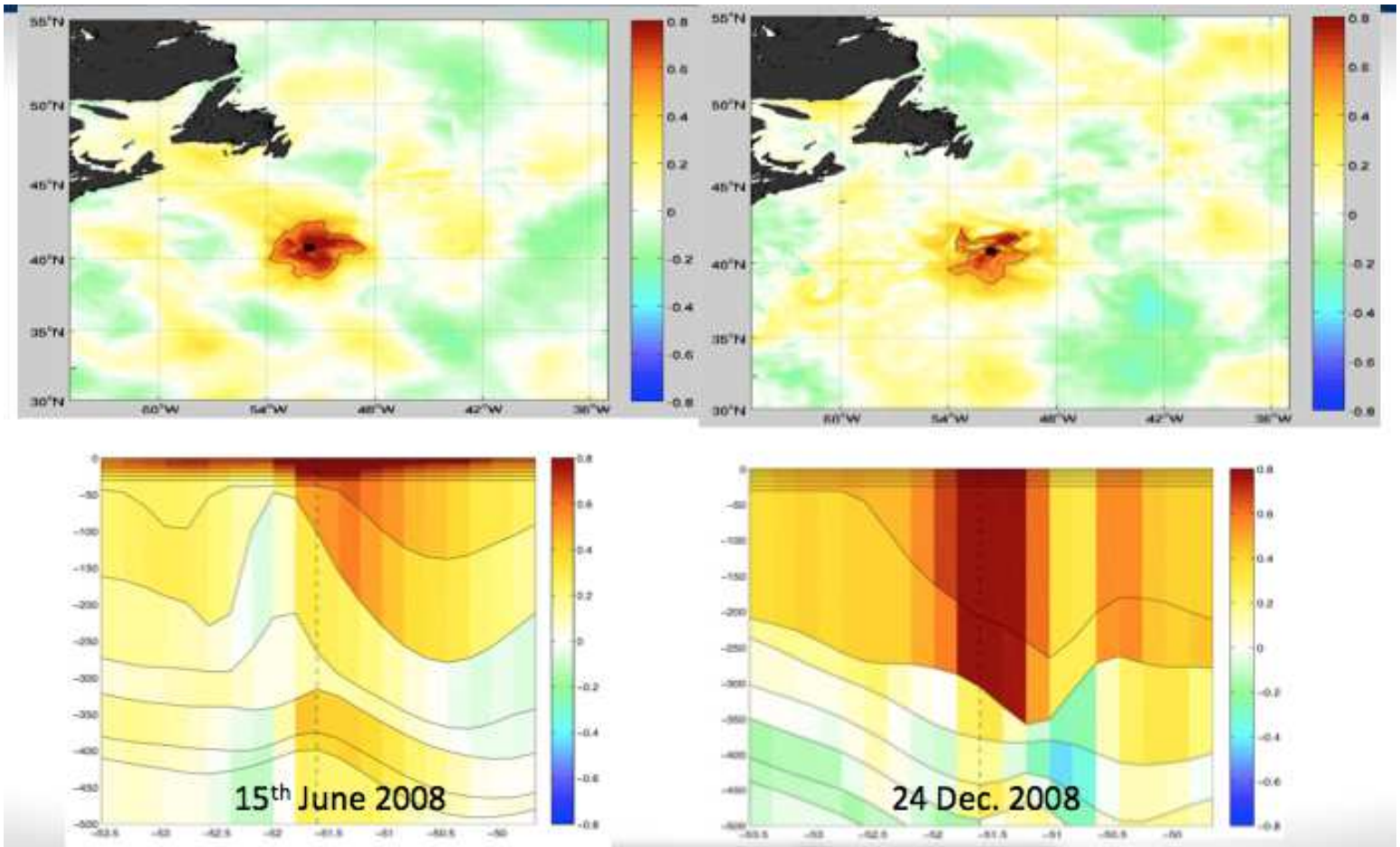
- DEnKF, asynchronous
  - 100 members
  - Local analysis (~90 km radius)
  - Ensemble inflation by 1%
- Observations:
  - Sea Level Anomalies (CLS)
  - SST (NOAA, then UK Met)
  - Sea Ice Concentr. (OSI-SAF)
  - Sea ice drift (CERSAT)
  - T/S profiles (Coriolis)
  - **400.000 observations** per week
  - ~100 in each local radius

$$\mathrm{SRF} = \sqrt{\frac{\mathrm{tr}(\mathbf{H}\mathbf{P}^f\mathbf{H}^T\mathbf{R}^{-1})}{\mathrm{tr}(\mathbf{H}\mathbf{P}^a\mathbf{H}^T\mathbf{R}^{-1})}} - 1$$

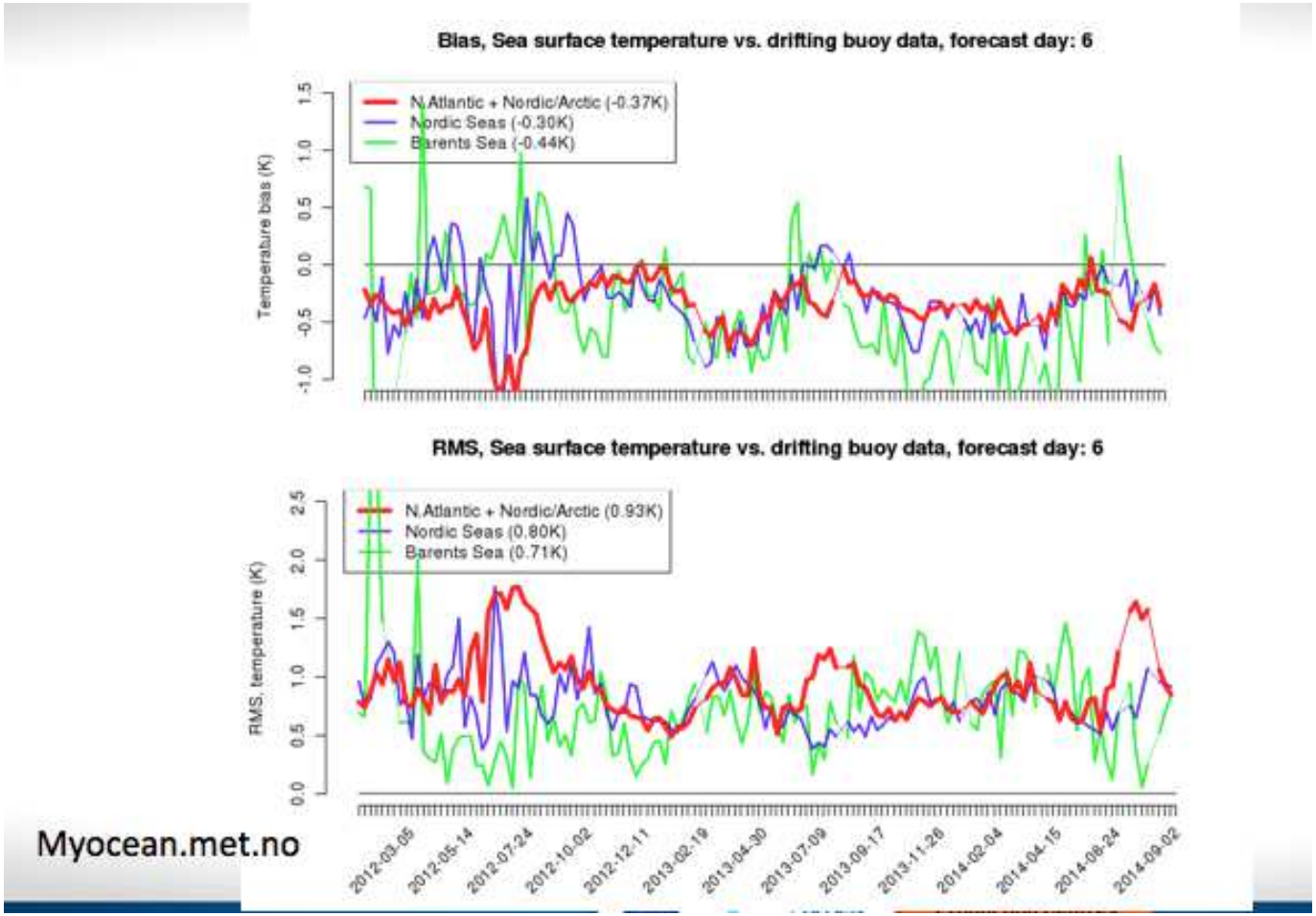
SRF: local spread reduction factor



TOPAZ System - EnKF correlations SST

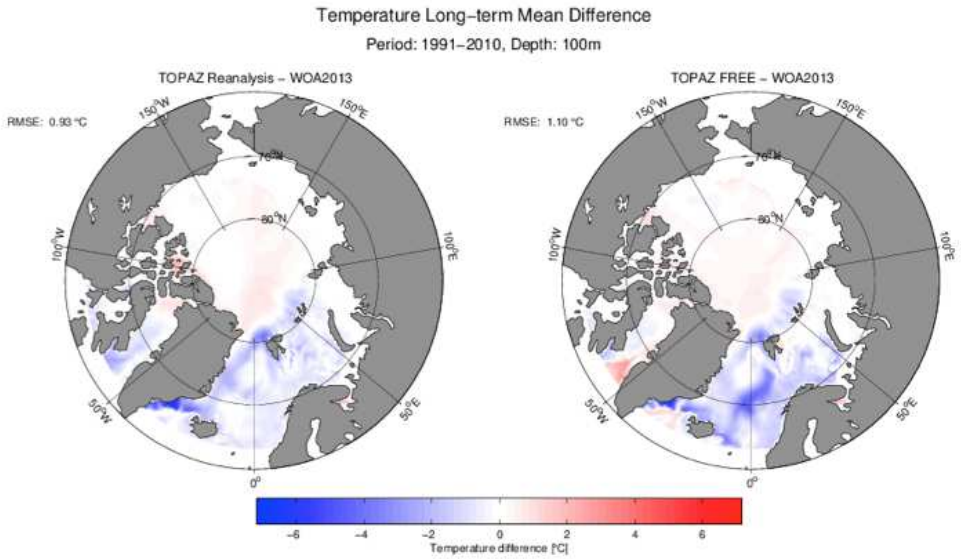


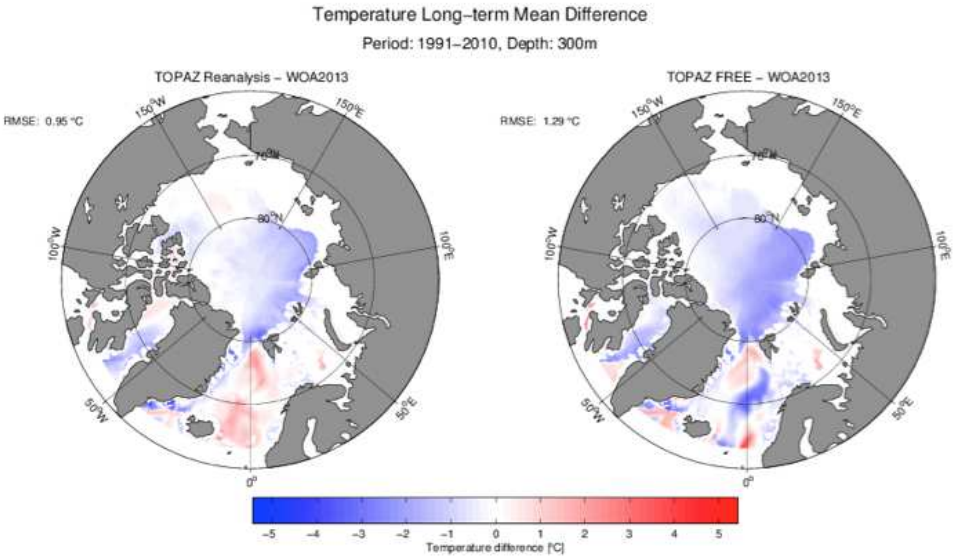
TOPAZ System - SST forecast in real time



The system has a cold bias !

TOPAZ System - Reanalysis





(some of) New Frontiers and Nowadays Challenges in DA

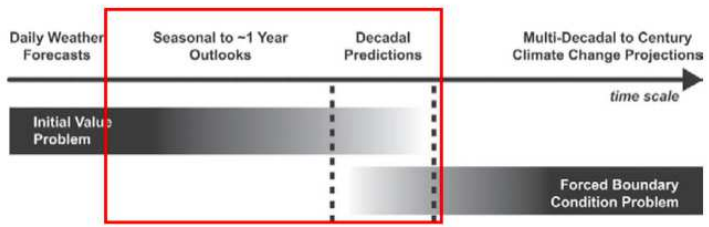
4 (some of) New Frontiers and Nowadays Challenges in DA

Seasonal-to-Decadal Predictions - S2D

Seasonal-to-Decadal is an optimal time window to prepare for:

1. disease treatment and mitigation
2. population move
3. agricultural planning
4. energy policy

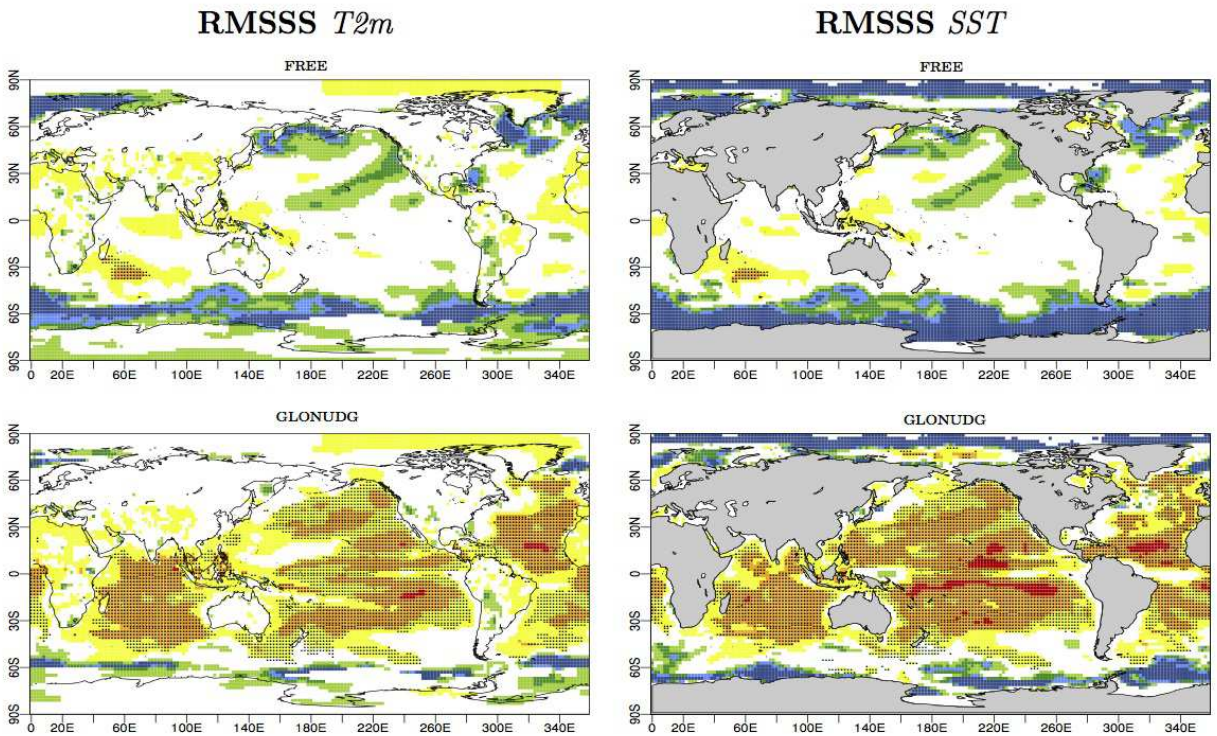
- s2d prediction is both an initial and boundary condition problem
- Longer Term Predictability rests on slow-varying components of the climate system (Soil moisture, Snow cover and Sea-ice, SST ...)
- S2D are currently initialized using ad-hoc empirical methods: *Full-Field* or *Anomaly Initialization*
- DA is nowadays seen with much interest in this field to improve the initial condition representation



Meehl et al. (2009)

Nudging Experiments with Ec-Earth climate model

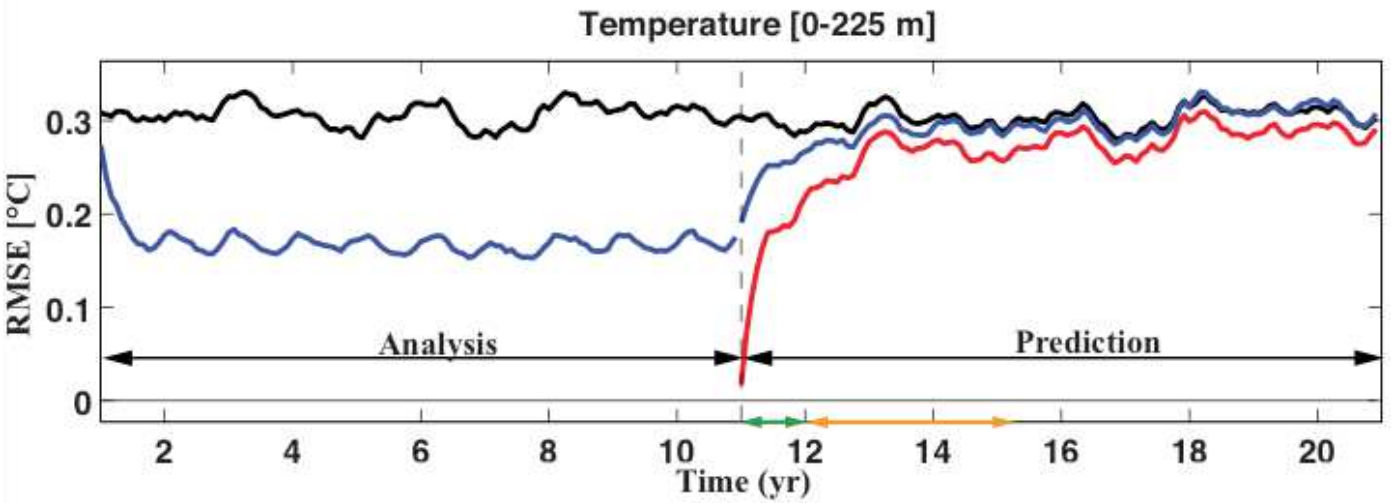
- NUDGING is a practical/empirical DA method based on adding a forcing term to the prognostic equations: $\frac{d\mathbf{x}}{dt} = \mathbf{F}(\mathbf{x}) + \mathbf{H}^T\mathbf{N}(\mathbf{y}^{obs} - \mathbf{H}\mathbf{x})$
- Root Mean Square Skill Score (RMSSS) of the ensemble-mean near surface temperature (left) and SST (right) anomalies
- Observations used for Nudging: Ocean Temperature & Salinity (from NEMOVAR REANALYSIS)
- $RMSSS = (1 - \frac{RSME^{nudg}}{RMSE^{clim}})$
- New initialization methods based on DA have been proposed in Carrassi *et al.*, 2014 NPG.
- A study on the rationale behind the choice of anomaly vs full-field initialization can be found in Weber *et al.*, 2015 MWR



From Carrassi et al., 2016

S2D with NorESM and EnKF

RMSE calculated over the full model domain (averaged over the 10 prediction cycles)



For all model variables at 1-year lead average; 2-5 lead year average

- Analyze reduction of RMSE in **EnKF-SST** relative to **Free**
- Compare the improvements relative to **Perfect**

From Counillon et al. (2014) Tellus

Going beyond Gaussianity ⇒ Particle Filters (PF)

- PF is a technique for implementing a recursive Bayesian filter by Monte Carlo simulations (see Bain and Crisan 2008).
- The state-estimate representation is afforded using an ensemble of members (i.e. particles)
- The analysis update is fully Bayesian and this makes PF particularly accurate in nonlinear filtering
- The key idea is to represent the required PDFs by a set of random samples with associated weights and to compute estimates based on these samples and weights:

$$\mathcal{P}(\mathbf{x}_k) \approx \sum_{i=1}^{N} \omega_{k-1}^i \delta(\mathbf{x}_k - \mathbf{x}_k^i) \qquad \omega_k^i \propto \omega_{k-1}^i \mathcal{P}(\mathbf{y}_k^o | \mathbf{x}_k)$$

- The sampling is done using the principle of importance sampling and the choice of the importance PDF plays a central role.
- A common problem with PF is the degeneracy phenomenon, where after a few iterations, all but a few particles will have negligible weight

Going beyond Gaussianity ⇒ Particle Filters (PF)

- The brute force approach to reducing its effect is to use a very large ensemble size (clearly impractical in geosciences).
- Nowadays the future of PF in this domain relies on two other approaches: (1) the good choice of importance density and (2) the use of resampling.
- The proposal distribution represents our approximation of the unknown filtering distribution.
- A common choice is to use the prior PDF, the transition probability, since this allows for a straightforward way to draw particles (they comes from the model forward integration) and associated weights.
- Significant advancements have been done recently toward a proposal density that is c̈loser ẗo observations.
- The basic idea of resampling is to eliminate particles that have small weights and to concentrate on particles with large weights.
- Successful application of the particle filter for the problem of past climate reconstruction has been done using climate model of intermediate complexity (Goose *et al.*, 2012).

4.4 Coupled Data Assimilation

Scientific Challenge: DA in coupled dynamics - CDA

- use Earth System Simulators (ESS) as the unified modelling instrument across all forecast timescales from days to decades
- better exploit the new generation of Earth observations (Argo, SMOS ...)
- improve the forecast capabilities of coupled phenomena (hurricanes, costal weather, ENSO, MJO)
- produce coupled reanalysis
- reconstruct the climate of areas for which adequate measurements are still unavailable
- assessment of climate change in connection with external factors (detection and attribution problem)

4.4 Coupled Data Assimilation

State-of-the-art: DA in coupled dynamics - CDA

- Decoupled DA in coupled system —— *i.e.* prediction are done using coupled atmosphere-ocean models where the ocean is forced with a wind stress output of independent atmospheric data; the two components are then coupled and used to make the prediction of interest.
- this raises problems, particularly at the boundary between the ocean and the atmosphere, where unwanted dynamical initial shocks can be introduced.
- Weakly-coupled DA —— *i.e.* the background field is obtained through the evolution of the full coupled model, but the different model compartments are then subject to an independent analysis

First Attempts:

1. weakly coupled reanalysis at the NCEP (Saha et al., 2010) - marked improvement over the standard uncoupled DA in recovering the MJO
2. 4DVar a coupled global ocean-atmosphere model at JAMS Technology. Weather modes are considered as noise, and the control variable includes the ocean i.c. plus a set of parameters of the sea-air fluxes (Sugiura et al., 2008).
3. At the ECMWF ocean and atmosphere are currently run separately (a 3DVar and 4DVar respectively) but research is ongoing to weakly couple the two schemes.
4. The ensemble-based approach has been implemented at the GFDL (Zhang et al., 2005) using the EAKF.