Data Assimilation in Geophysics - From Weather to Climate Prediction

Alberto Carrassi
alberto.carrassi@ic3.cat

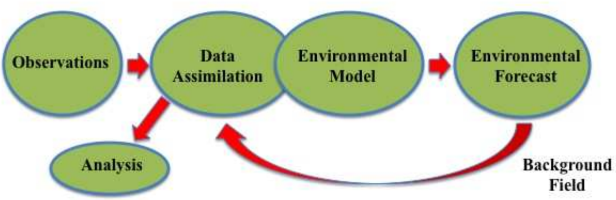
Climate Forecasting Unit - **CFU**
Catalan Institute for Climate Science - **IC3**
Spain

Nonlinear Filtering and Data Assimilation
International Centre for Theoretical Sciences, TIFR - Bangalore
8^{th} January 2014

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 prediction
- hydrology
- reanalysis

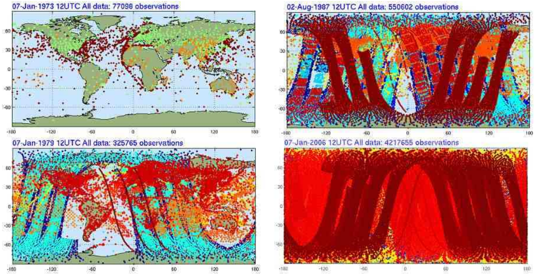


Data Assimilation - Overview

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**



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: $\mathbf{X}_k = \{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_k\}$ and $\mathbf{Y}^0_k = \{\mathbf{y}^0_0, \mathbf{y}^0_1, \dots, \mathbf{y}^0_k\}$

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

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

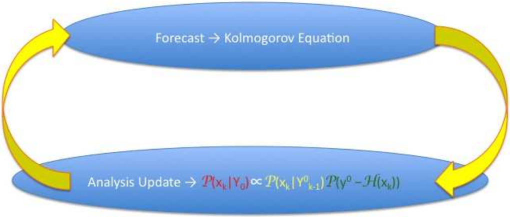
Probabilistic Approach

With Bayes's rules....

SMOOTHING

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

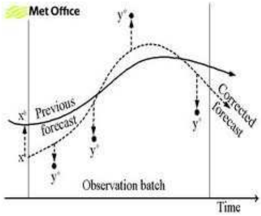
FILTERING



In high-dimensional nonlinear systems the full Bayesian formulation is not affordable

Note: The Particle Filters attempt to solve this problem and their potential application in geoscience has received much attention in recent years. See more talks in this workshop & van Leuween, 2009 (MWR) for a review of PF in Geosciences.

4D-Variational Assimilation



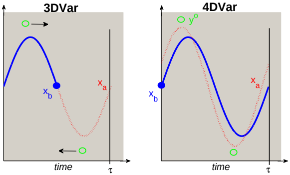
Initial condition, observational and model errors are all Gaussian and mutually uncorrelated $\Longrightarrow$ solving the SMOOTHING problem leads to the 4DVar formulation, *i.e.* minimise a penalty function as:

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

- **B** - Background error covariance matrix
- **R** - Observational error covariance matrix
- **Q** - Model error covariance matrix

4D-Variational Assimilation

- 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.* Tremolet, 2006 QJRMS; Bocquet, 2009 MWR)
- **B** 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 and ECMWF.



Sequential Assimilation

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

ANALYSIS UPDATE EQUATIONS

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

$$\mathbf{P}_k^a = [\mathbf{I} - \mathbf{K}_k \mathbf{H}_k]\mathbf{P}_k^f$$

$$\mathbf{K}_k = \mathbf{P}_k^f \mathbf{H}_k^T (\mathbf{H}_k \mathbf{P}_k^f \mathbf{H}_k^T + \mathbf{R}_k)^{-1}$$

- $\mathbf{x}_k^a$ - Analysis state at time t_k
- $\mathbf{x}_k^f = \mathcal{M}(\mathbf{x}_{k-1}^f)$ - Forecast state at time t_k
- $\mathbf{P}^f$ - Forecast error covariance matrix
- $\mathbf{R}$ - Observational error covariance matrix
- K - Kalman gain matrix
- 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

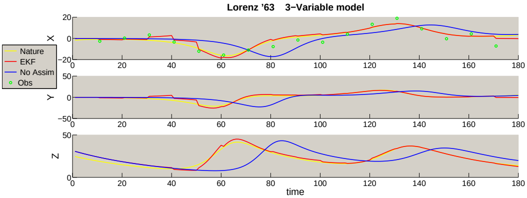
For linear dynamics and observational operator the KF provides a closed set of estimation equations (Kalman, 1960). The forecast step equations read:

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

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

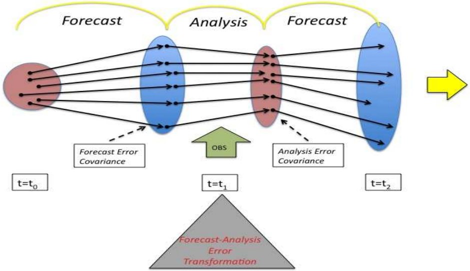
Extension to nonlinear dynamics - Extended Kalman Filter

- 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 (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



- Ensemble based covariances $\mathbf{P}^{f,a} = \frac{1}{M-1}\sum_{i=1}^{M}(\mathbf{x}_i^{f,a} - \bar{\mathbf{x}}^{f,a})(\mathbf{x}_i^{f,a} - \bar{\mathbf{x}}^{f,a})^T$
- For the approach to be suitable in geoscience $M \ll n$
- Flow dependent description of the forecast error
- The Kalman gain is computed in the observation space reducing the computational cost at a rate given by the ratio between the number of observations and the system size, p/n, $\mathbf{P}^f\mathbf{H}^T = \frac{1}{M-1}\sum_{i=1}^{M}(\delta\mathbf{x}_i^f \mathbf{H}\delta\mathbf{x}_i^f)\ldots$
- Provide automatically a set of initial conditions for ensemble prediction schemes.
- The choice of the forecast-analysis transformation characterises the ensemble-based algorithms.

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)

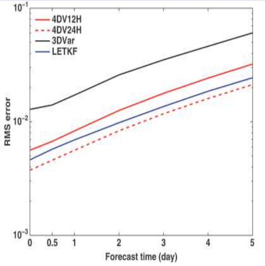
Deterministic (Square-Root approach)

- 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)

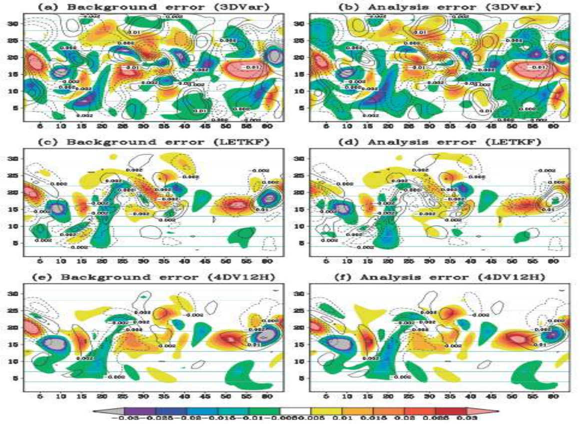
Ensemble-based or Variational: the comparison

- 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, 2009 MWR

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 **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..
- Huge dimension $\Rightarrow$ Computationally suitable solutions (see Fisher talk at WMO-DA Symposium 2013)

DA research issues

1. *Control of Chaos* $\Rightarrow$ DA methods for chaotic dynamics
2. *Model Error* $\Rightarrow$ DA methods accounting for model error
3. *Long Term Predictions* $\Rightarrow$ Initialization of climate models & Coupled DA

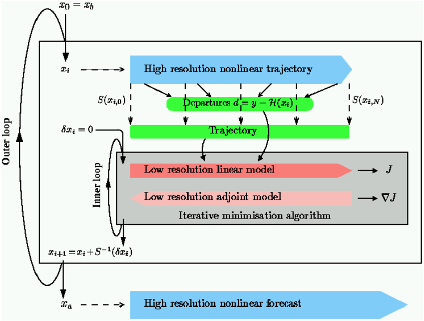
How to deal with geophysical systems: Variational

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

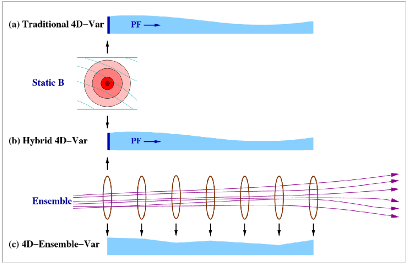
How to deal with geophysical systems: Variational

- 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 QJRMS)
- 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

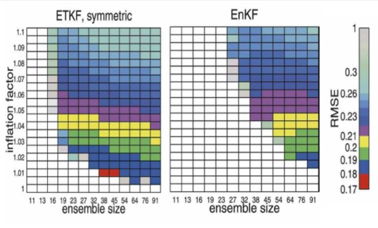
Proposed Solutions:

- 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)

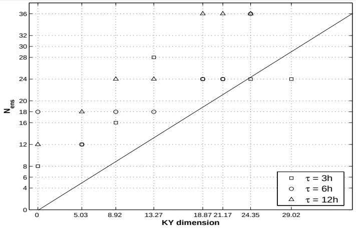
  Additive Inflation:
  - add random noise to $\mathbf{P}^f$ or $\mathbf{P}^a$
  - the process introduce new structures in the error space spanned by the ensemble covariances
  - a combined additive/multiplicative scheme has been proposed by Zhang *et al.*, 2004 MWR
  - an ensemble based algorithm without 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; Zhang and Zhang, 2012 MWR)
- Solution to problem (3) $\Rightarrow$ Particle Filters but ....

exploiting chaos... → optimal ensemble size

Can the ensemble size be designed based on the system dynamical properties ?



From Sakov and Oke, 2008 MWR



From Carrassi et al., 2009 TELLUS

- LEFT: The ETKF converges to low error level when $N_{ens} \geq KY_{dim}$ reaches the model subspace dimension
- RIGHT: Similar behaviour with another deterministic filter (MLEF, Zupanski, 2005 MWR), without inflation, we also observed error saturation when $N_{ens} \geq KY_{dim}$
- This behavior is deeply different from the EnKF whose performance is expected to improve indefinitely when $Ens_{size} \rightarrow \infty$
- Bocquet, 2011 (NPG) introduced a new deterministic filter (ETKF-N) that does not need inflation as long as $Ens_{size} > N^{+}$
- *INTERPRETATION:* In deterministic filters ensemble perturbations reflect the intrinsic system error dynamics and have to be intended as factorization of the system's error covariance rather than its Monte Carlo approximation as in the EnKF

Assimilation in the Unstable Subspace - AUS

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 **E**; the analysis solution reads:

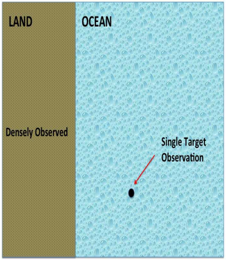
$$\mathbf{x}^{\mathbf{a}} = \mathbf{x}^{\mathbf{b}} + \mathbf{E}\Gamma\mathbf{E}^{T}\mathbf{H}^{T}(\mathbf{R} + \mathbf{H}\mathbf{E}\Gamma\mathbf{E}^{T}\mathbf{H}^{T})^{-1}(\mathbf{y}^{o} - \mathbf{H}\mathbf{x}^{\mathbf{b}})$$

- 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)

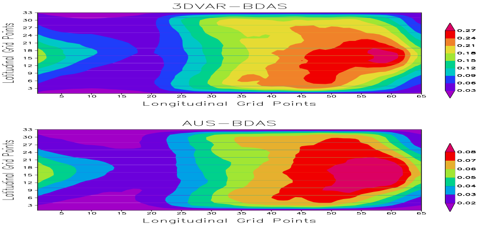
AUS and Target Observations

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



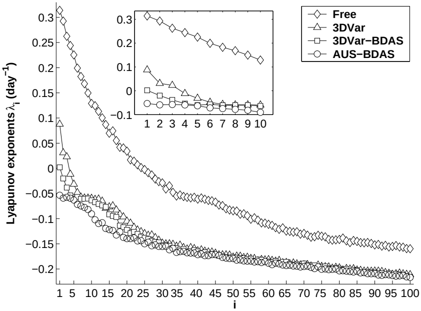
Experiment	Ocean Obs Type/Positioning/Assimilation	RMS Error
LO	-	0.462
FO	vert.Prof/fixed(in the max(err))/3DVar	0.338
RO	vert.Prof/random/3DVar	0.311
3DVar-BDAS	vert.Prof/BDAS/3DVar	0.184
AUS-BDAS	temp.1-Level/BDAS/AUS	0.060



adapted from *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 ?



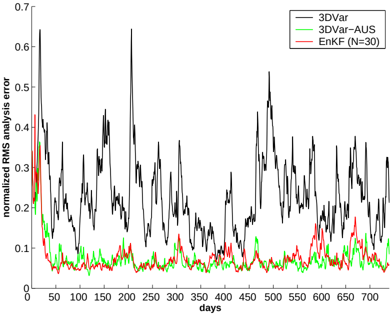
adapted from *Carrassi, Ghil, Trevisan & Uboldi, 2008 CHAOS*

- 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

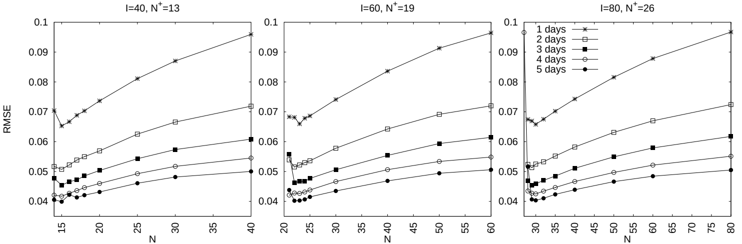


adapted from *Carrassi, Trevisan, Descamps, Talagrand & Uboldi, 2008 NPG*

- 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 ⇒ Reduced computational cost and implementation on a pre-existing 3DVar scheme

4DVar-AUS

4DVar-AUS: The analysis increment is confined in the unstable and neutral subspace by applying to 4DVar the AUS constraint



From Trevisan, D'Isidoro & Talagrand, 2010 QJRMS

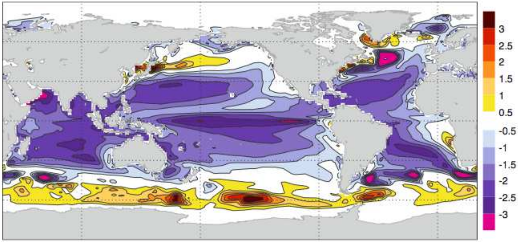
- Lorenz 40 variables
- The assimilation is performed in a subspace of dimension $N = N^0 + N^+$
- When $N = n$, the standard 4DVar is recovered
- The error of 4DVar-AUS is smaller than the error of 4DVar, particularly for short assimilation windows, when the errors in the stable directions are not yet damped
- It exists an optimal subspace dimension for the assimilation that is approximately equal to $N^+ + N^0$
- 4DVar-AUS does not need tangent/adjoint model
- See Trevisan & Palatella 2011 NPG for the EKF-AUS and Palatella, Carrassi & Trevisan, 2013 JPA for a review of the AUS algorithms

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

Model Error & Covariance

Estimation error evolution in the resolved scale

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

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 correlation between i.c. and model error neglected (standard hyp. in DA)
- the important factor controlling the error 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 fully correlated in time
- the short time evolution of $\mathbf{P}(t)$ reads:

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

- note that now the approximation involve neglecting the presence of the initial condition error

DA in the presence of model error

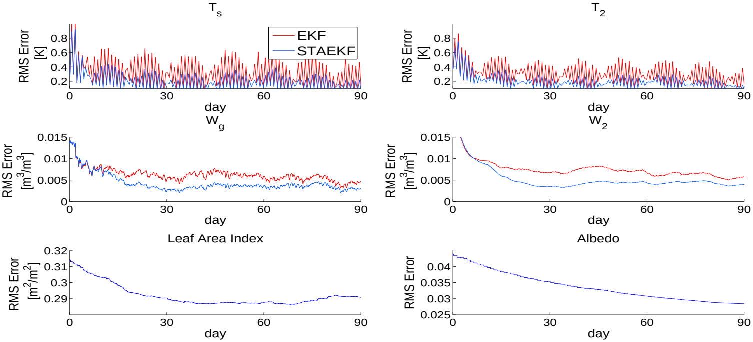
Can we incorporate the short-time approximation for the model error covariance in the context of DA procedures ?

Specific goals:

1. **Computation of the model error covariance in sequential data assimilationi - EKF & ETKF**
2. **Computation of the model error correlations in variational data assimilation - 4DVar**

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



adapted from *Carrassi et al., 2012 ASL*

Model Error from Unresolved Scales

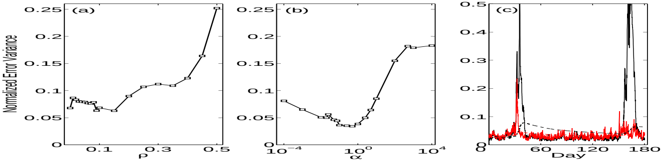
- Lorenz (1996) with two scales (*large scale* - **x**; *small scale* - **y**) - 12 observations of the large scale
- Use of the analysis increments of a reanalysis data-set :

$$f - \hat{f} = \frac{d\mathbf{x}}{dt} - \frac{d\hat{\mathbf{x}}}{dt} \approx \frac{\mathbf{x}_r^f(t+\tau_r) - \mathbf{x}_r^a(t)}{\tau_r} - \frac{\mathbf{x}_r^a(t+\tau_r) - \mathbf{x}_r^a(t)}{\tau_r} = \frac{\delta \mathbf{x}_r^a}{\tau_r} \Rightarrow \mathbf{P}^m(t) \approx < \delta \mathbf{x}_r^a \delta \mathbf{x}_r^{a\,T} > \frac{\tau}{\tau_r}$$

- τ_r reanalysis assimilation interval; τ current assimilation interval

Comparison with the EKF employing the inflation of the $\mathbf{P^f}$ as a tool to account for model error

- (a) - EKF; Inflation procedure on the $\mathbf{P}^f \rightarrow (1+\rho)\mathbf{P}^f$
- (b) - ST-EKF; Tuning of $\mathbf{P}^m \rightarrow \alpha \mathbf{P}^m$ ($\mathbf{P}^m$ estimated statistically and then kept fixed)
- (c) - Analysis Error Comparison ST-EKF ($\alpha = 0.5$ red line) and EKF ($\rho = 0.09$ black line)



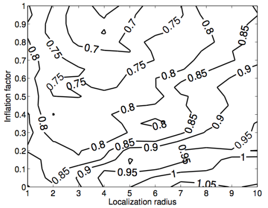
Adapted from *Carrassi & Vannitsem, 2011 IJBC*

Model Error from Unresolved Scales

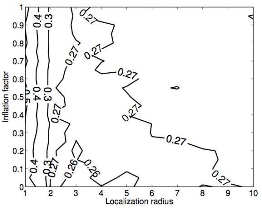
Ensemble Transform Kalman Filter - ETKF

- The performance of the ETKF is studied as a function of the inflation and localization factors
- The standard ETKF (left) is compared with ETKF employing model error treatment in the evolution equation of the ensemble members.
- Noise is added to each ensemble member.
- The noise is sampled by the analysis-increment statistics

Standard ETKF

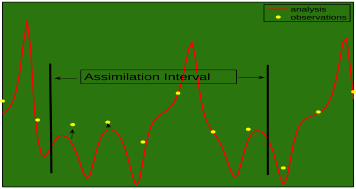


ETKF with model error treat



From Mitchell & Carrassi, 2014

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:

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

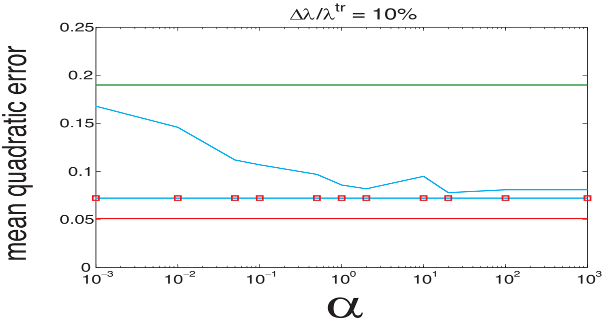
Estimate model error covariances/correlations using

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

Results Short Time 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



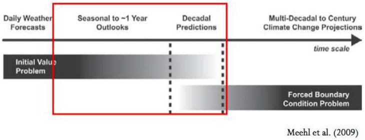
Carrassi & Vannitsem, 2010 MWR

Seasonal-to-Decadal Predictions - s2d

s2d is an optimal time window to prepare for:

- disease treatment and mitigation
- population move
- agricultural planning
- energy policy

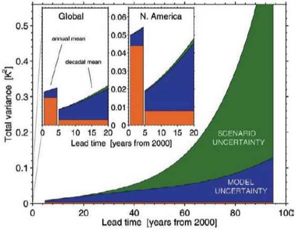
s2d prediction is both an initial and boundary condition problem



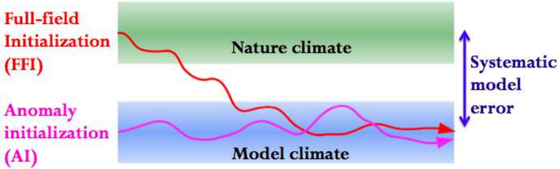
Longer Term Predictability rests on slow-varying components of the climate system:

- Soil moisture
- Snow cover and Sea-ice
- Sea surface temperature
- Response to external forcing / Boundary conditions (ex. GHG)

s2d prediction error sources: Internal Variability - Model error - Scenario Uncertainty. From Hawking and Sutton, 2009



Full-Field and Anomaly Initialization ⇒ A DA formulation



Full-Field Initialization - **FFI** $\mathbf{x}^a = \mathbf{x}^b + \mathbf{H}^T[\mathbf{y}^o - \mathbf{H}\mathbf{x}^b]$

Anomaly Initialization - **AI** Observed anomalies are added to model climatology
$\mathbf{x}^a = \mathbf{x}^b + \mathbf{H}^T[\mathbf{y}^{pso} - \mathbf{H}\mathbf{x}^b]$, $\mathbf{y}^{pso} = \mathbf{y}^o - (\bar{\mathbf{y}^o} - \mathbf{H}\bar{\mathbf{x}})$

with $\mathbf{x}^b$ a long control solution and $\mathbf{y}^{pso}$ the pseudo-obs.

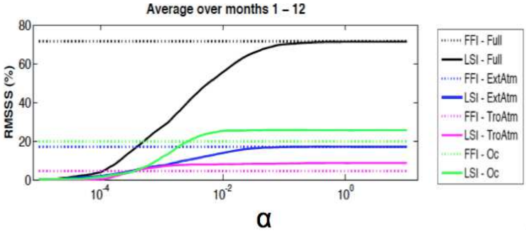
- FFI reduces RMSE in the short-term but it requires a "time dependent" bias-reduction approach to account for the drift (a posterior)
- AI maintains the model trajectory on its own attractor so that drift is reduced (not eliminated !) but bias is strong. The posterior bias-reduction approach is more robust.
- FFI is more prone to create initial shocks and unbalances

Least-Square-Initialization ⇒ FFI-LSI

- ***LSI*** *improves the fit to the observations allowing for their informational content to be propagated to the entire model domain.*
- ***LSI*** *merges observation and model, and the model error covariance is estimated using the statistics of the anomalies.*

$$\vec{x}^a = \vec{x}^b + BH^T\left[HBH^T + R\right]^{-1}\left[\vec{y}^o - H\vec{x}^b\right]$$

with: $B = \alpha\overline{\left(\vec{x}-\bar{\vec{x}}\right)\left(\vec{x}-\bar{\vec{x}}\right)^T}$



- LSI improves the performance of FFI in all the situations when only a portion of the systems state is observed.
- LSI operates an efficient propagation of information from data-covered to data-uncovered areas and to some extent reduces the initial error also far from the observational locations.

Exploring Parameter Uncertainty ⇒ FFI-EPU

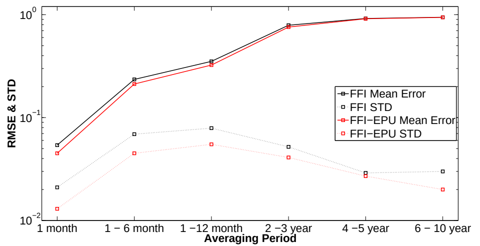
- *EPU provides an online correction of the drift based on a linear and short-time approximation of its evolution.*
- *Work Hyp: (1) Select Uncert. Param.; (2) Range of possible param.* $[\lambda_{\min}, \lambda_{\max}]$

$$\vec{x}^{un}(t_i) = \vec{x}(t_i) - \vec{b}(t_i) = \vec{x}(t_i) - \left.\frac{\partial F}{\partial \lambda}\right|_{\vec{x}(t_{i-1}),\vec{\lambda}} \delta\lambda_i \Delta T_{Bias}$$

- *The unknown parametric error is sampled using the assumed uniform distribution of possible parameter values*

- The use of EPU has clearly improved the skill of FFI within the first forecast year.
- Later the limits of accuracy of the linear and short-time assumptions at the basis of EPU are approached, and only minor advantages over FFI are recorded

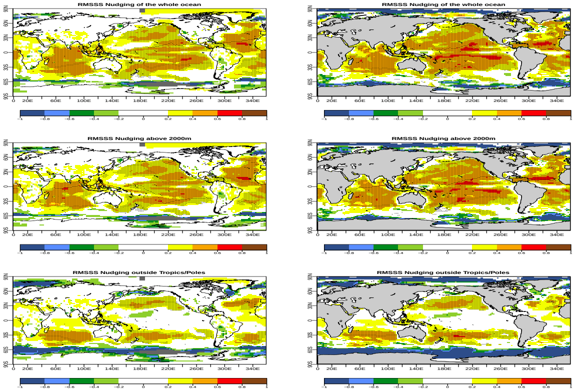
FFI vs FFI-EPU



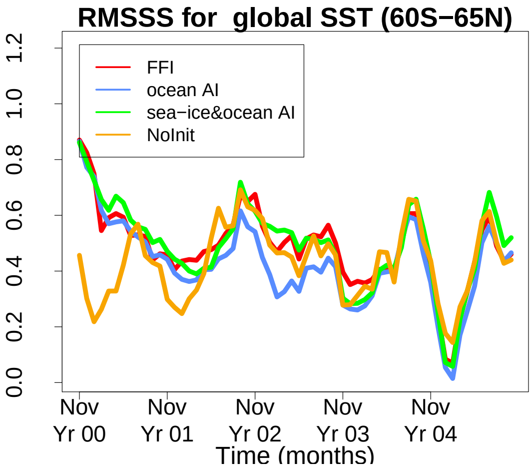
Nudging Experiments with Ec-Earth climate model

- Root Mean Square Skill Score (RMSSS) of the ensemble-mean near surface temperature (left) and SST (right) anomalies
- Three Exps: (1) GLOBAL NUDGING, (2) UPPER OCEAN NUDGING ($\leq$2000mt); (3) EXTRATROPICS NUDGING
- The black dots indicate the region where the ratio of the RMSSS is significant to the 95% level.

RMSSS *T2m* **RMSSS** *SST*



Nudging Experiments with Ec-Earth climate model



Prospects & New Challenges in DA

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)

Prospects & Problems in Coupled DA

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

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

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.

thanks