Data Assimilation Training Course
Final Discussion and Q&A

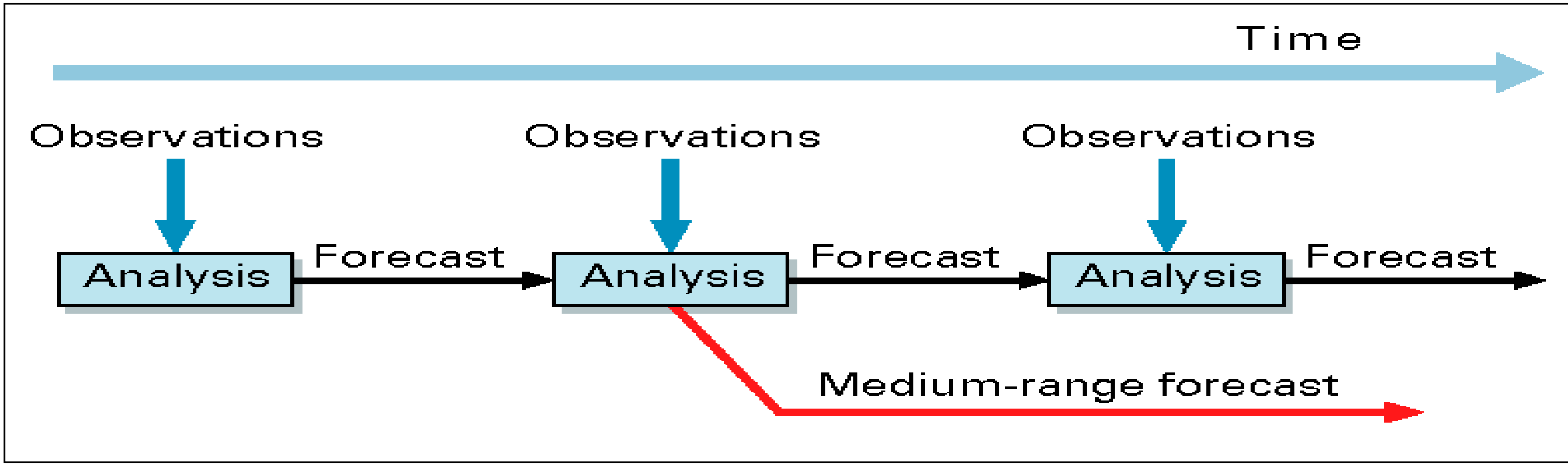
Data Assimilation

Data Assimilation has two main goals:

- Optimally blend information from observations and model to produce an accurate and physically consistent estimate of the initial state of the atmosphere and of the other components of the Earth System

- Quantify the uncertainty of our estimate of the initial state (this is necessary to be able to initialise an ensemble forecast!)

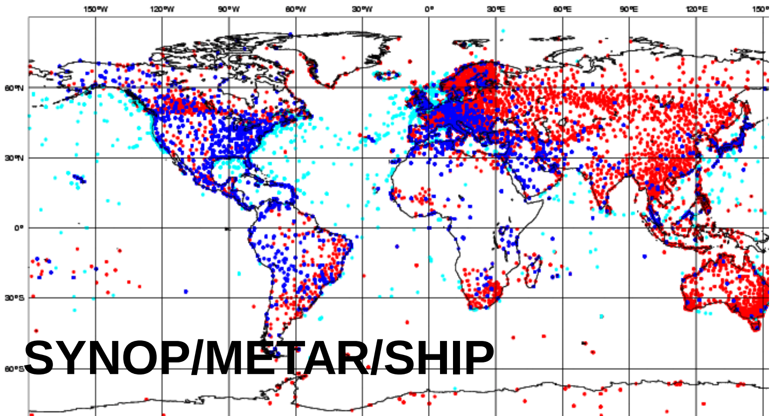
The Data assimilation cycle

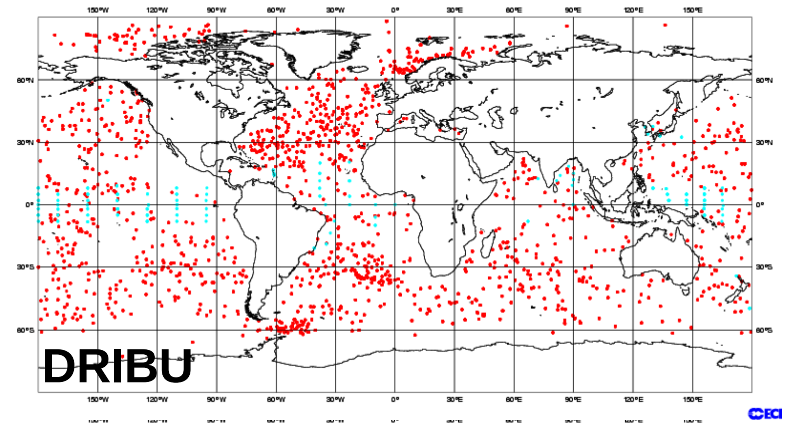


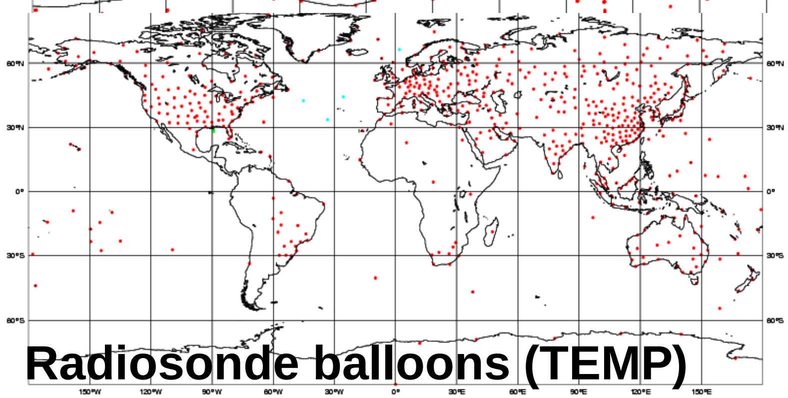
- An analysis is not produced by observations alone!
- The observations are used to correct errors in the short forecast from the previous analysis time (the background forecast).
- The background carries information from past observations into the current analysis
- The analysis is constructed so as to respect the physical and dynamical balances of the model ⇨ the model is an integral part of the analysis algorithm

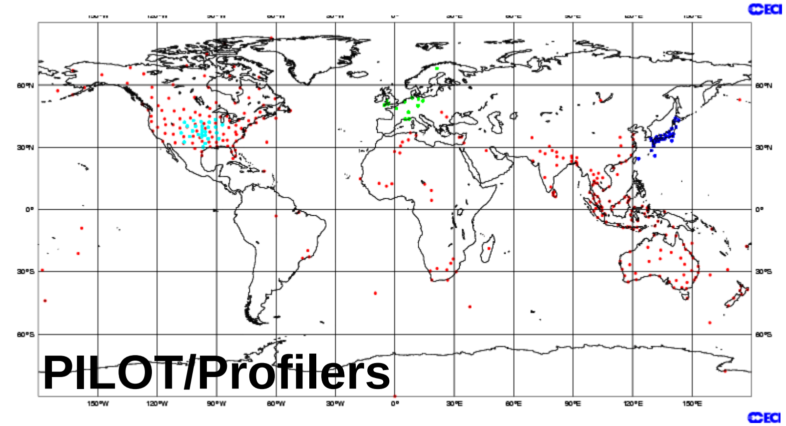
The observations

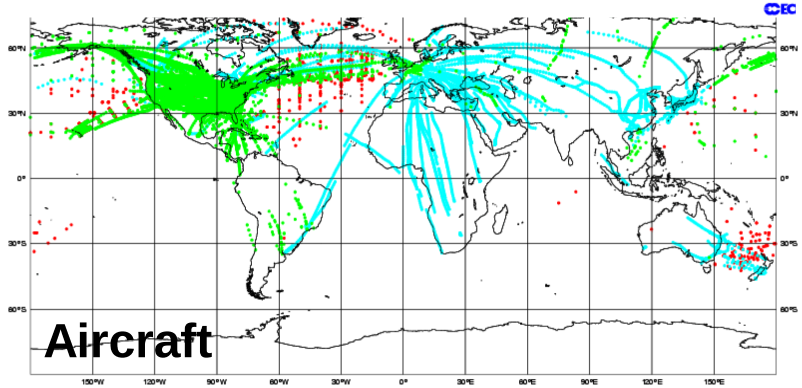
In situ observations







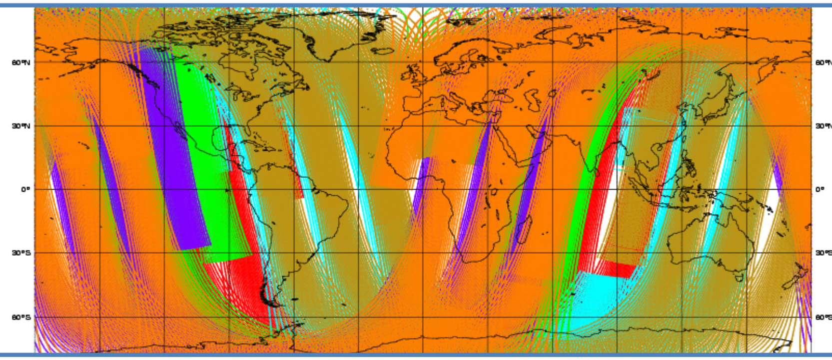




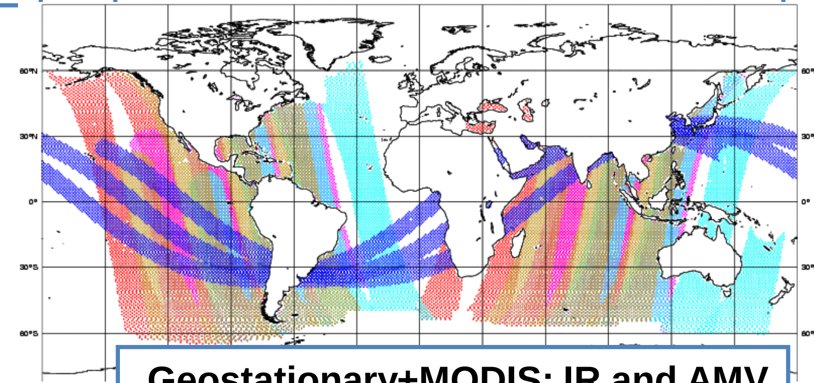
Lars Isaksen's talk

Satellite observations

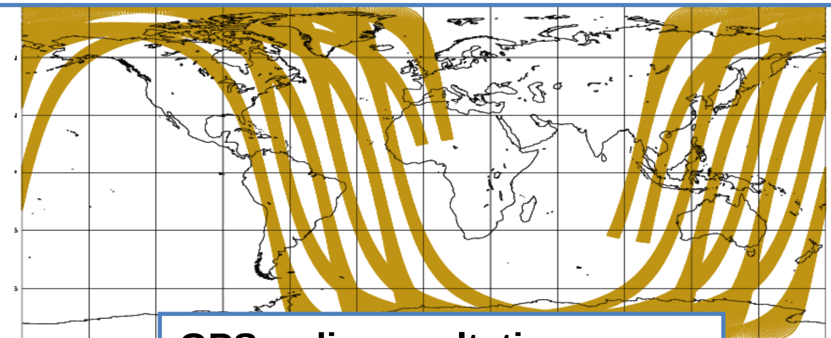
Sounders: NOAA AMSU-A/B, HIRS, AIRS, IASI, MHS



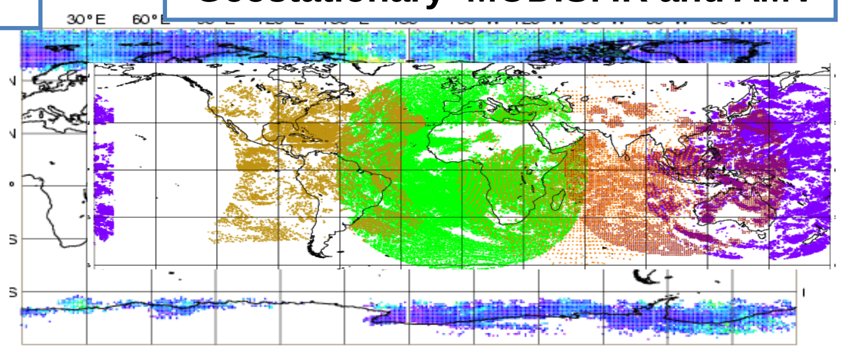
Imagers: SSMI, SSMIS, AMSR-E,



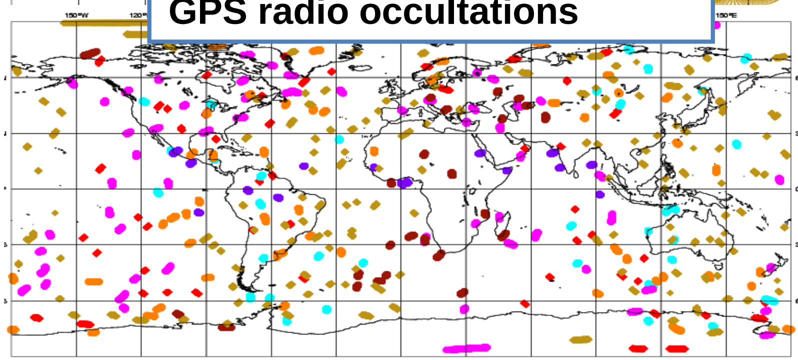
Scatterometer ocean low-level winds: ASCAT



Geostationary+MODIS: IR and AMV



GPS radio occultations



Tony McNally and Lars Isaksen talks

Observation errors

- Observations are affected by errors of different types
- Denoting $\mathbf{y}^*$ as the true observations of the model state ($\mathbf{y}^* = \mathcal{H}(\mathbf{x}^*)$):

$$\mathbf{y} - \mathbf{y}^* = \varepsilon_o = \varepsilon_G + \varepsilon_M + \varepsilon_R + \varepsilon_H$$

$\varepsilon_G =$ Gross errors (incorrect coding of observation, duplicates, incorrect location, wrong cloud clearing, etc.).

$\varepsilon_M =$ Measurement errors (instrument noise)

$\varepsilon_R =$ Representativity errors (e.g., in situ observations compared to grid point model value)

$\varepsilon_H =$ Observation operator (Forward model) errors (e.g., errors in the radiative transfer model, interpolation errors, etc.)

Observation errors

$$\mathbf{y} - \mathbf{y}^* = \varepsilon_o = \varepsilon_G + \varepsilon_M + \varepsilon_R + \varepsilon_H$$

- ε_G (gross errors) are dealt with by Observation Quality Control techniques (Variational Quality Control; Elias Holm's talk)

- Observations are assumed to be un-biased:

$$\langle \varepsilon_o \rangle = 0$$

- Biases are dealt with specific Bias Correction techniques: at ECMWF this is part of the analysis algorithm itself (e.g., Variational Bias Correction: Niels Bormann's talk)

Observation errors

- In common DA algorithms we require not only the observations to be un-biased but also the background forecast to be un-biased:

$$\langle \varepsilon_b \rangle = 0$$

- But our only source of information about observation and forecast errors are observation departures:

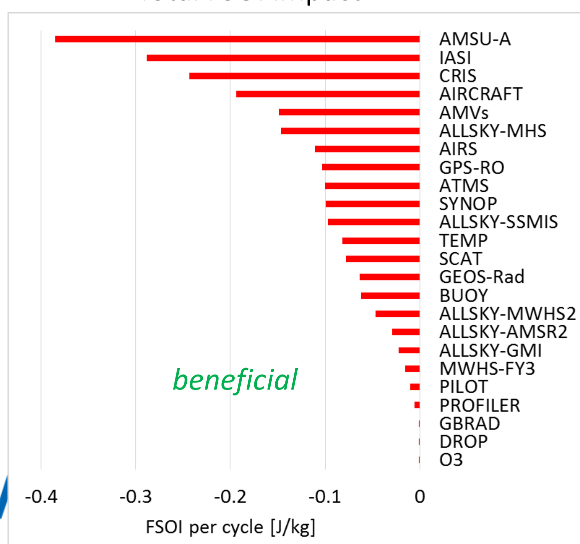
$$y - H(x_b)$$

- We need to make further assumptions in order to disentangle observation and model error (Niels Bormann and Patrick Laloyaux's talks)

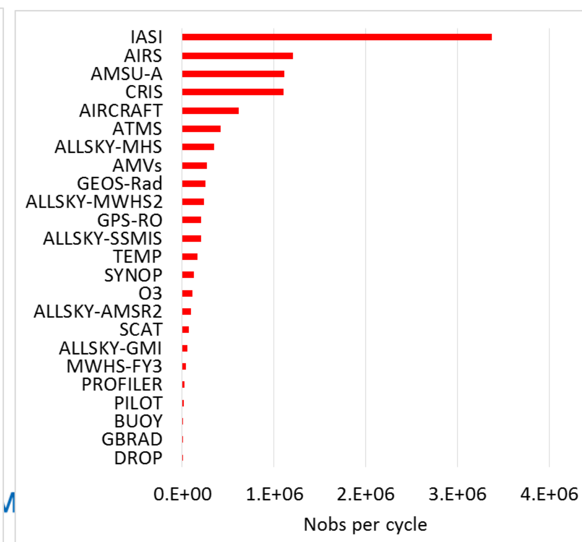
Observation impact

- It is also important to monitor and evaluate the impact different types of observations have on the quality of the analyses and forecasts
- To do this we routinely look at observation departures (with respect to both analysis and forecast fields: see Lars Isaksen's talk and practical sessions)
- We can also perform Observing System Experiments (OSEs: Cristina Lupu's talk)
- We routinely compute adjoint-based diagnostic quantities (Forecast Sensitivity to Observation Impact: Cristina Lupu's talk)

Total FSOI impact

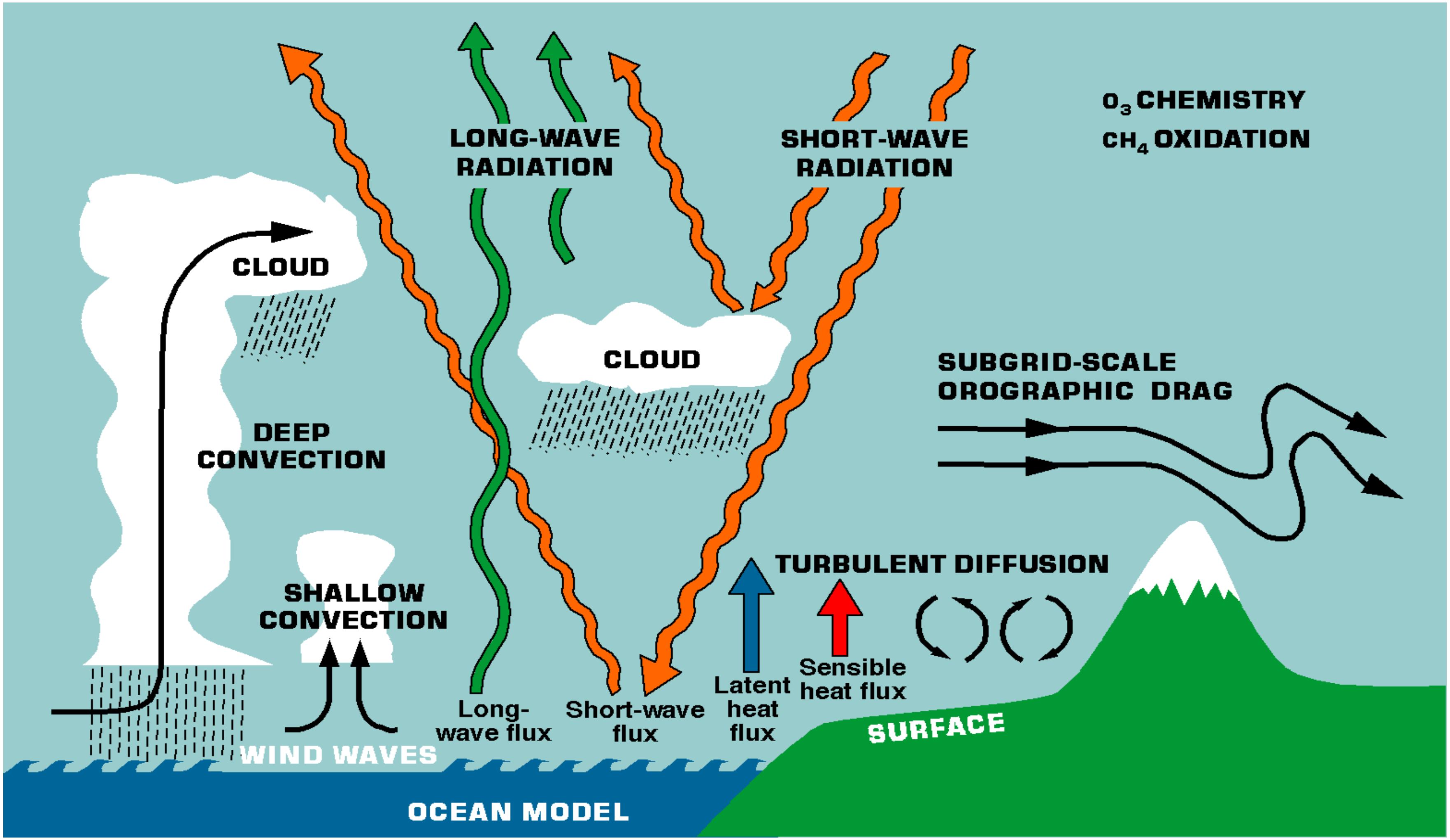


Total data count



The forecast model

The forecast model is a very important part of the data assimilation system



Most important physical processes in the ECMWF model

The forecast model

- A good model is able to effectively propagate information from past observations to the current analysis update => new batch of observations will only produce small corrections to the background => we are closer to the conditions of linearity of errors where current DA algorithms work best

- In incremental 4D-Var we not only require the full non-linear model to advance the state in time

- We also need its linearised versions (Tangent Linear and Adjoint) to propagate increments with respect to a linearisation trajectory forward and backwards in time during the assimilation window (update of J and computation of ∇J)

- Developing and maintaining TL and ADJ codes is a complex task (Philippe Lopez and Angela Benedetti's talks): but the availability of sophisticated TL and ADJ models is one of the main reasons for ECMWF success

Model errors

- Despite their increasing complexity and sophistication models are far from perfect!
- Many sources of model error: missing physical processes, errors in parametrizations of physical processes, discretisation errors (from continuous PDEs to discrete formulation), etc.,
- We represent model errors in two ways:

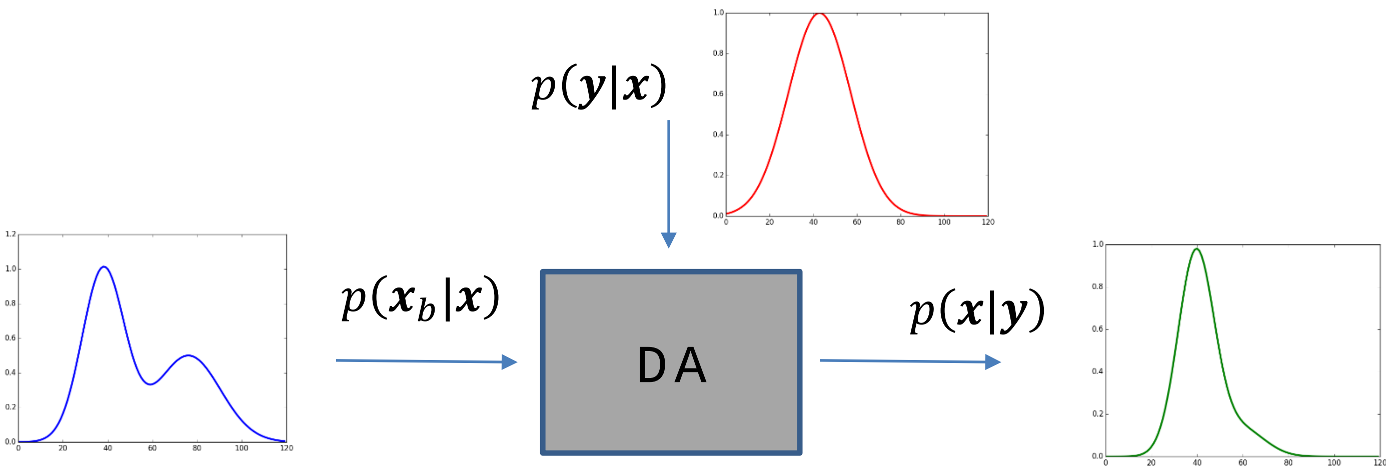
  1. Stochastic errors: explicitly perturbing the model integration in our ensemble data assimilation system (EDA; see Massimo Bonavita's talk – Assimilation Algorithms (5))
  2. Model biases: Using an explicit model error term in the 4D-Var cost function (weak constraint 4D-Var: see Sebastien Massart talk on 4D-Var and Patrick Laloyaux's talk on Model Error)

- A lot of work still needs to be done in this area, especially in terms of diagnosing the model error statistics

Blending observations and model information: the Bayes perspective

The Bayes perspective

- At an abstract level, we can think of the analysis process as updating our prior knowledge about the state, represented by a background forecast and the pdf of its errors, with new observations, represented by their values and the pdf of their errors:



$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}_b|\boldsymbol{x})}{p(\boldsymbol{y})} \propto p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}_b|\boldsymbol{x})$$

- $p(\boldsymbol{x}_b|\boldsymbol{x})$ = prior pdf (encapsulate our knowledge about the state before new observations)
- $p(\boldsymbol{y}|\boldsymbol{x})$ = observations likelihood (pdf of the observations conditioned on the state)
- $p(\boldsymbol{x}|\boldsymbol{y})$= posterior pdf (updated pdf of the state after the analysis)
- $p(\boldsymbol{y})$= marginal pdf of the observations (does not depend on $\boldsymbol{x}$: normalising constant in Bayes' law)

Particle Filters

$$p(\boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}_b|\boldsymbol{x}) \qquad (1)$$

- In principle an analysis update requires being able to compute the product pdf of the random variables $\boldsymbol{y}$, $\boldsymbol{x}_b$. This is usually not possible to do unless we choose very specific functional forms for the pdfs
- We thus need to make approximations
- One idea is to use Monte Carlo methods to sample and propagate the pdfs in (1) by an ensemble of states: Particle Filters
- This does not work (yet!) for high dimensional systems as in NWP
- Need to make further assumptions on (1)

Kalman Filter methods

- Need to make further assumptions on (1)
- Gaussian error pdfs => Gaussian posterior pdf

$$p(\mathbf{x}_a|\boldsymbol{y}) = \mathcal{N}(\mathbf{x}_a, \mathbf{P}^a)$$

$$\boldsymbol{x_a} = \boldsymbol{x_b} + \mathbf{K}(\boldsymbol{y} - \mathbf{H}(\boldsymbol{x_b}))$$

$$\mathbf{P}^a = (\mathbf{I} - \mathbf{KH})\mathbf{P}^b(\mathbf{I} - \mathbf{KH})^T + \mathbf{KRK}^{\boldsymbol{T}} = (\mathbf{I} - \mathbf{KH})\mathbf{P}^b$$

$$\mathbf{K} = \mathbf{P}^b\mathbf{H}^T\left(\mathbf{HP}^b\mathbf{H}^T + \mathbf{R}\right)^{-1} = \left(\left(\mathbf{P}^b\right)^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{R}^{-1}$$

- Solving directly these equations lead to Kalman Filter type DA methods: Optimum Interpolation, Kalman Filter, Extended KF, Ensemble KF (Massimo Bonavita's talk on KF and EnKF)
- These methods work well with low dimensional systems or small number of observations (O.I. in Snow analysis; Extended KF for soil moisture analysis, e.g. Patricia De Rosnay's talk on Land Data assimilation)
- For high-dim systems they require localisation which can limit the amount of information we are able to extract from non-local observations like satellite radiances

Variational methods

- The Kalman Filter analysis update equation can be formulated as an equivalent minimization problem:

$$J(\mathbf{x}_0) = (\mathbf{x}_b - \mathbf{x}_o)^T (\mathbf{P}^b)^{-1} (\mathbf{x}_b - \mathbf{x}_o) + \sum_{t=0}^{T} \left(\mathbf{y}_t - H_t M_{0\rightarrow t}(\mathbf{x}_0)\right)^T \mathbf{R}_t^{-1} \left(\mathbf{y}_t - H_t M_{0\rightarrow t}(\mathbf{x}_0)\right)$$

- This is the basis of Variational methods (3D-Var, 3D-Var FGAT, 4D-Var: see Sebastien Massart's lectures)
- Solving the KF update through iterative algorithms (conjugate gradient, Newton's methods)
- These methods do not require direct access to the elements of the error covariance matrices. We can represent error covariances by operators (i.e., pieces of code) acting on increments (see Elias Holm talk on background error modelling)
- Variational methods work well on high dimensional systems and are generally used in global NWP

Hybrid Data Assimilation methods

- The Kalman Filter equations require estimating and advancing in time not only the state but also its error covariance:

$$\mathbf{P}^a{}_t = (\mathbf{I} - \mathbf{KH})\mathbf{P}^b{}_t\,(\mathbf{I} - \mathbf{KH})^T + \mathbf{KRK}^T$$

$$\mathbf{P}^b{}_{t+1} = \mathbf{M}\,\mathbf{P}^a{}_t\,\mathbf{M}^T + \mathbf{Q}_{t+1}$$

- 4D-Var can implicitly do this but only inside the assimilation window (12 hours at ECMWF)

- The idea of Hybrid DA methods is to combine a variational DA system to estimate the state with an ensemble data assimilation system (EnKF/EDA) to estimate and cycle the errors of the state (see Massimo Bonavita's talk on Hybrid data assimilation)

- Ensemble DA systems also provide the initial conditions for Ensemble Prediction

- All major global NWP Centres run Hybrid DA systems for Atmospheric DA

Earth System Data Assimilation

- We have discussed Data Assimilation methods with an emphasis on global Atmospheric NWP applications
- The DA methods presented are however general: which one to apply to a given problem depends on the characteristics of the problem (size of the state vector, number and quality of observations, available computing resources, available manpower,…)
- You have seen applications in Atmospheric Composition DA (4D-Var: Melanie Ades's talk); in Ocean Data Assimilation (3D-Var FGAT: Hao Zuo's talk); in Land Data Assimilation (O.I., Simplified Extended KF: Patricia de Rosnay's talk)
- In current ECMWF DA the Earth system's components are at most only weakly coupled (through a coupled model background forecast)
- Phil Browne's talk has given you a sense of some of the challenges and the potential benefits of a stronger coupling in the data assimilation for the different components of the Earth System

Earth System Data Assimilation

- We have discussed Data Assimilation methods for the Earth System with an emphasis on producing the best initial state estimate for forecasting at short, extended and even seasonal timescales

- An increasingly important application of Earth System DA is to help to reconstruct the past climate and weather (see Dinand Scheper's talk on Reanalysis methods)

- As DA methods have dramatically improved over the years we are able to make better use of past observational records and more robustly estimate climatic trends

Earth System DA: Challenges

- We have tried to provide you with a description of the state of the art in data assimilation methods for Earth System applications

- More advanced material about current topics and challenges:

ECMWF Seminar 2018
On Earth System Assimilation

Reading, 10-13 September 2018

https://www.ecmwf.int/en/learning/workshops/annual-seminar-2018

Thank you for being such an attentive audience!

Thank you for being such an attentive audience!

Bayesian brain teaser

“Your favourite anti-spam software has 98% accuracy in discovering spam emails. On average 1% of the email we receive are spam.
If an email you have received is labelled as spam, is it more likely to actually be spam or not?”

Answers to Massimo.Bonavita@ecmwf.int (No spam please!)