Ensemble methods for data assimilation

Stefano Migliorini

National Centre for Earth Observation

University of Reading

Contents

- The weather prediction problem
- Need for data assimilation
- The Kalman filter
- Data assimilation techniques for NWP
- Ensemble methods: features and issues
- Examples with idealized models

Numerical weather prediction

- The mass conservation equation, the momentum equation, the energy equation, the concentration equation for humidity and the equation of state give expressions for the rates of change of $\mathbf{x}(\mathbf{r},t) = (\mathbf{v}, T, p, \rho, q)$
- all the properties of the system can *in principle* be calculated from nonlinear partial differential equations (PDE's) $\mathbf{x}_{t_1} = M(\mathbf{x}_{t_0})$ and boundary and initial conditions.

Predictability

- Consider two initial atmospheric states $\mathbf{x}_{t_0}$ and $\bar{\mathbf{x}}_{t_0}$ which differ by an infinitesimally small quantity (or "error") $\boldsymbol{\varepsilon}_0 = \mathbf{x}_{t_0} - \bar{\mathbf{x}}_{t_0}$
- A initial difference (or "error") between two solutions (or "analogs") to the equations for the evolution of the (atmospheric) state may grow with time

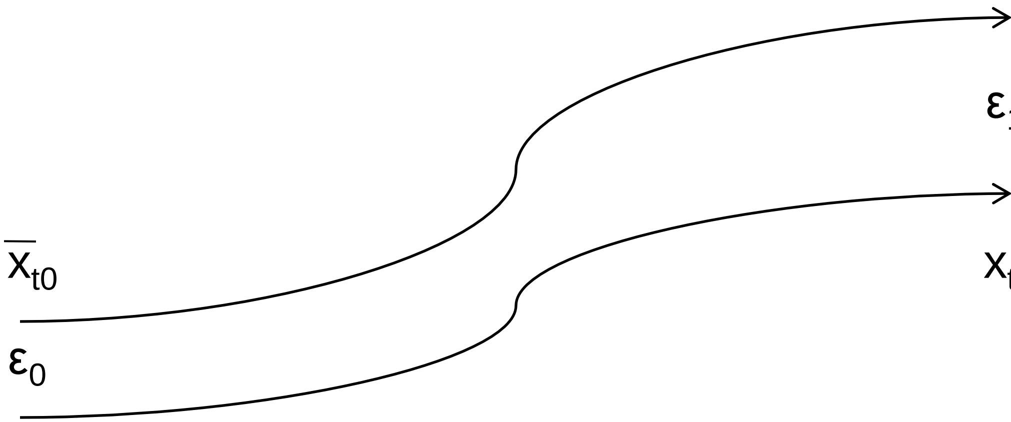
$$\mathbf{x}_{t_1} = M(\mathbf{x}_{t_0}) = M(\bar{\mathbf{x}}_{t_0} + \boldsymbol{\varepsilon}_0) = M(\bar{\mathbf{x}}_{t_0}) + \mathbf{M}(\bar{\mathbf{x}}_{t_0})\boldsymbol{\varepsilon}_0$$

$$\boldsymbol{\varepsilon}_1 = M(\mathbf{x}_{t_0}) - M(\bar{\mathbf{x}}_{t_0}) = \mathbf{M}(\bar{\mathbf{x}}_{t_0})\boldsymbol{\varepsilon}_0$$

$$\boldsymbol{\varepsilon}_n = \mathbf{M}(\bar{\mathbf{x}}_{t_{n-1}})\cdots\mathbf{M}(\bar{\mathbf{x}}_{t_0})\boldsymbol{\varepsilon}_0 \equiv \mathbf{M}(t_n : t_0)\boldsymbol{\varepsilon}_0$$

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

University of Reading



$\bar{x}_{t1}=M(\bar{x}_{t0})$

ε_1

$\bar{x}_{t0}$

$x_{t1}=M(x_{t0})=M(\bar{x}_{t0})+\mathbf{M}(\bar{x}_{t0})\varepsilon$

ε_0

x_{t0}

$x_{t1}=\bar{x}_{t1}+\varepsilon_1$

$x_{t0}=\bar{x}_{t0}+\varepsilon_0$

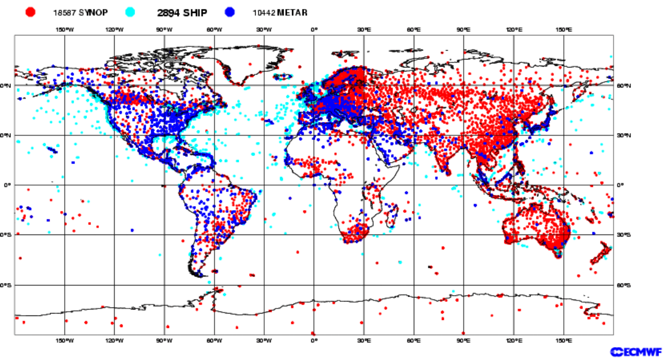
The need for data assimilation

- Weather and climate forecasts are uncertain due to uncertainty in
  - Initial and boundary conditions
  - Models (e.g., hydrostatic assumption or shallow water approximation; discretization)
  - Parameters (e.g., CO_2 sources and sinks)
- Not only we need accurate models but also good knowledge of initial (and boundary) conditions
- It is essential that good quality, widespread and frequent measurements (e.g., from satellites) are assimilated in NWP models in order to achieve good quality forecasts

Data coverage at ECMWF

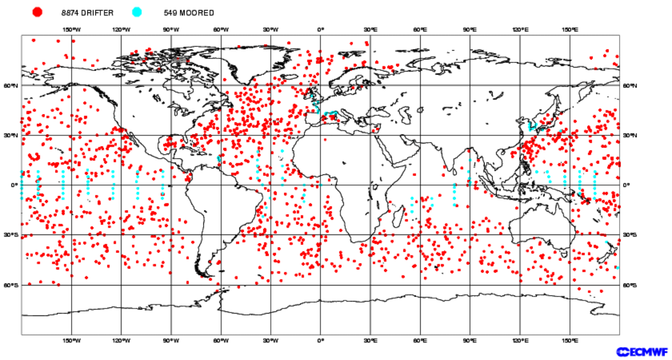
surface stations

ECMWF Data Coverage (All obs DA) - SYNOP/SHIP
12/NOV/2010; 00 UTC
Total number of obs = 31923



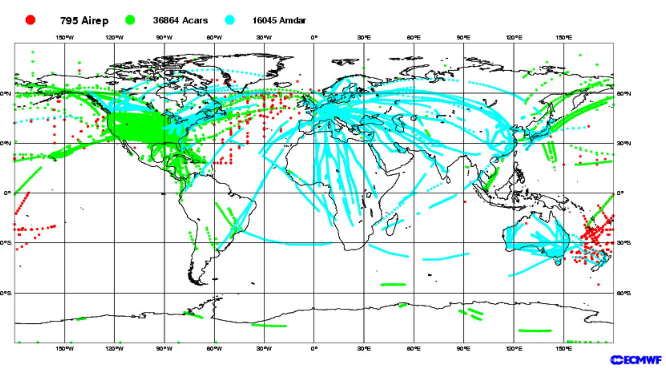
buoys

ECMWF Data Coverage (All obs DA) - BUOY
12/NOV/2010; 00 UTC
Total number of obs = 9423



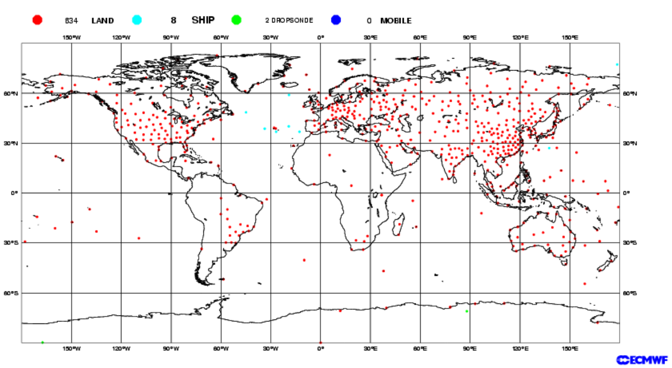
aircraft

ECMWF Data Coverage (All obs DA) - AIRCRAFT
12/NOV/2010; 00 UTC
Total number of obs = 53704



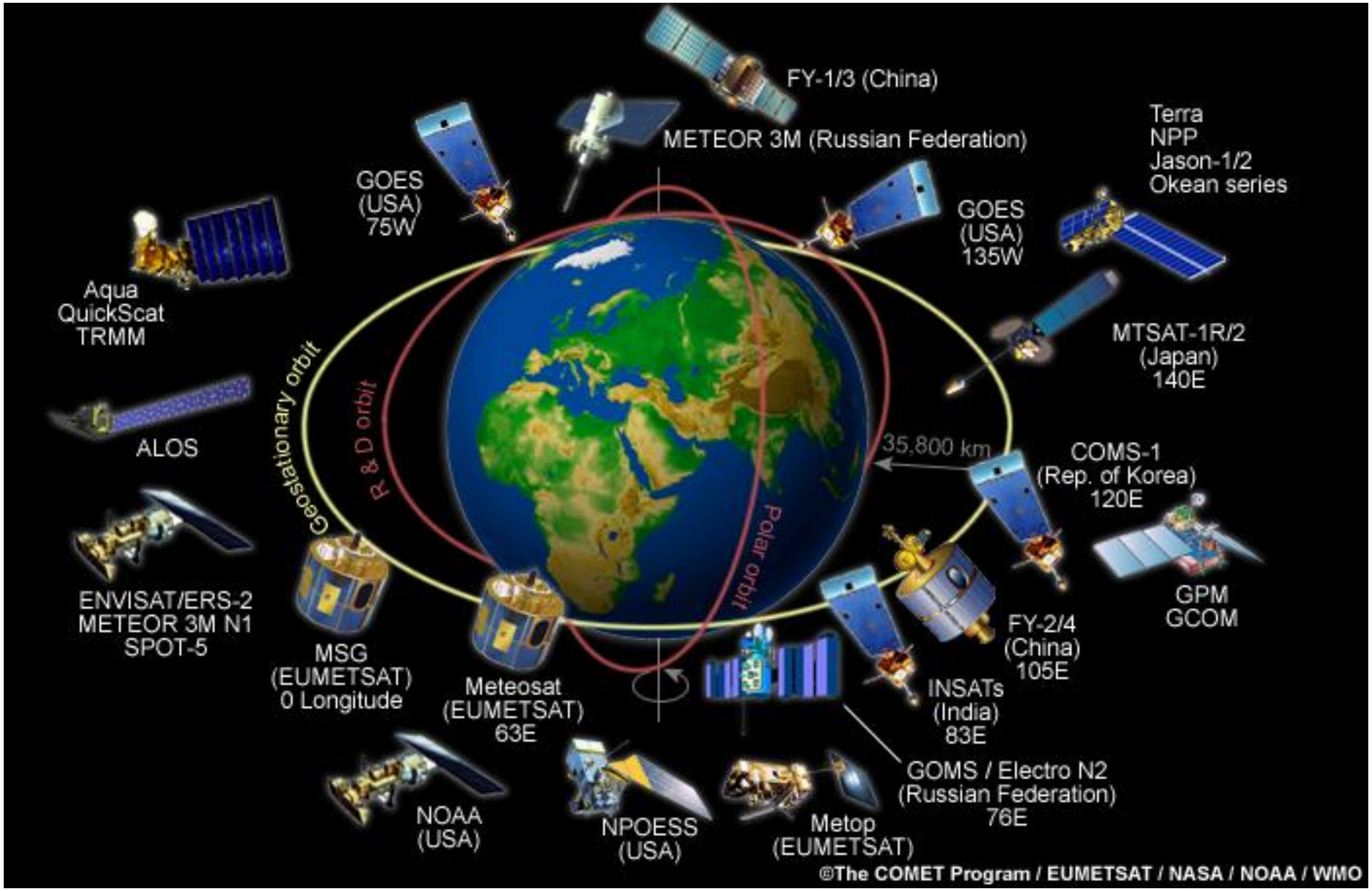
radiosondes

ECMWF Data Coverage (All obs DA) - TEMP
12/NOV/2010; 00 UTC
Total number of obs = 644



Courtesy

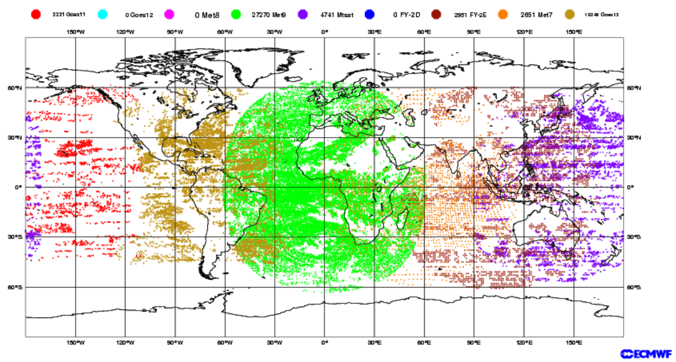
The global satellite operational observing system



Data coverage at ECMWF

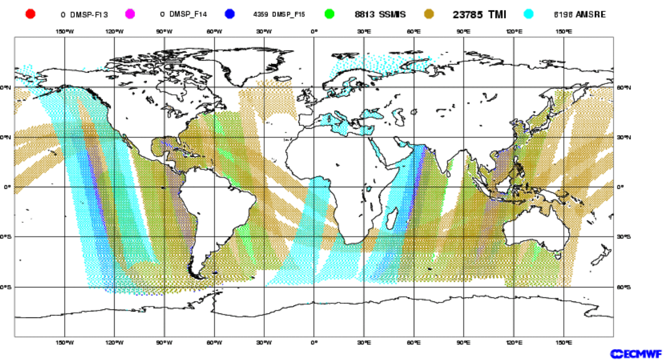
Infrared AMVs

ECMWF Data Coverage (All obs DA) - AMV IR
12/NOV/2010; 00 UTC
Total number of obs = 50083



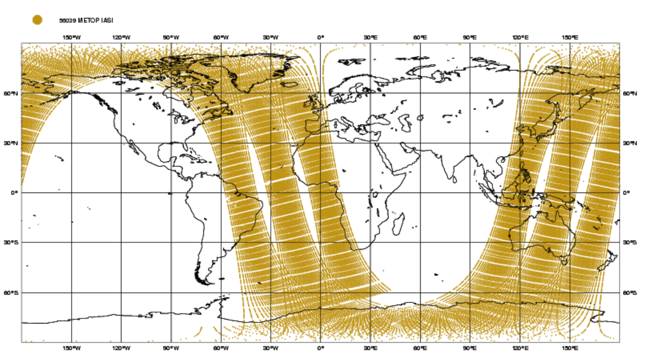
Microwave imagers

ECMWF Data Coverage (All obs DA) - Microwave imager
12/NOV/2010; 00 UTC
Total number of obs = 43153



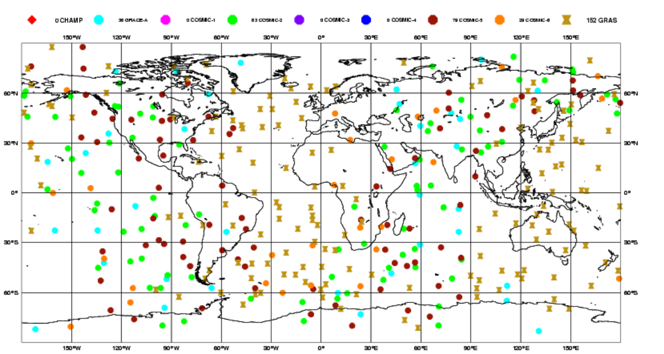
IASI

ECMWF Data Coverage (All obs DA) - IASI
12/NOV/2010; 00 UTC
Total number of obs = 56039



Bending angle

ECMWF Data Coverage (All obs DA) - GPSRO
12/NOV/2010; 00 UTC
Total number of obs = 379



Courtesy ECMWF

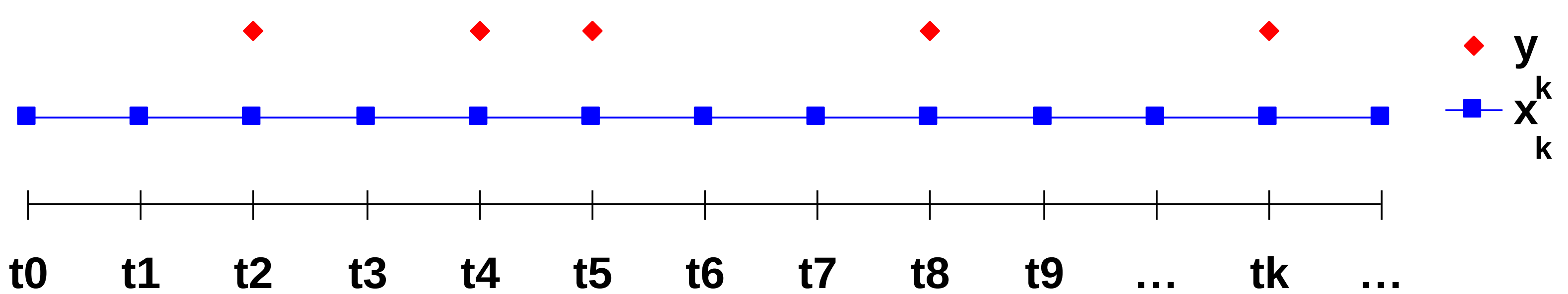
Observation model

- Discrete stochastic observation model

$$\mathbf{y}_k = H(\mathbf{x}_k) + \boldsymbol{\varepsilon}_k$$

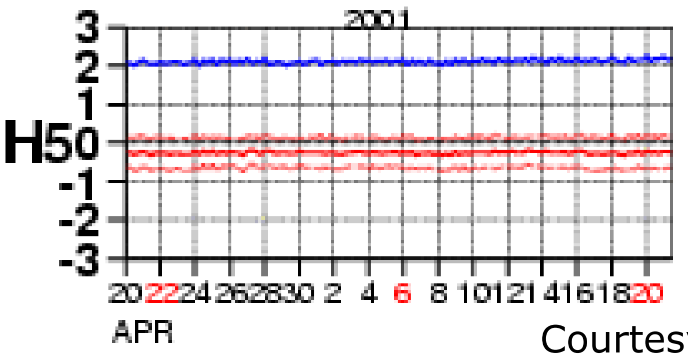
- Where we assume:
- $E\{\varepsilon_k\}=0$ for all k
- $E\{\mathbf{x}_0\varepsilon_k^T\} = 0$
- $E\{\varepsilon_k\eta_j^T\} = 0$

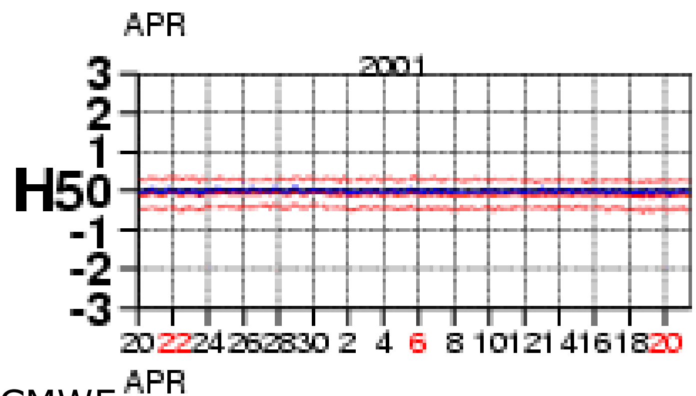
$$E\{\boldsymbol{\varepsilon}_k\boldsymbol{\varepsilon}_j^T\} = \begin{cases} \mathbf{R}_k & j = k \\ \mathbf{0} & j \neq k \end{cases}$$



Observational issues

- Bias monitoring





Courtesy Dick Dee ECMWF

HIRS channel 5 on NOAA-14 satellite HIRS channel 5 on NOAA-16 satellite

- Quality control

obs rejected when

$$\mathrm{cov}(\mathbf{y}-\mathbf{Hx}_b)=\mathbf{R}+\mathbf{HBH}^T \qquad y-hx_b \geq \lambda(\sigma_o^2+h^2\sigma_b^2)^{1/2}$$

- Correlated observation error

Whithening filter: $$\mathbf{R}^{-1/2}\mathbf{y}=\mathbf{R}^{-1/2}\mathbf{Hx}_t+\mathbf{R}^{-1/2}\boldsymbol{\varepsilon}_o$$

Stochastic-dynamic model

- In the presence of model error $\boldsymbol{\eta}_k$ the dynamical system is described by $\mathbf{x}_k = M(\mathbf{x}_{k-1}) + \boldsymbol{\eta}_k$
- Where $\mathbf{x}_0$ random with $E\{\mathbf{x}_0\} = \mathbf{m}_0$ and $E\{(\mathbf{x}_0 - \mathbf{m}_0)(\mathbf{x}_0 - \mathbf{m}_0)^T\} = \mathbf{P}_0$
- We assume (simplifying assumptions!):
- $E\{\boldsymbol{\eta}_k\} = 0$ for all k and $E\{\boldsymbol{\eta}_k \boldsymbol{\eta}_j^T\} = \begin{cases} \mathbf{Q}_k & j = k \\ \mathbf{0} & j \neq k \end{cases}$
- Model error uncorrelated with the initial state: $E\{\boldsymbol{\eta}_k \mathbf{x}_0^T\} = 0$
- This means $\mathbf{x}_k$ is random with probability density $p(\mathbf{x}_k)$
- $\mathbf{x}_k$ (or, equivalently, $p(\mathbf{x}_k)$) is to be determined

The data assimilation problem

- Consider a set of realizations of observations

$$Y_l = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_l\}$$

- When $k \le l$ (i.e., for $t_k \le t_l$), the *conditional* probability density $p(\mathbf{x}_k|Y_l)$ provides the solution of the data assimilation (or filtering) problem
- From $p(\mathbf{x}_k|Y_l)$ we could calculate $E\{\mathbf{x}_k|Y_l\}$ (the estimate of $\mathbf{x}_k$ at time t_k or analysis) and the covariance matrix

Data assimilation for NWP

- Is data assimilation (or probabilistic forecasting) a **solved** problem?? Certainly not for NWP applications!
- If we sample the probability to find a random variable between 0 and 1 with a 0.01 resolution, we need 101 sample points (say 100 for simplicity)
- How many sample points do we need to map the probability space in the case of a random vector with 7 components? (think of $\mathbf{x}(\mathbf{r},t) = (\mathbf{v}, T, p, \rho, q)$)

The "curse of dimensionality"

- We need, you guessed it, $100^7 = 10^{14}$ sample points. And to store such a number on a computer we need a hundred terabyte's memory space, for each grid point! This is a lot of hard-disks…
- Note that the current Met Office operational global NWP model considers 1024 x 769 grid points over 70 vertical levels (~55 million grid points) at a given time!
- This means that, by today standards, it is not possible to sample the pdf of the state vector used in NWP

The Kalman filter

- The good news is: under certain conditions we may not need to keep track of the whole joint state pdf
- If all relevant errors (i.e., model, initial conditions and observations errors) are Gaussian, their pdfs are completely known if their means and covariances are known
- When the model and the observation operator are linear, an optimal estimate of the mean and the covariance of $p(\mathbf{x}_k|Y_l)$ are given by the Kalman filter, that is the solution of the data assimilation problem.
- In the presence of (moderate) nonlinearities (typical case in NWP): extended Kalman filter

The extended Kalman filter

- Let now consider the nonlinear case with Gaussian errors

$$\boldsymbol{\varepsilon}_{k+1} \sim N(\mathbf{0}, \mathbf{R}_{k+1}) \quad \mathbf{x}_k \sim N(\mathbf{x}_k^a, \mathbf{P}_k^f) \quad \boldsymbol{\eta}_{k+1} \sim N(\mathbf{0}, \mathbf{Q}_{k+1})$$

- Forecast step

$$\mathbf{x}_{k+1} = M(\mathbf{x}_k) + \boldsymbol{\eta}_{k+1} \qquad \mathbf{M} = \left.\frac{\partial M(\mathbf{x}_k)}{\partial \mathbf{x}_k}\right|_{\mathbf{x}_k = \mathbf{x}_k^a}$$

$$\mathbf{x}_{k+1}^f = M(\mathbf{x}_k^a) \qquad \mathbf{P}_{k+1}^f = \mathbf{M}\mathbf{P}_k^a\mathbf{M}^T + \mathbf{Q}_{k+1}$$

- Data assimilation step

$$\mathbf{y}_k = H(\mathbf{x}_k) + \boldsymbol{\varepsilon}_k \qquad \mathbf{H} = \left.\frac{\partial H(\mathbf{x}_k)}{\partial \mathbf{x}_k}\right|_{\mathbf{x}_k = \mathbf{x}_k^a}$$

$$\mathbf{x}_{k+1}^a = \mathbf{x}_{k+1}^f + \mathbf{K}(\mathbf{y}_{k+1} - H(\mathbf{x}_{k+1}^f)) \qquad \mathbf{K} = \mathbf{P}_{k+1}^f\mathbf{H}^T(\mathbf{H}\mathbf{P}_{k+1}^f\mathbf{H}^T + \mathbf{R}_{k+1})^{-1}$$

$$\mathbf{P}_{k+1}^a = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}_{k+1}^f$$

Example: scalar and linear case

- Updating at t_{k+1} prior information valid at t_k with y_{k+1}

$$x_{k+1} = a x_k + \eta_{k+1} \qquad y_{k+1} = x_{k+1} + \varepsilon_{k+1}$$

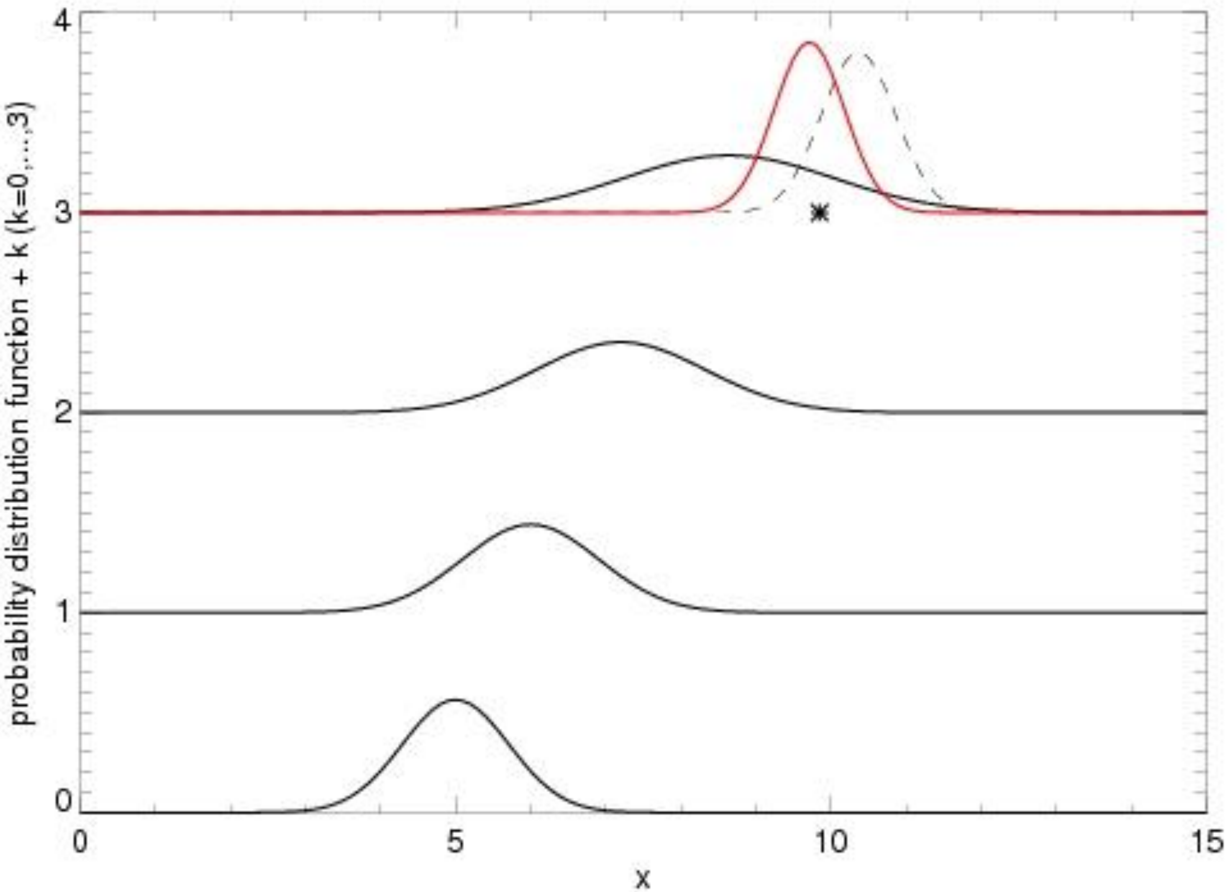
$$x_{k+1}^f = a x_k^f \qquad \mathbf{P}_{k+1}^f \equiv (\sigma_{k+1}^f)^2 = a^2 (\sigma_k^f)^2 + (\sigma_\eta^2)_k$$

$$\mathbf{K} = \frac{(\sigma_{k+1}^f)^2}{(\sigma_{k+1}^f)^2 + (\sigma_\varepsilon^2)_k} \qquad x_{k+1}^a = a x_k^f + \frac{(\sigma_{k+1}^f)^2}{(\sigma_{k+1}^f)^2 + (\sigma_\varepsilon^2)_k} (y_{k+1} - a x_k^f)$$

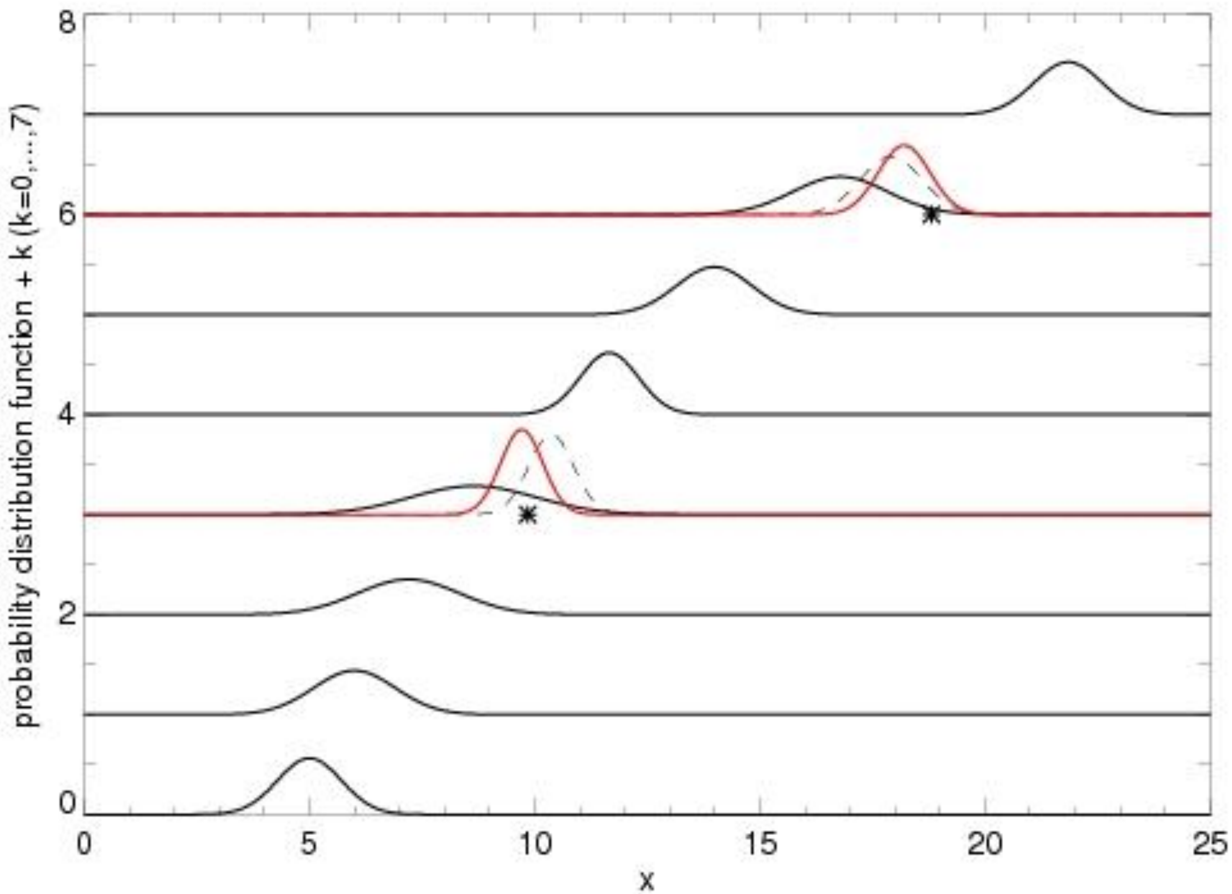
$$\mathbf{P}_{k+1}^a \equiv (\sigma_{k+1}^a)^2 = \frac{(\sigma_\varepsilon^2)_k}{(\sigma_{k+1}^f)^2 + (\sigma_\varepsilon^2)_k} (\sigma_{k+1}^f)^2 = \frac{(\sigma_\varepsilon^2)_k}{(\sigma_{k+1}^f)^2 + (\sigma_\varepsilon^2)_k} \left(a^2 (\sigma_k^f)^2 + (\sigma_\eta^2)_k \right)$$

Evolution of the pdf (with update)

$a = 1.2$
$x_0 = 5$
$(\sigma_0^f)^2 = 0.5$
$(\sigma^2_\eta)_0 = 0.1$



Evolution of the pdf: cycling



Use of Kalman filters in NWP

- To compute (not just store) the covariance matrix update $\mathbf{P}_{k+1}^{a}$, it can be shown that we need ~ n^3 Flops ($n \sim 10^7$ at the Met Office), i.e. 10^{21} Flops (floating-point operations)
- The IBM supercomputer at ECMWF has a maximal achieved performance of ~100 TFlops / second ~10^{14} Flops / s. We need ~10^7 s to compute $\mathbf{P}_{k+1}^{a}$, that is about five months!
- Clearly, we need to resort to some approximations and/or alternative strategies

Data assimilation techniques for NWP

- The most widespread data assimilation approach used for NWP is based on variational techniques
- 4D-Var calculates the optimal model trajectory over a given time window (typically 6h or 12h) by taking into account observations taken within the same window and some prior information.
- In the linear case, it is equivalent to a Kalman smoother initialized with the same prior information used in 4D-Var
- We now discuss an approximation of the Kalman filter based on a set of ensemble members (i.e., based on Monte Carlo techniques)

Ensemble methods

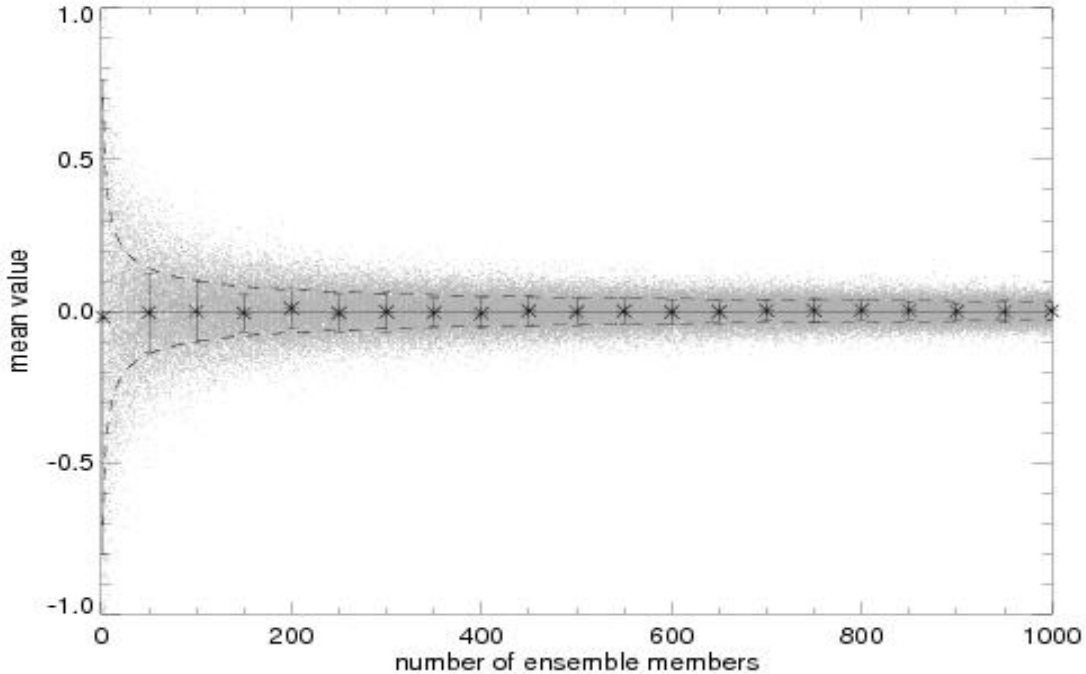
- Consider a random variable x distributed according f(x), with unknown mean μ and variance σ^2
- From f(x) we can generate a ensemble of N independent realizations of x: $x_1, x_2, \ldots, x_n$.
- We can estimate μ and σ^2 via the sample (or ensemble) mean $\bar{x}$ and sample variance *s*:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

- They are both unbiased estimators and their variance is

$$\mathrm{Var}(\bar{x}) = \frac{\sigma^2}{N} \qquad \mathrm{Var}(s^2) = \frac{2\sigma^4}{N-1}$$

Standard deviation of the mean: empirical vs. expected values



- Gray dots: mean values of N ensemble members drawn from a Gaussian with zero mean and unit variance, for a total of 100 mean values for each N value
- asterisks: sample mean of 100 means and sample standard deviation of the mean
- Solid line: expected value of the mean (= 0) and (dashed) of the standard deviation of the mean (= 1/sqrt(n))

Ensemble Kalman filter

- Replace $\mathbf{x}^t$ with $\overline{\mathbf{x}^a}$, the mean of an ensemble of forecasts $(\mathbf{x}_1^a, \mathbf{x}_2^a, \ldots, \mathbf{x}_i^a, \ldots, \mathbf{x}_N^a)$
- Replace $\mathbf{P}^f$ and $\mathbf{P}^a$ with $\mathbf{P}^f_e$ and $\mathbf{P}^a_e$ calculated from an ensemble of forecasts (error ~1/N)

$$\mathbf{P}^f = E\{(\mathbf{x}^f - \mathbf{x}^t)(\mathbf{x}^f - \mathbf{x}^t)^T\} \simeq \overline{(\mathbf{x}_i^f - \overline{\mathbf{x}_i^f})(\mathbf{x}_i^f - \overline{\mathbf{x}_i^f})^T} \equiv \mathbf{P}_e^f$$

$$\mathbf{P}^a = E\{(\mathbf{x}^a - \mathbf{x}^t)(\mathbf{x}^a - \mathbf{x}^t)^T\} \simeq \overline{(\mathbf{x}_i^a - \overline{\mathbf{x}_i^a})(\mathbf{x}_i^a - \overline{\mathbf{x}_i^a})^T} \equiv \mathbf{P}_e^a$$

where $\mathbf{x}_i^f(t_{k+1}) = M(\mathbf{x}_i^a(t_k)) + \boldsymbol{\eta}_i(t_k)$ with, e.g., $\boldsymbol{\eta}_i : N(\mathbf{0},\mathbf{Q})$

- We define $(\mathbf{P}^f\mathbf{H}^T)_e \equiv \overline{(\mathbf{x}_i^f - \overline{\mathbf{x}_i^f})(H(\mathbf{x}_i^f) - \overline{H(\mathbf{x}_i^f)})^T} \simeq \mathbf{P}^f\mathbf{H}^T$

$$(\mathbf{H}\mathbf{P}^f\mathbf{H}^T)_e \equiv \overline{(H(\mathbf{x}_i^f) - \overline{H(\mathbf{x}_i^f)})(H(\mathbf{x}_i^f) - \overline{H(\mathbf{x}_i^f)})^T} \simeq \mathbf{H}\mathbf{P}^f\mathbf{H}^T$$

$$\mathbf{x}_i^a = \mathbf{x}_i^f + (\mathbf{P}^f\mathbf{H}^T)_e((\mathbf{H}\mathbf{P}^f\mathbf{H}^T)_e + \mathbf{R})^{-1}(\mathbf{y}_i^o - H(\mathbf{x}_i^f))$$

with $\mathbf{y}_i^o = \mathbf{y}^o + \boldsymbol{\varepsilon}_i$

and, e.g., $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0},\mathbf{R})$

EnKF features

- Unlike the EKF, the EnKF makes use of the full nonlinear model M: more accurate determination of error growth (and of error saturation)
- The Kalman gain $\mathbf{K}$ is computed without the need to calculate $\mathbf{P}^f$. This is particularly advantageous when observations are processed serially
- No need to linearize the observation operator H
- Each ensemble member can be propagated forward in time independently: highly parallel algorithm
- Model error term can be included as a perturbation of the deterministic forecast

Covariance localization

- When forecast error covariance is misspecified (e.g., due to neglecting model error, or when $N << n$), it may include spurious correlations between very distant grid points
- A common solution is to multiply each $\mathbf{P}^f_e$ element by an appropriate weight that reduces long-distance correlations
- This ensures that only the components of $\mathbf{P}^f_e$ believed to represent the corresponding components of $\mathbf{P}^f$ accurately are retained

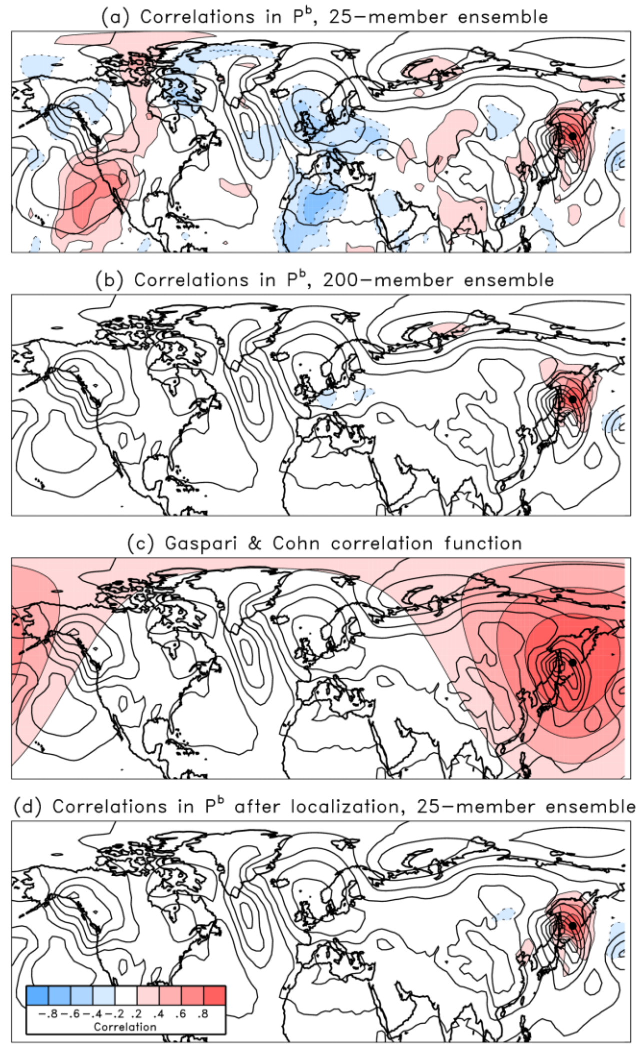
Covariance localization: an example

(a) $\mathbf{P}^{f}_{e}$ (N=25)

(b) $\mathbf{P}^{f}_{e}$ (N=100)

(c) Correlation function with compact support

(d) localized $\mathbf{P}^{f}_{e}$ (N=25)



From Fig. 6.4 of Hamill, 2006

Other issues

- In the extra-tropics at large scale the atmosphere is in near-geostrophic balance. An inappropropriate (e.g., too narrow) localization may give rise to unbalanced increments (e.g., between height gradient and wind increments)
- Analysis update equations are optimal only when errors are Gaussian (not necessarily the case)
- Unless model error is properly taken into account, forecast errors may be underestimated
- Computational expense is proportional to number of observations: problems when obs are “too many”

EnKF variants

- To estimate $\mathbf{P}^a_e$ correctly, the EnKF algorithm requires the observations to be treated as random variables: stochastic or fully-Monte Carlo algorithm
- An alternative strategy consists in requiring $\mathbf{P}^a_e$ to be consistent with the analysis error covariance prescribed by the Kalman filter: deterministic algorithms (nonuniqueness).
- An advantage of deterministic algorithms is to avoid misrepresentation of observation error (due to finite sampling) statistics
- Deterministic algorithms prescribe the expression of the matrix of the ensemble member perturbations (from the mean), that is a square-root (or, more precisely, Cholesky factor) of $\mathbf{P}^a_e$: also called square-root algorithms (e.g., EnSRF)

Conclusions

- Ensemble-based algorithms are imposing themselves as viable alternative to the more established variational techniques for NWP data assimilation
- Their easy method of implementation makes it easier to scientists (including PhD students!) to contribute to research in data assimilation
- They provide a natural framework to generate probabilistic forecasts, i.e. forecasts with a quantitative estimate of their errors