École Doctorale des Sciences de l'Environnement d'Île-de-France

Année Universitaire 2014-2015

Modélisation Numérique de l'Écoulement Atmosphérique et Assimilation de Données

Olivier Talagrand

Cours 5

27 Avril 2015

Best Linear Unbiased Estimate

State vector x, belonging to *state space* $\mathcal{S}$ ($\dim \mathcal{S} = n$), to be estimated.

Available data in the form of

- A '*background*' estimate (*e. g.* forecast from the past), belonging to *state space*, with dimension n

  $x^b = x + \zeta^b$

- An additional set of data (*e. g.* observations), belonging to *observation space*, with dimension p

  $y = Hx + \varepsilon$

  H is known linear *observation operator.*

Assume probability distribution is known for the couple (ζ^b, ε).

Assume $E(\zeta^b) = 0$, $E(\varepsilon) = 0$, $E(\zeta^b \varepsilon^T) = 0$ (not restrictive)

Set $E(\zeta^b \zeta^{bT}) = P^b$ (also often denoted B), $E(\varepsilon\varepsilon^T) = R$

Best Linear Unbiased Estimate (continuation 2)

Apply formulæ for Optimal Interpolation

$$\boldsymbol{x}^a = \boldsymbol{x}^b + P^b H^{\mathrm{T}} [HP^bH^{\mathrm{T}} + R]^{-1} (\boldsymbol{y} - H\boldsymbol{x}^b)$$
$$P^a = P^b - P^b H^{\mathrm{T}} [HP^bH^{\mathrm{T}} + R]^{-1} HP^b$$

x^a is the *Best Linear Unbiased Estimate* (*BLUE*) of x from x^b and y.

Equivalent set of formulæ

$$\boldsymbol{x}^a = \boldsymbol{x}^b + P^a H^{\mathrm{T}} R^{-1} (\boldsymbol{y} - H\boldsymbol{x}^b)$$
$$[P^a]^{-1} = [P^b]^{-1} + H^{\mathrm{T}} R^{-1} H$$

Matrix $K \equiv P^b H^{\mathrm{T}} [HP^bH^{\mathrm{T}} + R]^{-1} = P^a H^{\mathrm{T}} R^{-1}$ is *gain matrix*.

If probability distributions are *globally* gaussian, *BLUE* achieves bayesian estimation, in the sense that $P(\boldsymbol{x} \mid \boldsymbol{x}^b, \boldsymbol{y}) = \mathcal{N}[\boldsymbol{x}^a, \boldsymbol{P}^a]$.

Best Linear Unbiased Estimate (continuation 4)

Variational form of the *BLUE*

BLUE x^a minimizes following scalar *objective function*, defined on state space

$\xi \in \mathcal{S} \rightarrow$

- $$\mathcal{J}(\xi) \equiv (1/2)\,(x^b - \xi)^T\,[P^b]^{-1}\,(x^b - \xi) + (1/2)\,(y - H\xi)^T\,R^{-1}\,(y - H\xi)$$

$$\equiv \qquad \mathcal{J}_b \qquad + \qquad \mathcal{J}_o$$

'*3D-Var*'

Can easily, and heuristically, be extended to the case of a nonlinear observation operator H.

Used operationally in USA, Australia, China, …

Question. How to introduce temporal dimension in estimation process ?

- Logic of Optimal Interpolation and of *BLUE* can be extended to time dimension.

- But we know much more than just temporal correlations. We know explicit dynamics.

Real (unknown) state vector at time k (in format of assimilating model) x_k. Belongs to state space $\mathcal{S}$ (dim $\mathcal{S} = n$)

Evolution equation

$$x_{k+1} = M_k(x_k) + \eta_k$$

M_k is (known) model, η_k is (unknown) model error

Sequential Assimilation

- Assimilating model is integrated over period of time over which observations are available. Whenever model time reaches an instant at which observations are available, state predicted by the model is updated with new observations. In the jargon of the trade, *Optimal Interpolation* designates an algorithm for sequential assimilation in which the matrix P^b is constant with time, and *3D-Var* an algorithm in which, in addition, the analysis $\boldsymbol{x}^a$ is obtained through a variational algorithm.

Variational Assimilation

- Assimilating model is globally adjusted to observations distributed over observation period. Achieved by minimization of an appropriate scalar *objective function* measuring misfit between data and sequence of model states to be estimated.

Sequential Assimilation

Optimal Interpolation

- Observation vector at time k

$$y_k = H_k x_k + \varepsilon_k \qquad k = 0, \ldots, K$$

$$E(\varepsilon_k) = 0 \quad ; \quad E(\varepsilon_k \varepsilon_j^{\mathrm{T}}) = R_k\, \delta_{kj}$$

H_k linear

- Evolution equation

$$x_{k+1} = M_k(x_k) + \eta_k \qquad k = 0, \ldots, K\text{-}1$$

Optimal Interpolation (2)

At time k, background $x^b{}_k$ and associated error covariance matrix P^b known, assumed to be independent of k.

- Analysis step

$$x^a{}_k = x^b{}_k + P^b H_k{}^{\mathrm{T}} [H_k P^b H_k{}^{\mathrm{T}} + R_k]^{-1} (y_k - H_k x^b{}_k)$$

In *3D-Var*, $x^a{}_k$ is obtained by (iterative) minimization of associated objective function

- Forecast step

$$x^b{}_{k+1} = M_k(x^a{}_k)$$

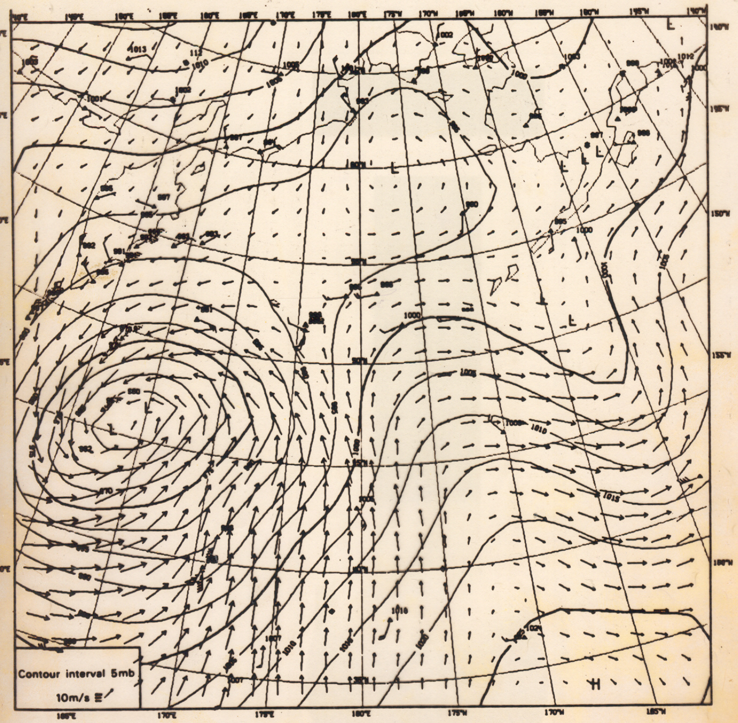


Fig. 4 Sea level pressure and wind forecast corresponding to the central area of Fig. 1, with plotted surface observations of pressure and wind (each fleche = 5 m/s).

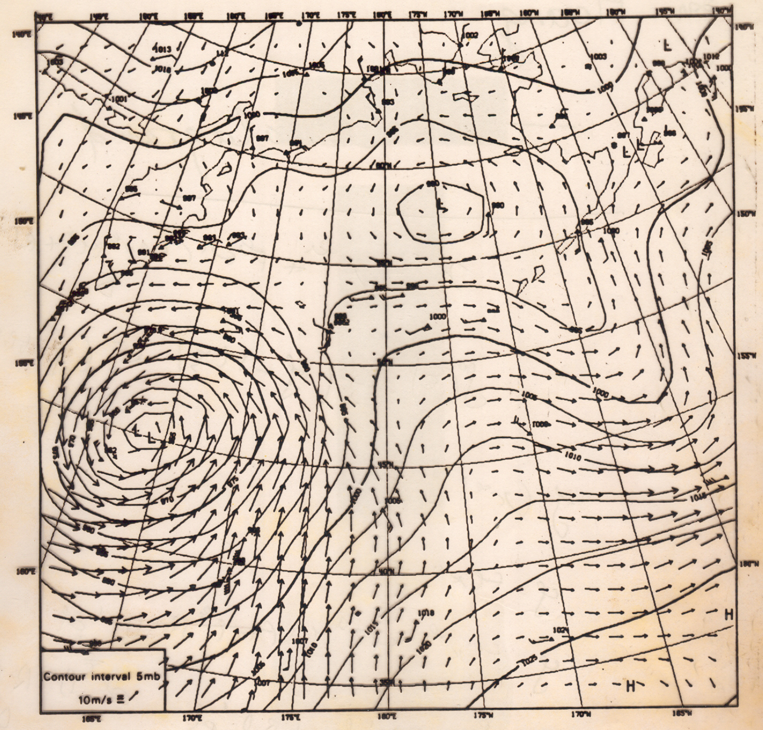


Fig. 5 As Fig. 4 for the analysis in the data assimilation cycle

After A. Lorenc

Sequential Assimilation. *Kalman Filter*

- Observation vector at time k

$$y_k = H_k x_k + \varepsilon_k \qquad\qquad k = 0, \ldots, K$$

$$E(\varepsilon_k) = 0 \quad ; \quad E(\varepsilon_k \varepsilon_j^{\mathrm{T}}) = R_k\, \delta_{kj}$$

H_k linear

- Evolution equation

$$x_{k+1} = M_k x_k + \eta_k \qquad\qquad k = 0, \ldots, K\text{-}1$$

$$E(\eta_k) = 0 \quad ; \quad E(\eta_k \eta_j^{\mathrm{T}}) = Q_k\, \delta_{kj}$$

M_k linear

- $E(\eta_k \varepsilon_j^{\mathrm{T}}) = 0$ (errors uncorrelated in time)

At time k, background $x^b{}_k$ and associated error covariance matrix $P^b{}_k$ known

- Analysis step

$$x^a{}_k = x^b{}_k + P^b{}_k H_k{}^T [H_k P^b{}_k H_k{}^T + R_k]^{-1} (y_k - H_k x^b{}_k)$$
$$P^a{}_k = P^b{}_k - P^b{}_k H_k{}^T [H_k P^b{}_k H_k{}^T + R_k]^{-1} H_k P^b{}_k$$

- Forecast step

$$x^b{}_{k+1} = M_k x^a{}_k$$
$$P^b{}_{k+1} = E[(x^b{}_{k+1} - x_{k+1})(x^b{}_{k+1} - x_{k+1})^T] = E[(M_k x^a{}_k - M_k x_k - \eta_k)(M_k x^a{}_k - M_k x_k - \eta_k)^T]$$
$$= M_k E[(x^a{}_k - x_k)(x^a{}_k - x_k)^T] M_k{}^T - E[\eta_k (x^a{}_k - x_k)^T] - E[(x^a{}_k - x_k)\eta_k{}^T] + E[\eta_k \eta_k{}^T]$$
$$= M_k P^a{}_k M_k{}^T + Q_k$$

At time k, background $x^b{}_k$ and associated error covariance matrix $P^b{}_k$ known

- Analysis step

$$x^a{}_k = x^b{}_k + P^b{}_k H_k{}^{\mathrm{T}} [H_k P^b{}_k H_k{}^{\mathrm{T}} + R_k]^{-1} (y_k - H_k x^b{}_k)$$
$$P^a{}_k = P^b{}_k - P^b{}_k H_k{}^{\mathrm{T}} [H_k P^b{}_k H_k{}^{\mathrm{T}} + R_k]^{-1} H_k P^b{}_k$$

- Forecast step

$$x^b{}_{k+1} = M_k x^a{}_k$$
$$P^b{}_{k+1} = M_k P^a{}_k M_k{}^{\mathrm{T}} + Q_k$$

Kalman filter (KF, Kalman, 1960)

Must be started from some initial estimate $(x^b{}_0, P^b{}_0)$

If all operators are linear, and if errors are uncorrelated in time, Kalman filter produces at time k the *BLUE* $x^b{}_k$ (resp. $x^a{}_k$) of the real state x_k from all data prior to (resp. up to) time k, plus the associated estimation error covariance matrix $P^b{}_k$ (resp. $P^a{}_k$).

If in addition errors are gaussian, the corresponding conditional probability distributions are the respective gaussian distributions

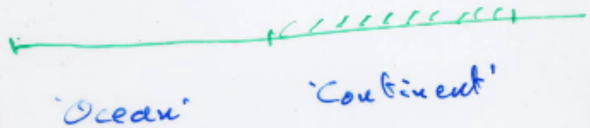
$\mathcal{N}[x^b{}_k, P^b{}_k]$ and $\mathcal{N}[x^a{}_k, P^a{}_k]$.

A didactic example (Ghil et al.)

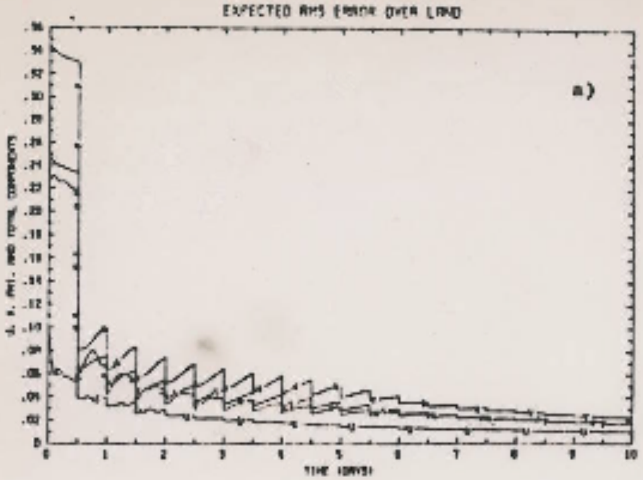
Barotropic model

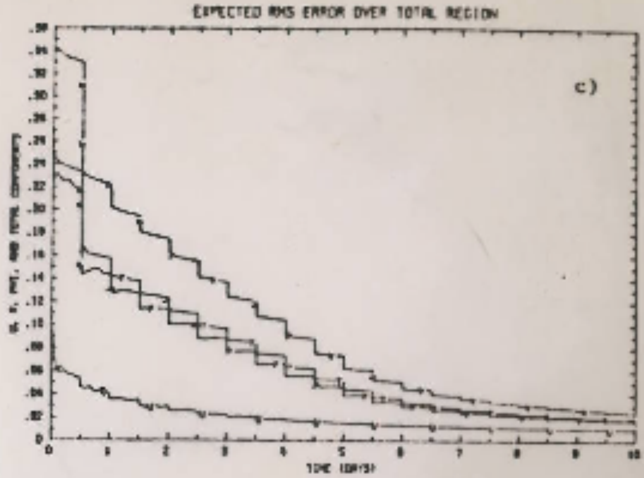
$$\begin{cases} \dfrac{\partial \varphi}{\partial t} + \operatorname{div}(\varphi \, \underline{U}) = 0 \\[2ex] \dfrac{\partial \underline{U}}{\partial t} + \underline{\operatorname{grad}}(\varphi + \frac{1}{2} U^2) + \underline{k} \times (f + \xi) \, \underline{U} = 0 \end{cases}$$

One dimension, periodic



Linearized (conserves energy)





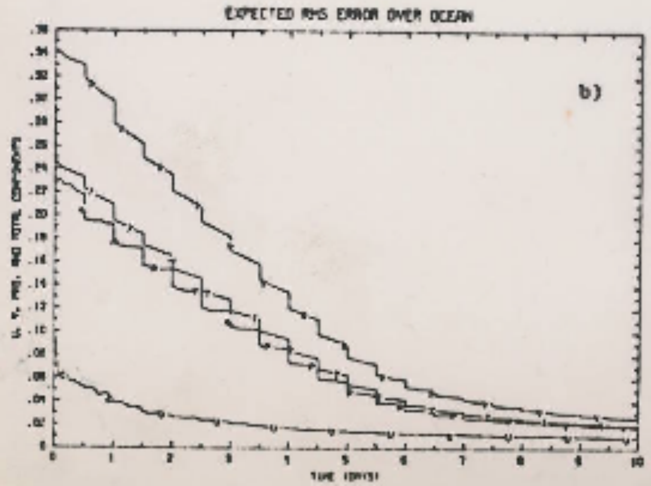


Fig. 2

The components of the total expected rms error (Erms), (trace: $P_k)^{1/2}$, in the estimation of solutions to the stochastic-dynamic system (Ψ, H), with Ψ given by (3.6) and H = (I 0). System noise is absent, Q = 0. The filter used is the standard K-B filter (2.11) for the model.

a) Erms over land; b) Erms over the ocean; c) Erms over the entire L-domain

In each one of the figures, each curve represents one component of the total Erms error. The curves labelled U, V, and P represent the u component, v component and ϕ component, respectively. They are found by summing the diagonal elements of P_k which correspond to u, v, and ϕ, respectively, dividing by the number of terms in the sum, and then taking the square root. In a) the summation extends over land points only, in b) over ocean points only, and in c) over the entire L-domain. The vertical axis is scaled in such a way that 1.0 corresponds to an Erms error of v_{max} for the U and V curves, and of ϕ_0 for the P curve. The observational error level is 0.089 for the U and V curves, and 0.080 for the P curve. The curves labelled T represent the total Erms error over each region. Each T curve is a weighted average of the corresponding U, V, and P curves, with the weights chosen in such a way that the T curve measures the error in the total energy $u^2 + v^2 + \phi^2/\bar{\phi}$, conserved by the system (3.1). The observational noise level for the T curve is then 0.088. Notice the immediate error decrease over land and the gradual decrease over the ocean. The total estimation error tends to zero.

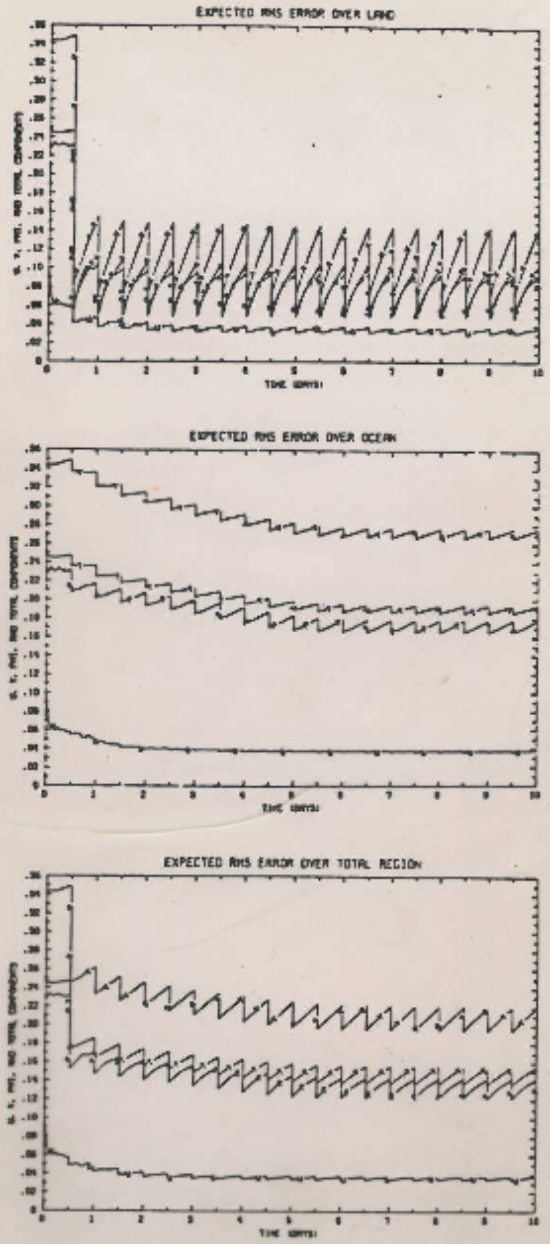


Fig. 6 This figure and the following ones show the properties of the estimated algorithm (2.11) in the presence of system noise, $Q \neq 0$. This figure gives the Erms estimation error, and is homologous to Fig. 2. Notice the sharper increase of error over land between synoptic times, and the convergence of each curve to a periodic, nonzero function.

Nonlinearities ?

Model is usually nonlinear, and observation operators (satellite observations) tend more and more to be nonlinear.

- Analysis step

$$x^a_k = x^b_k + P^b_k H_k'^{\mathrm{T}} [H_k' P^b_k H_k'^{\mathrm{T}} + R_k]^{-1} [y_k - H_k(x^b_k)]$$
$$P^a_k = P^b_k - P^b_k H_k'^{\mathrm{T}} [H_k' P^b_k H_k'^{\mathrm{T}} + R_k]^{-1} H_k' P^b_k$$

- Forecast step

$$x^b_{k+1} = M_k(x^a_k)$$
$$P^b_{k+1} = M_k' P^a_k M_k'^{\mathrm{T}} + Q_k$$

Extended Kalman Filter (EKF, heuristic !)

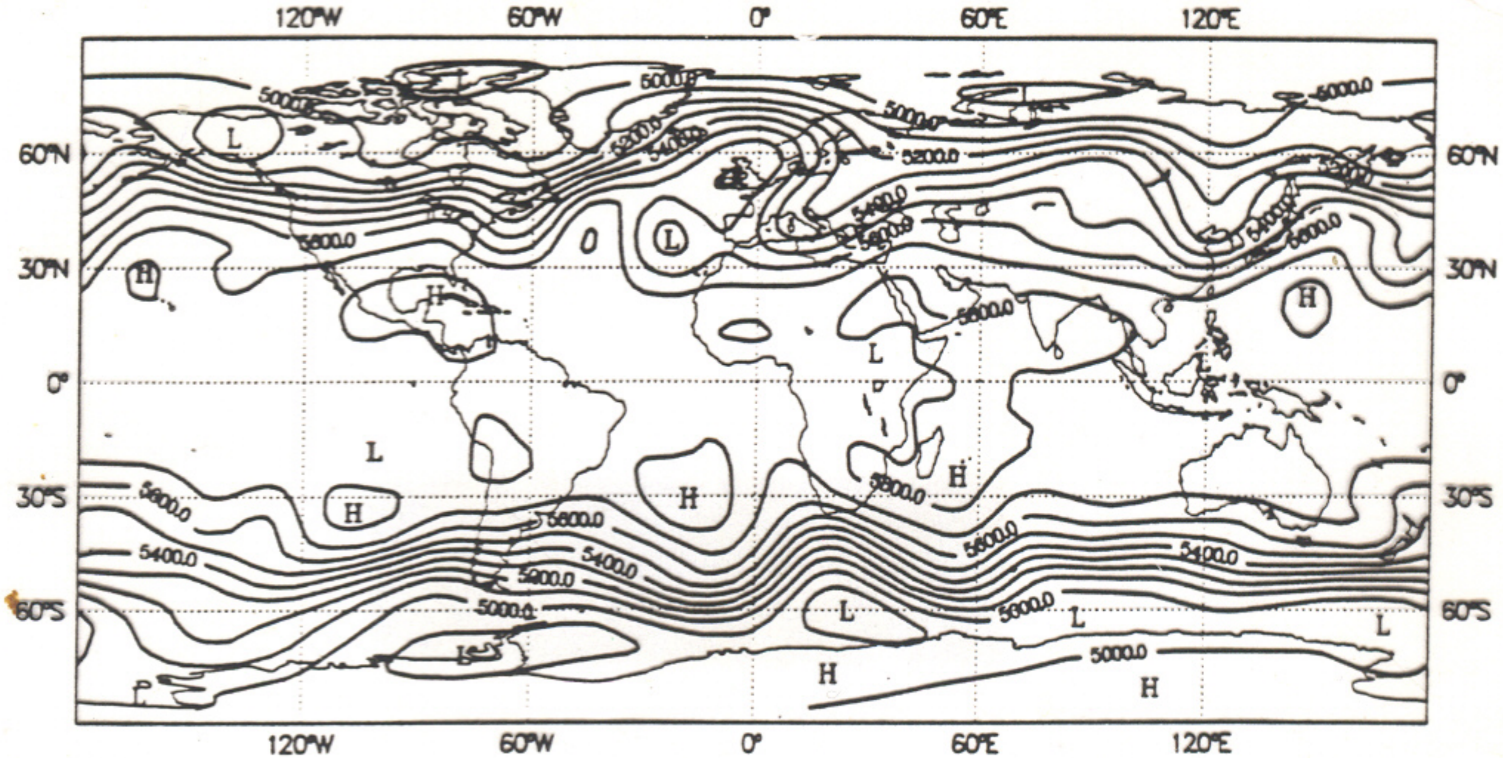
Costliest part of computation

$$P^b_{k+1} = M_k P^a_k M_k^{\mathrm{T}} + Q_k$$

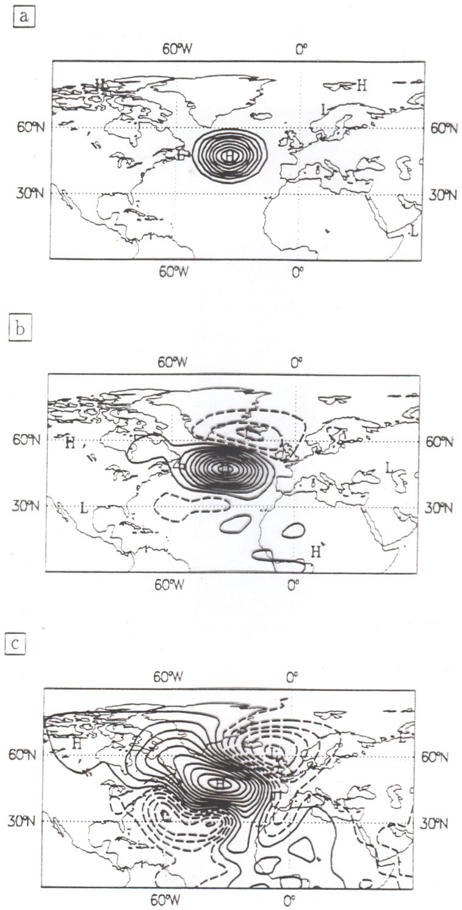
Multiplication by M_k = one integration of the model between times k and $k+1$.

Computation of $M_k P^a_k M_k^{\mathrm{T}} \approx 2n$ integrations of the model

Need for determining the temporal evolution of the uncertainty on the state of the system is the major difficulty in assimilation of meteorological and oceanographical observations



Analysis of 500-hPa geopotential for 1 December 1989, 00:00 UTC (ECMWF, spectral truncation T21, unit *m*. After F. Bouttier)



Temporal evolution of the 500-hPa geopotential autocorrelation with respect to point located at 45N, 35W. From top to bottom: initial time, 6- and 24-hour range. Contour interval 0.1. After F. Bouttier.

Two solutions :

- *Low-rank filters*

  Use low-rank covariance matrix, restricted to modes in state space on which it is known, or at least assumed, that a large part of the uncertainty is concentrated (this requires the definition of a norm on state space).

Reduced Rank Square Root Filters (*RRSQRT*, Heemink)

Singular Evolutive Extended Kalman Filter (*SEEK*, Pham)

….

Second solution :

- *Ensemble filters*

  Uncertainty is represented, not by a covariance matrix, but by an ensemble of point estimates in state space that are meant to sample the conditional probability distribution for the state of the system (dimension $L \approx O(10\text{-}100)$).

  Ensemble is evolved in time through the full model, which eliminates any need for linear hypothesis as to the temporal evolution.

  Ensemble Kalman Filter (*EnKF*, Evensen, Anderson, …)

How to update predicted ensemble with new observations ?

Predicted ensemble at time k : $\{x^b{}_l\}$, $\quad l = 1, \ldots, L$

Observation vector at same time : $y = Hx + \varepsilon$

- Gaussian approach

Produce sample of probability distribution for real observed quantity Hx

$y_l = y - \varepsilon_l$

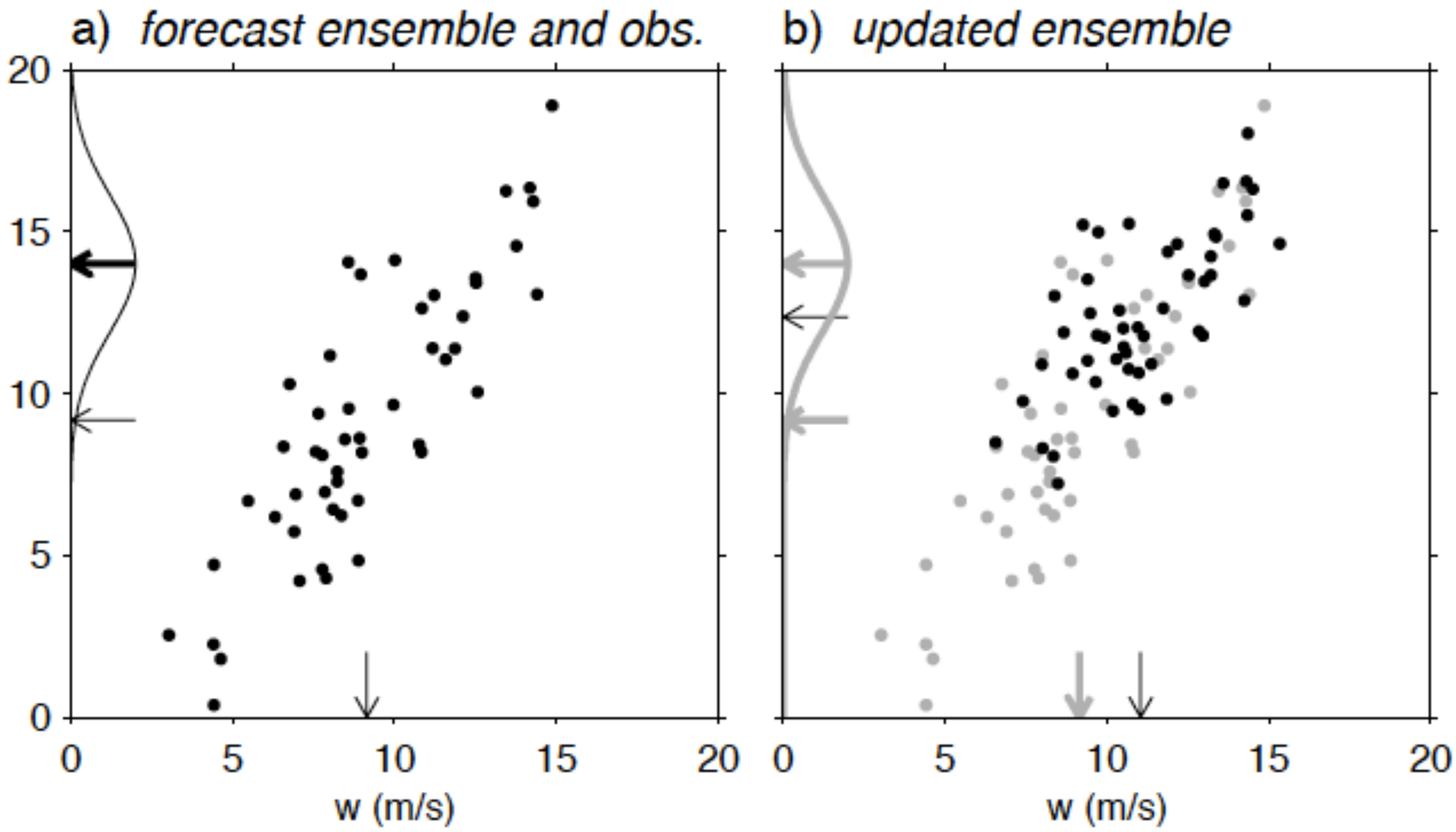
where ε_l is distributed according to probability distribution for observation error ε.

Then use Kalman formula to produce sample of 'analysed' states

$$x^a{}_l = x^b{}_l + P^b H^T [HP^bH^T + R]^{-1} (y_l - Hx^b{}_l) , \qquad l = 1, \ldots, L \qquad (2)$$

where P^b is the sample covariance matrix of predicted ensemble $\{x^b{}_l\}$.

Remark. In case of Gaussian errors, if P^b was exact covariance matrix of background error, (2) would achieve Bayesian estimation, in the sense that $\{x^a{}_l\}$ would be a sample of conditional probability distribution for x, given all data up to time k.

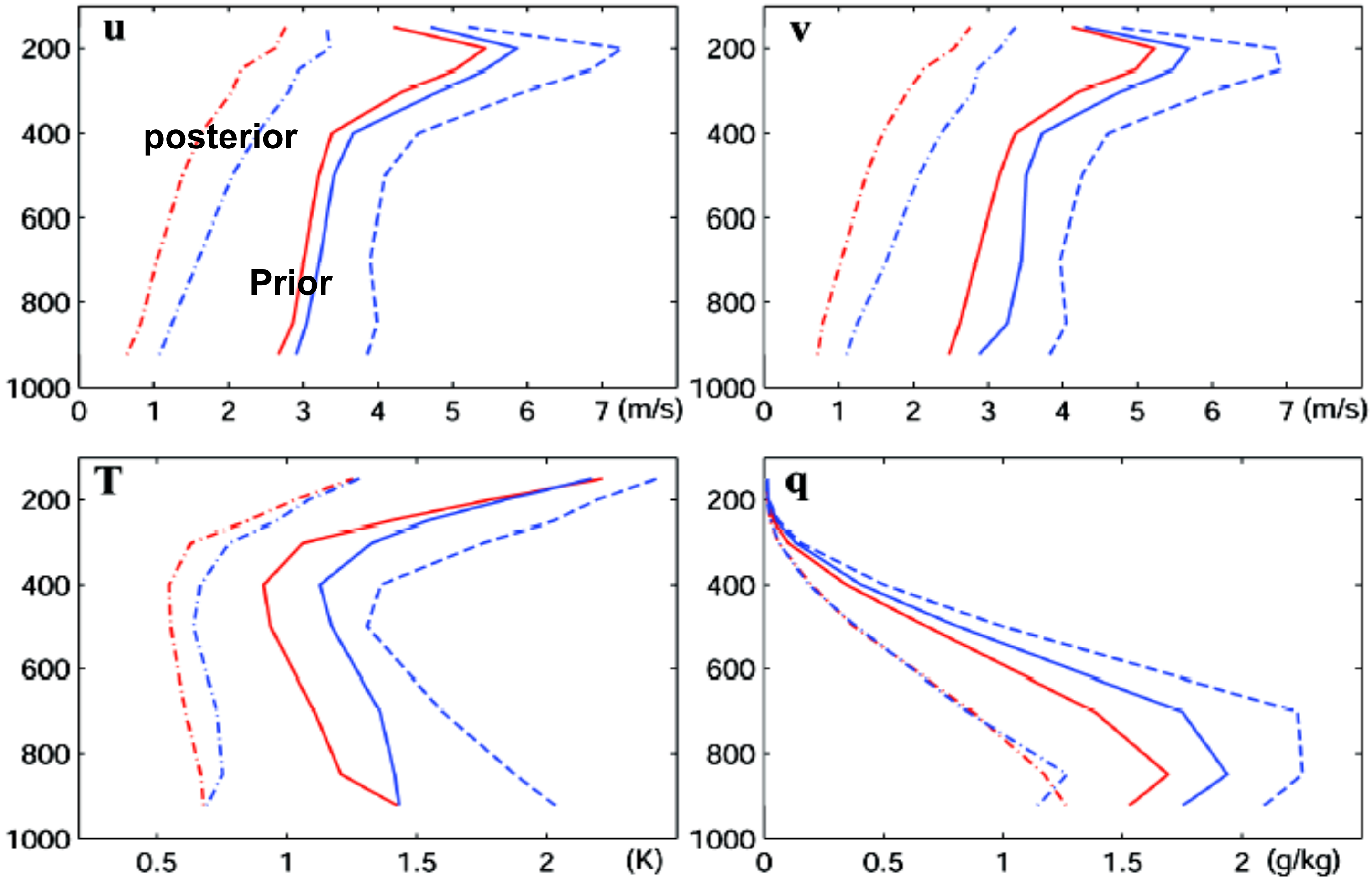


a) forecast ensemble and obs.
b) updated ensemble
20
15
10
5
0
0
5
10
15
20
w (m/s)
w (m/s)

C. Snyder

Month-long Performance of EnKF vs. 3Dvar with WRF

— EnKF — 3DVar (prior, solid; posterior, dotted)



Better performance of EnKF than 3DVar also seen in both 12-h forecast and posterior analysis in terms of root-mean square difference averaged over the entire month

(Meng and Zhang 2007c, MWR, in review)