Introduction to Data Assimilation

Olivier Talagrand

Summer School
Advanced Mathematical Methods
to Study Atmospheric Dynamical Processes and Predictability

Banff International Research Station
for Mathematical Innovation and Discovery

Banff, Canada
13 July 2011

Fig. 1: Members of day 7 forecast of 500 hPa geopotential height for the ensemble originated from 25 January 1993.

Figure 6 Hurricane Katrina mean-sea-level-pressure (MSLP) analysis for 12 UTC of 29 August 2005 and t+84h high-resolution and EPS forecasts started at 00 UTC of 26 August:

1st row: 1st panel: MSLP analysis for 12 UTC of 29 Aug
2nd panel: MSLP t+84h T_L511L60 forecast started at 00 UTC of 26 Aug
3rd panel: MSLP t+84h EPS-control T_L255L40 forecast started at 00 UTC of 26 Aug
Other rows: 50 EPS-perturbed T_L255L40 forecast started at 00 UTC of 26 Aug.

The contour interval is 5 hPa, with shading patters for MSLP values lower than 990 hPa.

Assimilation of observations, as it is known in meteorology, originated from the need of defining initial conditions (ICs) for numerical weather prediction. Difficulties progressively arose

- Need for defining ICs with appropriate spatial scales ⇒ '*structure functions*' (now incorporated in background error covariance matrices)

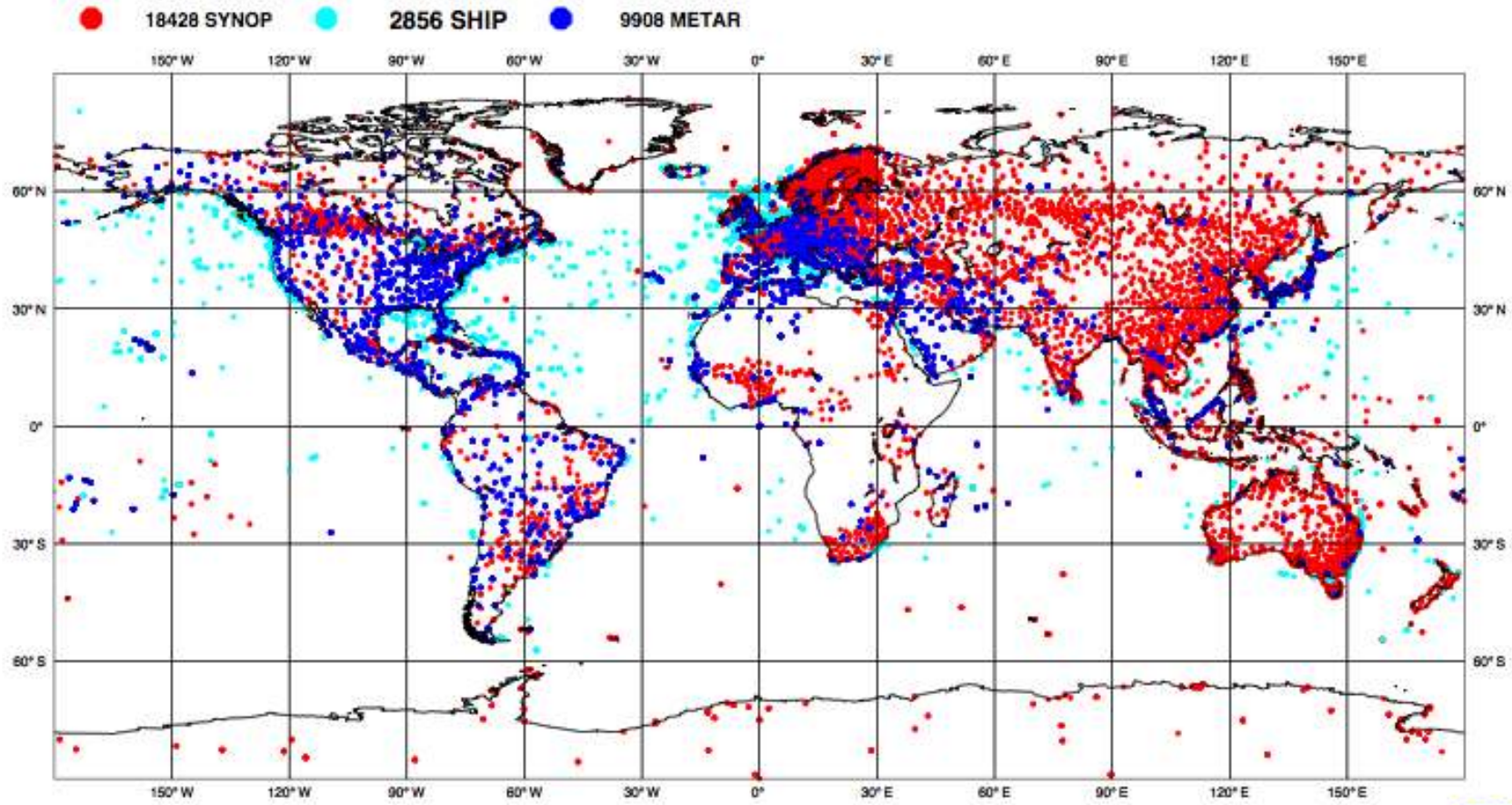
- Need for defining ICs in approximate geostrophic balance ⇒ '*initialization*' (now also incorporated, at least partially, in background error covariance matrices; lecture 2 by P. Lynch)

- Realization that meteorological forecasts are very sensitive to initial conditions (Lorenz, 1963).

- Realization that useful information was present in recent forecast ⇒ use of a background, to be defined with associated uncertainty (word *assimilation* was coined in 1967-68)

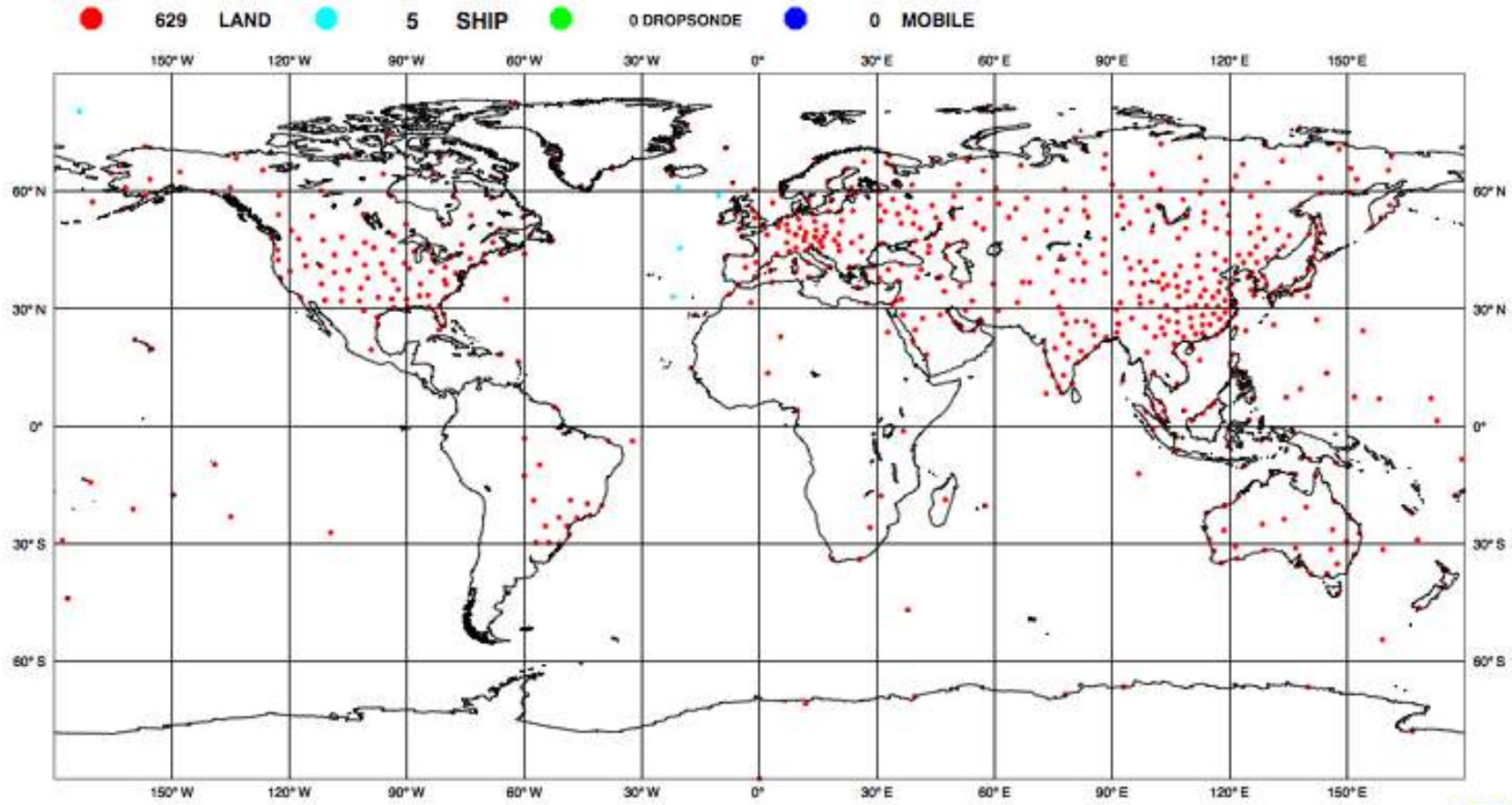
- Use of satellite observations, which are
  - distributed continuously in time
  - indirect ⇒ need for some form of 'inversion'

ECMWF Data Coverage (All obs DA) - SYNOP/SHIP
23/APR/2011; 00 UTC
Total number of obs = 31192



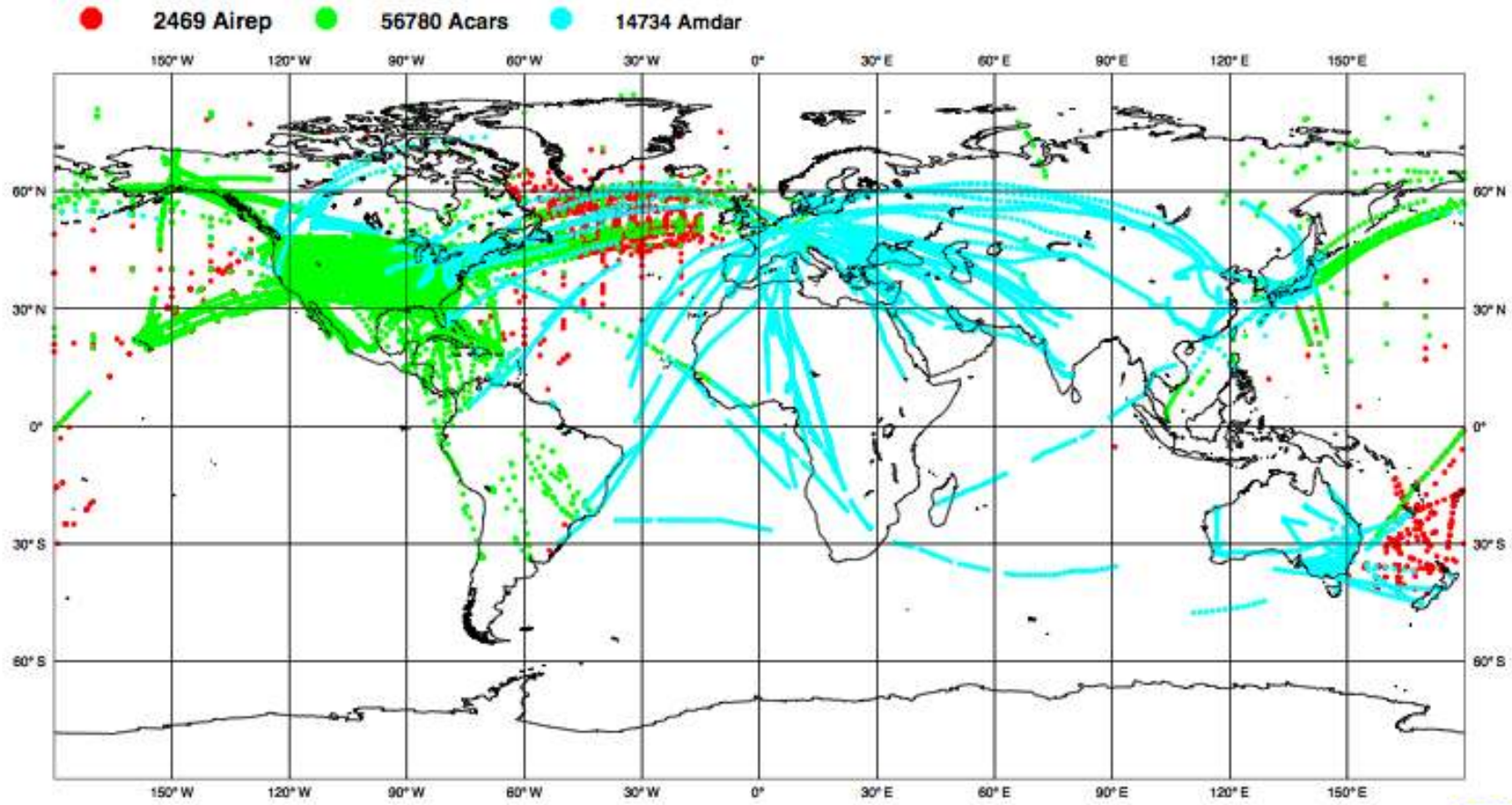
ECMWF

ECMWF Data Coverage (All obs DA) - TEMP
23/APR/2011; 00 UTC
Total number of obs = 634



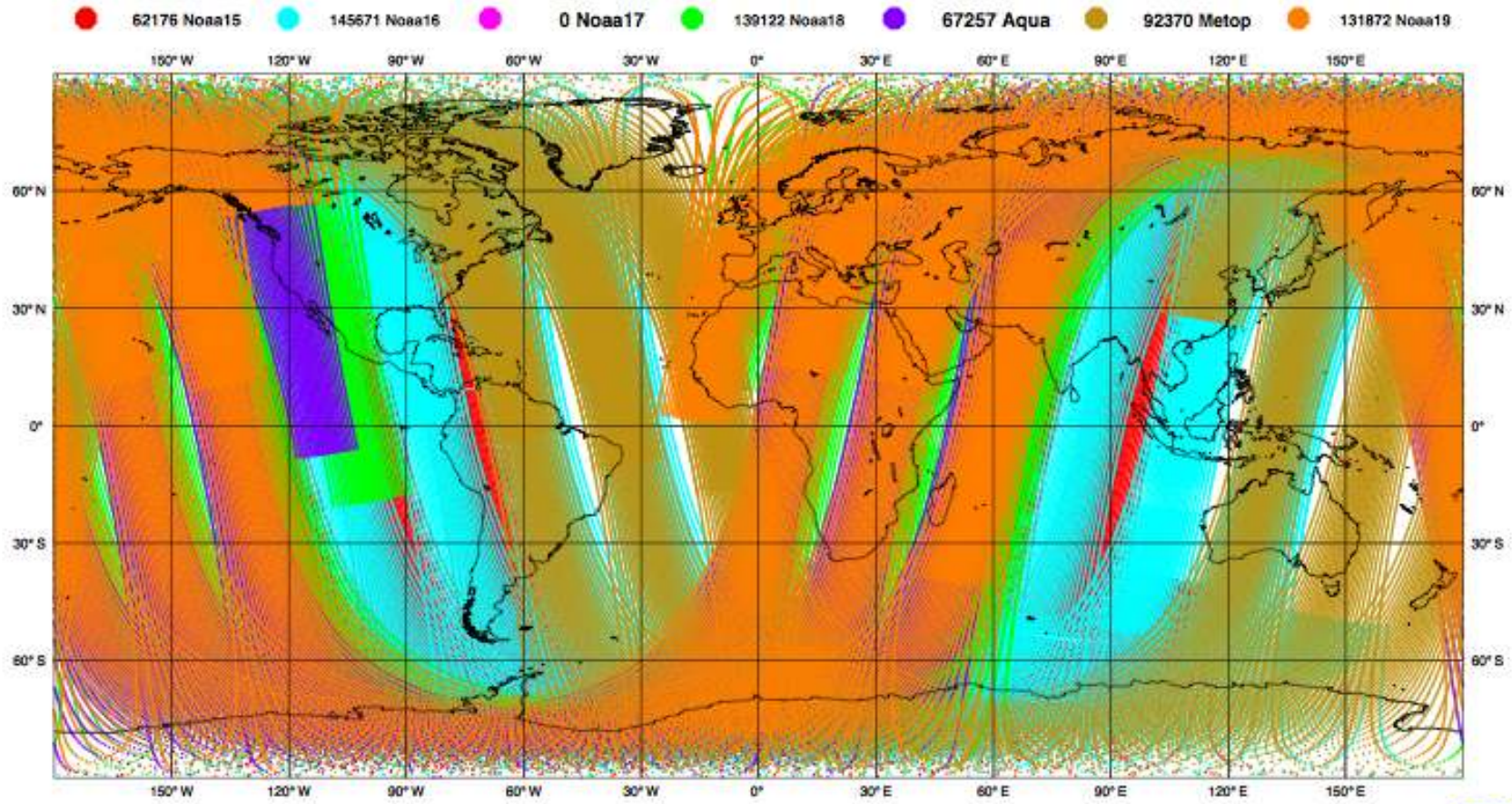
ECMWF

ECMWF Data Coverage (All obs DA) - AIRCRAFT
23/APR/2011; 00 UTC
Total number of obs = 73983



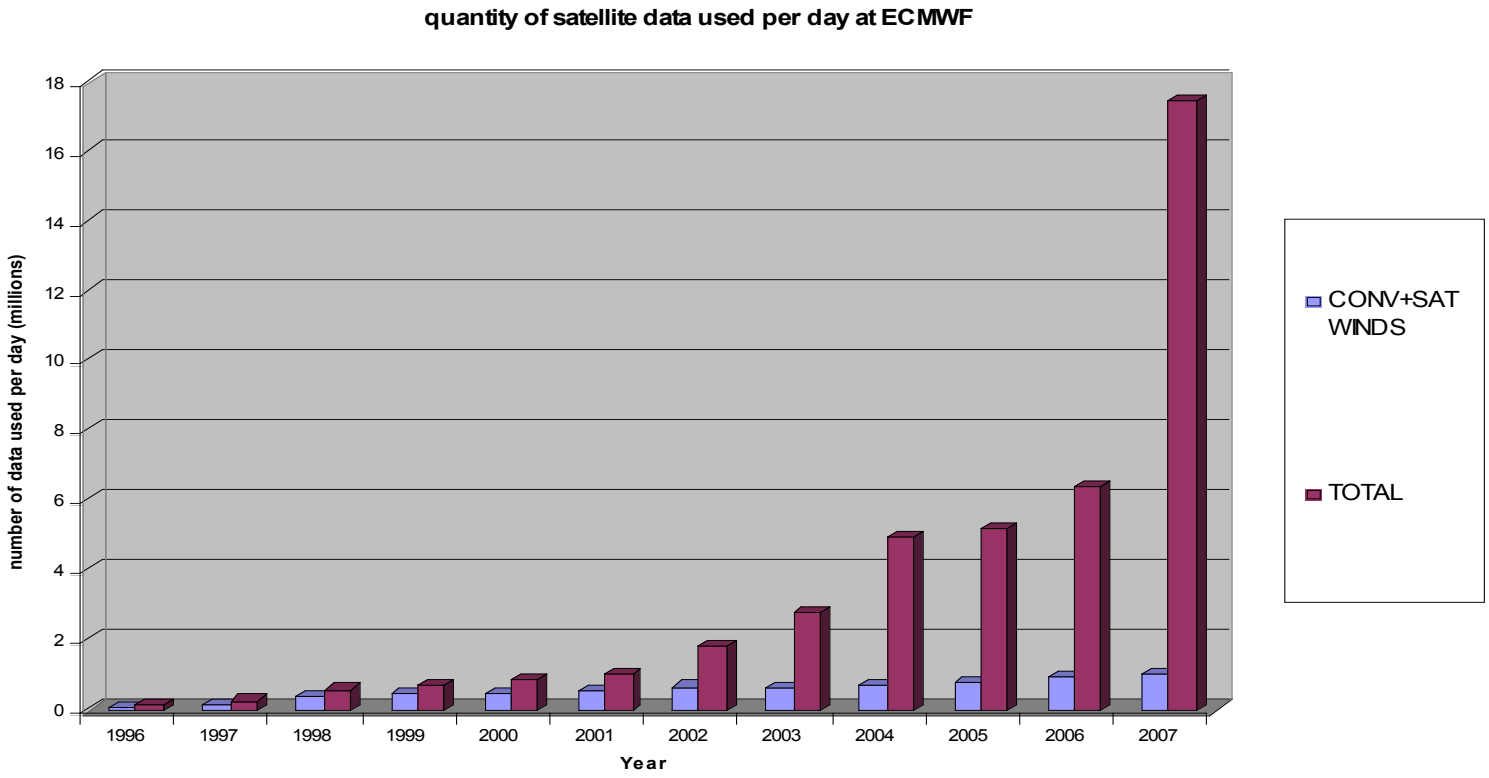
ECMWF

ECMWF Data Coverage (All obs DA) - AMSU-A
23/APR/2011; 00 UTC
Total number of obs = 638468



ECMWF

December 2007: Satellite data volumes used: around 18 millions per day



Value as of March 2010 : 25 millions per day

S. Louvel, Doctoral Dissertation, 1999

Physical laws governing the flow

- Conservation of mass

  $D\rho/Dt + \rho \, \mathrm{div}\underline{U} = 0$

- Conservation of energy

  $De/Dt - (p/\rho^2) \, D\rho/Dt = Q$

- Conservation of momentum

  $D\underline{U}/Dt + (1/\rho) \, \underline{\mathrm{grad}}p - \underline{g} + 2 \, \underline{\Omega} \wedge \underline{U} = \underline{F}$

- Equation of state

  $f(p, \rho, e) = 0$ $\qquad\qquad$ $(p/\rho = rT, e = C_v T)$

- Conservation of mass of secondary components (water in the atmosphere, salt in the ocean, chemical species, …)

  $Dq/Dt + q \, \mathrm{div}\underline{U} = S$

Physical laws available in practice in the form of a discretized (and necessarily imperfect) numerical model

European Centre for Medium-range Weather Forecasts

(ECMWF, Reading, UK)

Horizontal spherical harmonics triangular truncation T1279 (horizontal resolution ≈ 16 kilometres)

91 levels on the vertical (0 - 80 km)

Dimension of state vector $n \approx 1.5\ 10^9$

Timestep = 10 minutes

Purpose of assimilation : reconstruct as accurately as possible the state of the atmospheric or oceanic flow, using all available appropriate information. The latter essentially consists of

- The observations proper, which vary in nature, resolution and accuracy, and are distributed more or less regularly in space and time.

- The physical laws governing the evolution of the flow, available in practice in the form of a discretized, and necessarily approximate, numerical model.

- 'Asymptotic' properties of the flow, such as, *e*. *g*., geostrophic balance of middle latitudes. Although they basically are necessary consequences of the physical laws which govern the flow, these properties can usefully be explicitly introduced in the assimilation process.

Assimilation is one of many '*inverse problems*' encountered in many fields of science and technology

- solid Earth geophysics
- plasma physics
- 'nondestructive' probing
- navigation (spacecraft, aircraft, ….)
- …

Solution most often (if not always) based on bayesian, or probabilistic, estimation. 'Equations' are fundamentally the same.

Difficulties specific to assimilation of meteorological and oceanographical observations :

- Very large numerical dimensions ($n \approx 10^7$-10^9 parameters to be estimated, $p \approx 2.10^7$ observations per 24-hour period). Difficulty aggravated in Numerical Weather Prediction by the need for the forecast to be ready in time.

- Non-trivial underlying dynamics.

Both observations and ‘model’ are affected with some uncertainty ⇒ uncertainty on the estimate.

For some reason, uncertainty is conveniently described by probability distributions (don’t know too well why, but it works) (lecture by C. Bishop to-night)

Assimilation is a problem in bayesian estimation.

Determine the conditional probability distribution for the state of the system, knowing everything we know (unambiguously defined if a prior probability distribution is defined; see Tarantola, 2005).

Bayesian Estimation

Determine conditional probability distribution of the state of the system, given the probability distribution of the uncertainty on the data

$z_1 = x + \zeta_1$ $\qquad$ $\zeta_1 = \mathcal{N}[0, s_1]$

density function $p_1(\zeta) \propto \exp[- (\zeta^2)/2s_1]$

$z_2 = x + \zeta_2$ $\qquad$ $\zeta_2 = \mathcal{N}[0, s_2]$

density function $p_2(\zeta) \propto \exp[- (\zeta^2)/2s_2]$

ζ_1 and ζ_2 mutually independent

What is the conditional probability $P(x = \xi \mid z_1, z_2)$ that x be equal to some value ξ ?

$z_1 = x + \zeta_1$ density function $p_1(\zeta) \propto \exp[- (\zeta^2)/2s_1^2]$

$z_2 = x + \zeta_2$ density function $p_2(\zeta) \propto \exp[- (\zeta^2)/2s_2^2]$

$x = \xi \;\Leftrightarrow\; \zeta_1 = z_1 - \xi$ and $\zeta_2 = z_2 - \xi$

$$P(x = \xi \mid z_1, z_2) \propto p_1(z_1 - \xi)\, p_2(z_2 - \xi)$$
$$\propto \exp[- (\xi - x^a)^2/2s]$$

where $1/s = 1/s_1 + 1/s_2$, $x^a = s\,(z_1/s_1 + z_2/s_2)$

Conditional probability distribution of x, given z_1 and z_2 : $\mathcal{N}([x^a, s]$

$s < (s_1, s_2)$ independent of z_1 and z_2

$z_1 = x + \xi_1$
$z_2 = x + \xi_2$

Same as before, but ξ_1 and ξ_2 are now distributed according to exponential law with parameter a, *i. e.*

$$p(\xi) \propto \exp[-|\xi|/a] \quad ; \quad \mathrm{Var}(\xi) = 2a^2$$

Conditional probability density function is now uniform over interval $[z_1, z_2]$, exponential with parameter $a/2$ outside that interval

$E(x \mid z_1, z_2) = (z_1+z_2)/2$

$\mathrm{Var}(x \mid z_1, z_2) = a^2 (2\delta^3/3 + \delta^2 + \delta + 1/2) / (1 + 2\delta)$, with $\delta = |z_1 - z_2|/(2a)$
Increases from $a^2/2$ to ∞ as δ increases from 0 to ∞. Can be larger than variance $2a^2$ of original errors (probability 0.08)

(Entropy $-\int p \ln p$ always decreases in bayesian estimation)

Bayesian estimation

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

Data vector z, belonging to *data space* $\mathcal{D}$ ($\dim \mathcal{D} = m$), available.

$$z = F(x, \zeta) \qquad (1)$$

where ζ is a random element representing the uncertainty on the data (or, more precisely, on the link between the data and the unknown state vector).

For example

$$z = \Gamma x + \zeta$$

Probability that $x = \xi$ for given ξ ?

$$x = \xi \quad \Rightarrow \quad z = F(\xi, \zeta)$$

$$P(x = \xi \mid z) = P[z = F(\xi, \zeta)] \,/\, \int_{\xi'} P[z = F(\xi', \zeta)]$$

Unambiguously defined iff, for any ζ, there is at most one x such that (1) is verified.

$\Leftrightarrow$ data contain information, either directly or indirectly, on any component of x. *Determinacy* condition.

Bayesian estimation is however impossible in its general theoretical form in meteorological or oceanographical practice because

- It is impossible to explicitly describe a probability distribution in a space with dimension even as low as $n \approx 10^3$, not to speak of the dimension $n \approx 10^{7\text{-}9}$ of present Numerical Weather Prediction models.

- Probability distribution of errors on data very poorly known (model errors in particular).

One has to restrict oneself to a much more modest goal. Two approaches exist at present

- Obtain some 'central' estimate of the conditional probability distribution (expectation, mode, …), plus some estimate of the corresponding spread (standard deviations and a number of correlations).

- Produce an ensemble of estimates which are meant to sample the conditional probability distribution (dimension $N \approx O(10\text{-}100)$).

Proportion of resources devoted to assimilation in Numerical Weather Prediction has steadily increased over time.

At present at ECMWF, the cost of 24 hours of assimilation is half the global cost of the 10-day forecast *(i. e.*, including the ensemble forecast).

Random vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T = (x_i)$ (*e. g.* pressure, temperature, abundance of given chemical compound at n grid-points of a numerical model)

- Expectation $E(\boldsymbol{x}) \equiv [E(x_i)]$; centred vector $\boldsymbol{x}' \equiv \boldsymbol{x} - E(\boldsymbol{x})$

- Covariance matrix

$$E(\boldsymbol{x}'\boldsymbol{x}'^T) = [E(x_i'x_j')]$$

dimension nxn, symmetric non-negative (strictly definite positive except if linear relationship holds between the x_i'' s with probability 1).

- Two random vectors

$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$

$\boldsymbol{y} = (y_1, y_2, \ldots, y_p)^T$

$$E(\boldsymbol{x}'\boldsymbol{y}'^T) = E(x_i'y_j')$$

dimension nxp

Random function $\varphi(\xi)$ (field of pressure, temperature, abundance of given chemical compound, … ; ξ is now spatial and/or temporal coordinate)

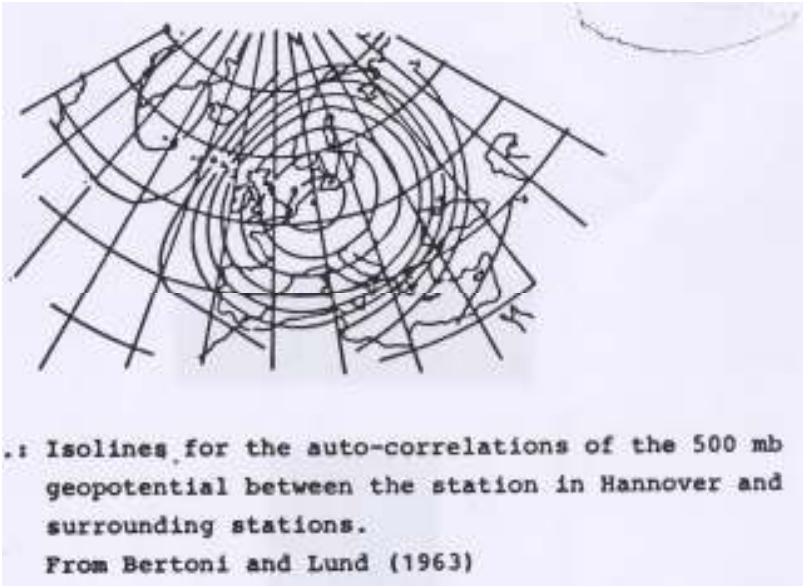
- Expectation $E[\varphi(\xi)]$; $\varphi'(\xi) \equiv \varphi(\xi) - E[\varphi(\xi)]$
- Variance $Var[\varphi(\xi)] = E\{[\varphi'(\xi)]^2\}$

- Covariance function

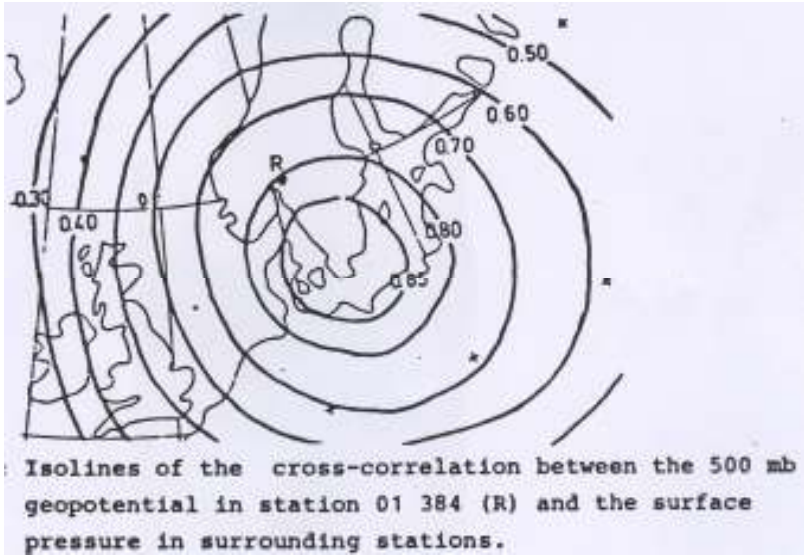
$$(\xi_1, \xi_2) \rightarrow C_\varphi(\xi_1, \xi_2) \equiv E[\varphi'(\xi_1)\, \varphi'(\xi_2)]$$

- Correlation function

$$Cor_\varphi(\xi_1, \xi_2) \equiv E[\varphi'(\xi_1)\, \varphi'(\xi_2)] \,/\, \{Var[\varphi(\xi_1)]\, Var[\varphi(\xi_2)]\}^{1/2}$$



.: Isolines for the auto-correlations of the 500 mb geopotential between the station in Hannover and surrounding stations.
From Bertoni and Lund (1963)



Isolines of the cross-correlation between the 500 mb geopotential in station 01 384 (R) and the surface pressure in surrounding stations.

After N. Gustafsson

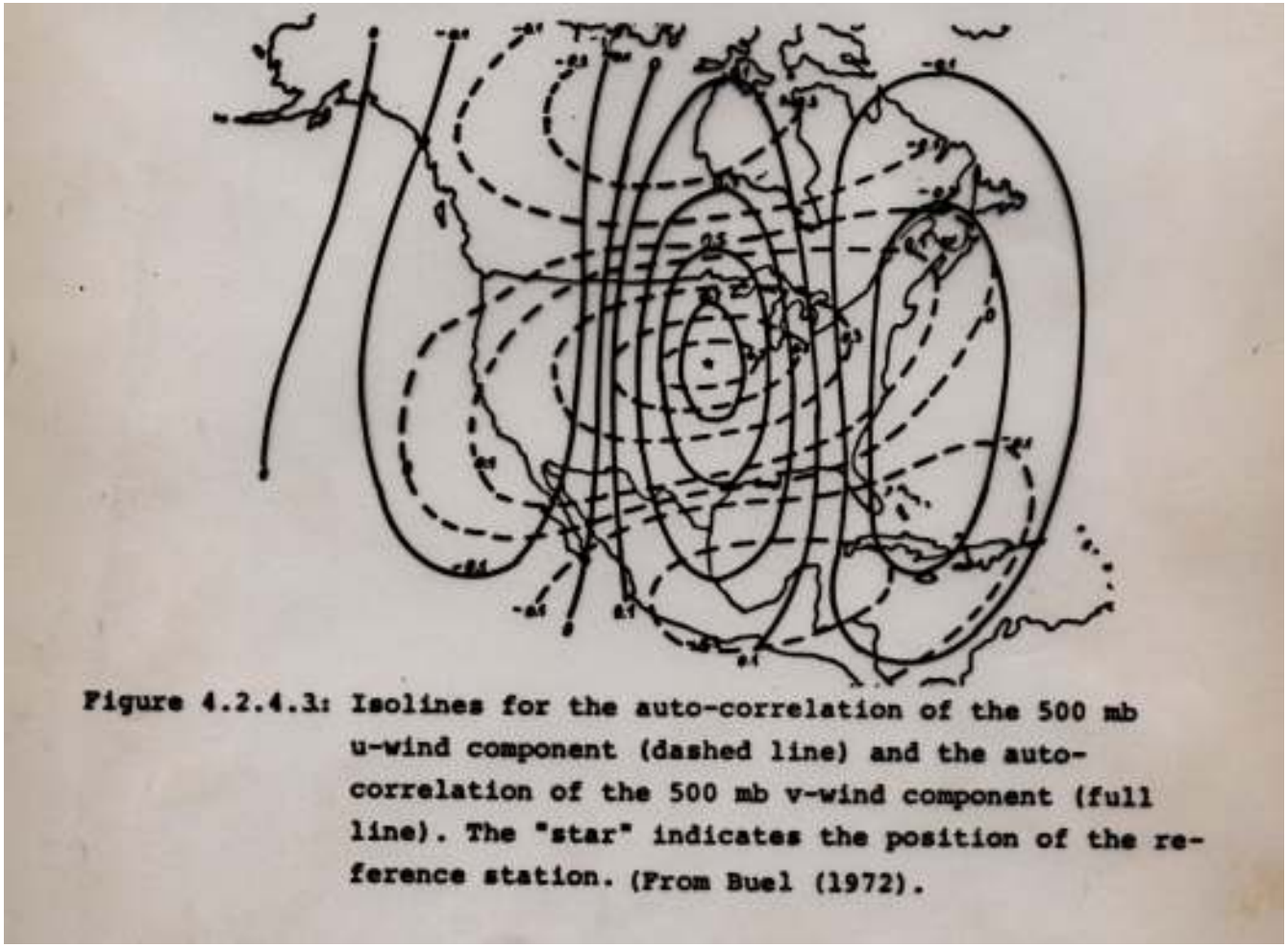


Figure 4.2.4.3: Isolines for the auto-correlation of the 500 mb u-wind component (dashed line) and the auto-correlation of the 500 mb v-wind component (full line). The "star" indicates the position of the reference station. (From Buel (1972).

After N. Gustafsson

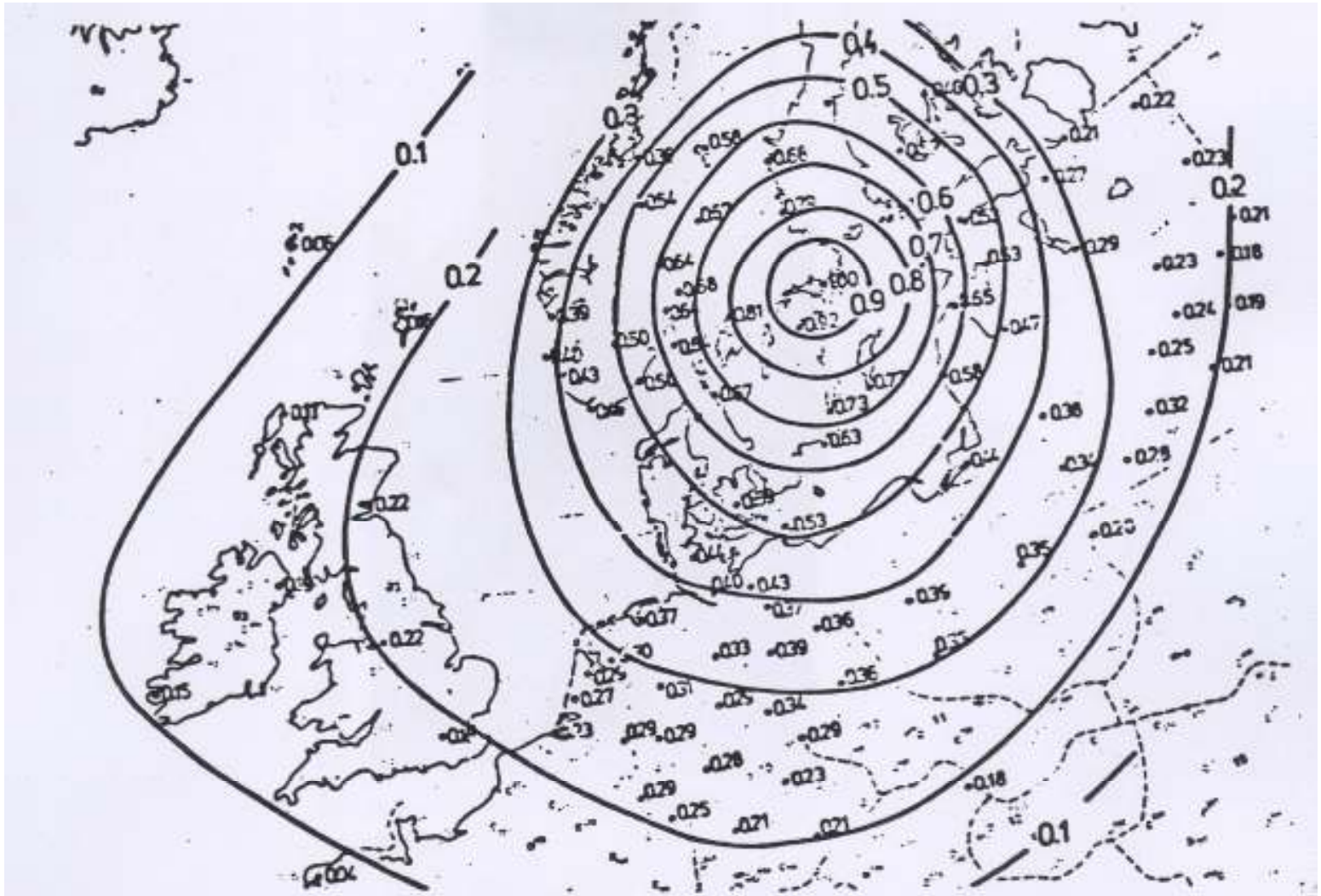


Figure 5.1.1.4.1 Auto-correlation of errors in 12h numerical forecasts of surface pressure in a reference station (Stockholm) and other stations.

After N. Gustafsson

Optimal Interpolation

Random field $\varphi(\xi)$

Observation network $\xi_1, \xi_2, \ldots, \xi_p$

For one particular realization of the field, observations

$$y_j = \varphi(\xi_j) + \varepsilon_j \quad , \; j = 1, \ldots, p \qquad , \qquad \text{making up vector } \boldsymbol{y} = (y_j)$$

Estimate $x = \varphi(\xi)$ at given point ξ, in the form

$$x^a = \alpha + \Sigma_j \, \beta_j \, y_j = \alpha + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{y} \qquad , \qquad \text{where } \boldsymbol{\beta} = (\beta_j)$$

α and the β_j's being determined so as to minimize the expected quadratic estimation error $E[(x\text{-}x^a)^2]$

Optimal Interpolation (continued 1)

Solution

$$x^a = E(x) + E(x'\boldsymbol{y}'^{\mathrm{T}})\,[E(\boldsymbol{y}'\boldsymbol{y}'^{\mathrm{T}})]^{-1}\,[\boldsymbol{y} - E(\boldsymbol{y})]$$

i. e.,

$$\boldsymbol{\beta} = [E(\boldsymbol{y}'\boldsymbol{y}'^{\mathrm{T}})]^{-1}\,E(x'\boldsymbol{y}')$$

$$\alpha = E(x) - \boldsymbol{\beta}^{\mathrm{T}}E(\boldsymbol{y})$$

Estimate is unbiased $E(x-x^a) = 0$

Minimized quadratic estimation error

$$E[(x-x^a)^2] = E(x'^2) - E(x'\boldsymbol{y}'^{\mathrm{T}})\,[E(\boldsymbol{y}'\boldsymbol{y}'^{\mathrm{T}})]^{-1}\,E(\boldsymbol{y}'x')$$

Estimation made in terms of deviations from expectations x' and $\boldsymbol{y}'$.

Optimal Interpolation (continued 2)

$$x^a = E(x) + E(x'\mathbf{y}'^{\mathrm{T}})\,[E(\mathbf{y}'\mathbf{y}'^{\mathrm{T}})]^{-1}\,[\mathbf{y} - E(\mathbf{y})]$$

$$y_j = \varphi(\xi_j) + \varepsilon_j$$

$$E(y_j'y_k') = E[(\varphi'(\xi_j) + \varepsilon_j')(\varphi'(\xi_k) + \varepsilon_k')]$$

If observation errors ε_j are mutually uncorrelated, have common variance r, and are uncorrelated with field φ, then

$$E(y_j'y_k') = C_\varphi(\xi_j, \xi_k) + r\delta_{jk}$$

and

$$E(x'y_j') = C_\varphi(\xi, \xi_j)$$

Optimal Interpolation (continued 3)

$$x^a = E(x) + E(x'\mathbf{y}'^{\mathrm{T}})\,[E(\mathbf{y}'\mathbf{y}'^{\mathrm{T}})]^{-1}\,[\mathbf{y} - E(\mathbf{y})]$$

Vector

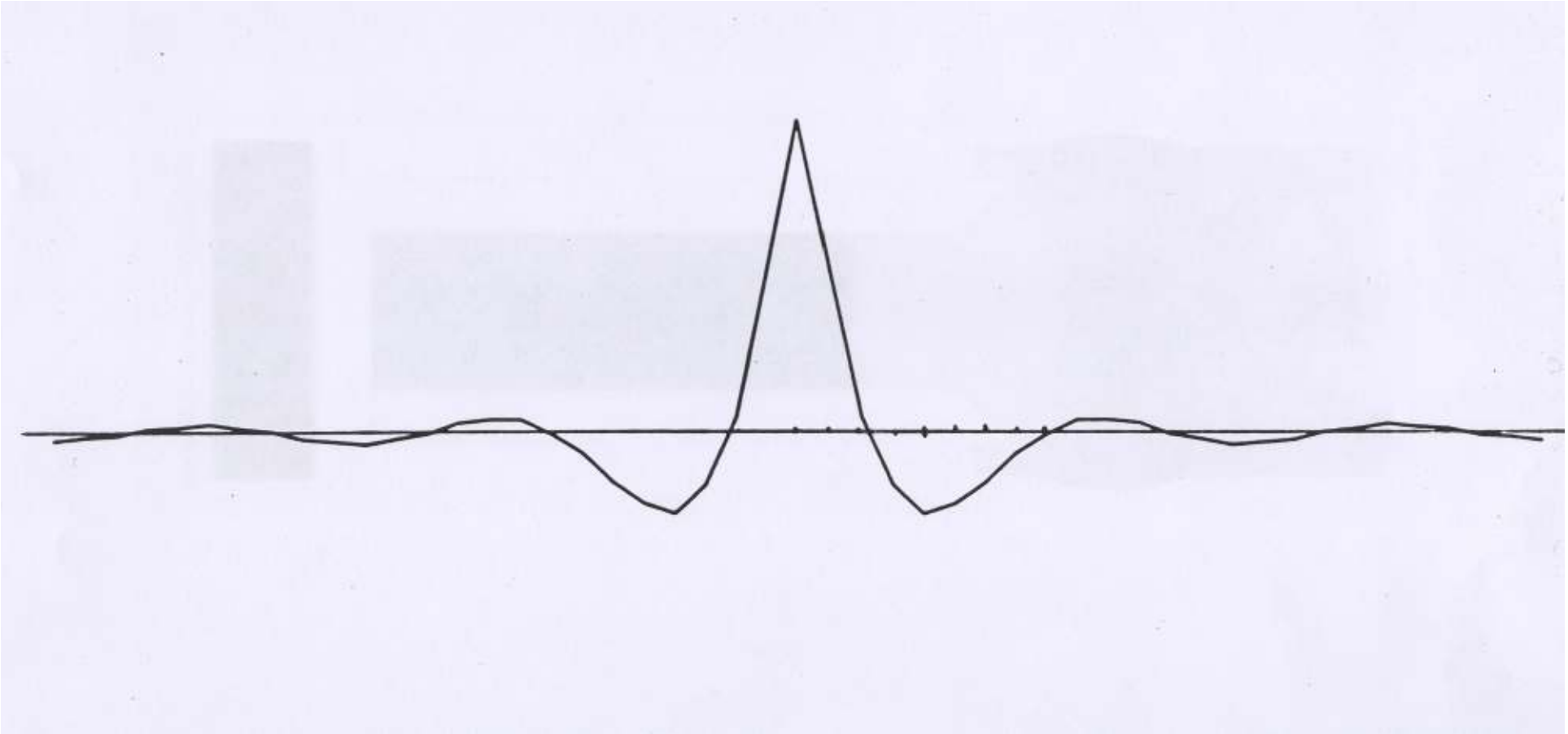
$$\boldsymbol{\mu} = (\mu_j) \equiv [E(\mathbf{y}'\mathbf{y}'^{\mathrm{T}})]^{-1}\,[\mathbf{y} - E(\mathbf{y})]$$

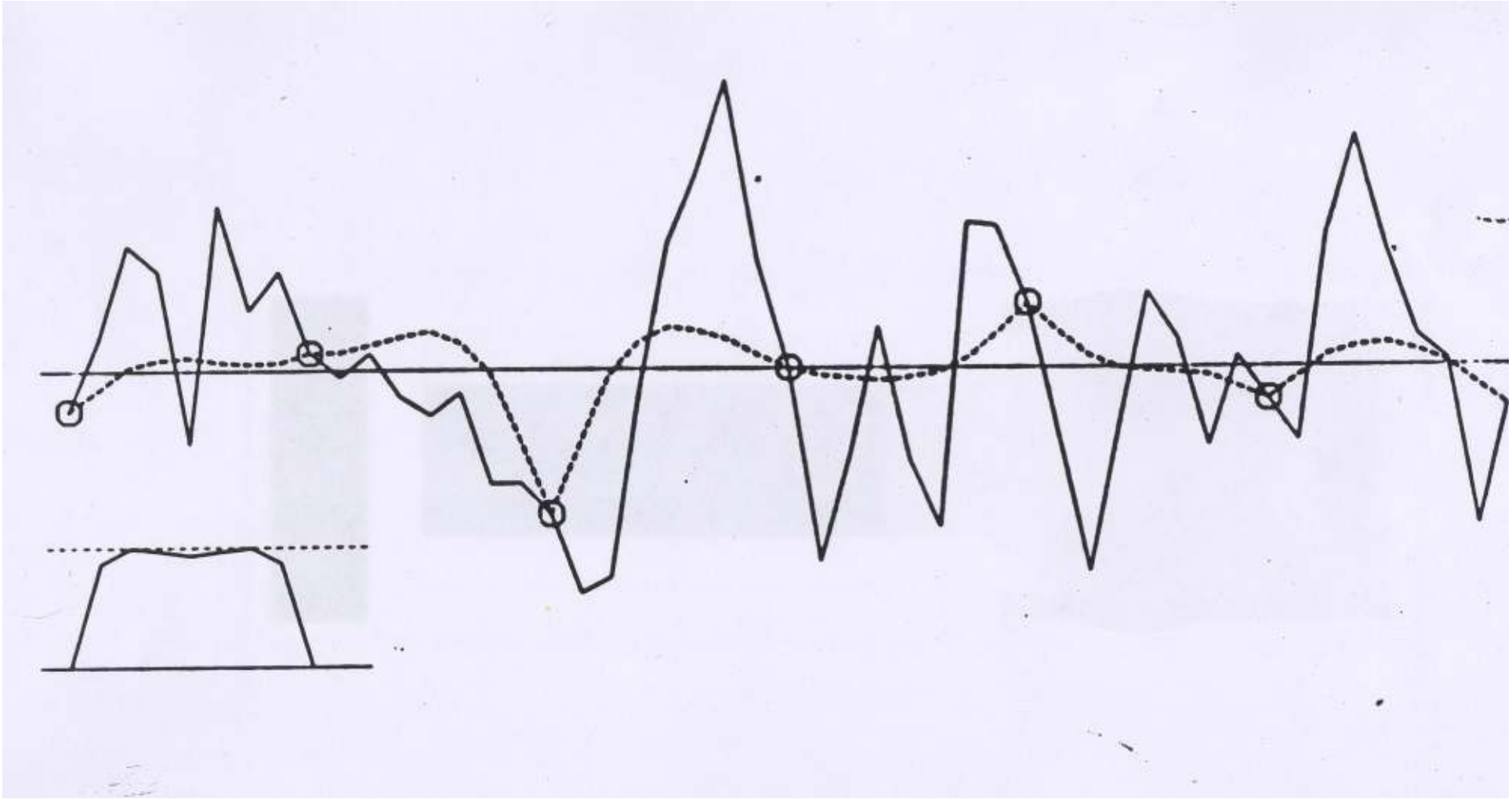
is independent of variable to be estimated

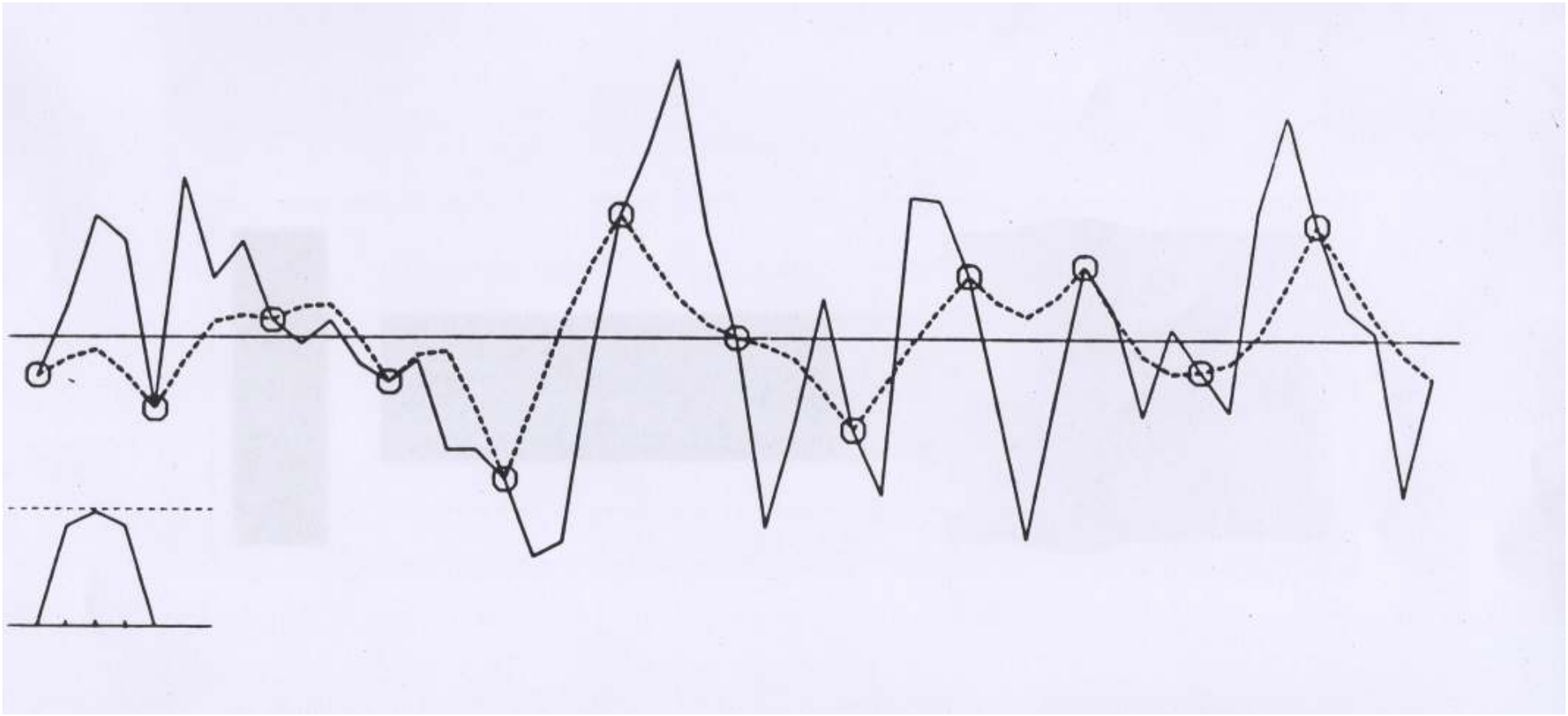
$$x^a = E(x) + \Sigma_j\, \mu_j\, E(x' y_j')$$

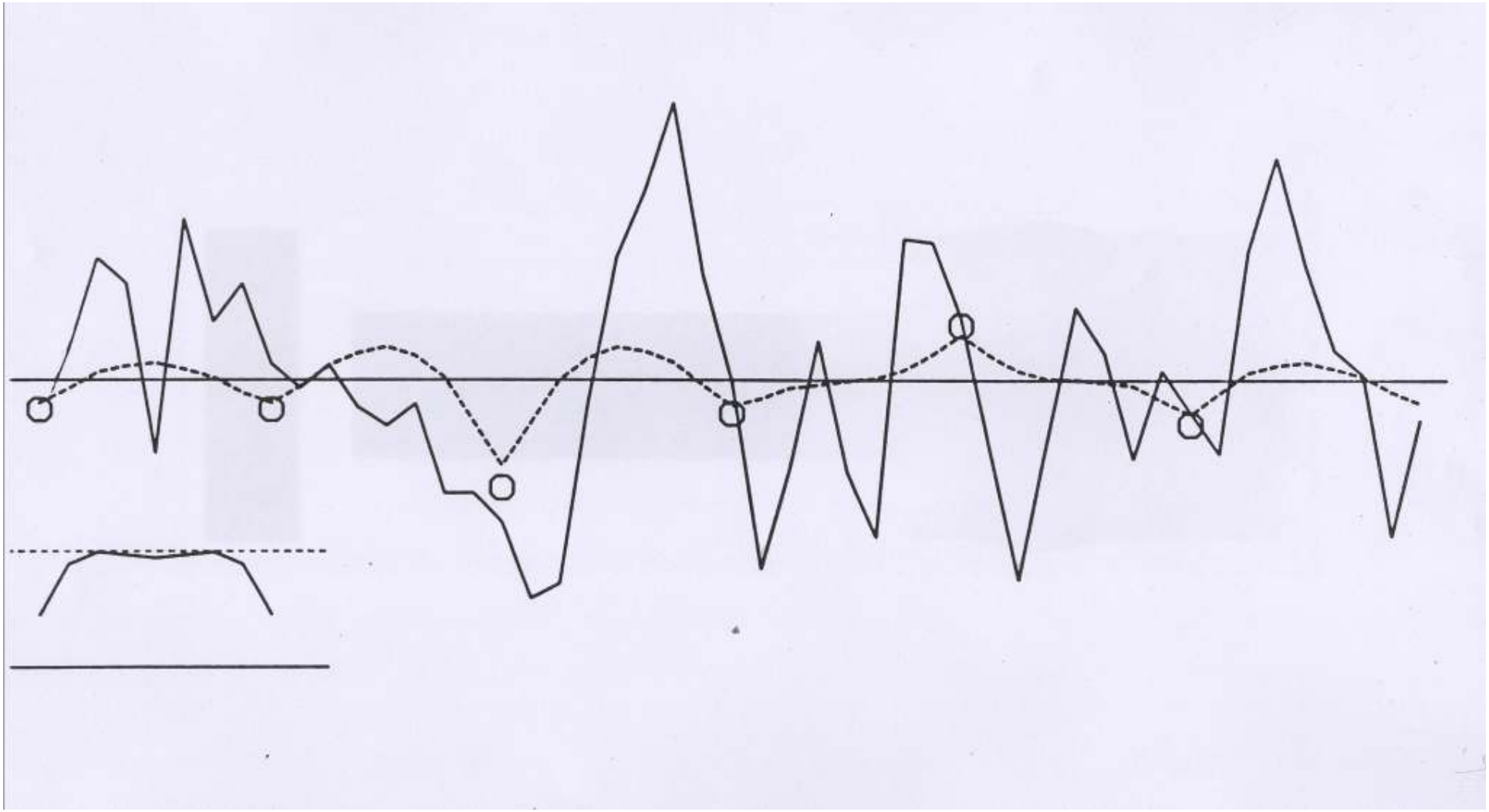
$$\begin{aligned}\varphi^a(\xi) &= E[\varphi(\xi)] + \Sigma_j\, \mu_j\, E[\varphi'(\xi)\, y_j'] \\ &= E[\varphi(\xi)] + \Sigma_j\, \mu_j\, C_\varphi(\xi, \xi_j)\end{aligned}$$

Correction made on background expectation is a linear combination of the p functions $E[\varphi'(\xi)\, y_j']$. $E[\varphi'(\xi)\, y_j']$ [$= C_\varphi(\xi, \xi_j)$], considered as a function of estimation position ξ, is the *representer* associated with observation y_j.









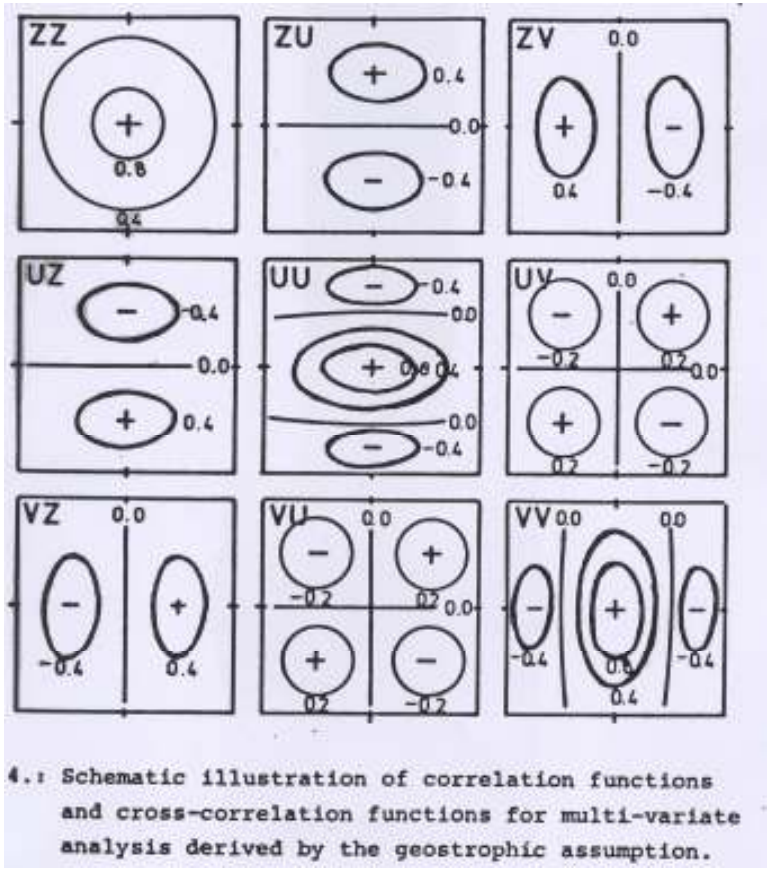
Optimal Interpolation (continued 4)

Univariate interpolation. Each physical field (*e. g.* temperature) determined from observations of that field only.

Multivariate interpolation. Observations of different physical fields are used simultaneously. Requires specification of cross-covariances between various fields.

Cross-covariances between mass and velocity fields can simply be modelled on the basis of geostrophic balance.

Cross-covariances between humidity and temperature (and other) fields still a problem.



4.: Schematic illustration of correlation functions and cross-correlation functions for multi-variate analysis derived by the geostrophic assumption.

After N. Gustafsson

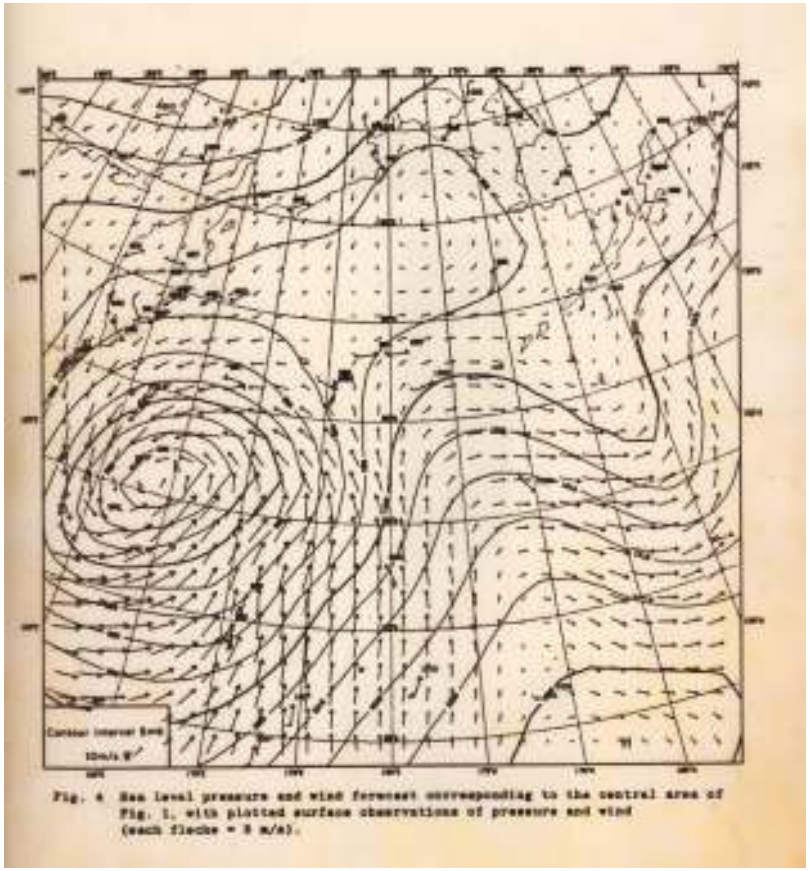


Fig. 4 Sea level pressure and wind forecast corresponding to the central area of Fig. 1, with plotted surface observations of pressure and wind (each fleche = 5 m/s).

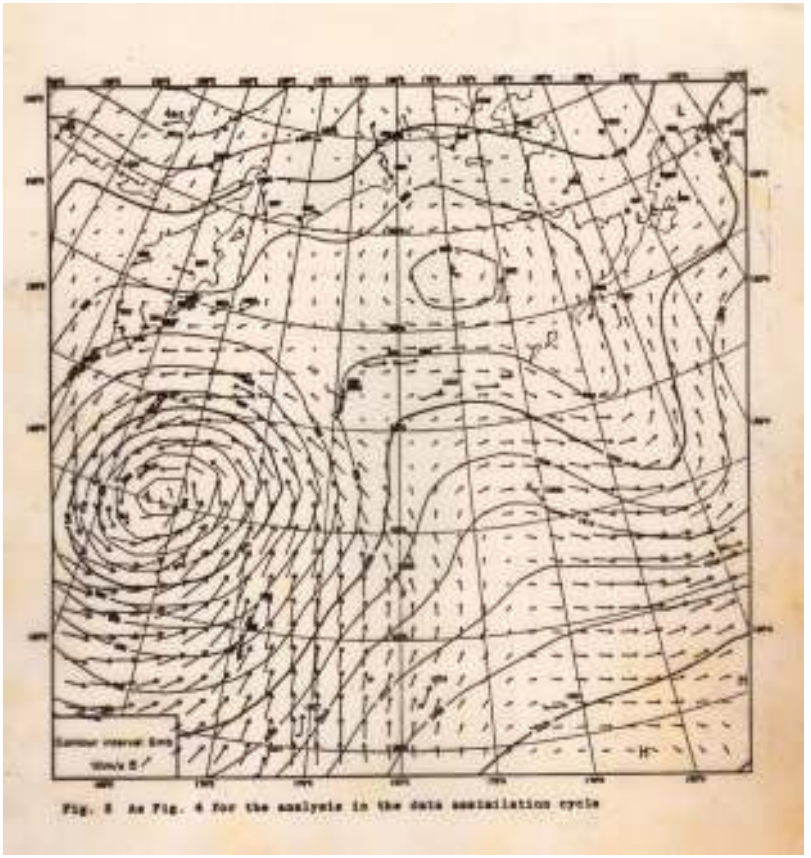


Fig. 5 As Fig. 4 for the analysis in the data assimilation cycle

After A. Lorenc

Optimal Interpolation (continued 5)

$$x^a = E(x) + E(x'\boldsymbol{y}'^{\mathrm{T}})\,[E(\boldsymbol{y}'\boldsymbol{y}'^{\mathrm{T}})]^{-1}\,[\boldsymbol{y} - E(\boldsymbol{y})] \qquad (1)$$

$$E[(x-x^a)^2] = E(x'^2) - E(x'\boldsymbol{y}'^{\mathrm{T}})\,[E(\boldsymbol{y}'\boldsymbol{y}'^{\mathrm{T}})]^{-1}\,E(\boldsymbol{y}'x') \qquad (2$$

If n-vector $\boldsymbol{x}$ to be estimated (*e*. *g*. meteorological at all grid-points of numerical model)

$$\boldsymbol{x}^a = E(\boldsymbol{x}) + E(\boldsymbol{x}'\boldsymbol{y}'^{\mathrm{T}})\,[E(\boldsymbol{y}'\boldsymbol{y}'^{\mathrm{T}})]^{-1}\,[\boldsymbol{y} - E(\boldsymbol{y})] \qquad (3)$$

$$\boldsymbol{P}^a \equiv E[(\boldsymbol{x}-\boldsymbol{x}^a)(\boldsymbol{x}-\boldsymbol{x}^a)^{\mathrm{T}}] = E(\boldsymbol{x}'\boldsymbol{x}'^{\mathrm{T}}) - E(\boldsymbol{x}'\boldsymbol{y}'^{\mathrm{T}})\,[E(\boldsymbol{y}'\boldsymbol{y}'^{\mathrm{T}})]^{-1}\,E(\boldsymbol{y}'\boldsymbol{x}'^{\mathrm{T}}) \qquad (4)$$

Eq. (3) says the same as eq. (1), but eq. (4) says more than eq. (2) in that it defines off-diagonal entries of estimation error covariance matrix $\boldsymbol{P}^a$.

If probability distributions are *globally* gaussian, eqs (3-4) achieve bayesian estimation, in the sense that $P(\boldsymbol{x} \mid \boldsymbol{y}) = \mathcal{N}[\boldsymbol{x}^a, \boldsymbol{P}^a]$.

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

Best Linear Unbiased Estimate (continuation 1)

$$\boldsymbol{x}^b = \boldsymbol{x} + \boldsymbol{\xi}^b \qquad (1)$$

$$\boldsymbol{y} = H\boldsymbol{x} + \boldsymbol{\varepsilon} \qquad (2)$$

A probability distribution being known for the couple $(\boldsymbol{\xi}^b, \boldsymbol{\varepsilon})$, eqs (1-2) define probability distribution for the couple $(\boldsymbol{x}, \boldsymbol{y})$, with

$E(\boldsymbol{x}) = \boldsymbol{x}^b$, $\boldsymbol{x}' = \boldsymbol{x} - E(\boldsymbol{x}) = -\boldsymbol{\xi}^b$

$E(\boldsymbol{y}) = H\boldsymbol{x}^b$, $\boldsymbol{y}' = \boldsymbol{y} - E(\boldsymbol{y}) = \boldsymbol{y} - H\boldsymbol{x}^b = \boldsymbol{\varepsilon} - H\boldsymbol{\xi}^b$

$\boldsymbol{d} \equiv \boldsymbol{y} - H\boldsymbol{x}^b$ is called the *innovation vector*.

Best Linear Unbiased Estimate (continuation 2)

$$E(\boldsymbol{x}'\boldsymbol{y}'^{\mathrm{T}}) = E[-\boldsymbol{\xi}^b(\boldsymbol{\varepsilon}-H\boldsymbol{\xi}^b)^{\mathrm{T}}] = -E(\boldsymbol{\xi}^b\boldsymbol{\varepsilon}^{\mathrm{T}}) + E(\boldsymbol{\xi}^b\boldsymbol{\xi}^{b\mathrm{T}})H^{\mathrm{T}}$$

$$E(\boldsymbol{y}'\boldsymbol{y}'^{\mathrm{T}}) = E[(\boldsymbol{\varepsilon}-H\boldsymbol{\xi}^b)\,(\boldsymbol{\varepsilon}-H\boldsymbol{\xi}^b)^{\mathrm{T}}] = HE(\boldsymbol{\xi}^b\boldsymbol{\xi}^{b\mathrm{T}})H^{\mathrm{T}} + E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}) - E(\boldsymbol{\varepsilon}\boldsymbol{\xi}^{b\mathrm{T}}) - E(\boldsymbol{\xi}^b\boldsymbol{\varepsilon}^{\mathrm{T}})$$

Assume $E(\boldsymbol{\xi}^b\boldsymbol{\varepsilon}^{\mathrm{T}}) = 0$ (not mathematically restrictive)

and set $E(\boldsymbol{\xi}^b\boldsymbol{\xi}^{b\mathrm{T}}) \equiv P^b$ (also often denoted B), $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}) \equiv R$

Best Linear Unbiased Estimate (continuation 3)

Apply formulæ for Optimal Interpolation

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

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 $K = P^b H^{\mathrm{T}} [H P^b H^{\mathrm{T}} + R]^{-1} = P^a H^{\mathrm{T}} R^{-1}$ 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)

H can be any linear operator

Example : (scalar) satellite observation

$\boldsymbol{x} = (x_1, \ldots, x_n)^T$ temperature profile

Observation $y = \Sigma_i h_i x_i + \varepsilon = \boldsymbol{H}\boldsymbol{x} + \varepsilon$, $\boldsymbol{H} = (h_1, \ldots, h_n)$, $E(\varepsilon^2) = r$

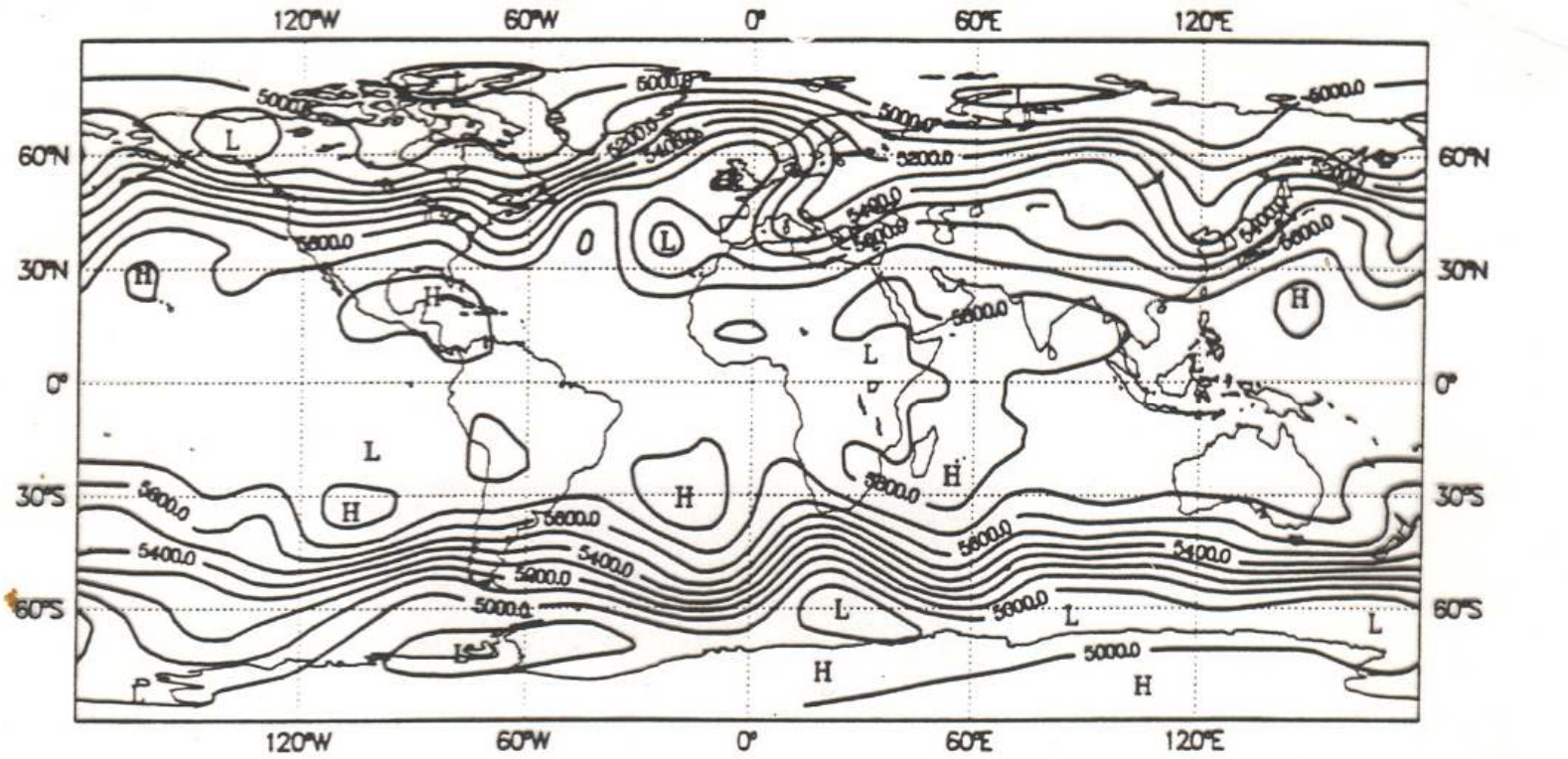
Background $\boldsymbol{x}^b = (x_1^b, \ldots, x_n^b)^T$, error covariance matrix $P^b = (p_{ij}^b)$

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

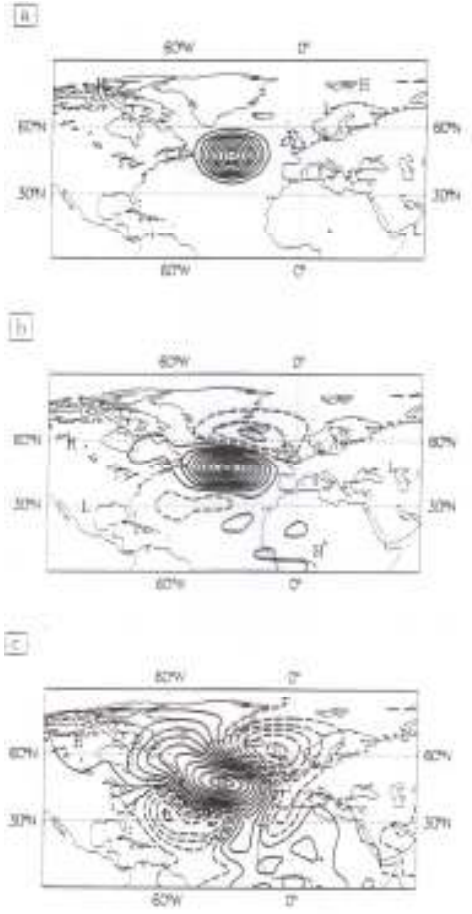
$[HP^b H^T + R]^{-1} (\boldsymbol{y} - H\boldsymbol{x}^b) = (y - \Sigma_\iota h_\iota x_\iota^b) / (\Sigma_{ij} h_i h_j p_{ij}^b + r)^{-1} \equiv \mu$ scalar !

- $P^b = p^b \boldsymbol{I}_n$ $x_i^a = x_i^b + p^b h_i \mu$
- $P^b = \mathrm{diag}(p_{ii}^b)$ $x_i^a = x_i^b + p_{ii}^b h_i \mu$
- General case $x_i^a = x_i^b + \Sigma_j p_{ij}^b h_j \mu$

Each level i is corrected, not only because of its own contribution to the observation, but because of the contribution of the other levels to which its background error is correlated.



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.

How to introduce temporal dimension and, in particular, temporal evolution of uncertainty on the state of the system ?

From an algorithmic point of view, two approaches (which can both be derived from the theory of the BLUE)

Variational Assimilation

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

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 (Kalman Filter, lecture by I. Szunyogh).

Exact bayesian estimation ?

Particle filters

Predicted ensemble at time t : $\{x^b_n, n = 1, \ldots, N\}$, each element with its own weight (probability) $P(x^b_n)$

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

Bayes' formula

$$P(x^b_n|y) \sim P(y|x^b_n)\, P(x^b_n)$$

Defines updating of weights

Bayes' formula

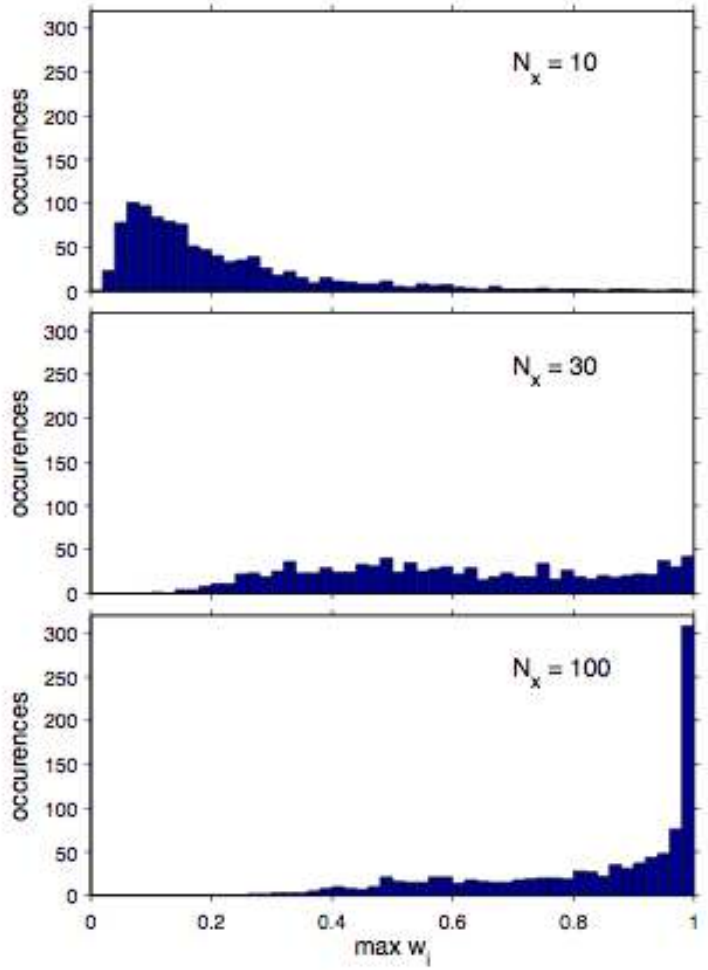
$$P(x^b{}_n|y) \sim P(y|x^b{}_n)\, P(x^b{}_n)$$

Defines updating of weights; particles are not modified. Asymptotically converges to bayesian pdf. Very easy to implement.

Observed fact. For large state dimension, ensemble tends to collapse.

Behavior of $\max w^i$

- $N_e = 10^3$; $N_x = 10, 30, 100$; 10^3 realizations



average squared error of posterior mean = 5.5

... = 25

... = 127

C. Snyder,
http://www.cawcr.gov.au/staff/pxs/wmoda5/Oral/Snyder.pdf

Problem originates in the 'curse of dimensionality' Large dimension pdf's are very diffuse, so that very few particles (if any) are present in areas where conditional probability ('*likelihood*') $P(y|x)$ is large.

Bengtsson *et al*. (2008) and Snyder *et al*. (2008) evaluate that stability of filter requires the size of ensembles to increase exponentially with space dimension.

Alternative possibilities (review in van Leeuwen, 2009, *Mon. Wea. Rev.*, 4089-4114)

Resampling. Define new ensemble.

Simplest way. Draw new ensemble according to probability distribution defined by the updated weights. Give same weight to all particles. Particles are not modified, but particles with low weights are likely to be eliminated, while particles with large weights are likely to be drawn repeatedly. For multiple particles, add noise, either from the start, or in the form of 'model noise' in ensuing temporal integration.

Random character of the sampling introduces noise. Alternatives exist, such as *residual sampling* (Lui and Chen, 1998, van Leeuwen, 2003). Updated weights w_n are multiplied by ensemble dimension N. Then p copies of each particle n are taken, where p is the integer part of Nw_n. Remaining particles, if needed, are taken randomly from the resulting distribution.

Importance Sampling.

Use a *proposal density* that is closer to the new observations than the density defined by the predicted particles (for instance the density defined by EnKF, after the latter has used the new observations). Independence between observations is then lost in the computation of likelihood $P(y|x)$ (or is it not ?)

In particular, *Guided Sequential Importance Sampling* (van Leeuwen, 2002). Idea : use observations performed at time k to resample ensemble at some timestep anterior to k, or 'nudge' integration between times k-1 and k towards observation at time k.

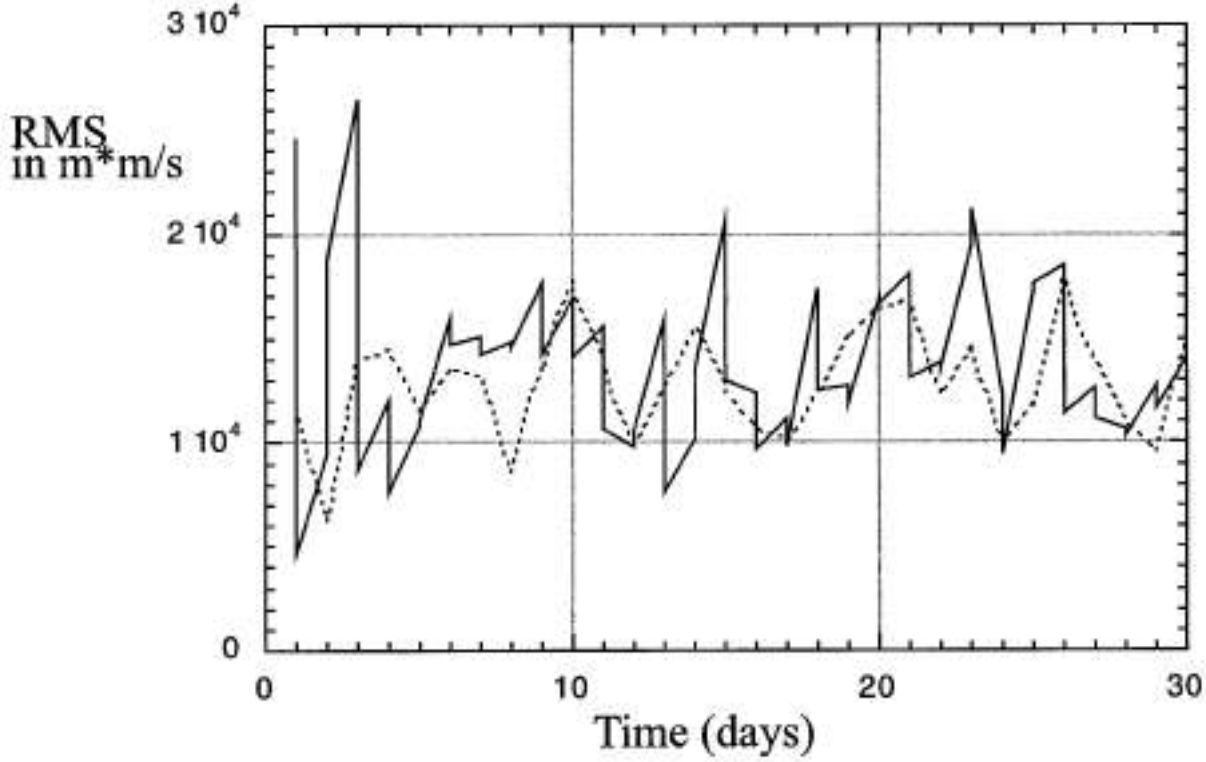


FIG. 12. Comparison of rms error ($m^2\ s^{-1}$) between ensemble mean and independent observations (dotted line) and the std dev in the ensemble (solid line). The excellent agreement shows that the SIRF is working correctly.

van Leeuwen, 2003, *Mon. Wea. Rev.*, **131**, 2071-2084

Conclusions (partial)

Assimilation, which originated from the need of defining initial conditions for numerical weather forecasts, has progressively extended to many diverse applications

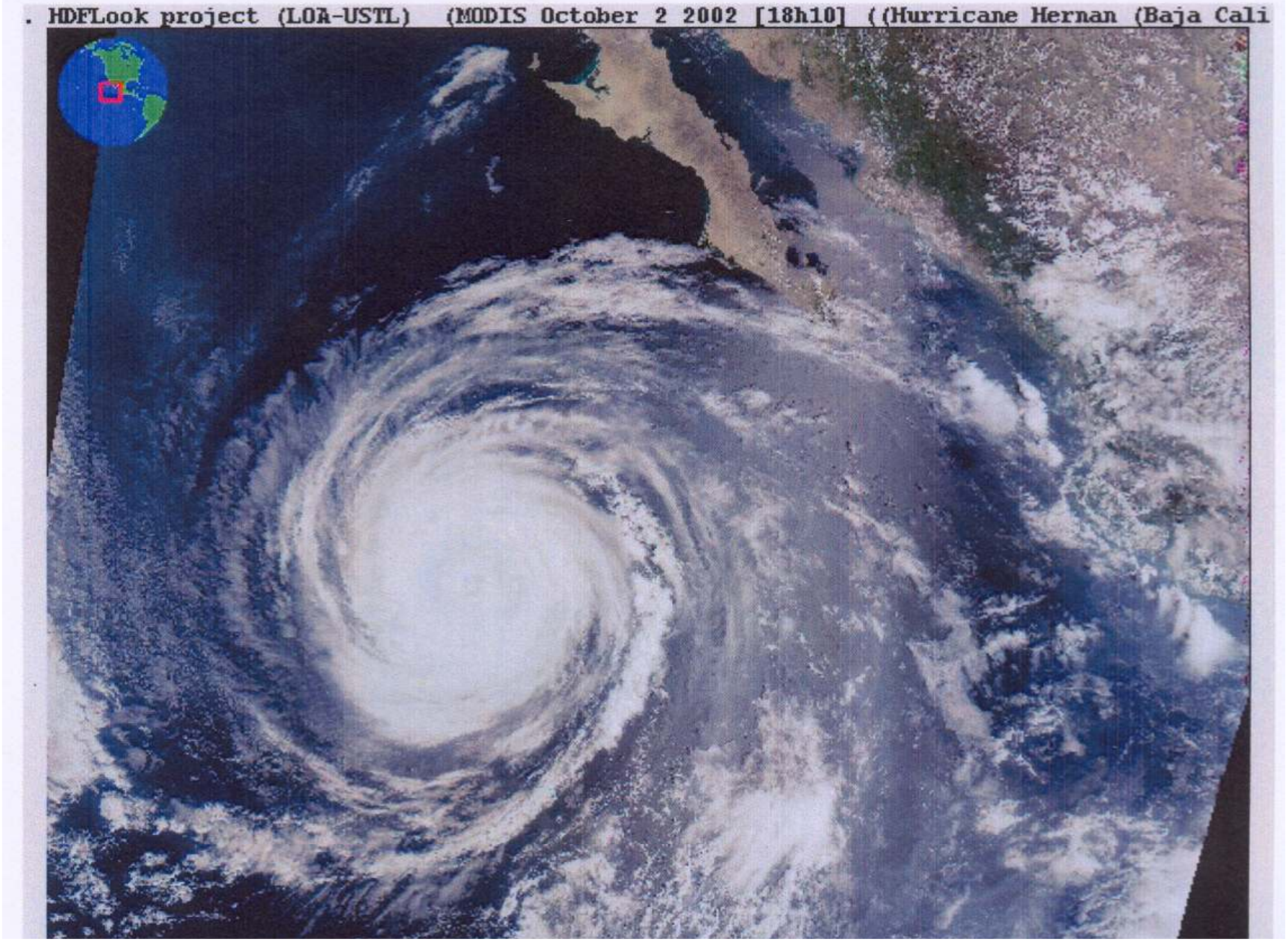
- Oceanography
- Atmospheric chemistry (both troposphere and stratosphere)
- Oceanic biogeochemistry
- Ground hydrology
- Terrestrial biosphere and vegetation cover
- Glaciology
- Magnetism (both planetary and stellar)
- Plate tectonics
- Planetary atmospheres (Mars, …)
- Reassimilation of past observations (mostly for climatological purposes, ECMWF, NCEP/NCAR)
- Identification of source of tracers
- Parameter identification
- *A priori* evaluation of anticipated new instruments
- Definition of observing systems *(Observing Systems Simulation Experiments)*
- Validation of models
- Sensitivity studies (adjoints)
- …

Assimilation is related to

- Estimation theory
- Probability theory
- Atmospheric and oceanic dynamics
- Atmospheric and oceanic predictability
- Instrumental physics
- Optimisation theory
- Control theory
- Algorithmics and computer science
- …

A few of the (many) remaining problems :

- Observability (data are noisy, system is chaotic !)
- More accurate identification and quantification of errors affecting data particularly the assimilating model (will always require independent hypotheses)
- Assimilation of images
- Particle Filters may define the way to fully bayesian assimilation algorithms
- …


. HDFLook project (LOA-USTL) (MODIS October 2 2002 [18h10] ((Hurricane Hernan (Baja Cali