The linear case.
Best Linear Unbiased Estimation.
Simple examples.
Part 2

Olivier Talagrand
School *Data Assimilation*
Nordic Institute for Theoretical Physics (NORDITA)
Stockholm, Sweden
26 April 2011

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)

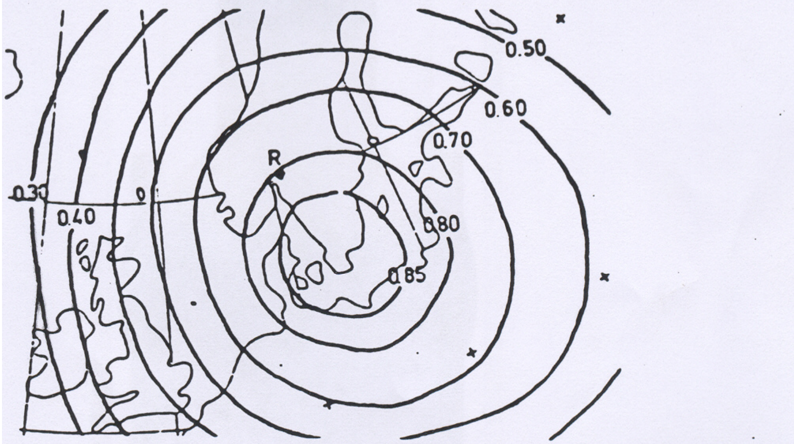
- Expectation $E[\varphi(\xi)]$; $\varphi'(\xi) \equiv \varphi(\xi) - E[\varphi(\xi)]$
- Variance $Var[\varphi(\xi)] = E\{[\varphi'(\xi)]^2\}$
- Covariance function

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

- 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

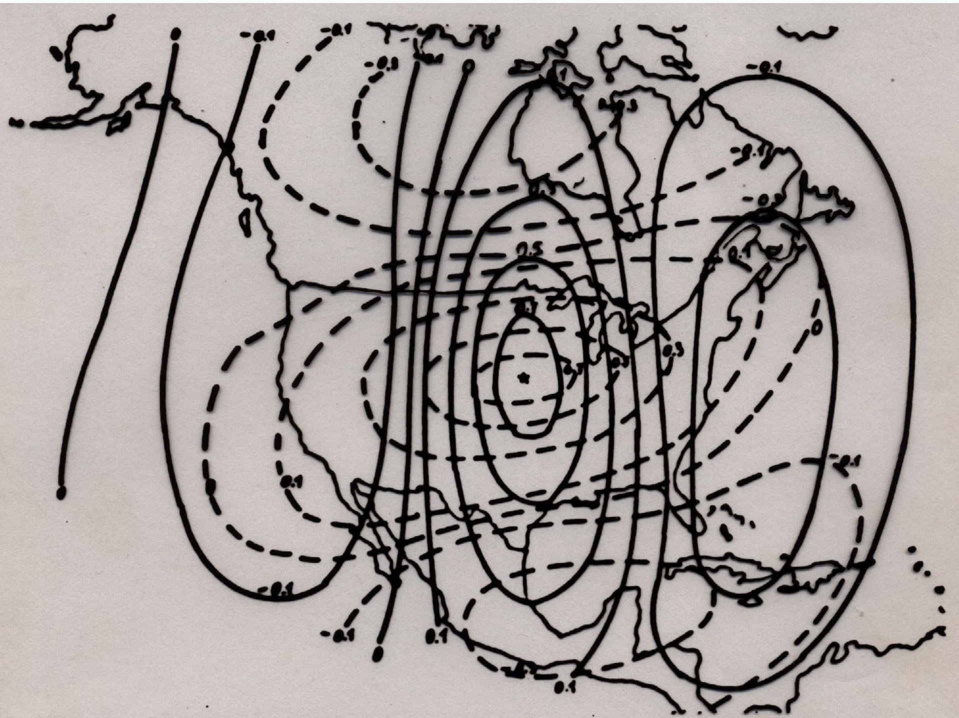


Figure 4.2.4.3: Isolines for the auto-correlation of the 500 mb u-wind component (dashed line) and the auto-correlation of the 500 mb v-wind component (full line). The "star" indicates the position of the reference station. (From Buel (1972).

After N. Gustafsson

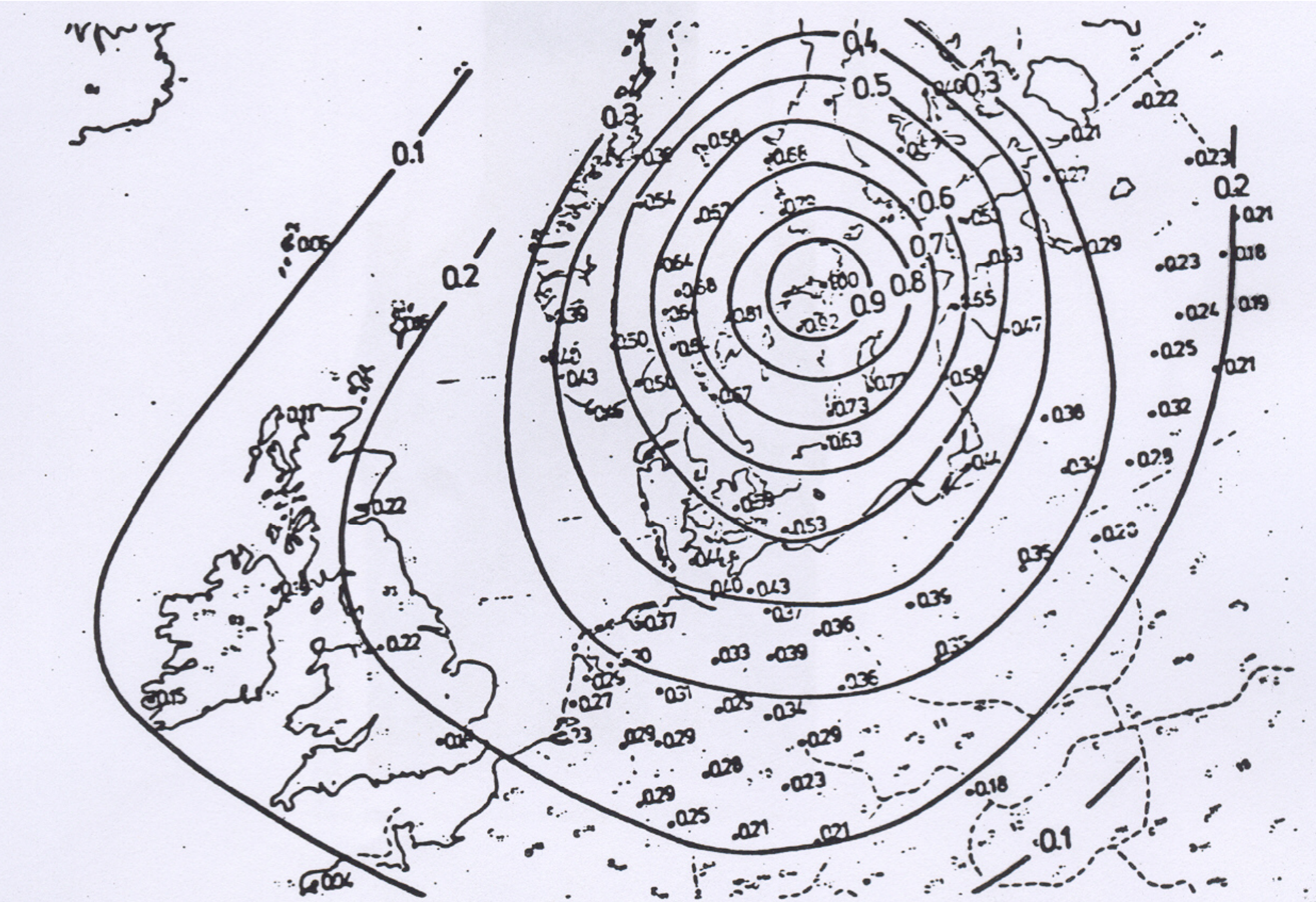


Figure 5.1.1.4.1 Auto-correlation of errors in 12h numerical forecasts of surface pressure in a reference station (Stockholm) and other stations.

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

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

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

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

i. e.,

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

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

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

Minimized quadratic estimation error

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

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

Optimal Interpolation (continued 2)

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

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

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

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

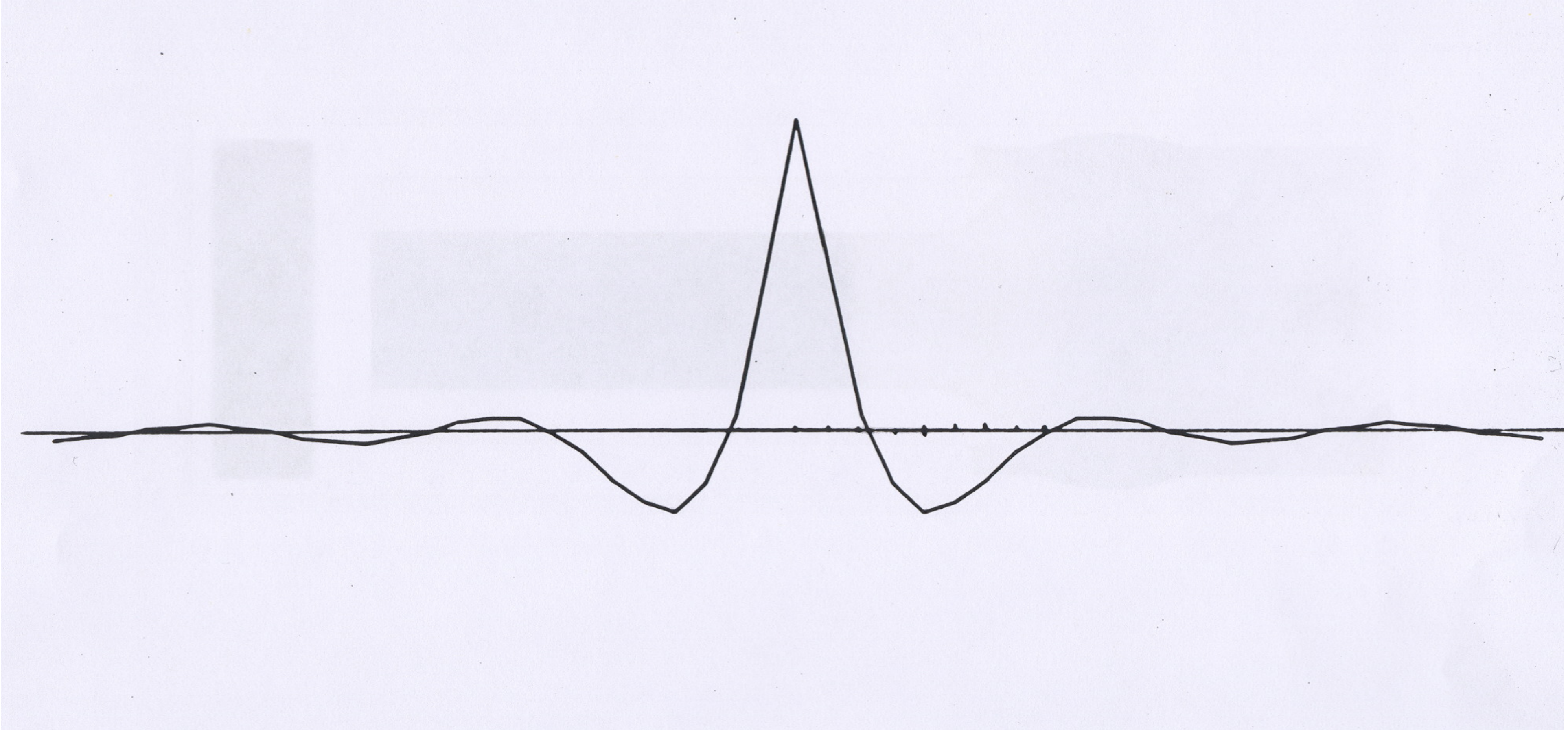
$$\boldsymbol{\mu} = (\mu_j) \equiv [E(\mathbf{y}'\mathbf{y}'^{\mathrm{T}})]^{-1}\,[\mathbf{y} - E(\mathbf{y})]$$

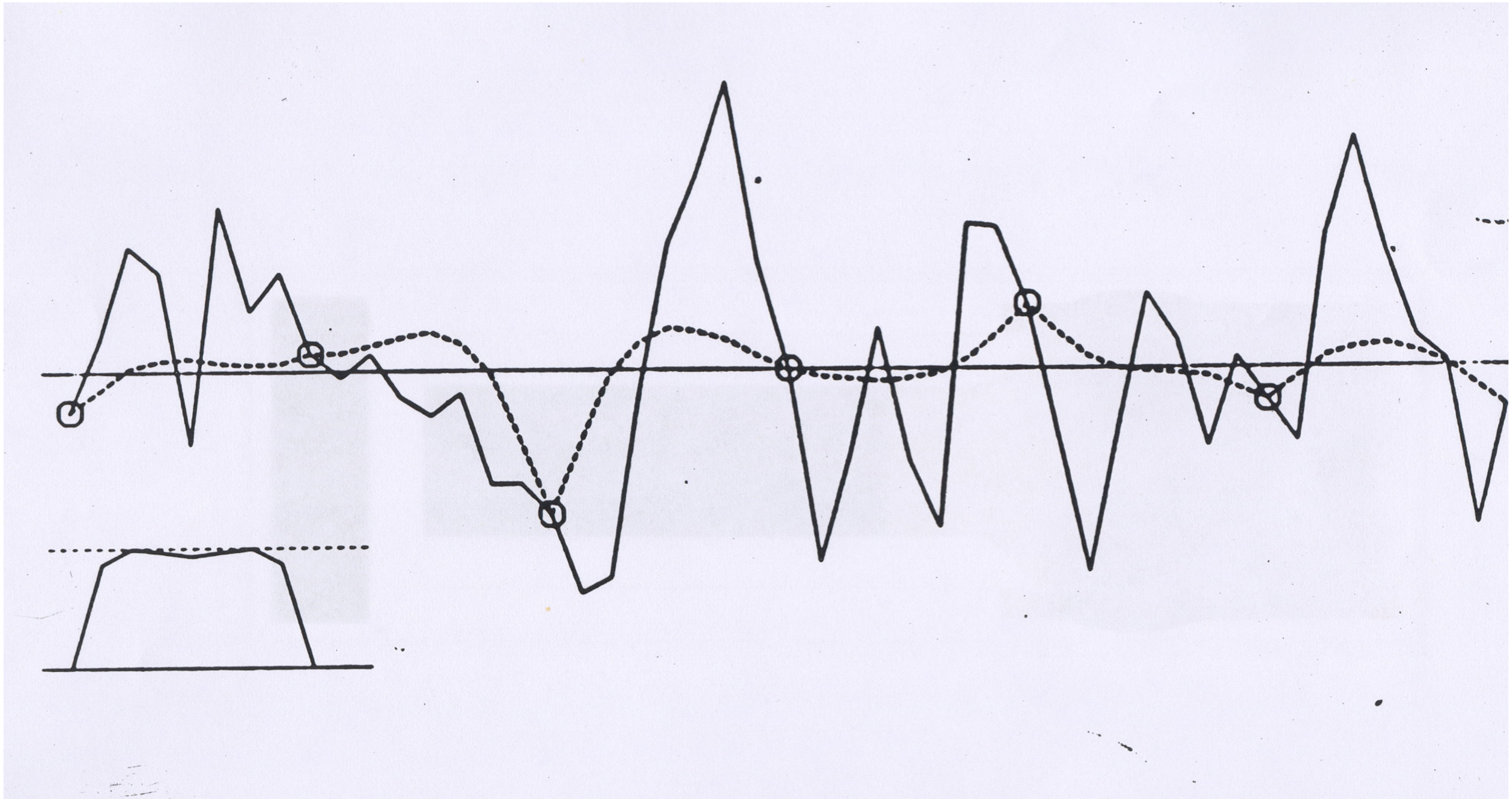
is independent of variable to be estimated

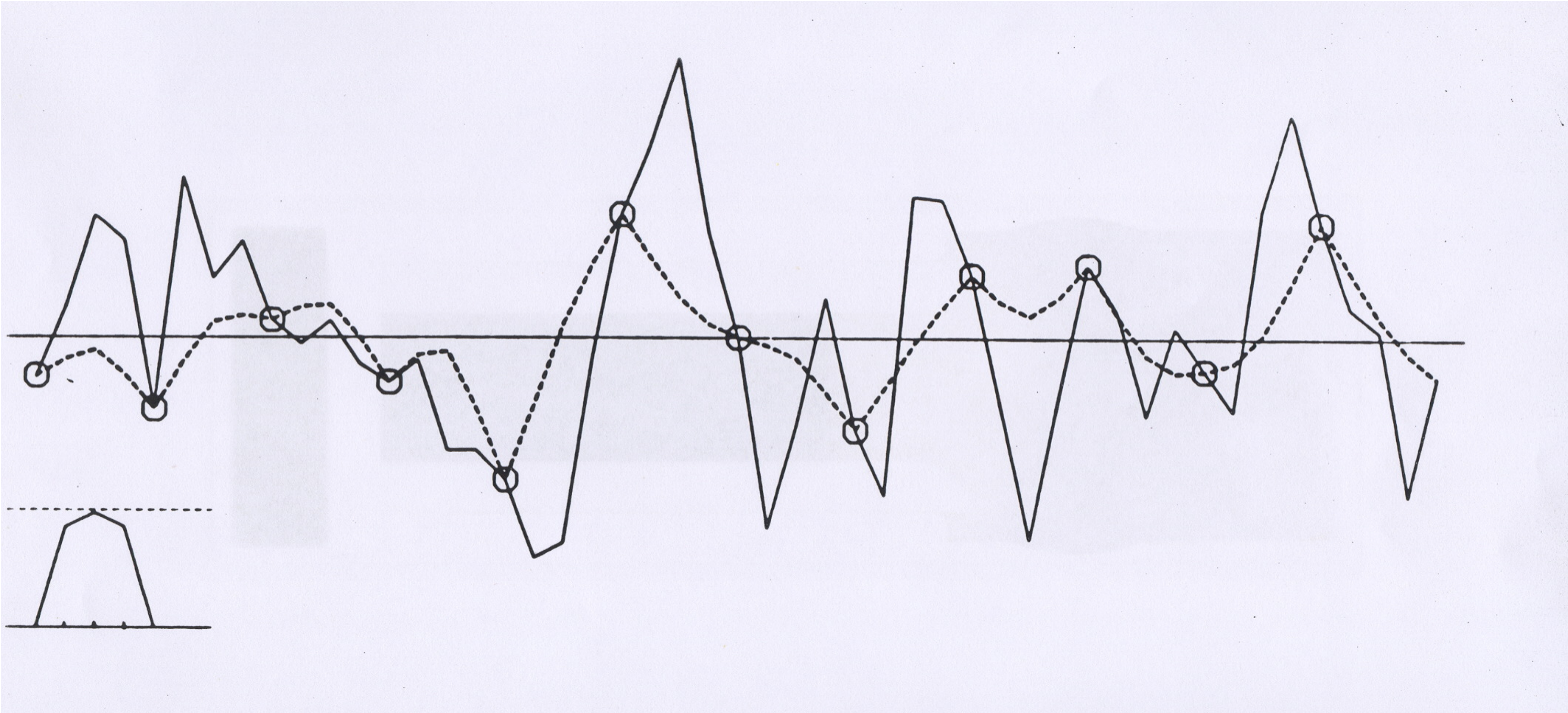
$$x^a = E(x) + \Sigma_j\, \mu_j\, E(x'y_j')$$

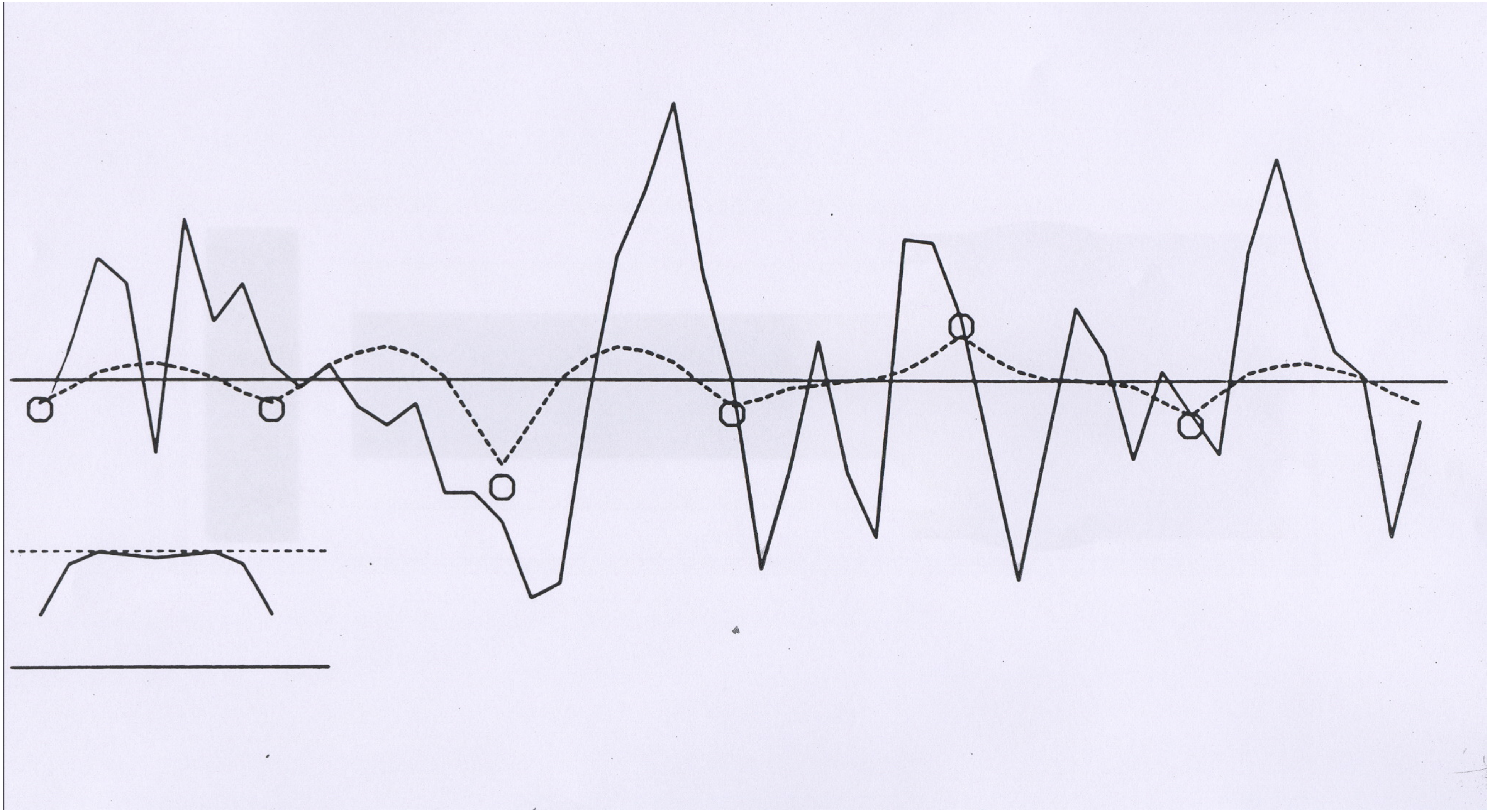
$$\begin{aligned}\varphi^a(\xi) &= E[\varphi(\xi)] + \Sigma_j\, \mu_j\, E[\varphi'(\xi)\, y_j'] \\ &= E[\varphi(\xi)] + \Sigma_j\, \mu_j\, C_\varphi(\xi, \xi_j)\end{aligned}$$

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.









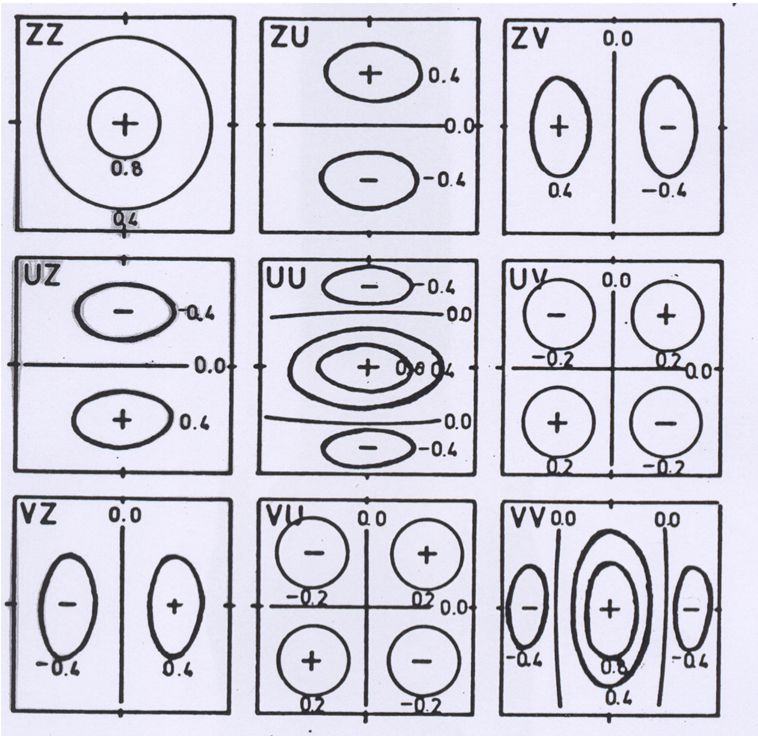
Optimal Interpolation (continued 4)

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

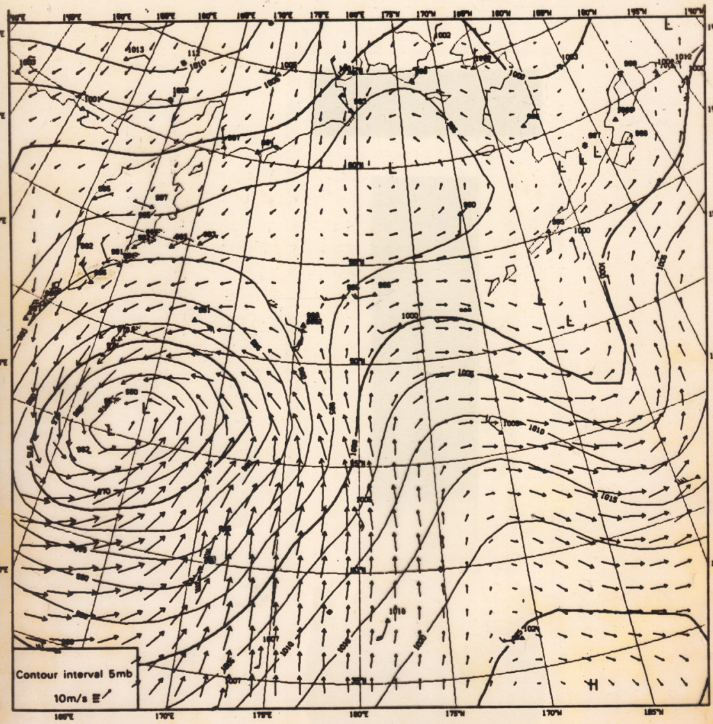


Fig. 4 Sea level pressure and wind forecast corresponding to the central area of Fig. 1, with plotted surface observations of pressure and wind (each fleche = 5 m/s).

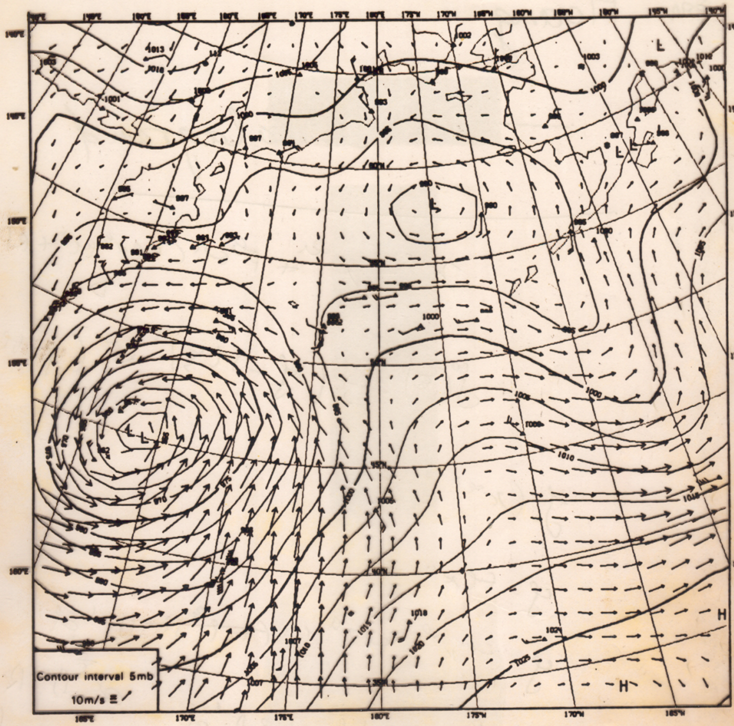


Fig. 5 As Fig. 4 for the analysis in the data assimilation cycle

After A. Lorenc

Optimal Interpolation (continued 5)

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

$$E[(x-x^a)^2] = E(x'^2) - E(x'\mathbf{y}'^T) [E(\mathbf{y}'\mathbf{y}'^T)]^{-1} E(\mathbf{y}'x') \quad (2$$

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

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

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

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 $P(\boldsymbol{x} \mid \mathbf{y}) = \mathcal{N}[\boldsymbol{x}^a, \boldsymbol{P}^a]$.

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)

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

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

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

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

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

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

Best Linear Unbiased Estimate (continuation 2)

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

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

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

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

Best Linear Unbiased Estimate (continuation 3)

Apply formulæ for Optimal Interpolation

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

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

Equivalent set of formulæ

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

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

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