Validacion Objetiva

Olivier Talagrand Curso Intensivo en Asimilacion de Datos Buenos Aires, Argentina

3 Noviembre 2008

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.

Validation must therefore aim primarily at determining, as far as possible, the accuracy with which assimilation reconstructs the state of the flow. In particular, if one wants to compare two different assimilation procedures, the ultimate test lies in the comparison of the accuracies with which those two procedures reconstruct the flow.

Now, the state of the flow is not perfectly known, and the output of an assimilation process is precisely meant to be the best possible estimate of that state. So there is a circular argument there. But there is another aspect to evaluation of an assimilation process. Does the process make the best possible use of the available information? This is totally distinct from the accuracy with which the process reconstructs the state of the flow. We will distinguish *numerical accuracy* from *optimality* (the property that the process makes the best possible use of the available information). Assimilation aims at minimizing statistical error on estimated fields (reasonable since uncertainty on data can be described only statistically). This means that objective validation can be only *statistical*. Validation on individual situations can be misleading.

Example

 $x^{b} = x + \zeta^{b}$ ('background') $y = x + \varepsilon$ ('observation')

where unknown *x* is scalar, and errors ζ^b and ε are centred ($E(\zeta^b) = E(\varepsilon) = 0$), have respective variances $E[(\zeta^b)^2] = p^b$ and $E(\varepsilon^2) = r$, and are uncorrelated, $E(\varepsilon\zeta^b) = 0$. Least-variance linear estimate of *x* from x^b and *y*

 $x^{a} = (rx^{b} + p^{b}y) / (r + p^{b})$

One can expect the squared differences $(x^a-x^b)^2$ and $(x^a-y)^2$ to be smaller than p^b and r respectively.

However, if errors ζ^{b} and ε are gaussian, and if $r \ll p^{b}$, the probability is 0.32 that $(x^{a}-x^{b})^{2} \Rightarrow p^{b}$, and 0.14 if $r=p^{b}/2$.

Validation against observations used in the assimilation ?

That cannot do, since best fit to observation z will be provided by best estimate of z from data. That best estimate is z itself.

It can be shown that fit of assimilation to any observation z can be made arbitrarily close by assuming that the corresponding observation error is small enough.

In a perfectly accurate assimilation system, the difference between an observation and the corresponding estimated value will statistically be equal to the corresponding observation error. If fit of analysis to observation is different from corresponding observation error (assuming the latter is exactly known), that will be a sign of imperfect accuracy.

Consequence. Objective validation of assimilation can be made only against observations that have not been used in the assimilation. If validation is made on forecasts, no problem (such 'external' observations will exist). But if validation is made on analysed fields, this may mean that some data will have to be kept for validation (and therefore not used in the assimilation), thus possibly decreasing the accuracy of the assimilation. Validation may be possible only on sub-optimal assimilation procedures.

Validation observation must be unbiased, and affected with errors that are independent of errors affecting data used in the assimilation

 $x^b = x + \zeta^b$ $y = x + \mathcal{E}$

where errors ζ^{b} and ε now have same variance *r*.

 $x^{a} = (x^{b} + y) / 2$

Validating observation

 $y' = x + \mathcal{E}'$

where ε has expectation 0, variance *r*, but is correlated with ε , $E(\varepsilon \varepsilon') = cr$ (and uncorrelated with ζ^{b}). Best fit to analysis will be provided by linear least-variance estimate of *y* from x^{b} and *y*. That estimate is

 $x^{'a} = [(1-c)x^b + (1+c)y)] / 2$

different from x^a .

Objective validation of assimilation can only be statistical, and must be made against observations (or data) that are unbiased, and are affected by errors that are statistically independent of the errors affecting the data used in the assimilation.

Amplitude of forecast error, if estimated against observations that are really independent of observations used in assimilation, is an objective measure of accuracy of assimilation.

But neither the unbiasedness nor the 'independence' of the verifying observations can be objectively verified, at least within the world of data and assimilation. External knowledge must be used.

Remark. Accuracy of verifying observations is secondary.

Analysed value

 $x^a = x + \zeta^a$

Verifying observation

 $z = x + \zeta$

If z is independent of data used for producing x^a , then $E(\zeta^a \zeta) = 0$, and

 $E[(z-x^{a})^{2}] = b^{2} + E(\zeta^{2}) + E(\zeta^{a2})$

where *b* is statistical bias in x^a ($b = E(x^a-z)$). Best fit of x^a to *z* will always be obtained by least-variance estimate (*i. e.*, *b*=0 and $E(\zeta^{a2})$ minimum).

Variance $E(\zeta^{a2})$ of assimilation error will be known only if variance $E(\zeta^{2})$ is known.

Best Linear Unbiased Estimate

State vector x, belonging to state space $S(\dim S = n)$, to be estimated. Available data in the form of

- A '*background*' estimate, 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*.

Best Linear Unbiased Estimate (continuation 2)

Assume $E(\zeta^b) = 0$, $E(\varepsilon) = 0$ Set $d \equiv y - Hx^b$ (innovation vector)

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

Assume $E(\zeta^{b}\varepsilon^{T}) = 0$ (not restrictive). Set $E(\zeta^{b}\zeta^{bT}) = P^{b}$ (also often denoted **B**), $E(\varepsilon^{T}) = R$

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

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

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

Determination of the *BLUE* requires (at least apparently) the *a priori* specification of the expectation and covariance matrix, *i. e.* the statistical moments of orders 1 and 2, of the errors. The expectation is required for unbiasing the data in the first place.

Questions

- Is it possible to objectively evaluate the quality of an assimilation system ?
- Is it possible to objectively evaluate the first- and second-order statistical moments of the data errors, whose specification is required for determining the *BLUE* ?
- Is it possible to objectively determine whether an assimilation system is optimal ?
- More generally, how to make the best of an assimilation system ?

$$x^{b} = x + \zeta^{b}$$
$$y = Hx + \varepsilon$$

The only combination of the data that is a function of only the error is the innovation vector

 $d = y - Hx^b = \varepsilon - H\zeta^b$

Innovation is the only objective source of information on errors. Now innovation is a combination of background and observation errors, while determination of the *BLUE* requires explicit knowledge of the statistics of both observation and background errors.

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

Innovation alone will never be sufficient to determine the required statistics.

With hypotheses made above

 $E(d) = 0 \quad ; \quad E(dd^{\mathrm{T}}) = HP^{b}H^{\mathrm{T}} + R$

Possible to check statistical *consistency* between *a priori* assumed and *a posteriori* observed statistics of innovation.

Consistency, which will be considered now, is a different quality than either accuracy or optimality.

Consider assimilation scheme of the form

$$\boldsymbol{x}^a = \boldsymbol{x}^b + \boldsymbol{K}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}^b) \tag{1}$$

with any (*i. e.* not necessarily optimal) gain matrix K.

(1) \Leftrightarrow if data are perfect, then so is the estimate x^a .

Data-minus-Analysis (DmA) difference

$$\delta \equiv \begin{pmatrix} x^b - x^a \\ y - Hx^a \end{pmatrix} = \begin{pmatrix} -Kd \\ (I_p - HK)d \end{pmatrix}$$

For given gain matrix *K*, one-to-one correspondance $d \Leftrightarrow \delta$ It is exactly equivalent to compute statistics on either the innovation *d* or on the *DmA* difference δ . For perfectly consistent system (*i. e.*, system that uses the exact error statistics):

 $E(d) = 0 \ (\iff E(\delta) = 0)$

Any systematic bias in either the innovation vector or the DmA difference is the signature of an inappropriately taken into account bias in either the background or the observation (or both).

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

A perfectly consistent analysis statistically fits the data to within their own accuracy.

If new data are added to (removed from) an optimal analysis system, *DmA* difference must increase (decrease).

Assume inconsistency has been found between *a priori* assumed and *a posteriori* observed statistics of innovation or DmA difference.

- What can be done ?

or, equivalently

- Which bounds does the knowledge of the statistics of innovation put on the error statistics whose knowledge is required by the *BLUE* ?

Data assumed to consist of a vector z, belonging to data space D (dimD = m), in the form

 $z = \Gamma x + \zeta$

where Γ is a known (*mxn*)-matrix, and ζ an unknown 'error'

For instance

$$z = \begin{pmatrix} x^b = x + \zeta^b \\ y = Hx + \varepsilon \end{pmatrix}$$

which corresponds to

$$\Gamma = \begin{pmatrix} I_n \\ H \end{pmatrix} \qquad \qquad \zeta = \begin{pmatrix} \zeta^b \\ \varepsilon \end{pmatrix}$$

20

Look for estimated state vector x^a of the form

 $x^a = \alpha + Az$

subject to

- invariance in change of origin in state space $\Rightarrow A\Gamma = I_m$
- quadratic estimation error $E[(x_i^a x_i)^2]$ minimum for any component x_i .

Solution

 $\boldsymbol{x}^{a} = (\boldsymbol{\Gamma}^{\mathrm{T}} S^{-1} \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^{\mathrm{T}} S^{-1} [\boldsymbol{z} - \boldsymbol{\mu}]$ $P^{a} \equiv E[(\boldsymbol{x}^{a} - \boldsymbol{x}) (\boldsymbol{x}^{a} - \boldsymbol{x})^{\mathrm{T}}] = (\boldsymbol{\Gamma}^{\mathrm{T}} S^{-1} \boldsymbol{\Gamma})^{-1}$

where
$$\mu \equiv \zeta$$
, $S \equiv E(\zeta \zeta^{T})$, $\zeta \equiv \zeta - \mu$

Requires (at least apparently) *a priori* explicit knowledge of $E(\zeta)$ and $E(\zeta \zeta^T)$

Unambiguously defined iff rank $\Gamma = n$. Determinacy condition. Requires $m \ge n$. We shall set m = n + p.

Invariant in any invertible linear change of coordinates, either in data or state space.

In case ζ is gaussian, $\zeta = N [\mu, S]$, *BLUE* achieves bayesian estimation in the sense that $P(x \mid z) = N [x^a, P^a]$

If determinacy condition is verified, it is always possible to decompose data vector z into

 $x^{b} = x + \zeta^{b}$ $y = Hx + \varepsilon$ with $E(\zeta^{b}) = 0$; $E(\varepsilon) = 0$; $E(\varepsilon\zeta^{bT}) = 0$

 x^a is the same estimate (*BLUE*) as before, *viz.*,

$$x^{a} = x^{b} + P^{a} H^{T} R^{-1} (y - Hx^{b})$$

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

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

$$P^{a} = P^{b} - P^{b} H^{T} [HP^{b}H^{T} + R]^{-1} HP^{b}$$

Variational form.

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

 $J(\boldsymbol{\xi}) \equiv (1/2) \left[\boldsymbol{\Gamma}\boldsymbol{\xi} - (\boldsymbol{z} - \boldsymbol{\mu}) \right]^{\mathrm{T}} S^{-1} \left[\boldsymbol{\Gamma}\boldsymbol{\xi} - (\boldsymbol{z} - \boldsymbol{\mu}) \right]$

(Mahalanobis <u>S</u>-metric)

$$J(\xi) = (1/2) [\Gamma \xi - (z - \mu)]^{T} S^{-1} [\Gamma \xi - (z - \mu)]$$



Minimizing $J(\xi)$ amounts to

- unbias *z*
- project orthogonally onto space $\Gamma(S)$ according to Mahalanobis *S*-metric
- take inverse through
 (inverse unambiguously defined through determinacy condition)

Many assimilation methods

- Optimal Interpolation
- 3DVar, either primal or dual
- Kalman Filter, either in its simple or Extended form
- Kalman Smoother
- 4DVar, either in its strong- or weak-constraint form, or in its primal or dual form

are particular cases of that general scheme.

Only exceptions (so far)

- Ensemble Kalman Filter (which is linear however in its 'updating' phase)

- Particle Filters

Decompose data space D into image space $\Gamma(S)$ (index 1) and its S-orthogonal space $\perp \Gamma(S)$ (index 2)

$$\Gamma = \begin{pmatrix} \Gamma_1 \\ 0 \end{pmatrix} \qquad \Gamma_1 \text{ invertible} \qquad z = \begin{pmatrix} z_1 = \Gamma_1 x + \zeta_1 \\ z_2 = \zeta_2 \end{pmatrix}$$
Assume
$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

Then

$$\boldsymbol{x}^{a} = \boldsymbol{\Gamma}_{1}^{-1} \left[\boldsymbol{z}_{1} - \boldsymbol{\mu}_{1} \right]$$

$$\boldsymbol{x}^{a} = \boldsymbol{\Gamma}_{1}^{-1} \left[\boldsymbol{z}_{1} - \boldsymbol{\mu}_{1} \right]$$

The probability distribution of the error

 $x^{a} - x = \Gamma_{1}^{-1} [\zeta_{1} - \mu_{1}]$

depends on the probability distribution of ζ_1 .

On the other hand, the probability distribution of

$$\boldsymbol{\delta} = (\boldsymbol{z} - \boldsymbol{\mu}) - \boldsymbol{\Gamma} \boldsymbol{x}^a = \begin{pmatrix} 0 \\ \zeta_2 - \boldsymbol{\mu}_2 \end{pmatrix}$$

depends only on the probability distribution of ζ_2 .

- *DmA* difference, *i. e.* $(z-\mu) \Gamma x^a$, is in effect 'rejected' by the assimilation. Its expectation and covariance are irrelevant for the assimilation.
- Consequence. Any assimilation scheme (*i. e.*, *a priori* subtracted bias and gain matrix *K*) is compatible with any observed statistics of either *DmA* or innovation. Not only is not consistency between *a priori* assumed and *a posteriori* observed statistics of innovation (or *DmA*) sufficient for optimality of an assimilation scheme, it is not even necessary.

Example

 $z_1 = x + \zeta_1$ $z_2 = x + \zeta_2$

Errors ζ_1 and ζ_2 assumed to be centred ($E(\zeta_1) = E(\zeta_2) = 0$), to have same variance *s* and to be mutually uncorrelated. Then

 $x^a = (1/2) (z_1 + z_2)$

with expected quadratic estimation error

 $E[(x^a - x)^2] = s/2$

Innovation is difference $z_1 - z_2$. With above hypotheses, one expects to observe

 $E(z_1 - z_2) = 0$; $E[(z_1 - z_2)^2] = 2s$

Assume one observes

 $E(z_1 - z_2) = b$; $E[(z_1 - z_2)^2] = b^2 + 2\gamma$

Inconsistency if $b \neq 0$ and/or $\gamma \neq s$

Inconsistency can always be resolved by assuming that

$$E(\zeta_{1}) = -E(\zeta_{2}) = -b/2$$
$$E(\zeta_{1}^{2}) = E(\zeta_{2}^{2}) = (s+\gamma)/2$$
$$E(\zeta_{1}^{2}\zeta_{2}) = (s-\gamma)/2$$

This alters neither the *BLUE* x^a , nor the corresponding quadratic estimation error $E[(x^a-x)^2]$.

Explanation. It is not necessary to know explicitly the complete expectation μ and covariance matrix S in order to perform the assimilation. It is necessary to know the projection of μ and S onto the subspace $\Gamma(S)$. As for the subspace that is *S*-orthogonal to $\Gamma(S)$, it suffices to know what it is, but it is not necessary to know the projection of μ and S onto it. A number of degrees of freedom are therefore useless for the assimilation. The parameters determined by the statistics of d are equal in number to those useless degrees of freedom, to which any inconsistency between *a priori* and *a posteriori* statistics of the innovation can always mathematically be attributed.

However it may be that resolving the inconsistency in that way requires conditions that are (independently) known to be very unlikely, if not simply impossible. For instance, in the above example, consistency when $\gamma \neq s$ requires the errors ζ_1 and ζ_2 to be mutually correlated, which may be known to be very unlikely.

$$J(\xi) = (1/2) [\Gamma \xi - (z - \mu)]^{T} S^{-1} [\Gamma \xi - (z - \mu)]$$



That result, which is purely mathematical, means that the specification of the error statistics required by the assimilation must always be based, in the last resort, on external hypotheses, *i. e.* on hypotheses that cannot be validated on the basis of the innovation alone. Now, such knowledge always exists.

Problem. Identify hypotheses

- That will not be questioned (errors on observation perfomed a long distance apart by radiosondes made by different manufacturers are uncorrelated)
- That sound reasonable, but may be questioned (observation and background errors are uncorrelated)
- That are undoubtedly questionable (model errors are negligible)

Ideally, define a minimum set of hypotheses such that all remaining undetermined error statistics can be objectively determined from observed statistics of innovation.

Objective function

 $J(\xi) \equiv (1/2) [\Gamma \xi - z]^{T} S^{-1} [\Gamma \xi - z]$ $J_{min} \equiv J(x^{a}) = (1/2) [\Gamma x^{a} - z]^{T} S^{-1} [\Gamma x^{a} - z]$ $= (1/2) d^{T} [E(dd^{T})]^{-1} d$ $\Rightarrow E(J_{min}) = p/2 \qquad (p = \dim y = \dim d)$

If *p* is large, a few realizations are sufficient for determining $E(J_{min})$ Often called χ^2 criterion.

Remark. If in addition errors are gaussian $Var(J_{min}) = p/2$

Results for ECMWF (January 2003, $n = 8 \ 10^6$)

- Operations ($p = 1.4 \ 10^6$, has almost doubled since then)

 $2J_{min}/p = 0.40 - 0.45$

Innovation is significantly smaller than implied by P^b and R (a residual bias in d would make J_{min} too large).

- Assimilation without satellite observations ($p = 2 - 3 \ 10^5$)

 $2J_{min}/p = 1. - 1.05$

Similar results obtained at other NWP centres (C. Fischer, W. Sadiki with Aladin model, T. Payne at Meteorological Office, UK).

Probable explanation: error variance of satellite observations overestimated in order to compensate for ignored spatial correlation.

Informative content

Objective function

 $J(\boldsymbol{\xi}) = \Sigma_k J_k(\boldsymbol{\xi})$

where

$$\boldsymbol{J}_{k}(\boldsymbol{\xi}) \equiv (1/2) (\boldsymbol{H}_{k}\boldsymbol{\xi} - \boldsymbol{y}_{k})^{\mathrm{T}} \boldsymbol{S}_{k}^{-1} (\boldsymbol{H}_{k}\boldsymbol{\xi} - \boldsymbol{y}_{k})$$

with $\dim y_k = m_k$

Accuracy of analysis

 $P^a = (\Gamma^{\mathrm{T}} S^{-1} \Gamma)^{-1}$

$$[P^{a}]^{-1} = \Sigma_{k} H_{k}^{T} S_{k}^{-1} H_{k}$$

$$1 = (1/n) \Sigma_{k} \operatorname{tr}(P^{a} H_{k}^{T} S_{k}^{-1} H_{k})$$

$$= (1/n) \Sigma_{k} \operatorname{tr}(S_{k}^{-1/2} H_{k} P^{a} H_{k}^{T} S_{k}^{-1/2})$$

Informative content (continuation 1)

$(1/n) \Sigma_k \operatorname{tr}(S_k^{-1/2} H_k P^a H_k^{\mathrm{T}} S_k^{-1/2}) = 1$

 $I(y_k) \equiv (1/n) \operatorname{tr}(S_k^{-1/2} H_k P^a H_k^T S_k^{-1/2})$ is a measure of the relative contribution of subset of data y_k to overall accuracy of assimilation. Invariant in linear change of coordinates in data space \Rightarrow valid for *any* subset of data.

In particular

$$I(\mathbf{x}^b) = (1/n) \operatorname{tr}[P^a(P^b)^{-1}] = 1 - (1/n) \operatorname{tr}(KH)$$

$$I(\mathbf{y}) = (1/n) \operatorname{tr}(KH)$$

Rodgers, 2000, calls those quantities *Degrees of Freedom for Signal*, or *for Noise*, depending on whether considered subset belongs to 'observations' or 'background'.



Informative content of subsets of observations (Arpège Assimilation System, Météo-France)

Chapnik et al., 2006, QJRMS, 132, 543-565

QuickTime™ et un décompresseur TIFF (LZW) sont requis pour visionner cette image.

Informative content per individual (scalar) observation (courtesy B. Chapnik)

Objective function

$$J(\boldsymbol{\xi}) = \Sigma_k J_k(\boldsymbol{\xi})$$

where

$$J_{k}(\boldsymbol{\xi}) \equiv (1/2) (H_{k}\boldsymbol{\xi} - \boldsymbol{y}_{k})^{\mathrm{T}} S_{k}^{-1} (H_{k}\boldsymbol{\xi} - \boldsymbol{y}_{k})$$

with $\dim y_k = m_k$

For a perfectly consistent system

$$E[J_k(x^a)] = (1/2) [m_k - tr(S_k^{-1/2} H_k P^a H_k^T S_k^{-1/2})]$$

(in particular, $E(J_{min}) = p/2$)

For same vector dimension m_k , more informative data subsets lead at the minimum to smaller terms in the objective function.

Equality

$$E[\mathbf{J}_{k}(x^{a})] = (1/2) [m_{k} - \operatorname{tr}(S_{k}^{-1/2} H_{k} P^{a} H_{k}^{T} S_{k}^{-1/2})]$$
(1)

can be objectively checked.

Chapnik *et al.* (2004, 2005). Multiply each observation error covariance matrix S_k by a coefficient α_k such that (1) is verified simultaneously for all observation types.

System of equations for the α_k 's solved iteratively.



Chapnik *et al.*, 2006, *QJRMS*, **132**, 543-565

Figure 9. Difference between tunned rms (tuned geopotential forecasts - geopotential TEMP observations) and the operational rms computed over 21 situations. the x axis is the forecast term and the y axis is the vertical pressure level. Dashed lines mean that the tuned forecast is further from the observations than the operational one (degradation), on the contrary the solid lines mean that the tuned forecast is better than the operational. the difference between two colored line is 1 m. Subpanel a is for the northern hemisphere, subpanel b for the southern

Informative content (continuation 2)

$$I(\mathbf{y}_k) \equiv (1/n) \operatorname{tr}(S_k^{-1/2} H_k P^a H_k^{\mathrm{T}} S_k^{-1/2})$$

Two subsets of data z_1 and z_2

If errors affecting z_1 and z_2 are uncorrelated, then $I(z_1 \cup z_2) = I(z_1) + I(z_2)$

If errors are correlated

 $I(z_1 \cup z_2) \neq I(z_1) + I(z_2)$

Informative content (continuation 3)

Example 1

 $z_1 = x + \zeta_1$ $z_2 = x + \zeta_2$

Errors ζ_1 and ζ_2 assumed to centred, to have same variance and correlation coefficient c.

 $I(z_1) = I(z_2) = (1/2)(1+c)$

Example 2

State vector \mathbf{x} evolving in time according to

 $x_2 = \alpha x_1$

Observations are performed at times 1 and 2. Observation errors are assumed centred, uncorrelated and with same variance. Information contents are then in ratio $(1/\alpha, \alpha)$. For an unstable system $(\alpha > 1)$, later observation contains more information (and the opposite for a stable system).

Informative content (continuation 4)

- Subset u_1 of analyzed fields, $dimu_1 = n_1$. Define relative contribution of subset y_k of data to accuracy of u_1 ?
- u_2 : component of x orthogonal to u_1 with respect to Mahalanobis norm associated with P^a (analysis errors on u_1 and u_2 are uncorrelated).

 $\boldsymbol{x} = (\boldsymbol{u}_1^{\mathrm{T}}, \boldsymbol{u}_2^{\mathrm{T}})^{\mathrm{T}}$. In basis $(\boldsymbol{u}_1, \boldsymbol{u}_2)$

$$P^a = \begin{pmatrix} P^a_1 & 0 \\ 0 & P^a_2 \end{pmatrix}$$

Informative content (continuation 5)

Observation operator H_k decomposes into

 $H_k = (H_{k1}, H_{k2})$

and expression of estimation error covariance matrix into

 $[P_{1}^{a}]^{-1} = \Sigma_{k} H_{k1}^{T} S_{k}^{-1} H_{k1}$ $[P_{2}^{a}]^{-1} = \Sigma_{k} H_{k2}^{T} S_{k}^{-1} H_{k2}$

Same development as before shows that the quantity

 $(1/n_1) \operatorname{tr}(S_k^{-1/2} H_{k1} P_1^a H_{k1}^T S_k^{-1/2})$

is a measure of the relative contribution of subset y_k of data to analysis of subset u_1 of state vector.

But can it be computed in practice for large dimension systems (requires the explicit decomposition $\mathbf{x} = (\mathbf{u}_1^T, \mathbf{u}_2^T)^T$)?

Other possible diagnostics (Desroziers et al., 2006, submitted to QJRMS)

For a consistent system

 $E[H(x^{a}-x^{b})(y-Hx^{b})^{T}] = E[H(x^{a}-x^{b})d^{T}] = HP^{b}H^{T}$ $E[(y-Hx^{a})(y-Hx^{b})^{T}] = E[(y-Hx^{a})d^{T}] = R$

Optimality

Equation

$$\boldsymbol{x}^{a} = \boldsymbol{x}^{b} - E(\boldsymbol{\zeta}^{b}\boldsymbol{d}^{\mathrm{T}}) [E(\boldsymbol{d}\boldsymbol{d}^{\mathrm{T}})]^{-1} (\boldsymbol{y} - H\boldsymbol{x}^{b})$$

means that estimation error $x - x^a$ is uncorrelated with innovation $y - Hx^b$ (if it was not, it would be possible to improve on x^a by statistical linear estimation).

Independent unbiased observation

 $v = Cx + \gamma$

Fit to analysis

$$\boldsymbol{v} - C\boldsymbol{x}^a = C(\boldsymbol{x} - \boldsymbol{x}^a) + \boldsymbol{\gamma}$$

$$E[(\boldsymbol{v} - \boldsymbol{C}\boldsymbol{x}^{a}) \, \boldsymbol{d}^{\mathrm{T}}] = \boldsymbol{C}E[(\boldsymbol{x} - \boldsymbol{x}^{a}) \, \boldsymbol{d}^{\mathrm{T}}] + E(\boldsymbol{\gamma}\boldsymbol{d}^{\mathrm{T}})$$

First term is 0 if analysis is optimal, second is 0 if observation v is independent from previous data.

Daley (1992)

Conclusions

Absolute evaluation of analysis schemes, and comparison between different schemes

Can be evaluated only against independent unbiased data (independence and unbiasedness cannot be objectively checked). Fundamental, but not much to say.

Determination of required statistics

Impossible to achieve in a purely objective way. Will always require physical knowledge, educated guess, interaction with instrumentalists and modelers, and the like.

Inconsistencies in specification of statistics can be objectively diagnosed, and can help in improving assimilation.

For given error statistics, possible to quantify relative contribution of each subset of data to analysis of each subset of state vector.

(and also Generalized Cross-Validation, Adaptive Filtering)

Optimality of analysis schemes

Optimality in the sense of least error variance can be objectively checked against independent unbiased data. 51

Sequential assimilation cannot be optimal if errors are correlated in time.

$$z_1 = x + \zeta_1$$

$$z_2 = x + \zeta_2$$

$$E(\zeta_1) = E(\zeta_2) = 0 \quad ; \quad E(\zeta_1^2) = E(\zeta_2^2) = s \quad ; \quad E(\zeta_1\zeta_2) = 0$$

$$BLUE \text{ of } x \text{ from } z_1 \text{ and } z_2 \text{ gives equal weights to } z_1 \text{ and } z_2.$$

Additional observation then becomes available

$$z_3 = x + \zeta_3$$

 $E(\zeta_3) = 0$; $E(\zeta_3^2) = s$; $E(\zeta_1\zeta_3) = cs$; $E(\zeta_2\zeta_3) = 0$

BLUE of *x* from (z_1, z_2, z_3) has weights in the proportion (1, 1+c, 1)

Time-correlated Errors (continuation 1)

Example of time-correlated model errors

Evolution equation

 $x_{k+1} = x_k + \eta_k \qquad \qquad E(\eta_k^2) = q$

Observations

`

 $y_k = x_k + \varepsilon_k$, k = 0, 1, 2 $E(\varepsilon_k^2) = r$, errors uncorrelated in time

Sequential assimilation. Weights given to y_0 and y_1 in analysis at time 1 are in the ratio r/(r+q). That ratio will be conserved in sequential assimilation. All right if model errors are uncorrelated in time.

Assume $E(\eta_0 \eta_1) = cq$

Weights given to y_0 and y_1 in estimation of x_2 are in the ratio

$$\rho = \frac{r - qc}{r + q + qc}$$

Time-correlated Errors (continuation 2)

Moral. If data errors are correlated in time, it is not possible to discard observations as they are used while preserving optimality of the estimation process. In particular, if model error is correlated in time, all observations are liable to be reweighted as assimilation proceeds.

Variational assimilation can take time-correlated errors into account.

Example of time-correlated observation errors. Global covariance matrix

 $\boldsymbol{R} = (\boldsymbol{R}_{kk'} = \boldsymbol{E}(\boldsymbol{\varepsilon}_k \boldsymbol{\varepsilon}_{k'}^{\mathrm{T}}))$

Objective function

 $\begin{aligned} \xi_0 \in \ S \ \to \\ \mathsf{J}(\xi_0) \ &= \ (1/2) \ (x_0^{\ b} - \xi_0)^{\mathrm{T}} \ [P_0^{\ b}]^{-1} \ (x_0^{\ b} - \xi_0) + (1/2) \ \Sigma_{kk'} [y_k - H_k \xi_k]^{\mathrm{T}} \ [\mathbf{R}^{-1}]_{kk'} \ [y_{k'} - H_{k'} \xi_{k'}] \end{aligned}$

where $[\mathbf{R}^{-1}]_{kk'}$ is the kk'-sub-block of global inverse matrix \mathbf{R}^{-1} .

Similar approach for time-correlated model error.

Time-correlated Errors (continuation 3)

Time correlation of observational error has been introduced by ECMWF (Järvinen *et al.*, 1999) in variational assimilation of high-frequency surface pressure observations (correlation originates in that case in representativeness error).

Identification and quantification of temporal correlation of errors, especially model errors?