Cross-validation methods for quality control, cloud screening, etc.

Olaf Stiller, Deutscher Wetterdienst

Are observations consistent

- with the other observations?

given the
- background
- assumed error covariances
- observation operator

Which observations are affected by the cloud???
Diagnose clouds from the observations  [ i.e., from \((\text{obs-fg})\) ]

1. Look whether a FoV is cloudy: \((\text{obs-fg})\) threshold
2. Find upper edge of cloud : gradient criterion
Cross-validation

Diagnose clouds from the observations [i.e., from (obs-fg)]

Question: Can we do this more systematically?

Aim: Identify observations which are not consistent with ……

Are observations consistent

- with the other observations?

given the
- background
- assumed error covariances
- observation operator

Which observations are affected by the cloud???
**Assumed uncertainties in data assimilation**

Assumptions about Obs and FG errors:

\[ J(x) = \frac{1}{2} \left[ x^T B^{-1} x + (y^o - Hx)^T R^{-1} (y^o - Hx) \right] \]

Are FG departures consistent with these assumptions?

- \( y^o = Y^o - Y^b \)
  obs - first guess

\[ \langle (y^o)^T y^o \rangle = H^T B H + R \]

Checking diagonal:

\[ \langle (y^o_k)^2 \rangle = [H^T B H + R]_{kk} = \sigma_k^2 \]

Conditional probability of the observations \( y^o_k \) (given the background):

\[ P(y^o_k | X^b) \propto \exp \left( -\frac{1}{2} \left( \frac{y^o_k}{\sigma_k} \right)^2 \right) \]

Cross-Validation with background (standard Quality Control check):

\[ n \text{ sigma check: } \left| \frac{y^o_k}{\sigma_k} \right| < n \]
**Assumed uncertainties in data assimilation**

**Assumptions about Obs and FG errors:**

\[
J(x) = \frac{1}{2} \left[ x^T B^{-1} x + (y^o - Hx)^T R^{-1} (y^o - Hx) \right]
\]

Are FG departures consistent with these assumptions?

- \( y^o = Y^o - Y^b \)
  - obs - first guess

\[
\langle (y^o)^T y^o \rangle = H^T B H + R
\]

**Checking diagonal:**

\[
\langle (y_k^o)^2 \rangle = \left[ H^T B H + R \right]_{kk} = \sigma_k^2
\]

Conditional probability of the observations \( y_k^o \) (given the background):

\[
P(y_k^o | X^b) \propto \exp \left( -\frac{1}{2} \left( \frac{y_k^o}{\sigma_k} \right)^2 \right)
\]

**Decompose observations:** \( \{ y^o_\tau^C, y^o_\tau \} \)

Conditional probability of observations \( y_\tau^o \)
(given the background and observations \( y_\tau^o_\tau^C \)):

\[
P(y_\tau^o | y_\tau^o_\tau^C, X^b) \propto \exp \left( -\frac{1}{2} \left\{ (y_\tau^o - \bar{y}_\tau)^T D_\tau (y_\tau^o - \bar{y}_\tau) \right\} \right)
\]
Special case:
Observations can be ordered

\[ P(y_k | y_{l<k}, x^b) = N^{-1} \exp \left( -\frac{1}{2} \{ Y_k \}^2 \right) \]

\[ Y = (T_l^T)^{-1} y \]

Cholesky decomposition:

\[
T_l T_U = \begin{pmatrix}
t_{11} & 0 & \cdots & 0 \\
t_{21} & t_{22} & \cdots & 0 \\
t_{k1} & t_{k2} & \cdots & 0 \\
t_{p1} & \cdots & 0 & t_{pp}
\end{pmatrix} \begin{pmatrix}
t_{11} & t_{21} & \cdots & t_{p1} \\
t_{21} & t_{22} & \cdots & t_{p2} \\
t_{k1} & t_{k2} & \cdots & t_{kk} \\
t_{p1} & \cdots & 0 & t_{pp}
\end{pmatrix} = [R + HBH^T]
\]

\[ Y_k = \frac{y_k - y_k^{*}}{\sqrt{\epsilon_k^{obs} + \epsilon_k^{a*}}} \]

analysis considering only obs \( y_l \) with \( l < k \)

error of this analysis
**Application: IASI cloud screening**

(see also 8P.05)

**Problem:** Standard deviation dominated by obs error
Single observation not sensitive enough

**Need to detect systematic perturbations**

Consider joint probability:

\[
P(y_k, y_{k+1}, \ldots, y_{k+s} | y_{\{l<k\}}, x^l) \propto \exp \left( - \frac{1}{2} \left( Y_k^2 + Y_{k+1}^2 + \ldots + Y_{k+s}^2 \right) \right)
\]

\[
Y_k^n = \sum_{j=k}^{k+n} Y_j / \sqrt{n}
\]

is also: stochastic variable with variance 1

**Generalization (for any vector \( \vec{h}_l \)):**

\[
\overrightarrow{\vec{Y}} \rightarrow \overrightarrow{\vec{Y}}_l \equiv \frac{\vec{h}_l \ast \overrightarrow{\vec{Y}}}{||\vec{h}_l||}
\]

stochastic variable with variance 1

**Targeted approach:** project on most relevant directions \( \vec{h}_l \)
Application: IASI cloud screening

Project on $H_{cfr}$

Generalization (for any vector $\vec{h}_l$):

$$\vec{Y} \to \vec{Y}_l \equiv \frac{\vec{h}_l \cdot \vec{Y}}{\|\vec{h}_l\|}$$

stochastic variable with variance 1

Let: $c_{fr}$ be a model state variable for cloud fraction in a layer corresponding part of observation operator matrix

$$H_F = \begin{pmatrix} H & H_{cfr} \\ \end{pmatrix} \quad B_T = \begin{pmatrix} B \\ \vdots \\ \vdots \\ \vdots \\ 0 \\ \vdots \\ \vdots \\ \sigma_{cfr} \end{pmatrix}$$

Then, in the limit of large $\sigma_{cfr}$, one finds:

$$c_{fr}^a \rightarrow \left[ H_{cfr}^T \left[ R + H B H^T \right]^{-1} H_{cfr} \right]^{-1} H_{cfr}^T \left[ R + H B H^T \right]^{-1} \left[ y^o - H(\vec{x}_c^h) \right] = \frac{h^T \vec{Y}}{h^T h}$$

cloud fraction in layer $k$

$$\frac{c_{fr}^a[k]}{\sqrt{\left( c_{fr}^a[k] \right)^2}} = \frac{\vec{h}_k \vec{Y}}{\|\vec{h}_k\|}$$

stochastic variable with variance 1

$\mathbf{Y} = T_{L}^{-1} \left[ y^o - H(\vec{x}_c^h) \right]$

$h = T_{L}^{-1} H_{cfr}$
Discussion

- **A cross validation method** for observations has been developed which

  - works within the probabilistic framework of the DA system: \( H^T B H + R \)
    - **disadvantage**: employed error matrixes are far from perfect
    - **advantage**: method will develop and improve systematically with improved DA systems

  - is **cheap enough** to be run in **preprocessing step**
  - requires that observation operators sufficiently overlap
    - good for IASI

- Diagnostics have to be taylored for systematic perturbations

  - project on relevant directions \( \hat{h} \)
    - employed error matrixes are (*probably*) not good enough to flag more generally perturbed observations

- **Which influences can be diagnosed from obs-fg increments?**

  - impact has to be generally strong (scale separation – weak signal must be rare)
  - \( ||\hat{h}|| \) must be large for typical signal
    - very low clouds can not be detected from IASI radiances
The cross validation method

- is planned to be run as a preprocessing system
  - flagging of bad observation **before** they enter into the analysis

- possibly within a **1D Var** preprocessing step
  (important for strongly nonlinear observation as, e.g.,
  the water vapor channels of IASI)

- will profit from improved \( B \) matrix from Ensemble Kalman Filter

The cross validation method may be useful for testing also other influences

- which the observation operator does not represent properly
  - like, e.g., surface emissivity

CV diagnostics good for comparing compatibility of different observation types

- collecting statistics of targeted diagnostics
Thank you for listening