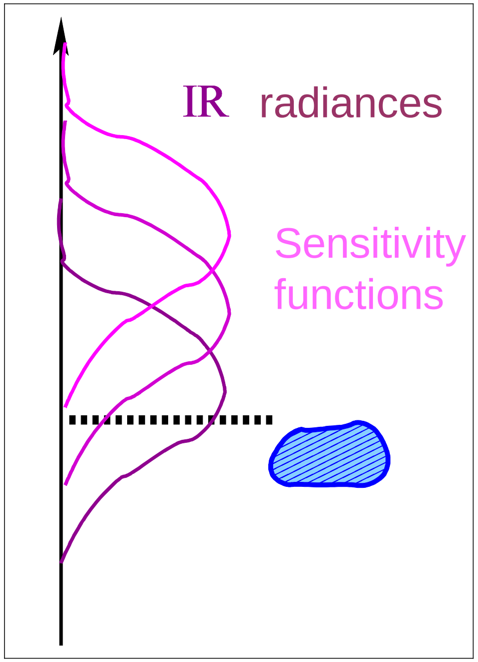
Cross-validation methods for quality control, cloud screening, etc.

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???

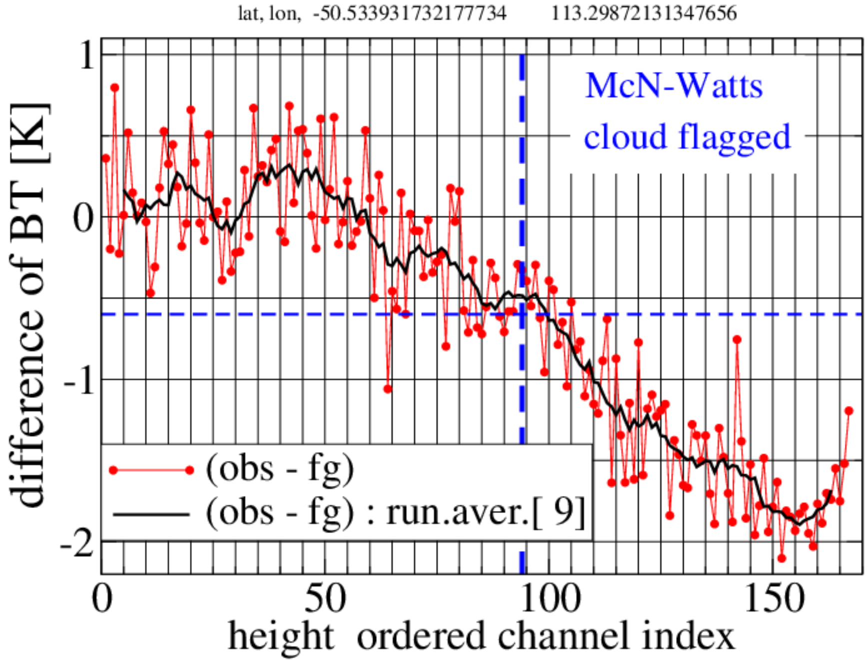
Olaf.Stiller@dwd.de

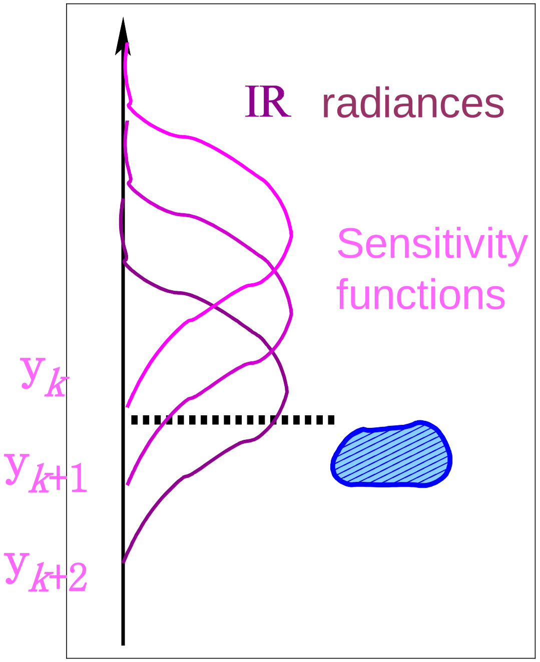
Inspiration:

McNally&Watts scheme

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

1.Look whether a FoV is cloudy: (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?

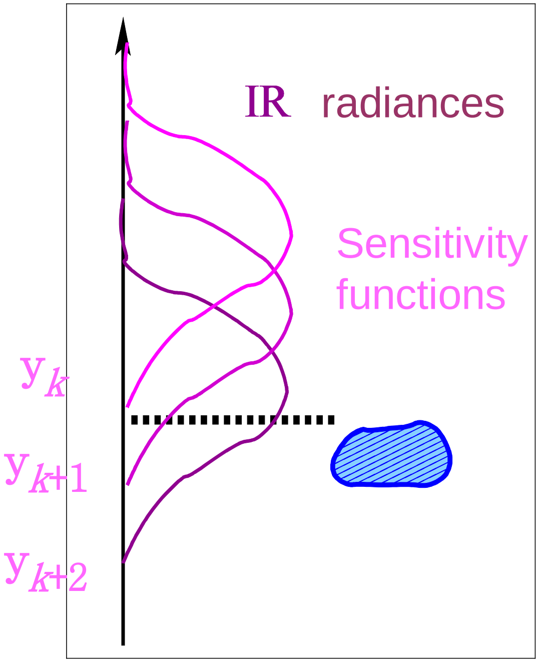
Aim : Identify observations which are not consistent with ……

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

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

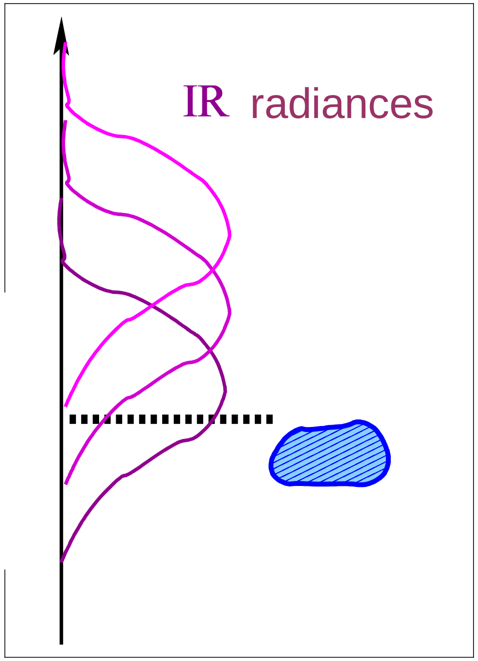
Assumptions about Obs and FG errors:

$$J(\mathbf{x}) \;=\; \frac{1}{2}\left[\mathbf{x}^T\mathbf{B}^{-1}\mathbf{x} + (\mathbf{y}^o - \mathbf{H}\mathbf{x})^T\,\mathbf{R}^{-1}\,(\mathbf{y}^o - \mathbf{H}\mathbf{x})\right]$$

Are FG departures consistent with these assumptions?

- $\mathbf{y}^o = \mathbf{Y}^o - \mathbf{Y}^b$
  obs - first guess

$$\boxed{\left\langle (\mathbf{y}^o)^T\,\mathbf{y}^o \right\rangle = \quad \mathbf{H}^T\mathbf{B}\mathbf{H} + \mathbf{R}}$$



Checking diagonal:

$$\boxed{\left\langle \left(\mathbf{y}_k^o\right)^2 \right\rangle \;=\; \left[\mathbf{H}^T\mathbf{B}\mathbf{H} + \mathbf{R}\right]_{kk} = \sigma_k^2}$$

Conditional probability of the observations $\mathbf{y}_k^o$ (given the background):

$$P(\mathbf{y}_k^o|\mathbf{X}^b) \;\propto\; \exp -\frac{1}{2}\left(\frac{\mathbf{y}_k^o}{\sigma_k}\right)^2$$

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

$$\boxed{n \text{ sigma check:} \qquad \left|\frac{\mathbf{y}_k^o}{\sigma_k}\right| < n}$$

Assumed uncertainties in data assimilation

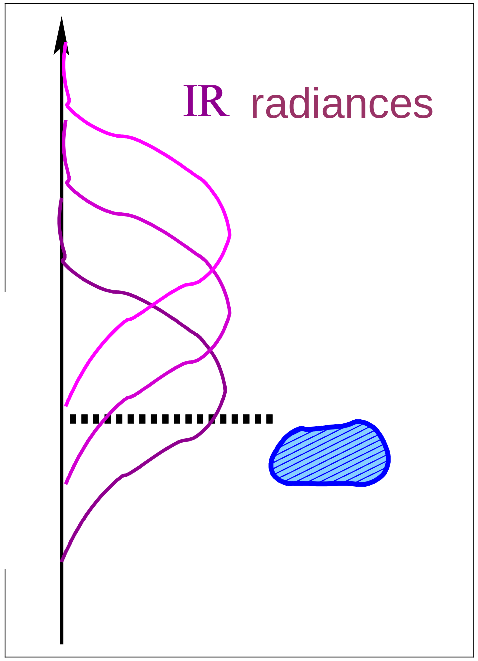
Assumptions about Obs and FG errors:

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

Are FG departures consistent with these assumptions?

- $\mathbf{y}^o = \mathbf{Y}^o - \mathbf{Y}^b$
  obs - first guess

$$\boxed{\langle (\mathbf{y}^o)^T \mathbf{y}^o \rangle = \mathbf{H}^T\mathbf{B}\mathbf{H} + \mathbf{R}}$$



Checking diagonal:

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

Conditional probability of the observations $\mathbf{y}^o_k$ (given the background):

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

Decompose observations:

$$\{\mathbf{y}^o_{\tau^C}, \mathbf{y}^o_\tau\}$$

Conditional probability of observations $\mathbf{y}^o_\tau$
(given the background and observations $\mathbf{y}^o_{\tau^C}$):

$$P(\mathbf{y}^o_\tau | \mathbf{y}^o_{\tau^C}, \mathbf{X}^b) \propto \exp -\frac{1}{2}\left\{(\mathbf{y}^o_\tau - \overline{\mathbf{y}_\tau})^T \mathbf{D}_\tau (\mathbf{y}^o_\tau - \overline{\mathbf{y}_\tau})\right\}$$

Special case: Observations can be ordered

$$P(\mathbf{y}_k|\mathbf{y}_{\{l<k\}},\mathbf{x}^b) \;=\; \mathbb{N}^{-1}\exp-\frac{1}{2}\left\{\mathbf{Y}_k{}^2\right\}$$

$$\mathbf{Y} \;=\mathbf{T}_l^{-1}\mathbf{y}$$

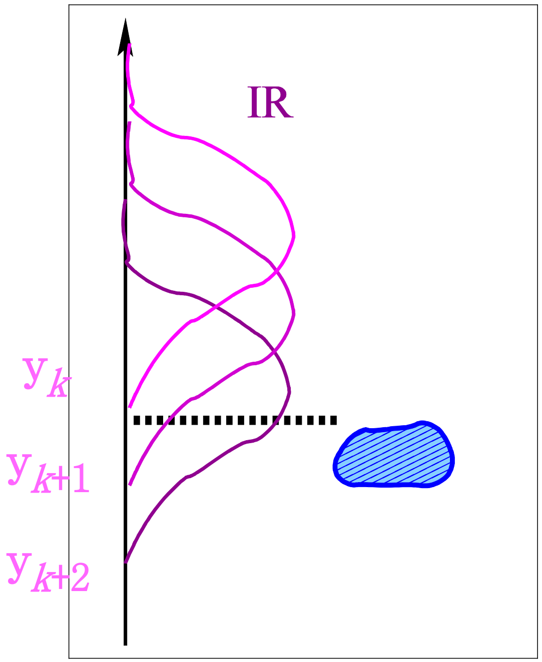
$$\mathbf{Y}_k \;=\; \frac{\mathbf{y}_k-\mathbf{y}_k^{a*}}{\sqrt{\epsilon_k^{obs}+\epsilon_k^{a*}}}$$

analysis considering only obs $\mathbf{y}_l$ with $l < k$

error of this analysis

Cholesky decomposition:

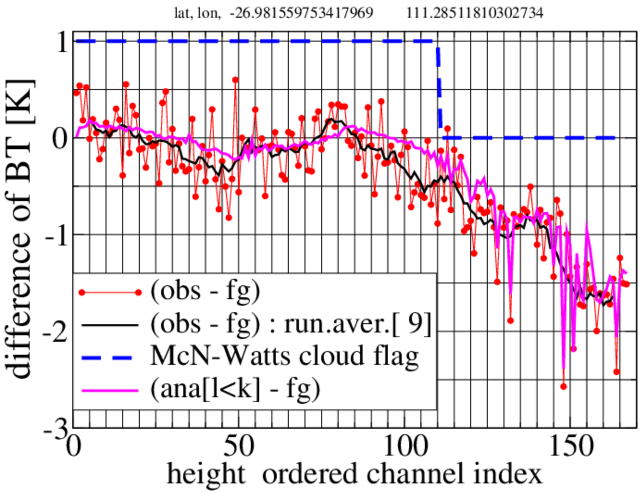
$$\mathbf{T}_L\mathbf{T}_U \;=\; \begin{pmatrix} t_{11} & 0 & & & & & \\ t_{21} & t_{22} & 0 & & & 0 & \\ & . & . & 0 & & & \\ t_{k1} & t_{k2} & & t_{kk} & 0 & & \\ & & & . & . & . & \\ & & & & & . & 0 \\ t_{p1} & & & & & & t_{pp} \end{pmatrix}\begin{pmatrix} t_{11} & t_{21} & & t_{k1} & & t_{p1} \\ 0 & t_{22} & & t_{k2} & & \\ & 0 & . & . & & \\ & & 0 & t_{kk} & & \\ & & & 0 & . & \\ & & & & . & . \\ & & & & 0 & t_{pp} \end{pmatrix} = \left[\mathbf{R}+\mathbf{HBH}^T\right]$$



Application: IASI cloud screening

(see also 8P.05)

Deutscher Wetterdienst
Wetter und Klima aus einer Hand



$$\mathbf{Y}_k = \frac{\mathbf{y}_k - \mathbf{y}_k^{a*}}{\sqrt{\epsilon_k^{obs} + \epsilon_k^{a*}}}$$

analysis considering only obs $\mathbf{y}_l$ with $l < k$

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

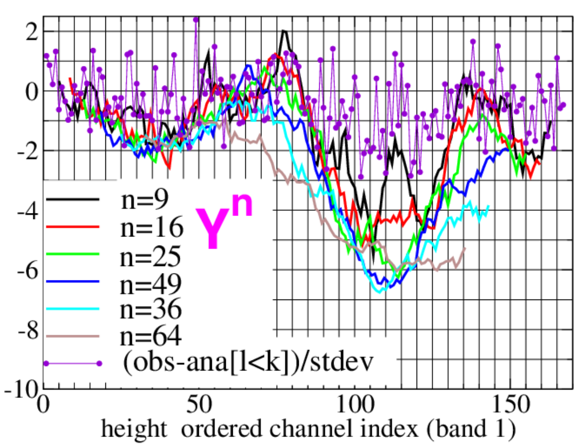
Need to detect systematic perturbations

Consider joint probability:

$$P(\mathbf{y}_k, \mathbf{y}_{k+1}, ., \mathbf{y}_{k+s} | \mathbf{y}_{\{l<k\}}, \mathbf{x}^b) \propto \exp -\frac{1}{2}\{\mathbf{Y}_k{}^2 + \mathbf{Y}_{k+1}{}^2 + . + \mathbf{Y}_{k+s}{}^2\}$$

$$\mathbf{Y}_k{}^n = \sum_{j=k}^{k+n} \mathbf{Y}_j / \sqrt{n}$$

is also: stochastic variable with variance 1



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

$$\vec{\mathbf{Y}} \rightarrow \tilde{\mathbf{Y}}_l \equiv \frac{\vec{h}_l * \vec{\mathbf{Y}}}{\|\vec{h}_l\|}$$

stochastic variable with variance 1

Targeted approach: project on most relevant directions $\vec{h}_l$

Application: IASI cloud screening

Project on $\mathbf{H}_{cfr}$

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

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

$$\vec{\mathbf{Y}} \rightarrow \tilde{\mathbf{Y}}_l \equiv \frac{\vec{h}_l * \vec{\mathbf{Y}}}{\left\| \vec{h}_l \right\|}$$

stochastic variable
with variance 1

Let: c_{fr} be a **model** state variable for **cloud fraction** in a layer
$\mathbf{H}_{cfr}$ corresponding part of **observation operator** matrix

$$\mathbf{H}_T = \left(\begin{array}{c|c} \mathbf{H} & \begin{array}{c} \cdot \\ \mathbf{H}_{cfr} \\ \cdot \end{array} \end{array} \right) \qquad \mathbf{B}_T = \left(\begin{array}{c|c} \mathbf{B} & \begin{array}{c} \cdot \\ 0 \\ \cdot \end{array} \\ \hline \cdot \quad 0 \quad \cdot & \sigma_{cfr} \end{array} \right)$$

Then, in the limit of large σ_{cfr}, one finds:

$$c^a_{fr} \rightarrow \left[\mathbf{H}^T_{cfr}\left[\mathbf{R}+\mathbf{HBH}^T\right]^{-1}\mathbf{H}_{cfr}\right]^{-1}\mathbf{H}^T_{cfr}\left[\mathbf{R}+\mathbf{HBH}^T\right]^{-1}\left[\mathbf{y}^o - H\left(\mathbf{x}^b_c\right)\right] = \frac{\mathbf{h}^T\mathbf{Y}}{\left[\mathbf{h}^T\mathbf{h}\right]}$$

cloud fraction
in layer k

$$\frac{c^a_{fr}[k]}{\sqrt{\left\langle \left(c^a_{fr}[k]\right)^2 \right\rangle}} = \frac{\vec{h}_k \vec{\mathbf{Y}}}{\left\| \vec{h}_k \right\|}$$

stochastic variable
with variance 1

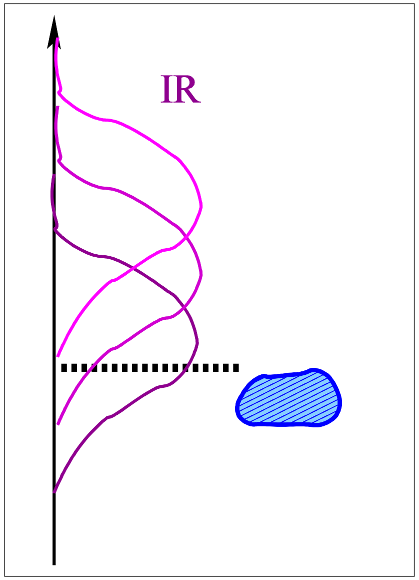
$$\mathbf{Y} = \mathbf{T}_L^{-1}\left[\mathbf{y}^o - H\left(\mathbf{x}^b_c\right)\right]$$

$$\mathbf{h} = \mathbf{T}_L^{-1}\mathbf{H}_{cfr}$$

Discussion

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

- A **cross validation method** for observations has been developed which
  - ➢ works within the probabilistic framework of the DA system: $\boxed{\mathbf{H}^T\mathbf{B}\mathbf{H} + \mathbf{R}}$
    - – **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



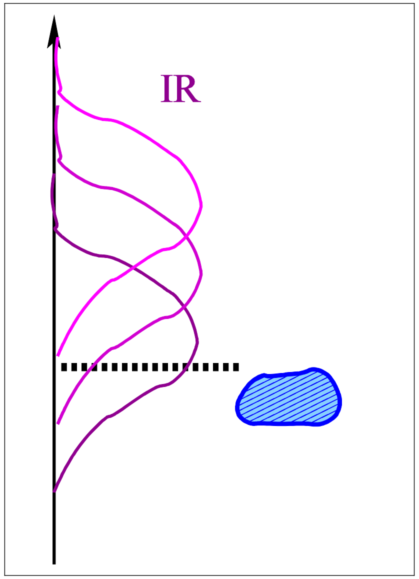
- Diagnostics have to be taylored for systematic perturbations
  - ➢ project on relevant directions $\vec{h}_l$
    - – employed error matrixes are (*probably*) not good enough to flag more generally perturbed observations

- Which influences can be diagnosed from obs-fg increments?
  - ➢ impact has to be generally strong (scale separation – weak signal must be rare)
  - ➢ $||\vec{h}_l||$ must be large for typical signal
    - – very low clouds can not be detected from IASI radiances

Outlook

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

- 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 compatability of different observation types
  - collecting statistics of targeted diagnostics

Thank you for listening

Olaf.Stiller@dwd.de