Influence matrix diagnostic to monitor the assimilation system

Carla Cardinali

Monitoring Assimilation System

- ECMWF 4D-Var system handles a large variety of space and surface-based observations. It combines observations and atmospheric state a priori information by using a linearized and non-linear forecast model

- Effective monitoring of a such complex system with 10^7 degree of freedom and 10^6 observations is a necessity. No just few indicators but a more complex set of measures to answer questions like
  - How much influent are the observations in the analysis?
  - How much influence is given to the a priori information?
  - How much the estimate depends on one single influential obs?

- Diagnostic methods are available for monitoring multiple regression analysis to provide protection against distortion by anomalous data

Influence Matrix: Introduction

- Unusual or influential data points not necessarily are bad observations but they may contain some of most interesting sample information

- In OLS quantitatively the information is available in the *Influence Matrix* which gives each fitted value $\hat{y}_i$ as a linear combination of y_i

$$\hat{\mathbf{y}} = \mathbf{S}\,\mathbf{y}$$

Hat Matrix

Leverage

Influence

Tuckey 63, Hoaglin and Welsch 78, Velleman and Welsch 81

Outline

- Influence matrix diagnostic in Ordinary Least Square
- Influence matrix application in data assimilation in NWP
- Influence matrix approximation
- Results and findings related to data influence and information content
- Conclusion

Influence Matrix in OLS

- The OLS regression model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Y (*mx1*) observation vector

X (*mxq*) predictors matrix, full rank q

$ (*qx1*) unknown parameters

, (*mx1*) error $E(\boldsymbol{\varepsilon}) = 0, Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$

- OLS provide the solution $\boldsymbol{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$

- The fitted response is

$$\hat{\mathbf{y}} = \mathbf{S}\mathbf{y} \qquad \mathbf{S} = \mathbf{X}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}^{T}$$

Influence Matrix Properties

$$\hat{\mathbf{y}} = \mathbf{S}\,\mathbf{y}$$

S (*mxm*) symmetric, idempotent and positive definite matrix

- **The diagonal element satisfy** $0 \le S_{ii} \le 1 \quad Tr(\mathbf{S}) = q$
- **It is seen**

$$\mathbf{S} = \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{y}}$$

Cross-Sensitivity

$$S_{ij} = \frac{\partial \hat{y}_i}{\partial y_j}$$

Self-Sensitivity

$$S_{ii} = \frac{\partial \hat{y}_i}{\partial y_i}$$

Average Self-Sensitivity=***q/m***

Influence Matrix Properties

$$\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$$

- Error covariance in ŷ and covariance of the residual r=y- ŷ are related

$$\mathrm{var}(\hat{\mathbf{y}}) = \sigma^2 \mathbf{S}$$

$$\mathrm{var}(\mathbf{r}) = \sigma^2 (\mathbf{I} - \mathbf{S})$$

- The change in estimate occurring when the *i*-th observation is deleted

$$\hat{y}_i - \hat{y}_i^{(-i)} = \frac{S_{ii}}{(1 - S_{ii})} r_i$$

Influence Matrix Properties

$$Tr(\mathbf{S}) = \sum_{i=1}^{m} S_{ii} = q$$

- The trace of S is the amount of *information extracted* from the observations
- A related result is with the leaving-out-one Cross Validation score

$$\sum_{i=1}^{m} (y_i - \hat{y}_i^{(-i)})^2 = \sum_{i=1}^{m} \frac{(\hat{y}_i - y_i)^2}{(1 - S_{ii})^2}$$

  - CV score can be computed knowing ŷ and S_{ii} without performing the *m* separate LS regression on the leaving-out-one samples

- In non parametric statistics Tr(S) measure the *degrees of freedom for signal*

Influence Matrix and Self-sensitivity

- Definition of *Influence Matrix* $\mathbf{S} = \dfrac{\partial \hat{\mathbf{y}}}{\partial \mathbf{y}}$ and *Self-sensitivity* $S_{ii} = \dfrac{\partial \hat{y}_i}{\partial y_i}$

are general and can be applied to non-linear prediction problems.

Interpretation remain the same as in LS and most the results as the CV

leaving-out-one theorem still apply

Analysis Solution

- The BLUEstimate of x given y and x_b in the LS sense

$$\mathbf{x}_a = \mathbf{K}\mathbf{y} + (\mathbf{I}_q - \mathbf{K}\mathbf{H})\mathbf{x}_b$$

4D-Var analysis is the solution of an appropriate Generalized LS minimization problem

$$\mathbf{K} = (\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{R}^{-1}$$

K: (q*x*p) gain matrix

H: (p*x*q) Jacobian matrix

B = Var(x_b): (qxq)

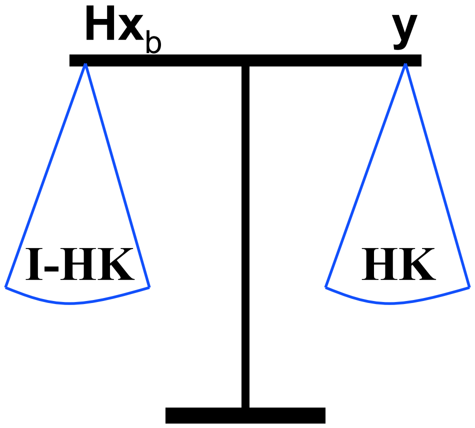
R = Var(y): (pxp)

Solution in the Observation Space

- The analysis projected at the observation location

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{x}_a = \mathbf{H}\mathbf{K}\mathbf{y} + (\mathbf{I}_p - \mathbf{H}\mathbf{K})\mathbf{H}\mathbf{x}_b$$

The estimation ŷ is a weighted mean



ECMWF

Influence Matrix

$$\hat{\mathbf{y}} = \mathbf{HKy} + (\mathbf{I}_p - \mathbf{HK})\mathbf{Hx}_b$$

$$\mathbf{S} = \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{y}} = \mathbf{K}^T\mathbf{H}^T$$

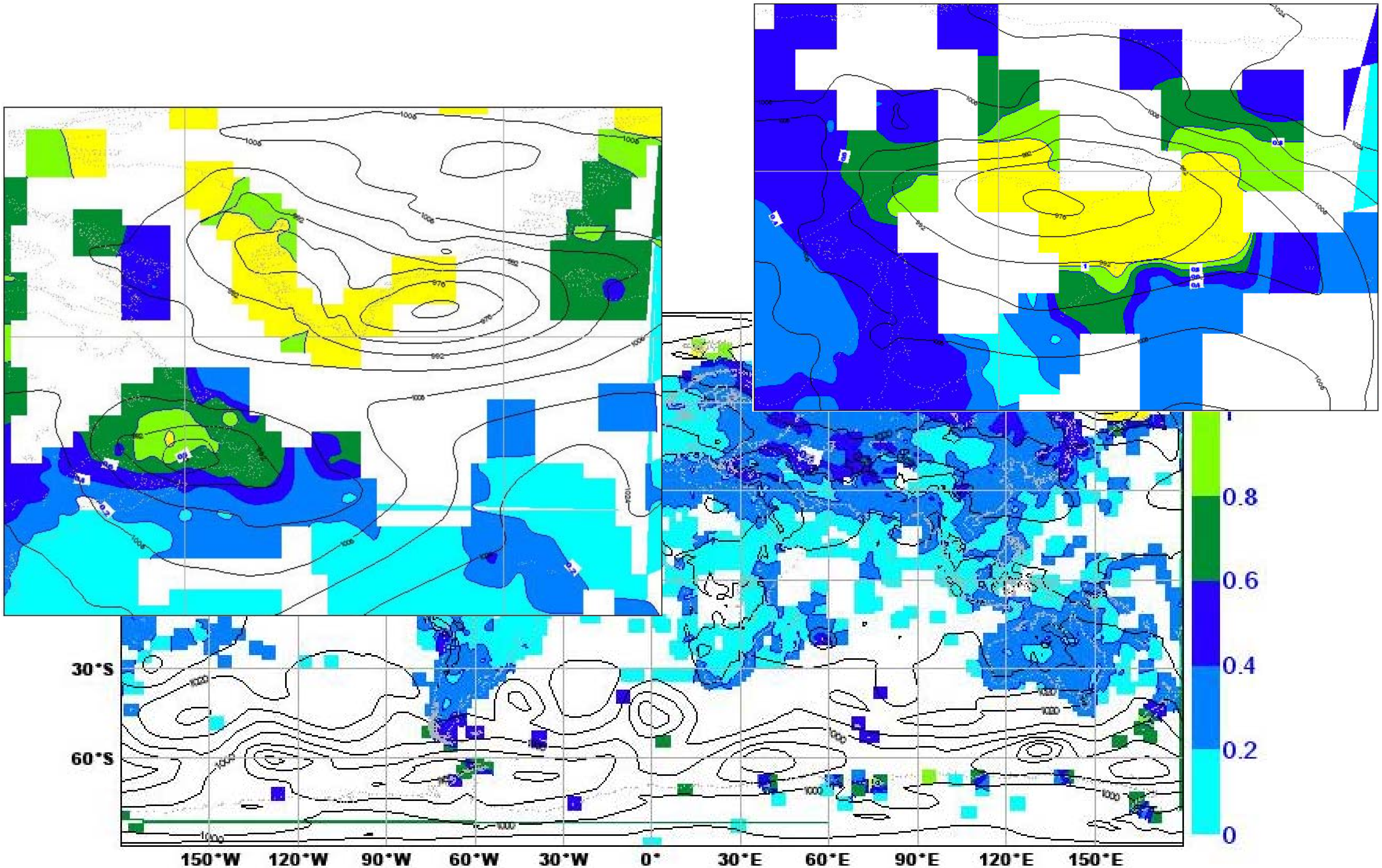
$$\mathbf{S} = \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{Hx}_b} = \mathbf{I} - \mathbf{K}^T\mathbf{H}^T$$

$$\mathbf{S} = \mathbf{R}^{-1}\mathbf{H}(\mathbf{B}^{-1} + \mathbf{HR}^{-1}\mathbf{H}^T)^{-1}\mathbf{H}^T$$

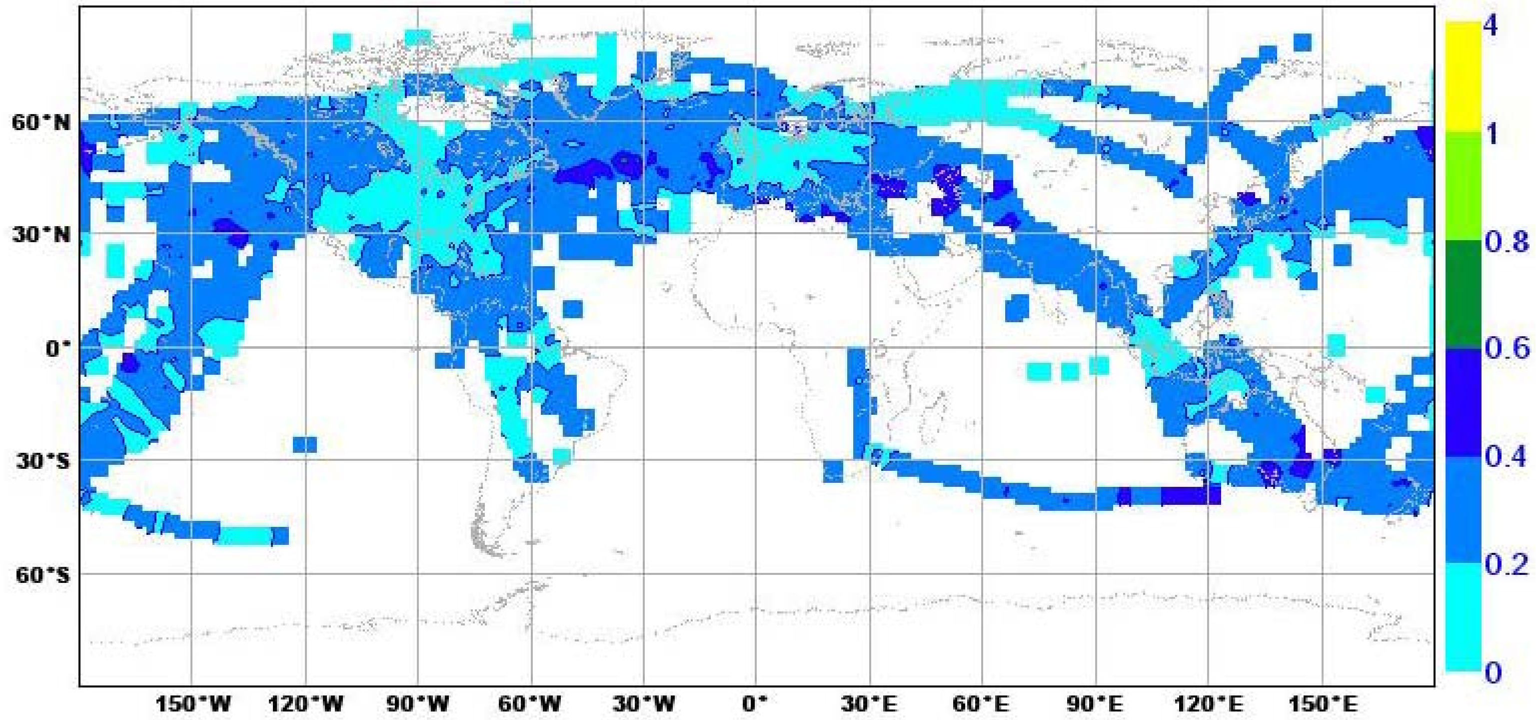
$$\mathbf{A} = (\mathbf{B}^{-1} + \mathbf{HR}^{-1}\mathbf{H}^T)^{-1}$$

$$\mathbf{S} = \mathbf{R}^{-1}\mathbf{H}(\mathbf{J}'')^{-1}\mathbf{H}^T$$

Synop Surface Pressure Influence

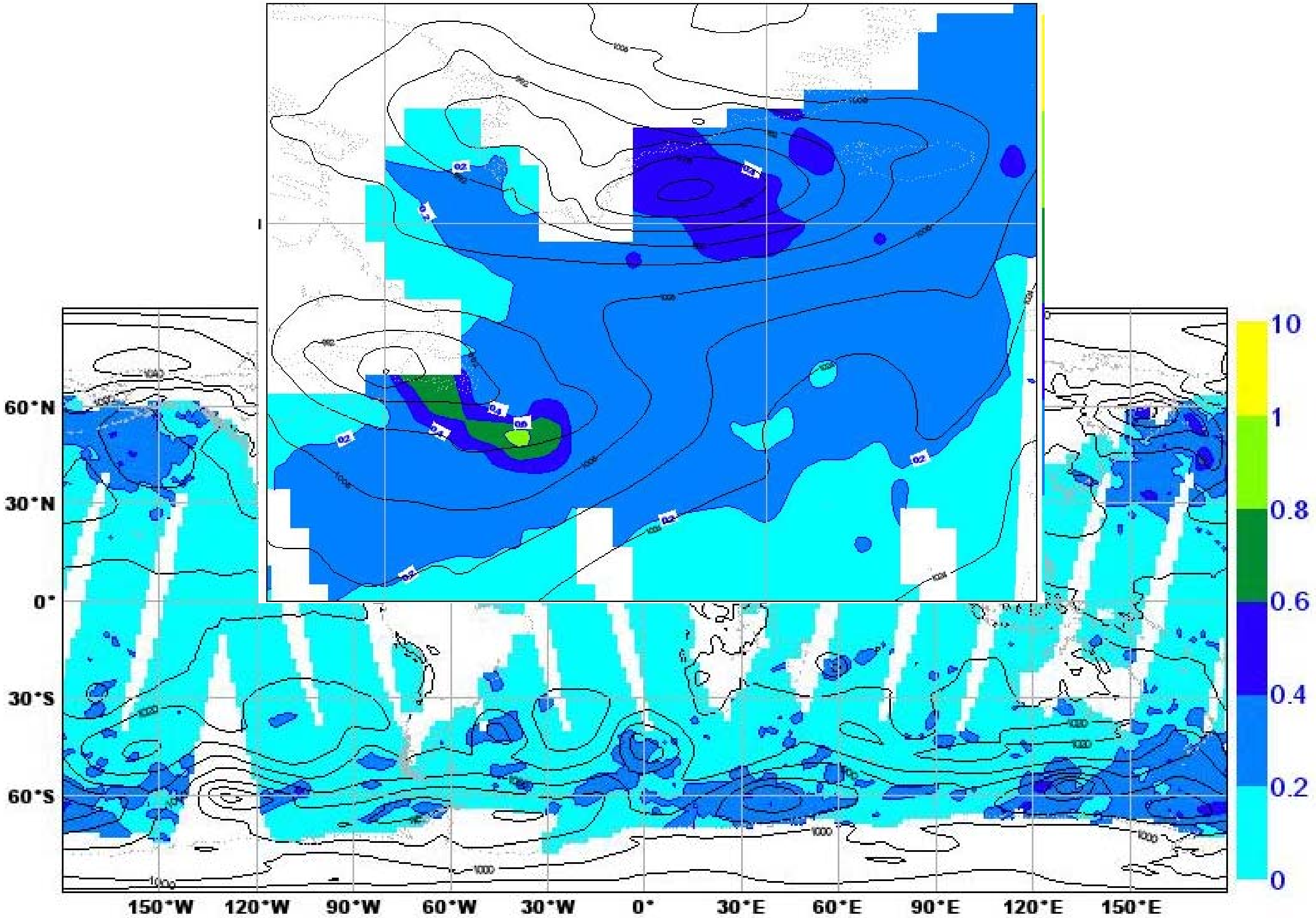


Airep 250 hPa U-Comp Influence



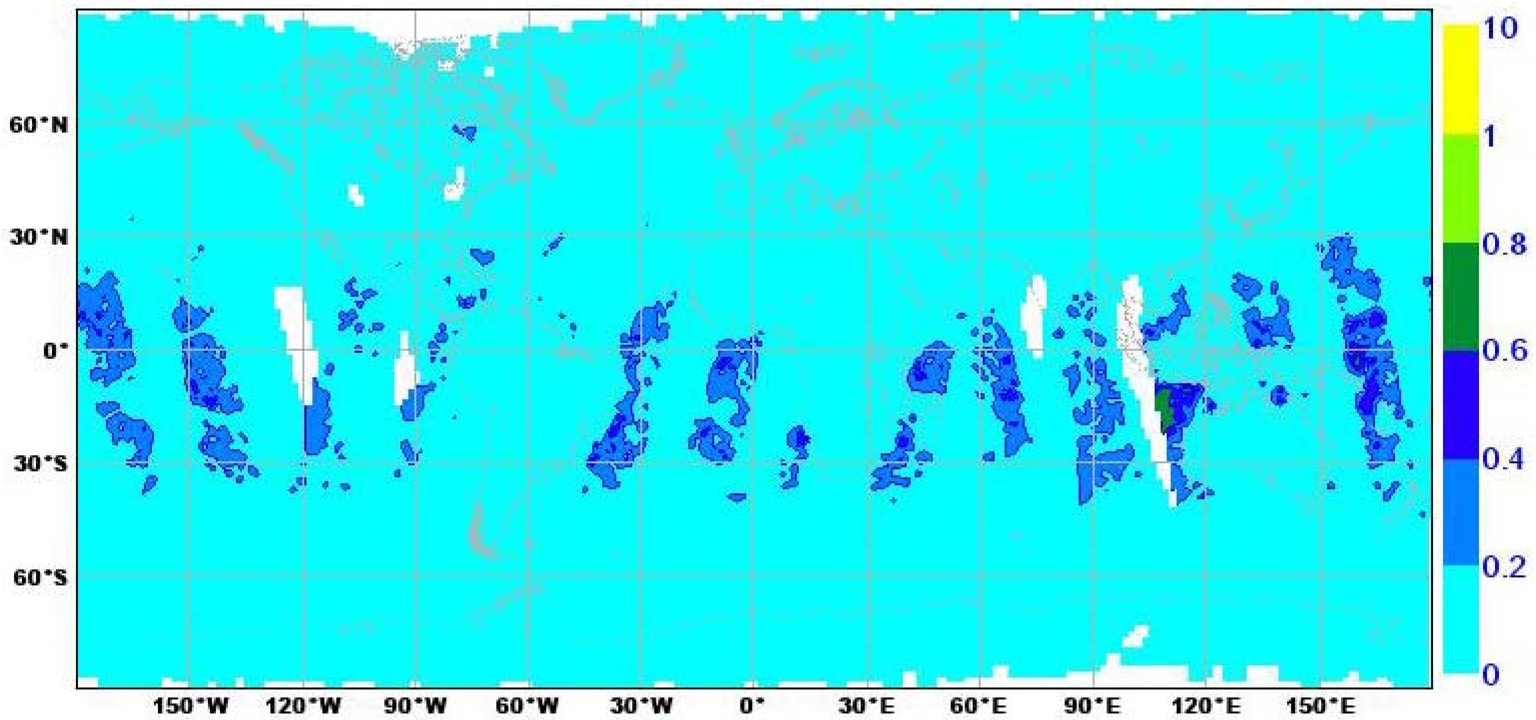
ECMWF

QuikSCAT U-Comp Influence



ECMWF

AMSU-A channel 13 Influence



Hessian in P variable

$$\mathbf{A} \simeq (\mathbf{J}'')^{-1}$$

$$\chi = \mathbf{L}^{-1}\mathbf{x} \qquad \mathbf{B}^{-1} = \mathbf{L}^T\mathbf{L}$$

$$\mathbf{J}''(\chi) = \mathbf{I} + \mathbf{L}^T\mathbf{H}\mathbf{R}^{-1}\mathbf{H}^T\mathbf{L} = \mathbf{L}^T(\mathbf{B}^{-1} + \mathbf{H}\mathbf{R}^{-1}\mathbf{H}^T)\mathbf{L} = \mathbf{L}^T\mathbf{J}''(\mathbf{x})\mathbf{L}$$

$$\mathbf{J}''(\chi) = \mathbf{L}^T\mathbf{J}''(\mathbf{x})\mathbf{L}$$

Influence Matrix Computation

$$\mathbf{S} = \mathbf{R}^{-1}\mathbf{H}(\mathbf{J}'')^{-1}\mathbf{H}^{T}$$

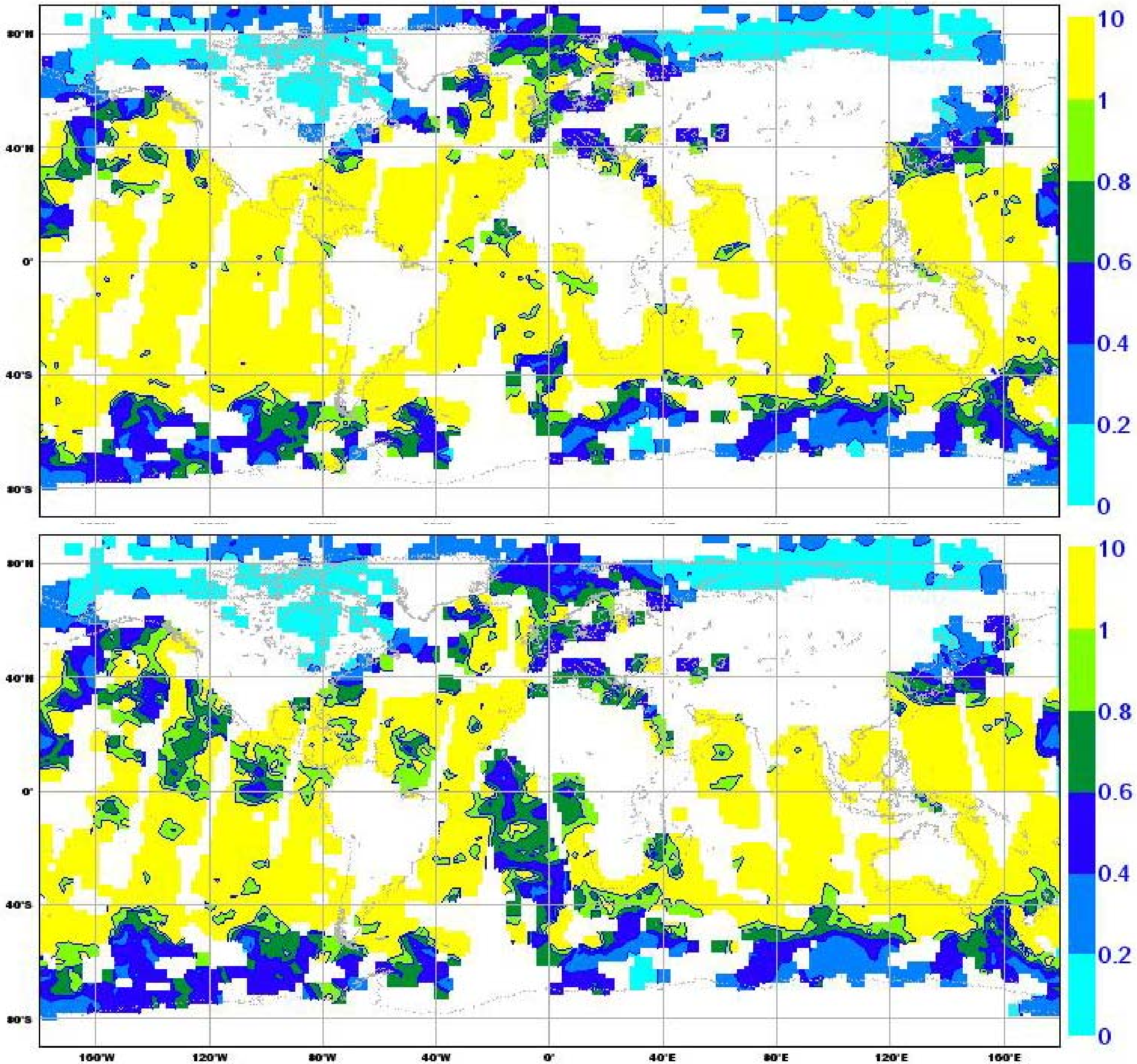
$$(\mathbf{J}'')^{-1} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{L}u_i)(\mathbf{L}u_i)^{T} - \sum_{i=1}^{M}\frac{1-\lambda_i}{\lambda_i}(\mathbf{L}v_i)(\mathbf{L}v_i)^{T}$$

B

A sample of N=50 random vectors from ù(0,1)

Truncated eigenvector expansion with vectors obtained through the combined Lanczos/conjugate algorithm. M=40

HIRS channel 11 radiances Influence

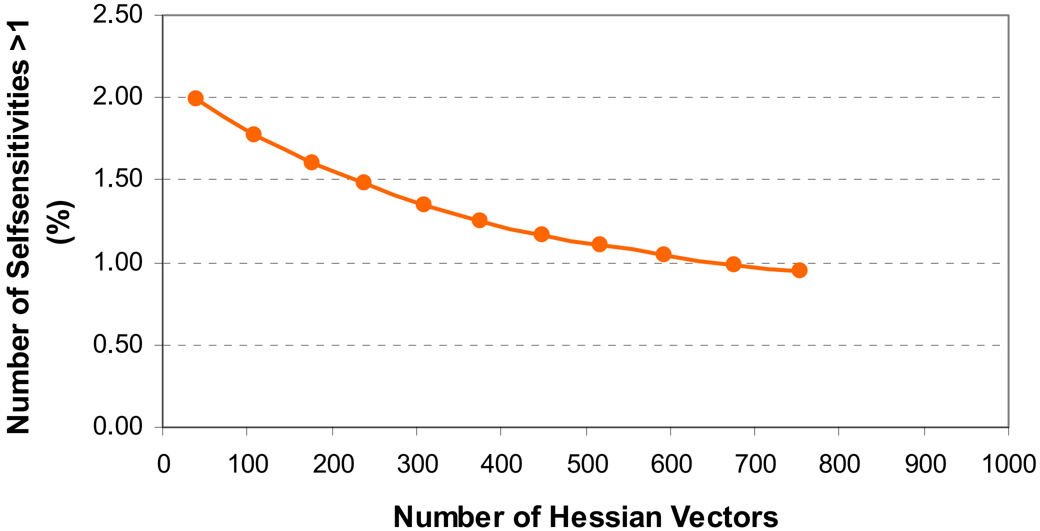


Random B Vector N=50
Hessian Vector M=40

Random B Vector N=500
Hessian Vector M=753

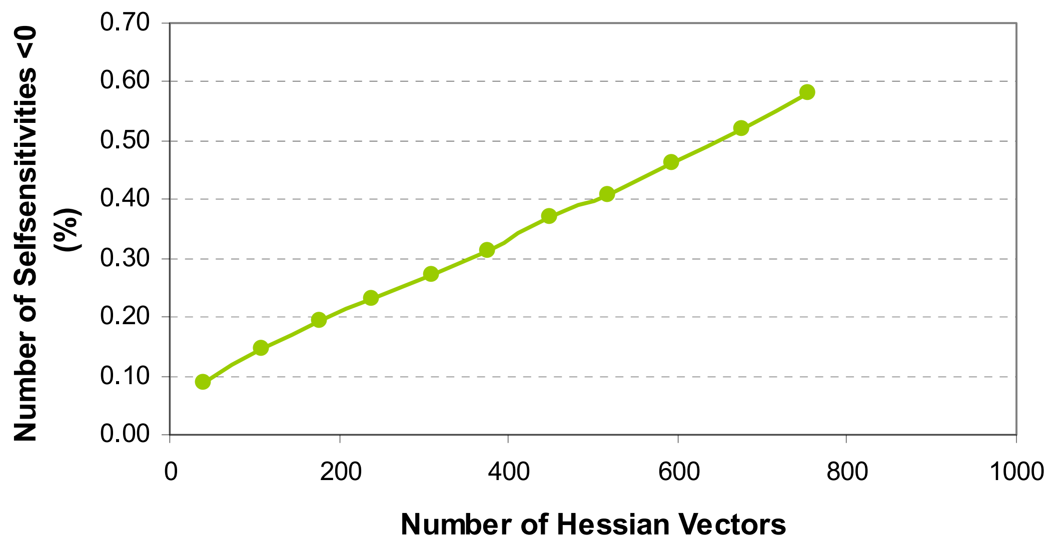
ECMWF

Hessian Approximation ➟ B-A



$$\sum_{i=1}^{M} \frac{1-\lambda_i}{\lambda_i} (\mathbf{L}\, v_i)(\mathbf{L}\, v_i)^T$$

$$\frac{1}{N} \sum_{i=1}^{N} (\mathbf{L} u_i)(\mathbf{L} u_i)^T$$



Global and Partial Influence

$$\text{Global Influence} = \text{GI} = \frac{Tr(\mathbf{S})}{p}$$

⇨ 100% only Obs Influence

⇨ 0% only Model Influence

$$\text{Partial Influence} = \text{PI} = \frac{\sum_{i \in I} \mathbf{S}_{ii}}{p_I}$$

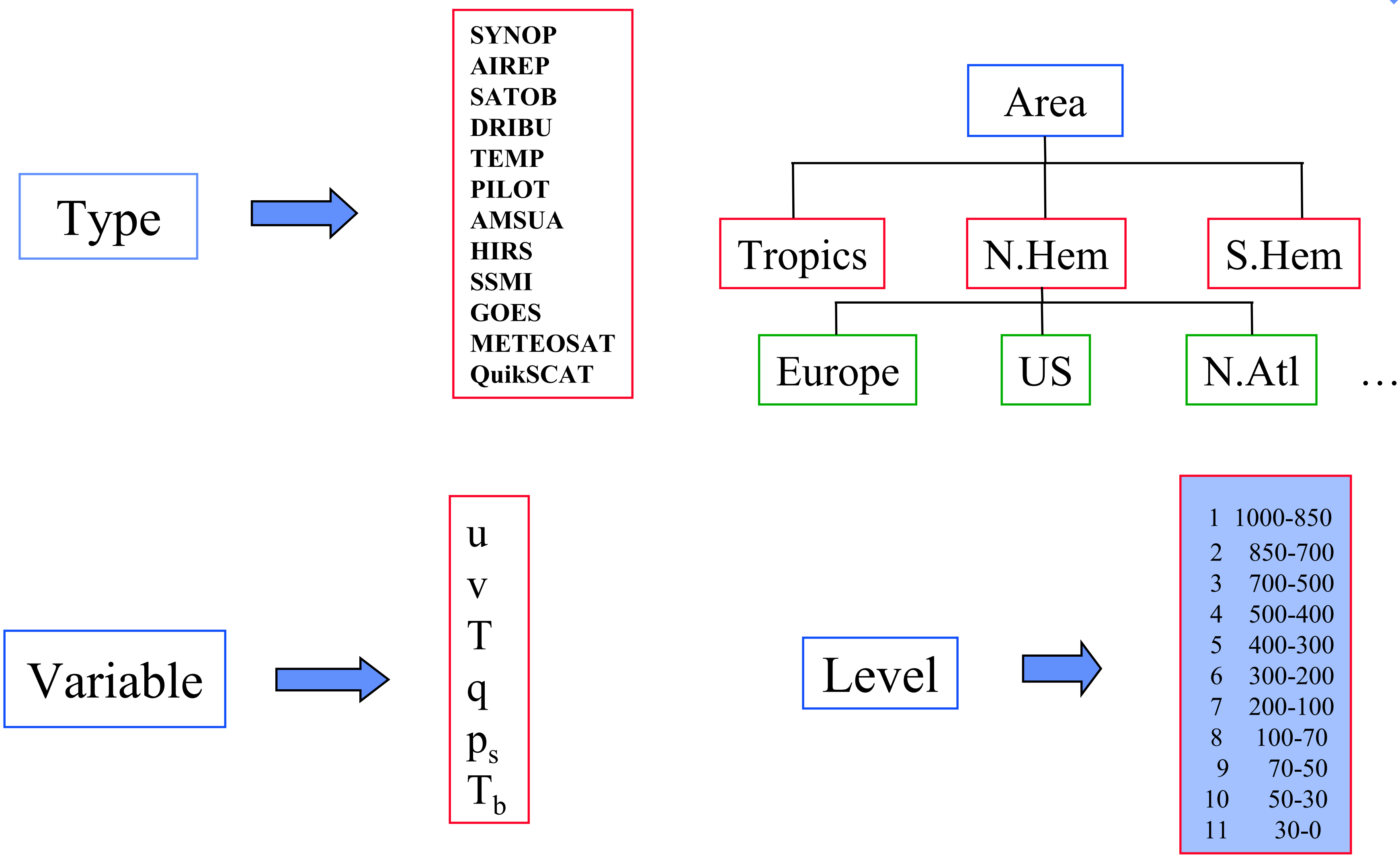
Type

Area

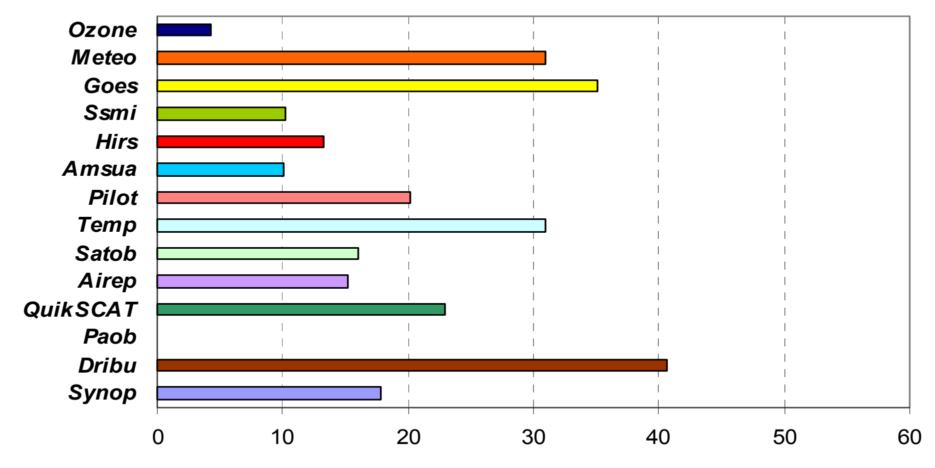
Variable

Level

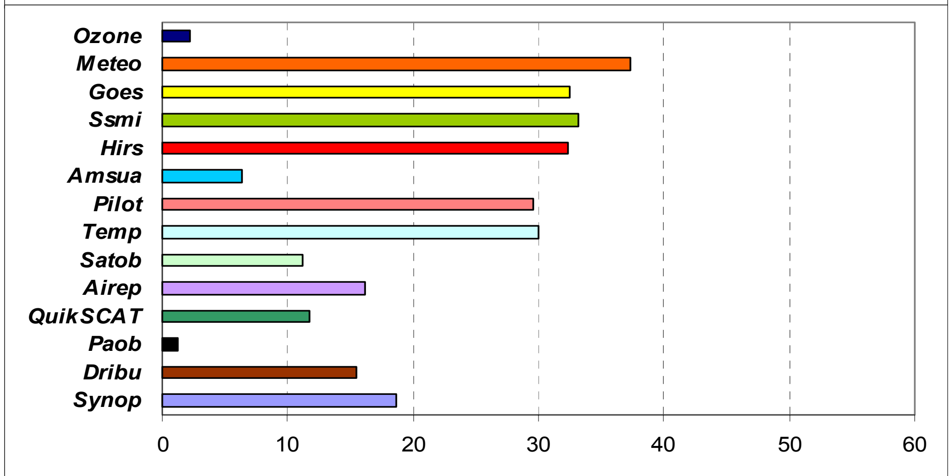
Global and Partial Influence



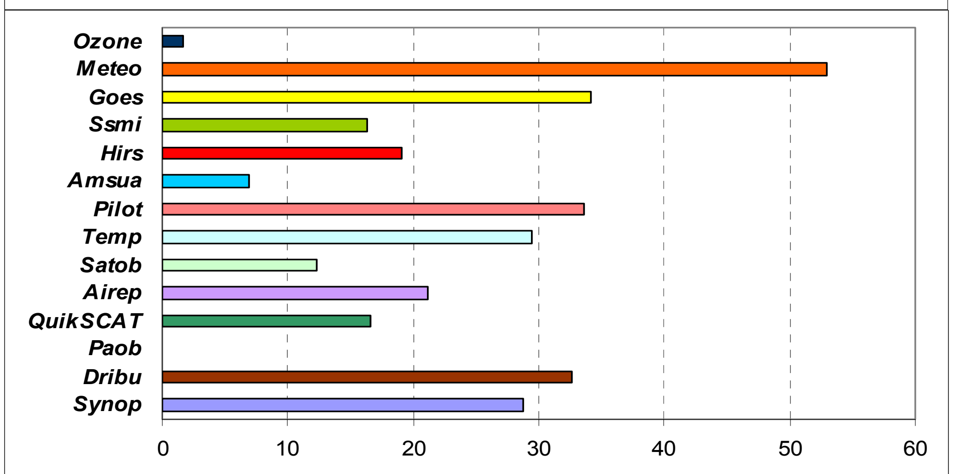
GI = 15.3%



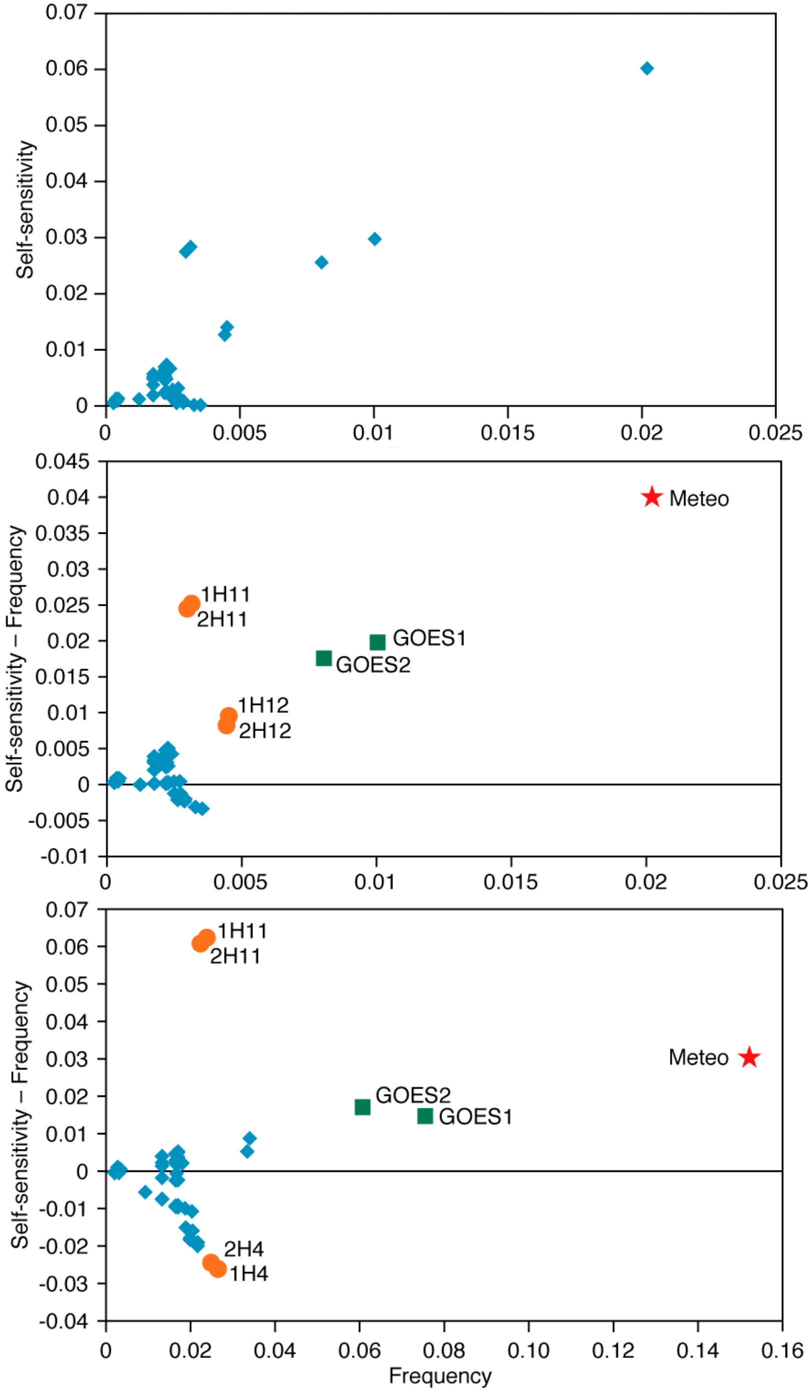
N.Hemisphere
PI = 15%



Tropics
PI = 17.5%



S.Hemisphere
PI = 12%



Humidity over Tropics: Hirs, Ssmi, Goes, Meteo, Temp, Synop

GI = 15.3% **PI = 33.4%**

$$Self = \frac{S_{HIRS}}{Tr(\mathbf{S})}, \frac{S_{SSMI}}{Tr(\mathbf{S})}.....$$

$$Freq = \frac{\mathbf{N}}{p} = \frac{n_{HIRS}}{p}, \frac{n_{SSMI}}{p}..$$

Values relative to Global Amounts

Intercept	Slope	F-test Y=X
0	3.1	p-value=0%

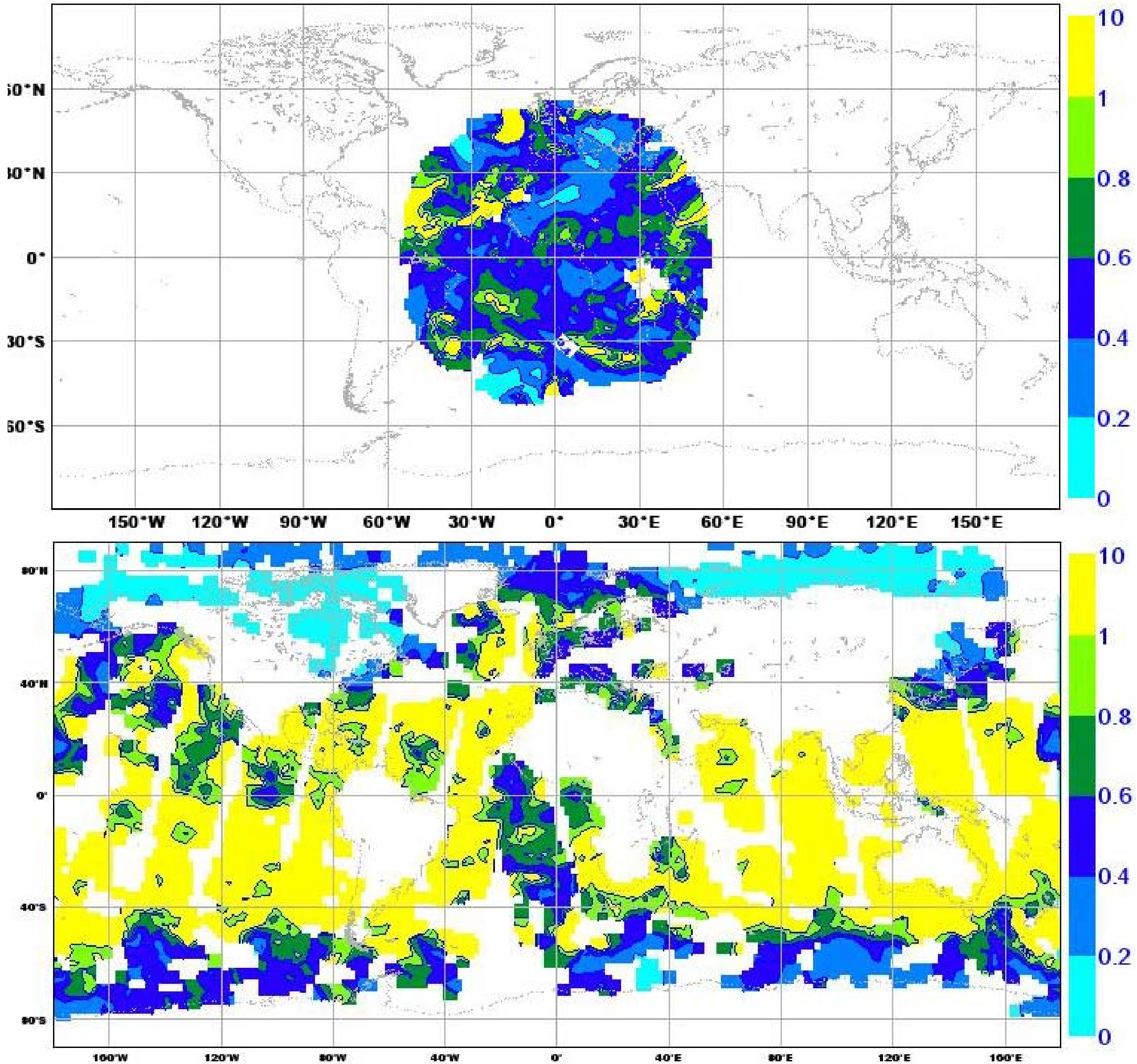
$$Self = \frac{S_{\mathbf{HIRS}}}{S_{\mathbf{N}}}, \frac{S_{SSMI}}{S_{\mathbf{N}}}...$$

$$Freq = \frac{n_{\mathbf{HIRS}}}{\mathbf{N}}, \frac{n_{\mathbf{SSMI}}}{\mathbf{N}}...$$

Values relative to Partial Amounts

Intercept	Slope	F-test Y=X
0	1.3	p-value=8%

METEOSAT and HIRS-11 radiances Influence



Ill-Condition Problem

- A set of linear equation is said to be *ill-conditioned* if small variations in X=(HK I-HK) have large effect on the exact solution ŷ, e.g matrix close to singularity

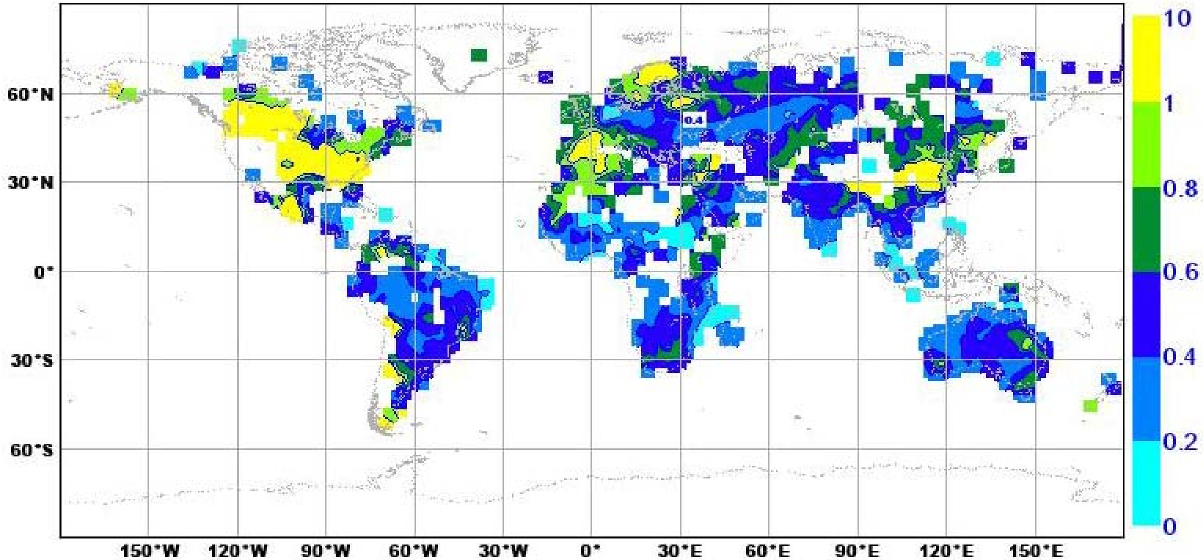
- A Ill-conditioning has effects on the stability and solution accuracy . A measure of ill-conditioning is

$$\kappa(\mathbf{X}) = \frac{\lambda_{max}}{\lambda_{min}}$$

- A different form of ill-conditioning can results from collinearity: XX^T close to singularity

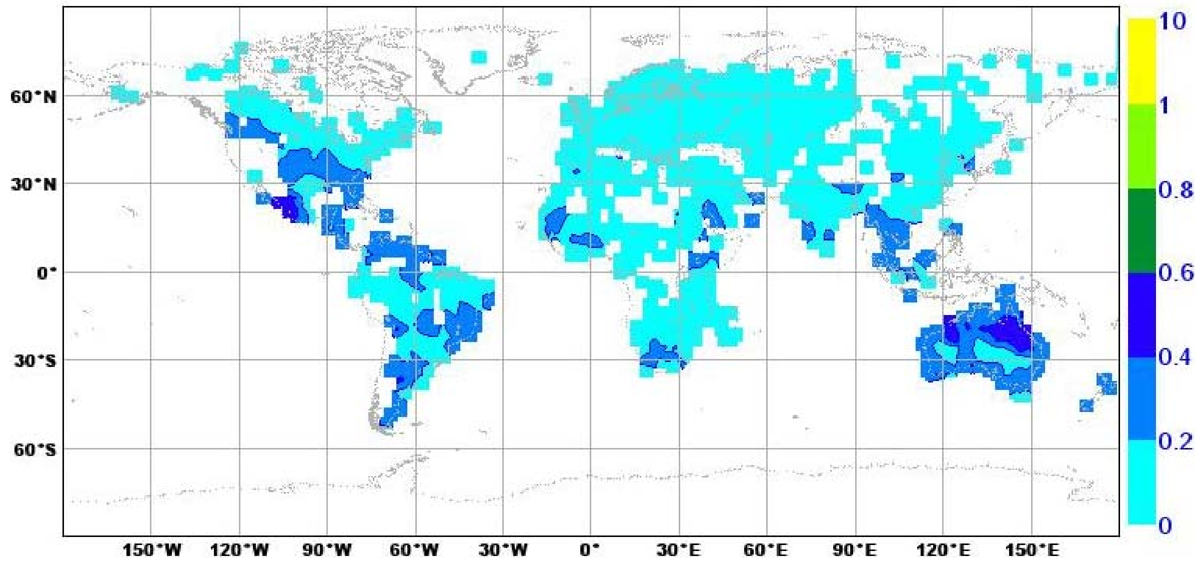
- Large difference between the background and observation error standard deviation and high dimension matrix

SYNOP RH 2m Influence



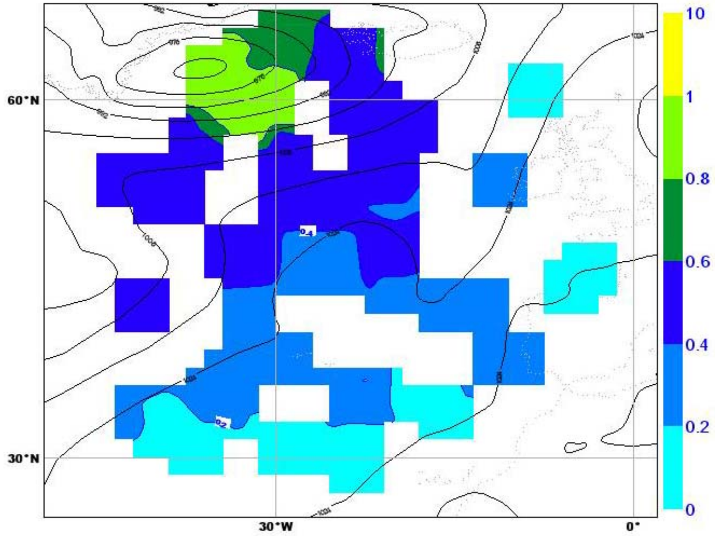
Background Error Variances Depending on T and Q variables

Computed at every cycle

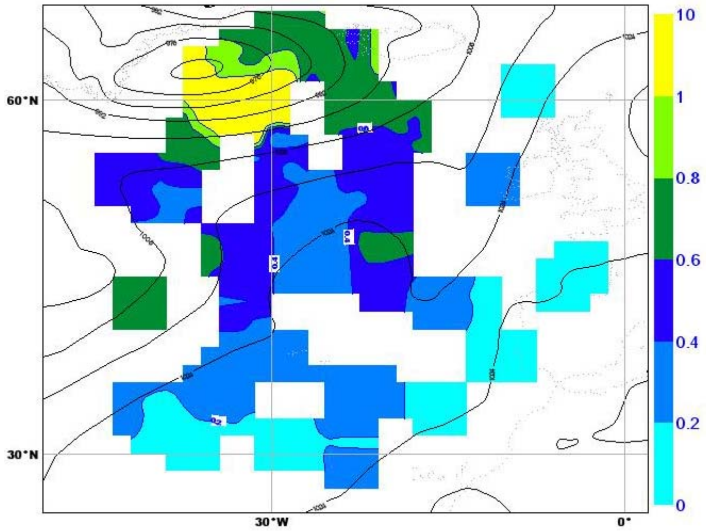


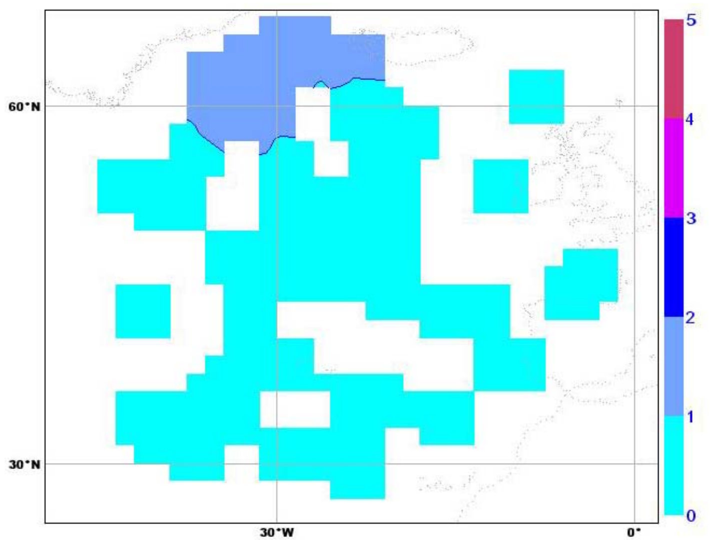
Use of Standardized Humidity variable

Flow Dependent F_b: MAM^T+Q

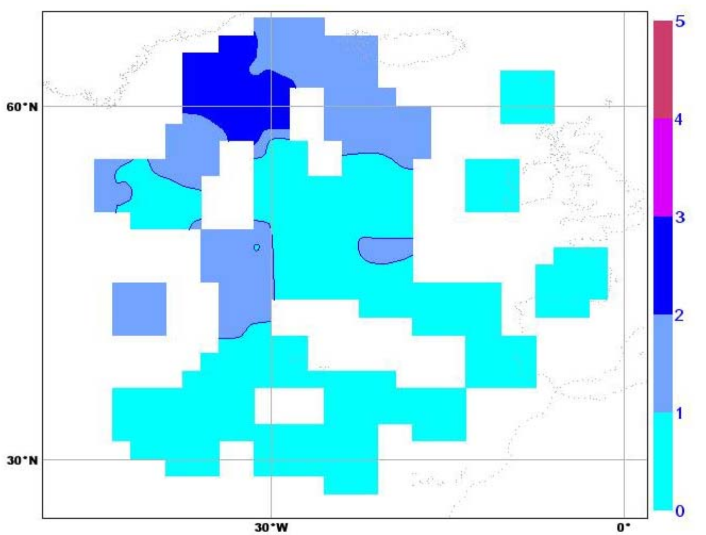


DRIBU ps
Influence

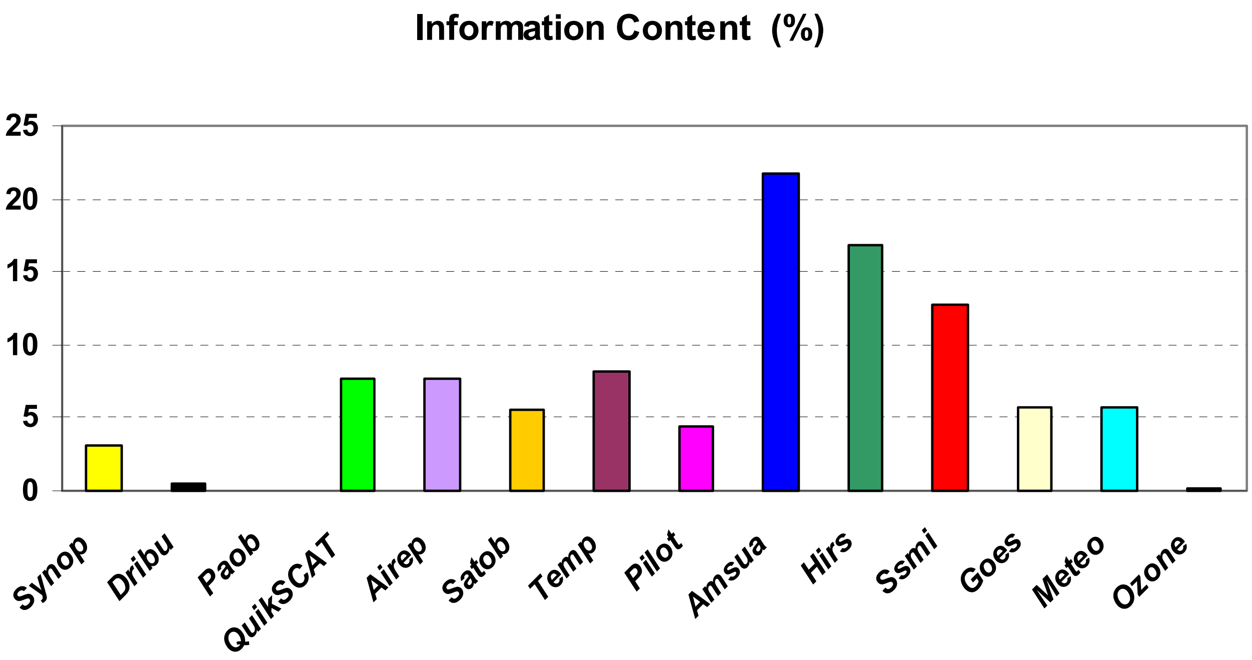




$$\frac{\sigma_b}{\sigma_o}$$



Information Content



ECMWF

Conclusions

- The Influence Matrix is well-known in multi-variate linear regression. It is used to identify influential data and to predict the impact on the estimate of removing individual observations

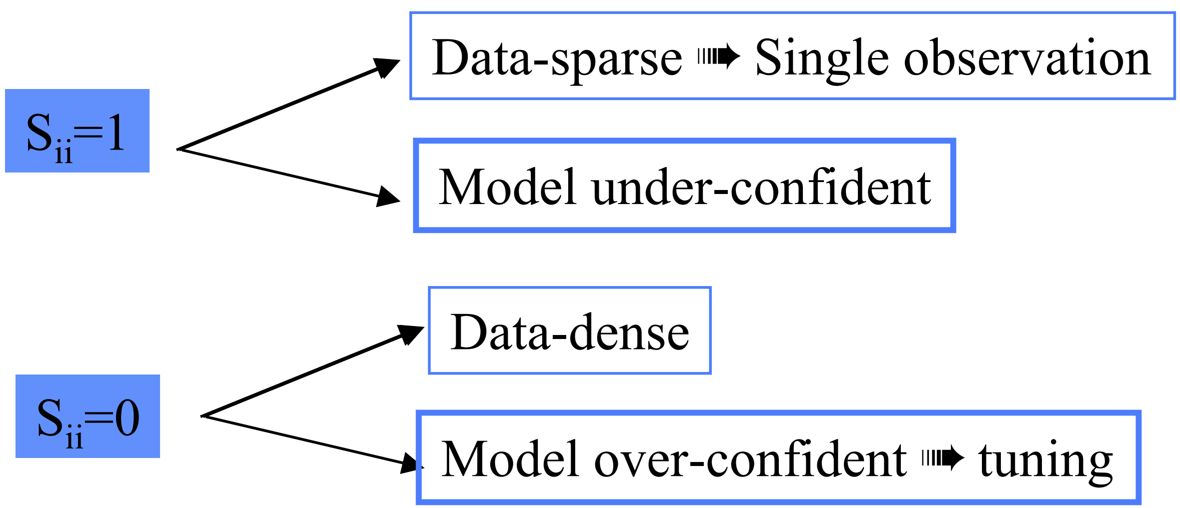
- An approximate method to compute the diagonal elements, self-sensitivities, of the influence matrix in 4D-Var has been shown. The approximation is necessary due to the large dimension of the estimation problem (10^6)

- Influence patterns are not part of the estimates of the model but rather are part of the conditions under which the model is estimated

- It is expected that the data have a similar influence. Disproportionate influence can be due to:
  - incorrect data
  - legitimately extreme observations occurrence
    - → to which extent the estimate depends on these data

Conclusions

- **For the same observation type the influence is significantly larger in data-sparse regions than in data-dense regions**
  - **the former have much larger impact on the local analysis error variance**
    - ➔ Utility of observations in data-sparse region
    - ➔ Redundancy of additional observations in a well-observed region



Conclusions

- **Observational Influence pattern would provide information on different observation system**
  - **New observation system**
  - **Special observing field campaign**

- **Thinning is mainly performed to reduce the spatial correlation but also to reduce the analysis computational cost**
  - **Knowledge of the observations influence helps in selecting appropriate data density**

- **Diagnose the impact of improved physics representation in the linearized forecast model in terms of observation influence**

What about Background and Observation Tuning in ECMWF 4D-Var?