Influence matrix diagnostic to monitor the assimilation system

Carla Cardinali

Office N 1140

Monitoring the 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 10^8 degree of freedom and 10^7 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 does the estimate depend on one single influential obs?

Influence Matrix: Introduction

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

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

- In Ordinary Least-Square the information is quantitatively available in the *Influence Matrix*

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

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

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

β (*qx1*) unknown parameters

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

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

- 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** $0 \le S_{ii} \le 1 \qquad Tr(\mathbf{S}) = q$
- **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 Related Findings

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

- The change in the estimate that occur when the i-th is deleted

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

$$r_i = y_i - \hat{y}_i$$

- CV score can be computed by relying on the all data estimate $\hat{\mathbf{y}}$ and S_{ii}

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

Outline

- Generalized Least Square method
- Observation and background Influence
- Findings related to data influence and information content
- Toy model: 2 observations
- Conclusion

Solution in the Observation Space

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

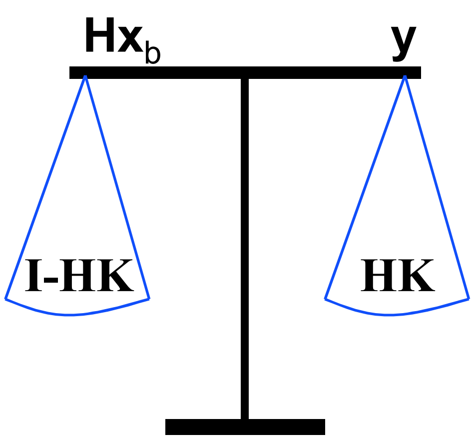
- The analysis projected at the observation location

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

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

K(q$_x$p) gain matrix
H(p$_x$q) Jacobian matrix

B(q$_x$q)=Var(x_b)
R(p$_x$p)=Var(y)



The estimation $\hat{y}$ is a weighted mean

Influence Matrix

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

$$\mathbf{S} = \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{y}} = (\mathbf{HK})^T = \mathbf{K}^T\mathbf{H}^T = Observation - Influence$$

$$\mathbf{I} - \mathbf{S} = \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{H}\mathbf{x}_b} = Background - Influence$$

Observation Influence is complementary to Background Influence

$$\hat{\mathbf{y}} = \mathbf{Sy} + (\mathbf{I} - \mathbf{S})\mathbf{H}\mathbf{x}_b$$

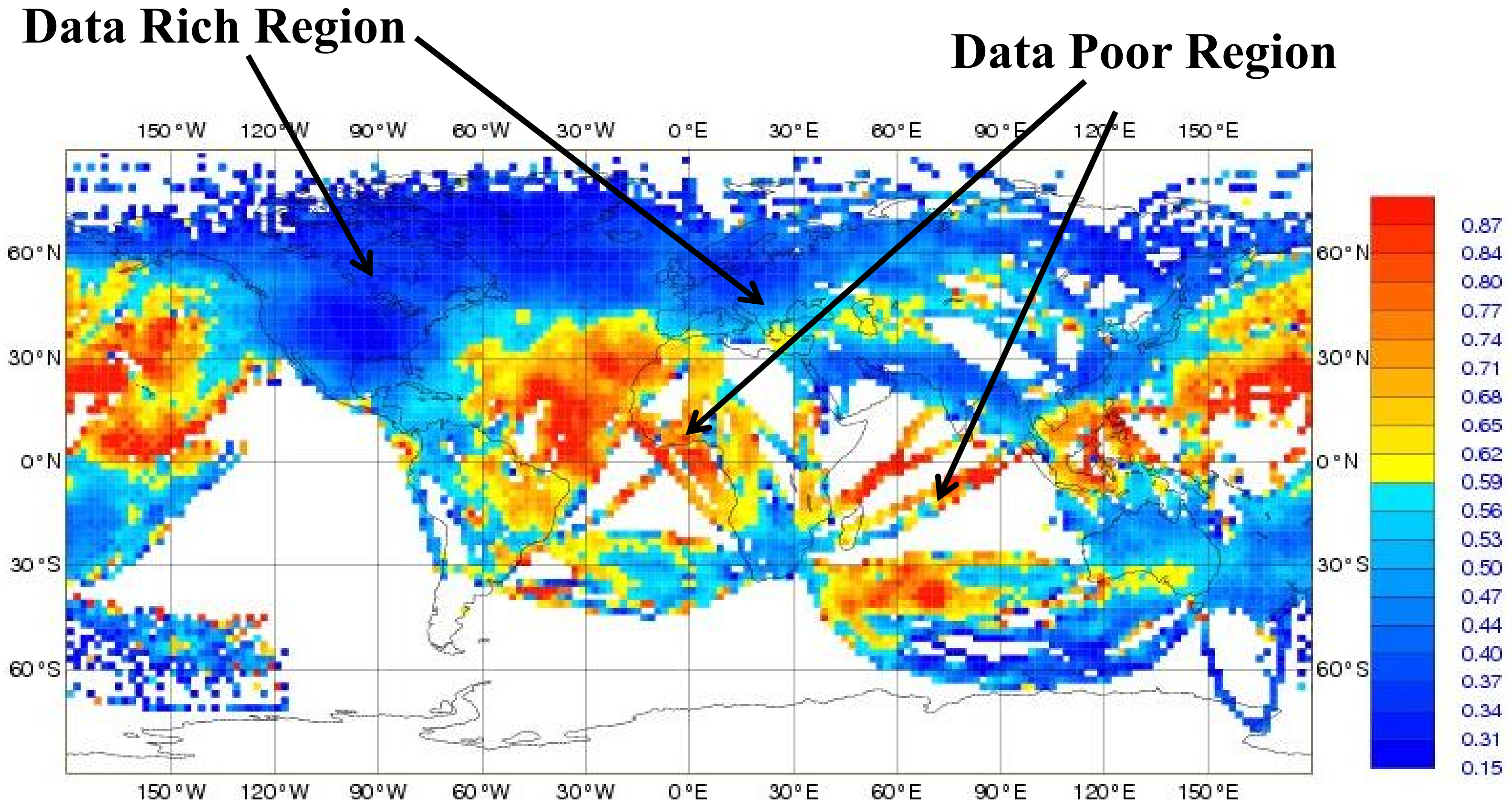
Influence Matrix Properties

The diagonal element satisfy $0 \le S_{ii} \le 1$

$$\sum_{i=1}^{N} S_{ii} = Total\_Information\_Content$$

$$\frac{\sum_{i=1}^{N} S_{ii}}{Tot.Obs.Number} = Average\_Influence$$

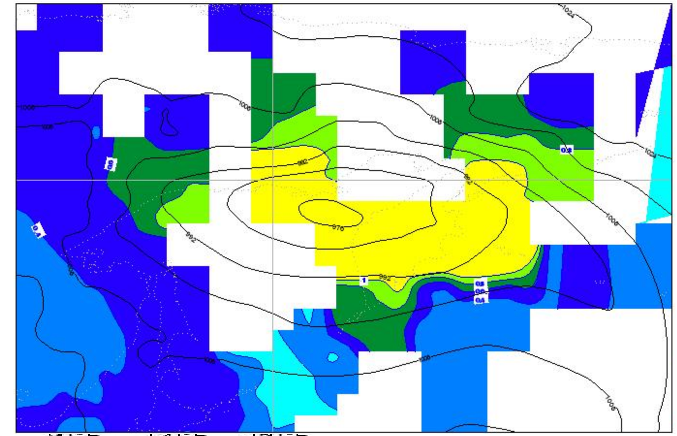
Aircraft above 400 hPa U-Comp Influence

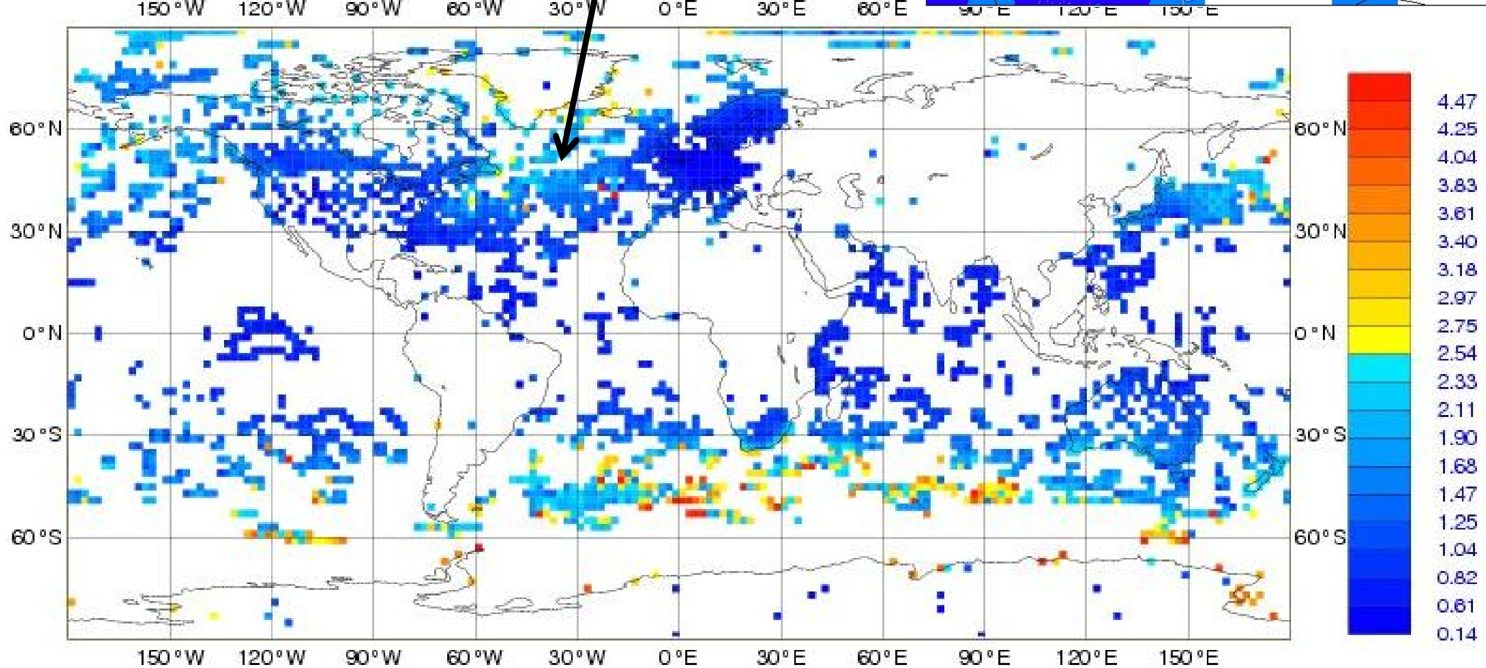


Synop&DRIBU Surface Pressure Influence

$S_{ii}>1$
due to
the numerical
approssimation

Dynamical active area

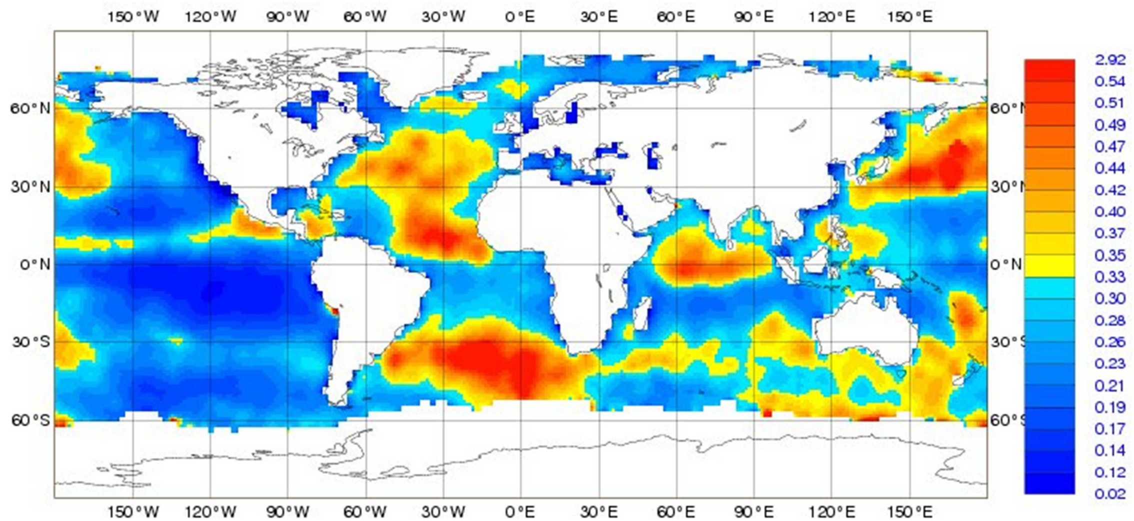




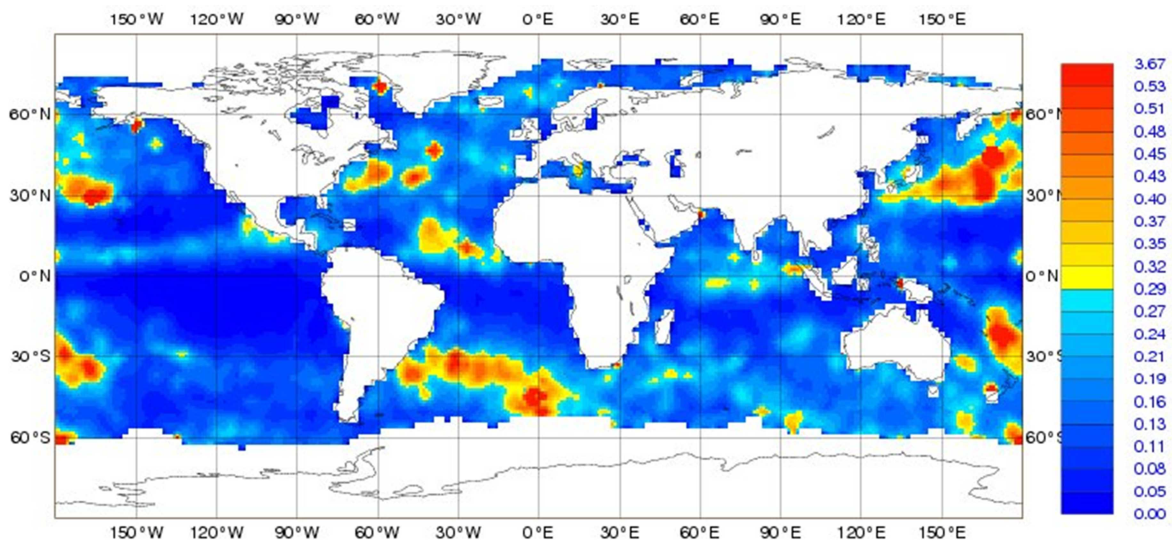
ECMWF

ASCAT U-Comp Influence

Mean



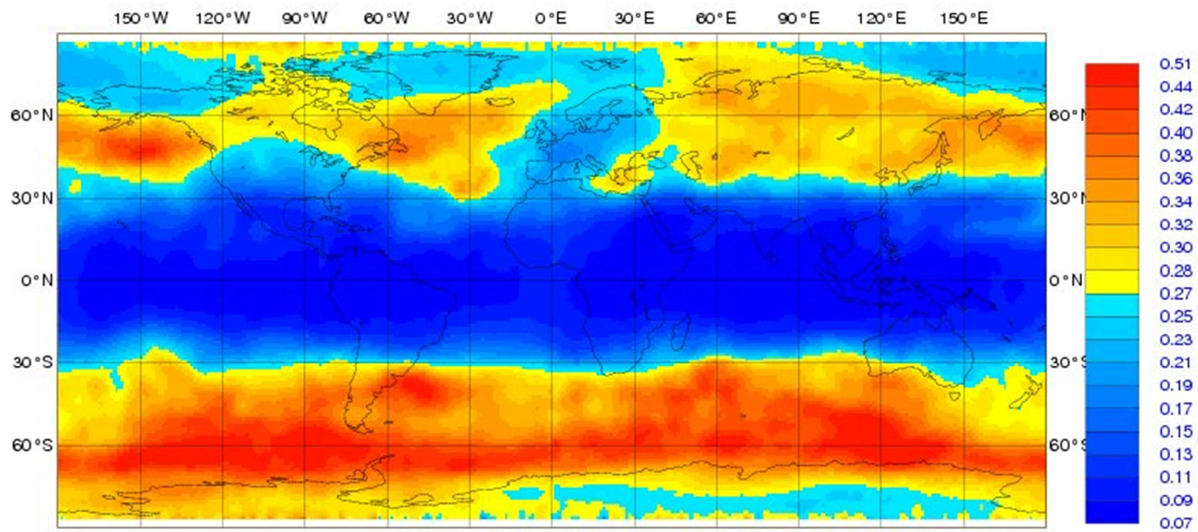
STD



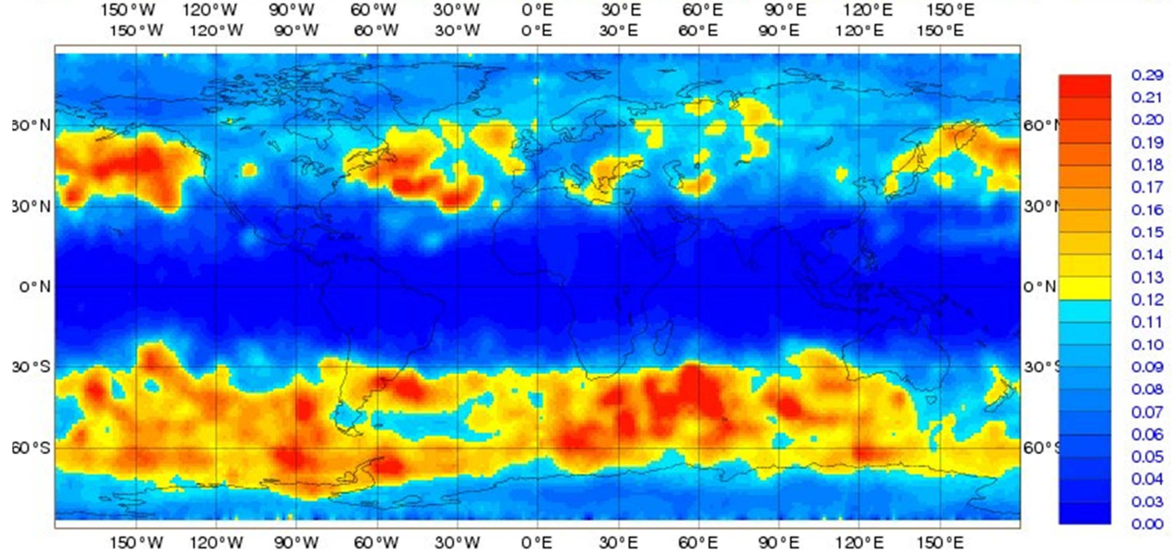
ECMWF

AMSU-A channel 8 Influence

Mean

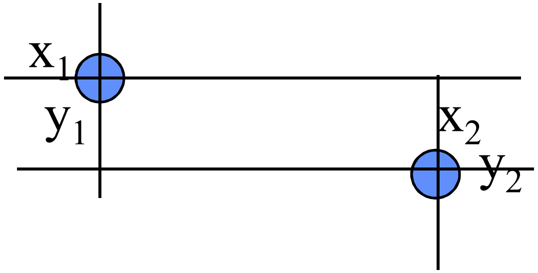


STD



Toy Model: 2 Observations

Find the expression for S as function of r and the expression of $\hat{\mathbf{y}}$ for α=0 and ~1 given the assumptions:

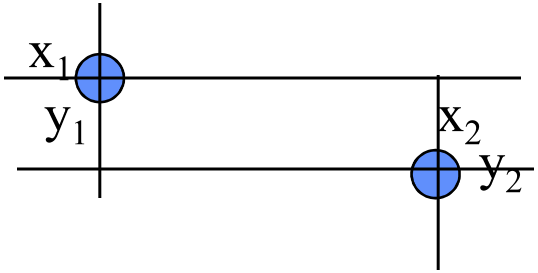


$$\mathbf{H} = \mathbf{I} \qquad \mathbf{R} = \sigma_o^{\,2}\mathbf{I} \qquad \mathbf{B} = \sigma_b^{\,2}\begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix} \qquad r = \frac{\sigma_o^{\,2}}{\sigma_b^{\,2}}$$

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

$$\boxed{\hat{\mathbf{y}} = \mathbf{S}\mathbf{y} + (\mathbf{I} - \mathbf{S})\mathbf{x}_b}$$

Toy Model: 2 Observations



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

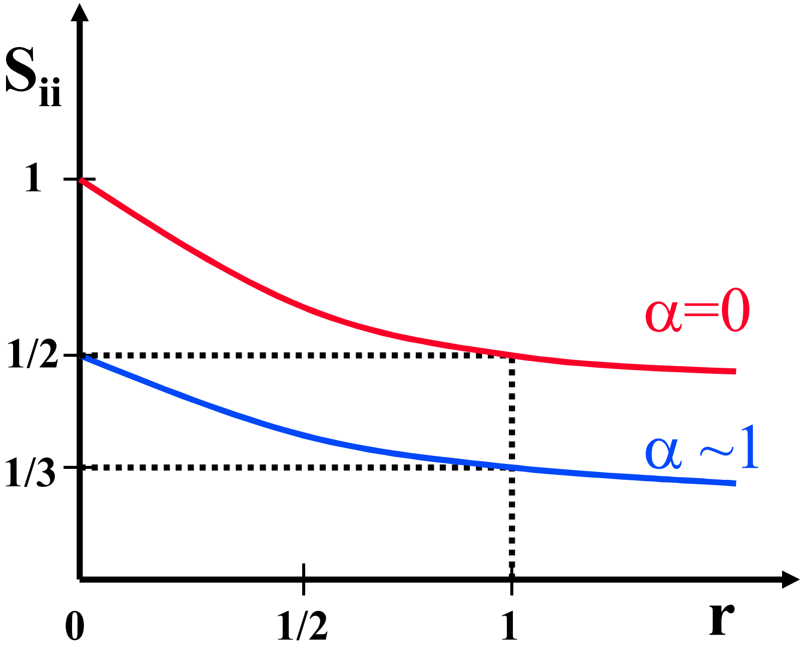
$$\mathbf{H} = \mathbf{I} \qquad \mathbf{R} = \sigma_o^{\ 2}\mathbf{I} \qquad \mathbf{B} = \sigma_b^{\ 2}\begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}$$

$$\boxed{r = \frac{\sigma_o^{\ 2}}{\sigma_b^{\ 2}}}$$

$$S = \begin{pmatrix} \dfrac{r+1-\alpha^2}{r^2+2r+1-\alpha^2} & \dfrac{\alpha r}{r^2+2r+1-\alpha^2} \\ \dfrac{\alpha r}{r^2+2r+1-\alpha^2} & \dfrac{r+1-\alpha^2}{r^2+2r+1-\alpha^2} \end{pmatrix}$$

$$\alpha \cong 1 \quad \rightarrow \quad S_{11} = S_{22} = S_{12} = S_{21} = \frac{1}{r+2}$$

$$\alpha = 0 \quad \rightarrow \quad S_{11} = S_{22} = \frac{1}{r+1}$$



ECMWF

Consideration (1)

- **Where observations are dense S_{ii} tends to be small and the background sensitivities tend to be large and also the surrounding observations have large influence (off-diagonal term)**

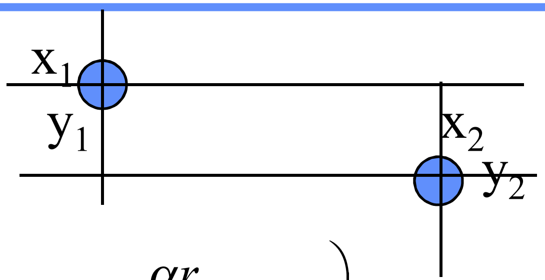
$$\alpha \cong 1 \quad \rightarrow \quad S_{11} = S_{22} = S_{12} = S_{21} = \frac{1}{r+2}$$

- **When observations are sparse S_{ii} and the background sensitivity are determined by their relative accuracies (r) and the surrounding observations have small influence (off-diagonal term)**

$$\alpha = 0 \quad \rightarrow \quad S_{11} = S_{22} = \frac{1}{r+1}$$

Toy Model: 2 Observations

$$\boxed{\hat{\mathbf{y}} = \mathbf{S}\mathbf{y} + (\mathbf{I} - \mathbf{S})\mathbf{x}_b}$$



$$\boxed{r = \frac{\sigma_o^2}{\sigma_b^2} = 1}$$

$$S = \begin{pmatrix} \dfrac{r+1-\alpha^2}{r^2+2r+1-\alpha^2} & \dfrac{\alpha r}{r^2+2r+1-\alpha^2} \\ \dfrac{\alpha r}{r^2+2r+1-\alpha^2} & \dfrac{r+1-\alpha^2}{r^2+2r+1-\alpha^2} \end{pmatrix}$$

$$\boxed{\hat{y}_1 = \frac{2-\alpha^2}{4-\alpha^2} y_1 + \frac{2}{4-\alpha^2} x_1 + \frac{\alpha}{4-\alpha^2}(y_2 - x_2)}$$

$\alpha \to\ =0$

$$\boxed{\hat{y}_1 = \frac{1}{2} y_1 + \frac{1}{2} x_1}$$

$\alpha \to\ \sim 1$

$$\boxed{\hat{y}_1 = \frac{1}{3} y_1 + \frac{2}{3} x_1 + \frac{1}{3}(y_2 - x_2)}$$

Consideration (2)

- **When observation and background have similar accuracies (r), the estimate $\hat{y}_1$ depends on y_1 and x_1 and an additional term due to the second observation. We see that if R is diagonal the observational contribution is devaluated with respect to the background because a group of correlated background values count more than the single observation ($2-\alpha^2 \rightarrow 2$). Also by increasing background correlation, the nearby observation and background have a larger contribution**

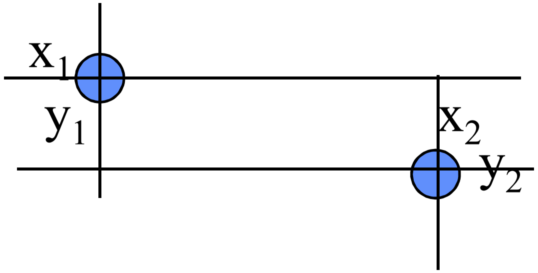
$$\alpha \rightarrow = 0 \qquad \boxed{\hat{y}_1 = \frac{1}{2} y_1 + \frac{1}{2} x_1}$$

$$\alpha \rightarrow \sim 1 \qquad \boxed{\hat{y}_1 = \frac{1}{3} y_1 + \frac{2}{3} x_1 + \frac{1}{3} (y_2 - x_2)}$$

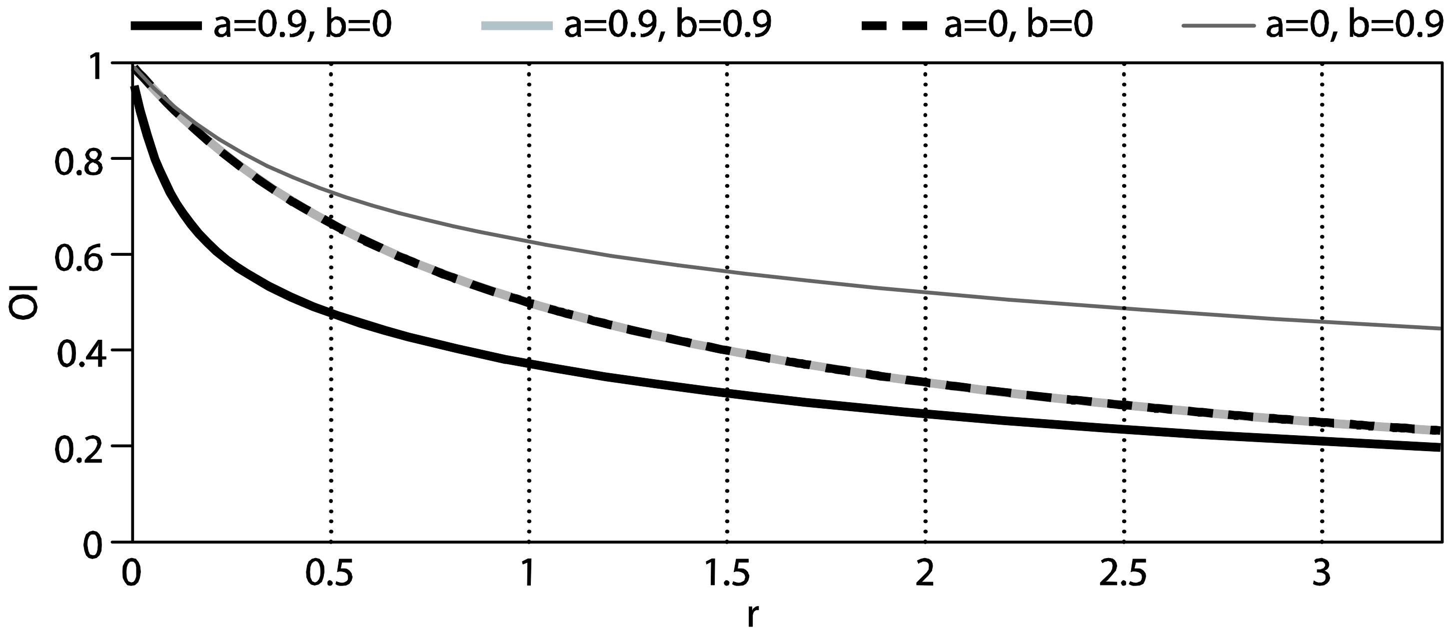
Toy Model: Correlated R

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

$$\hat{\mathbf{y}} = \mathbf{S}\mathbf{y} + (\mathbf{I} - \mathbf{S})\mathbf{x}_b$$



$$\mathbf{H} = \mathbf{I} \qquad \mathbf{R} = \sigma_o^{\ 2}\begin{pmatrix} 1 & \beta \\ \beta & 1 \end{pmatrix} \quad \mathbf{B} = \sigma_b^{\ 2}\begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix} \qquad r = \frac{\sigma_o^{\ 2}}{\sigma_b^{\ 2}}$$



ECMWF

Global and Partial Influence

Global Observation Influence $= \dfrac{\sum_{i=1}^{N} S_{ii}}{N} =$ ↗ 100% only Obs Influence ↘ 0% only Model Influence

Partial Observation Influence $= \dfrac{\sum_{i \in I} S_{ii}}{p_I}$

Type | Area | Variable | Level

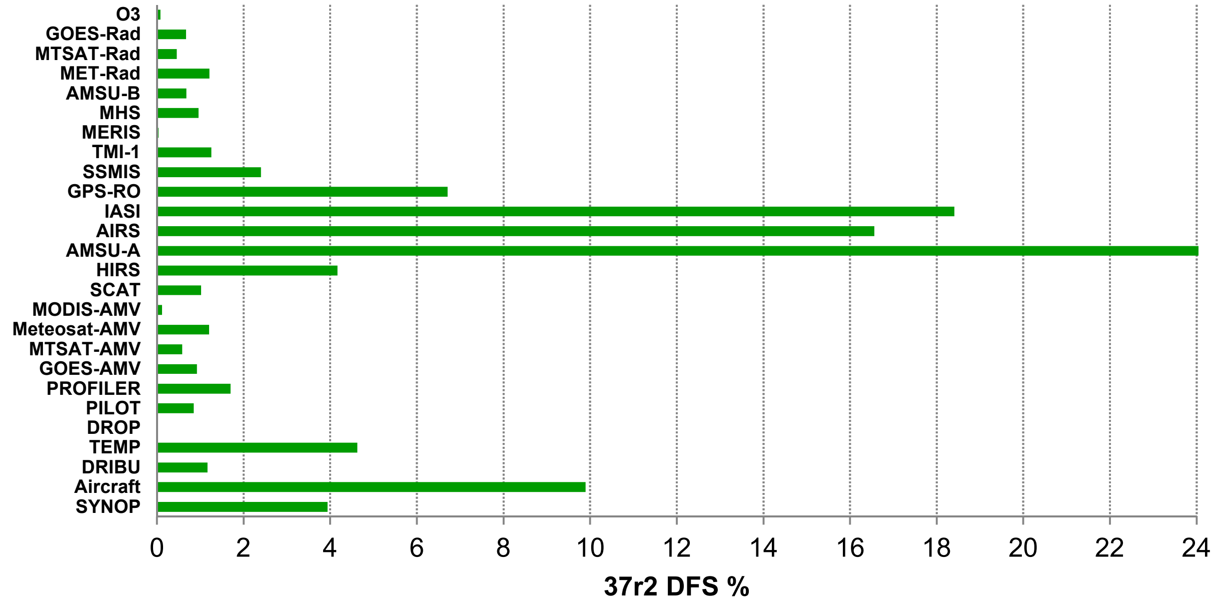
Type ⇒ **SYNOP, AIREP, SATOB, DRIBU, TEMP, PILOT, AMSUA, HIRS, SSMI, GOES, METEOSAT, QuikSCAT**

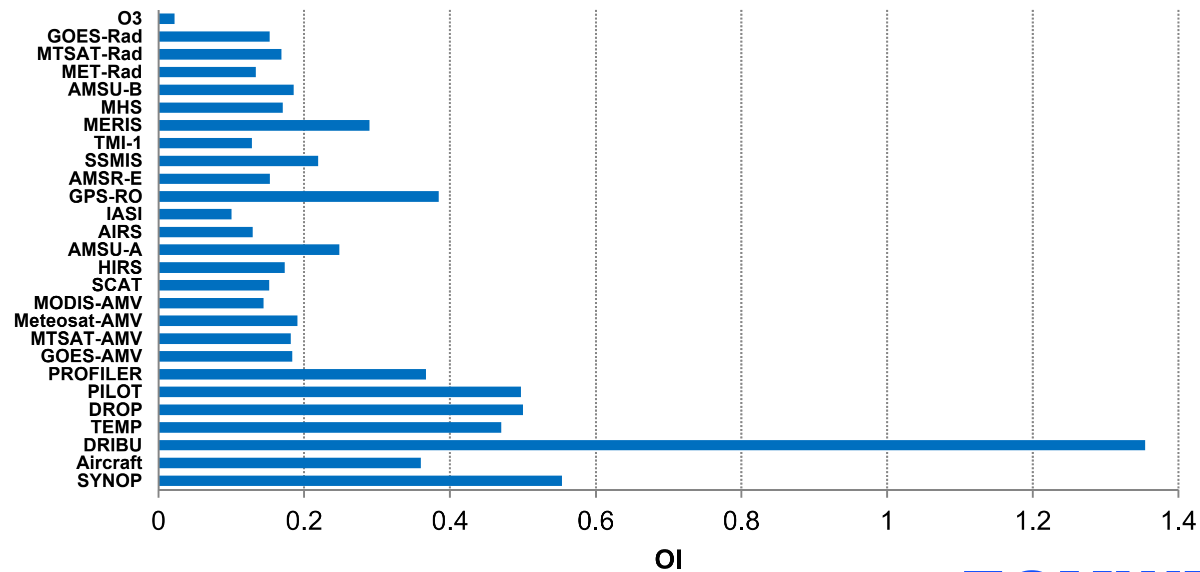
Variable ⇒ **u, v, T, q, p_s, T_b**

Level ⇒

1	**1000-850**
2	**850-700**
3	**700-500**
4	**500-400**
5	**400-300**
6	**300-200**
7	**200-100**
8	**100-70**
9	**70-50**
10	**50-30**
11	**30-0**

DFS and OI



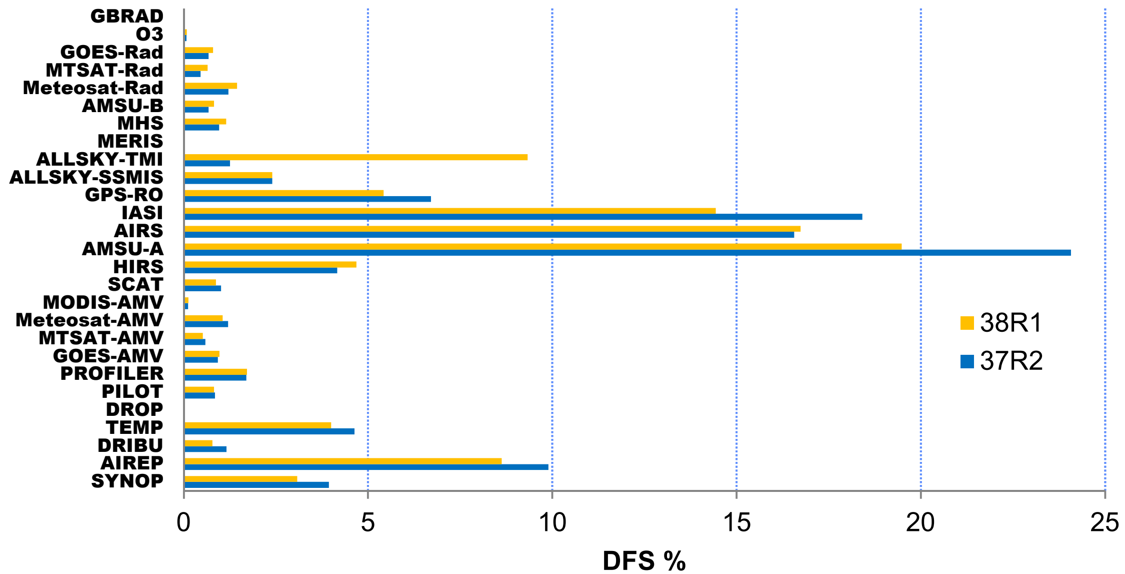


GOI =18%
GBI =82%

ECMWF

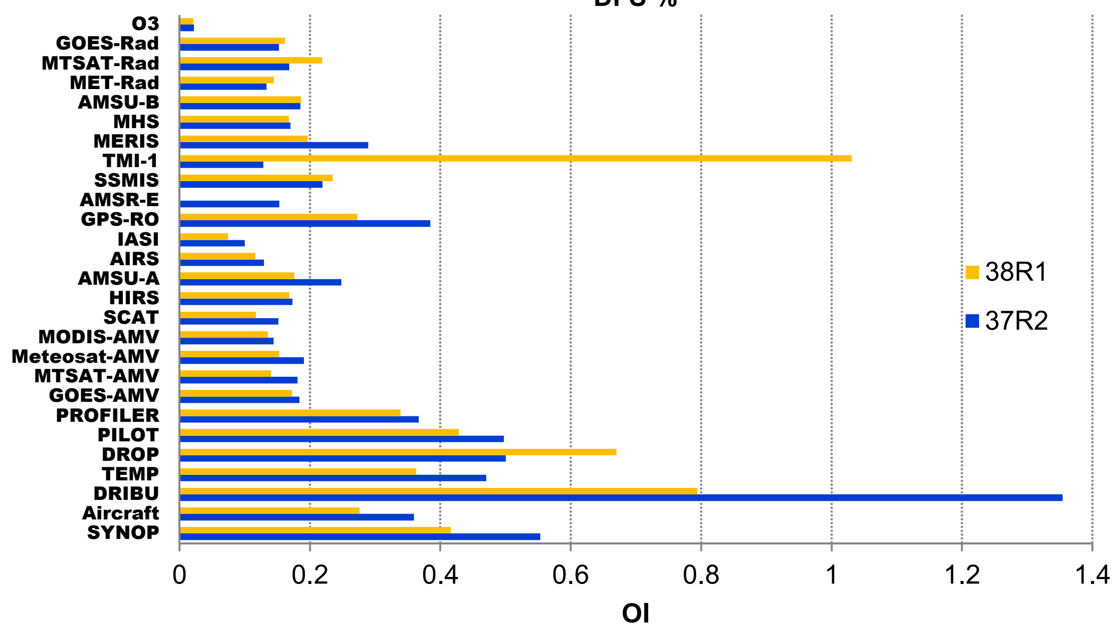
DFS and OI: 2012 Operational versus Next Oper Cycle

GOI =18%
GBI =82%



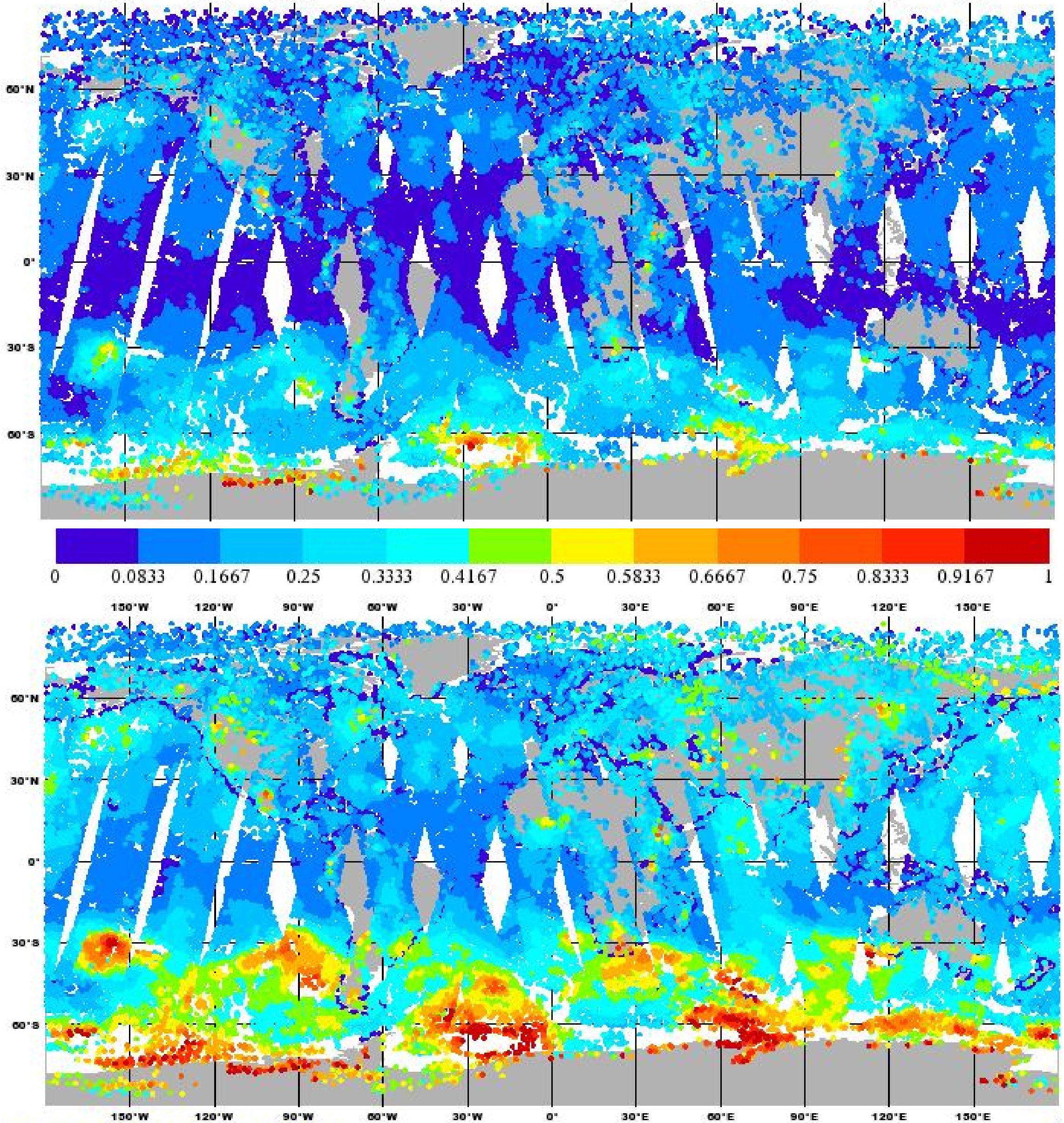
versus

GOI =16%
GBI =84%



ECMWF

Evolution of the B matrix: σ_b computed from EnDA



$$\begin{cases} X^t+\varepsilon^{\text{Stochastics}} \\ y+\varepsilon^o \\ SST+\varepsilon^{SST} \end{cases}$$

AMSU-A ch 6

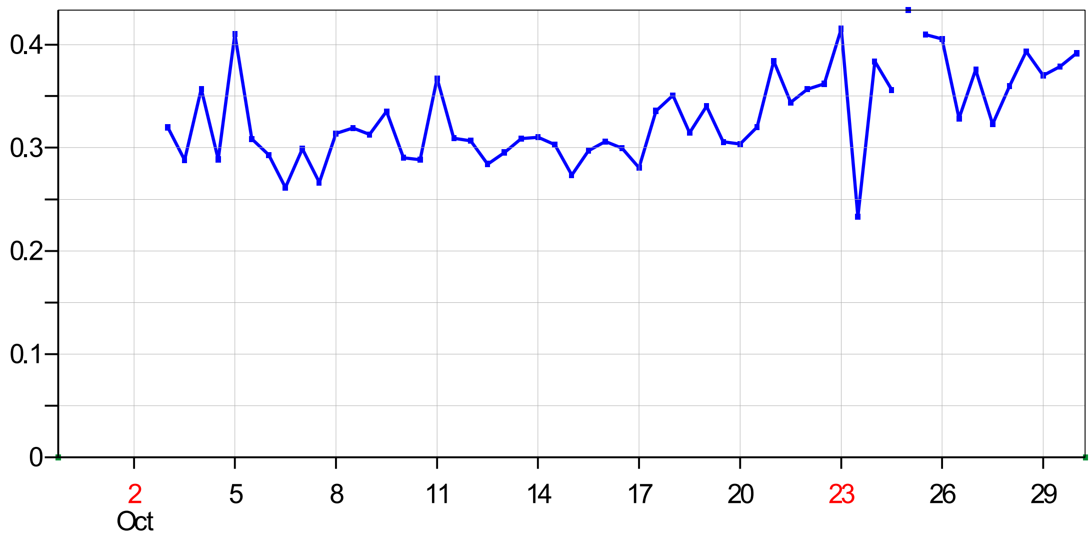
$$\begin{cases} X^b+\varepsilon^b \\ y+\varepsilon^o \\ SST+\varepsilon^{SST} \end{cases}$$

ECMWF

Evolution of the B matrix: σ_b from EnDA

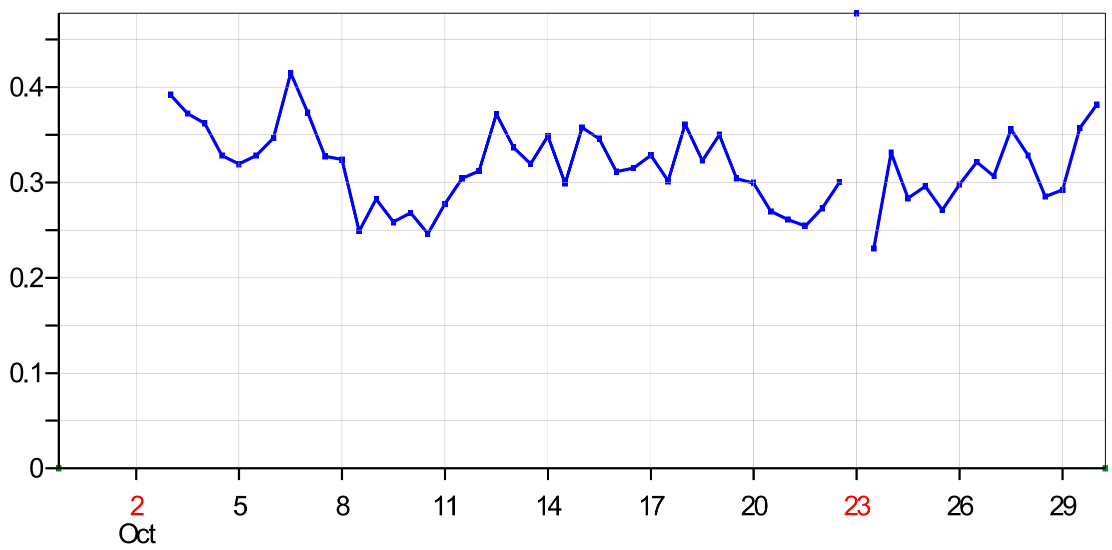
OI

N. Atlantic

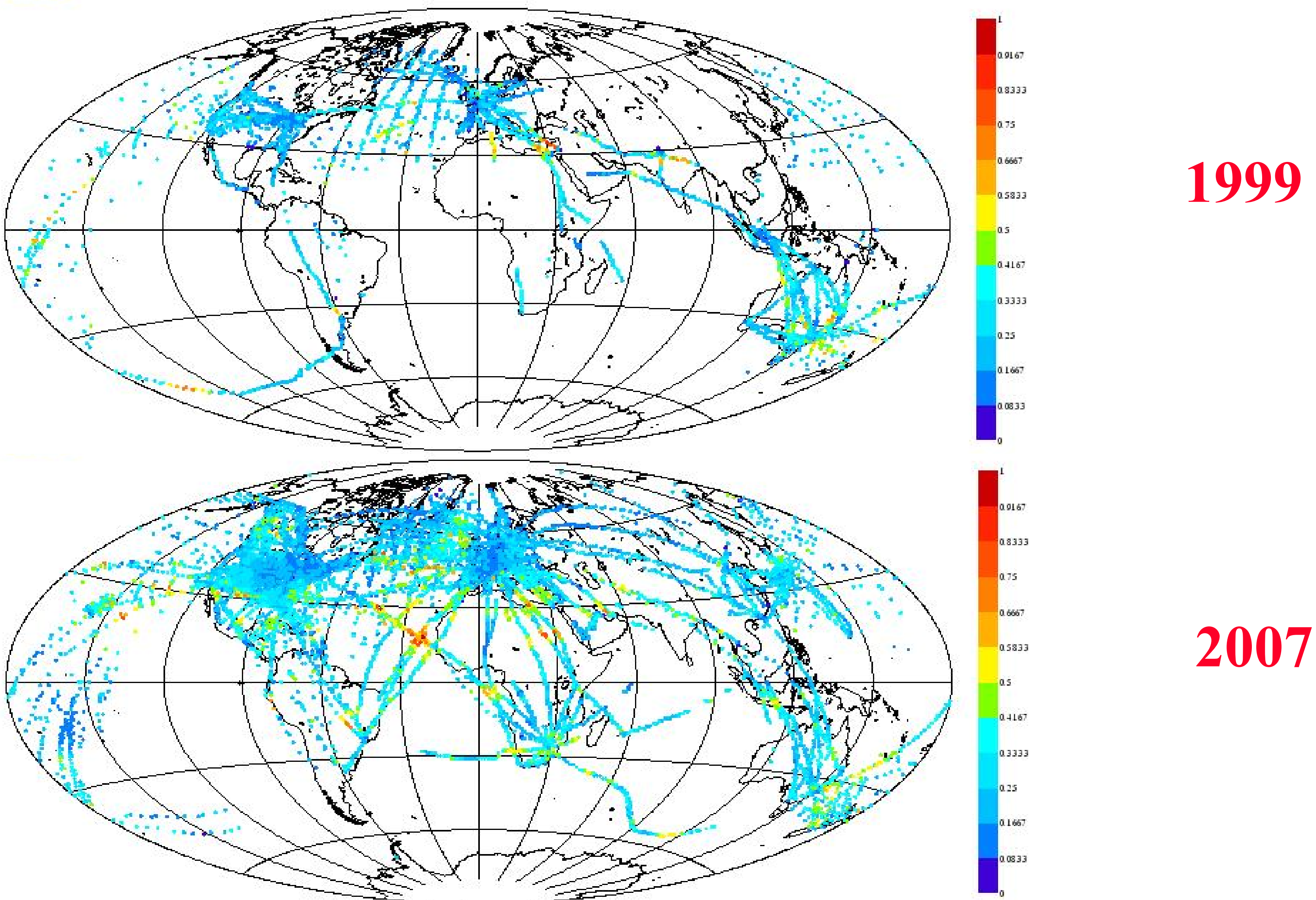


ASCAT

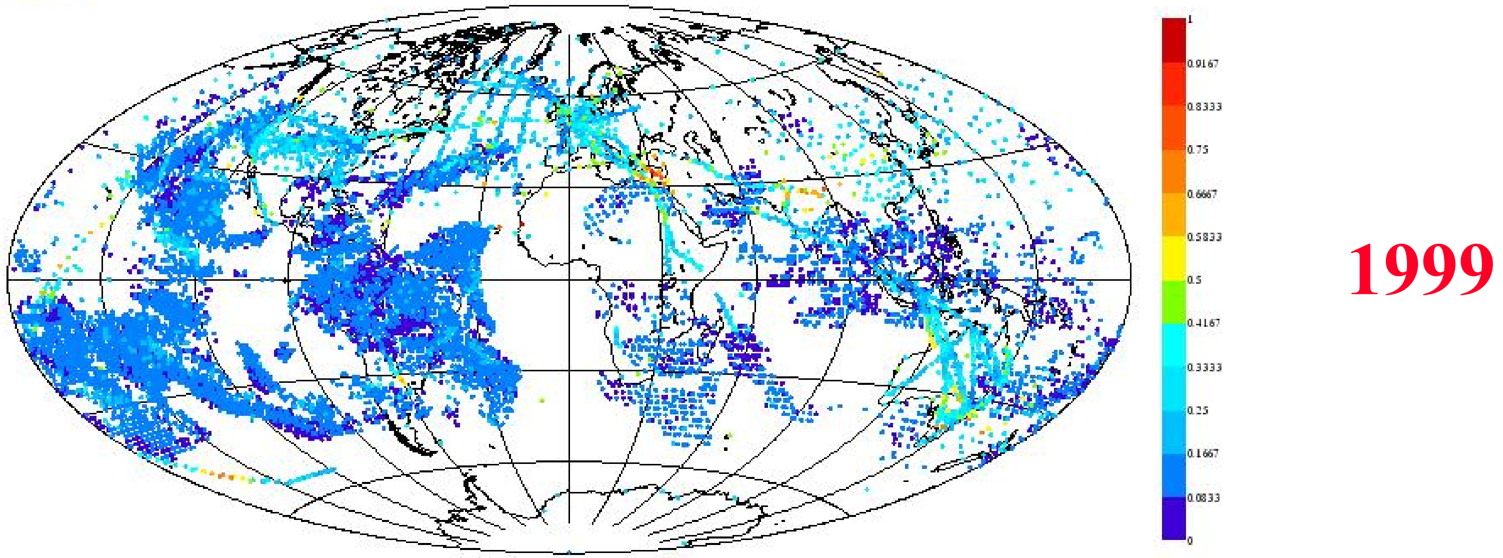
N. Pacific

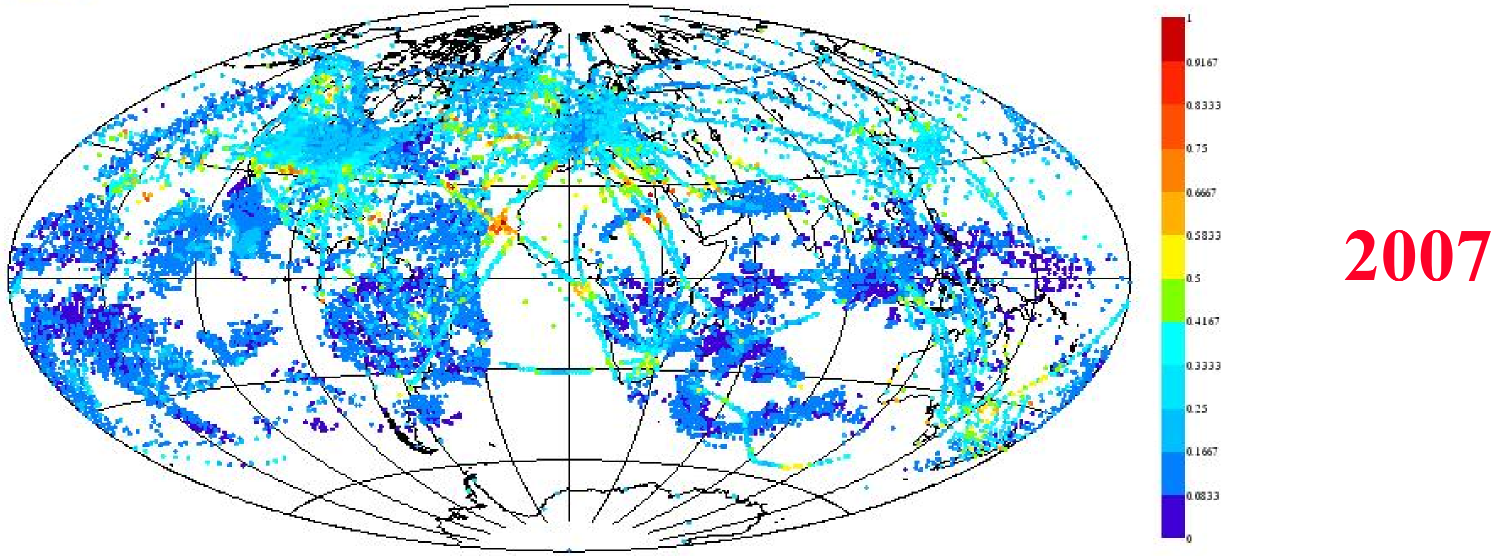


Evolution of the GOS: Interim Reanalysis   Aircraft above 400 hPa



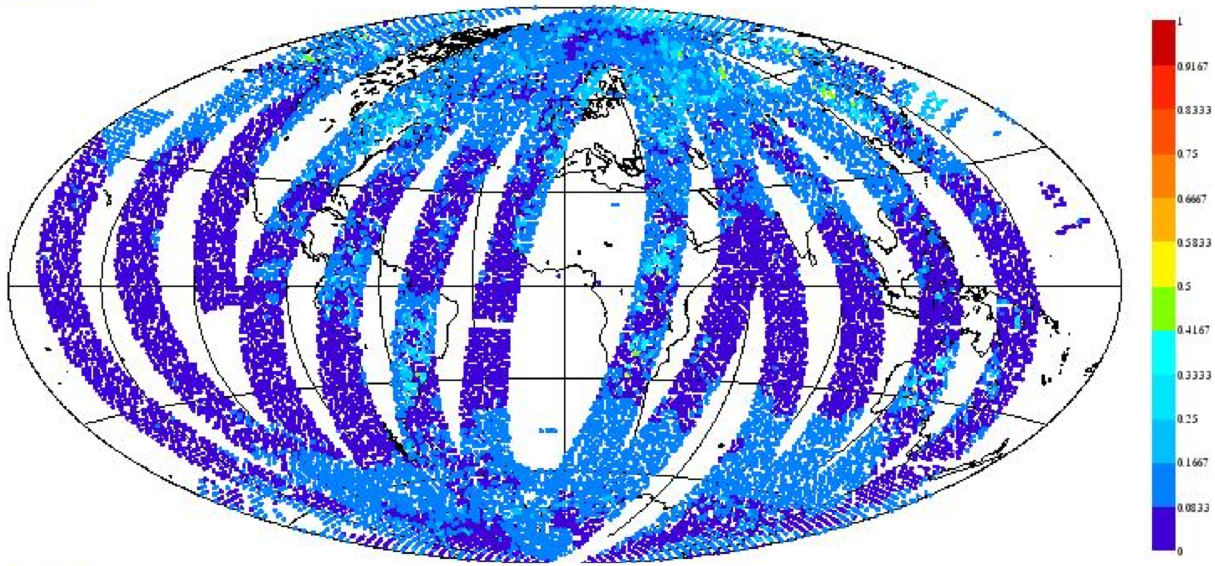
Evolution of the GOS: Interim Reanalysis
U-comp Aircraft, Radiosonde, Vertical Profiler, AMV



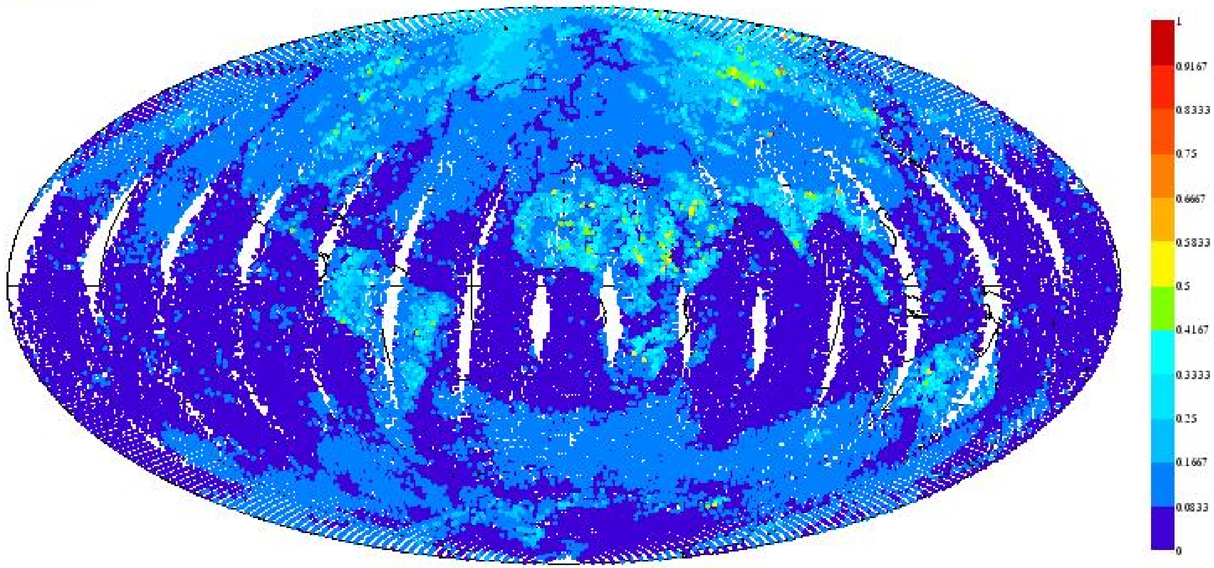


ECMWF

Evolution of the GOS: Interim Reanalysis AMSU-A



1999



2007

ECMWF

Conclusions

- The Influence Matrix is well-known in multi-variate linear regression. It is used to identify influential data. Influence patterns are not part of the estimates of the model but rather are part of the conditions under which the model is estimated

- Disproportionate influence can be due to:
  - **incorrect data (quality control)**
  - **legitimately extreme observations occurrence**
    - → to which extent the estimate depends on these data

$S_{ii}=1$
- Data-sparse → Single observation
- Model under-confident $(1-S_{ii})$

$S_{ii}=0$
- Data-dense
- Model over-confident → tuning $(1-S_{ii})$

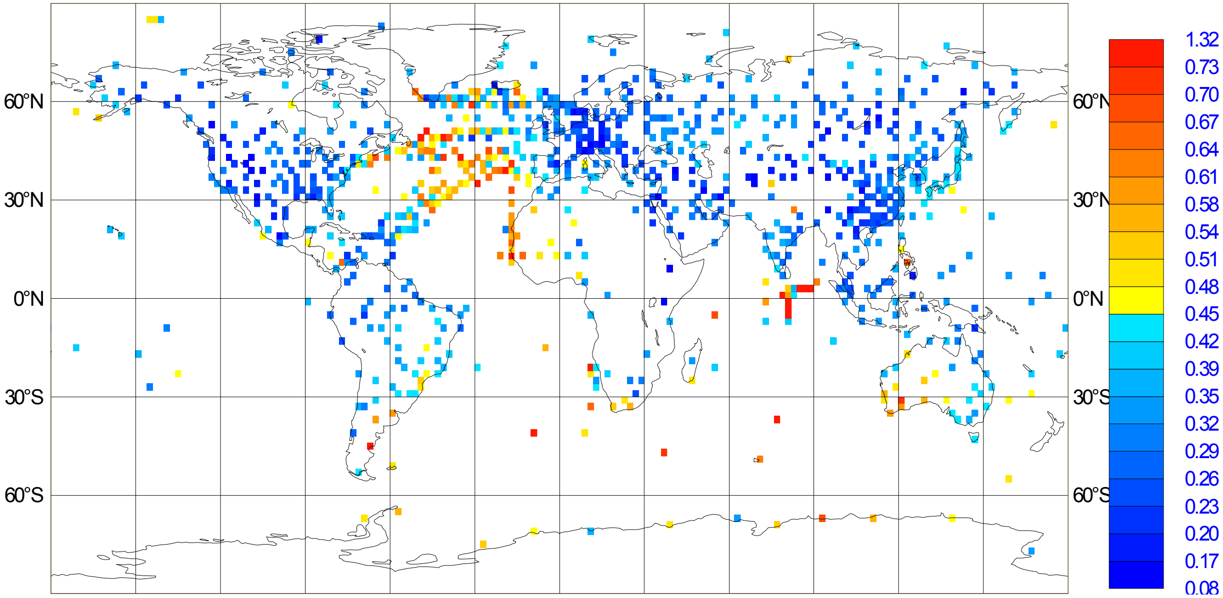
Conclusions

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

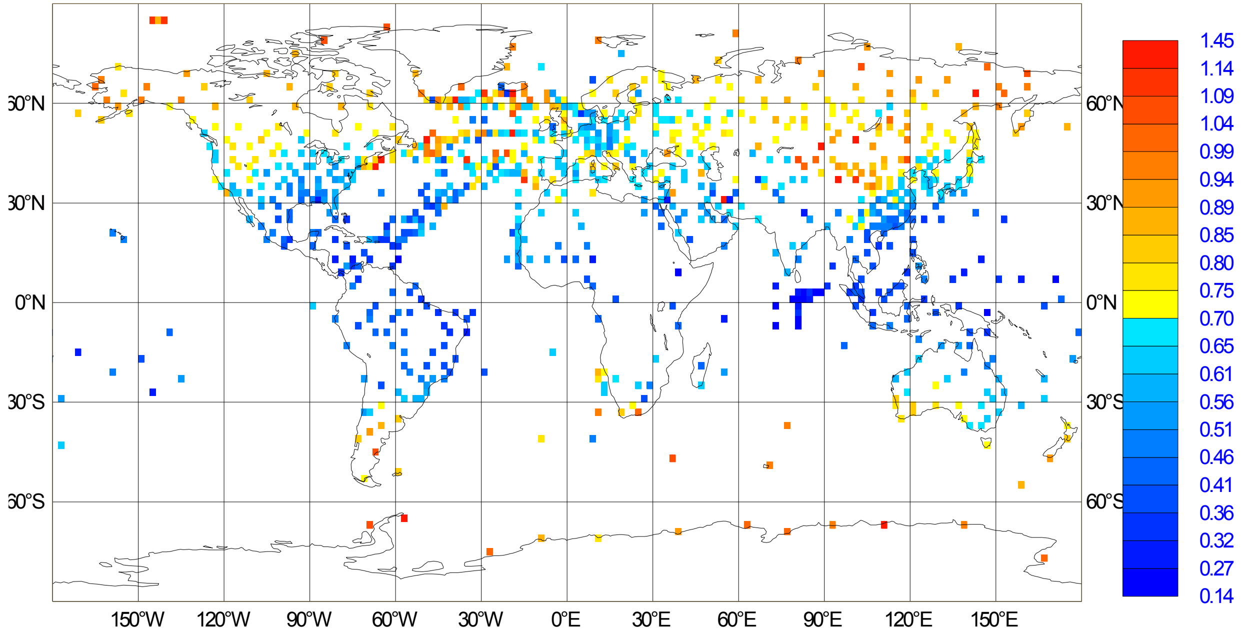
- Observational Influence pattern would provide information on different observation system
    - New observation system
    - Special observing field campaign

- 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

Data density: Radiosonde Observation Influence



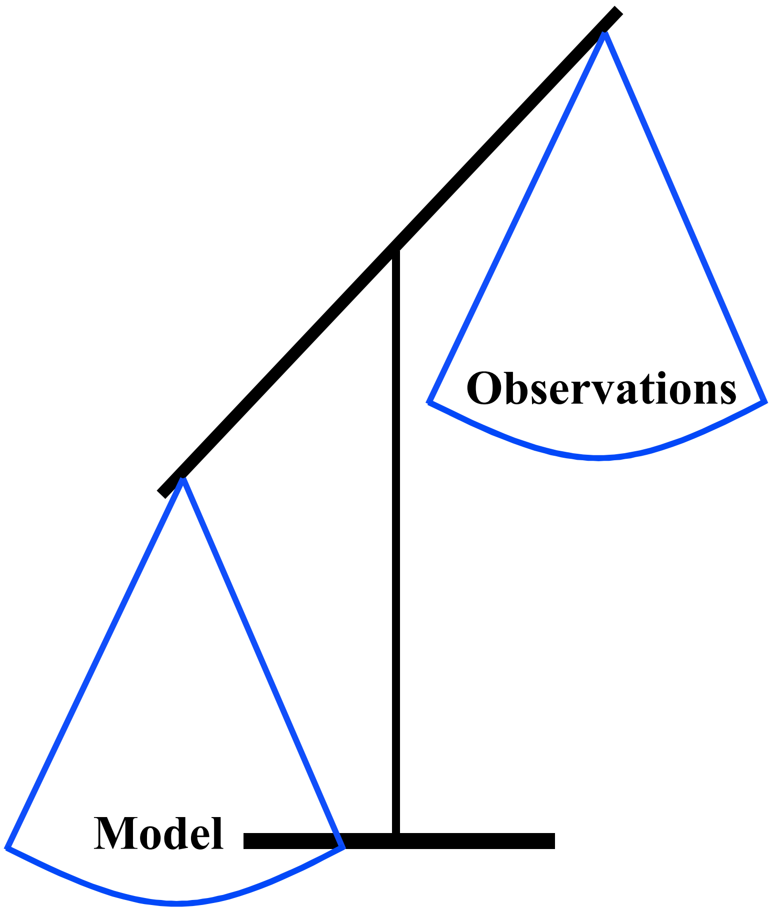
U-Comp
Below 700 hPa



Temperature
Below 700 hPa

ECMWF

Background and Observation Tuning in ECMWF 4D-Var



Influence Matrix Computation

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

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

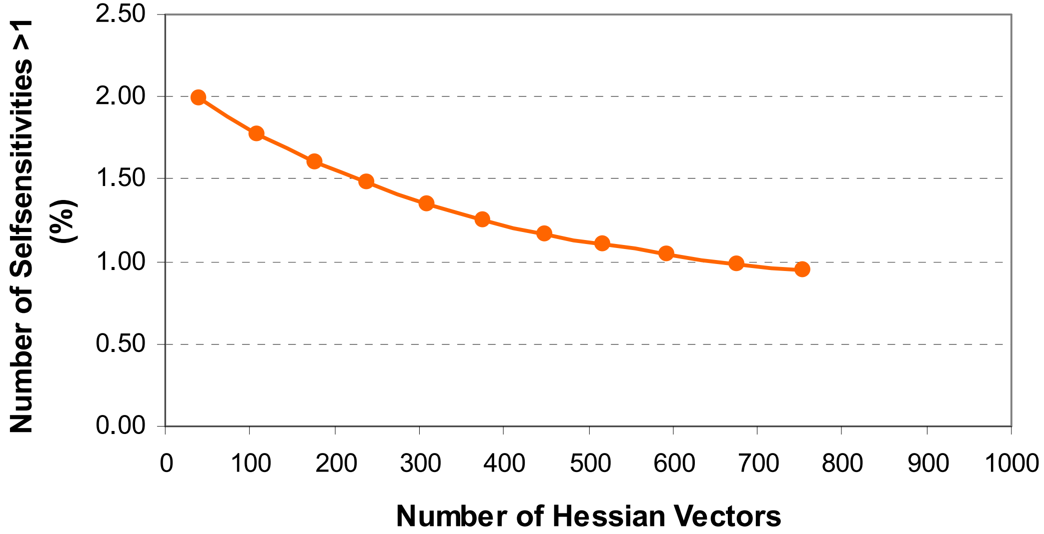
$$(\mathbf{J}'')^{-1} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{L}u_i)(\mathbf{L}u_i)^{T} - \sum_{i=1}^{M}\frac{1-\lambda_i}{\lambda_i}(\mathbf{L}\nu_i)(\mathbf{L}\nu_i)^{T}$$

B

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

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

Hessian Approximation → B-A



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

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

500 random vector to represent B

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

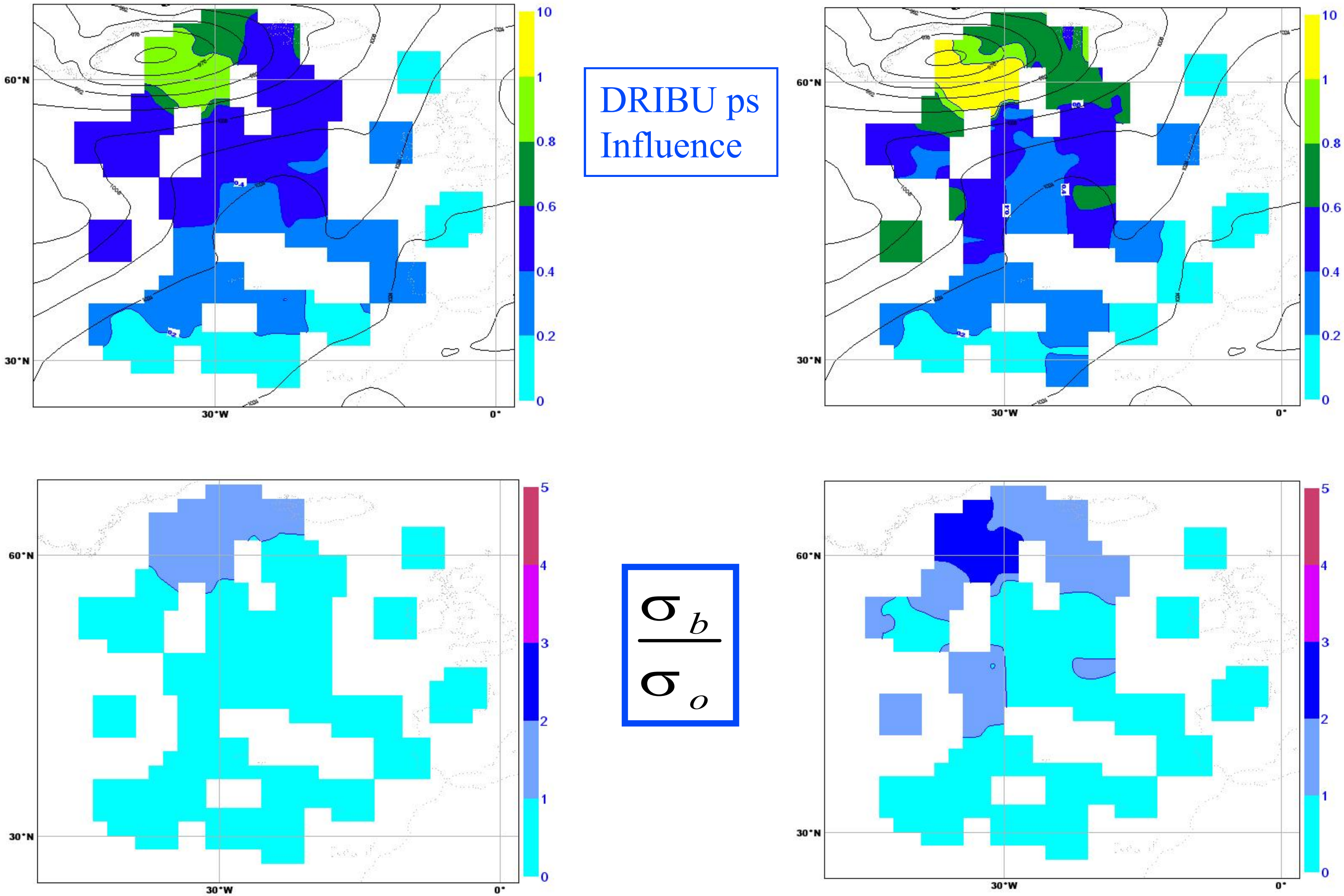
- A Ill-conditioning has effects on the stability and solution accuracy . A measure of ill-conditioning is

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

- 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

Flow Dependent σ_b: MAM^T+Q



DRIBU ps Influence

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

ECMWF