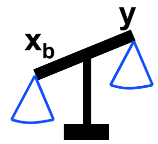
Adjoint diagnostics of data assimilation systems

Carla Cardinali

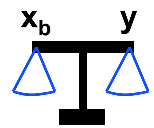
Monitoring The Assimilation System

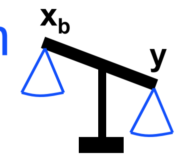
- ECMWF 4D-Var system handles a large variety of space and surface-based observations. It combines observations and atmospheric state a priori information by using a linearized and non-linear forecast model

- Effective monitoring of a such a complex system with 10^8 degrees of freedom and 10^7 observations is a necessity. Not just a few indicators but a more complex set of measures to answer questions like is needed:
  - 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?
  - How much is the observation impact on the forecast?



Analysis Solution





Model space

Observation space

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

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

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

K(qxp) gain matrix
H(pxq) Jacobian matrix

B(qxq)=Var(x_b)
R(pxp)=Var(y)

$$\frac{\partial \mathbf{x}_a}{\partial \mathbf{y}} = \mathbf{K}^T$$

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

Forecast Sensitivity to the Observation

Analysis Sensitivity to the Observation

Outline

- **Analysis Sensitivity to Observation or Observation Influence**
  - **Ordinary Least Square method**
  - **Findings related to data influence and information content**
  - **Toy model: 2 observations**
  - **Monitoring the observation influence**
- **Forecast Sensitivity to Observation or Observation Impact on Forecast**
  - **Equation**
  - **FSO Diagnostic Tool**
  - **Monitoring the forecast impact: ECMWF Operational configuration**

Observation Influence: Influence Matrix in OLS

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

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

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

- **The fitted response is** $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$

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

S (*mxm*)

$$\mathbf{S} = \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{y}} \quad \begin{cases} S_{ij} = \dfrac{\partial \hat{y}_i}{\partial y_j} \\ S_{ii} = \dfrac{\partial \hat{y}_i}{\partial y_i} \end{cases}$$

symmetric, idempotent and positive definite matrix

$$0 \leq S_{ii} \leq 1$$

Observation Influence: Influence Matrix Related Findings

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

- **The change in the estimate that occurs 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 ŷ and** S_{ii}

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

Whaba 1990

Observation Influence: Influence Matrix in a Generalized Least Square Method

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

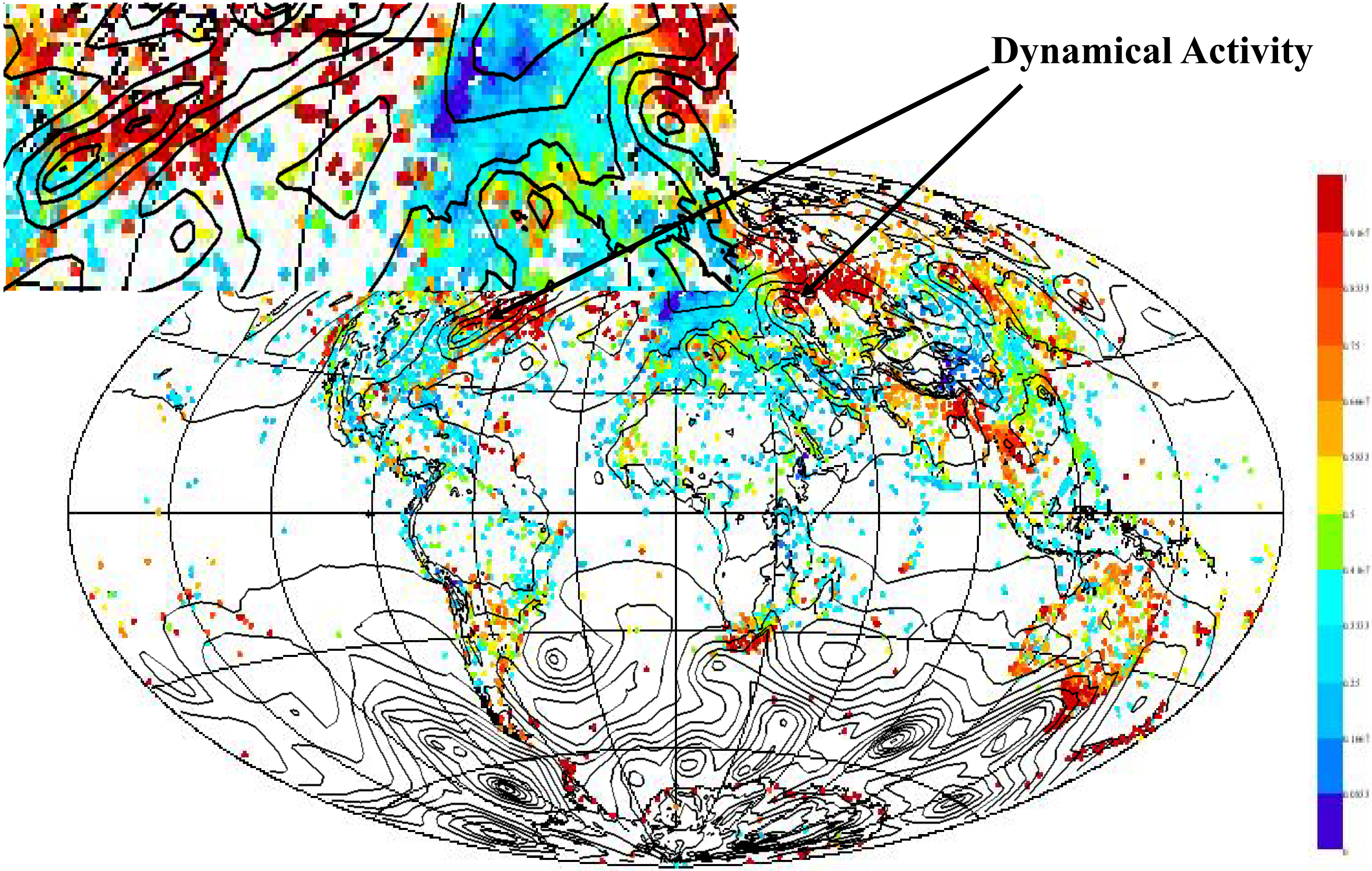
$$\mathbf{I} - \mathbf{S} = \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{Hx}_b} = Background - Influence$$

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

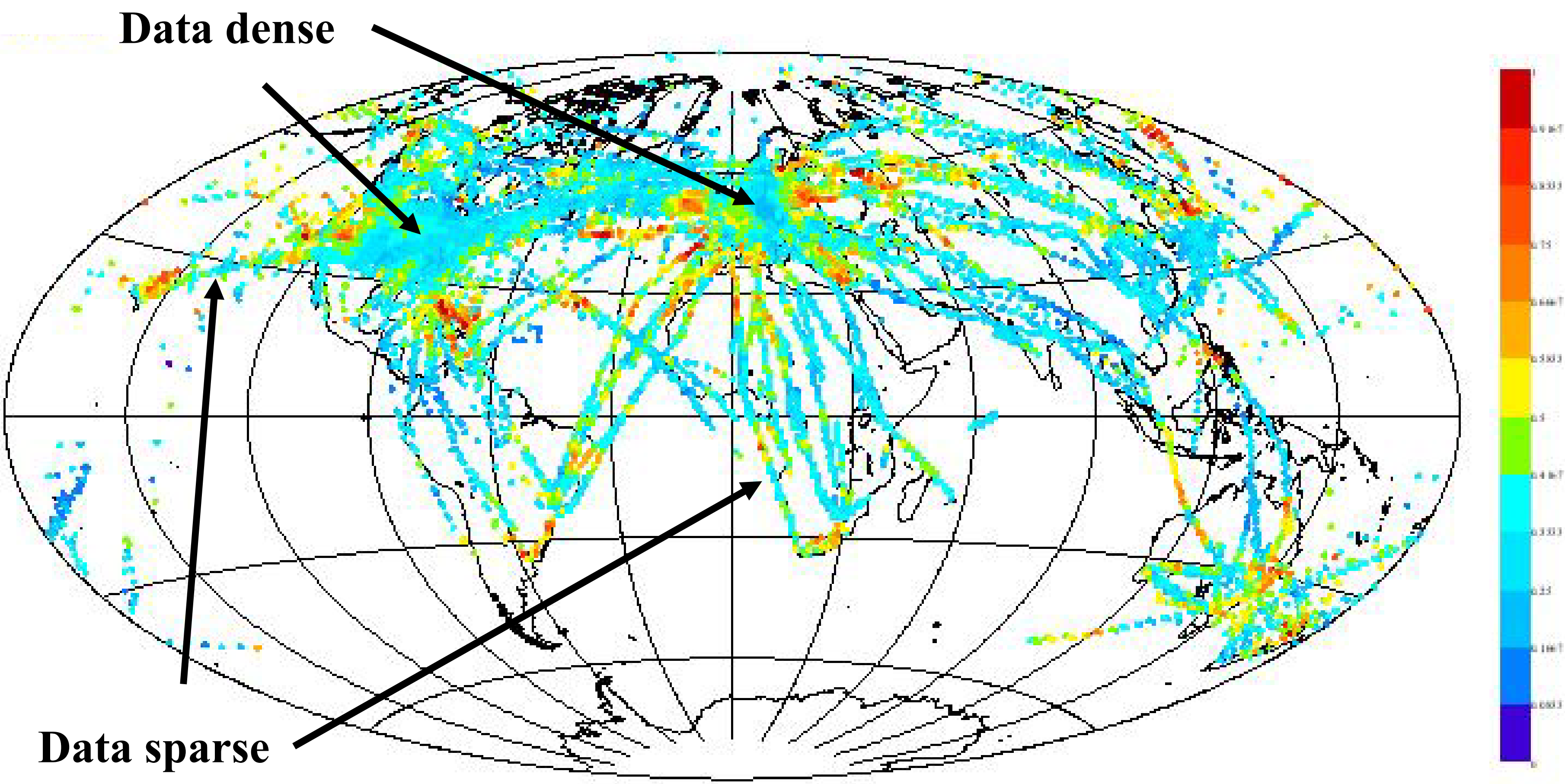
$$\sum_{i=1}^{N} S_{ii} = \text{Information Content or DFS}$$

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

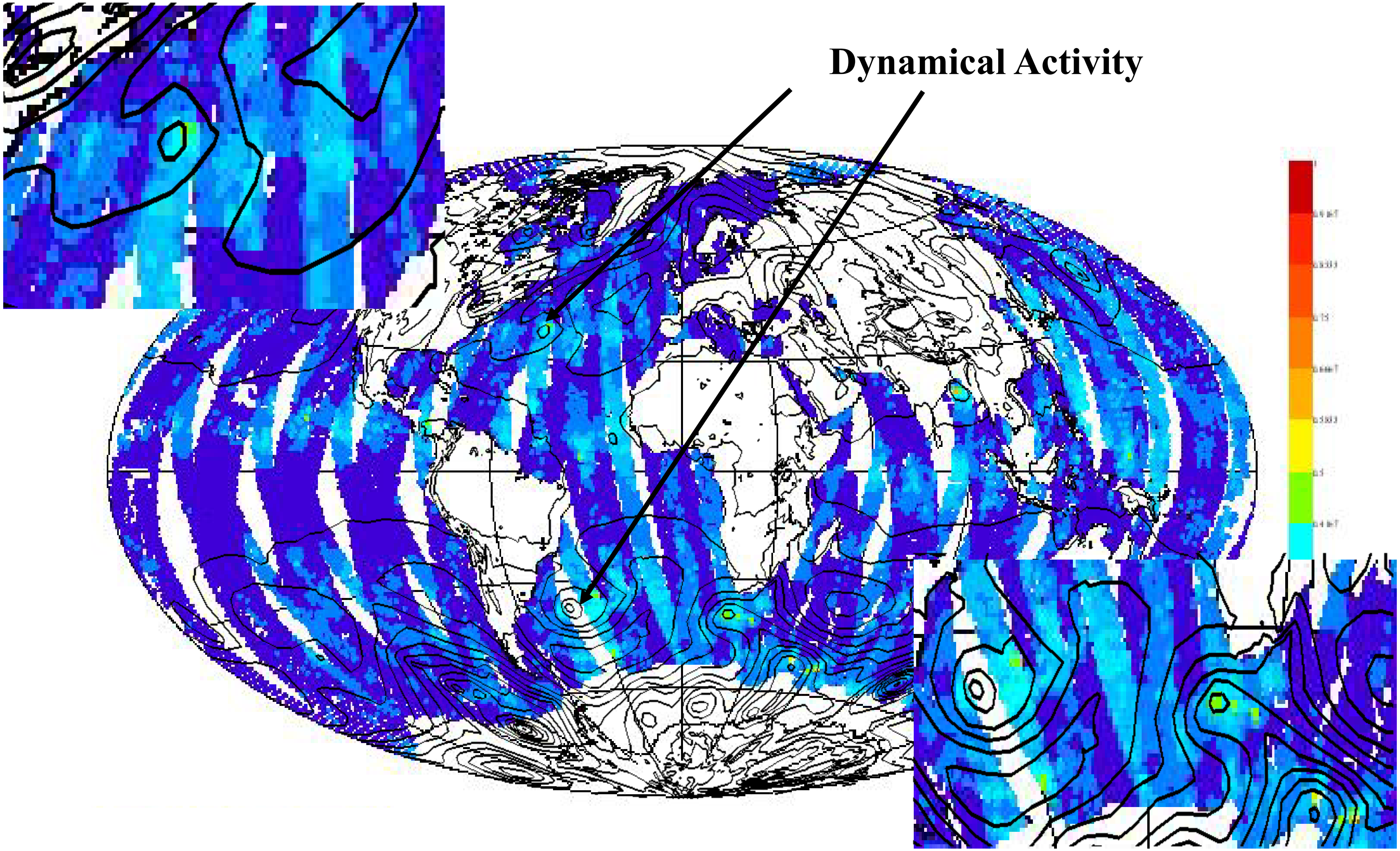
Synop Surface Pressure Influence



Aircraft 250 hPa U-Comp Influence



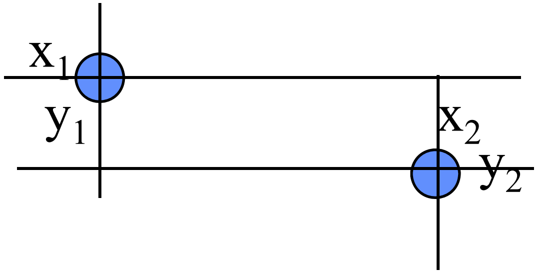
Scatterometer U-Comp Influence



ECMWF

Toy Model: 2 Observations

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



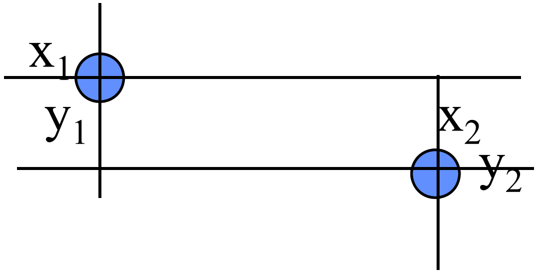
$$\mathbf{H}=\mathbf{I} \qquad \mathbf{R}=\sigma_o^2\mathbf{I} \qquad \mathbf{B}=\begin{pmatrix}\sigma_b^2 & \alpha \\ \alpha & \sigma_b^2\end{pmatrix} \qquad r=\frac{\sigma_o^2}{\sigma_b^2}$$

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

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

Toy Model: 2 Observations

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



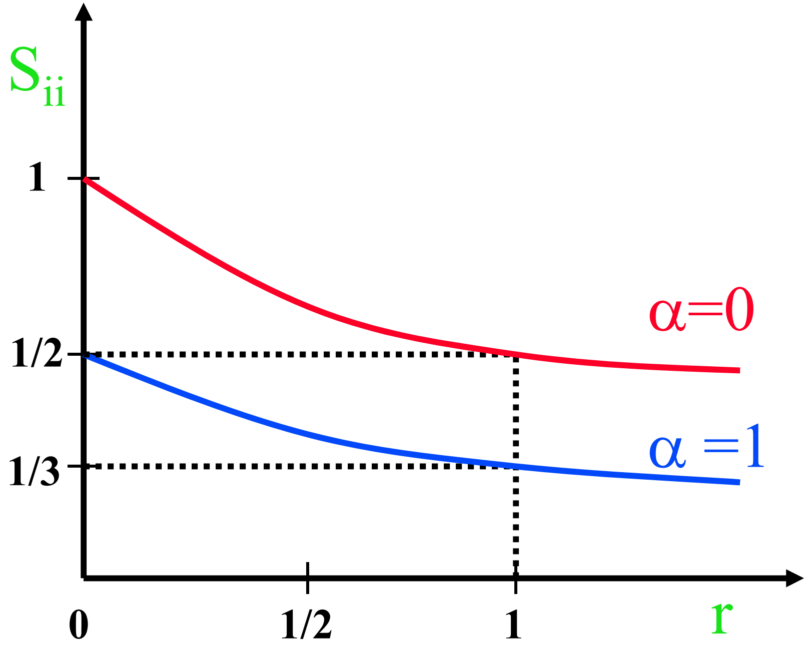
$$\mathbf{H} = \mathbf{I} \qquad \mathbf{R} = \sigma_o^2 \mathbf{I} \qquad \mathbf{B} = \begin{pmatrix} \sigma_b^2 & \alpha \\ \alpha & \sigma_b^2 \end{pmatrix}$$

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

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

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

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

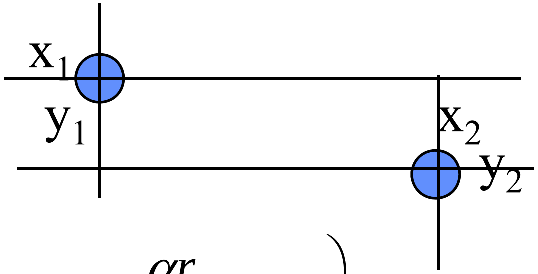


ECMWF

(1) Consideration

- **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)**
- **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)**

Toy Model: 2 Observations



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

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

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

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

$\alpha = 0$:

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

$\alpha = 1$:

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

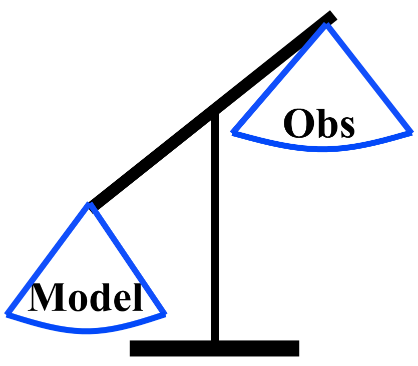
(2) Consideration

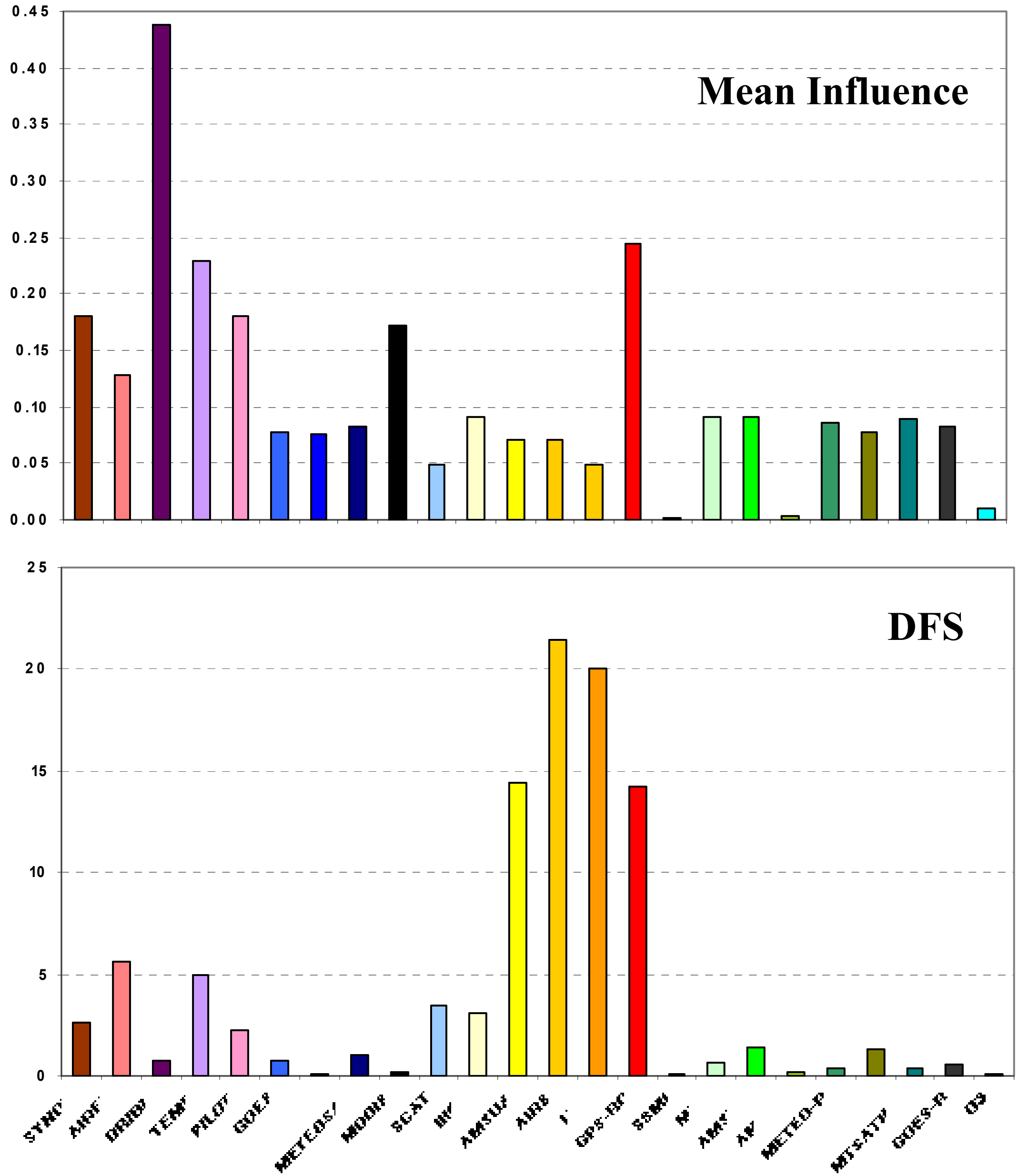
- **When observation and background have similar accuracies (r), the estimate $\hat{y}_1$ depends on y_1 and x_1 and an additional term due to the second observation. We see that if R is diagonal the observational contribution is devaluated with respect to the background because a group of correlated 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**

ECMWF Operational Average Influence and Information Content

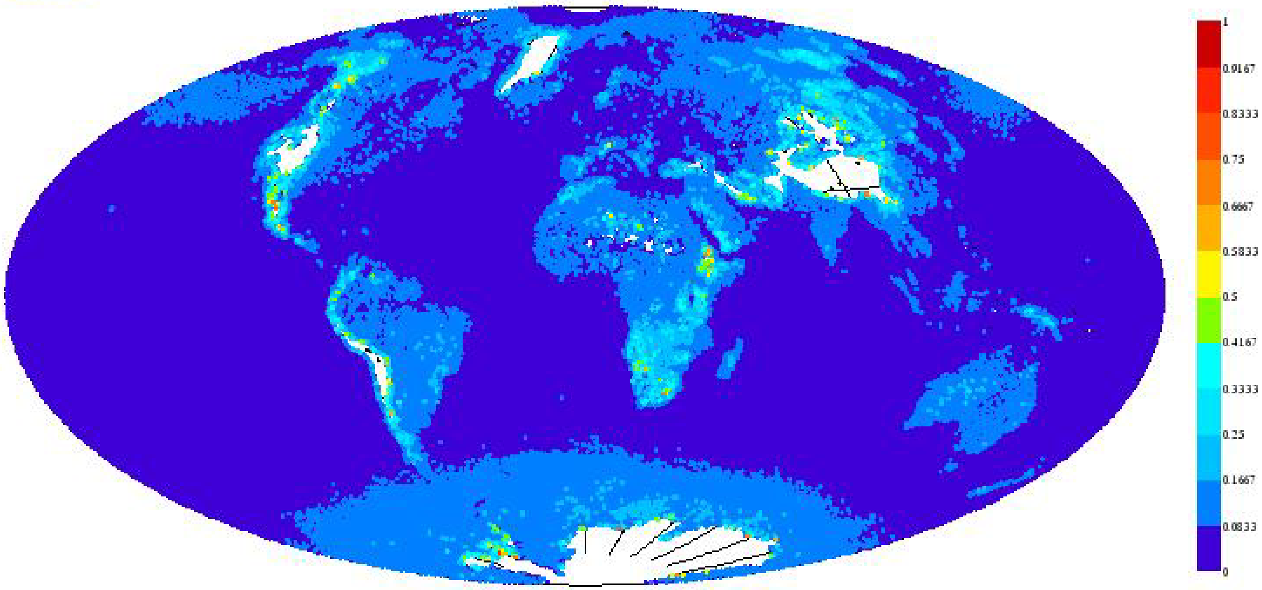
Global Observation Influence
GI=7%

Global Background Influence
I-GI=93%





Evolution of the B matrix: B computed from EnDA

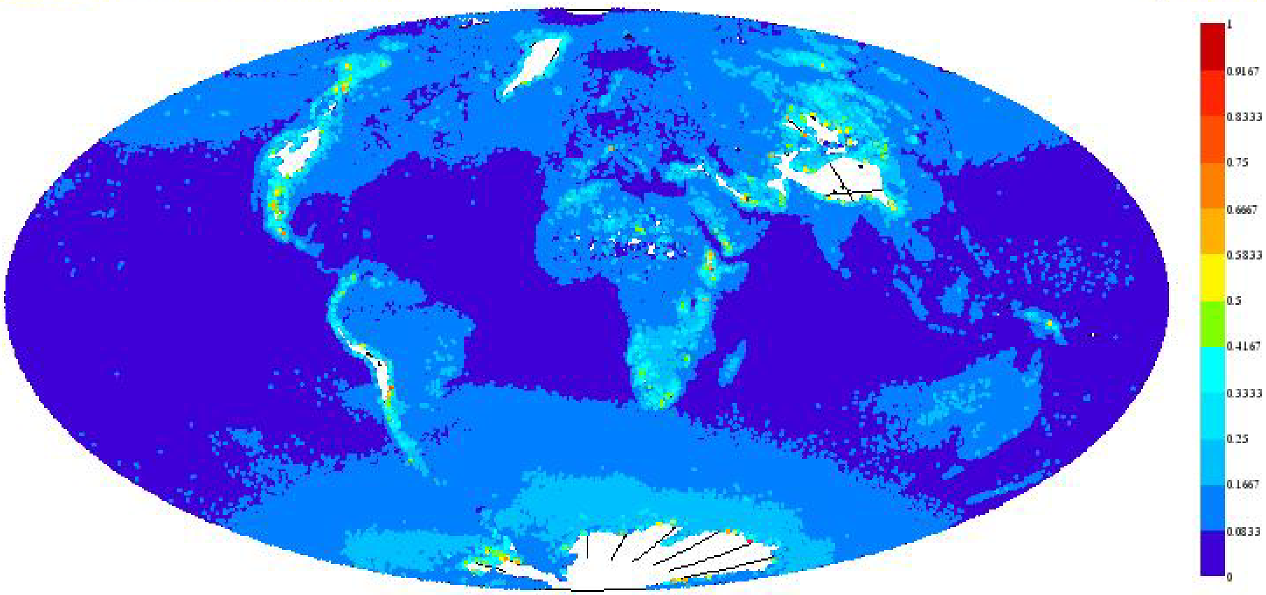


$X^t+\varepsilon^{Stochastics}$

$y+\varepsilon^o$

$SST+\varepsilon^{SST}$

AMSU-A ch6

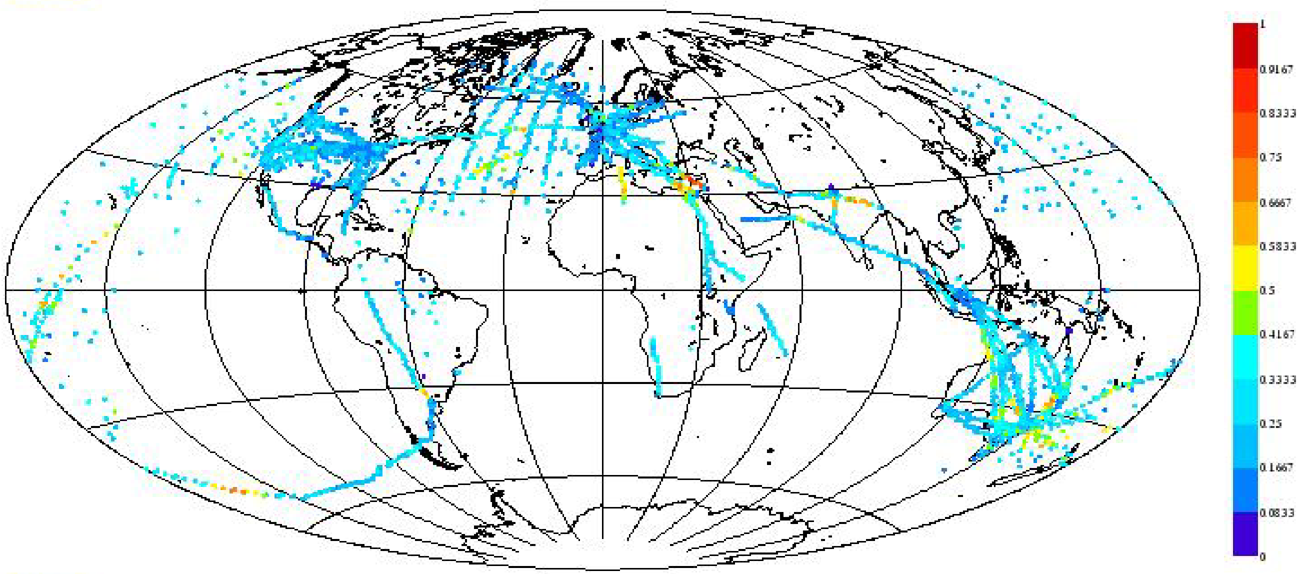


$X^b+\varepsilon^b$

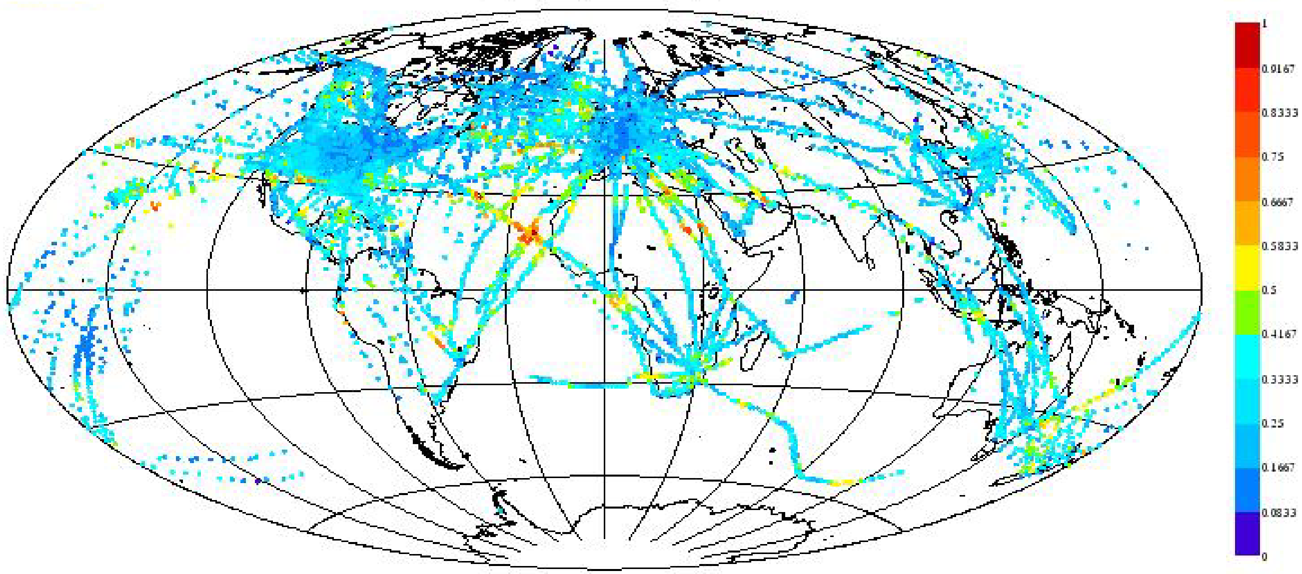
$y+\varepsilon^o$

$SST+\varepsilon^{SST}$

Evolution of the GOS: Interim Reanalysis
Aircraft 200-300 hPa

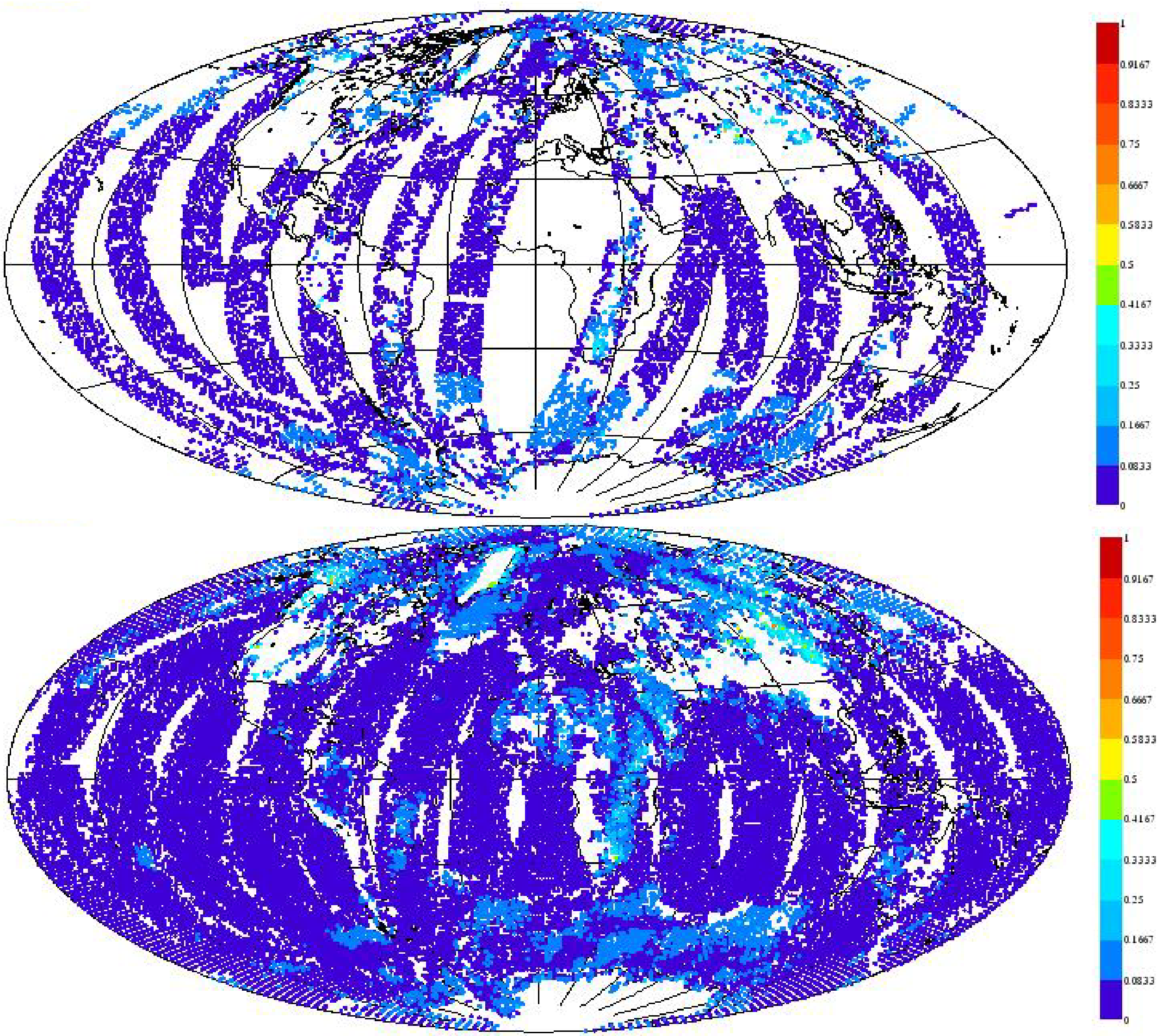


1999



2007

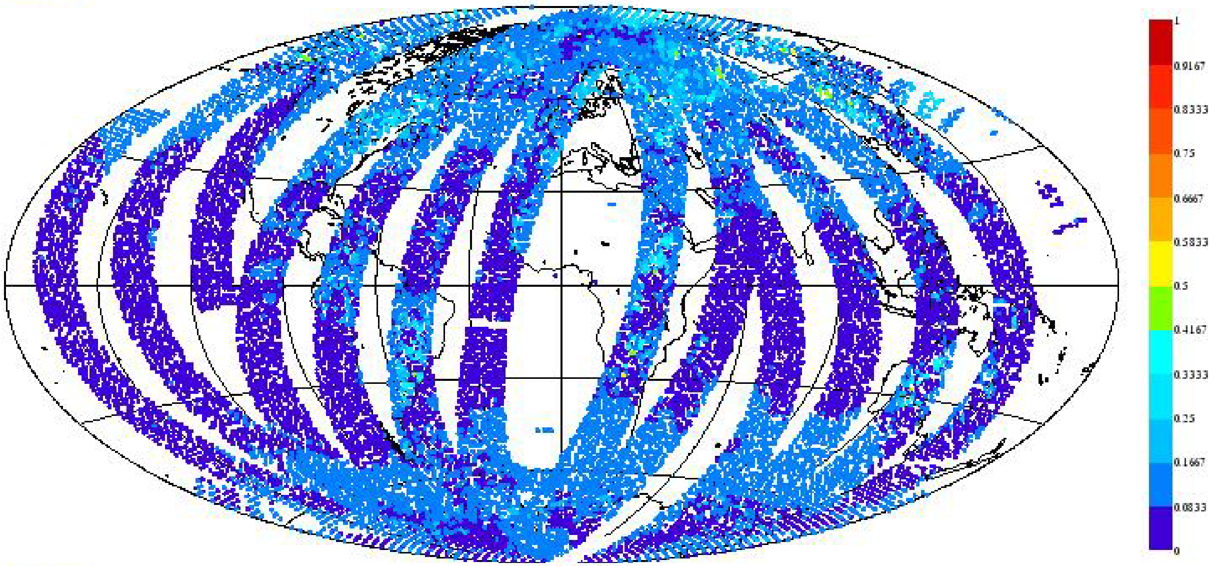
Evolution of the GOS: Interim Reanalysis
AMSU-A ch6



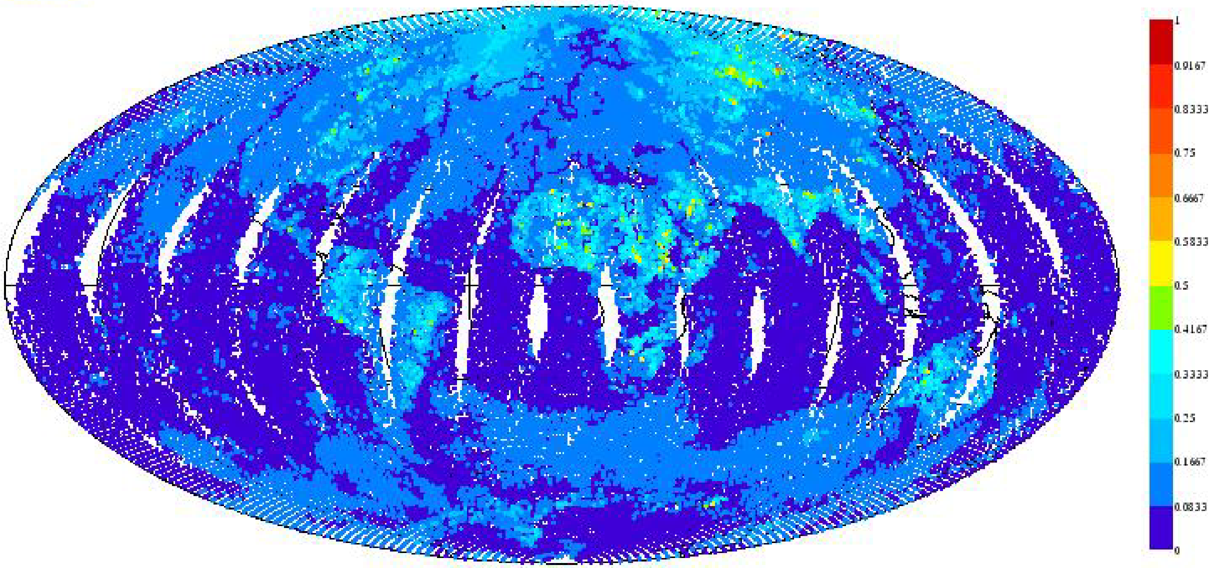
1999

2007

Evolution of the GOS: Interim Reanalysis AMSU-A



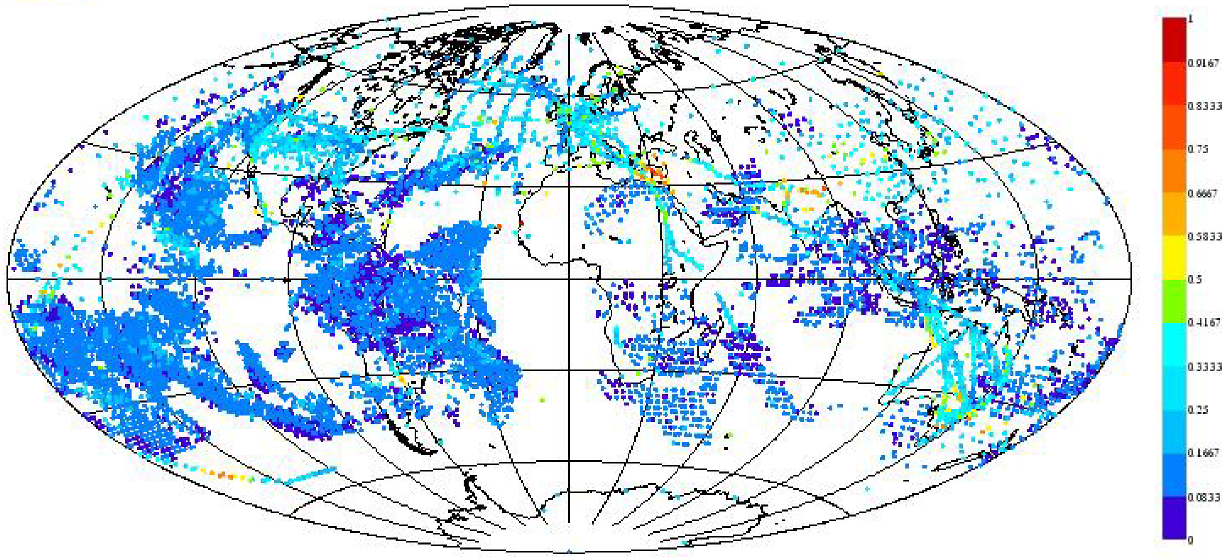
1999



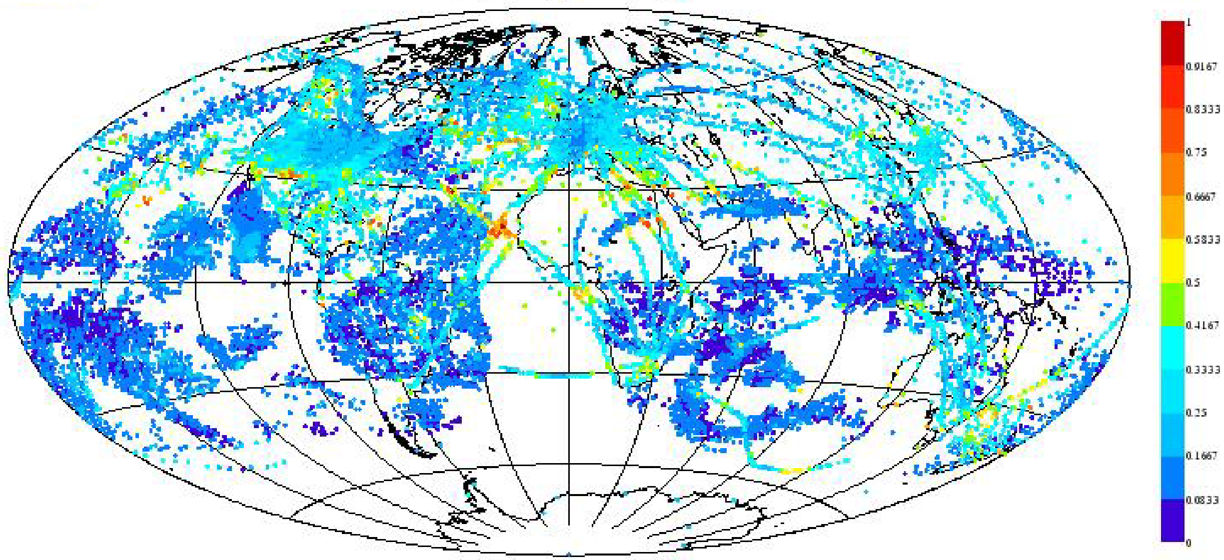
2007

ECMWF

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



1999



2007

Observation Influence Conclusion

- 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
$S_{ii}=1$ → Model under-confident $(1-S_{ii})$

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

- 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**

Forecast sensitivity to observation: Equations
from a Roger Daley idea

J is a measure of the forecast error e.g dry energy norm

$$\frac{\partial J}{\partial \mathbf{y}} = \frac{\partial \mathbf{x}_a}{\partial \mathbf{y}} \frac{\partial J}{\partial \mathbf{x}_a}$$

$$\frac{\partial J}{\partial \mathbf{x}_a}$$

Forecast error sensitivity to the analysis
Rabier F, *et al*. 1996

$$\frac{\partial \mathbf{x}_a}{\partial \mathbf{y}} = \mathbf{K}^T = \mathbf{R}^{-1}\mathbf{H}(\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}$$

$$\frac{\partial J}{\partial \mathbf{y}} = \mathbf{R}^{-1}\mathbf{H}\mathbf{A}\frac{\partial J}{\partial \mathbf{x}_a}$$

1) $\mathbf{A}^{-1}\mathbf{z} = \dfrac{\partial J}{\partial \mathbf{x}_a}$

2) $\dfrac{\partial J}{\partial \mathbf{y}} = \mathbf{R}^{-1}\mathbf{H}\mathbf{z}$

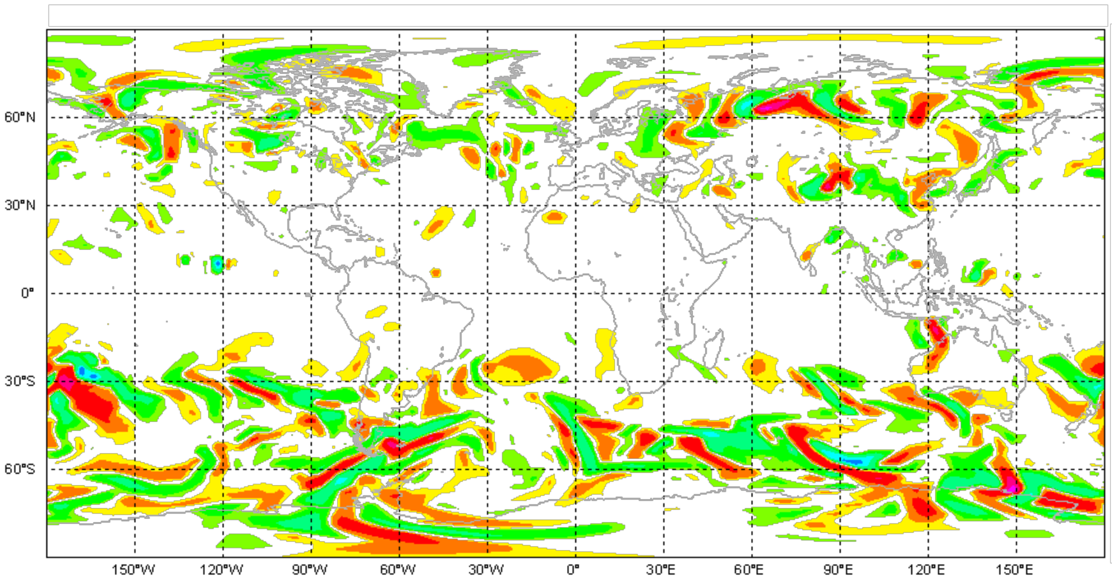
Krylov Subspace Method
Henk A. van der Vorst 2003

- **Compute the forecast impact or forecast error variation δJ**

$$< \frac{\partial J}{\partial \mathbf{x}_a}, \delta\mathbf{x}_a > = < \frac{\partial J}{\partial \mathbf{x}_a}, \mathbf{x}_a - \mathbf{x}_b > = < \frac{\partial J}{\partial \mathbf{x}_a}, \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}_b) > = < \mathbf{K}^T \frac{\partial J}{\partial \mathbf{x}_a}, (\mathbf{y} - \mathbf{H}\mathbf{x}_b) > = < \frac{\partial J}{\partial \mathbf{y}}, \delta\mathbf{y} >$$

$$\delta J = \frac{\partial J}{\partial \mathbf{y}}(\mathbf{y} - \mathbf{H}\mathbf{x}_b)$$

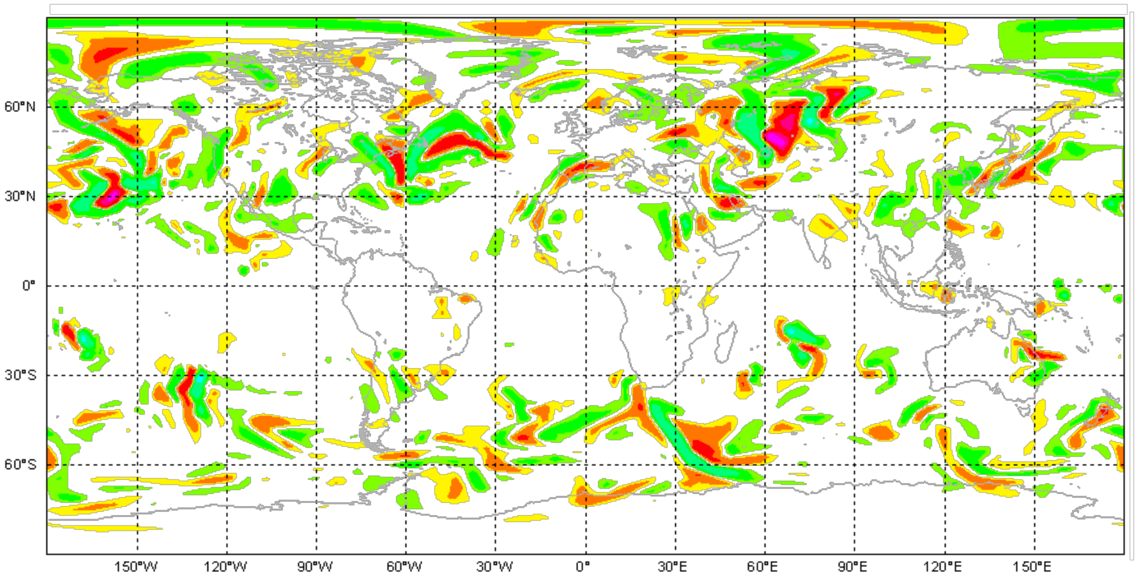
Forecast Sensitivity to Observation: Sensitivity Gradient



Summer

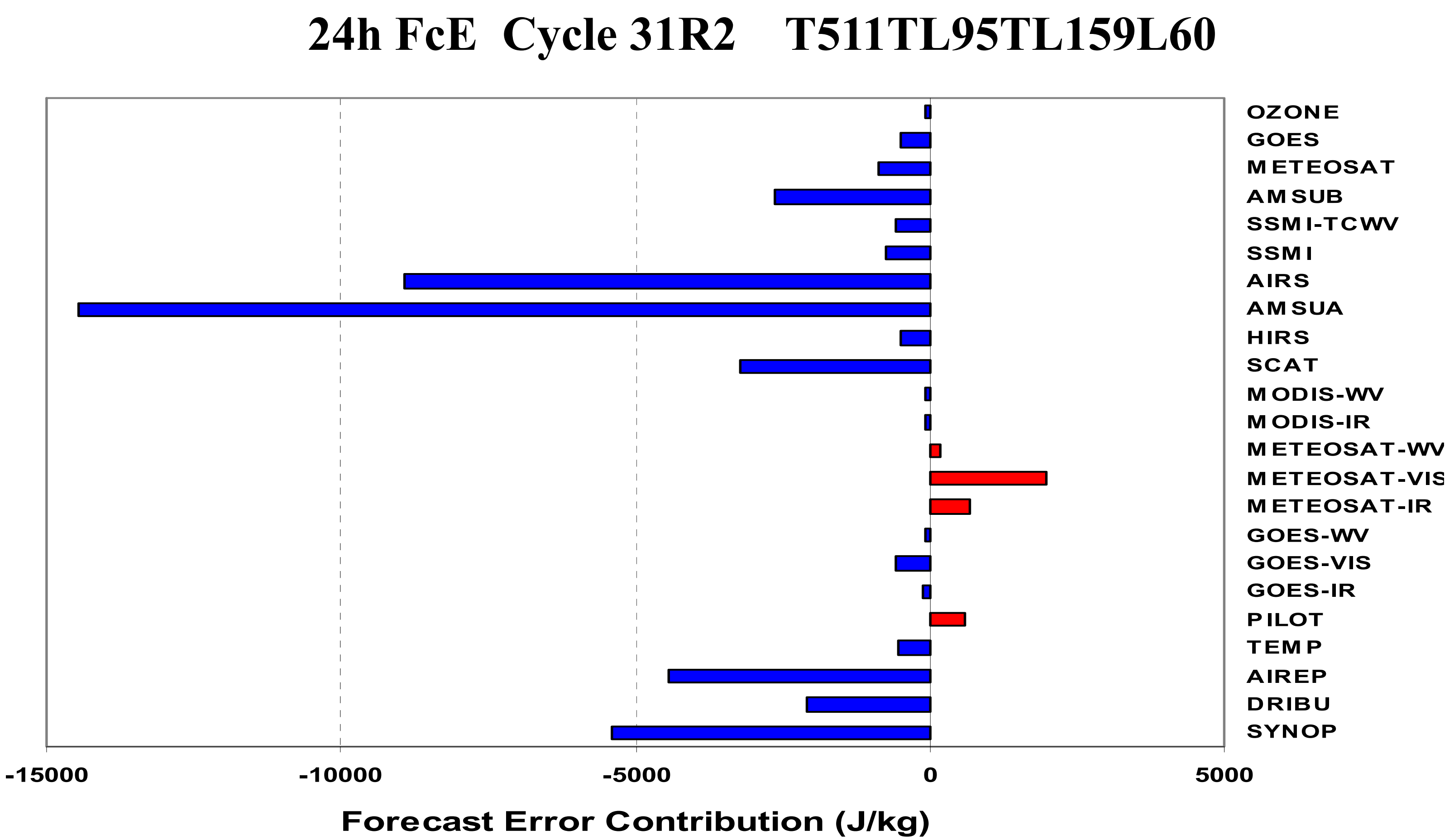
$$\frac{\partial J}{\partial \mathrm{T}_{39}}$$

Dry Energy norm



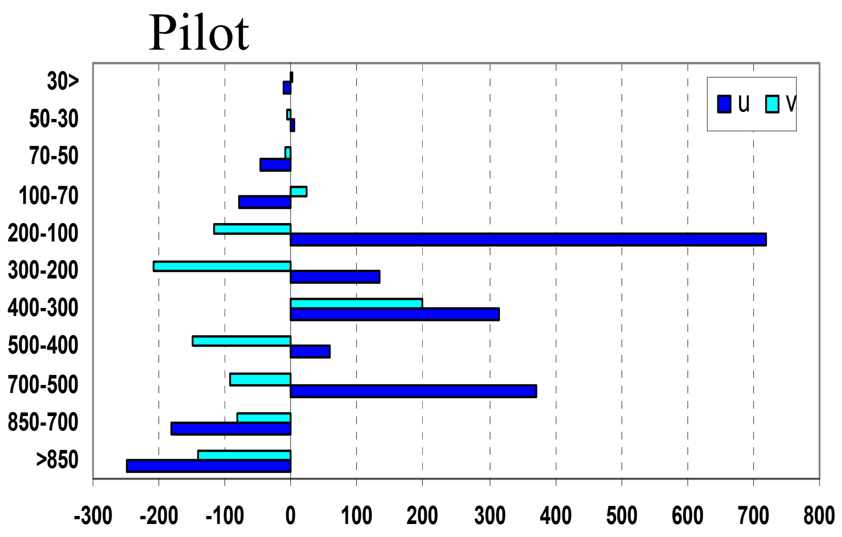
Winter

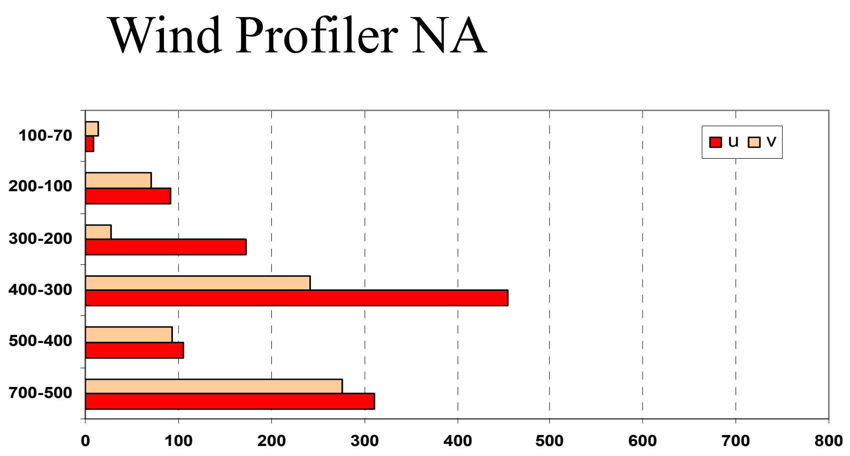
FSO Monitoring ECMWF System: Summer 2006

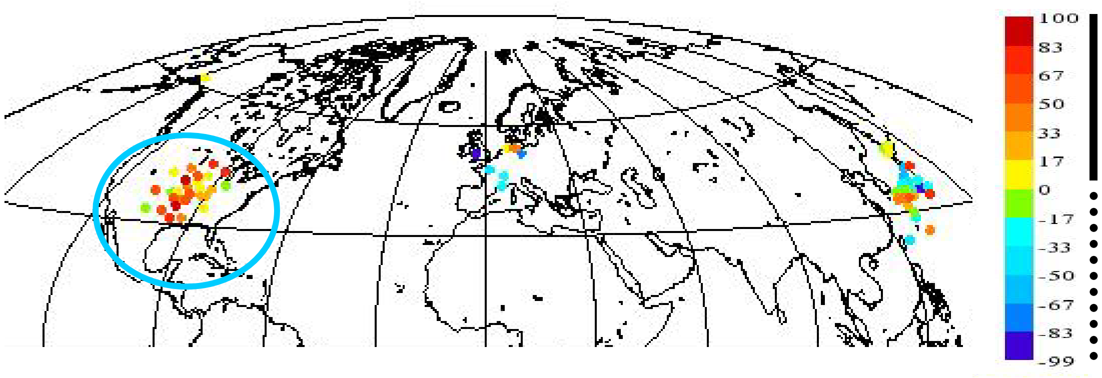


The tool provides information on the observation type, subtype, variable and level responsible for the forecast error variation

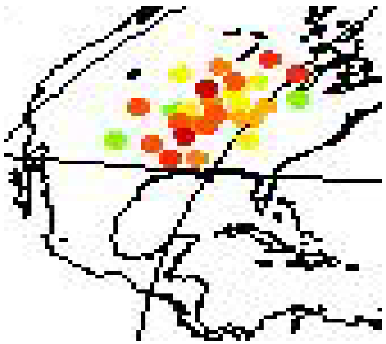
FSO: Pilot and Wind Profilers FcE contribution Summer 2006

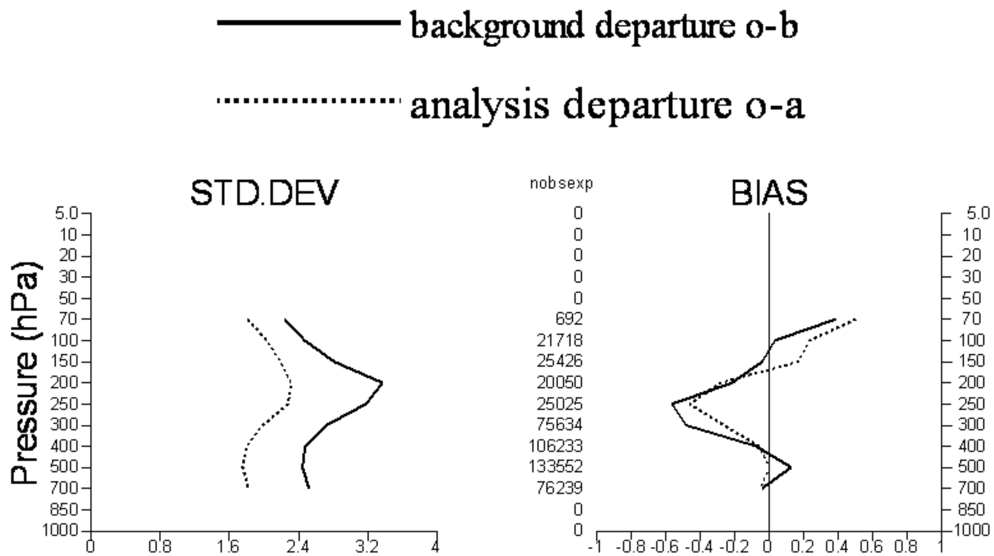






FSO: Wind Profilers North America Summer 2006

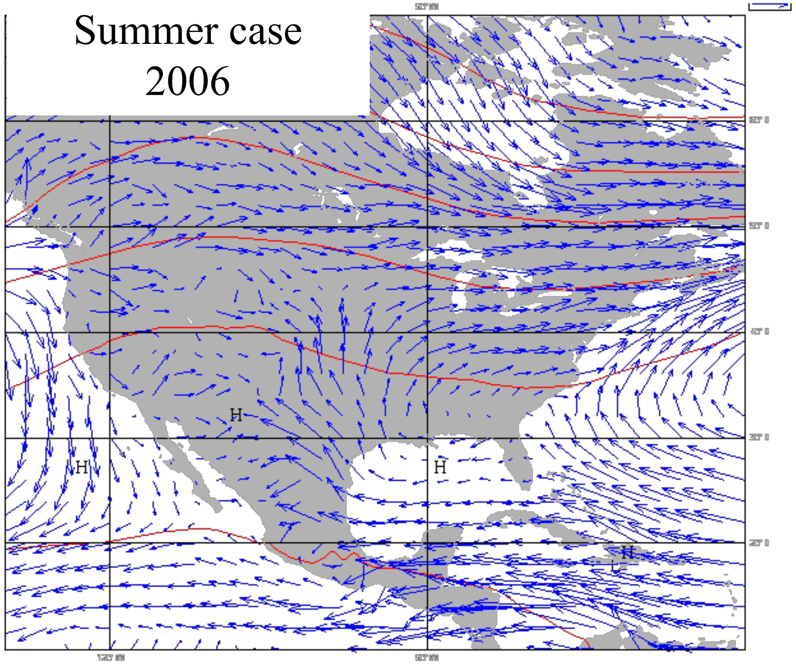


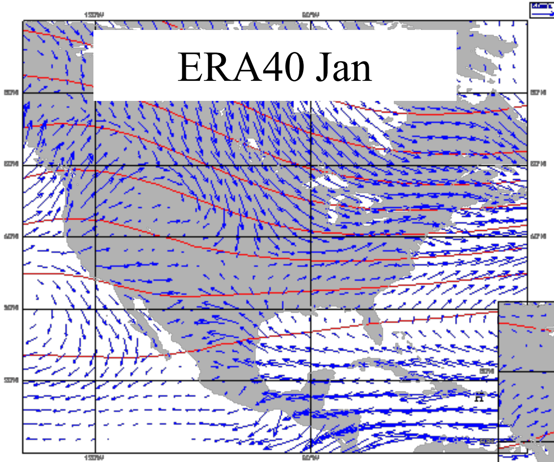


North America "Problem" (OD/RD special topic 2005)

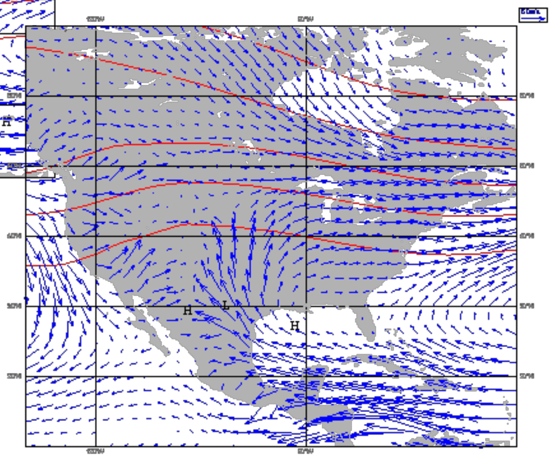
- **strong, moist warm flow from the Gulf of Mexico**
- **wind increments are huge and divergent at 150-250 hPa**
- **the conclusion was that "increments are not related to bad observations or a poor 4D-Var performance"**

… under certain meteorological conditions wind profilers measurements can be contaminated….(Ackley *et al*, 1998)





ERA40 Jun

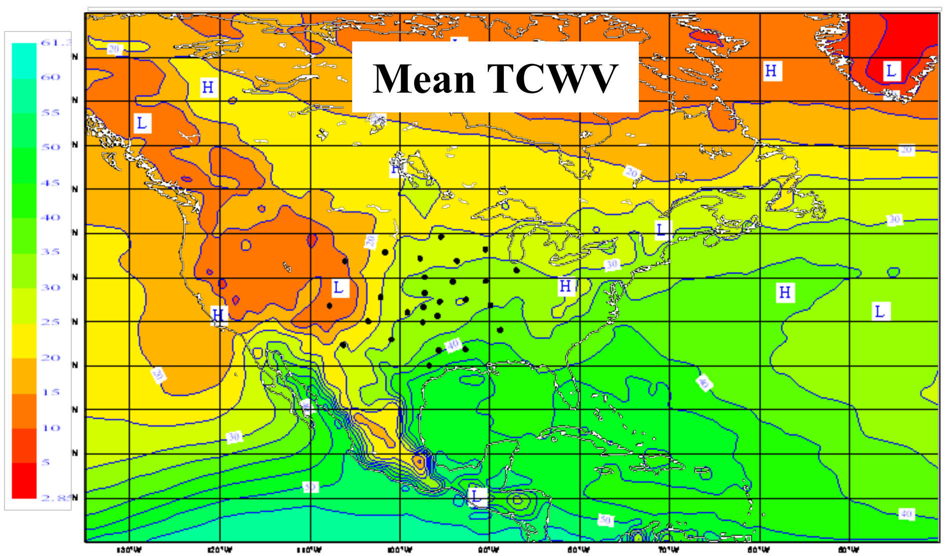


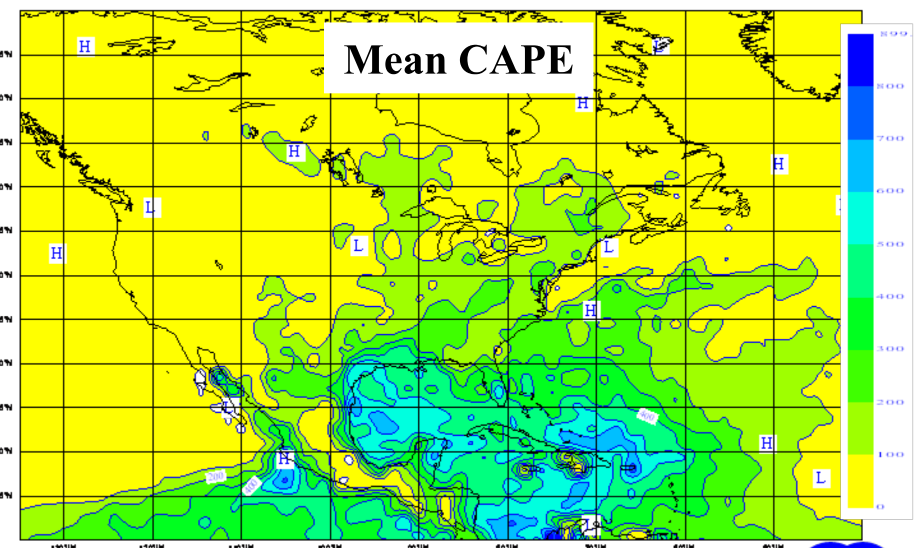
Mean 850-hPa Wind

&

Z500 hPa

courtesy by Fernando Prates



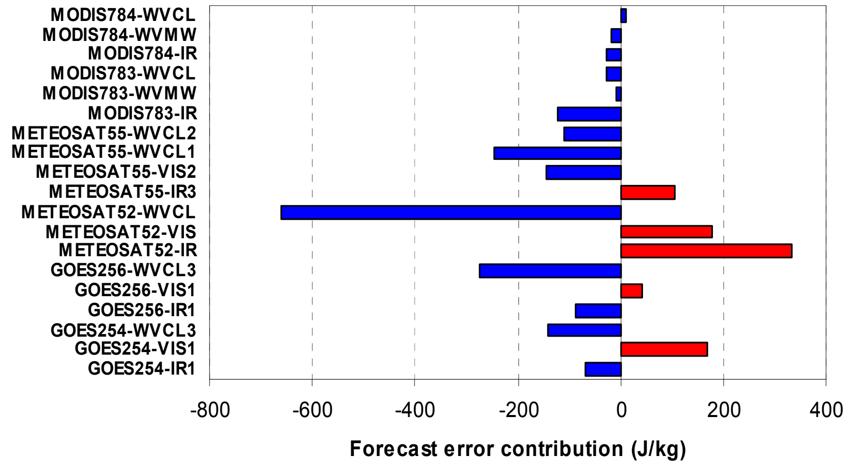


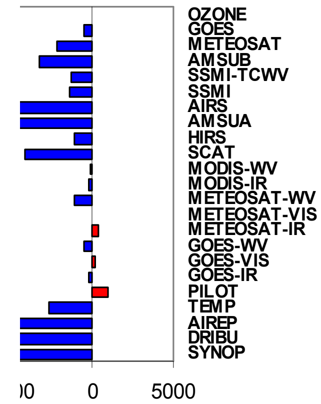
Summary FSO wind Profiler

- **FSO showed a Fc Error increase due to the American wind profiler observations.**
- **Southerly flow across SE USA bringing warm and moist air from Gulf of Mexico produced strong convective instability in the region, a typical situation at this time of the year.**
- **Following Ackley *et al* report (1998) on wind profiler measurements validity "in strong unstable conditions (turbulence) the measure of the mean horizontal wind is corrupted affecting the measurements". Suggesting that the forecast impact can change with the meteorological situation for the summer 2006 case.**

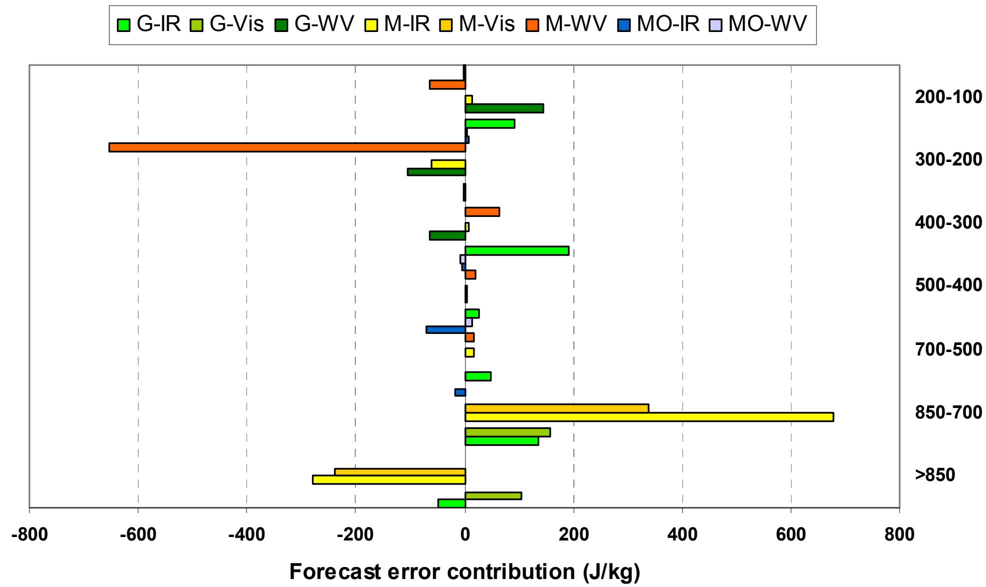
FSO: Atmospheric Motion Vector FcE Contribution Summer 2006

Forecast error contribution of the observed wind grouped by satellite types- positive corresponds to an **increase** of Fc Error



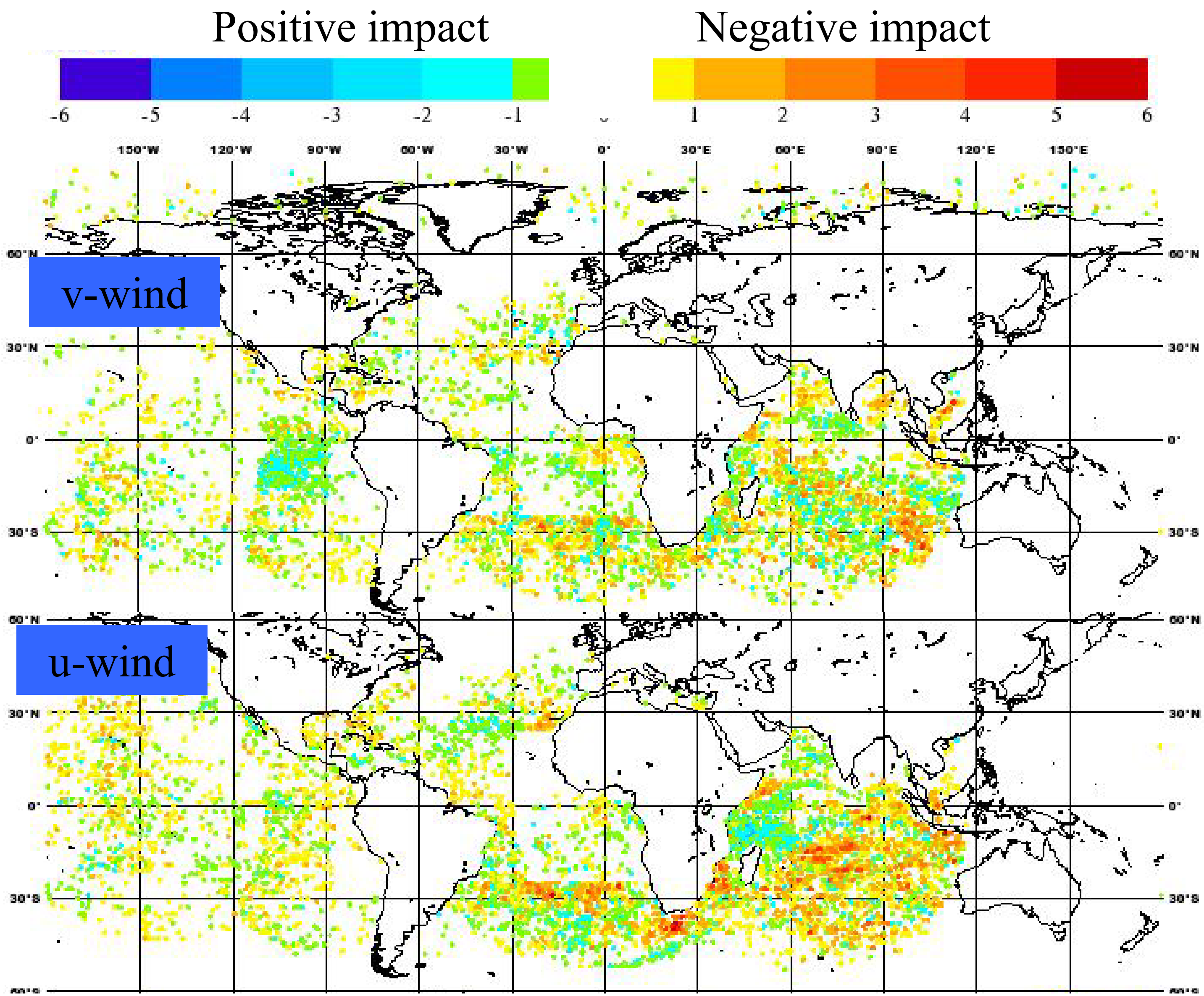


Forecast error contribution of the wind on **pressures levels** & grouped by satellite types- largest degradation comes from the lower troposphere

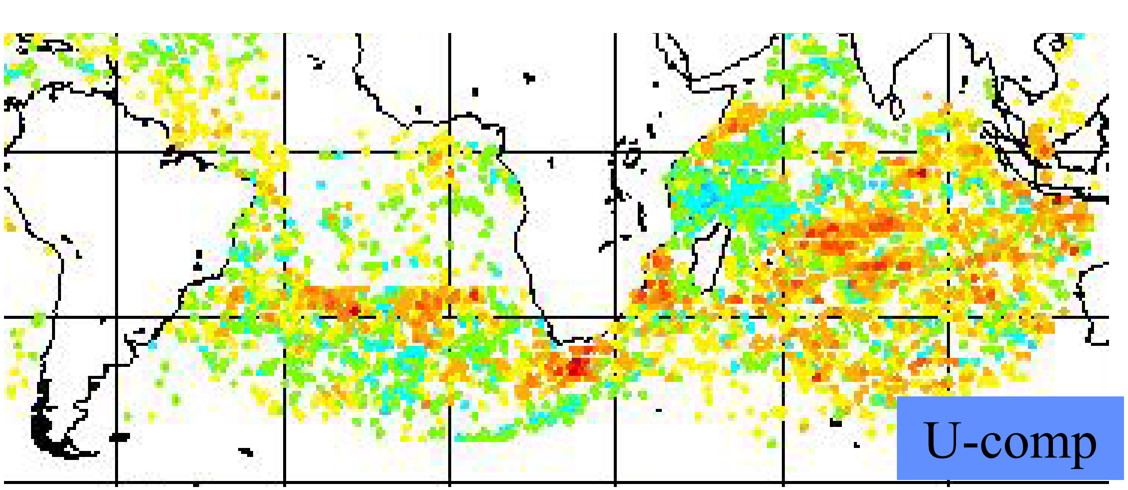


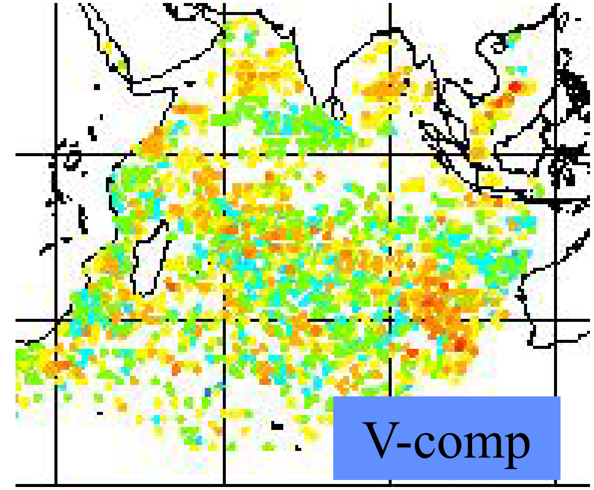
ECMWF

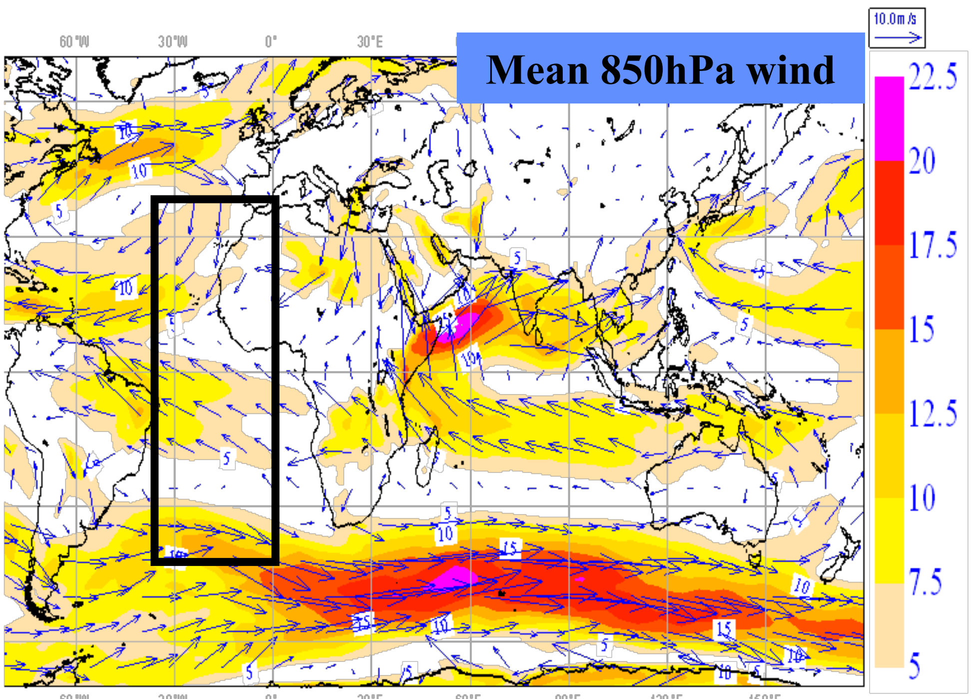
FSO AMV 700-1000 hPa: Summer 2006



FSO AMV 700-1000 hPa Summer 2006







Atlantic Ocean: transition between sub-tropical and extra-tropical from week to strong zonal flow

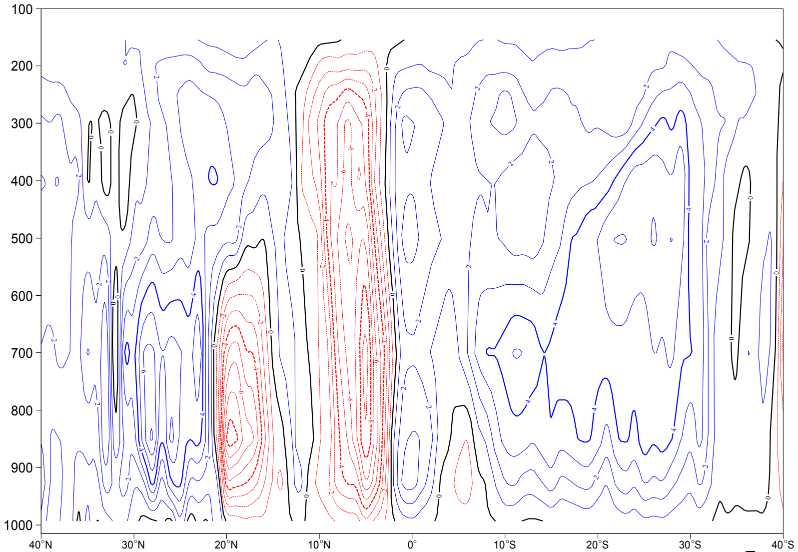
Indian Ocean: well established Monsoon circulation

courtesy by Fernando Prates

FSO Atlantic Ocean: Observation Quality

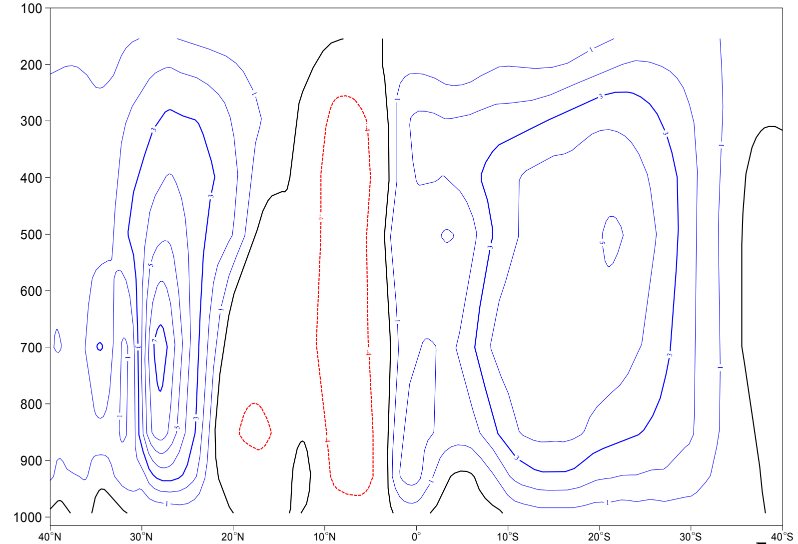
Cross Section [35W-0E]

AN mean vertical velocity (*0.01 Pa/s)



north **south**

ERA40



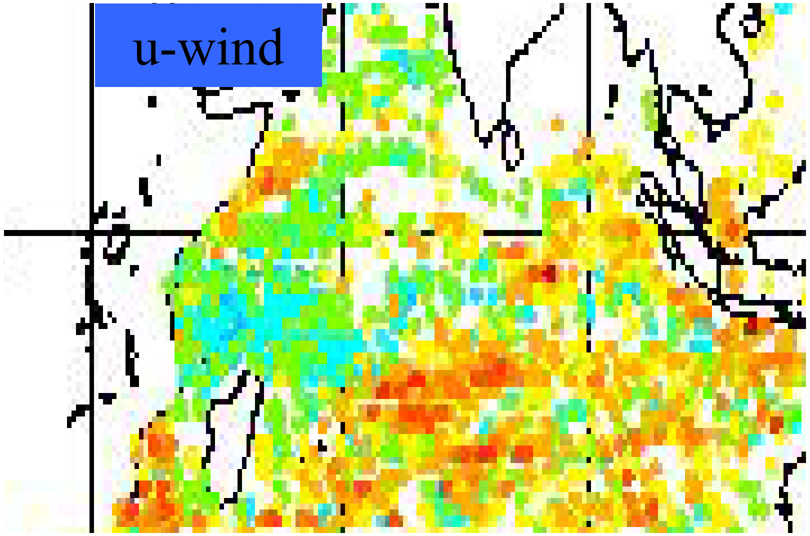
north **south**

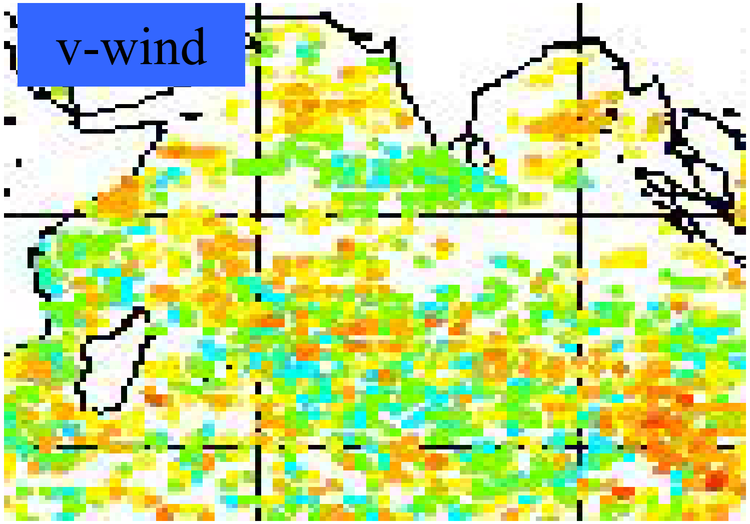
The strong sinking motion in SH near 30S represents the southern limit of the Hadley circulation where the subtropical high cell is located. Cloud suppression or low clouds.

AMV quality: difficult to assign the height of the cloud top

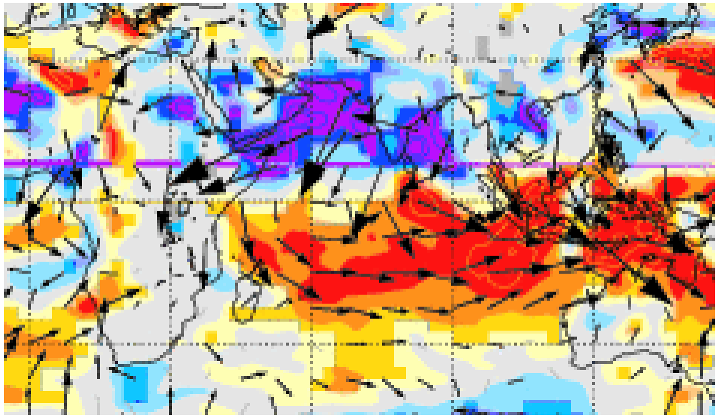
courtesy by Fernando Prates

FSO Indian Monsoon Summer 2006: Model bias





A too strong low level flow of Indian Summer Monsoon is a well known problem in the model as is indicated by the JJA mean analysis increments



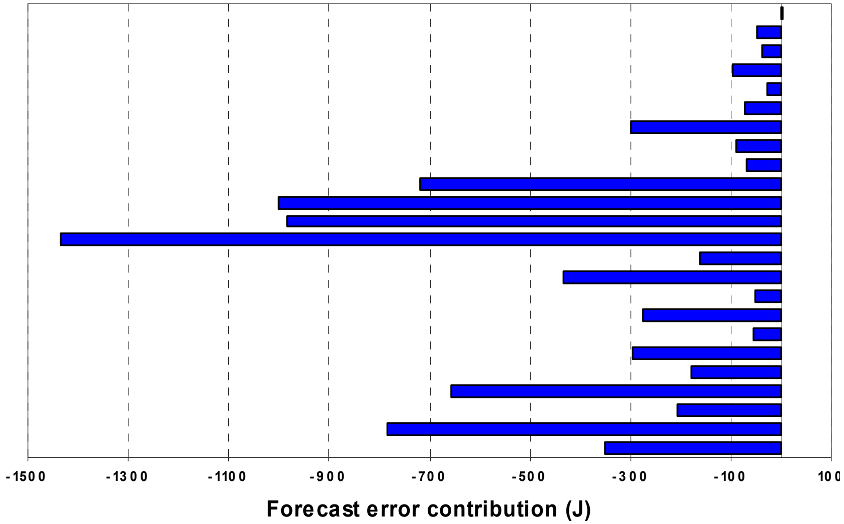
Mean An inc 925-hPa JJA 2006

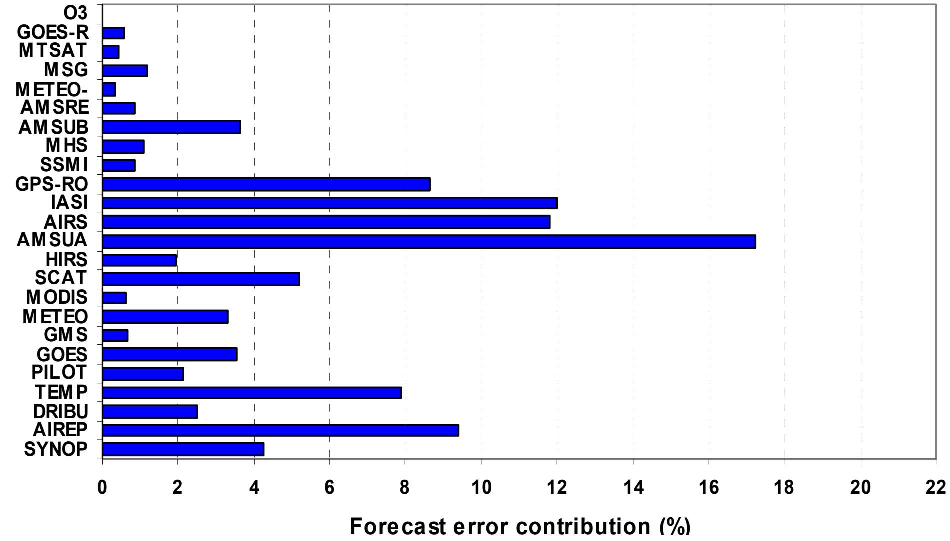
Diagnostic explorer

Summary FSO AMVs

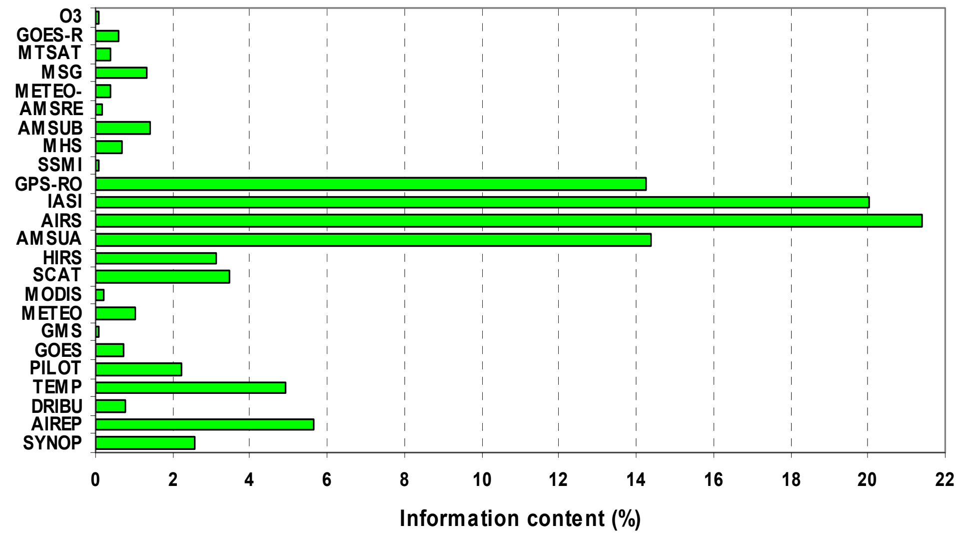
- **In Summer 2006 FSO showed a Fc error increase due to AMVs**
- **The location of the largest negative impact of the AMVs in Atlantic is found close to the region of strong sinking mean motion embedded in the Hadley circulation**
  - **Observation quality problem on the height assignment**
- **Detrimental effect is also observed in the Indian ocean associated with a too strong Indian monsoon circulation developed by the model**
  - **Model bias**

Operational ECMWF system
September to December 2008

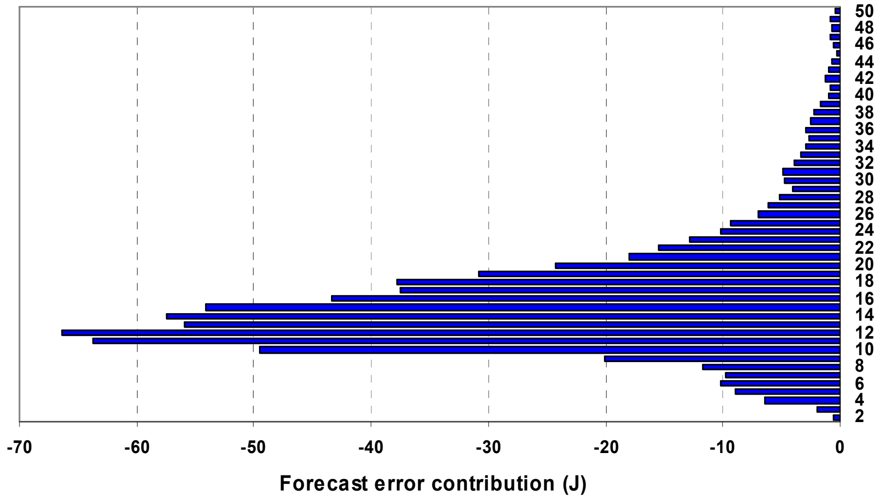


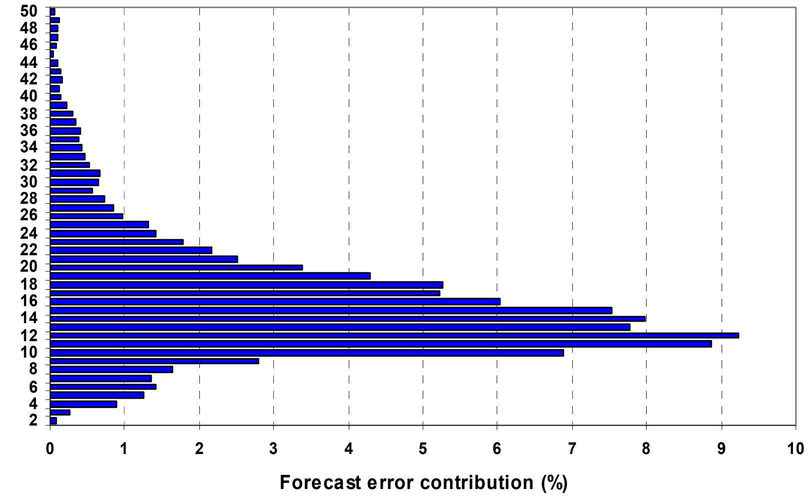


All Observations

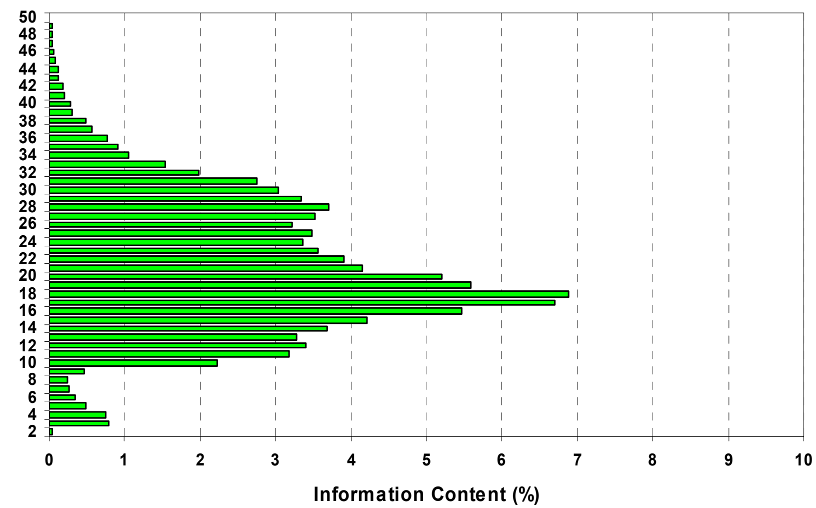


Operational ECMWF system
September to December 2008





GPS-Radio Occultation



ECMWF

Adjoint Diagnostic Conclusion

- Over the last decade the assessment of each observation contribution to analysis and forecast is among the most challenging diagnostics in data assimilation and NWP.

- These techniques show how the influence is assigned during the assimilation procedure and how is the forecast impact of each observation.

- Recently, Daescu (2008) derived a sensitivity equation of an unconstrained variational data assimilation system with respect to the main input parameters: observation, background and their error covariance matrices.

- Observation influence and forecast impact have also been developed in a non-adjoint context. Junjie Liu *et al* 2008 and Junjie Liu *et al* 2009 translated the concepts to EnKF system and also showed that the solution being very accurate, Cross Validation can straightforward be applied.

- In an operational context , the correct usage of the tools requires a close collaboration with synopticians and observation monitoring section.