Satellite data assimilation for NWP III

Estimating the impact of new observations with Ensemble techniques

Sean Healy
European Centre for Medium-range Weather Forecasts (ECMWF)

Special thanks to: Massimo Bonavita, Florian Harnisch …

Review of previous lectures

- In general, NWP has moved away from using satellite retrievals/products, to assimilating “raw” observations.

- Stressed the importance of understanding the error characteristics and limitations of the observations.

  - This means knowing/understanding H, R.

- In NWP, we accept that all observations have errors, but we can still use them provided we have a reasonable estimate of the observation error statistics, R.

- Observations where the errors are not characterised by R must be screened out in the quality control (QC).

Aside: Quality control (QC) Qu. After L2

- **Data assimilation systems usually include a QC step for satellite data of the form (*say for a radiance or bending angle*). REJECT IF:**

$$|y - \mathbf{Hx_b}| > \gamma(\sigma_o + \sigma_b)$$

or

$$|y - \mathbf{Hx_b}| > \gamma\sigma_o$$

where typically

$$\gamma \approx 5\text{-}8$$

It is a good idea to monitor the data that is being removed by QC.

The ozone hole was originally missed because of a QC step.(Alan O'Neill)

- **And some data is blacklisted meaning it doesn't enter into the DA system even before QC checks.**

Aim for this lecture

- The satellite component of the global observing system should evolve to reflect updated user requirements and the emergence of new measurement techniques and technologies. ONE OF YOU MAY PROPOSE A NEW MISSION. *NWP assimilation may be one goal.*

- But how can we estimate the impact or value of a new mission/observations to inform GOS decisions? What information will the new observations add to those already available?

- If we get a good forward model H, and a good estimate the observation error covariance matrix R, we can use variational and ensemble DA techniques to estimate the impact of the new observations.

Outline

- Estimating the “information content” using a 1D-Var approach. *Valid for linear and ~weakly non-linear problems*.

- Link this to Kalman Filter/4D-Var, and the need to approximate with ensemble techniques in NWP because of the size of the problem.

- The Ensemble of Data Assimilations (EDA).

- Assessing the impact of new data with the EDA. (not an OSSE)

- Summary.

Information content

- If we assume a linear problem, recall the 1D-Var solution from lecture 1. We minimize a cost function:

$$J(\mathbf{x}) = (\mathbf{x} - \mathbf{x_b})^{\mathrm{T}} \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x_b}) + (\mathbf{y_m} - \mathbf{H}[\mathbf{x}])^{\mathrm{T}} \mathbf{R}^{-1} (\mathbf{y_m} - \mathbf{H}[\mathbf{x}])$$

- The linear solution is:

$$\mathbf{x_a} = \mathbf{x_b} + \mathbf{BH}^{\mathrm{T}} (\mathbf{HBH}^{\mathrm{T}} + \mathbf{R})^{-1} (\mathbf{y_m} - \mathbf{Hx_b})$$

- And we obtain a *theoretical* estimate of the solution error covariance matrix:

$$\mathbf{S_a} = \mathbf{B} - \mathbf{BH}^{\mathrm{T}} (\mathbf{HBH}^{\mathrm{T}} + \mathbf{R})^{-1} \mathbf{HB}$$

- Note that the solution error cov. does not depend on the observation values, only H and the covariance estimates.

Information content (2)

- If the assumed covariance matrices are reasonable, the solution error covariance matrix should be a reasonable approximation of the actual solution error ***statistics***.

- We can use it to investigate the "information content" of the observation.

- "Information content": assume it is related to reduction of statistical uncertainty as result of making the observation. IE, how the error PDF changes.

- Uncertainty before making the observation: $\mathbf{B}$

- Uncertainty after: $\mathbf{S_a} = \mathbf{B} - \mathbf{B}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}\mathbf{H}\mathbf{B}$

Information content

- There are some more complex mathematical approaches to quantify the information content: Reduction in Shannon entropy; Degrees of Freedom of Signal.

$$\boxed{S_e = -\tfrac{1}{2}\ln\left|\mathbf{S_a}\mathbf{B}^{-1}\right|}$$

$$\boxed{DFS = Tr\left(\mathbf{I} - \mathbf{S_a}\mathbf{B}^{-1}\right)}$$

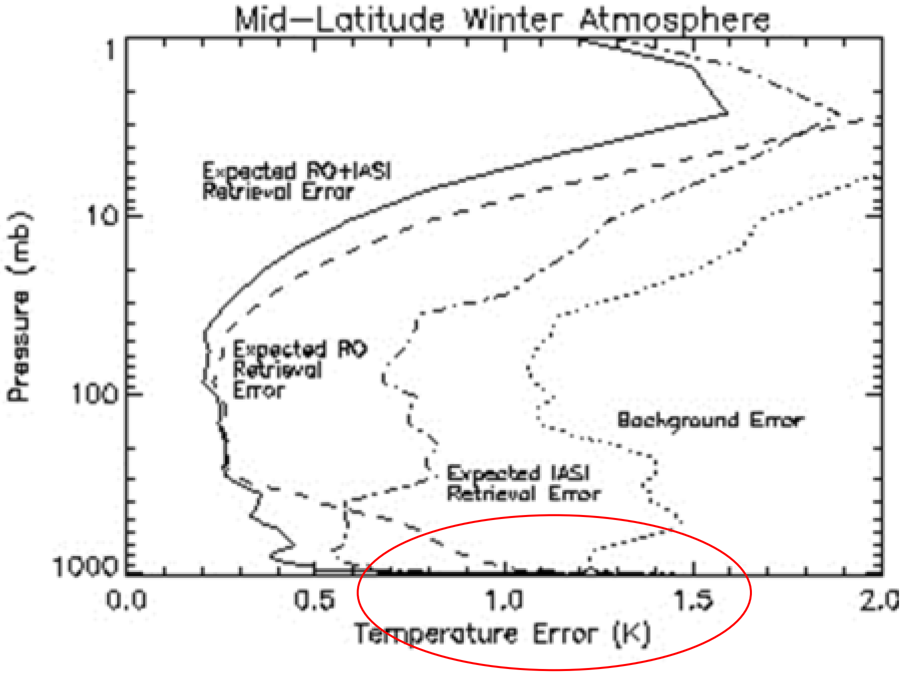
  - EG, see Rodgers: *Inverse Methods for Atmospheric Sounding: Theory and Practice (page 36).*

- Perhaps the easiest way is to compare the diagonal values of the covariance matrices ($\sqrt{S_a(i,i)}$ and $\sqrt{B(i,i)}$).

- This approach provides a good indication of where the observation will have the most influence

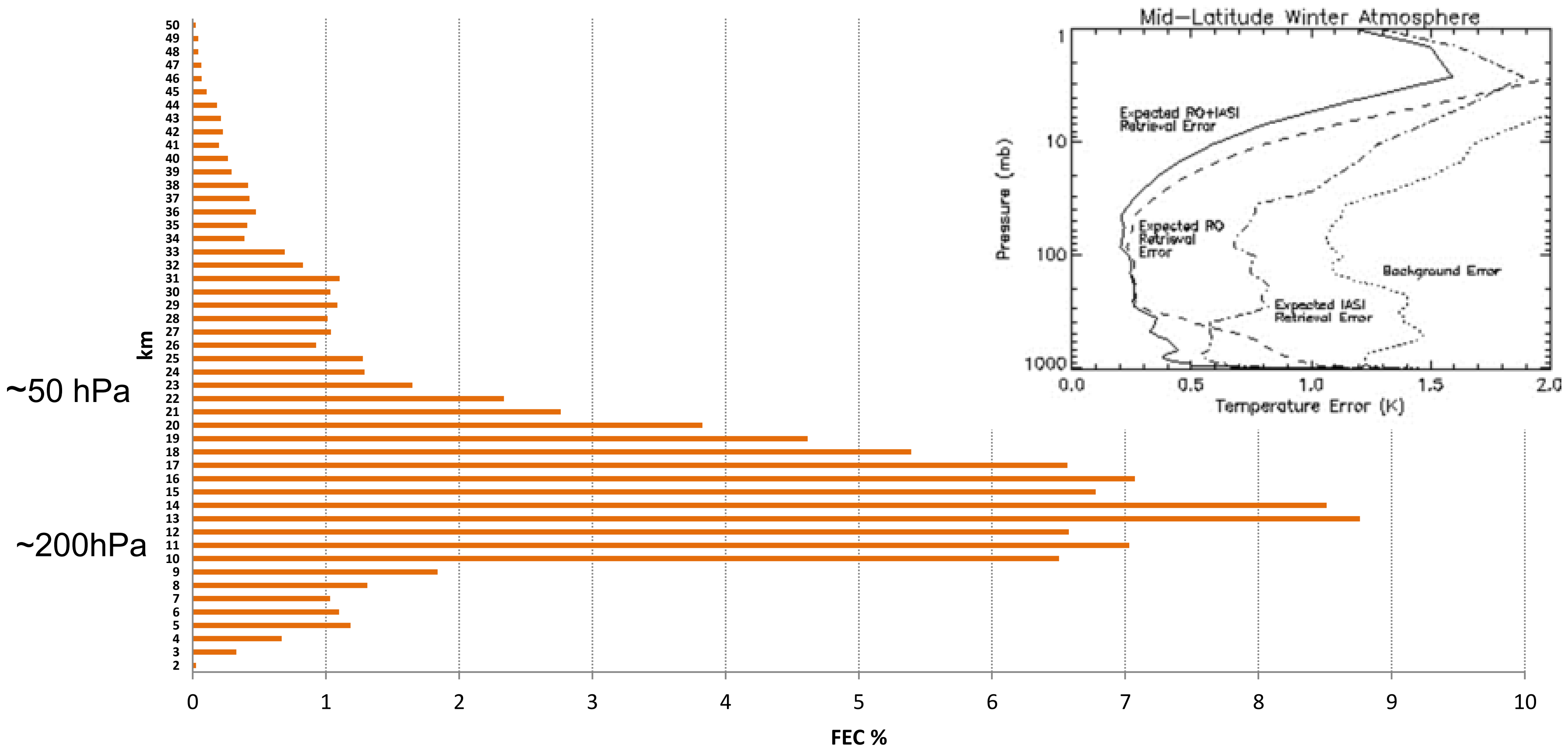
Example using GPS-RO and IASI: Do we need both? What will GPS-RO add?

- Compared the information content of GPS-RO and IASI measurements in 1D-Var (2003).



- Concluded measurements highly complementary. Suggested GPS-RO would provide the best temperature information in the 300-50 hPa interval.

Heights where GPS-RO is reducing the 24 hr forecast errors in ECMWF system using an adjoint approach

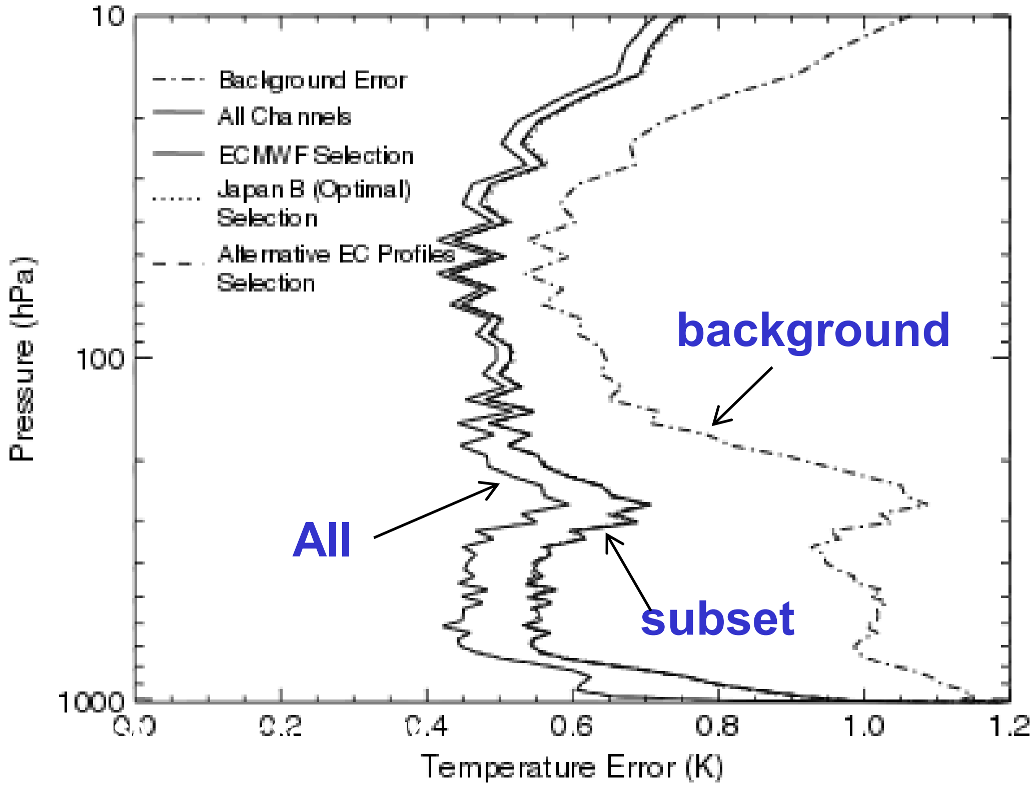


Remark: Agrees with early 1D-Var information content studies.

Example 2: IASI channel selection

The infrared sounder IASI provides 8461 channels. This is too many for assimilation into the NWP model**. In any case, 8461 channels DOES NOT mean 8461 pieces of information.**

We can use 1D-Var information content techniques to chose a subset of ~300 channels.



The subset of channels is chosen to minimize the loss of information, with respect to using all the available channels.

Again we need **H, R** and **B** for this computation.

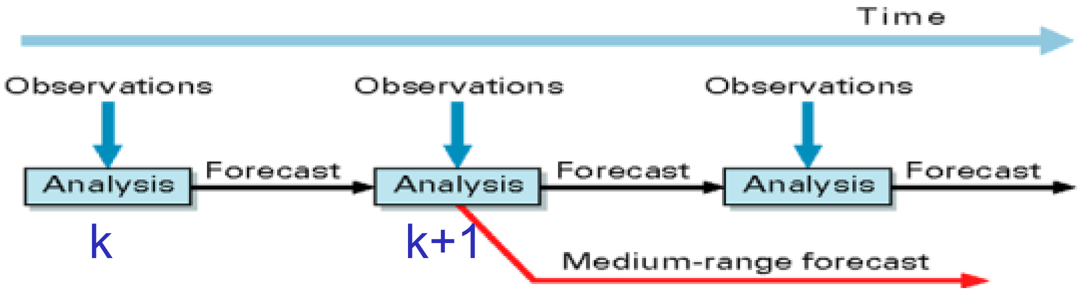
Can we generalise these 1D information content studies?

- Can we estimate the impact/information content of a set of new observations from a future mission, distributed in space/time, in the 4D-Var system?

- New ensemble techniques, developed by the data assimilation community, provide a framework for tackling this problem.

- Ensemble techniques have been developed to provide estimates flow dependent background error *statistics*, B.

Recap: Standard Kalman Filter



- The linear, unbiased analysis equation has the form:

$$\mathbf{x}^a_k = \mathbf{x}^b_k + \mathbf{K}_k\,(\mathbf{y}_k - \mathbf{H}_k(\mathbf{x}^b_k))$$

a = analysis; b = background
k = time index (t=0,1,…,k,…)

- The best linear unbiased analysis (a.k.a. Best Linear Unbiased Estimator, BLUE) is achieved when the matrix $\mathbf{K}_k$ (Kalman Gain Matrix) has the form:

$$\mathbf{K}_k = \mathbf{P}^b_k\,\mathbf{H}^T_k(\mathbf{H}_k\,\mathbf{P}^b_k\,\mathbf{H}^T_k + \mathbf{R}_k)^{-1} = ((\mathbf{P}^b_k)^{-1} + \mathbf{H}^T_k\,\mathbf{R}_k^{-1}\,\mathbf{H}_k\,)^{-1}\,\mathbf{H}^T_k\,\mathbf{R}_k^{-1}$$

$\mathbf{P}^b$ = covariance matrix of the background error
$\mathbf{R}$ = covariance matrix of the observation error

Standard Kalman Filter

- What is the error covariance matrix associated with this background?

$$\mathbf{x}^b{}_k = \mathbf{M}_{t_{k-1}\to t_k}(\mathbf{x}^a{}_{k-1})$$

- Subtract the true state $\mathbf{x}^*{}_k$ from both sides of the equation:

$$\boldsymbol{\varepsilon}^b{}_k = \mathbf{M}_{t_{k-1}\to t_k}(\mathbf{x}^a{}_{k-1}) - \mathbf{x}^*{}_k$$

- Since $\mathbf{x}^a{}_{k-1} = \mathbf{x}^*{}_{k-1} + \boldsymbol{\varepsilon}^a{}_{k-1}$ we have:

$$\boldsymbol{\varepsilon}^b{}_k = \mathbf{M}_{t_{k-1}\to t_k}(\mathbf{x}^*{}_{k-1} + \boldsymbol{\varepsilon}^a{}_{k-1}) - \mathbf{x}^*{}_k =$$

$$\mathbf{M}_{t_{k-1}\to t_k}(\mathbf{x}^*{}_{k-1}) + \mathbf{M}_{t_{k-1}\to t_k}\boldsymbol{\varepsilon}^a{}_{k-1} - \mathbf{x}^*{}_k =$$

$$\mathbf{M}_{t_{k-1}\to t_k}\boldsymbol{\varepsilon}^a{}_{k-1} + \eta_k$$

- Where we have defined the model error $\eta_k = \mathbf{M}_{t_{k-1}\to t_k}(\mathbf{x}^*{}_{k-1}) - \mathbf{x}^*{}_k$
- We will also assume that $< \boldsymbol{\varepsilon}^a{}_{k-1} > = < \eta_k > = 0 \; => \; < \boldsymbol{\varepsilon}^b{}_k >$
- The background error covariance matrix will then be given by:

Standard Kalman Filter

$$<\boldsymbol{\varepsilon}^b_k (\boldsymbol{\varepsilon}^b_k)^T> = \mathbf{P}^b_k = <(\mathbf{M}_{t_{k-1}\rightarrow t_k}\boldsymbol{\varepsilon}^a_{k-1} + \eta_k)\,(\mathbf{M}_{t_{k-1}\rightarrow t_k}\boldsymbol{\varepsilon}^a_{k-1} + \eta_k)^T> =$$

$$\mathbf{M}_{t_{k-1}\rightarrow t_k}<\boldsymbol{\varepsilon}^a_{k-1} (\boldsymbol{\varepsilon}^a_{k-1})^T> (\mathbf{M}_{t_{k-1}\rightarrow t_k})^T + <\eta_k (\eta_k)^T> =$$

$$\mathbf{M}_{t_{k-1}\rightarrow t_k}\,\mathbf{P}^a_{k-1}\,(\mathbf{M}_{t_{k-1}\rightarrow t_k})^T + \mathbf{Q}_k$$

- Here we have assumed $< \boldsymbol{\varepsilon}^a_{k-1} (\eta_k)^T> = 0$ and defined the model error covariance matrix $\mathbf{Q}_k = <\eta_k (\eta_k)^T>$
- We now have all the equations necessary to propagate and update the state and its error estimates:

The Kalman Filter

- The Kalman filter includes the covariance evolution, providing error statistics that vary in time and space.

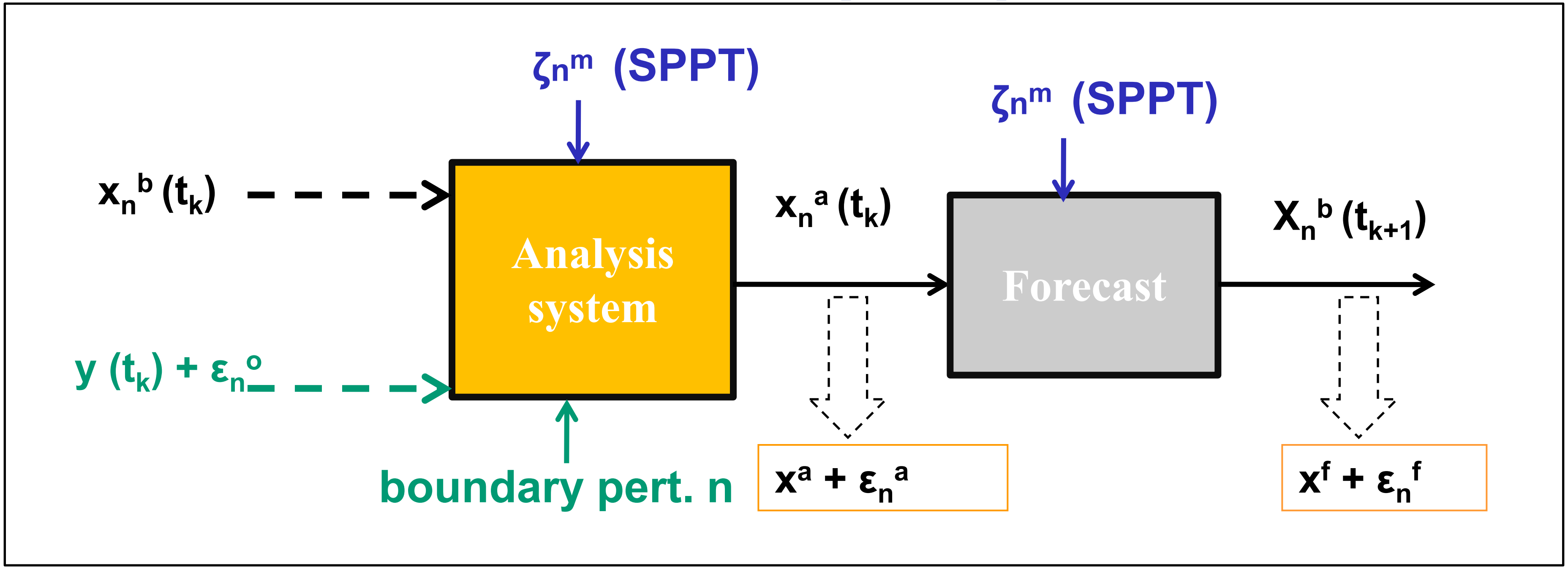
- In principle, it provides all the information we need for an information content study.

- But the NWP matrices are too large for practical application. It can approximated with ensemble techniques (EnKF, …).

- “Classic” 4D-Var: static background error covariance matrix, ignore model error (“strong constraint”).

- But 4D-Var can be combined with an ensemble approach to estimate flow dependent error statistics.

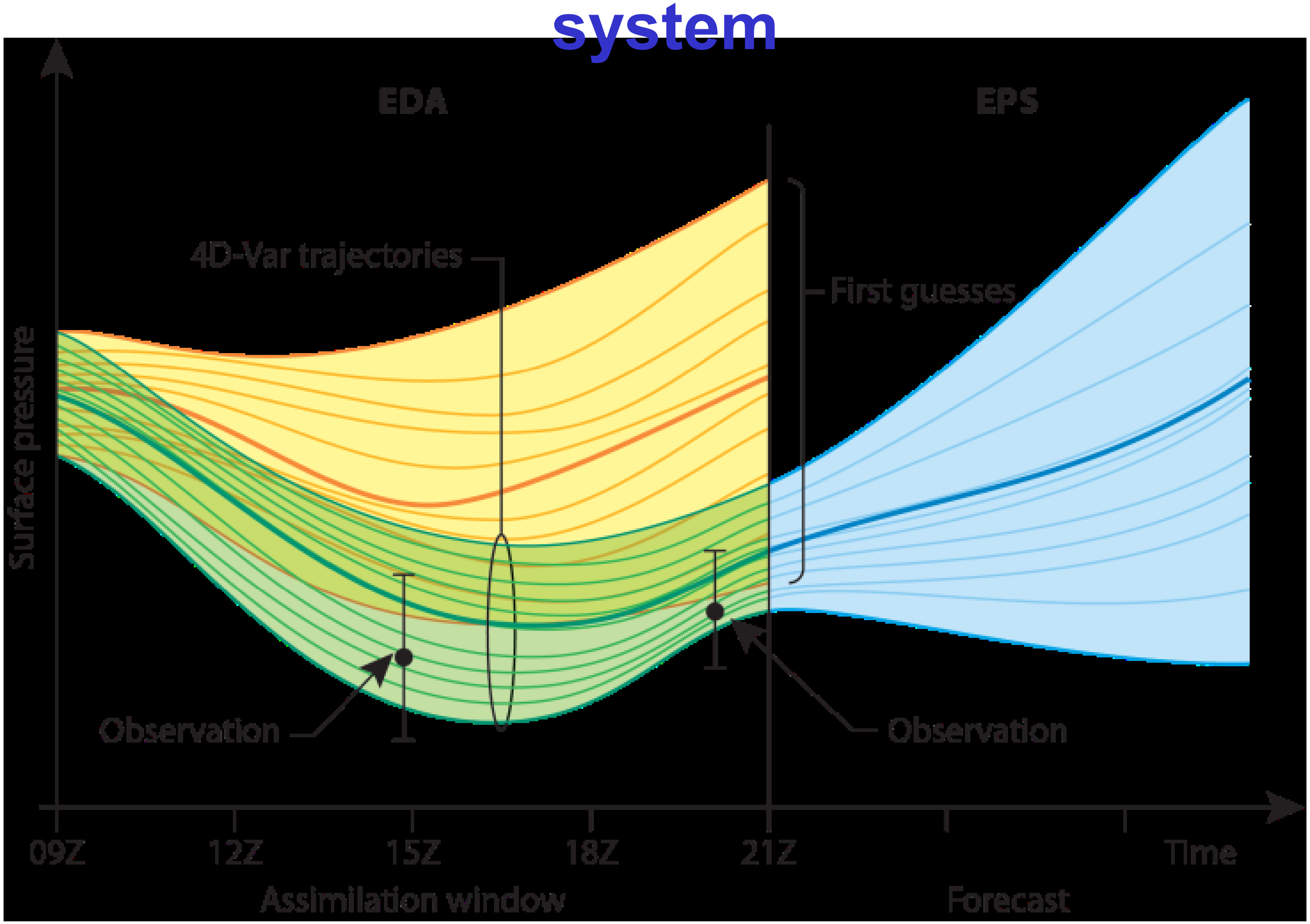
The Ensemble of Data Assimilations method (EDA)



•We cycle 10 (or 20) 4D-Vars in parallel using perturbed observations in each 4D-Var, plus a control experiment with no perturbations.

•The spread of the ensemble about the mean is related to the theoretical estimate of the analysis and short-range forecast *error statistics*.

Applications of the EDA: Ensemble prediction system



Remark: EDA method

$$\boxed{\begin{aligned} \mathbf{x}_a^k &= \mathbf{x}_b^k + \mathbf{K}_k\left(\mathbf{y}^k - \mathbf{H}_k\mathbf{x}_b^k\right) \\ \mathbf{x}_b^{k+1} &= \mathbf{M}_k\mathbf{x}_a^k \end{aligned}} \quad \boxed{\begin{aligned} \tilde{\mathbf{x}}_a^k &= \tilde{\mathbf{x}}_b^k + \mathbf{K}_k\left(\mathbf{y}^k + \eta^k - \mathbf{H}_k\tilde{\mathbf{x}}_b^k\right) \\ \tilde{\mathbf{x}}_b^{k+1} &= \mathbf{M}_k\tilde{\mathbf{x}}_a^k + \zeta^k \end{aligned}} \quad \begin{aligned} \eta &\sim \mathcal{N}\left(0, \mathbf{R}\right) \\ \zeta &\sim \mathcal{N}\left(0, \mathbf{Q}\right) \end{aligned}$$

$$\boxed{\varepsilon_a^k = \tilde{\mathbf{x}}_a^k - \mathbf{x}_a^k \qquad \varepsilon_b^{k+1} = \tilde{\mathbf{x}}_b^{k+1} - \mathbf{x}_b^{k+1}}$$

$$\boxed{\begin{aligned} &\mathbf{P}_k^a \\ &\mathbf{P}_k^b \end{aligned}} \quad \boxed{\begin{aligned} \overline{\varepsilon_a^k\left(\varepsilon_a^k\right)^T} &= \left(\mathbf{I} - \mathbf{K}_k\mathbf{H}_k\right)\overline{\varepsilon_b^k\left(\varepsilon_b^k\right)^T}\left(\mathbf{I} - \mathbf{K}_k\mathbf{H}_k\right)^T + \mathbf{K}_k\mathbf{R}_k\mathbf{K}_k^T \\ \overline{\varepsilon_b^{k+1}\left(\varepsilon_b^{k+1}\right)^T} &= \mathbf{M}_k\overline{\varepsilon_a^k\left(\varepsilon_a^k\right)^T}\mathbf{M}_k^T + \mathbf{Q}_k \end{aligned}}$$

→ State estimate cancels out and to first order only the perturbations are important for the EDA spread.

The EDA method – EDA spread

- The spread of the ensemble about the mean provides an estimate of the error variance of the analysis and short-range forecast – if the R matrices are realistic .

- spread *s* (variance) of *N*-member ensemble for EDA experiments:

  for each time d

$$s_d = \sqrt{\sigma_d^2} = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}\left(x^n - \overline{x}\right)^2}$$

for a period D (Expectation)

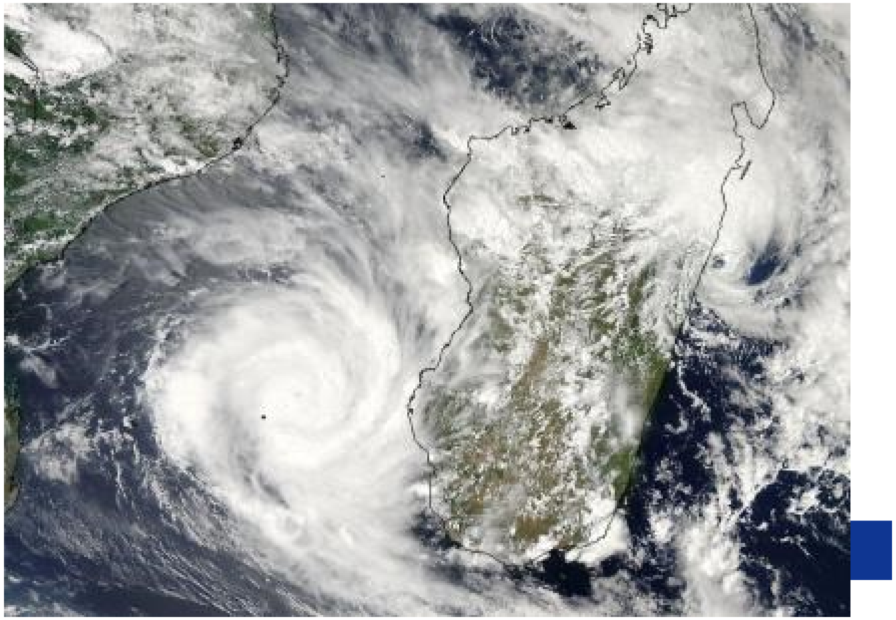
$$s = \sqrt{\mathbb{E}\left[\sigma_d^2\right]} = \sqrt{\frac{1}{D}\sum_{d=1}^{D}\left(\frac{1}{N-1}\sum_{n=1}^{N}\left(x^n - \overline{x}\right)^2\right)}$$

Applications of the EDA (M. Bonavita)

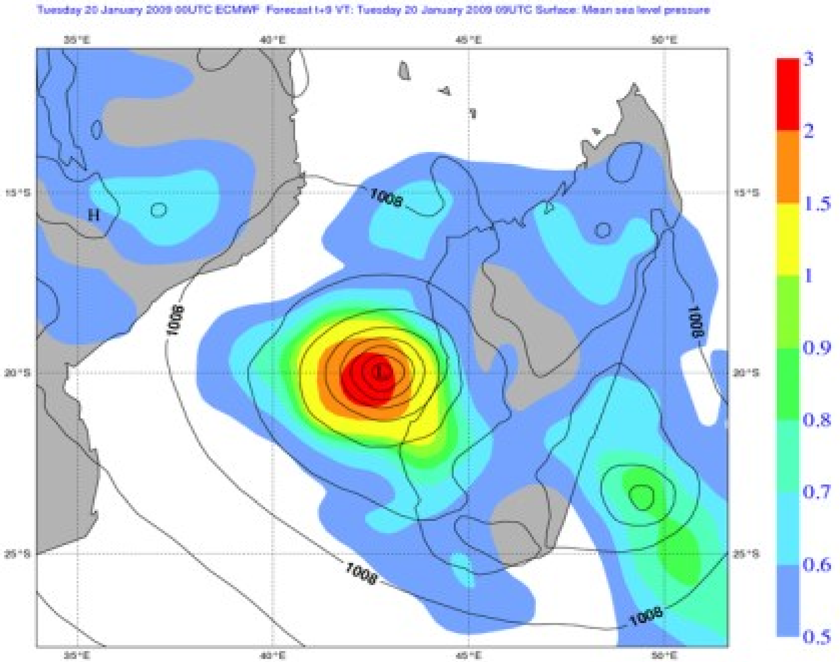
- We want to use EDA perturbations to simulate 4DVar flow-dependent background error covariance evolution

- We start with the EDA flow-dependent estimates of background error variances (

Hurricane Fanele, 20 January 2009



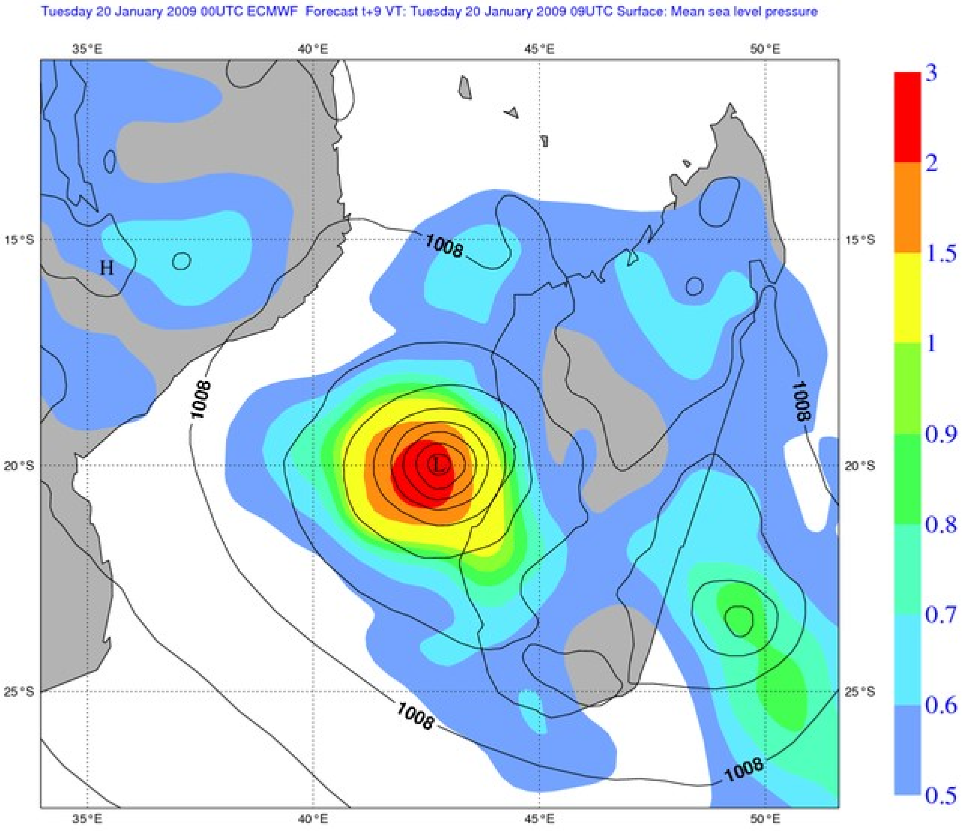
EDA based background error variance for surface pressure


Tuesday 20 January 2009 00UTC ECMWF Forecast t+9 VT: Tuesday 20 January 2009 09UTC Surface: Mean sea level pressure
35°E
40°E
45°E
50°E
15°S
20°S
25°S
H
L
1008
3
2
1.5
1
0.9
0.8
0.7
0.6
0.5

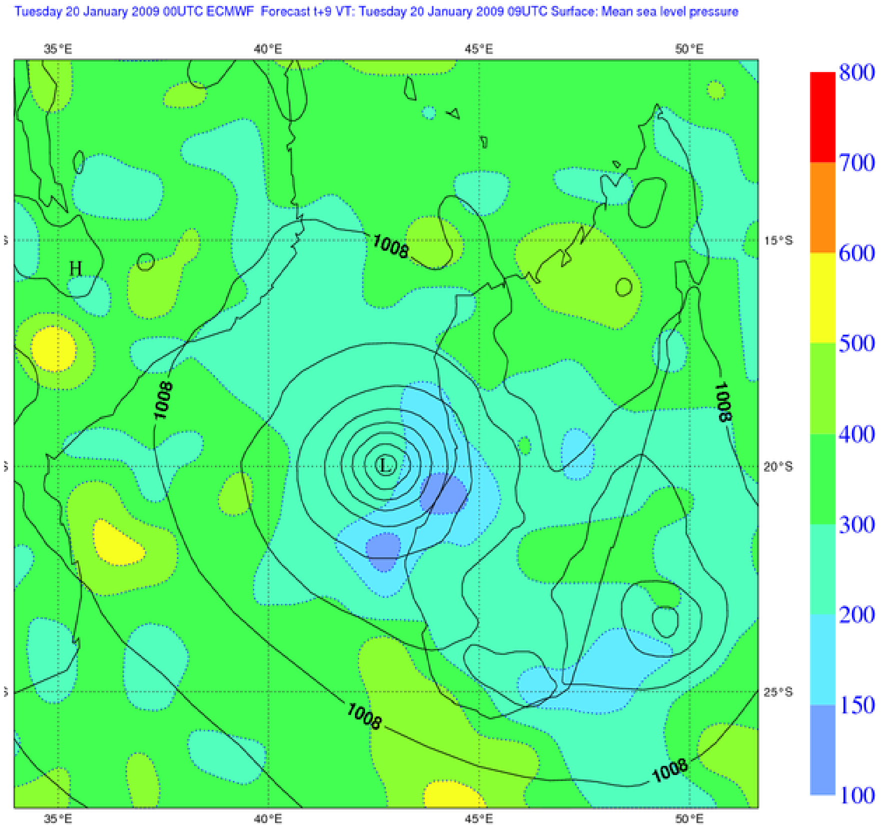
Use of EDA covariances in 4DVar

20 member EDA

Surf. Press. Background Err. **St.Dev.**


Tuesday 20 January 2009 00UTC ECMWF Forecast t+9 VT: Tuesday 20 January 2009 09UTC Surface: Mean sea level pressure
35°E
40°E
45°E
50°E
15°S
20°S
25°S
H
L
1008
1008
1008
1008
1008
3
2
1.5
1
0.9
0.8
0.7
0.6
0.5

Surf. Press. BG Err. **Correlation** L.


Tuesday 20 January 2009 00UTC ECMWF Forecast t+9 VT: Tuesday 20 January 2009 09UTC Surface: Mean sea level pressure
35°E
40°E
45°E
50°E
15°S
20°S
25°S
H
L
1008
1008
1008
1008
1008
800
700
600
500
400
300
200
150
100

The EDA method

The EDA spread

- estimates the analysis (forecast) uncertainty, which is related to the error statistics and not the error itself.

$$\mathbf{P}_k^a \quad \mathbf{P}_{k+1}^b$$

- depends on the assumed input error statistics and not the actual ones (→ R, B, Q)

- provides realistic estimate of uncertainty if, *and only if*, the assumed input error statistics are realistic.

- For a non-specialist *"Errors of the day"* is a confusing term. "Error statistics of the day" is more appropriate.

EDA and 4D-Var information content

- We can trick the EDA system into thinking we have a new set of observations, even if they contain no new information about the real atmospheric state.

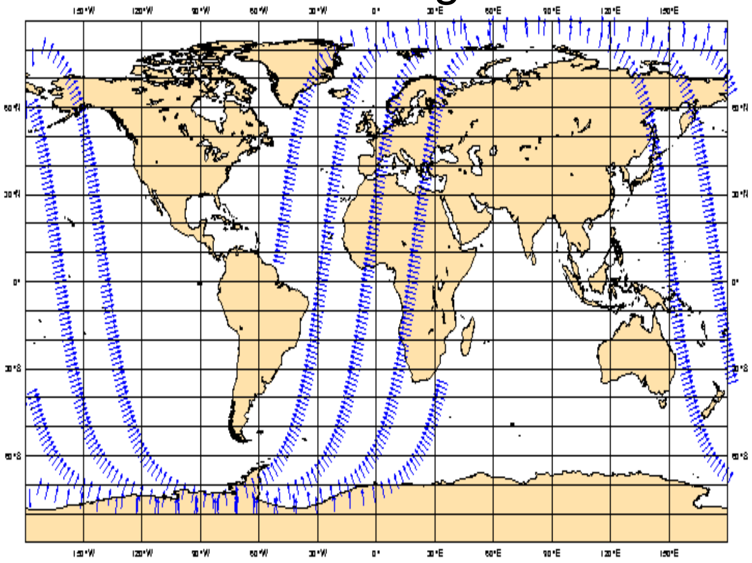
- We simulate a new observation set, using the new H and R.

- We then assimilate these simulated data into the EDA system, to see how the spread changes.

- This was initially used the estimate the impact of the Doppler Wind Lidar (DWL) shown lecture 2.

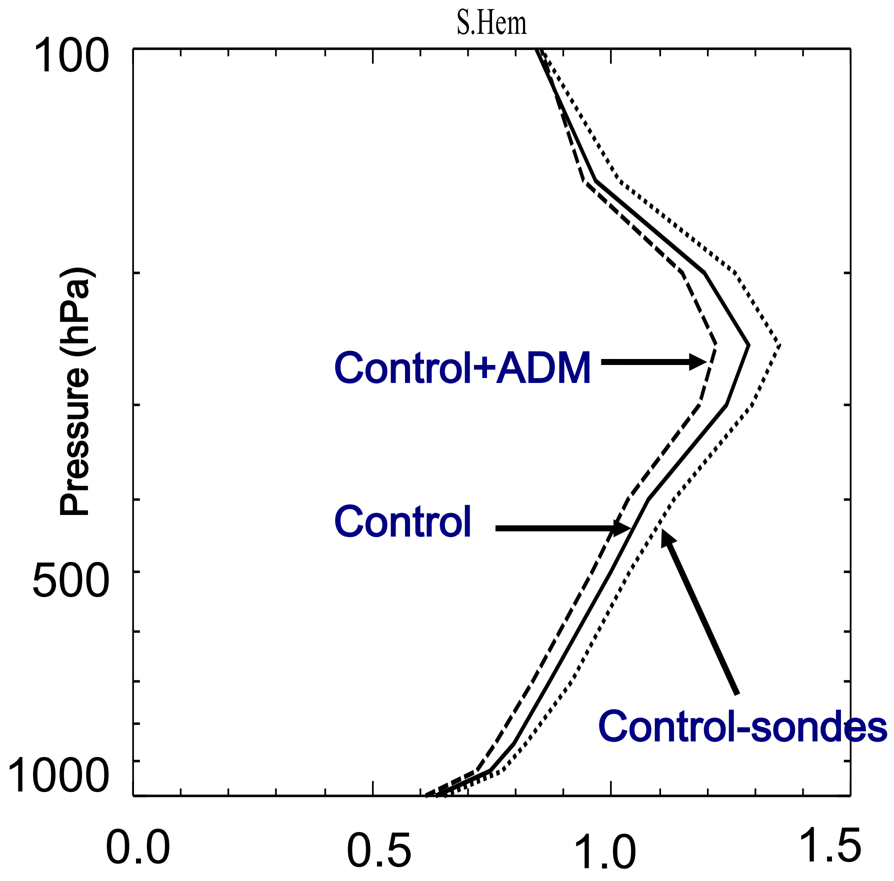
ADM-Aeolus: Simulated impact (Tan et al.)

6-hour data coverage:



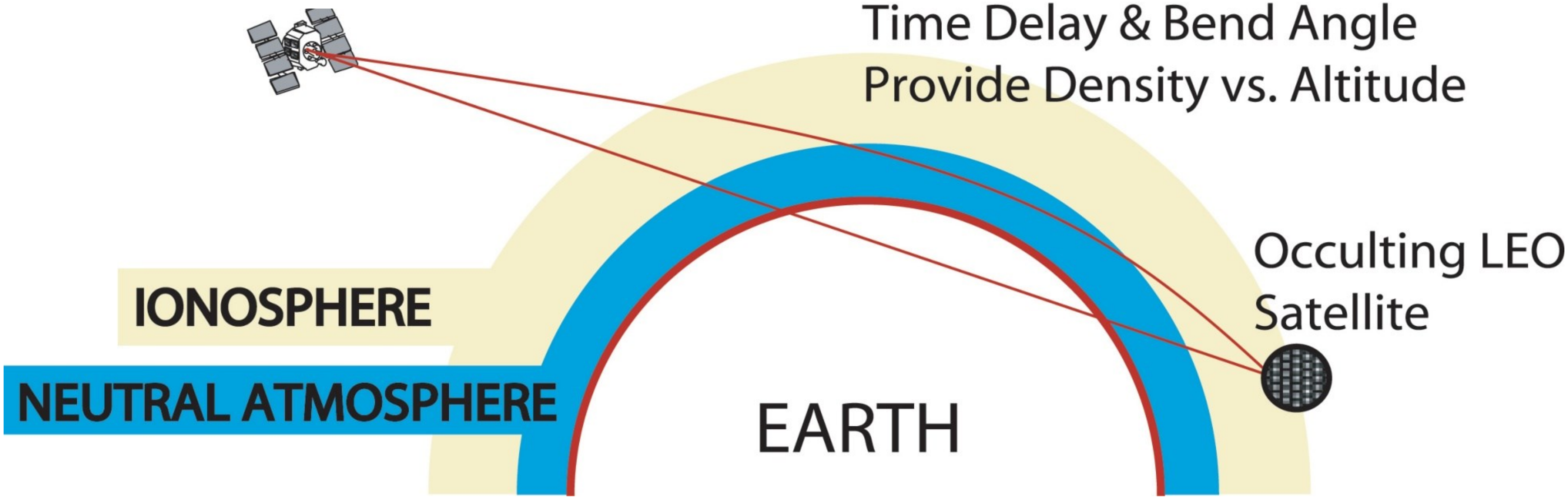
Expected forecast impact for ADM-Aeolus has been simulated using ensemble methods.

Simulated DWL data adds value at all altitudes and well into longer-range forecasts.

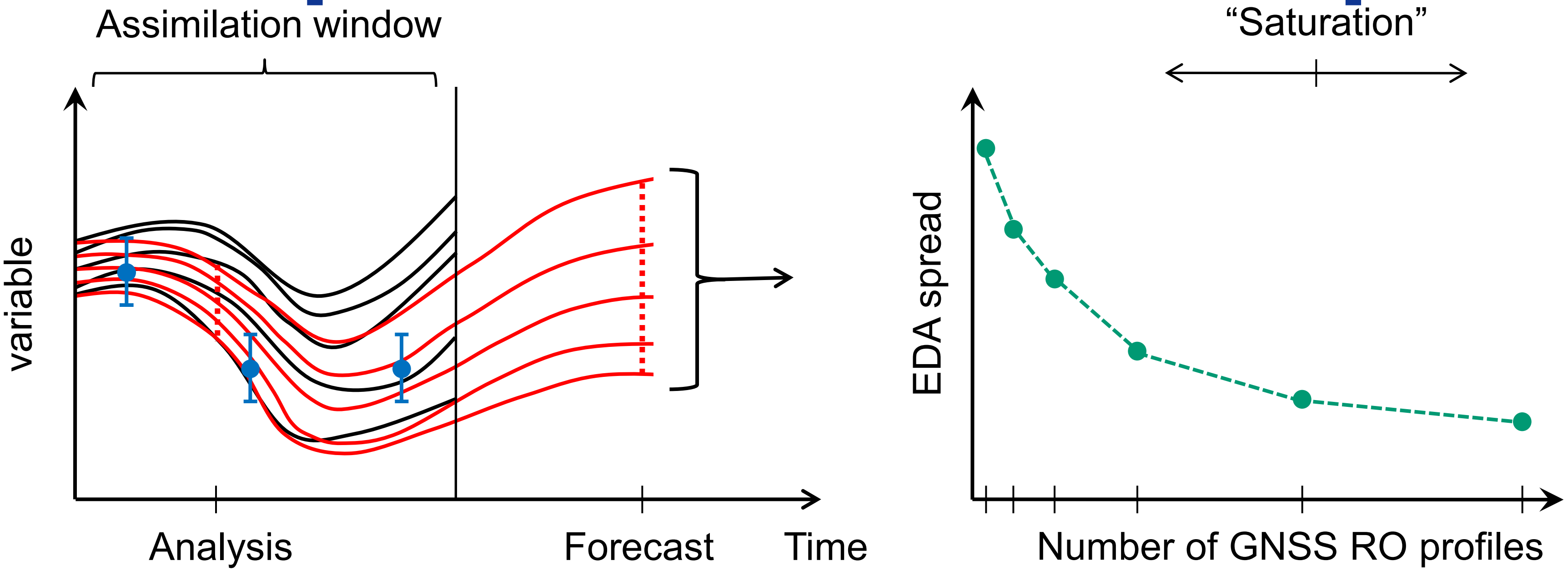


Zonal wind forecast error (m/s) for 12-hour forecast.

Example: GNSS radio occultation concept



Example: EDA based GNSS-RO impact



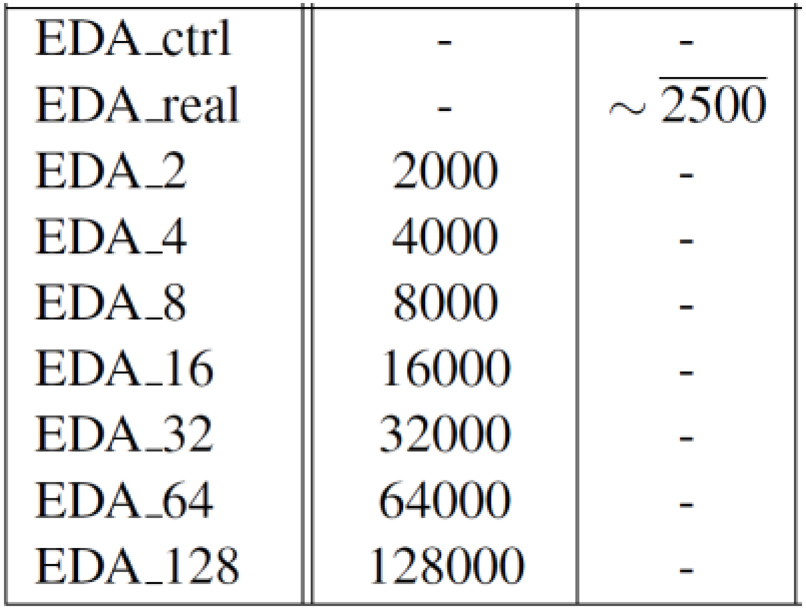
- Aim to investigate ensemble spread as a function of GNSS-RO number.
- Identify, if and when the impact begins to "saturate".

Setup of GNSS-RO experiments

- EDA experiments assimilate:

  - all operationally used GOS (apart from GNSS-RO data)

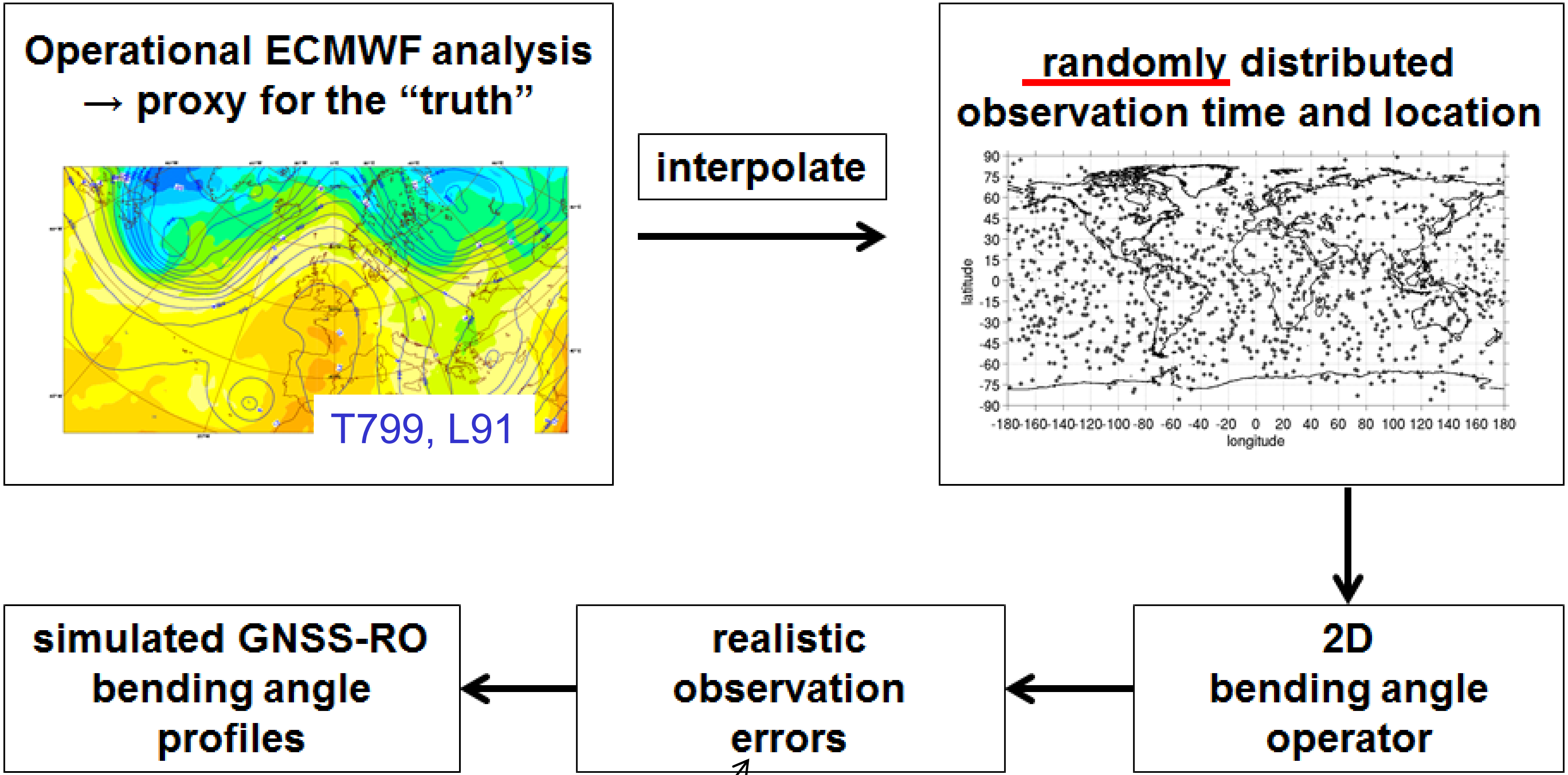
  - plus | simulated | real | GNSS-RO profiles per day



EDA_ctrl	-	-
EDA_real	-	~ 2500
EDA_2	2000	-
EDA_4	4000	-
EDA_8	8000	-
EDA_16	16000	-
EDA_32	32000	-
EDA_64	64000	-
EDA_128	128000	-

→ Total of nine EDA experiment that only differ in the number of assimilated GNSS RO data. 6 week period July-August 2008.

Simulation of GNSS-RO data



Operational ECMWF analysis → proxy for the "truth"

T799, L91

interpolate

randomly distributed observation time and location

2D bending angle operator

realistic observation errors

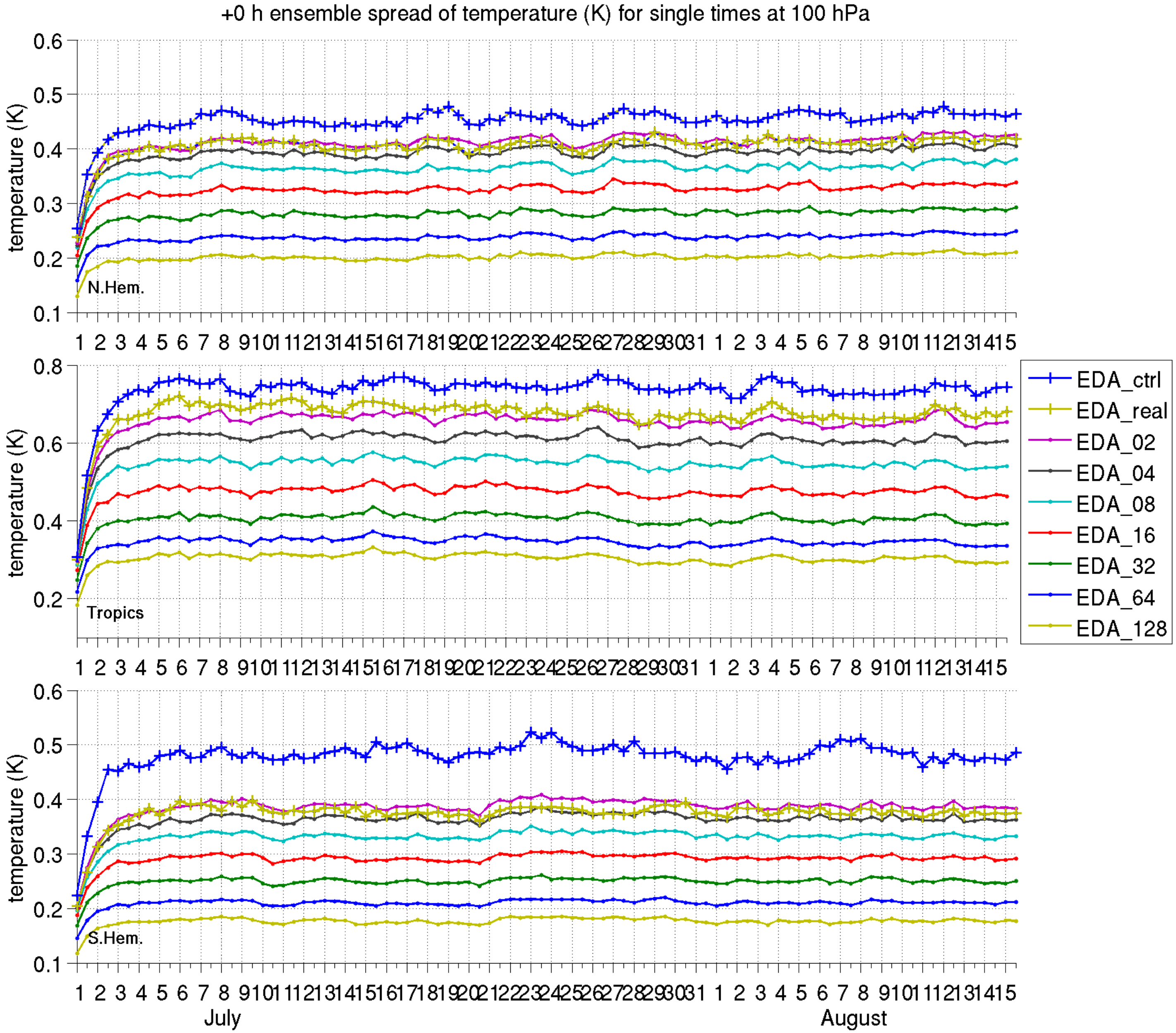
simulated GNSS-RO bending angle profiles

On 247 levels and looks like GRAS data

Adjusted to get reasonable (o-b)s

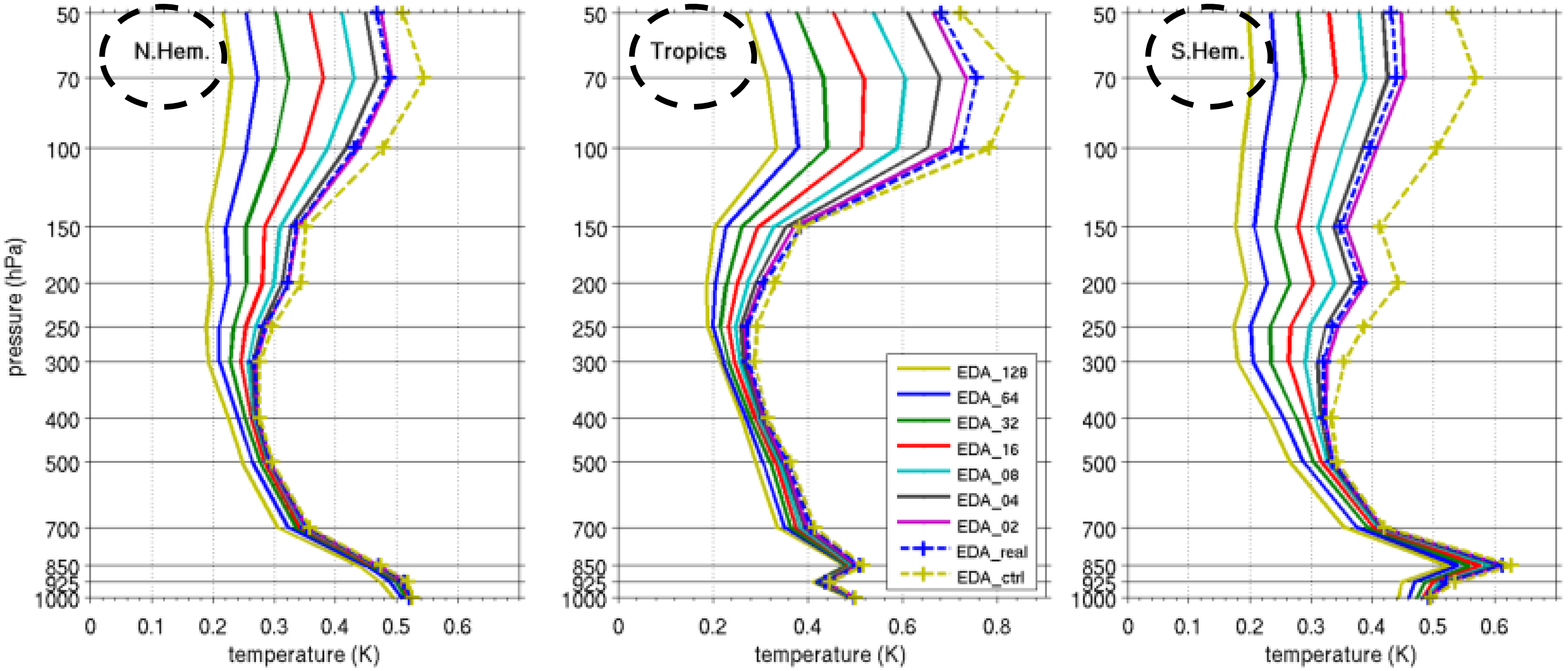
We use a 1D operator to assimilate this data.

Time series of EDA analysis spread



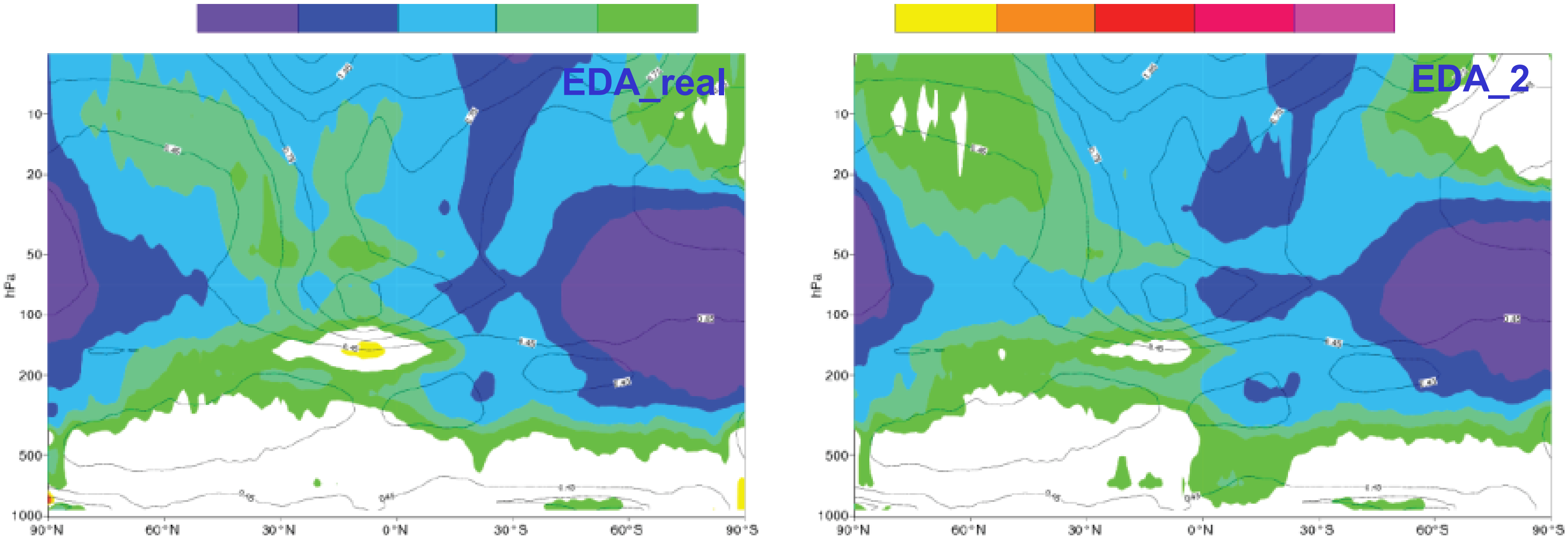
Vertical profiles of EDA spread T(K)

- Temperature uncertainty for the analysis
  → reduced with additional GNSS-RO profiles
- Very good agreement between EDA_real and EDA_2



Cross section of observation impact

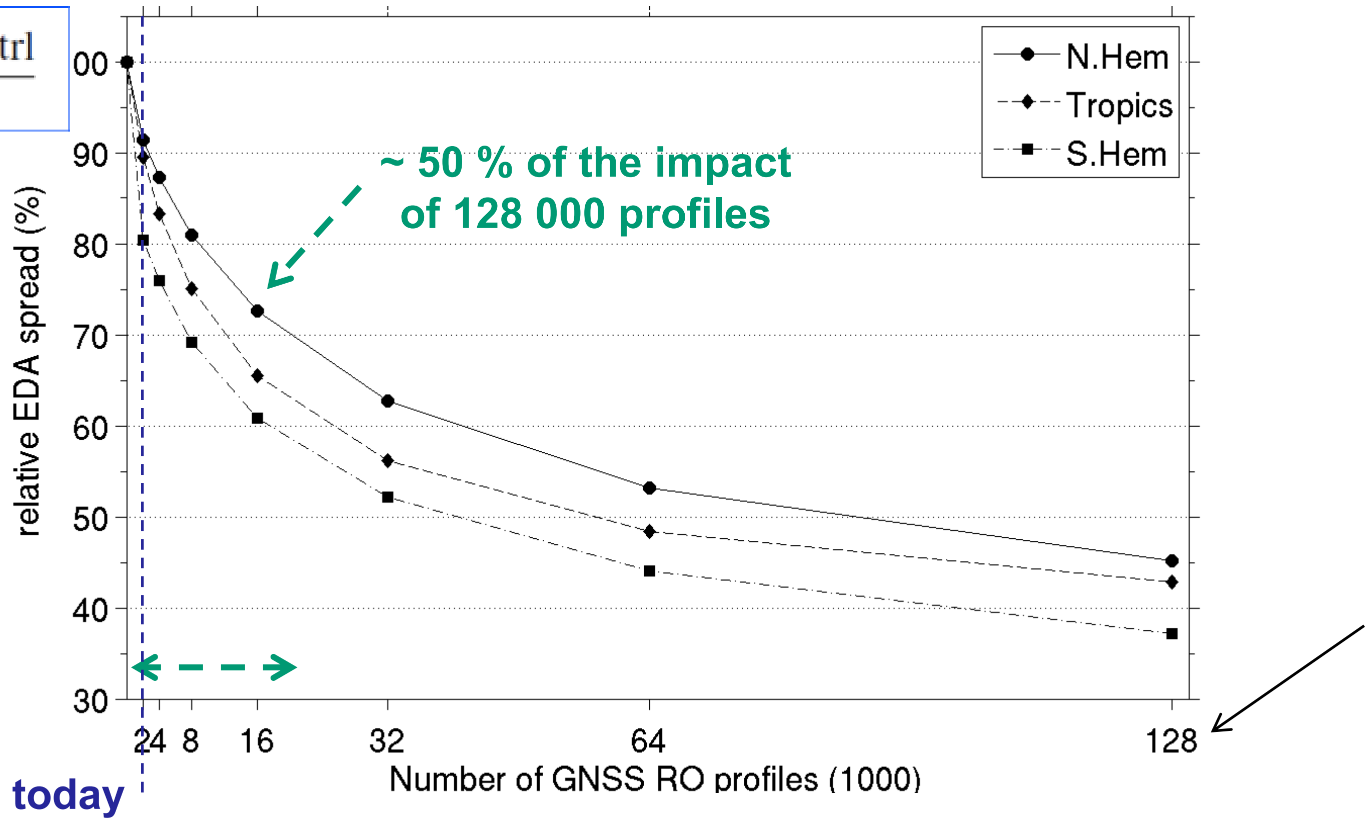
$$\frac{\mathrm{EDA\_}n - \mathrm{EDA\_ctrl}}{\mathrm{EDA\_ctrl}}$$



- Maximum impact on upper-tropospheric / middle-stratospheric temperatures
- Again, very good agreement between real and simulated GNSS RO data in the EDA system.
- Similar pattern for geopotential height

Scaling of GNSS RO impact - EDA

$$\frac{\mathrm{EDA\_}n - \mathrm{EDA\_ctrl}}{\mathrm{EDA\_ctrl}}$$



~ 50 % of the impact of 128 000 profiles

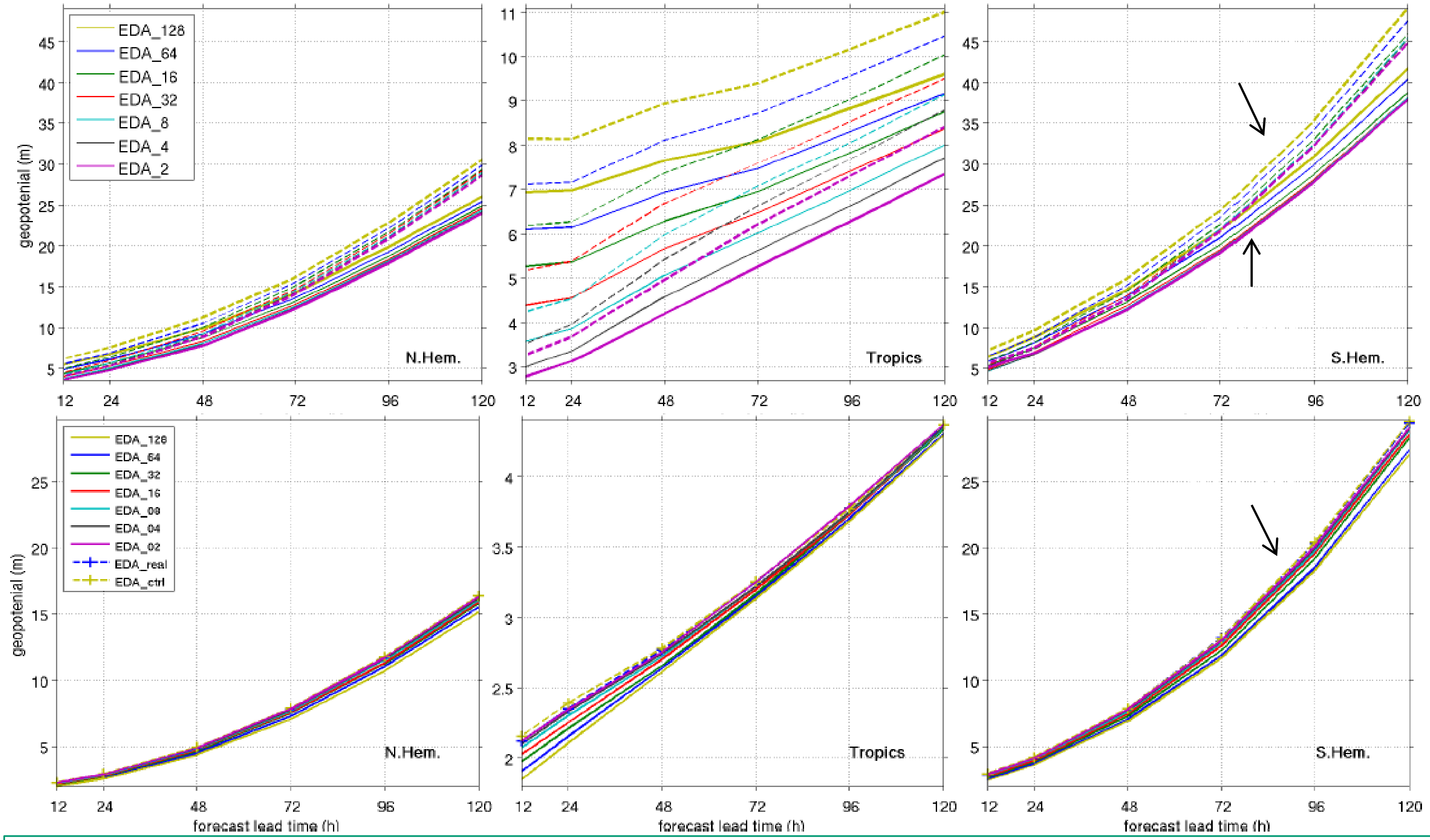
today

- Large improvements up to **16000 profiles** per day
- Even with 32000 – 128000 profiles still improvements possible

→ no evidence of saturated impact up to 128000 profiles.

EDA mean / control vs. EDA spread

geopotential height at 500 hPa

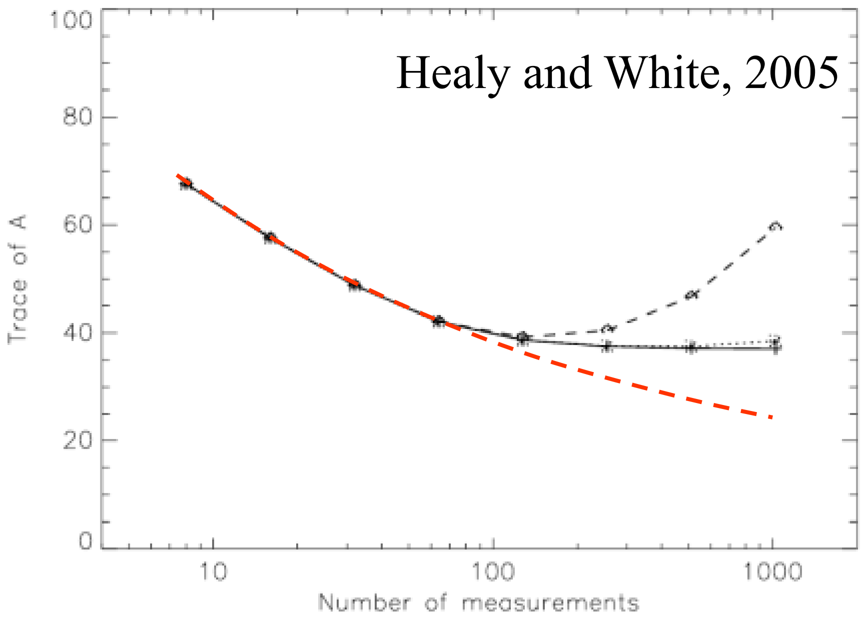


→ EDA mean / ctrl FC error not reduced, while EDA spread is reduced

Limitations: Scaling of GNSS RO impact - EDA

- Mis-specification of the input error covariance matrices can introduce additional uncertainty. We can see this in toy models.

→ incorrect specification of observation errors can lead to larger analysis std.devs as more observations are added.



Healy and White, 2005

actual uncertainty with
mis-specified
correct
observation error
covariance matrix

estimated uncertainty with
mis-specified observation
error covariance matrix

Summary

- New observations are most valuable if the provide us with new information.

- Information content studies are useful for estimating the impact of new observation types.

- The new ensemble data assimilation techniques provide a framework for estimating the impact of new missions on the 3D analysis.

  - The ensemble (EnkF, EDA) provide information about the error statistics NOT the errors

- Important tool for planning the future Global Observing System.

Summary for 3 lectures

- **Covered a lot of ground. More detail at:**
  - **http://old.ecmwf.int/newsevents/training/meteorological_presentations/2013/SAF2013/index.html**

- **Satellite data are now very important in NWP, but this was not always the case (problems in 1980's).**

- **Key point: The difference between the measurement and the retrieval product, and the need for *a priori data/constraints.***

$$\underset{\uparrow \atop \text{retrieval}}{\mathbf{x}_a} = (\mathbf{I}-\mathbf{KH})\underset{\uparrow \atop \text{prior}}{\mathbf{x}_b} + \mathbf{K}\underset{\uparrow \atop \text{measurement}}{\mathbf{y}_m}$$

Summary

- Always question what the "satellite temperature (or humidity or wind) measurement …" actually is, because the original problem was probably ill-posed.

- Variational assimilation/retrievals techniques can look daunting, but they are just a least-squares approach, written in matrix/vector form.

  - WE COULD WRITE FITTING y = (ax+ b) TO DATA LOOK LIKE THE 4D-VAR COST FUNCTION IF WE WANTED.

- If you have a good understanding of the forward problem, H (y=Hx), and the observation error statistics, R, you are more likely to interpret the data, y, correctly.