A Selective Look Back Over the Last Twenty Years of Variational Data Assimilation

Mike Fisher[1]

ECMWF

June 3, 2015

[1]Thanks to: Sarah Keely, Lars Isaksen, Massimo Bonavita, etc. from whom I shamelessly stole slides

Outline

Introduction

ECMWF implemented 3D-Var in January 1996 — nearly 20 years ago.

Since then, a lot has happened...

- 4D-Var (in November 1997)
- Several changes of horizontal and vertical resolution
- A big increase in the number an types of observations assimilated
- Improvements to model physics, including to the TL & adjoint models
- Better modelling of observation errors and biases, including:
  - Variational bias correction
  - Variational quality control
- Modelling of model bias (weak constraint 4D-Var)
- Better modelling of background error
  - Wavelet J_b
  - Flow-dependent covariances derived from an Ensemble of Data Assimilations (EDA)
- etc. etc. etc.

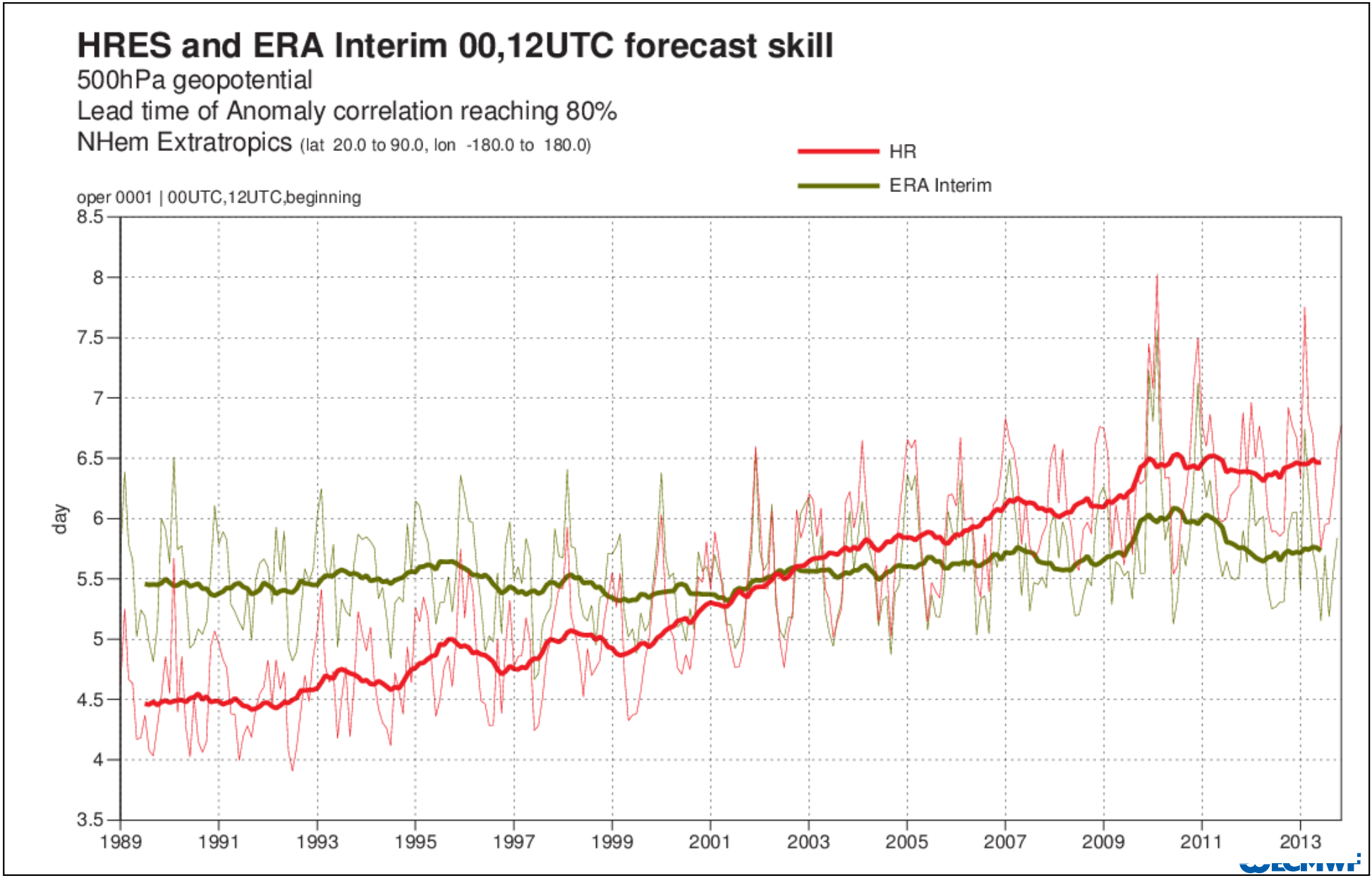
ECMWF

Introduction

As a result of all these changes, forecasts have gained almost 2 days of predictability compared with those of 20 years ago.

We can get an idea of how much of this improvement came from improvements in the observing system by comparing with a frozen (reanalysis) system.

Operational System versus Reanalysis

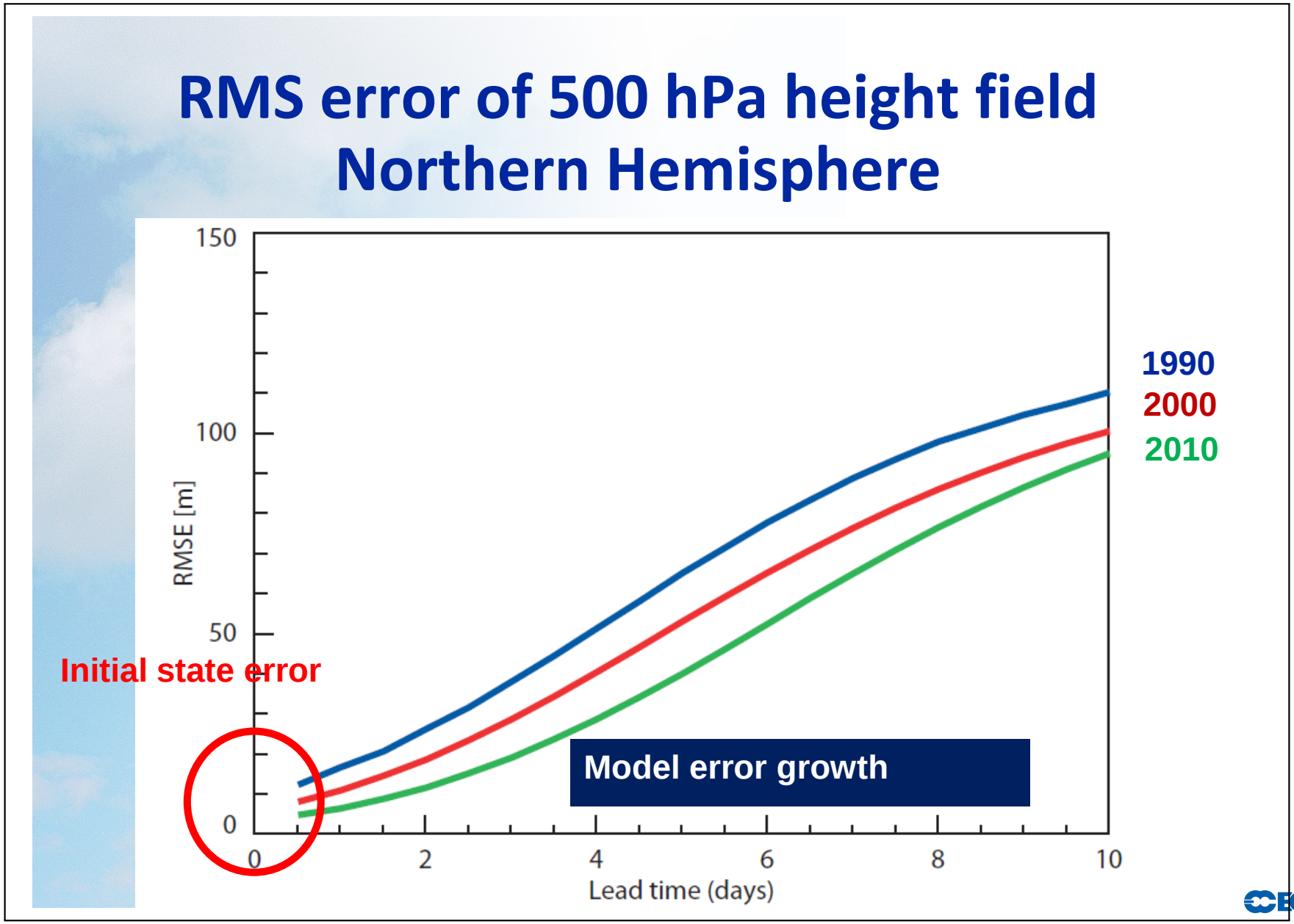


Introduction

It is difficult to separate the impact of improvements in the model from improvements in the data assimilation system, since the model is a component of the DA system.

But, it is clear that reduction in initial-state error is the main contributor to the improvement in forecast skill.

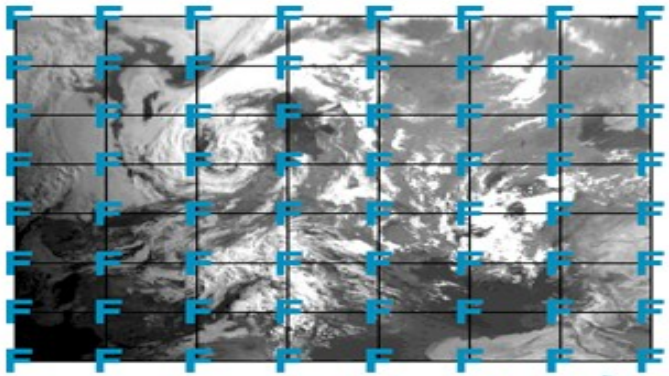
Introduction

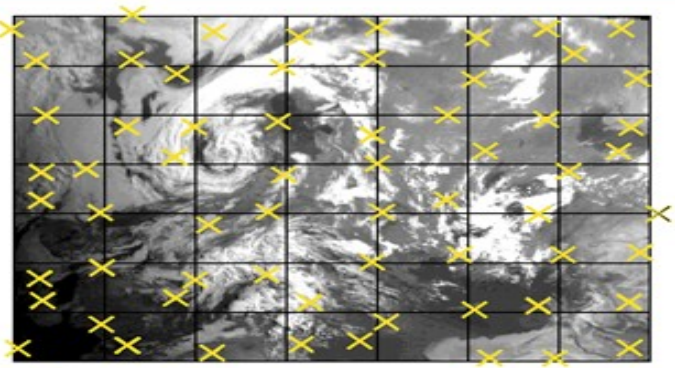


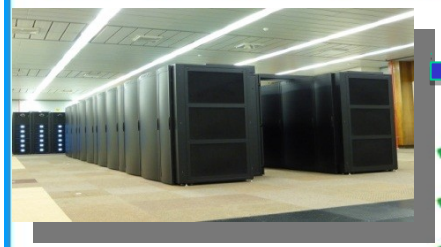
Introduction

Model Forecast (with errors)

Observations (with errors)

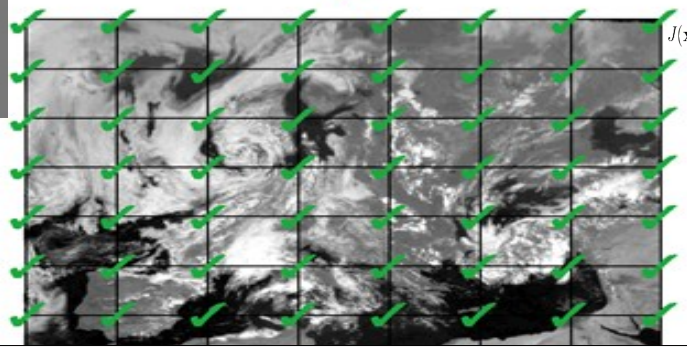






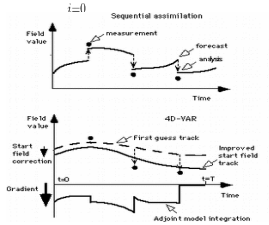
Computer (with a lot of CPUs)

People (with a lot of good ideas)

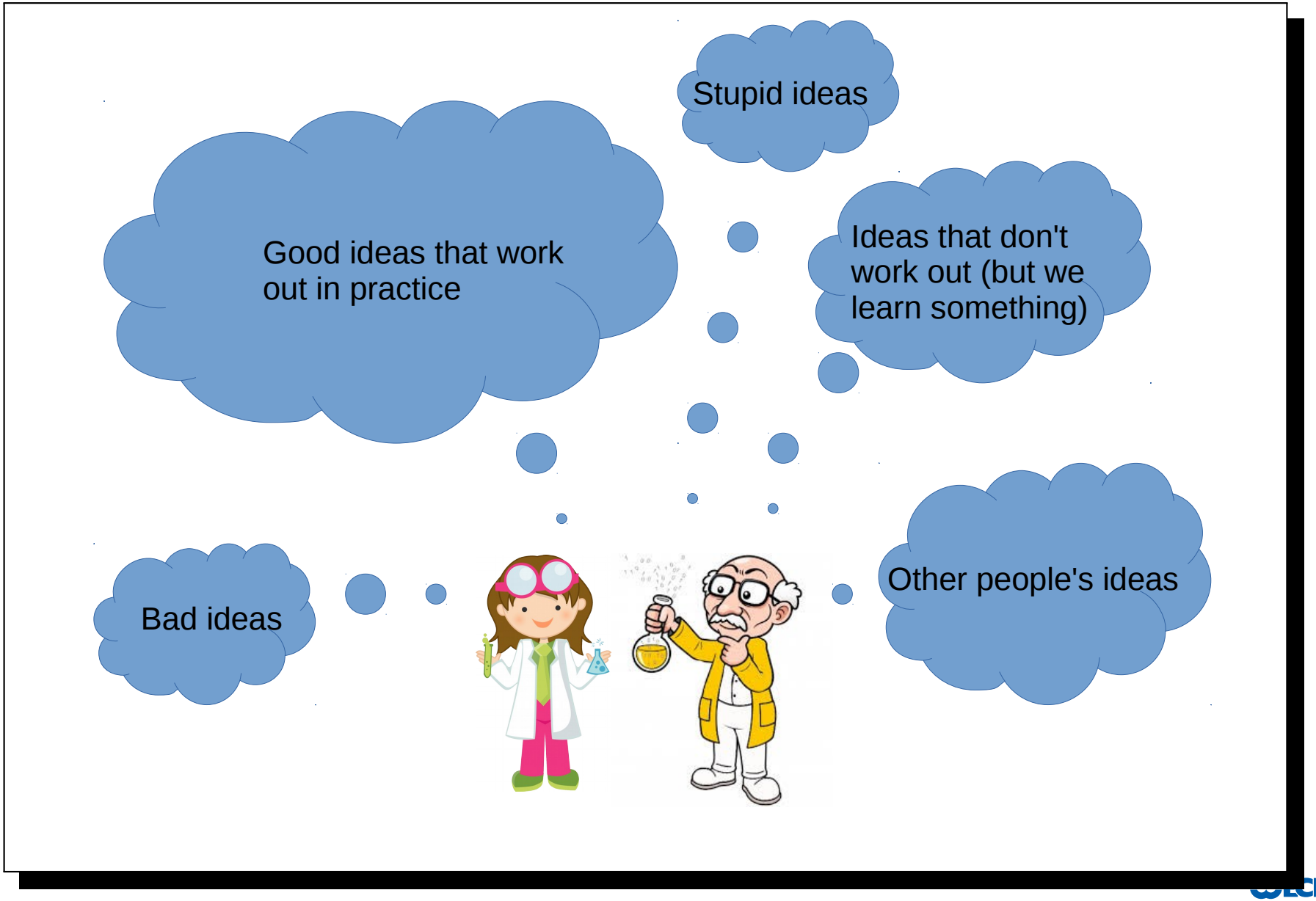


Analysis (with - smaller – errors)

$$J(\mathbf{x}) = (\mathbf{x}-\mathbf{x}_b)^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b) + \sum_{i=0}^{n}(\mathbf{y}_i - H_i[\mathbf{x}_i])^{\mathrm{T}}\mathbf{R}_i^{-1}(\mathbf{y}_i - H_i[\mathbf{x}_i])$$



Introduction



Some good ideas that worked out in practice

It would be impossible to cover all the good ideas that have contributed to the improvement in forecast skill over the last 20 years.

Instead, I have picked a few examples with which I have been personally involved, or which I find particularly interesting:

- The Derber-Bouttier J_b
- Wavelet covariance model for background errors
- Huber norm quality control

Derber-Bouttier J_b

When 3D-Var was first implemented at ECMWF, the background covariance model was based on a normal model decomposition:

$$J_b = \tfrac{1}{2}\, c_R \left(\frac{\Delta x_R}{\sigma_R}\right)^t (h_R^{-1/2})^t\, (V_R^{-1/2})^t\, V_R^{-1/2}\, h_R^{-1/2} \left(\frac{\Delta x_R}{\sigma_R}\right)$$
$$+ \tfrac{1}{2}\, c_G \left(\frac{\Delta x_G}{\sigma_G}\right)^t (h_G^{-1/2})^t\, (V_G^{-1/2})^t\, V_G^{-1/2}\, h_G^{-1/2} \left(\frac{\Delta x_G}{\sigma_G}\right)$$
$$+ \tfrac{1}{2} \left(\frac{\Delta x_U}{\sigma_U}\right)^t (h_U^{-1/2})^t\, (V_U^{-1/2})^t\, V_U^{-1/2}\, h_U^{-1/2} \left(\frac{\Delta x}{\sigma_U}\right)$$

It was rather complicated, and did not work particularly well.

In particular, it did not give good results in the tropics, and there were problems with the tides.

Derber-Bouttier J_b

Derber and Bouttier (1997) simplified the background covariance model to:

$$J_b = \frac{1}{2}\chi\chi^{\mathrm{T}}$$

where $x = x_b + B^{1/2}\chi$

The new formulation was cheaper (4 times fewer spectral transforms), simpler, and more elegant than the old scheme.

It also produced a better-conditioned minimization problem.

Derber-Bouttier J_b

The clever part of the Derber-Bouttier scheme was the choice of a control variable and a balance operator that gave up gracefully in the tropics:

$$\chi_{\text{bal}} \xrightarrow{B_\zeta^{1/2}} \zeta_{\text{bal}} \xrightarrow{\text{“linear balance”}} \Phi_{\text{bal}} \xrightarrow{\text{regression}} (T_{\text{bal}}, p_{\text{bal}})$$

Because of the way f appears in the linear balance equation, $\Phi_{\text{bal}} \longrightarrow 0$ in the tropics.

This avoids imposing inappropriate balances in the tropics.

(An inability to “fail gracefully” in the tropics would stymie later attempts to develop a potential-vorticity based balance operator for J_b.)

Derber-Bouttier J_b

The use of regression to compute $(T_{\text{bal}}, p_{\text{bal}})$ takes into account that Φ_{bal} is not the true geopotential, but a diagnostic quantity derived from the vorticity field.

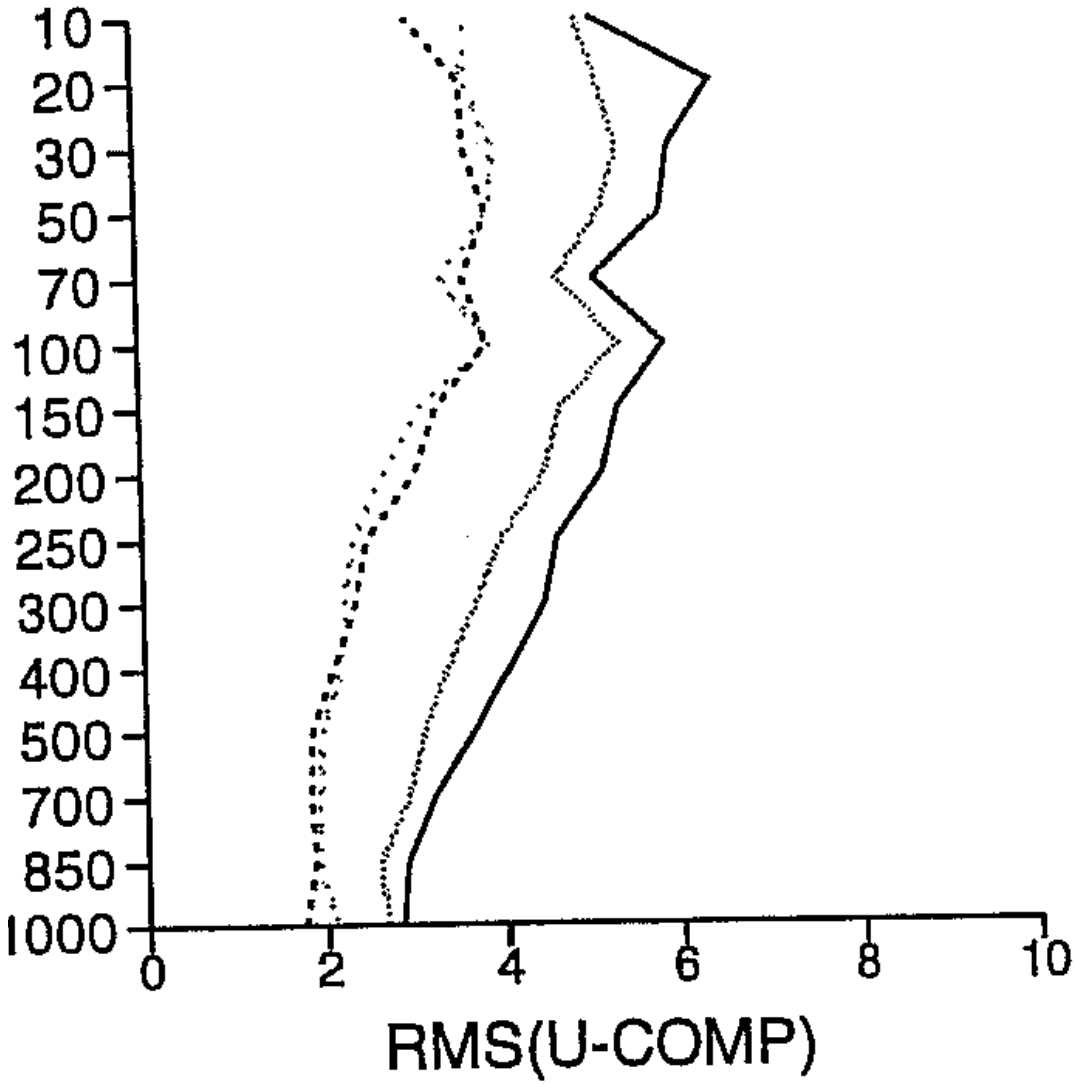
It also neatly sidesteps the fact that the analytical hydrostatic equation is not invertible for the vertical grid staggering used in the ECMWF model.

Overall, the Derber-Bouttier covariance model is a good example of the KISS[2] principle.

The impact on forecast scores (particularly in the tropics) was dramatic.

[2] Keep It Simple, Stupid

Derber-Bouttier J_b



from Derber and Bouttier 1997, ECMWF Technical Memorandum 238

Derber-Bouttier J_b

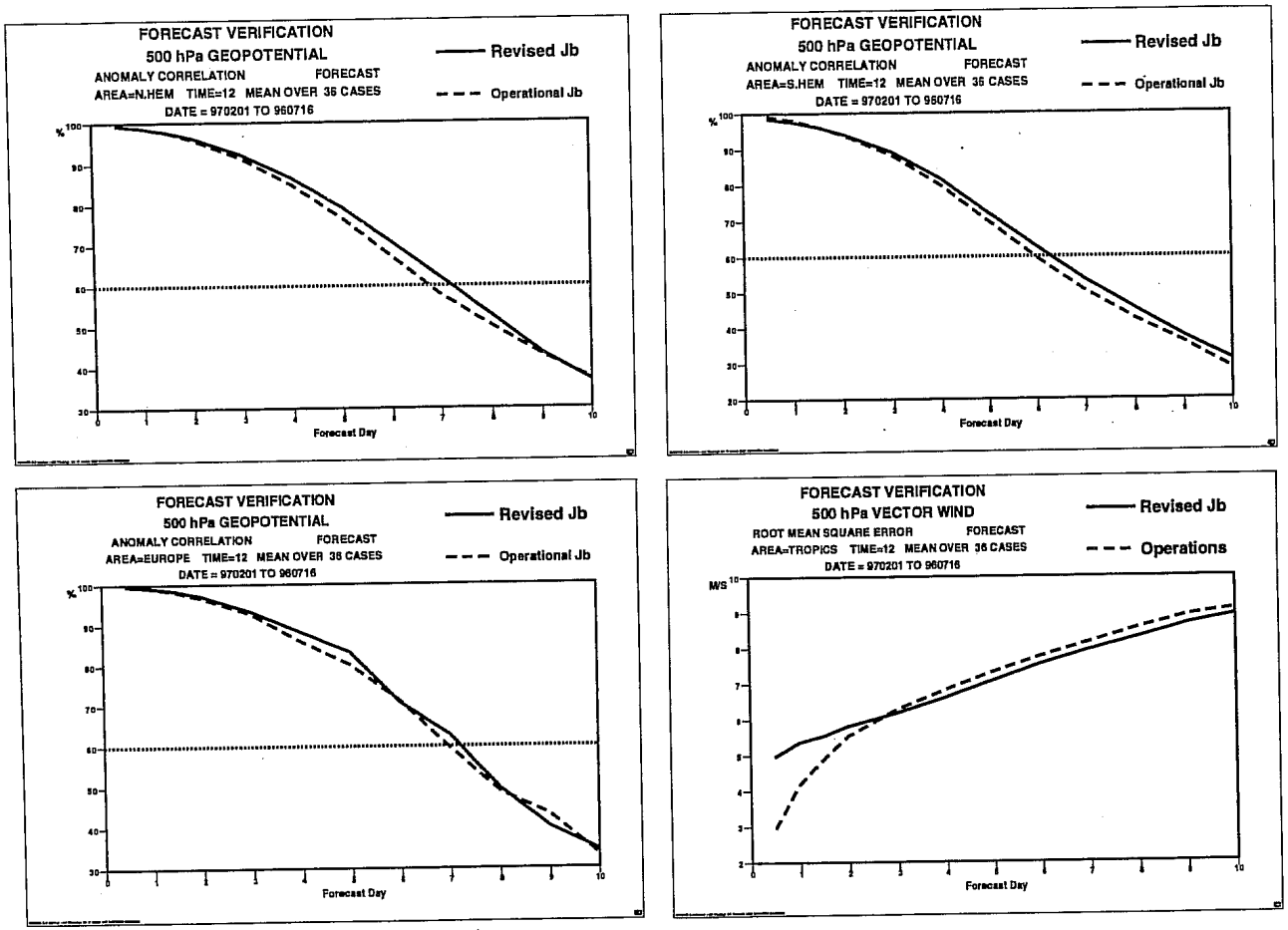


Figure 36: Impact on the forecast scores on the J_b revision (only) over 5 weeks of experimental assimilation (2 in February 1997, 3 in June/July 1996).

from Derber and Bouttier 1997, ECMWF Technical Memorandum 238

Wavelet Covariance Model

ECMWF Newsletter No. 106 – Winter 2005/06

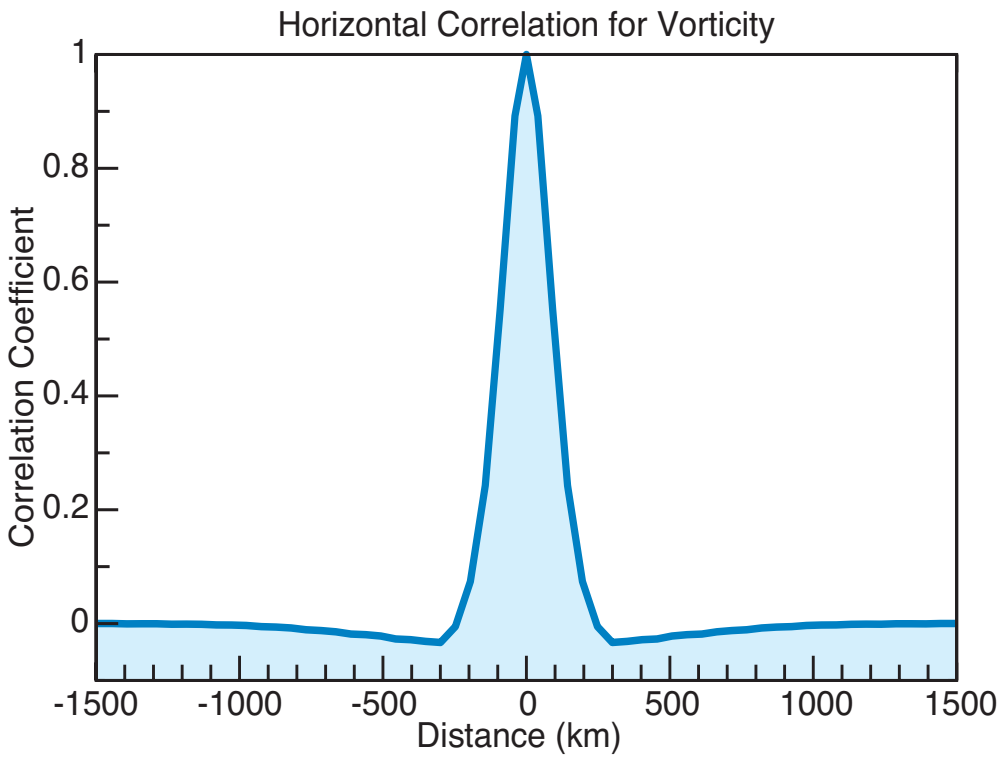


Figure 2 Mean horizontal correlation of vorticity at model level 49 (near 850 hPa) as a function of distance.

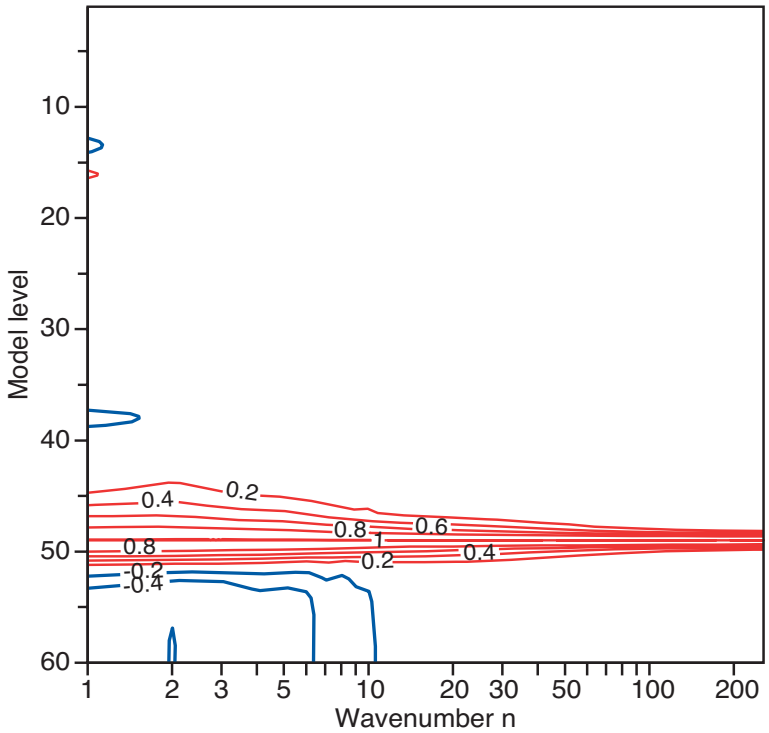


Figure 1 Mean correlation between background errors of temperature on model level 49 (near 850 hPa) and the corresponding errors on other model levels, as a function of spherical wave number.

Wavelet Covariance Model

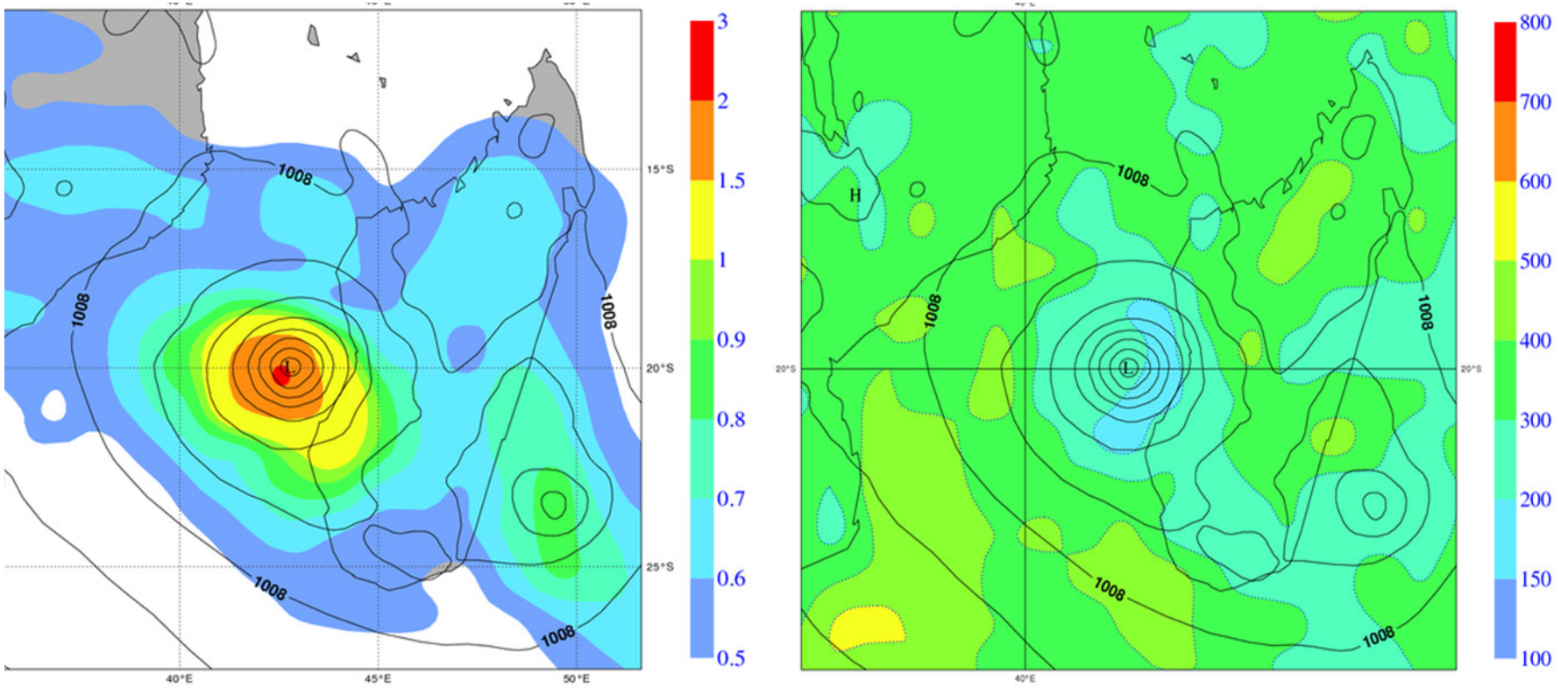


Figure 14: EDA estimate of surface pressure background error standard deviation (shaded contours, left panel) and EDA estimate of background error correlation length scale (shaded contours, right panel) in the area affected by hurricane Fanele. Plots are derived from a 50 member EDA, valid on 20 January 2009 at 09 UTC. Mean Sea Level Pressure contours (4 hPa interval) are superimposed.

Bonavita 2011, proc ECMWF Seminar on Data Assimilation

Wavelet Covariance Model

The spherical harmonics for a single wavenumber n have amplitude everywhere, so a covariance model that specifies vertical covariances for each n has no possibility for modelling the spatial variation of covariance.

Conversely, specifying the covariances at each gridpoint does not allow any variation of covariance with wavenumber.

The idea of the wavelet covariance model is to trade off some spectral resolution in exchange for some spatial resolution.

The result is a better model for the covariances, in the same way that the musical score describes the tune better than a list of frequencies (with no indication when they are to be played) or a list of times (with no indication which note is to be played).

Wavelet Covariance Model

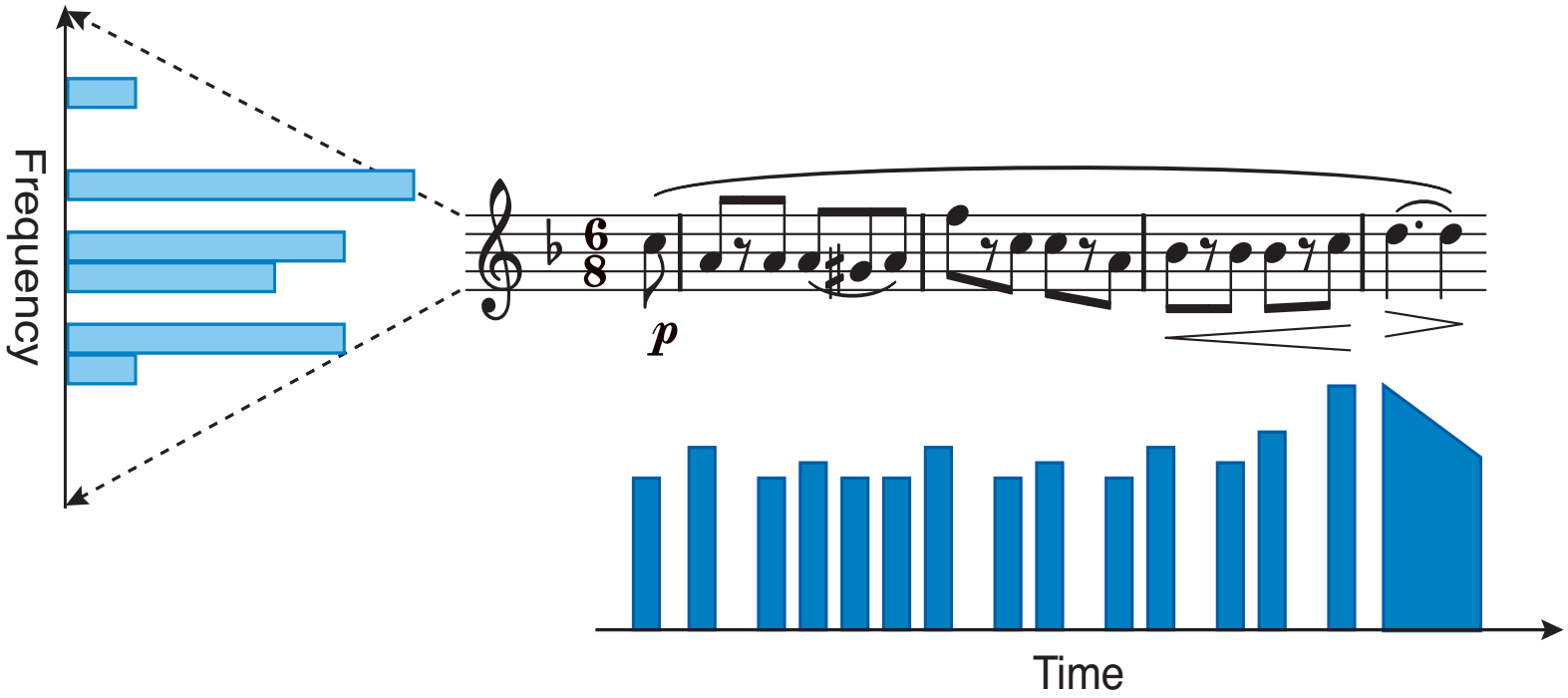


Figure 3 An illustration of the benefits of resolving both temporal and spectral information.

Fisher 2006, ECMWF Newsletter 106

Wavelet Covariance Model

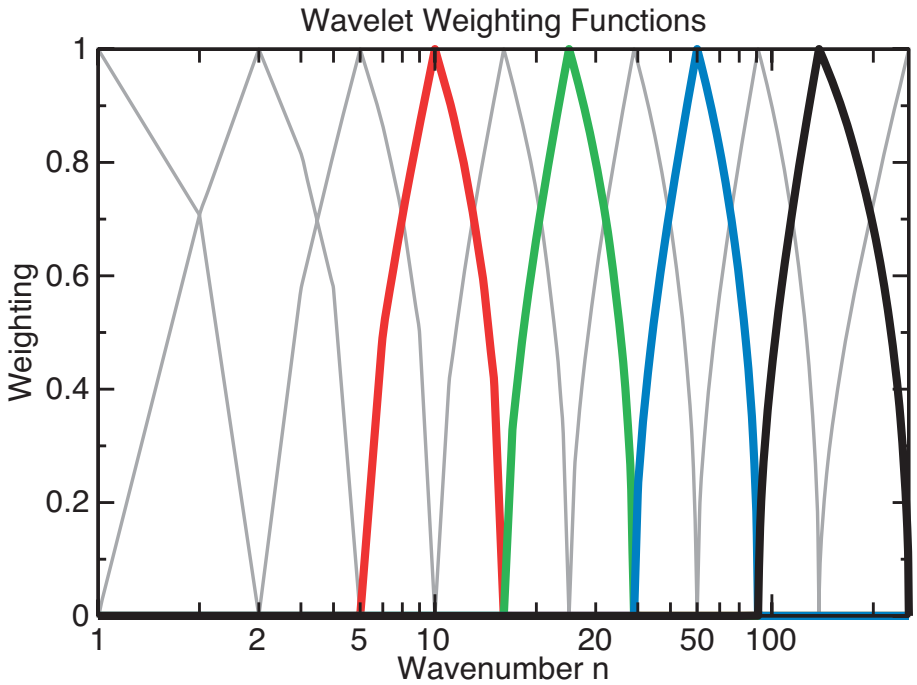


Figure 4 Weighting functions for the different wavenumber bands in "Wavelet" J_b. The coloured curves are referred to in Figure 5.

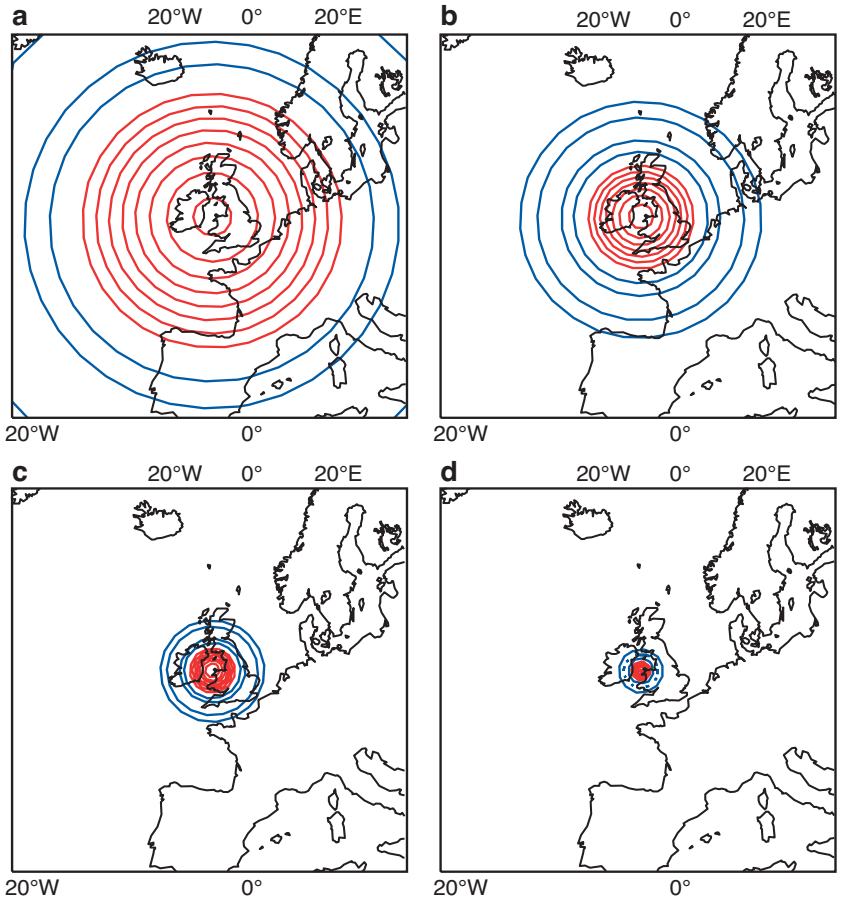


Figure 5 The spatial weighting functions corresponding to the functions of wavenumber highlighted in Figure 4, for a point over the Irish Sea. Red and blue contours represent positive and negative weights. The zero line is not plotted. Plots (a), (b), (c) and (d) refer to the red, green, blue and black curves of Figure 4.

Fisher 2006, ECMWF Newsletter 106

Wavelet Covariance Model

The wavelet covariance model allows us to specify vertical covariances that vary geographically, while retaining non-separability (i.e. that large/small horizontal scales correspond to deep/shallow vertical correlations).

By assigning different variances to each waveband, we can control the horizontal covariances, and allow them to vary spatially.

When originally introduced, we were only able to capture climatological spatial variation.

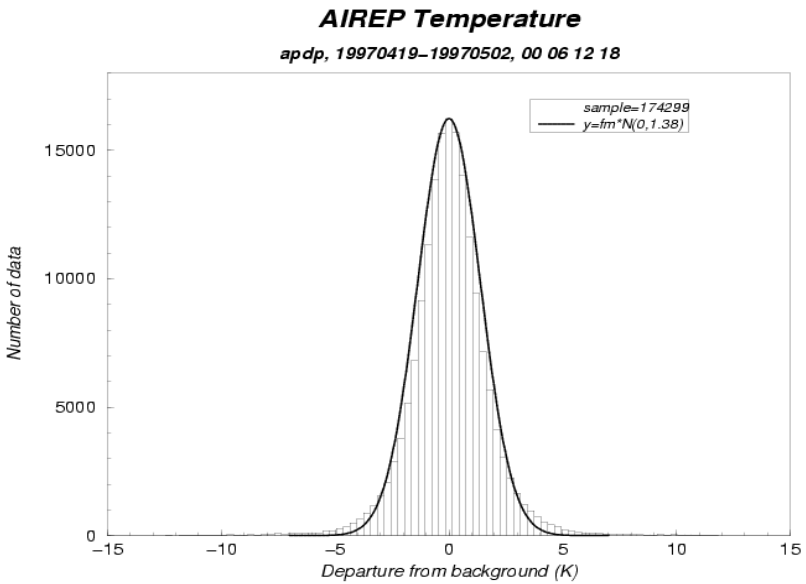
Now, we we can fully exploit the potential of the wavelet formulation by using an Ensemble of Data Assimilations to construct a flow-dependent covariance model.

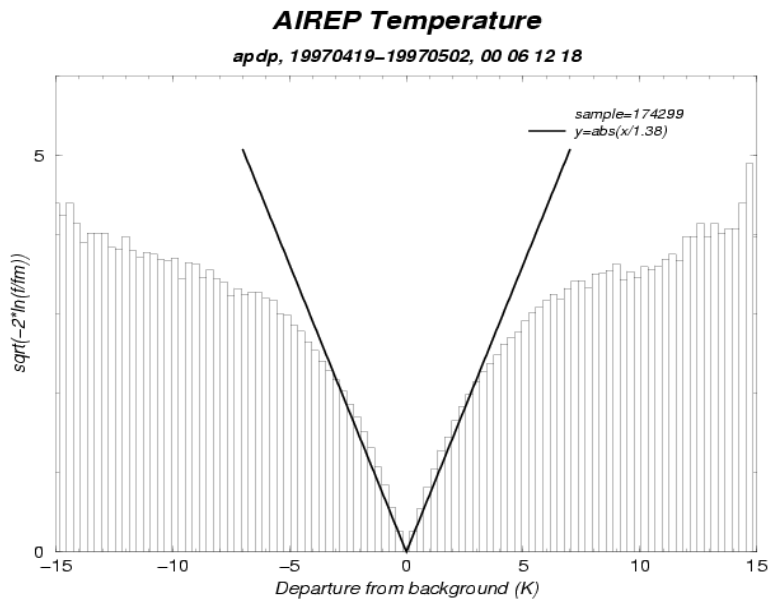
Huber norm quality control

Tuning the rejection limit

The left histogram on the left has been transformed into the right histogram such that the Gaussian part appears as a pair of straight lines forming a 'V' at zero.
The slope of the lines gives the standard deviation of the Gaussian.

The rejection limit can be chosen to be where the actual distribution is some distance away from the 'V' - around 6 to 7 K in this case, would be appropriate.





Huber norm quality control

The theory of variational quality control was well known (Lorenc and Hammon 1988; Lorenc and Ingleby 1993).

The idea is to replace the Gaussian observation error statistics with something that better describes the true statistics of model error (and, in particular, takes into account the fact that outliers are more likely than a Gaussian model would suggest).

VarQC was implemented in the ECMWF system (Andersson and Jarvinen 1999).

In principle, all that is required is to specify the correct (non-Gaussian) observation error statistics, and 4D-Var will weight the observations correctly.

In practice, there were occasions when large numbers of mutually supporting observations were rejected.

Huber norm quality control

27 Dec 1999 – French storm 18UTC

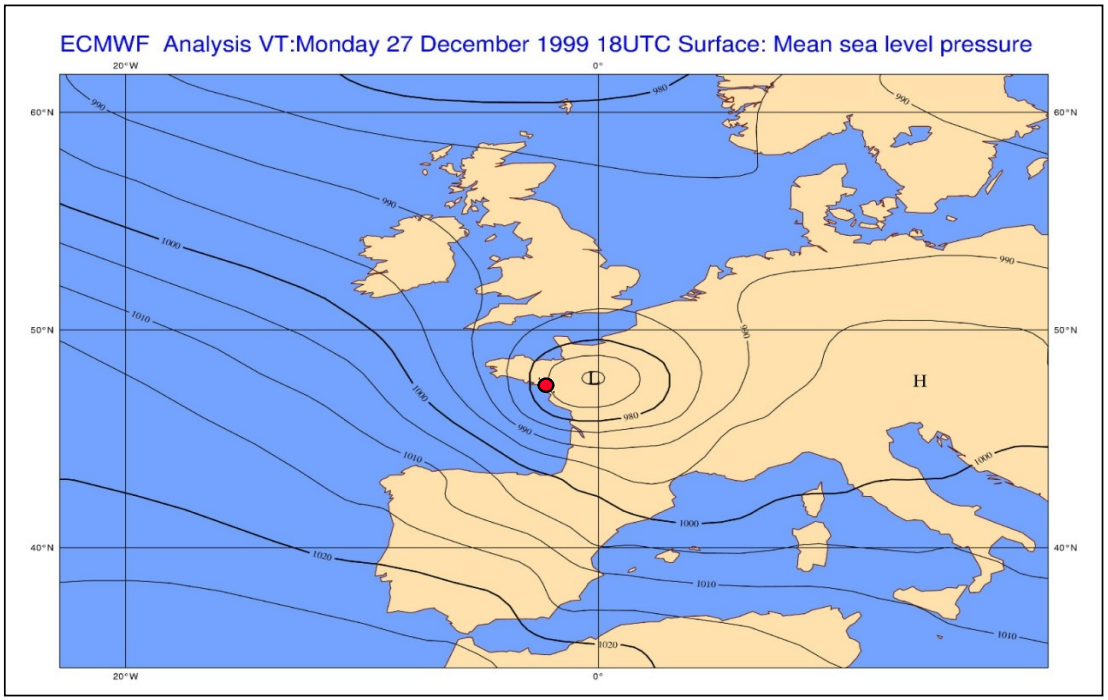
- Era interim analysis produced a low with min 970 hPa
- Lowest pressure observation (SYNOP: red circle)

  963.5 hPa (supported by neighbouring stations)

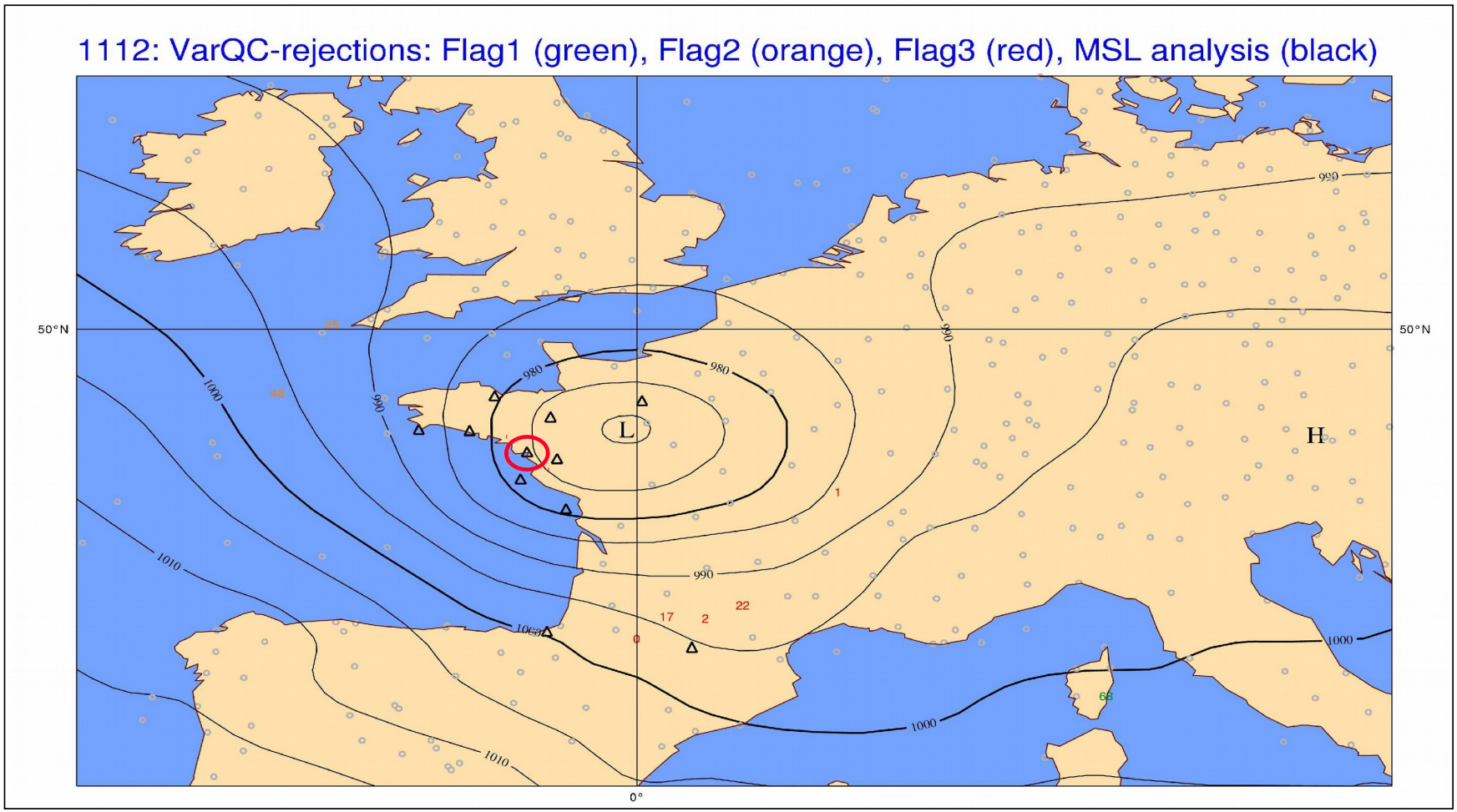
  At this station the analysis shows 977 hPa

  Analysis wrong by 16.5 hPa!

- High density of good quality surface data for this case

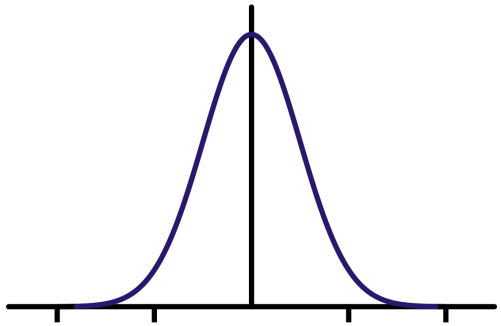


Huber norm quality control

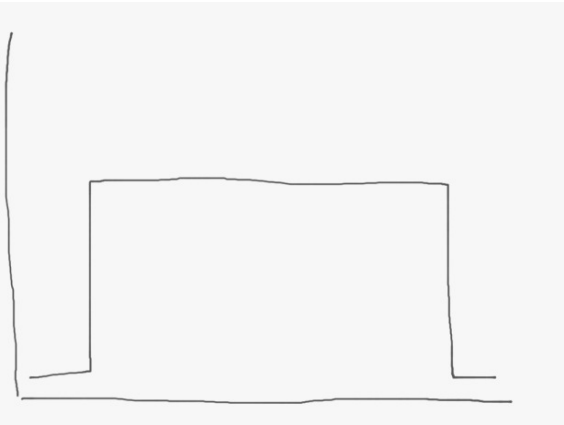


Huber norm quality control

Most observation errors look like this:



But, occasionally we get gross errors (when anything is possible):



$$p(\text{obs error}) = A \times p_{\text{normal}} + (1 - A) \times p_{\text{gross}}$$

Huber norm quality control

This error model seems entirely reasonable, so why does it allow large numbers of mutually supporting observations to be rejected?

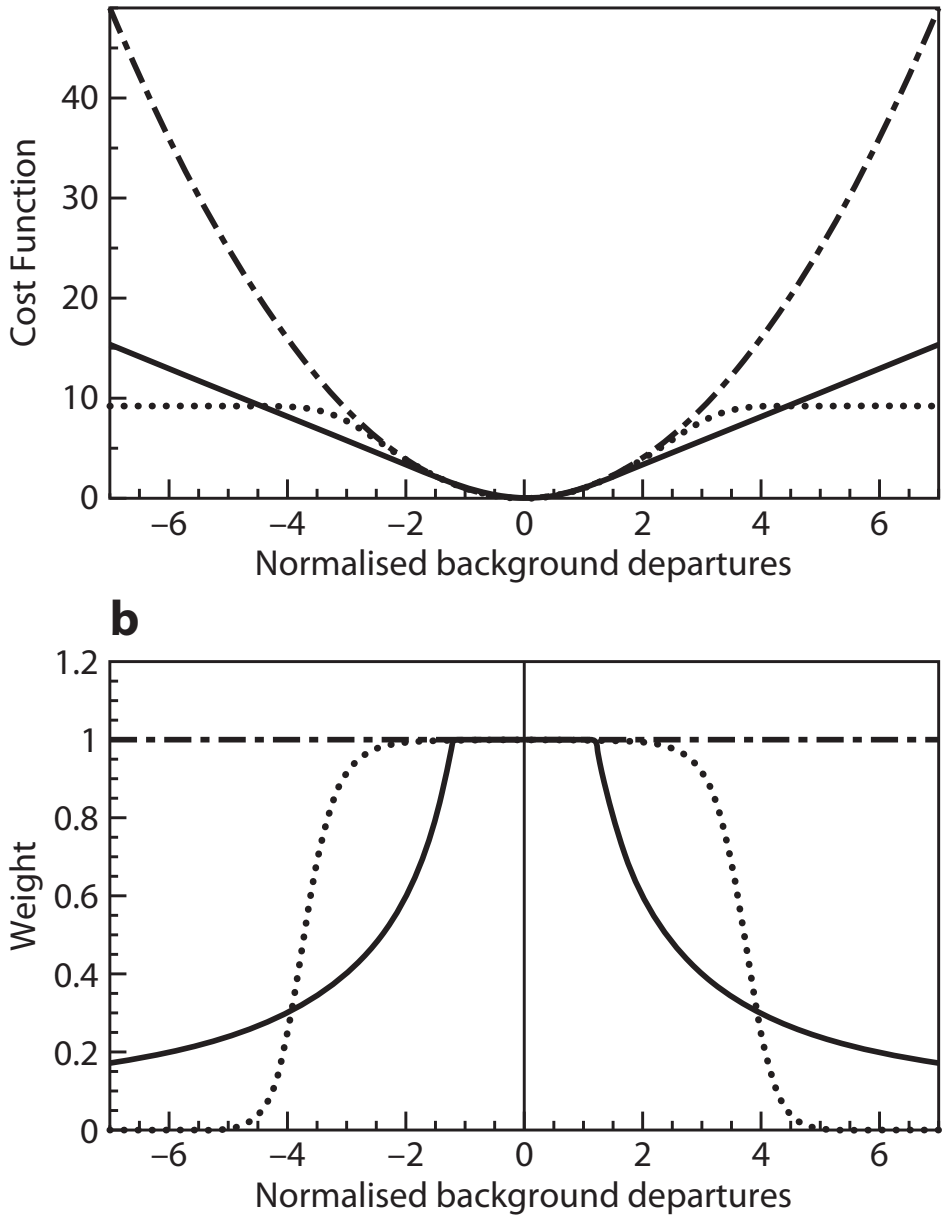
The problem is that the quality control is too "binary".

Observations outside the acceptance region are given zero weight, even if there are many of them.

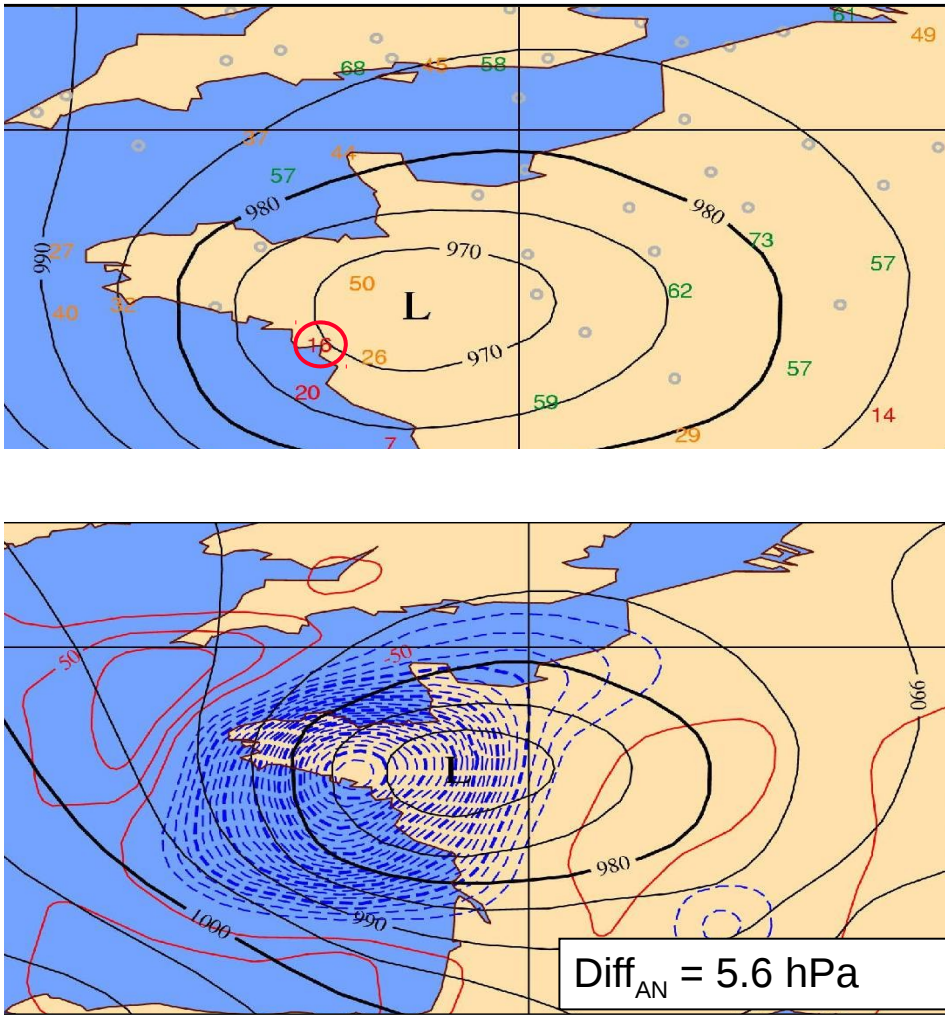
By giving such observations a reduced (but non-zero) weight, we allow rejected observations to have a small influence on the analysis.

If there are several such observations, their small weights will add up to a large influence on the analysis.

Huber norm quality control



Huber norm quality control



RRKF – A good idea that didn't work out

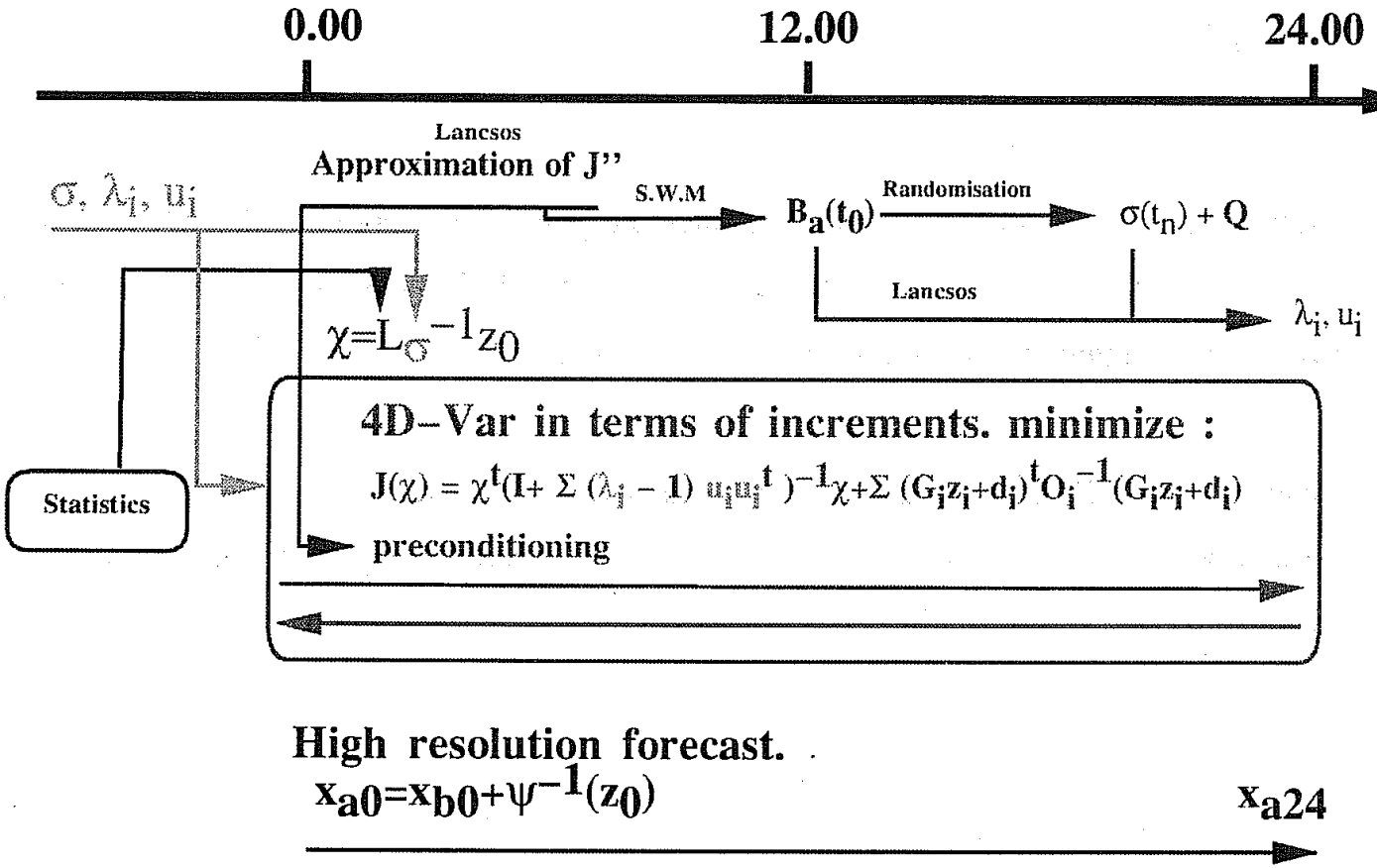


Fig. 2 Synopsis of the proposed algorithm for cycling 4D-Var.

from: Courtier 1993: "Introduction to Numerical Weather Prediction Data Assimilation Methods", proc. ECMWF/EUMETSAT Seminar on Developments in the Use of Satellite Data in Numerical Weather Prediction, 6-10 September 1993.

The Reduced-Rank Kalman Filter

The main ideas behind the RRKF were:

- Approximation of the Hessian $[J'' = (P^a)^{-1}]$ based on eigenvectors and eigenvalues calculated using a Lanczos algorithm.
- Use of the eigenvalues and eigenvectors to precondition the minimization.
- Use of the Shermann-Morrison-Woodbury formula to invert the approximation to get P^a.
- Randomization to produce samples drawn from $\mathcal{N}(0, P^a)$.

All these ideas proved useful individually, and still form core parts of the ECMWF DA system (and other systems), even though the RRKF as a whole failed to live up to expectations.

The Reduced-Rank Kalman Filter

For example, work on the Lanczos algorithm lead to CONGRAD:

- Still a widely used (and very efficient) minimization method for 3D/4D-Var.
- Generates eigenpairs for free while it minimizes
- Eigenpairs are very useful diagnostics (particularly in cases of slow or non-convergence)
- Allows an efficient spectral preconditioner to be built during the first incremental minimization without additional effort
- Allows information content to be calculated with strict mathematical bounds (Golub and Meurant 1993)

Although not the full RRKF, a method for cycling background errors came out of the RRKF work. The method was used at ECMWF between 1996 and 2005.

The Reduced-Rank Kalman Filter

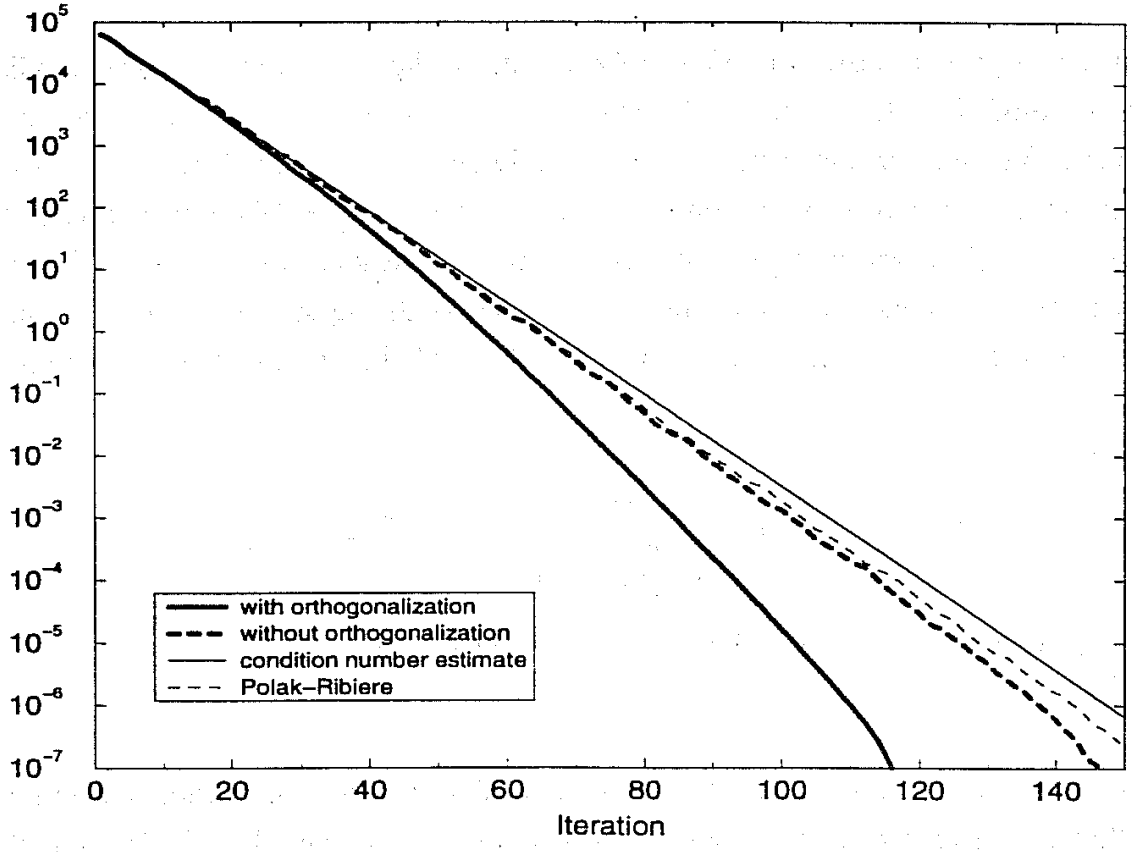


Figure 1: Convergence of conjugate gradients as a function of iteration for a 4dVar cost function. The abscissa is the square of the Hessian norm of the difference between the control vector at a given iteration and its value at the 118^{th} iteration of the orthogonalized algorithm. The thick dashed curve is for the standard conjugate gradient algorithm. The thick solid curve is for the version of the algorithm in which each gradient is explicitly orthogonalized against its predecessors. The thin dashed curve uses the Polak-Ribiere method to determine the descent direction. The thin solid line is the upper bound on the convergence rate defined by equation 12.

Fisher 1998, proc ECMWF Seminar

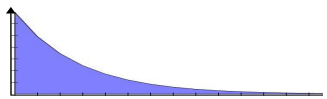
The Reduced-Rank Kalman Filter

Why did RRKF fail?

The main problem is that the RRKF relies on generating a small subspace that contains a significant amount of analysis error.

A reasonable fraction of forecast error can be captured in this way (e.g. Martin Leutbecher's talk).

But the spectrum of analysis error is flat:



Moreover, initial-time SVs are very sensitive to the choice of initial-time metric. In principle, we should use the analysis error covariance matrix to define the metric, but we don't know it well enough.

In addition, we didn't know about localization in the early 1990s...

Discussion

This has been my selection of a few of the significant developments (and failures) over the past 20 years.

What would be on your list?

What will be on the list 20 years from now?