

# Variational Quality Control

**Elias Holm**

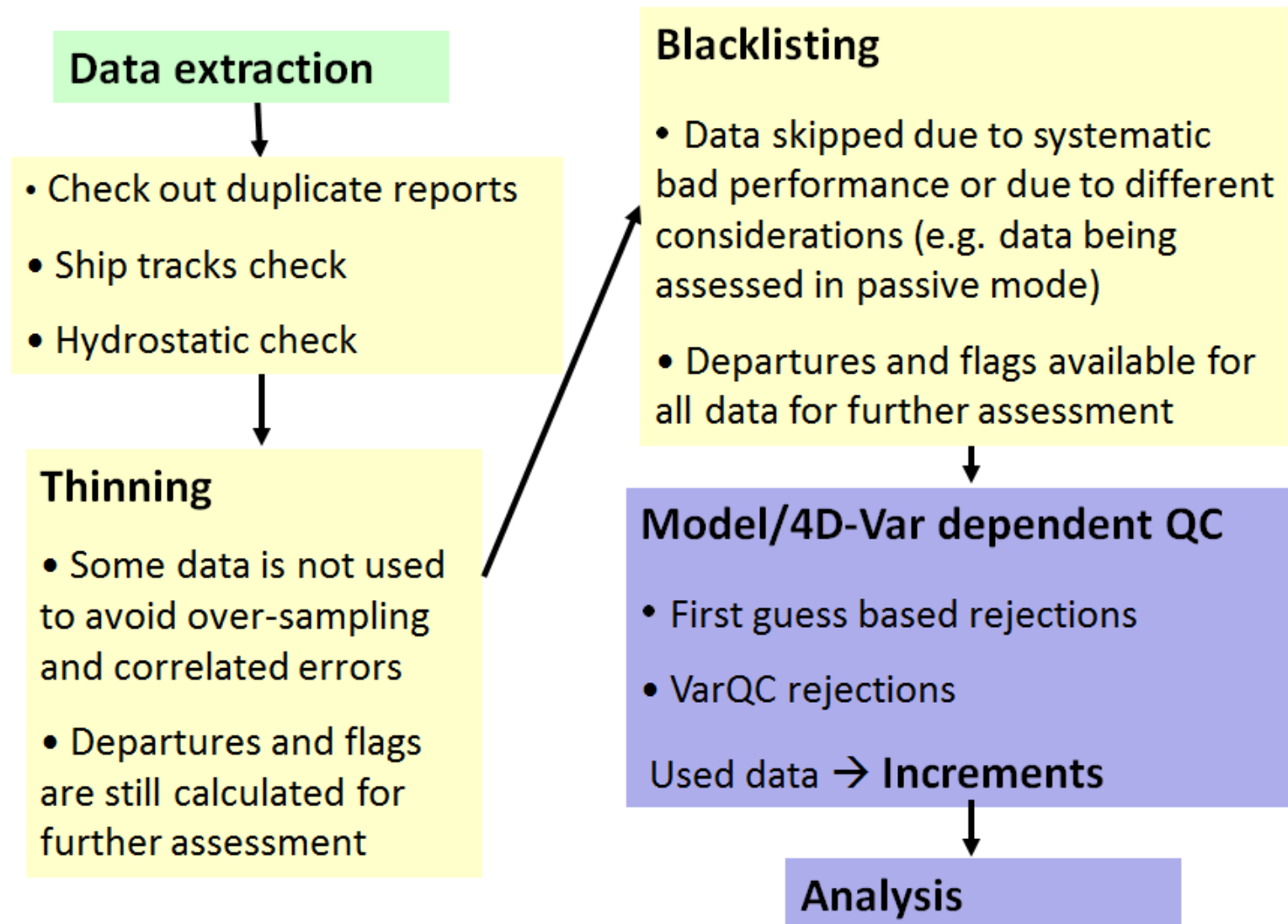
**Data Assimilation Section, ECMWF**

**Contributors: Lars Isaksen, Christina Tavolato,  
Erik Andersson and Elias Holm, ECMWF**

# Outline of Lecture

- Introduction
- VarQC formulation 1: Gaussian+constant
- Rejection limits and tuning
- VarQC formulation 2: Huber norm
- Example
- Summary

# Pre-check → Thin → Blacklist → FG-check → VarQC

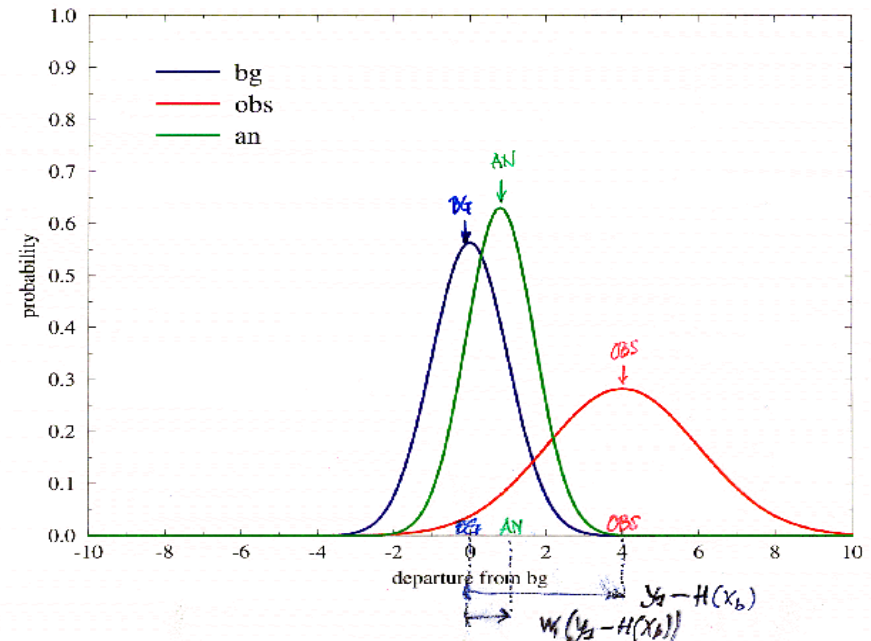


# Weight of observations in the analysis

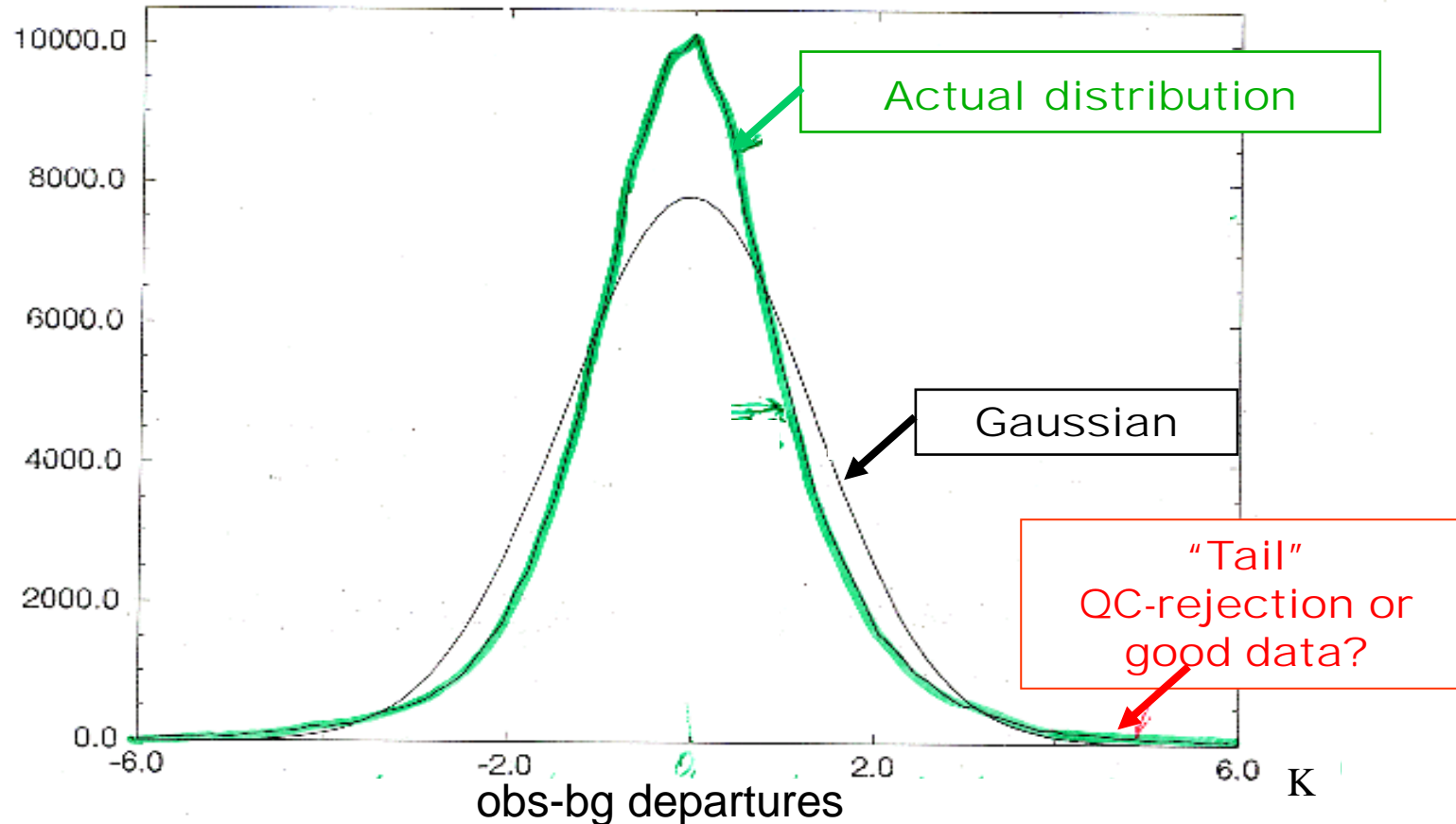
Assuming Gaussian statistics, the maximum likelihood solution to the linear estimation problem results in observation analysis weights ( $w$ ) that are independent of the observed value.

$$x_a - x_b = w(y - Hx_b)$$
$$w = \frac{\sigma_b^2}{\sigma_o^2 + \sigma_b^2}$$

Outliers will be given the same weight as good data, potentially corrupting the analysis



# Even good-quality data show significant deviations from the pure Gaussian form



- The real data distribution has fatter tails than the Gaussian
- Aircraft temperature observations shown here

## Observation cost function $J_o$ (1)

The general expression for the observation cost function is based on the probability density function (the pdf) of the observation error distribution (see Lorenc 1986):

$$J_o = -\ln p + \text{const}$$

$p$  = probability density function of observation error

Constant chosen such that  $J_o=0$  when  $y=Hx$

## Observation cost function $J_0$ (2)

When for  $p$  we assume the normal (Gaussian) distribution ( $N$ ):

$$N = \frac{1}{\sigma_o \sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{y - Hx}{\sigma_o} \right)^2 \right]$$

we obtain the expression

$$J_0^N = -\ln N + \text{const} = \frac{1}{2} \left( \frac{y - Hx}{\sigma_o} \right)^2$$

**y:** observation  
**x:** represents the model/analysis variables  
**H:** observation operators  
 **$\sigma_o$ :** observation error standard deviation

**Normalized departure**

In VarQC a **non-Gaussian** pdf will be used,  
resulting in a **non-quadratic** expression for  $J_0$ .

# Accounting for non-Gaussian effects in Jo

In an attempt to better describe the tails of the observed distributions, Ingleby and Lorenc (1993) suggested a modified pdf (probability density function), written as a sum of two distinct distributions:

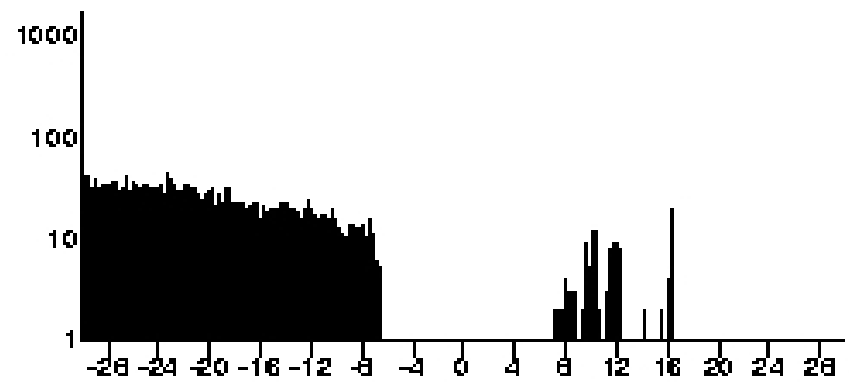
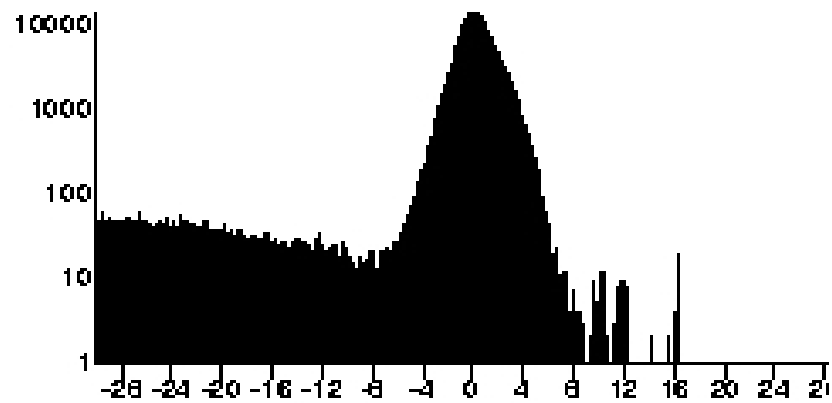
$$p^{QC} = (1 - A)N + Ap^G$$

Normal distribution (pdf),  
as appropriate for  
'good' data

pdf for data affected by  
gross errors

A is the prior probability of gross error





## Gross error pdf as flat distribution

Thus, a pdf for the data affected by gross errors ( $p^G$ ) needs to be specified. Several different forms could be considered.

In the ECMWF 1998-2009 implementation (Andersson and Järvinen 1999, QJRMS) a flat distribution was chosen.

$$p^G = \frac{1}{2d}$$

$2d$  is the width of the distribution

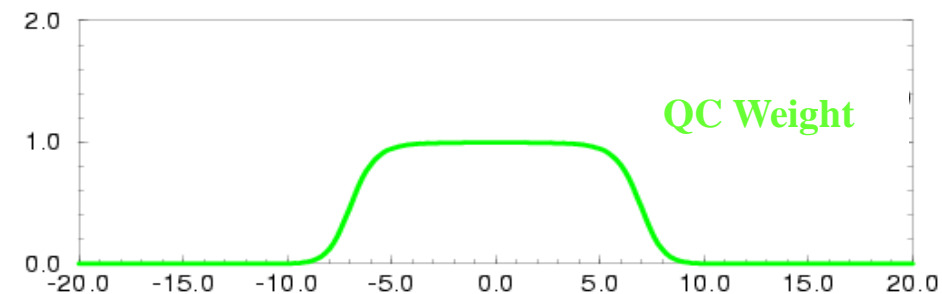
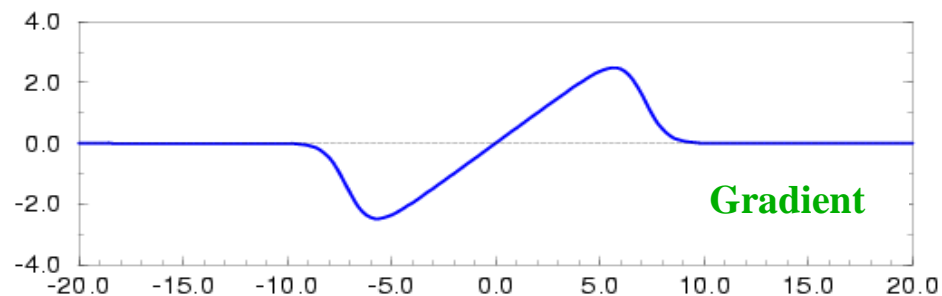
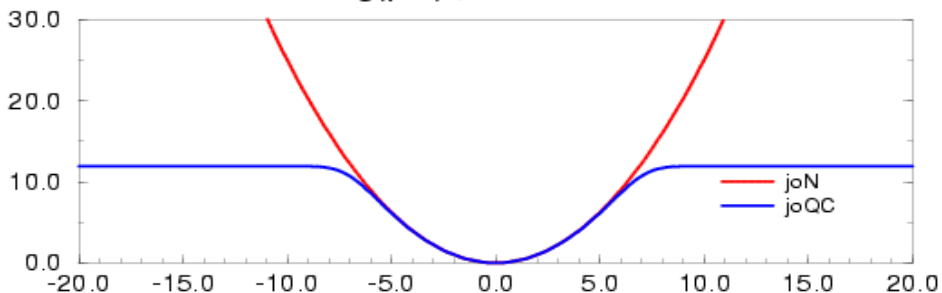
The consequence of this choice will become clear in the following

# Gaussian + flat PDF

# Sum of 2 Gaussians

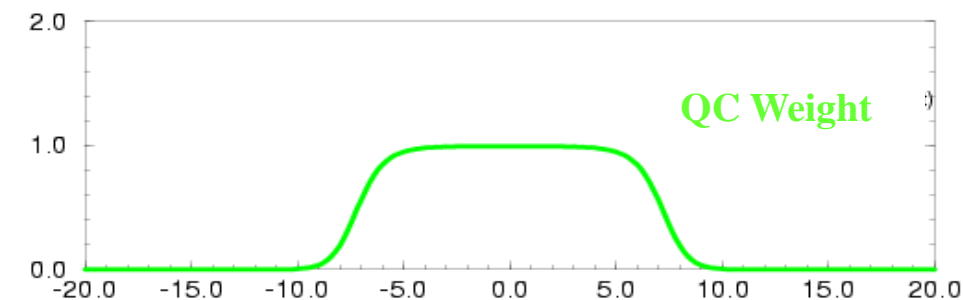
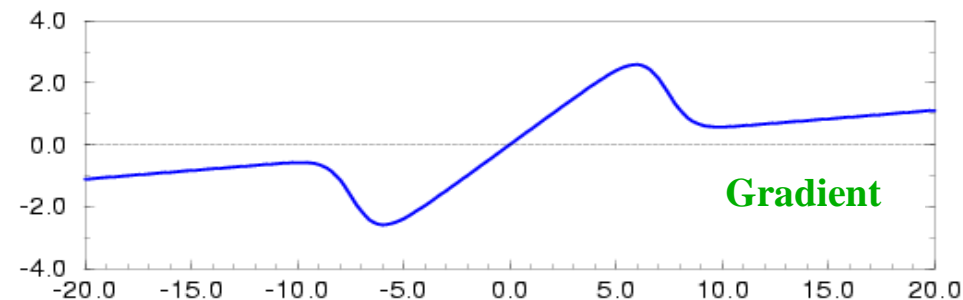
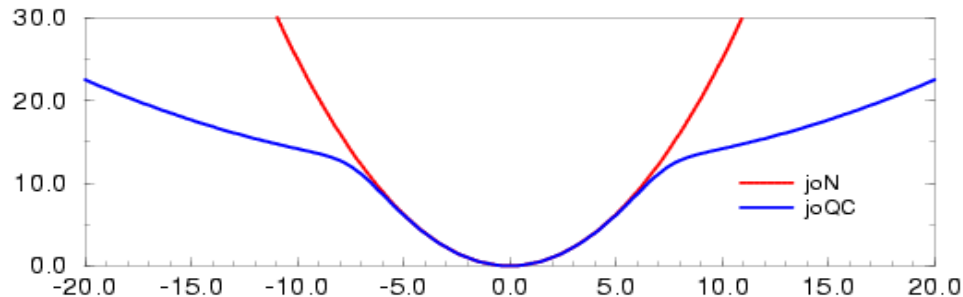
**VarQC: pdf=(1-A)\*N(0,so) + A/(2L\*so)**

*Jo=-log(pdf) ; A=1% L=5 so=2.*



**VarQC: pdf=(1-A)\*N(0,so) + A\*N(0,3\*so)**

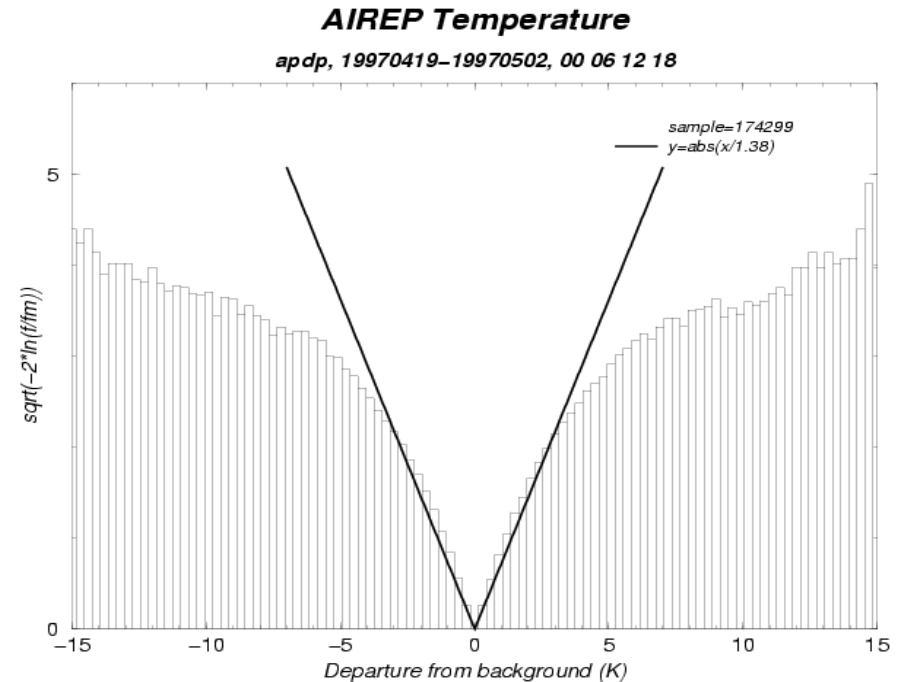
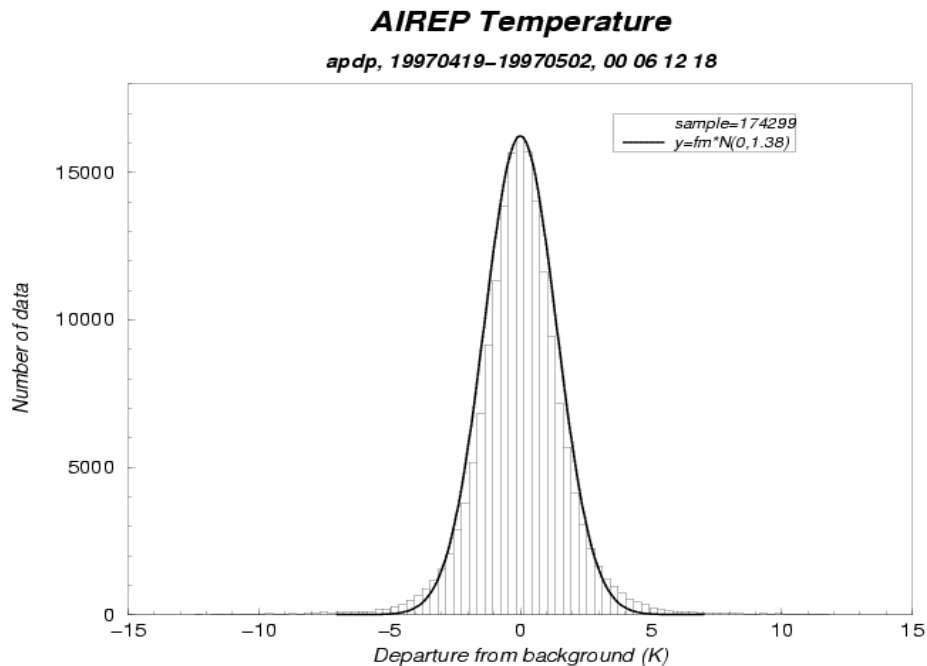
*Jo=-log(pdf) ; A=1% L=5 so=2.*



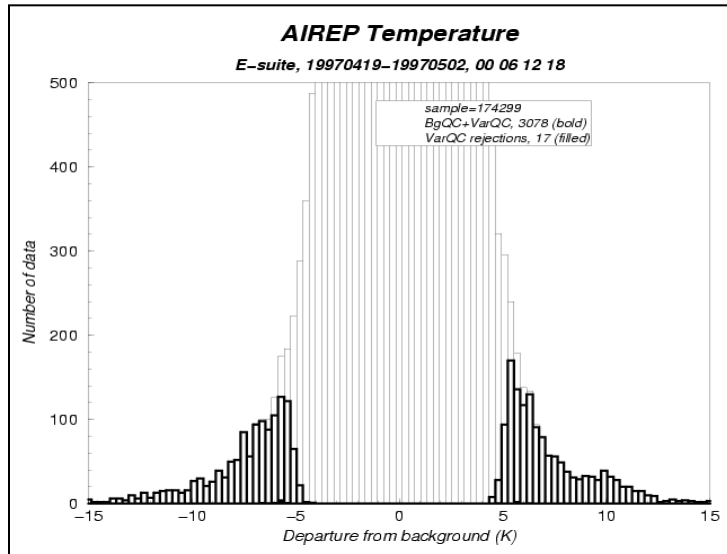
# Tuning the rejection limit

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.



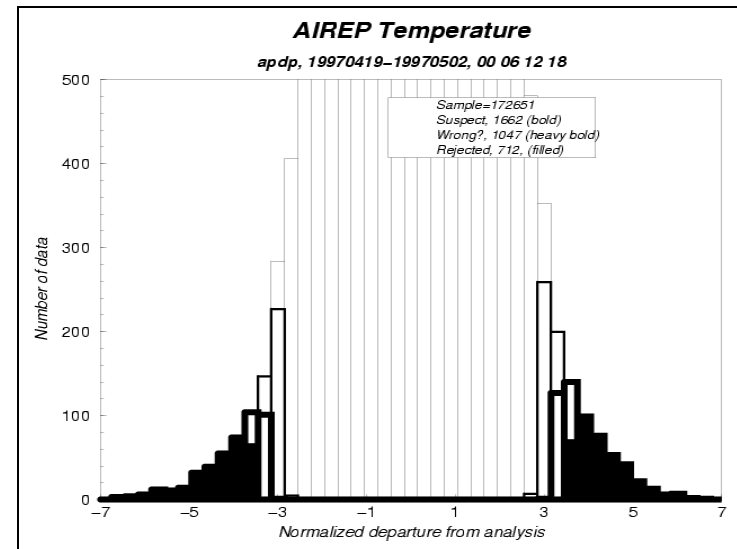
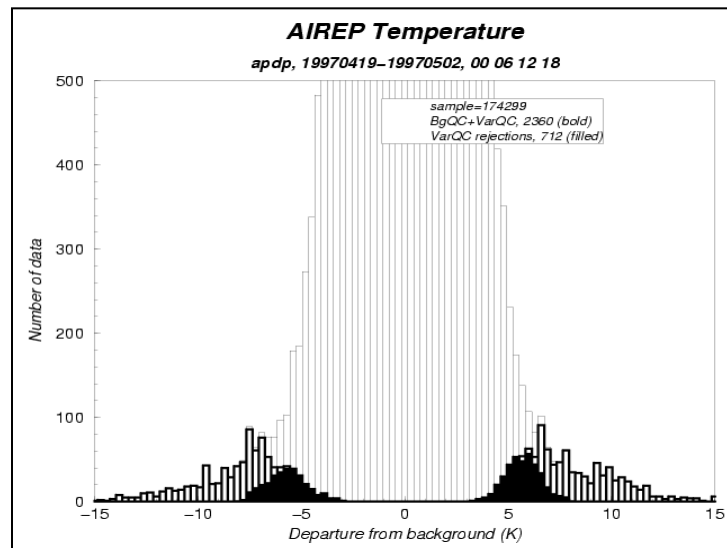
# Tuning example



BgQC too tough

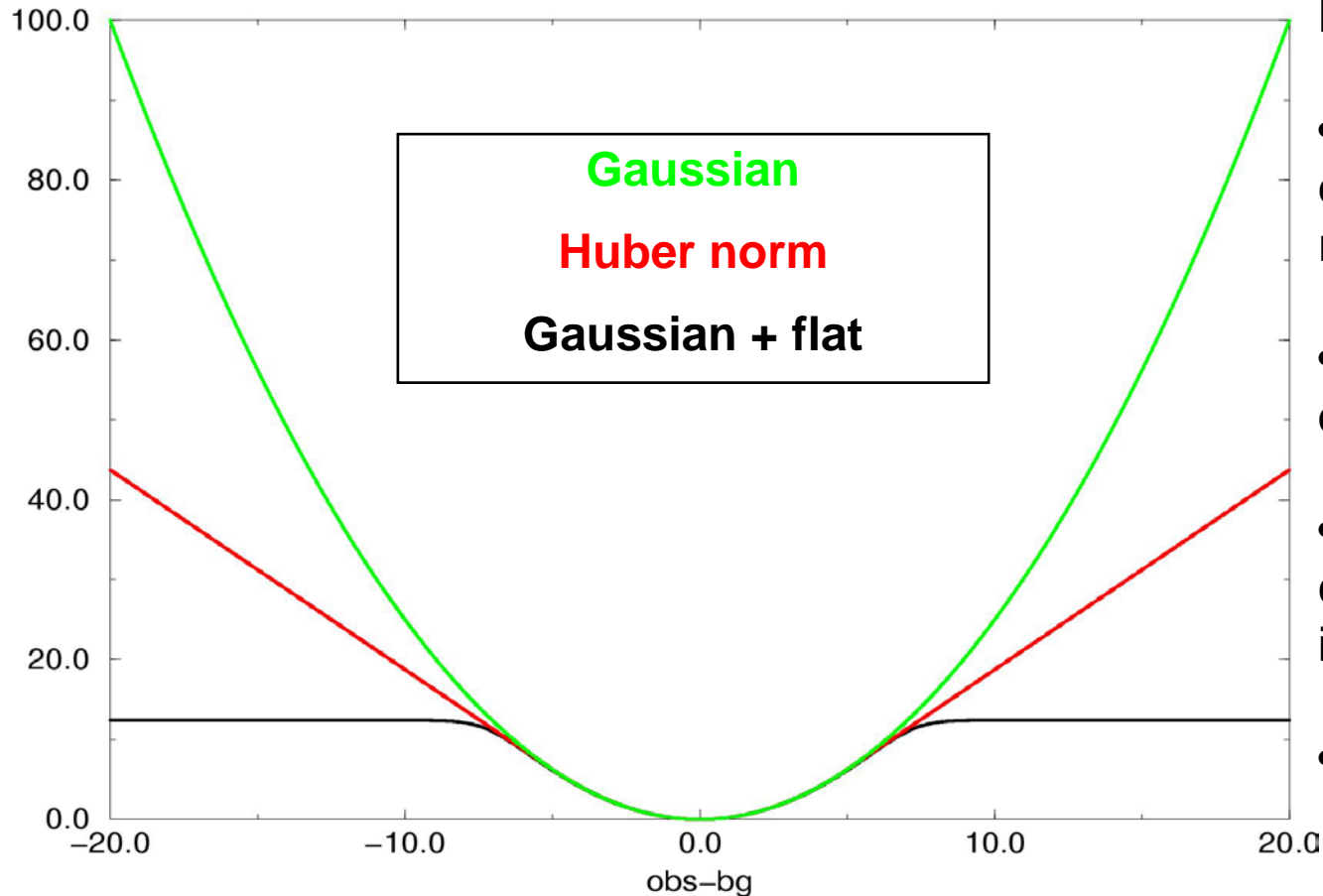
BgQC and VarQC correctly tuned

The shading reflects the value of P, the probability of gross error



# Huber-norm as alternative for non-Gaussian Jo

## A compromise between the $l_2$ and $l_1$ norms



Huber norm:

- Robust method: a few erroneous observations does not ruin analysis
- Adds some weight on observations with large departures
- A set of observations with consistent large departures will influence the analysis
- Concave cost function

# Huber norm variational quality control

The pdf for the Huber norm is:

$$p(y|x) = \begin{cases} \frac{1}{\sigma_o \sqrt{2\pi}} \exp\left(\frac{a^2}{2} - |a\delta|\right) & \text{if } a < \delta \\ \frac{1}{\sigma_o \sqrt{2\pi}} \exp\left[-\frac{1}{2}\delta^2\right] & a \leq \delta \leq b \\ \frac{1}{\sigma_o \sqrt{2\pi}} \exp\left(\frac{b^2}{2} - |b\delta|\right) & \text{if } \delta > b \end{cases} \quad \text{where } \delta = \frac{y - H(x)}{\sigma_o}$$

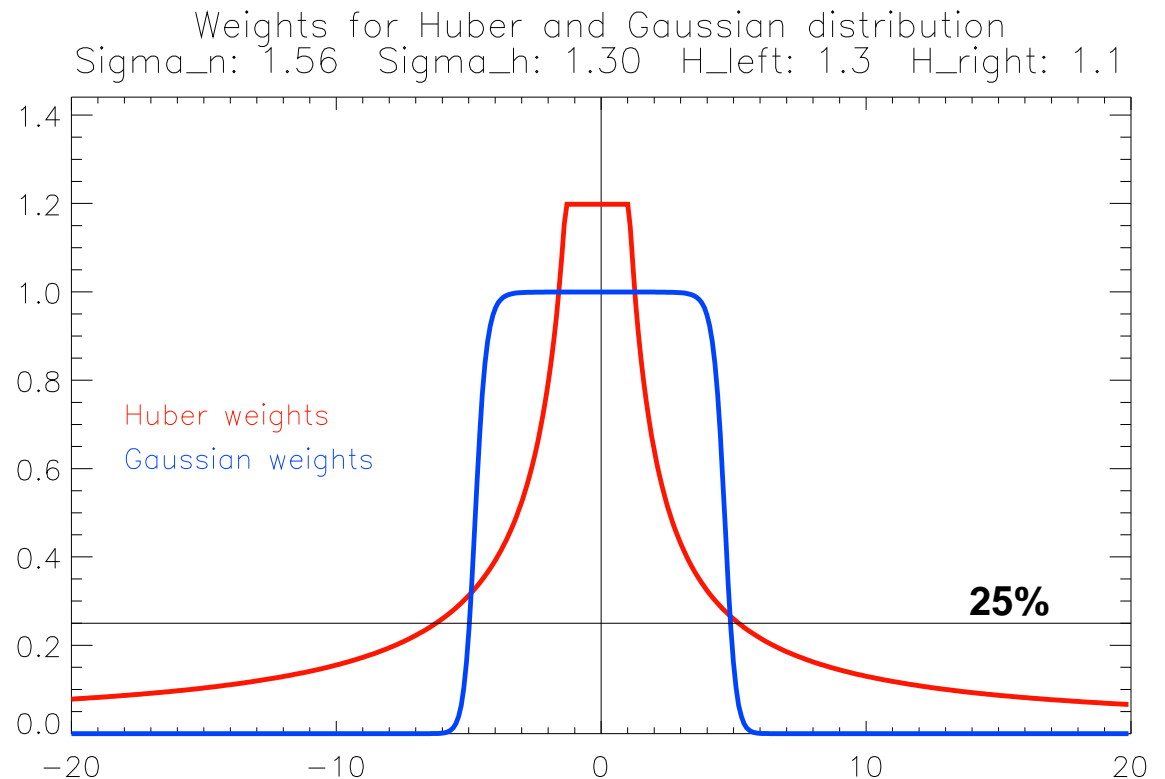
Equivalent to  $L_1$  metric far from  $x$ ,  $L_2$  metric close to  $x$ .

With this pdf, observations far from  $x$  are given less weight than observations close to  $x$ , but can still influence the analysis.

Many observations have errors that are well described by the Huber norm.

# Comparing observation weights: Huber-norm (red) versus Gaussian+flat (blue)

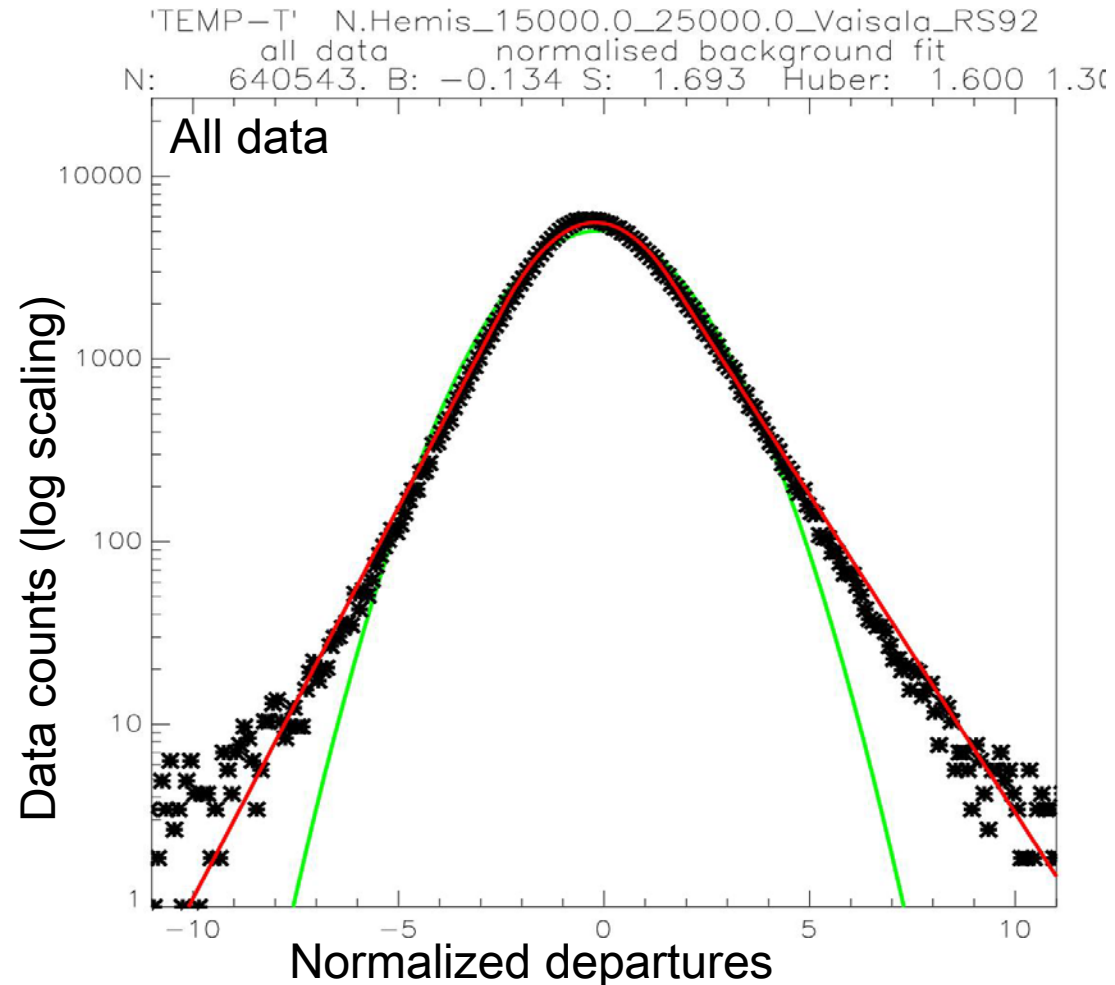
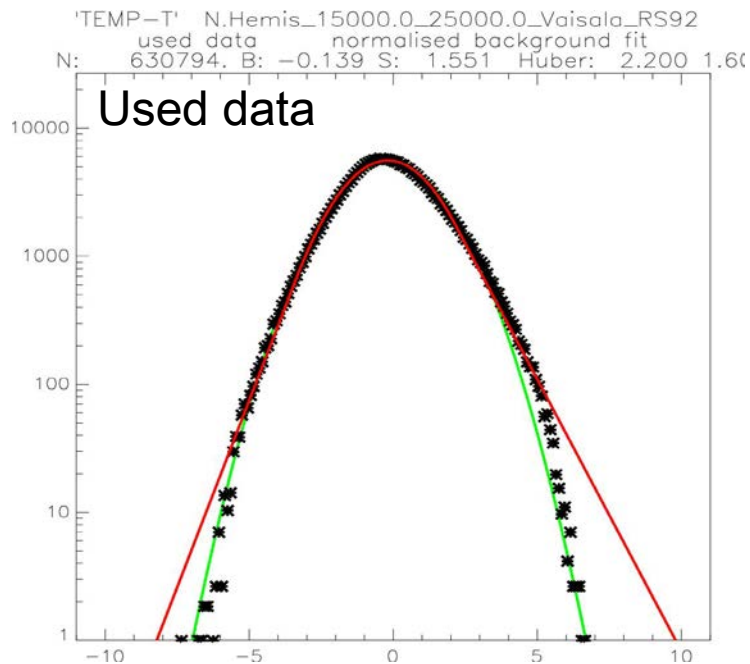
- More weight in the middle of the distribution
- More weight on the edges of the distribution
- More influence of data with large departures
  - Weights: 0 – 25%





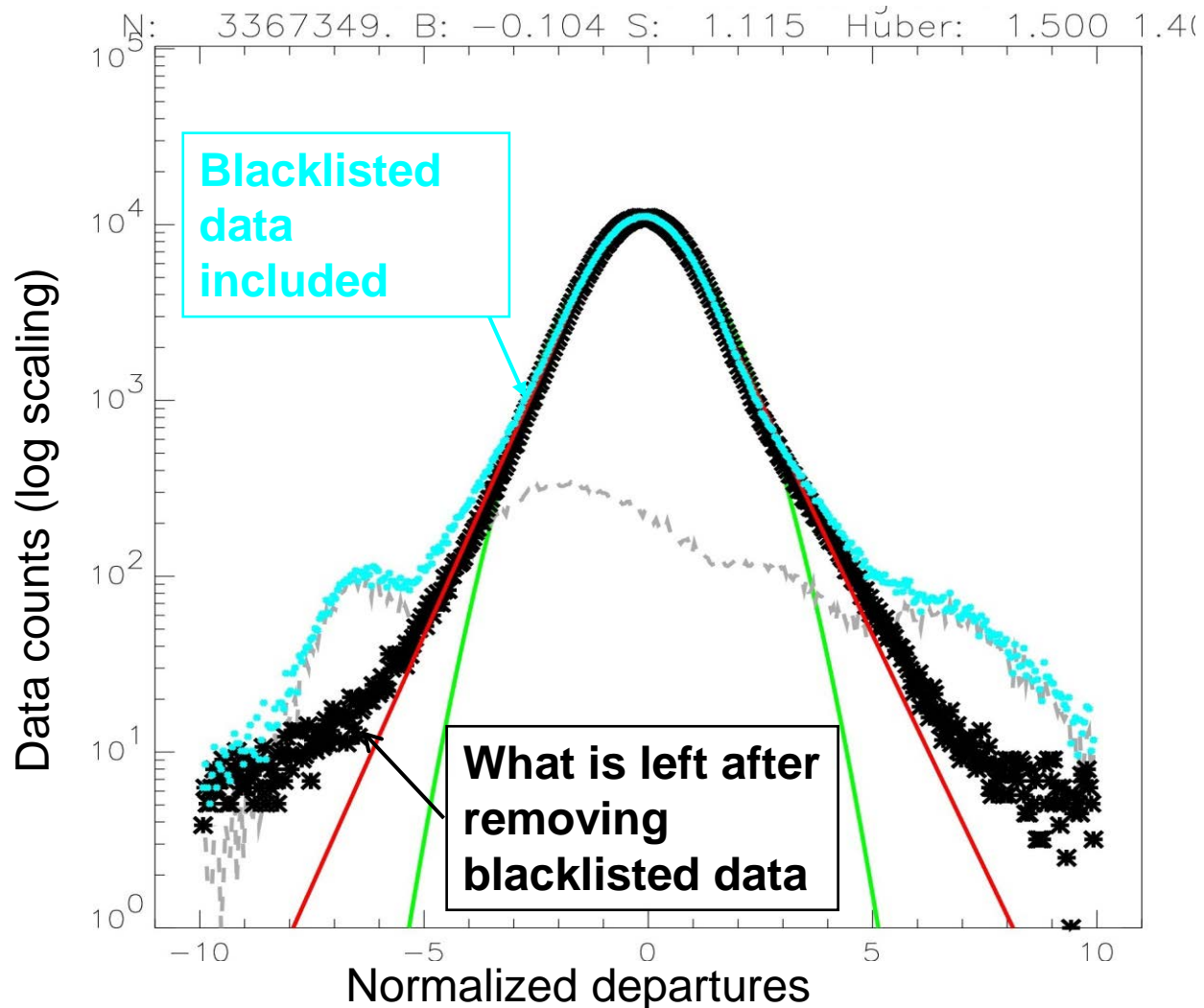
# Departure statistics for radiosonde temperatures is well described by a Huber-norm distribution

- Based on 18 months of data  
Feb 2006 – Sep 2007
- Normalised fit of pdf to data
  - Best Gaussian fit
  - Best Huber norm fit



# METAR surface pressure data (Tropics)

Blacklisting data is sometimes enough to limit gross errors



After removing the blacklisted data the departures (black crosses) are well described by a Huber norm (red line)

# 27 Dec 1999 – French storm 18UTC

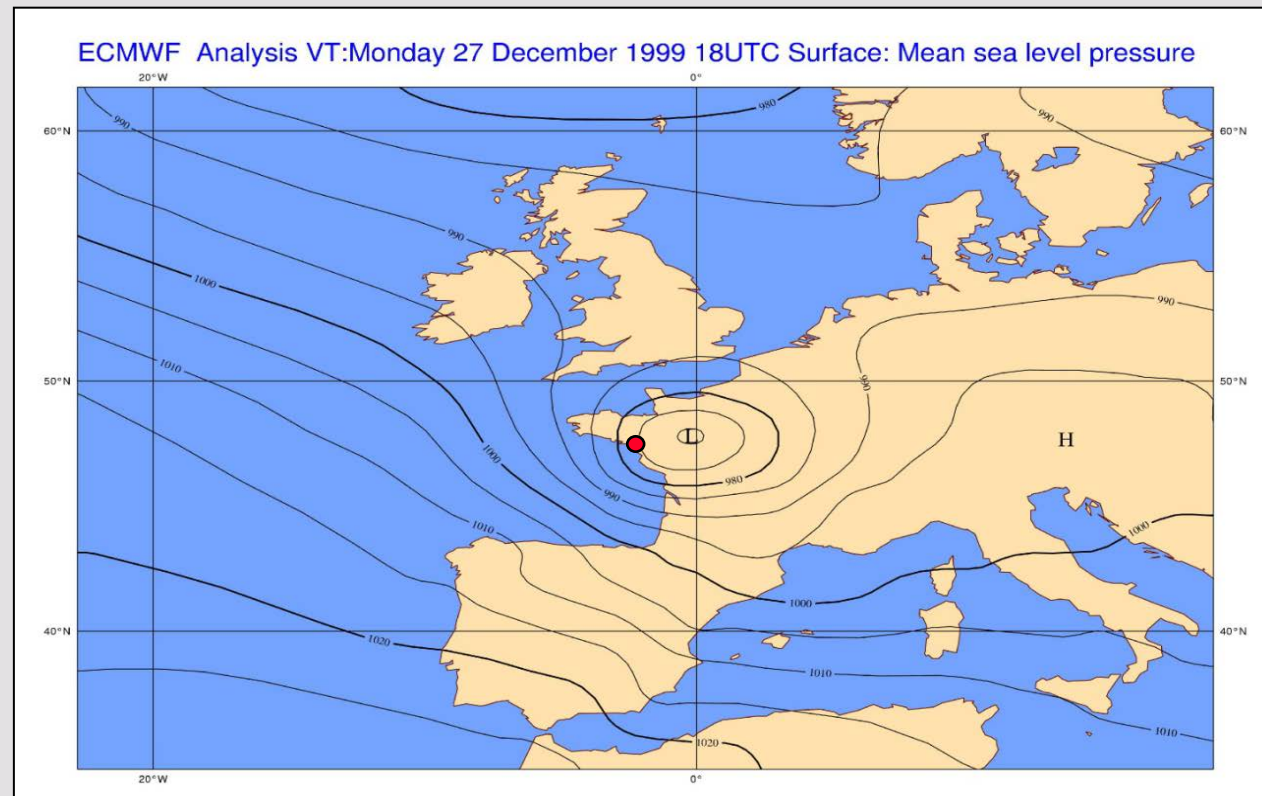
- Era interim analysis produced a low with min 970 hPa
- Lowest pressure observation (SYNOP: red circle)

963.5 hPa (supported by neighbouring stations)

At this station the analysis shows 977 hPa

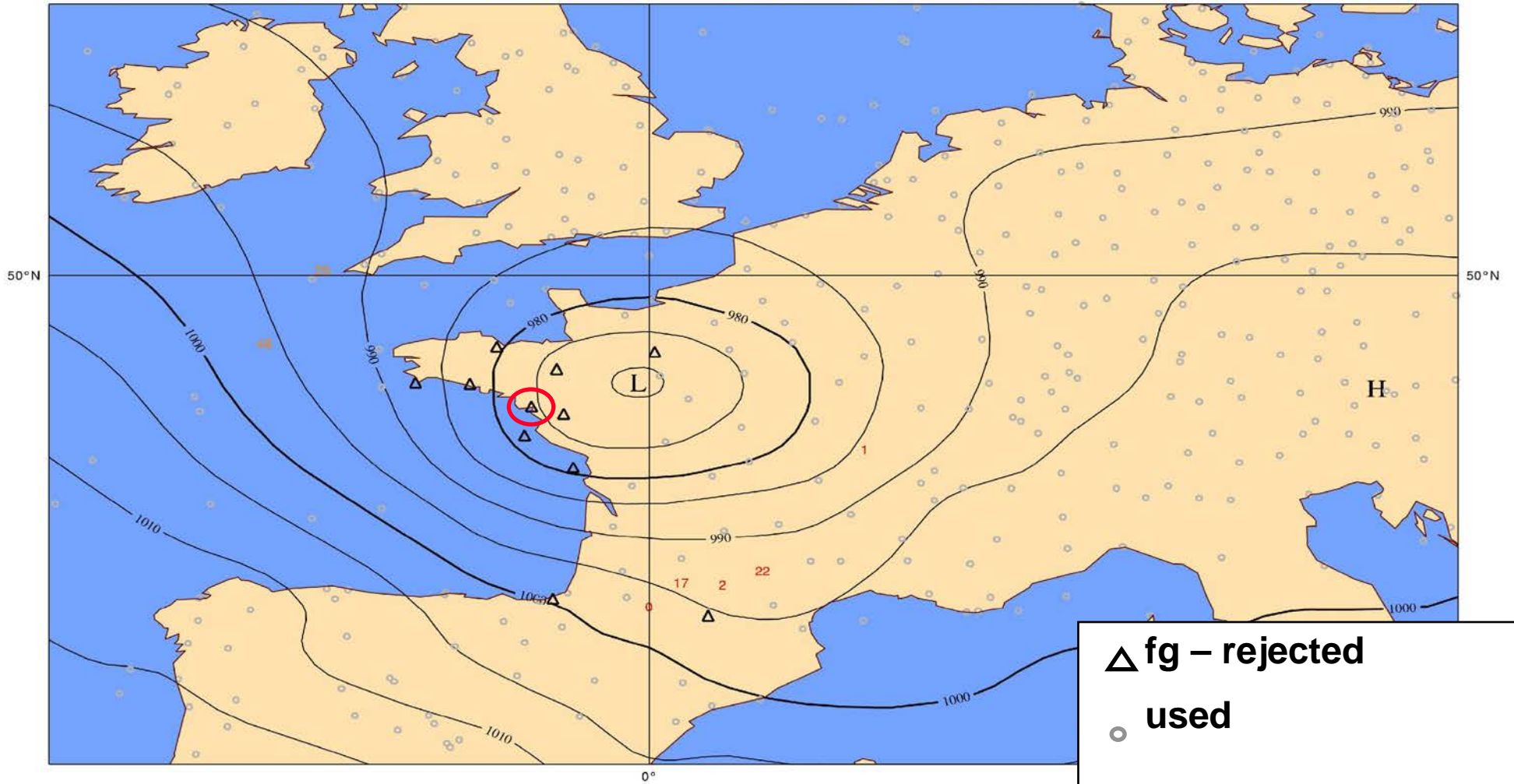
Analysis wrong by 16.5 hPa!

- High density of good quality surface data for this case



# Data rejection and VarQC weights – Era interim reanalysis 27.12.99 18UTC +60min

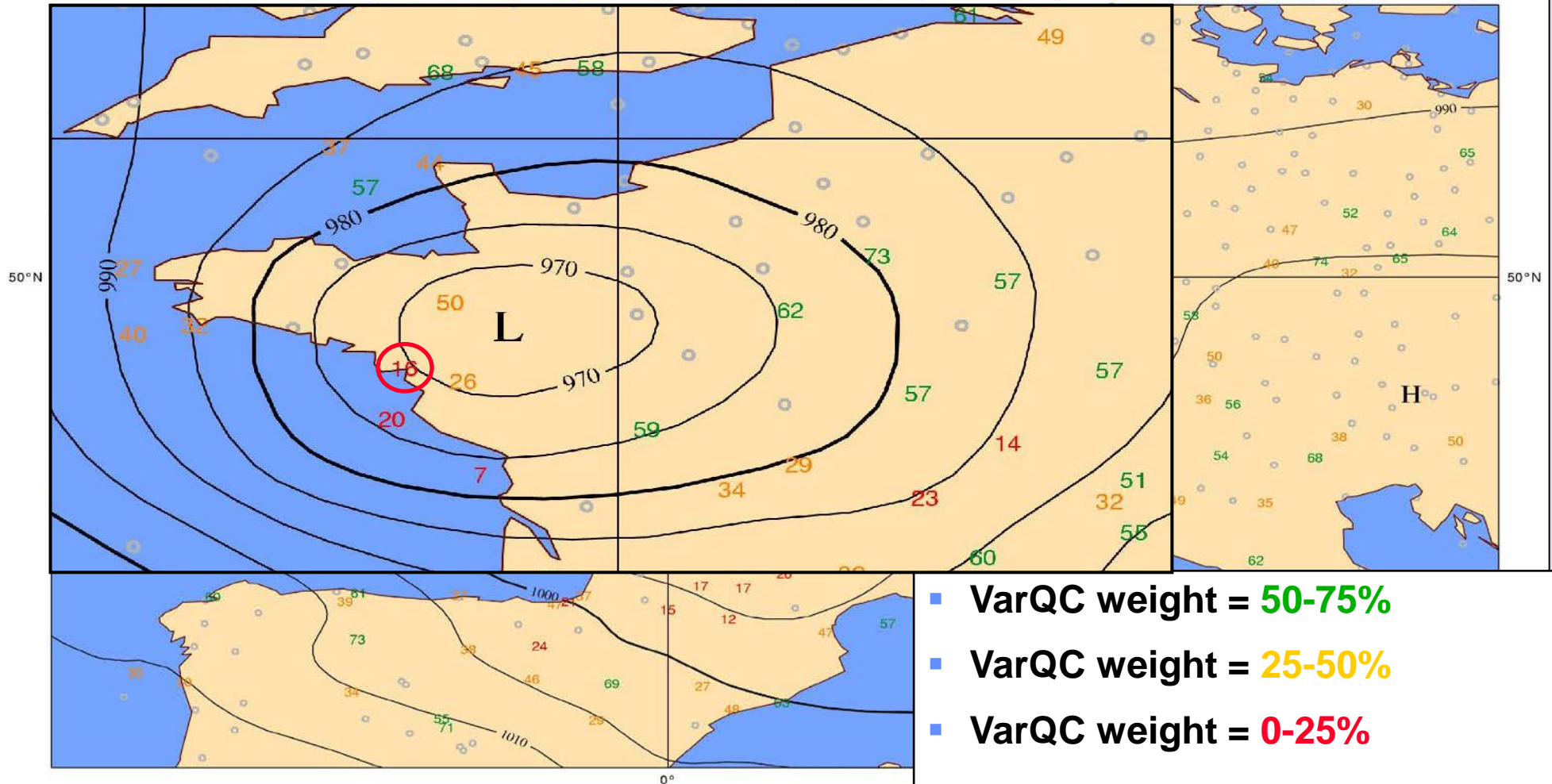
1112: VarQC-rejections: Flag1 (green), Flag2 (orange), Flag3 (red), MSL analysis (black)



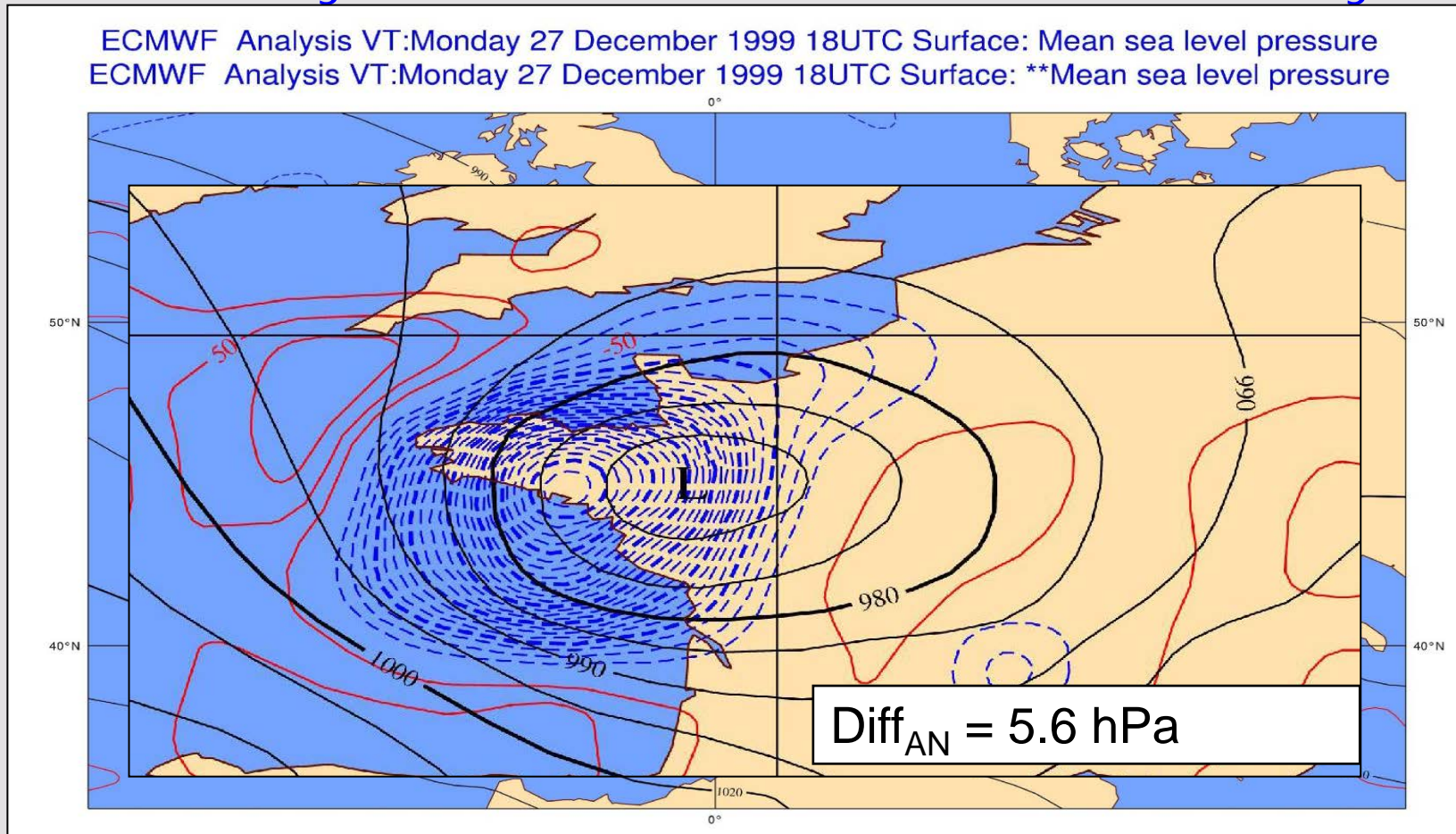


# Data rejection and VarQC weights – Huber exp.

1362: VarQC-rejections: Flag1 (green), Flag2 (orange), Flag3 (red), MSL analysis (black)



# MSL Analysis differences: Huber v. Reanalysis



- New min 968 hPa
- Low correctly shifted towards west and intensified in better agreement with surface pressure observations

# VarQC general summary

- VarQC provides a satisfactory and very efficient quality control mechanism - consistent with 3D/4D-Var.
- The implementation can be very straight forward (multiply observation departures by a factor).
- VarQC does not replace the pre-analysis checks - the checks against the background for example. **However, with Huber-norm these are relaxed significantly.**
- All observational data from all data types are quality controlled simultaneously, as part of the general 3D/4D-Var minimisation.

A good description of background errors is essential for effective, flow-dependent QC: background error lecture.