

Variational Quality Control

Elias Holm

Data Assimilation Section, ECMWF

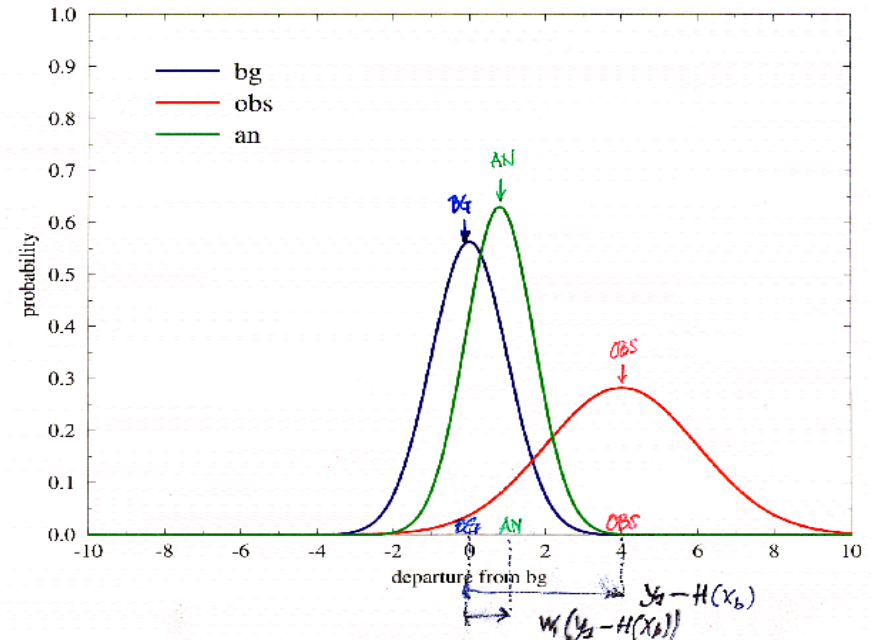
Acknowledgements to Lars Isaksen, Christina Tavolato and
Erik Andersson, ECMWF

Introduction

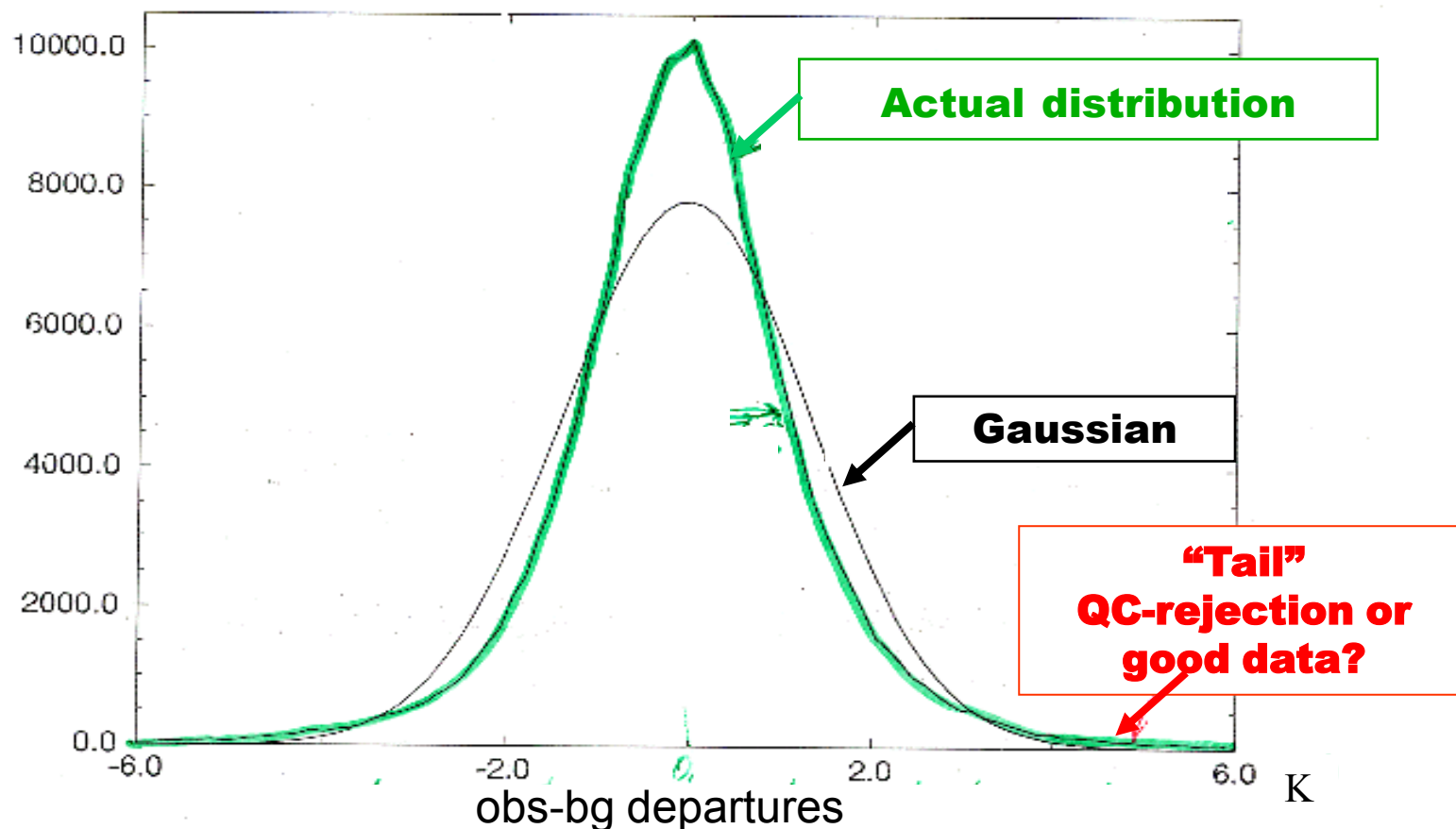
Assuming Gaussian statistics, the maximum likelihood solution to the linear estimation problem results in observation analysis weights (w) that are independent of the observed value.

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

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

The Normal 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 is the probability density function of observation error

arbitrary constant, chosen such that $J_o=0$ when $y=Hx$

The Normal 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
 σ_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 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

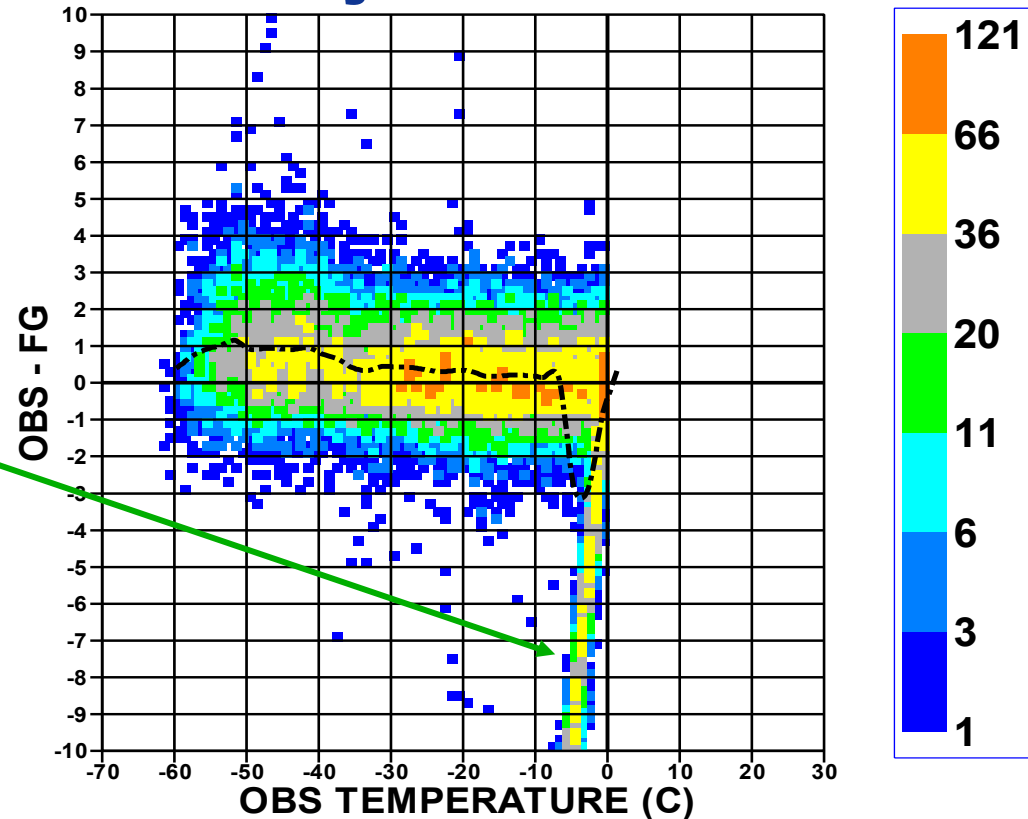
pdf for data affected by
gross errors

A is the prior probability of gross error

Gross errors of that type occur occasionally...

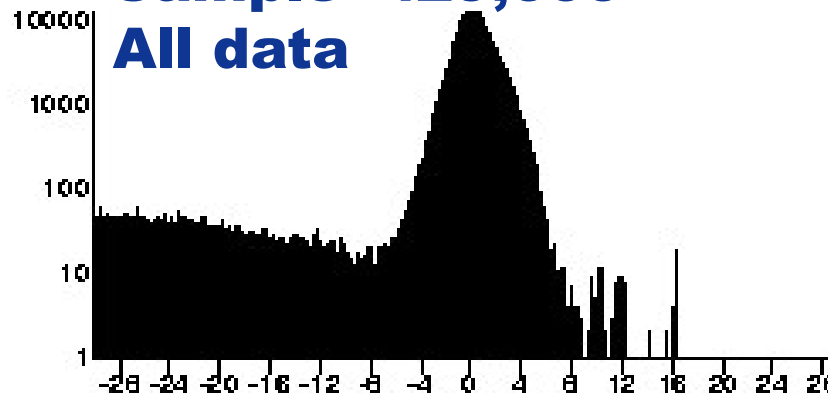
Positive observed temperatures ($^{\circ}\text{C}$) reported with wrong sign.

(Chinese aircraft data 1-21 May 2007)

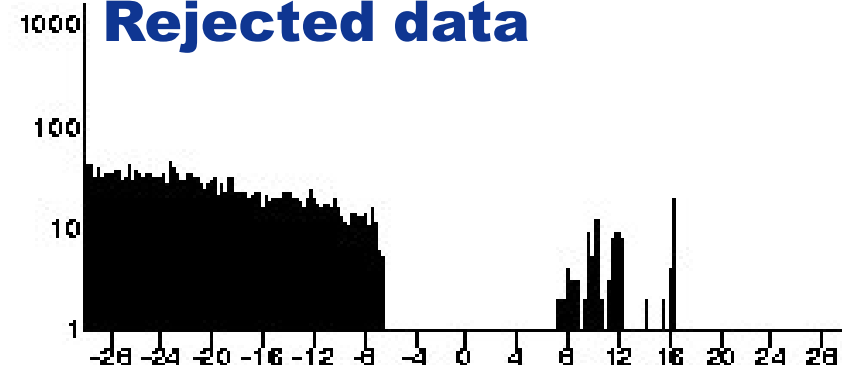


Innovation Statistics Sample=429,000

All data




Rejected data



Variational quality control

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

VarQC formulation

Inserting p^{QC} for p in the expression $J_o = -\ln p + \text{const}$, we obtain:

$$J_o^{\text{QC}} = -\ln \left[\frac{\gamma + \exp(-J_o^{\text{N}})}{\gamma + 1} \right]$$
$$\nabla J_o^{\text{QC}} = \nabla J_o^{\text{N}} \left[1 - \frac{\gamma}{\gamma + \exp(-J_o^{\text{N}})} \right]$$

with γ defined as : $\gamma = \frac{A\sqrt{2\pi}}{(1-A)2d}$

We can see how the presence of γ modifies the normal cost function and its gradient

Probability of gross error

The term modifying the gradient (on the previous slide) can be shown to be equal to:

the *a-posteriori* probability of gross error P , given x and assuming that Hx is correct (see Ingleby and Lorenc 1993)

$$P = \frac{\gamma}{\gamma + \exp(-J_o^N)}$$

Furthermore, we can define a VarQC weight W :

It is the factor by which the gradient's magnitude is reduced.

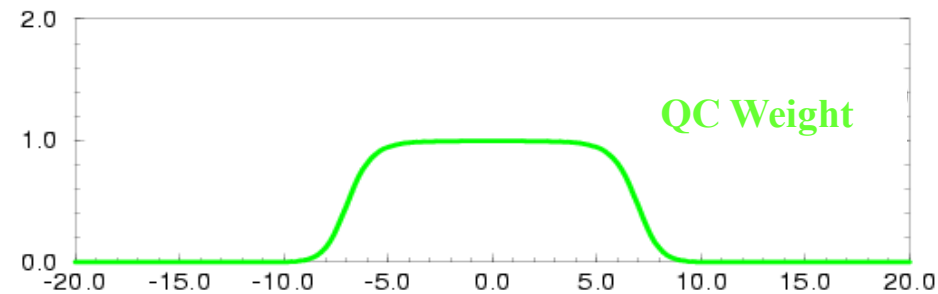
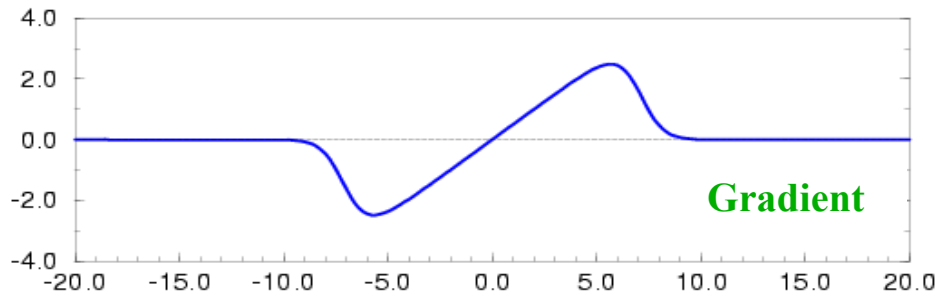
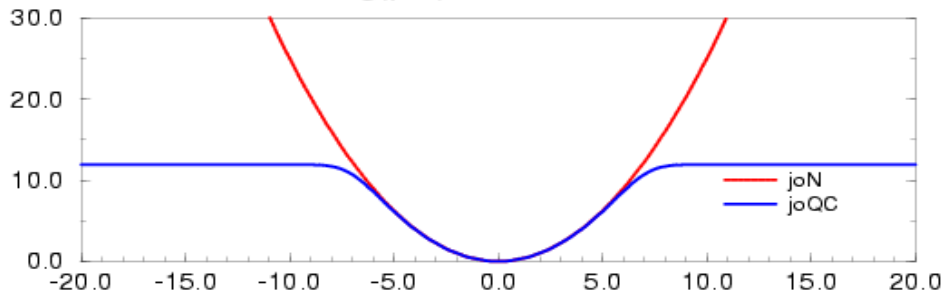
$$W = 1 - P$$

$$\nabla J_o^{QC} = W \nabla J_o^N$$

- Data which are found likely to be incorrect ($P \approx 1$) are given reduced weight in the analysis.
- Data which are found likely to be correct ($P \approx 0$) are given the weight they would have had using purely Gaussian observation error pdf.

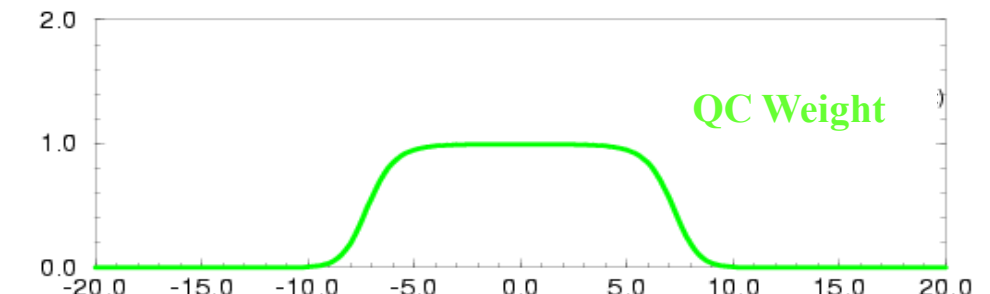
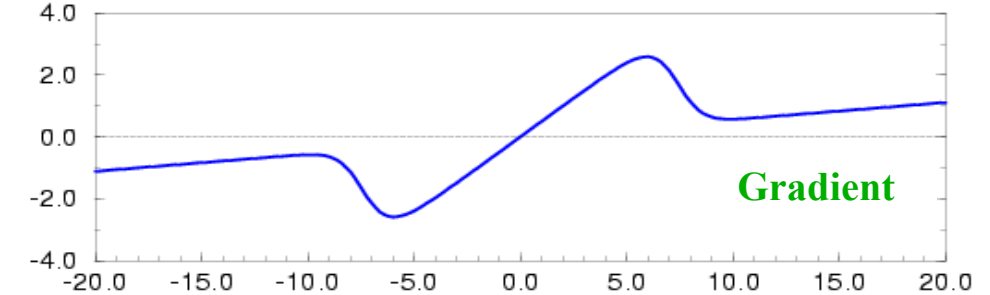
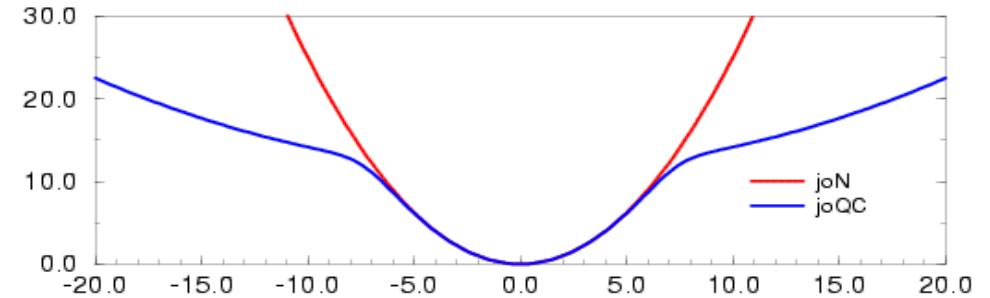
Gaussian + flat PDF

VarQC: pdf=(1-A)*N(0,so) + A/(2L*so)
Jo=-log(pdf) ; A=1% L=5 so=2.



Sum of 2 Gaussians

VarQC: pdf=(1-A)*N(0,so) + A*N(0,3*so)
Jo=-log(pdf) ; A=1% L=5 so=2.



Application

In the case of many observations, all with uncorrelated errors, J_o^{QC} is computed as a sum (over the observations i) of independent cost function contributions:

$$J_o^{\text{QC}} = -\ln \prod_i p_i^{\text{QC}} + \text{const} = -\sum_i \ln p_i^{\text{QC}} + \text{const} = \sum_i J_{oi}^{\text{QC}}$$

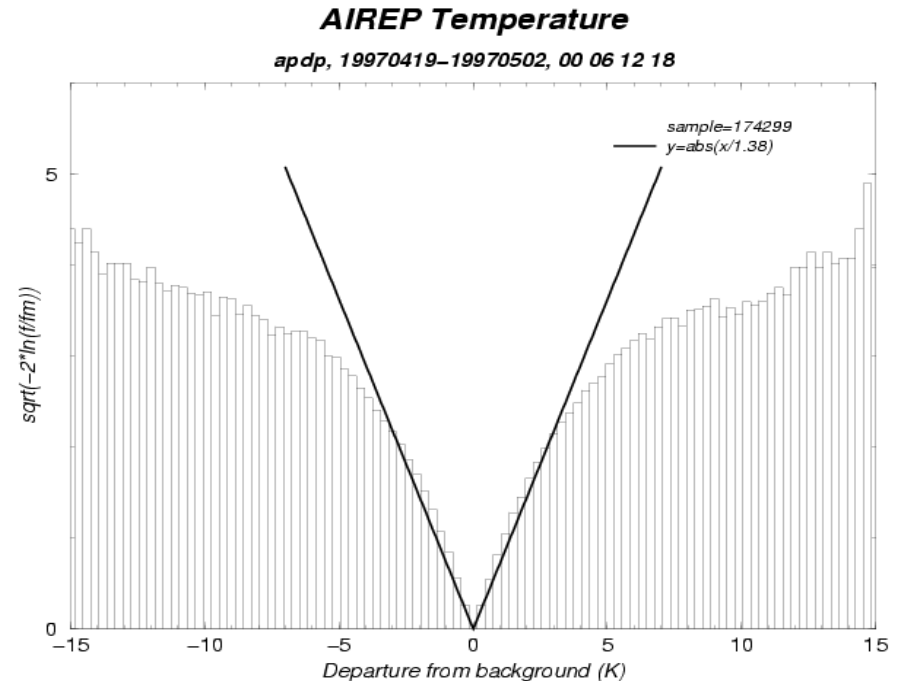
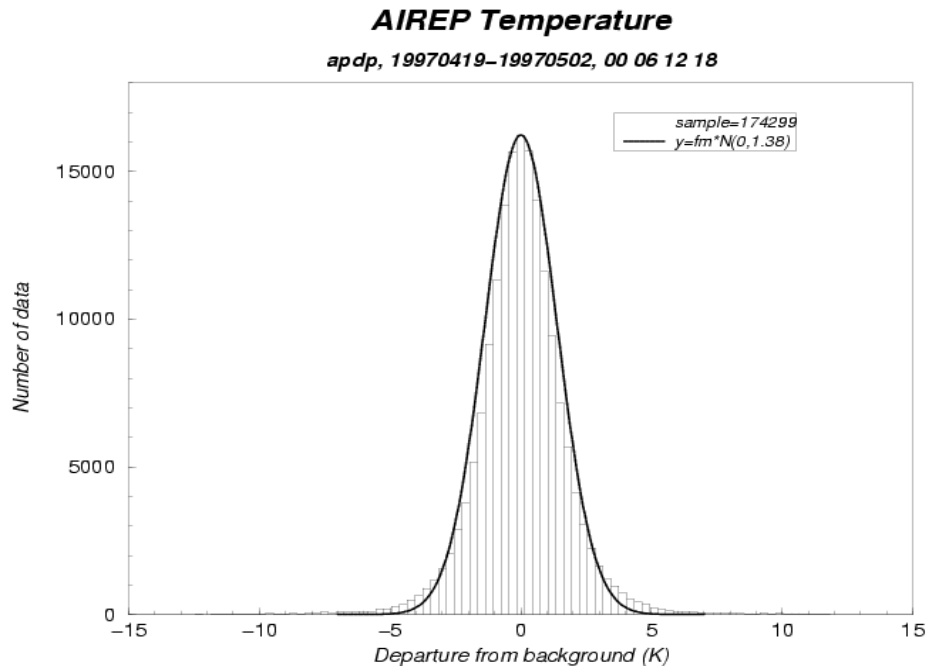
The global set of observational data includes a variety of observed quantities, as used by the variational scheme through their respective observation operators. All are quality controlled together, as part of the main 4D-Var estimation.

The application of VarQC is always in terms of the observed quantity.

Tuning the rejection limit

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

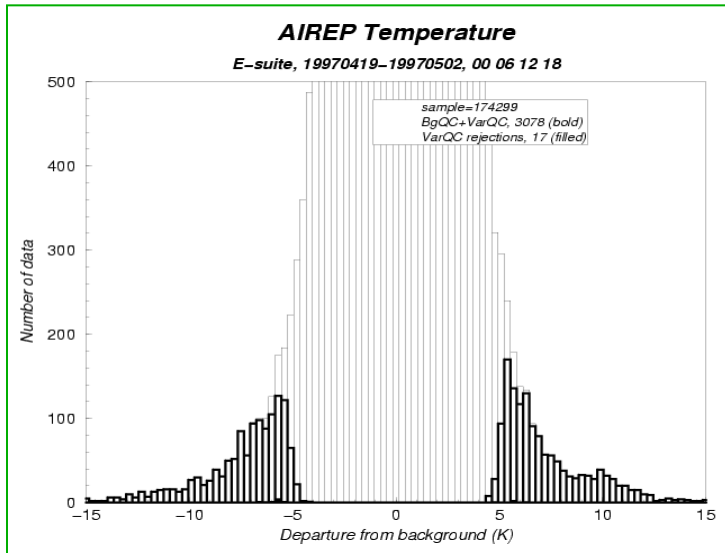


Transforming the Gaussian pdf

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

y: observation
x: represents the model/analysis variables
H: observation operators
 σ_o : observation error standard deviation

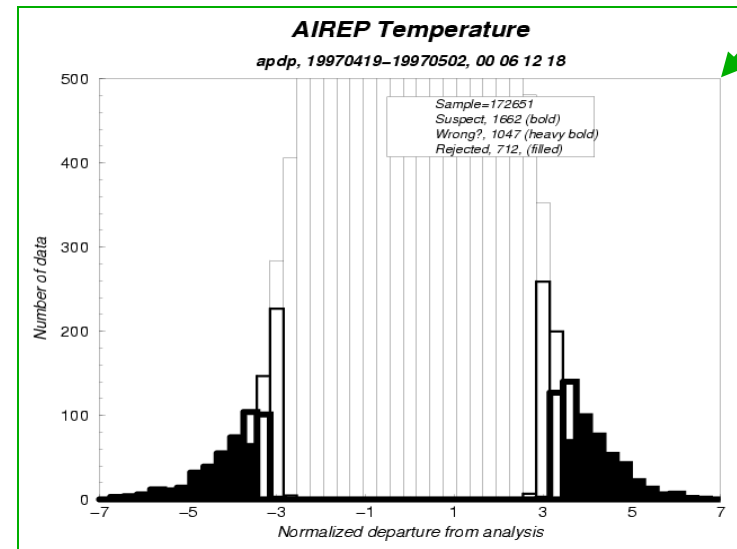
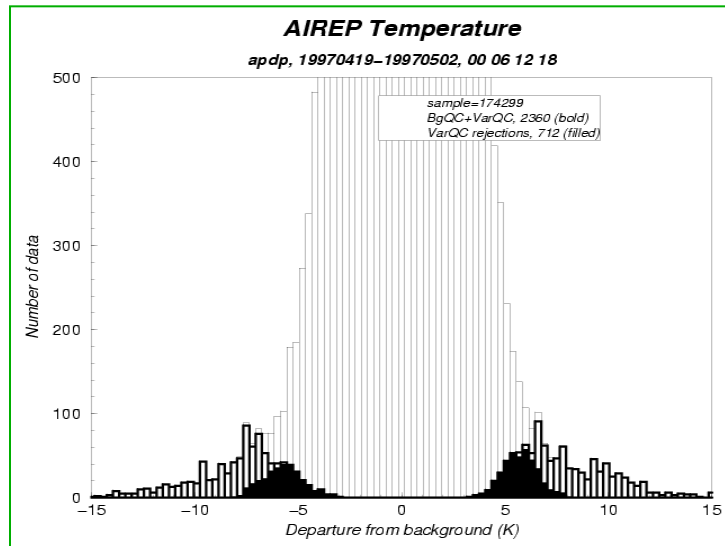
Tuning example



BgQC too tough

BgQC and VarQC correctly tuned

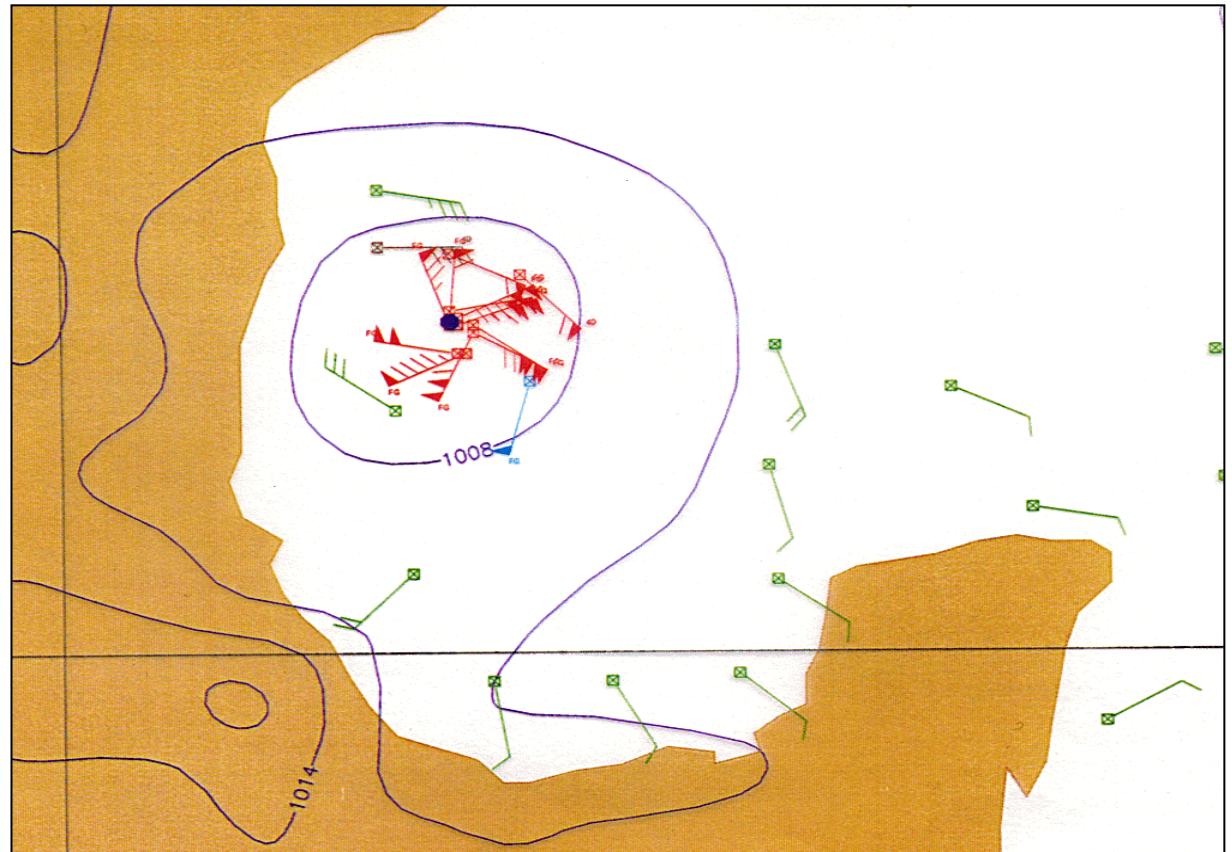
The shading reflects the value of P, the probability of gross error



Tropical Cyclone example

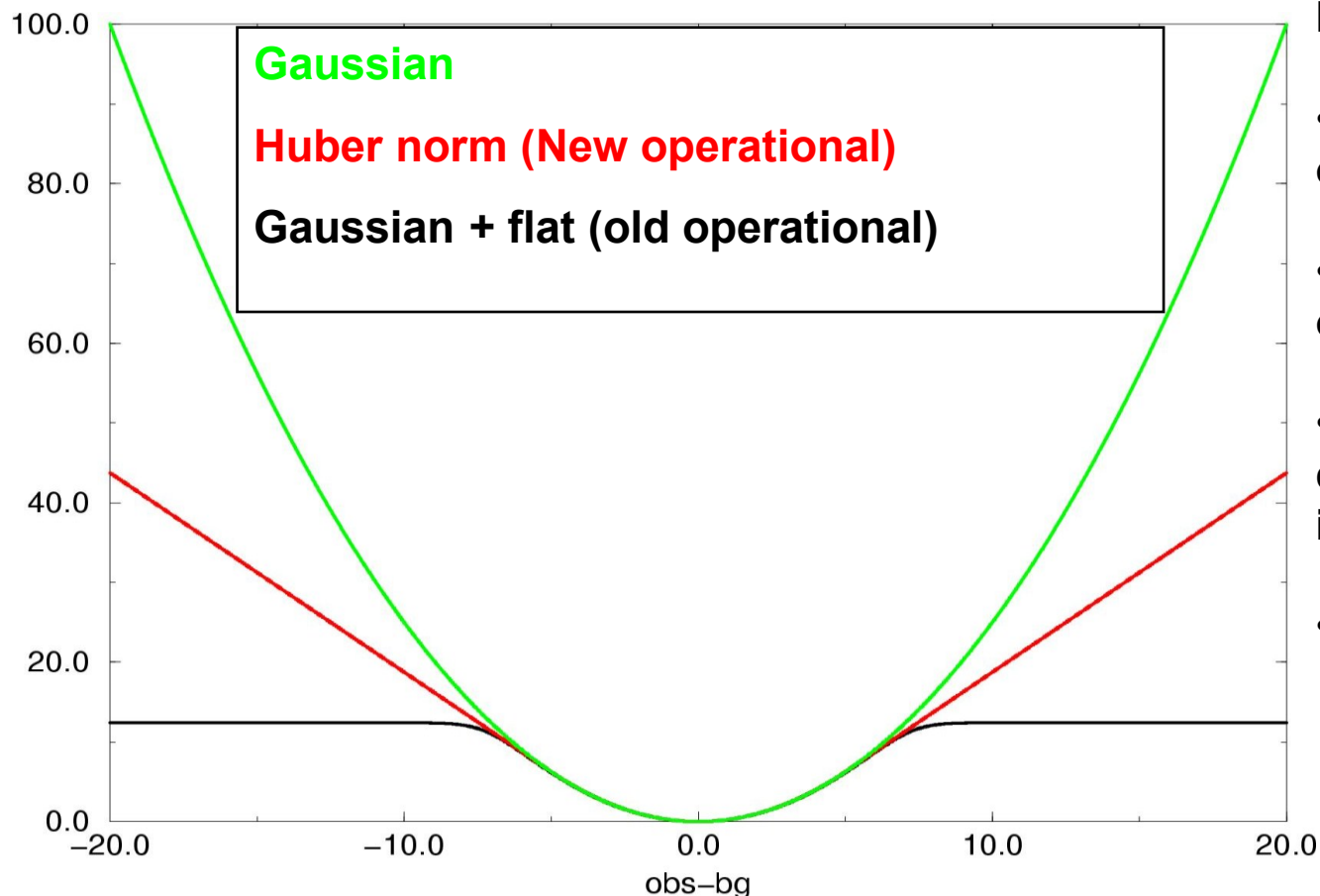
Observations of intense and small-scale features may be rejected although the measurements are correct.

The problem occurs when the resolution of the analysis system (as determined by the B-matrix and model resolution) is insufficient.



Huber-norm an alternative

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.

Transforming the Gaussian and exponential pdf

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

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

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

$$-2 \ln E = -const_1 + \left[2a \left(\frac{|y - Hx|}{\sigma_o} \right) \right]$$

y: observation

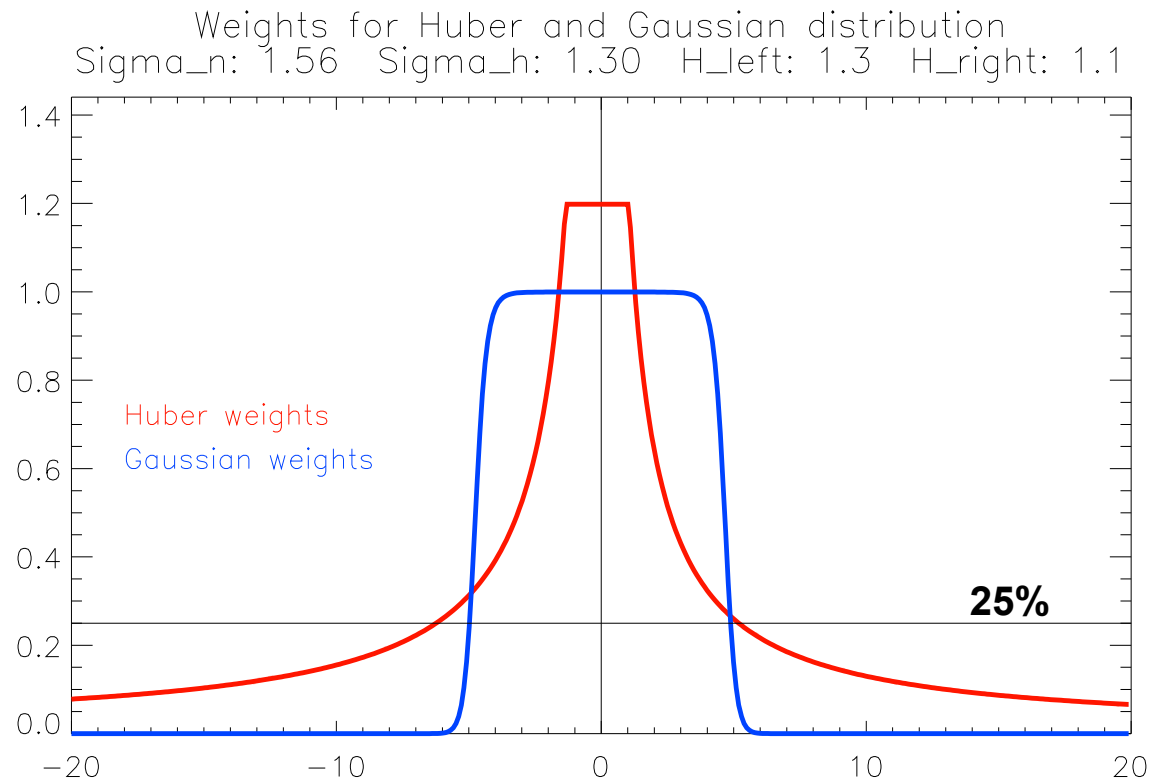
x: represents the model/analysis variables

H: observation operators

σ_o : observation error standard deviation

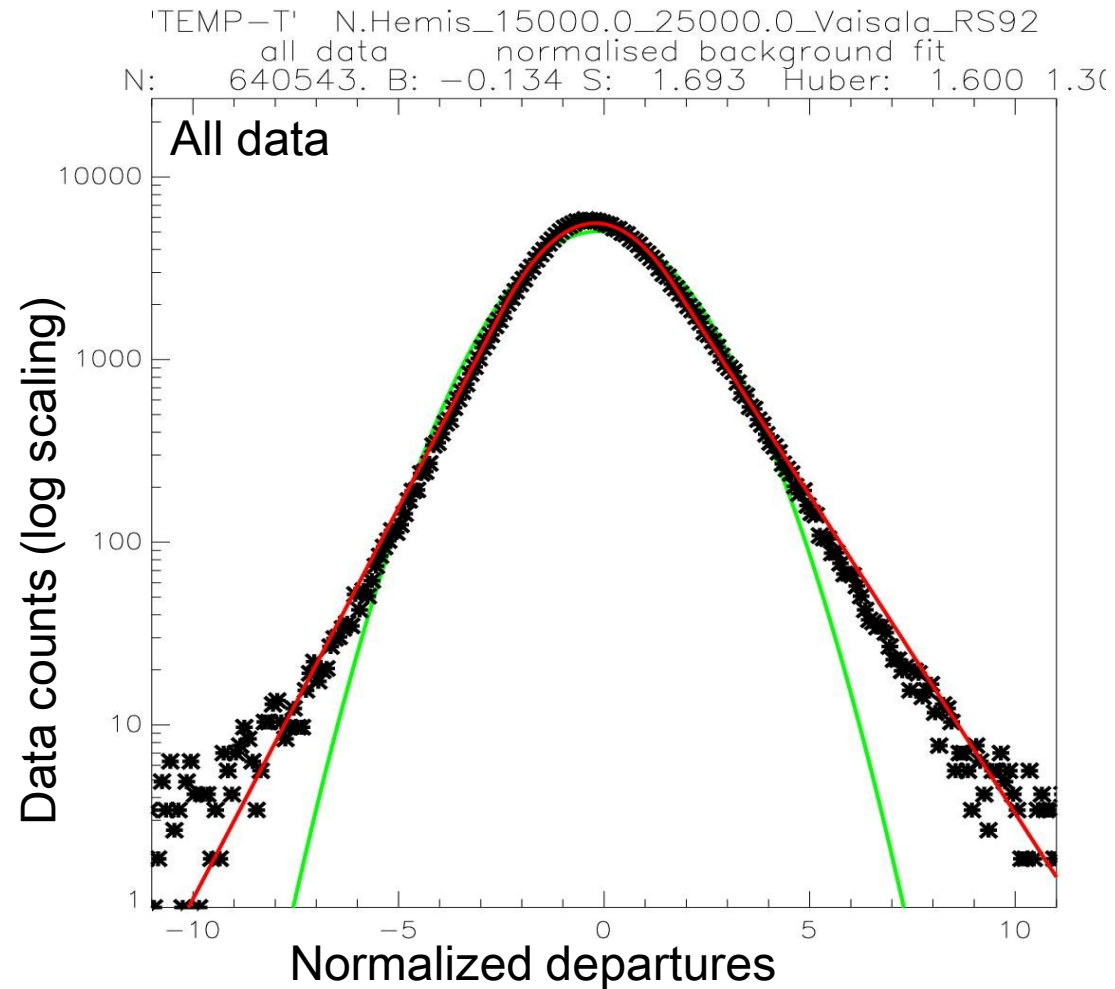
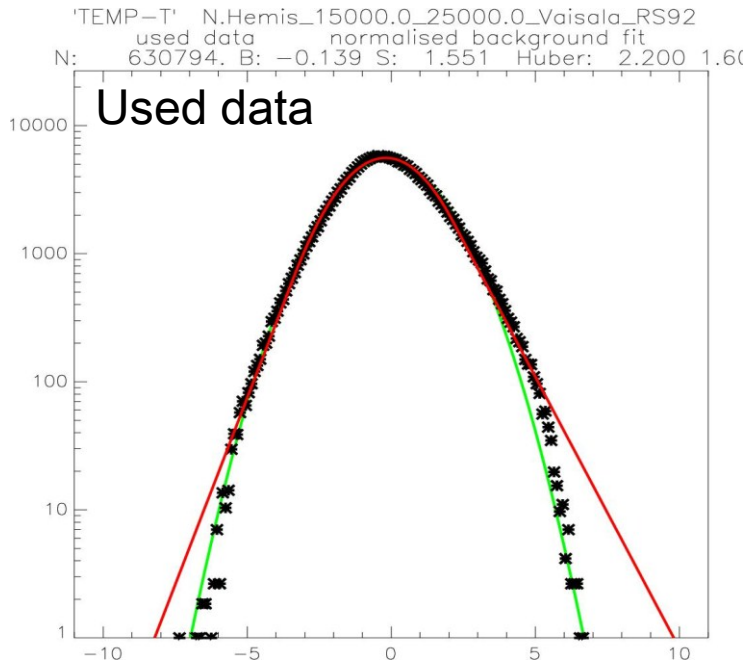
Comparing observation weights: Huber-norm (red) versus Gaussian+flat (blue)

- 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%

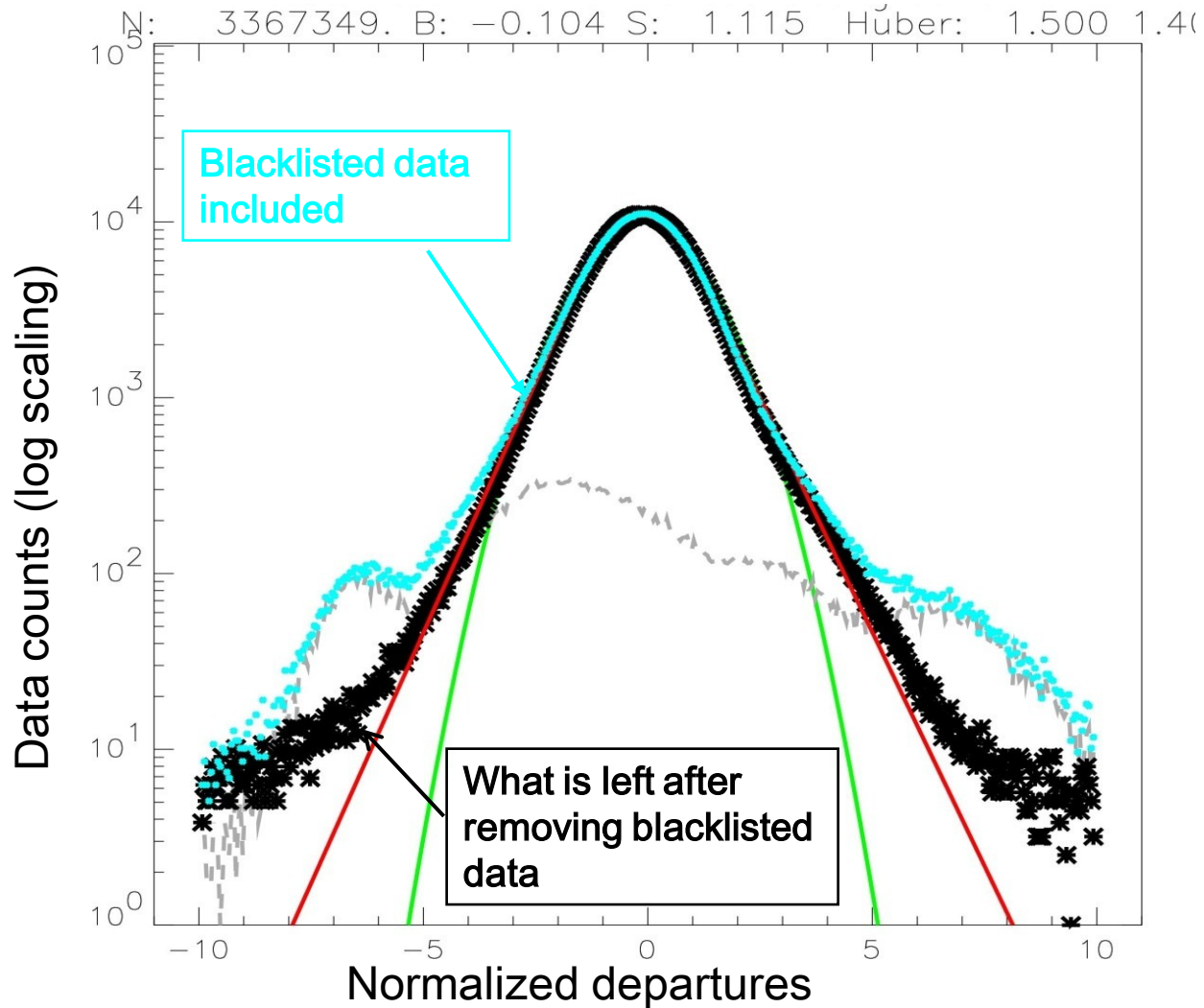


Departure statistics for radiosonde temperatures is well described by a Huber-norm distribution

- 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 may well contain gross errors



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

VarQC Summary

- VarQC is efficient quality control mechanism – all data types are quality controlled simultaneously as part of the 3D/4D-Var minimization.
- The implementation is very straight forward.
- VarQC does not replace the pre-analysis checks - the checks against the background for example. **However, with Huber-norm these are relaxed significantly.**