Implementation of a new nonlinear quality control scheme for the RTMA/URMA at NCEP

R. James Purser[1], Xiujuan Su[1], Manuel Pondeca[1], Steven Levine[2], Jacob Carley[1], Runhua Yang[1], and Geoff DiMego[3]

[1] IMSG at NOAA/NCEP/EMC
[2] SRG at NOAA/NCEP/EMC
[3] NOAA/NCEP/EMC

AMS Annual Meeting, January 6, 2015. IOAS-AOLS Conference.

Topics discussed

- The RTMA/URMA variables needing quality control.
- Risk of multiple optima in the assimilation.
- Generalizing Ob error distribution to a "Gaussian mixture" with both kurtosis and skewness.
- A special choice of mixture guaranteeing single optimum analysis.

Variables and Data Sources

The Real-Time Mesoscale Analysis, and its delayed version, the UnRestricted Mesoscale Analysis (which is able to access more late observations, and can therefore serve as the surface “analysis of record”), produce two dimensional analyses of surface, or near-surface variables.

RTMA and URMA are specialized 2D versions of NCEP’s Gridpoint Statistical Interpolation (GSI).

In addition to the more standard meteorological variables:

2m temperature,
2m specific humidity
10m horizontal wind components
surface pressure

We are also (now, or soon) analyzing variables quantifying:

10m wind gust
visibility
total cloud amount (sky cover analysis)

(Max and min T, ceiling, and significant wave ht are intended to be added in future, as discussed In Jacob Carley's talk 4.3 in this session.)

Quality control is needed for all the observations entering the analysis to ensure that their rare but destructive gross errors do not jeopardize the final result. Data sources include:

METAR
Mesonet
Ships
Buoys
Satwind
GOES Imager (soon)
METOP-B-ASCAT winds (later)

The Quality Control Dilemma; How to resolve it.

The dilemma we always face in dealing with the matter of quality control is deciding how to treat those observations in the analysis which seem to be at greater variance than expected from the background or evolving iterations of the analysis, and yet are not so greatly deviant that they can be rejected without qualm.

An objective resolution of this dilemma is provided by the Bayesian statistical way of thinking about the analysis problem, which derives from the non-Gaussianity of the distribution of the errors in the observations considered.

A reject/accept criterion is replaced by an optimal progressive down-weighting of the observation as Observation-minus-Analysis (O-A) increases.

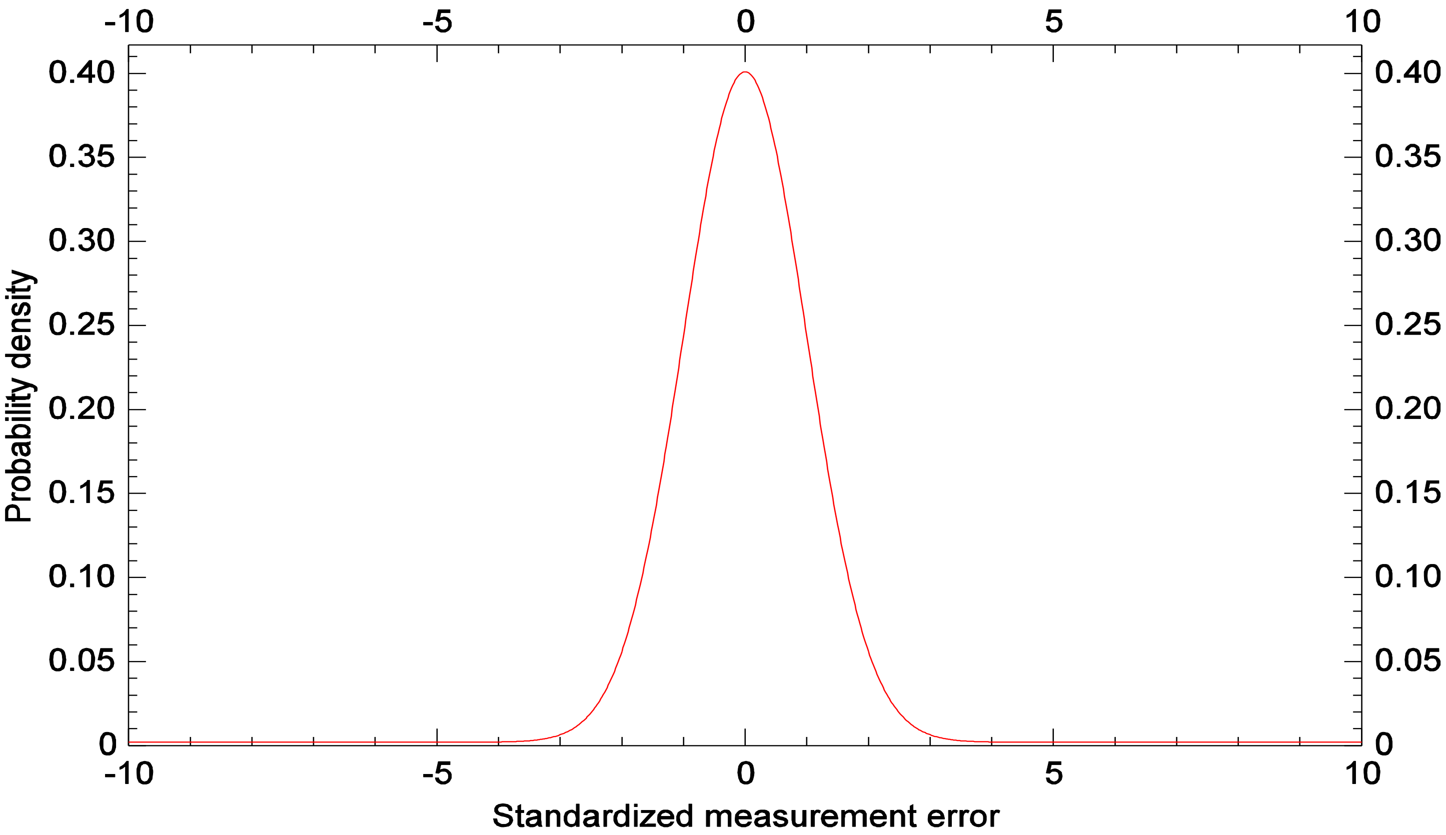
Typically, we believe that once the contributions of representativeness error and instrument quirks are all combined, *effective* observation errors have tails considerably fatter than those of a pure Gaussian (and peaks that are narrower).

If we knew just what that tail shape was, the Bayesian theory would essentially tell us how best to assign weight to that observation for each O-A.

The popular model for the probability density distribution of observation errors, which we believe to be too oversimplified, comprises:

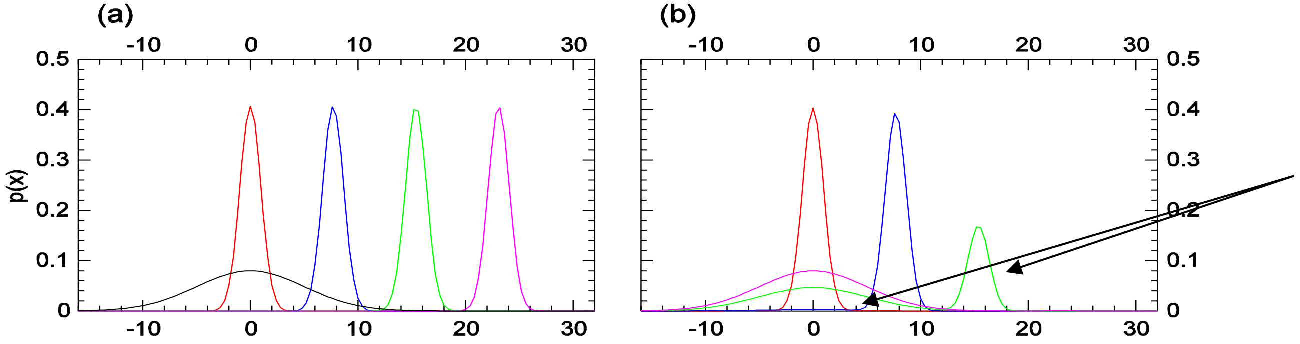
Gaussian + Uniform:

Gaussian-plus-uniform:



This is the kind of probability model used in the present NLQC in the GSI, but it cannot be safely "switched on" until 50 iterations have elapsed. (Note: "Uniform" makes a thin tail.)

Gaussian obs. Gaussian + Uniform obs.



Note the bimodality of the green posterior density.

The broad Gaussian prior density (black) multiplies the observation likelihood to produce the posterior density, alternative versions of which are shown for four different locations of the observation (close to their sharp peaks).

Theoretically, in a strictly Bayesian approach, the difficulty of multi-modality is resolved by convolving the posterior probability by a “loss function”, but this is not a practical possibility given the extremely high dimensionality of our problem.

The problem is avoided at the outset by a choice of fat-tailed distribution whose logarithm remains convex ---
i.e., it has no inflection points.

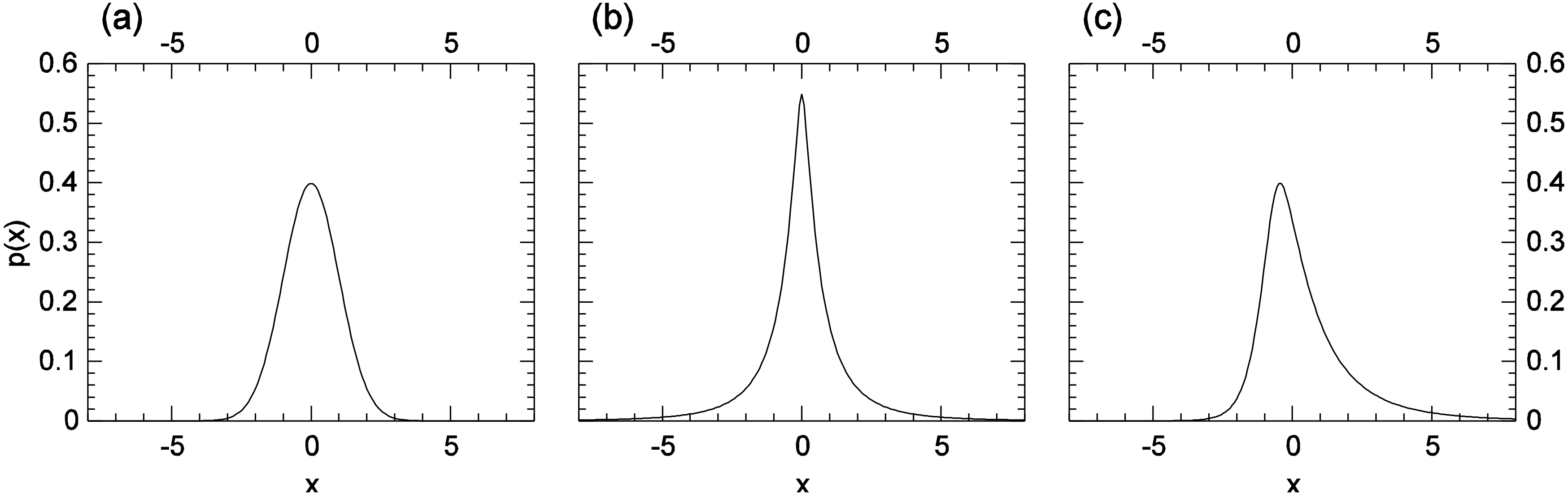
(Convex log-prior + convex log-likelihood
➔ Convex log-posterior➔ unimodality.)

Moreover, such a distribution is believed to better characterize typical real data. (See discussion in Tavolato and Isaksen, QJRMS 2014, whose “Huber norm” model also has this property.)

Gaussian mixture models

A more general model of observational error, reflecting the probable importance of fluctuating representativeness, (i.e., environmental effects) is obtained by a construction involving a positive, continuously-weighted “mixture” of Gaussians.

The contributing Gaussians certainly have a varying **scale** parameter, which will always lead to heavy tails. But it is not much harder to construct a mixture in which the contributing Gaussians also have varying **mean**s. Thus, we admit nontrivial skewness into the family of probability models.



In NOAA/NCEP Office Note 468, it was argued that a very natural symmetrical bell-shaped distribution expressible as a continuous Gaussian mixture is the classical “logistical distribution”

$$L(x) = .25\, \mathrm{sech}^2 [x/2]$$

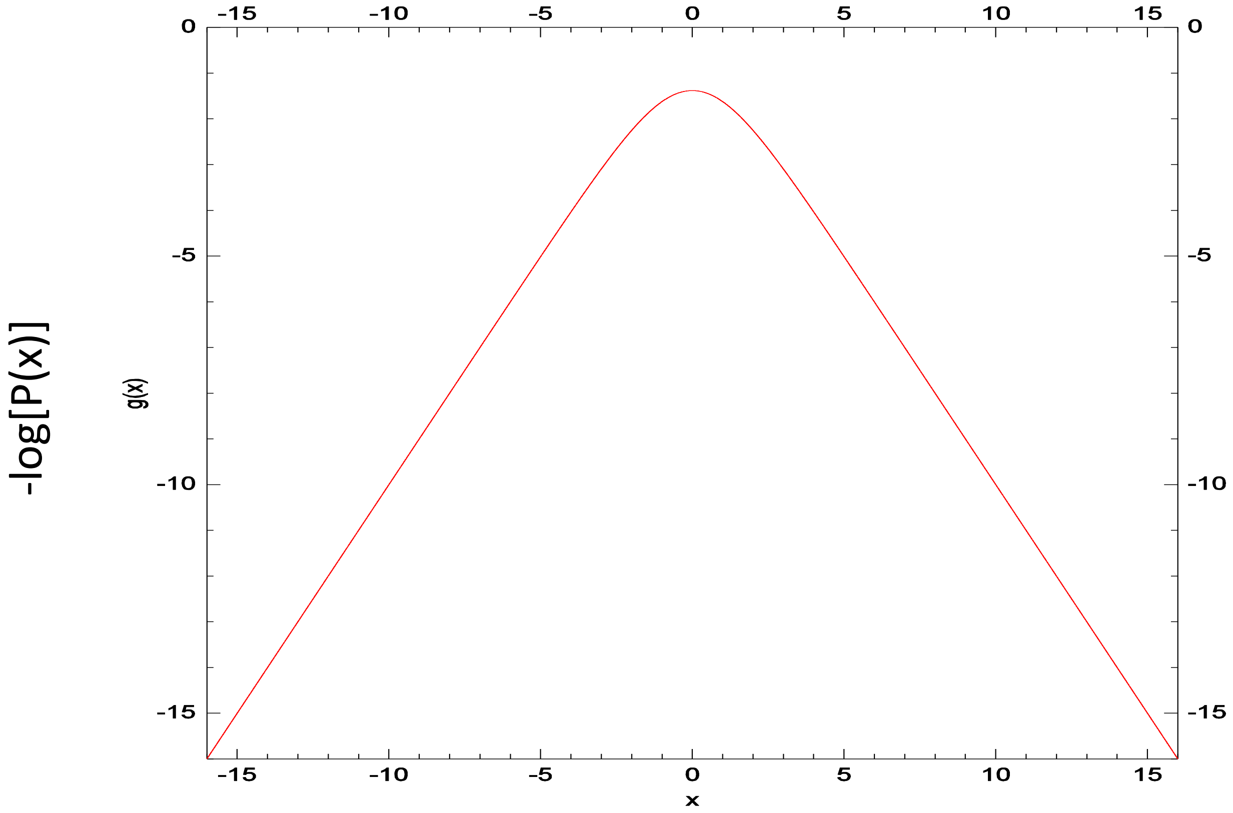
More general fat-tailed symmetric bell-shaped members come from this one simply by raising it to the positive power b:

$$f(x; b) = \mathrm{sech}^{2b} [x/2]$$

which can be shown (see the ON468 again) to preserve both convexity and the property of being a Gaussian mixture.

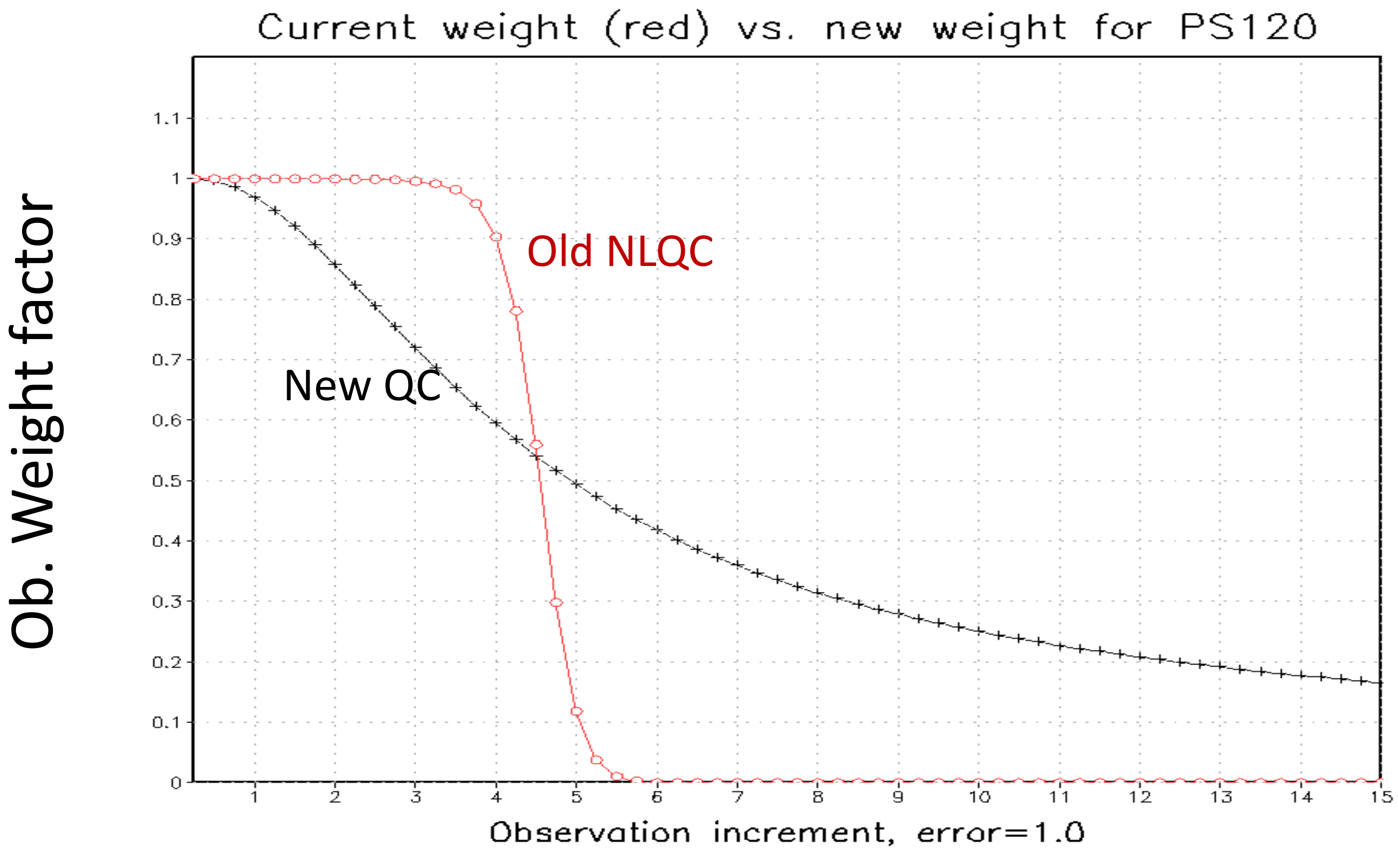
Further generalizations (i.e., additional parameters) to incorporate skewness and to provide additional fine control over the shape of the tails are available within the context of this proposed “Super-Logistic” model, though we are presently focusing on the implementation of the simpler, symmetrical model above, with its single shape parameter b.

The negative-log-probability for the ordinary logistic distribution has the form shown below; the effect of the shape parameter b is simply to change the slope of its asymptotes:



Comparison of the weight from new nonlinear QC with one from current operational GSI QC

- **Rawinsonde surface pressure**



Current status and future plans

The method of jointly estimating the shape parameters, b, and nominal standard deviation, sigma, is based on the statistical technique of “maximum likelihood”. This may need to be “regularized” for some data, e.g., by the inclusion of Bayesian priors, owing to the difficulty of solving problems of this kind. (And this ill-conditioning problem only tends to get worse as parameters are added, which is our intention for the future.)

Parallel tests currently under way to evaluate the new NLQC scheme in the 3D GSI appear to show improved convergence of the iterations. We are working on setting up corresponding tests in the RTMA/URMA.

Current status and future plans

As a general rule, the optimal value for the nominal sigma is reduced by a small percentage when nonlinear quality control of this type is applied.

Future extensions will include applications to correlated data. This should allow a large disparity between analysis and one observation to cause the automatic down-weighting of the neighboring observations made with the same instrument.

Thank you for your attention!