



A new observation error probability model for
Nonlinear Variational Quality Control and applications
within the NCEP Gridpoint Statistical Interpolation.

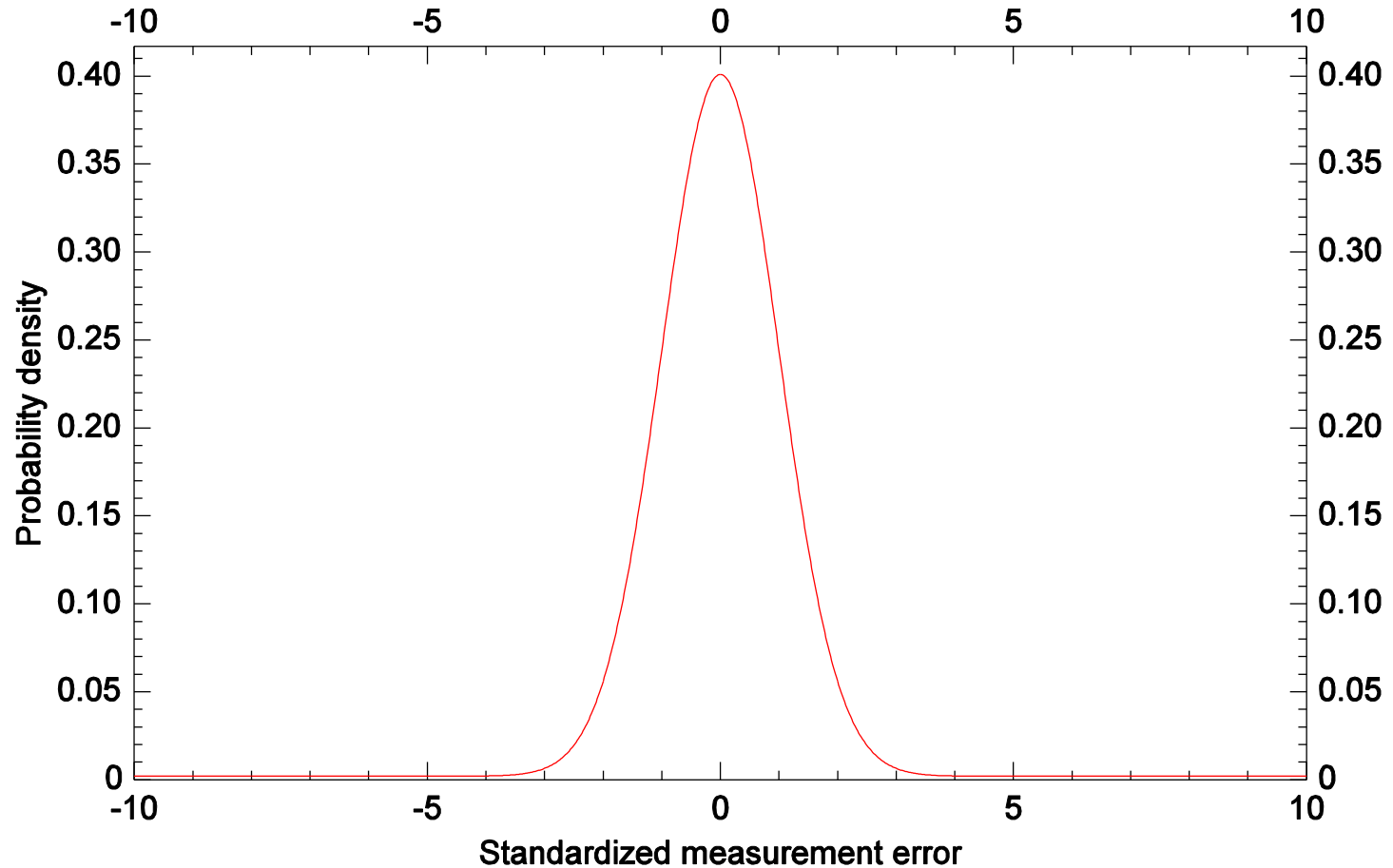
R. James Purser and Xiujuan Su

IMSG at NOAA/NCEP/EMC, College Park, MD, USA.



Nonlinear variational quality control provides a methodology for treating observations within a variational data assimilation in a way that accounts for the fact that, in practice, their effective errors are distributed according to probability densities whose tails are significantly heavier than those of a Gaussian.

In this way the rare but potentially damaging gross errors infecting typical atmospheric measurements can be recognized and down-weighted automatically within the iterative variational assimilation algorithm without recourse to a separate procedure.



One popular model of “realistic” observational error is the “Gaussian-plus-Uniform” distribution, which models the rare contingency in which the “normal” error is completely replaced by an essentially uniform random distribution conferring no information.

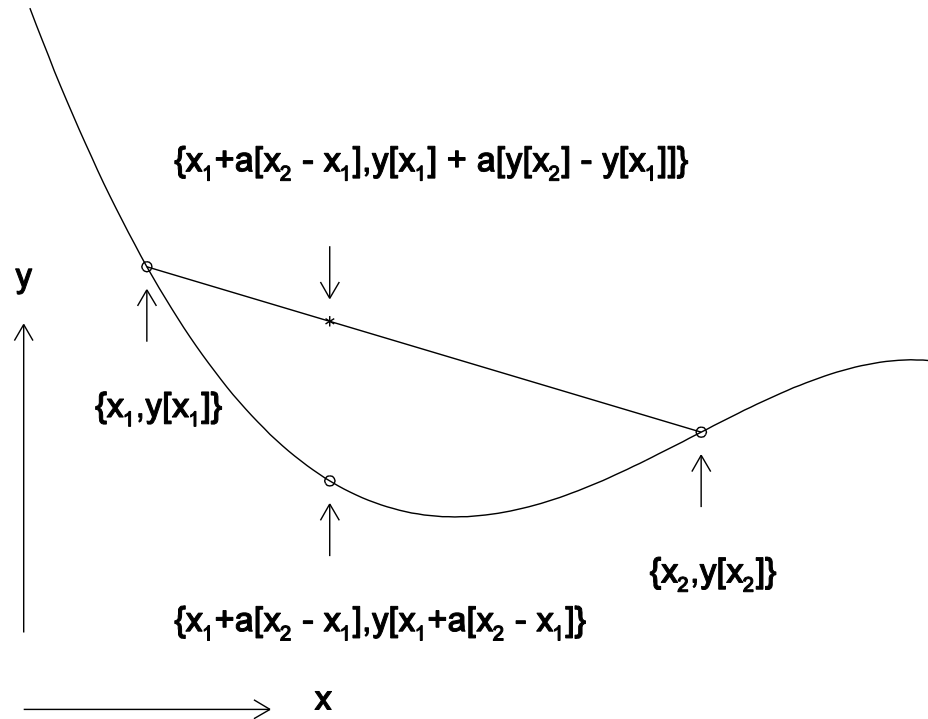


Unfortunately, an undesirable consequence of some of the simplest popular models of non-Gaussian error is an implied cost function characterized by multiple-minima.

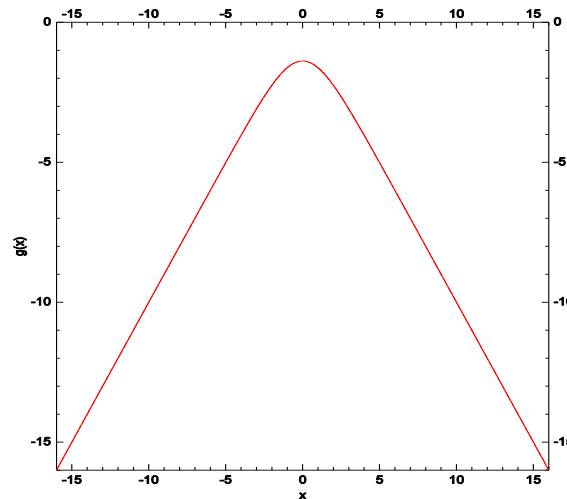
The reason for this is that these non-Gaussian probability models (such as Gaussian-plus-uniform) possess log-probability contributions that are not CONVEX functions and this nonconvexity is very often transferred to the cost function.

This means that it is easy for the assimilation to become effectively "locked" into false solutions close to the initial background state of the iterations that either accepts observations that should actually be severely down weighted, or fails to attribute sufficient weight to valid observations at variance from a badly misleading background to permit these good data from adequately correcting the bad background.

A graphical illustration of the notion of “convexity” in a function.



A new probability model for representing realistic measurement errors, which generalizes the "logistic" distribution, corrects the defective characteristics of traditional nonlinear quality control by ensuring that the negative-log-posterior distribution preserves the property of convexity possessed by the negative-log-prior, and is therefore free of multiple minima.



The figure shows the form of the logarithm of the Logistic, or “sech-squared”, distribution.

[Andrew Collard and Jeff Whitaker kindly brought to our attention the qualitative similarities of this distribution with the “Huber norm” model developed at ECMWF by Lars Isaksen and colleagues]



A technical discussion motivating the choice of this Logistic distribution and its generalizations as plausible models for realistic observational error is contained in the NOAA/NCEP Office Note 468 (available on line).

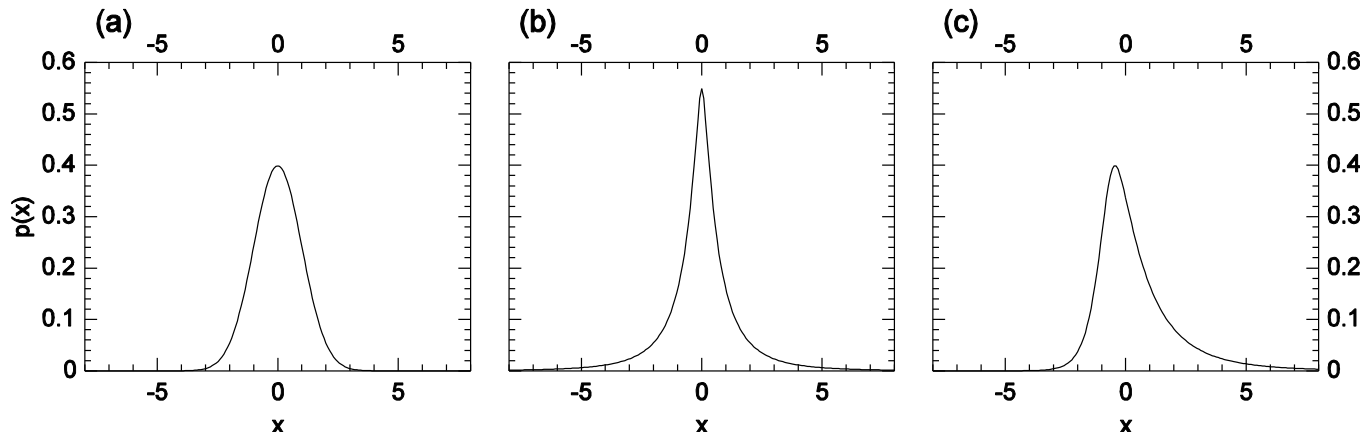
The simplest symmetrical generalization, which provides a control over the thickness of the distribution's tails, is to raise the basic Logistic to a power, b (the "broadness" parameter).

Another convexity-preserving generalization that allows control over the degree of asymmetry, and hence the distribution's skewness, modulates the density distribution with an exponential function, parameterized by an "asymmetry parameter", a . With this asymmetry, the asymptotes in the log-probability domain remain straight lines, but with unequal slopes.

Finally, abandoning the requirement of strict convexity, there is another, but more complicated generalization controlled by "convexity parameter", c , which causes the asymptotes in the log-domain to conform to curves proportional to $|x|^c$ (see ON 468 for details).

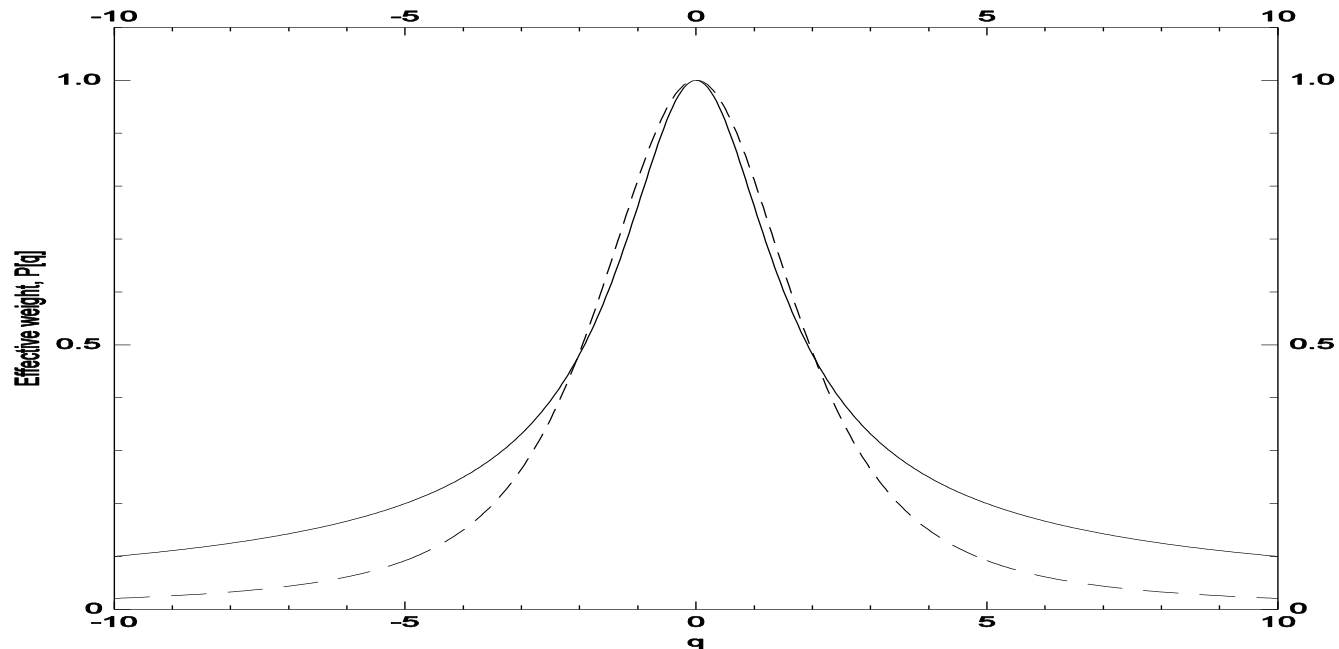
Members of this generalized **SUPER-LOGISTIC** family of density distributions share one attractive property:

Each is a continuous Gaussian mixture.



Panel (a) shows the Gaussian form; (b) shows an extreme example of a symmetric, but broad-tailed density (parameter **b** exercised); in the third panel, both parameters, **a** and **b**, are non-trivial and the result is a strongly skewed density model.

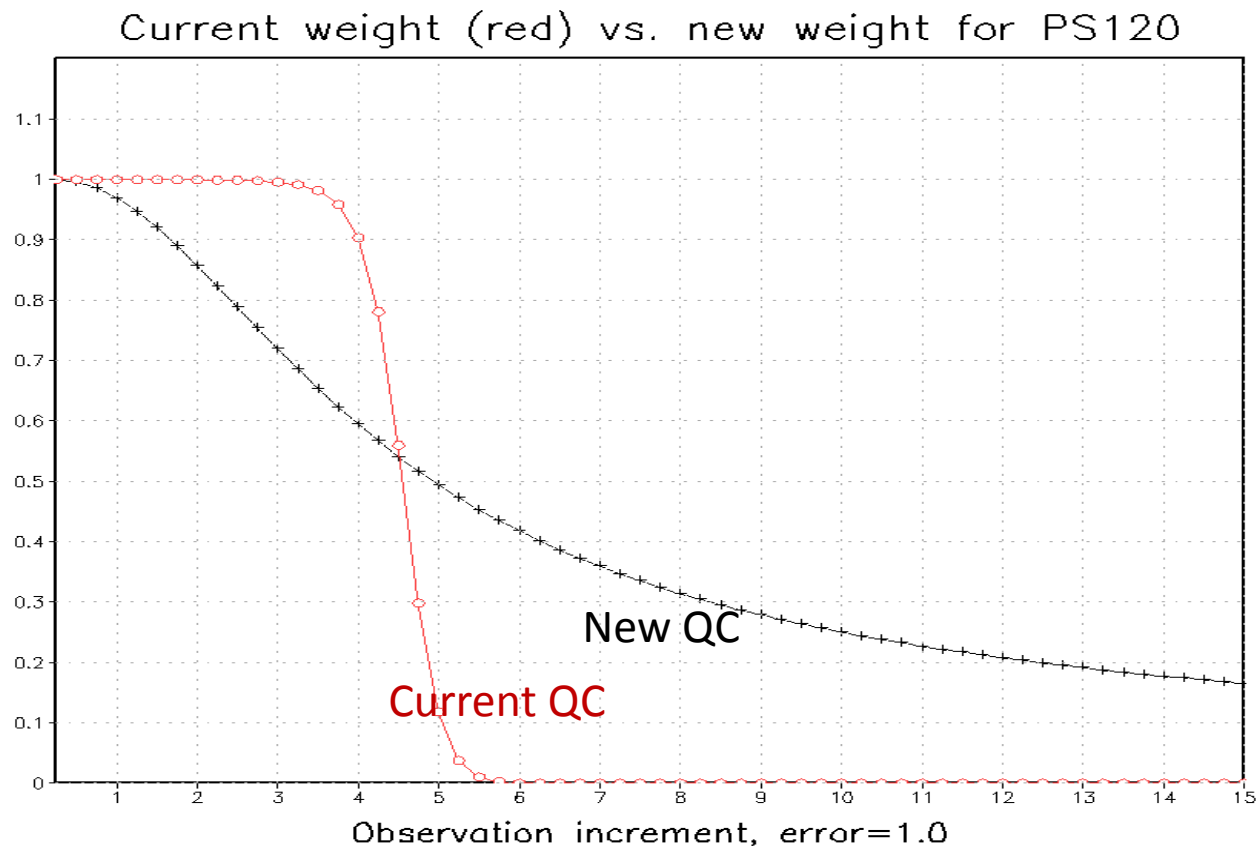
One way to look at the effect of substituting the new probability densities into the data assimilation scheme is to graph the effective weight factor that each model implies for the modulation of the observation's weight as a function of its O-A. For a Gaussian, the weighting is flat; for the new models the weight factor has profiles of the type shown below.



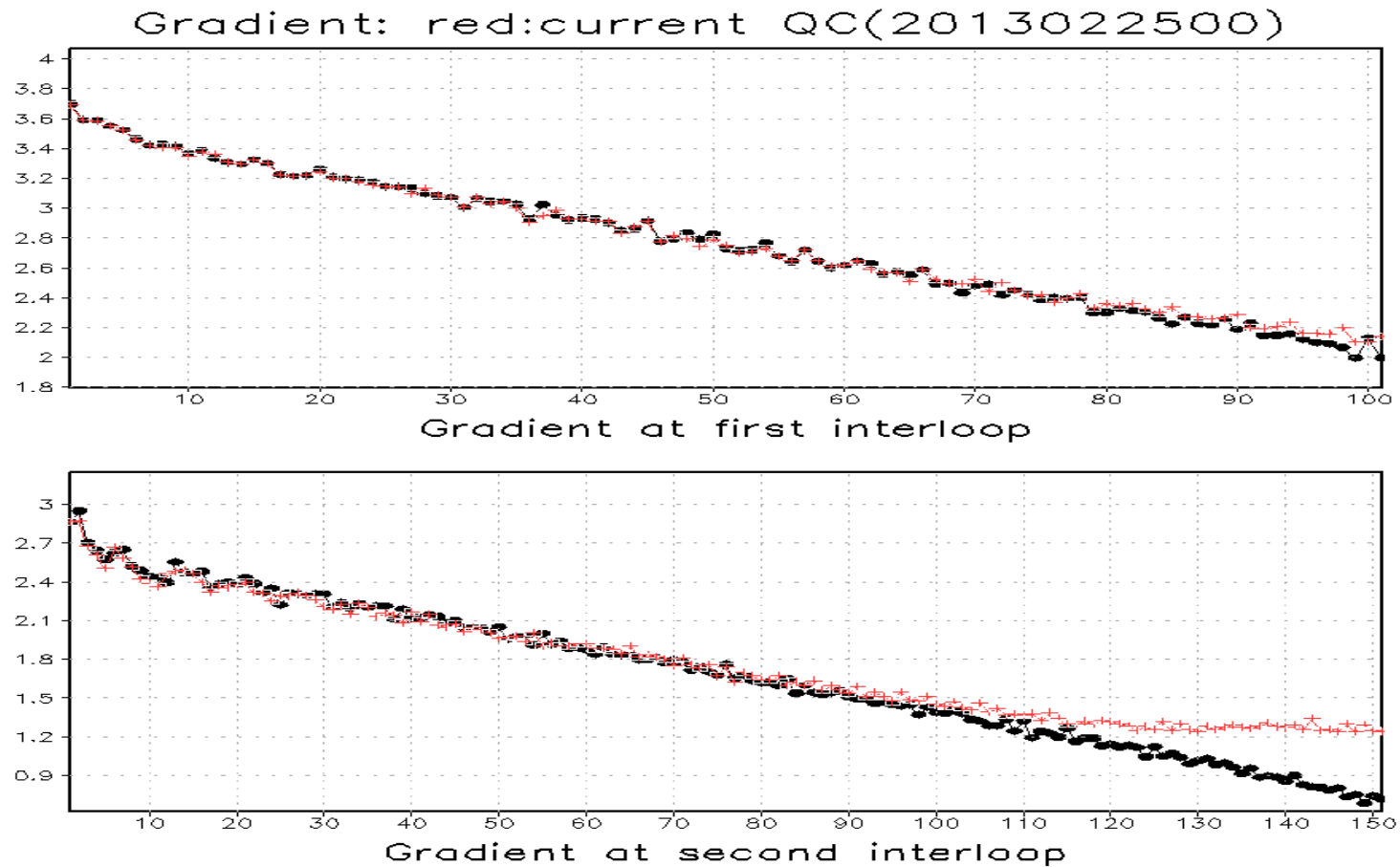
Both curves are symmetric and have significant b parameters, but the dashed curve also exercises the convexity control parameter, c , and then exhibits a more precipitous decline in the effective weight as $|O-A|$ increases.

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

- Rawinsonde surface pressure



Evolution of cost function gradient with iteration index, comparing current Operational nonlinear QC scheme (in red) with new one (black).





Concluding Remarks

We are able to exploit the convexity property of the new scheme to safely invoke it from the very first iterations (in the current scheme, owing to the multi-modal effects introduced into the cost function, the NLQC scheme is deferred for at least about 50 iterations, to make it more likely that the assimilation does not become “locked in” to an erroneous mode).

The challenging problem of estimating the appropriate parameters is complicated by the presence of either background or analysis error in the diagnostics (O-B) or (O-A) available from which to make an inference of the parameters. Also, in addition to the new “shape parameters”, the scaling of the effective spread (e.g., the standard deviation) of the model error and, in some cases, its bias, must also be included in the parameters to be freshly estimated (the standard deviation, σ , appropriate for a Gaussian+Uniform model is not the appropriate one for our super-logistic model, in general).

We are adopting an approach based on max-likelihood estimation. We are also looking at a slightly more convenient (for the user) parameterization of this super-logistic family than that described in the Office Note 468.