U.S. Department of Commerce
National Oceanic and Atmospheric Administration
National Weather Service
National Centers for Environmental Prediction
5830 University Research Court
College Park, MD 20740-3818

Office Note 495
https://doi.org/10.25923/kvmz-vf34

CONVENIENT PARAMETERIZATIONS OF SUPER-LOGISTIC PROBABILITY MODELS OF EFFECTIVE OBSERVATION ERROR

R. James Purser*
IM Systems Group, Rockville, Maryland

September 19, 2018

THIS IS AN UNREVIEWED MANUSCRIPT, PRIMARILY INTENDED FOR INFORMAL EXCHANGE OF INFORMATION AMONG THE NCEP STAFF MEMBERS

* email: jim.purser@noaa.gov

Abstract

This note describes a re-parameterization of the 'super-logistic' probability family, defined and described in a previous office note, and proposed as a model for the distribution of errors of typical scalar meteorological observations in which gross errors occasionally occur. The parameters of the distributions described in the earlier note (Office Note 468) were chosen to simplify the algebra in the mathematical investigation of the distributions proposed, but it is found that a re-parameterization of the identical distribution facilitates their practical application to real data. It is the purpose of this note to present this alternative re-parameterization.

1. Practical parameters for the super-logistic densities

The mathematical aspects of the proposed family of 'super-logistic' probability densities are described in a recent NOAA/NCEP office note (Purser 2011, henceforth denoted ON468). The motivation for developing these densities was to find, in the most compellingly natural way, generalizations of the Gaussian model that accommodate the broad tails characterizing typical error data, so that the method of nonlinear quality control (Purser 1984; Lorenc and Hammon 1988; Andersson and Järvinen, 1999; Tavolato and Isaksen 2014) could be confidently applied.

The principal generalization involved a 'broadness' (of the tails) positive parameter, called b, in ON468, which was essentially the power to which the logistic (or 'sech-squared') function is raised in order to match the intended degree of symmetrical broadening of the distribution's tails in the simplest and most symmetrical generalization. The Gaussian was only recovered asymptotically in the limit as $b \to \infty$, and the pure logistic density when $b = 1$. Although b was a convenient parameter choice for the algebraic investigations of the family of probability models carried out in ON468, it has been found not to be the most convenient choice from the point of view of the practitioner, partly because of its infinite range, and partly because, owing to the way it was formulated, a change in b led to an inconvenient change in the effective precision of the datum even when the observation was close to the background value.

From a practical perspective, it is desirable that, in expanding and generalizing a Gaussian probability model to incorporate broader tails, the Gaussian itself is recovered as the new broadness parameter, β, tends to zero (not infinity, as had been the case with the old broadness parameter, b). It is also desirable that the allowable range of this new parameter (and the others that we shall introduce) be finite, since this facilitates discrete tabulations within the parameter space. We have therefore chosen to let the pure logistic distribution be specified by $\beta = 1$, while the allowable range is limited to $\beta < 2$. In the new scheme, asymptotic broadness increases as β does, which conforms more directly to the intuitive notion that β represents the calibrated broadness.

In variational data assimilation, we recall that it is the negative-logarithm of the probability density function of the observational residual (observation minus analysis) that contributes to the overall cost function — the quantity which the variational scheme seeks to minimize. It is therefore desirable to avoid, as far as possible, introducing nonconvex contributions to this cost function in the vicinity of 'points' in the analysis state space visited by the steps in the minimization process, as there is a risk that the iterations could 'fall' into an unwanted secondary minimum (the implied probability density of the analysis space exhibits multimodality). The pure logistic, and the powers of it controlled by β, all have straight-line asymptotes in their negative-log-probability functions, which make them convex, but in only a borderline sense – the smallest smooth perturbation to the shape of the probability function could tip the balance into the nonconvex regime. For this reason, we like to calibrate the new 'convexity' parameter, which we call 'κ', to have a zero value for the pure logistic (and powers of it). For greater degrees of convexity, we have $\kappa > 0$, while for eventually concave asymptotes, we have $\kappa < 0$, but we confine κ to a finite allowable range, $\kappa \in (-1, 1)$. This convexity parameter replaces the previous parameter, c, of ON468. As $\kappa \to 1$, the shape of the asymptotes of the log-probability density become progressively more parabolic in shape; as $\beta \to 0$, the effective points of departure of the asymptotes from the 'osculating' parabola that best fits the central portion of the log-probability are pushed progressively further out from the center. Thus, although either changes $\kappa \to 1$ or $\beta \to 0$ tend to produce distributions more closely resembling the Gaussian form, the approach towards the Gaussian in these two cases occurs for subtly different reasons.

The asymmetry (a kind of skewness) is not typically needed, but a small degree of asymmetry might be useful for some unusual data. Again, the natural default for an asymmetry parameter should be zero in the case of pure symmetry and a theoretical range of the interval, $(-1, 1)$. Our new asymmetry parameter is named, 'α', but is defined slightly differently from the asymmetry parameter a of ON468.

In this brief note, we shall be concerned exclusively with the shapes of observation error probability density distributions described in their 'standardized' forms. By this we mean that, when the observation error is a dimensional quantity, x, while its assumed characteristic effective standard error of this kind of datum, in the same units, is σ in the special case of a vanishingly small departure of the measurement value from the background value, then the 'standardized distribution' of that kind of measurement is the probability density distribution of the non-dimensional error variable,

$$z = \frac{x}{\sigma}, \tag{1.1}$$

when the errors are distributed symmetrically. With this convention, the ith measurement should always be accorded an effective precision weight,

$$R_{i,i}^{-1} \equiv \frac{1}{\sigma^2}, \tag{1.2}$$

just as in the case of purely Gaussian models of observation error, at vanishingly small residuals, $z \to 0$. But for cases where the measurement does not accord with the collocated analysis (that must optimally blend the background with all the other observations as well) then this effective precision is modified, or 'modulated', by a 'weight factor', $W(z)$:

$$R_{i,i}^{-1} = W(z)\frac{1}{\sigma^2}. \tag{1.3}$$

For symmetric error distributions at least, this $W(z)$ generally decreases, on both sides of $z = 0$, monotonically with increasing $|z|$. In the cases where the error distributions are **not** symmetrical, the convention for standardization that we adopt here is to take the standardized error, z, to correspond to the error at which the probability density distribution achieves its 'mode', or maximum value, and to take the effective standard error, σ, to be such that the second derivative of the natural logarithm of the density at the mode with respect to x is $-\frac{1}{\sigma^2}$, thus ensuring that the equivalent derivative at the mode with respect to the 'standardized residual', z, remains negative-unity in every case.

To summarize, the main change we make here is the way we control the shape of the asymptotes of each probability function so that, regardless of the combination of the three new shape parameters, (α, β, κ), the standardized form of the probability density, $\hat{f}(z)$, is scaled and positioned such that it *always* peaks at $z = 0$, and its logarithm, $g(z) = \ln(\hat{f}(z))$ has a second derivative there:

$$\left.\frac{d^2 g}{dz^2}\right|_0 = -1. \tag{1.4}$$

This choice of standardization greatly simplifies the practical problem associated with the optimization of the choices of the various shape parameters from given sources of data for each data type. Section 2 will describe in greater detail how the parameters are defined. Section 3 will offer discussion and conclusions.

2. Re-parameterization of the super-logistic family

In a fairly general setting, a given type of data may be presumed to possess a unimodal probability density distribution of its *a priori* error that involves: a 'location' parameter, or bias, μ; a nominal 'scale' parameter in the same units, σ; and additional 'shape' parameters, such as our α, β and κ, that we have chosen to specify the remaining asymmetry, broadness and convexity attributes of the asymptotic 'tail' regions of the distributions. The standarization to a non-dimensional residual variable, z, as we described in the introduction, removes from direct consideration the location and scale aspects of the description in terms of the dimensional observation residual variable, x. However, these more complete densities, $f(\mu, \sigma, \alpha, \beta, \kappa; x)$ can be recovered trivially from the corresponding standardized density, $\hat{f}(\alpha, \beta, \kappa; z)$, simply by the identification:

$$f(\mu, \sigma, \alpha, \beta, \kappa; x) = \frac{1}{\sigma}\hat{f}(\alpha, \beta, \kappa; z), \quad z = \frac{x - \mu}{\sigma}. \tag{2.1}$$

Henceforth, we shall therefore consider only the probabilities in their standardized forms, and take the observation residuals in their non-dimensional representations, z.

The new style of parameterizing the probabilities uses explicitly an important property of the symmetrical distributions ($\alpha = 0$) mentioned in ON468 but not explicitly used there: each standardized log-probability, $g(0, \beta, \kappa; z)$, is a convolution of a power of $|z|$ times some negative constant, with a properly normalized (integral equals unity) and appropriately scaled (in width) logistic distribution in z. The width scale can always be found that will ensure the condition, (1.4). While it would not be practical to iteratively perform the requisite convolution to reconstruct the functional form each time we invoke a member of our super-logistic family in a data assimilation, we can certainly carry out the necessary calculations in advance for the construction of a discrete table of the attributes (effective weight multiplier and cost function contribution) associated with this family of functions. In fact, it has been found so convenient to do it this way that we have followed this kind of convolution procedure also in each case where there is a mild degree of asymmetry ($\alpha \neq 0$), even though the resulting density in such asymmetric cases does not exactly conform to the way it was proposed in ON468 to introduce asymmetry. Thus, although the new method may not result formally in the same kind of continuous Gaussian mixture model based on what, in ON468, we called a 'mobile heat transform' in the asymmetrical case, this Gaussian mixture property has never actually been exploited in practice, so its loss is not considered important.

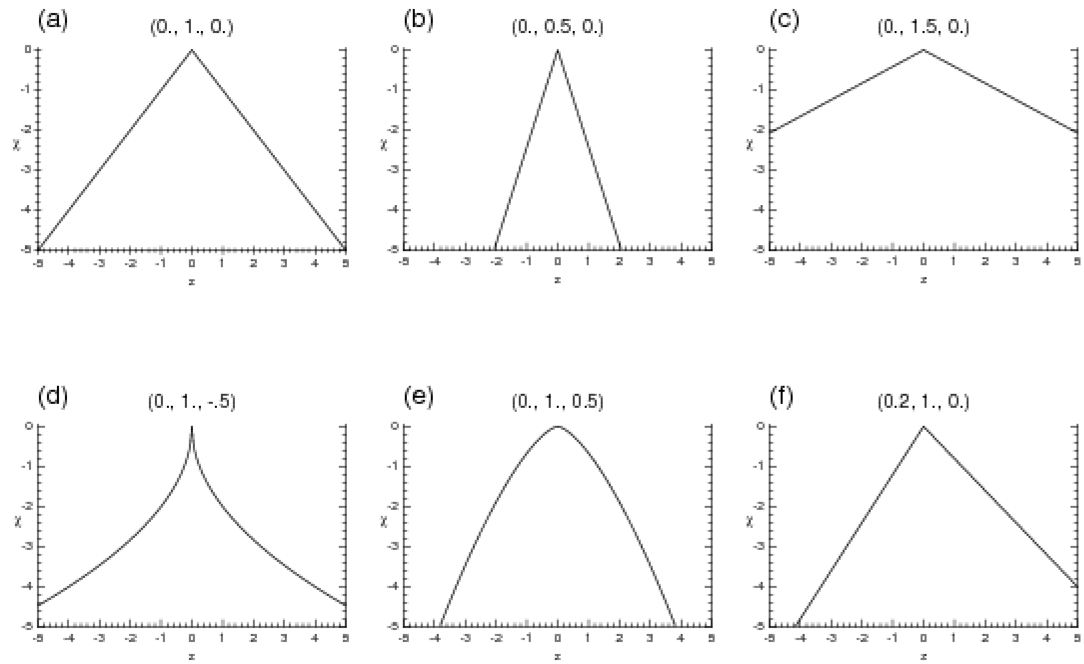

Figure 1. Assorted examples of the forms of the chevron functions, χ, for different triplets of shape parameters, (α, β, κ), as indicated above each panel.

We start by introducing what we might call an ideal power-law '*chevron function*', which is going to be used to determine the asymptotic behavior of the standardized log-probability at large $|z|$:

$$\chi(\alpha, \beta, \kappa; z) = \begin{cases} \frac{-\hat{b}(\beta)[1+\alpha]|z|^{\kappa+1}}{\kappa+1} & : \quad z < 0 \\ \frac{-\hat{b}(\beta)[1-\alpha]|z|^{\kappa+1}}{\kappa+1} & : \quad z > 0 \end{cases}, \tag{2.2}$$

where

$$\hat{b}(\beta) = \tan\left(\frac{\pi}{4}(2-\beta)\right). \tag{2.3}$$

Some illustrations for different parameter combinations are shown in Fig. 1. Note that the two branches of the chevron function differ only when $\alpha \neq 0$, and in a very simple way in any case. Ignoring the asymmetry, the rate, $|d\chi/dz|$ at which the function χ drops away with $|z|$ is constant when $\kappa = 0$, but even when there is a finite amount of convexity, $\kappa \neq 0$, we see that, at the standard unit of lateral displacement, $|z| = 1$, this rate is given exactly by $\hat{b}(\beta)$. When $\beta = 1$ (which, we recall, corresponds to the pure logistic case if $\alpha = \kappa = 0$), then $\hat{b}(1) = 1$. A negative κ clearly makes the inverted chevron function, $-\chi(z)$, concave. These attributes of steepness and convexity/concavity, as well as the asymmetry introduced by the α parameter, are all preserved with only minor alteration in the limbs of the function when it is convolved with a localized smoothing kernel. The logistic function is one such localized kernel and this is the one we employ to go from the chevron function to the standardized super-logistic, via convolution.

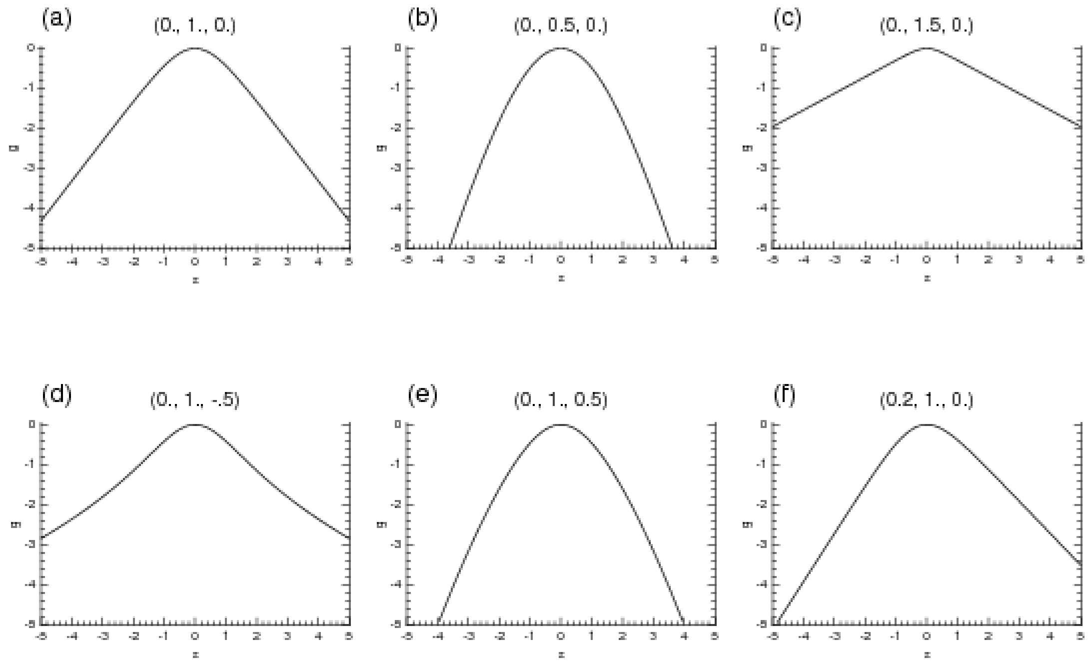

Figure 2. The centered profiles of the log-probability functions, $g(z)$, that correspond to the same triplets of shape parameters of the chevron functions illustrated in Fig. 1.

The normalized logistic distribution, a function of z, with its own location and scale parameters, z_0 and s, can be written:

$$\ell(z_0, s; z) = \frac{1}{4s}\mathrm{sech}^2\left(\frac{z - z_0}{2s}\right). \tag{2.4}$$

We simply define the log-probability of the general super-logistic function to be, apart from an additive constant that is only important when we need to normalize the implied probability density, the result of the convolution:

$$g(\alpha, \beta, \kappa; z) = g_0(\alpha, \beta, \kappa) + \int_{-\infty}^{\infty} \chi(\alpha, \beta, \kappa; z - z')\ell(z_0, s; z')dz', \tag{2.5}$$

where the additive constant, g_0, the location parameter, z_0, and the scale parameter, s, are adjusted jointly to ensure that:

$$g = 0, \quad z = 0, \tag{2.6a}$$

$$\frac{dg}{dz} = 0, \quad z = 0, \tag{2.6b}$$

$$\frac{d^2g}{dz^2} \;=\; -1, \quad z = 0. \tag{2.6c}$$

Fig. 2 shows the function, $g(z)$, plotted for all the cases corresponding to those shown in Fig. 1. A parabolic g would imply a Gaussian shape for the associated probability, $\hat{f}(z)$, since $\hat{f}(z) \propto \exp(g(z))$. The more parabolic-looking shapes within the plotted region exhibited in panels (b) and (e) confirm that the small β of panel (b) implies a relatively thin tail (small deviation from the Gaussian shape for small $|z|$), as does the positive κ of panel (e), though these two cases would show distinctly different shapes for their asymptotes at values of $|z|$ larger than are accommodated in these plots.

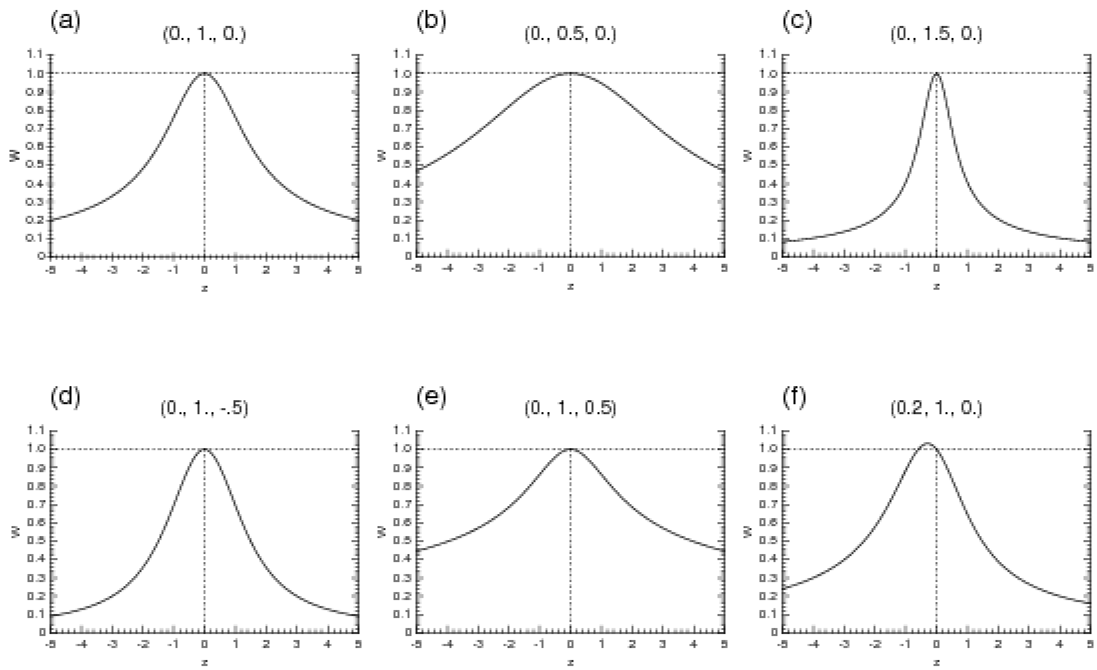

Figure 3. The profiles of modulating weight, $W(z)$, that correspond to the same triplets of shape parameters of the chevron functions illustrated in Fig. 1 or the log-probability functions, g, of Fig. 2 .

The convolutions can be carried out numerically by suitably and smoothly padding the far limbs of functions, χ and ℓ, in such a way that they both appear to be the central portions of broader periodic functions on a grid of dimensions compatible with efficient fast Fourier transforms (FFTs); this speeds up the convolution calculations on large grids. The sensitivity of the residual errors in (2.6a), (2.6b) and (2.6c), with respect to variations in g_0, z_0 and s, can then be used in a Newton-Raphson iteration to find the values of those parameters that ensure the fit, (2.6a), (2.6b) and (2.6c), is exact. This is carried out for each combination of α, β and κ in a discretization within the cuboid of theoretically allowable values.

For each tabulated combination of the shape parameters, we can now step through a discrete range of the argument, z, and tabulate, for each of its selected values, the effective contribution to the cost function. If we consider $\mathcal{L}_o$ to denote the part of the cost function that comes only from the observations, which we index by i, then we can write:

$$\mathcal{L}_o = \sum_i -g(\alpha_i, \beta_i, \kappa_i; z_i), \tag{2.7}$$

or simply,

$$\mathcal{L}_o = \sum_i -g_i(z_i), \tag{2.8}$$

and the effective weight multiplier for each datum, i, is:

$$W_i(z_i) = \begin{cases} 1 & : \quad z_i = 0 \\ \frac{-1}{z_i}\frac{dg_i(z_i)}{dz_i} & : \quad z_i \neq 0 \end{cases} . \tag{2.9}$$

In Fig. 3 we illustrate the weights, $W(z)$, that correspond to each of the cases we have exemplified in the previous figures. We see that, in all cases, the modulating weight factor, $W(z)$, goes to unity (the dotted horizontal line) at $z = 0$; for the symmetrical cases where $\alpha = 0$, this is the peak value of the weight and it decreases in a monotonic way as $|z|$ increases. But if there is asymmetry, $\alpha \neq 0$, as in the case of panel (f), the peak of the modulating weight is slightly offset from $z = 0$ (shown in each of these panels by the vertical dotted line).

3. Discussion and conclusions

We have replaced the three parameters, $\{a, b, c\}$ of the original formulation of the super-logistic distributions defined in ON468 by a new parameter set, $\{\alpha, \beta, \kappa\}$, designed to describe the probability family in a somewhat more convenient and rational manner for practical applications. With the new parameters, the simpler sub-family with only its broadness parameter, β, nonzero is the likely to be sufficient for most observations in applications to nonlinear quality control. The special case of the Gaussian distribution is recovered in the limit $\beta \to 0$. Since the convexity shape parameter, κ, can be kept zero in many typical cases, the convexity of the cost function is maintained.

Convexity is also preserved in the rare cases when the appropriate shape entails $\kappa > 0$. As κ increases towards its allowed upper bound of $+1$, the shape of the implied density distribution tends to become more approximately Gaussian regardless of the choice of β, so the objective optimization of parameters in this region of the parameter space becomes somewhat ambiguous in practice. In the better behaved cases of negative convexity, $\kappa < 0$, there remains the risk that multiple minima of the cost function might appear. To mitigate the risk of poor convergence that this can occasionally bring about, it is recommended that, at least for the first few iterations of the cost function minimization, the parameter, κ, be temporarily set to zero; after which, it is reset to its intended negative value. Then, even though secondary minima of the cost function might formally exist somewhere in the analysis state space, the first run of iterations should usually lead the estimated state to lie closest to the true intended minimum.

The asymmetry, corresponding qualitatively to some degree of skewness, is parameterized in the new convention by α. Although it is not expected that this parameter will be exercised except in very rare cases of data with unusual error characteristics, and when it exists at all, it is only expected to have a very small numerical magnitude, it is nevertheless accommodated in our probability model (although only for a very narrow numerical range of these values, in the practical tabulations).

The manner in which it is expected that this variational quality control scheme will be used is by way of a pre-selection of the shape parameters from a discrete tabular menu of them; this way, we avoid the need to perform costly multidimensional interpolations at each cost function iteration and, instead, interpolate the cost function penalty contribution and the weight multiplier for each datum only in the single dimension of variation in the standardized measurement residual, z; this should be computationally cheap enough.

Acknowledgments

The author is grateful to Drs. Yanqiu Zhu and Catherine Thomas for their reviews of this report.

References

Andersson, E., and H. Järvinen,	1999	Variational quality control. *Quart. J. Roy. Meteor. Soc.,* **125,** 697–722.
Lorenc, A. C., and O. Hammon	1988	Objective quality control of observations using Bayesian methods. Theory and a practical implementation. *Quart. J. Roy. Meteor. Soc.,* **114,** 515–543.
Purser, R. J.	1984	A new approach to the optimal assimilation of meteorological data by iterative Bayesian analysis. *Preprint, AMS 10th Conference on Weather Forecasting and Analysis,* Clearwater Beach, Florida, pp 102–105.
Purser, R. J.	2011	Mathematical principles of the construction and characterization of a parameterized family of Gaussian mixture distributions suitable to serve as models for the probability distributions of measurement errors in nonlinear quality control. NOAA/NCEP Office Note 468.
Tavolato, C., and L. Isaksen	2014	On the use of a Huber norm for observation quality control in the ECMWF 4D-Var. *Quart. J. Roy. Meteor. Soc.,* **141,** 1514–1527.