Revision of the ECMWF humidity analysis: Construction of a gaussian control variable

Elías Valur Hólm

ECMWF, Shinfield Park, Reading
GR2 9AX, United Kingdom

1 Introduction

In the ECMWF analysis specific humidity q is used as a control variable for the humidity. The background error covariances for humidity are determined partly statistically and partly empirically. The correlations are determined statistically from forecast differences. However, due to the spatial inhomogeneity of the humidity field, the global statistical variances are given by a local empiric function of relative humidity and temperature.

As a part of a revision of the ECMWF humidity analysis, we have looked for a new control variable for humidity with simpler errors characteristics than q. The main aim has been to find a control variable which has close to gaussian error statistics, which are the only error statistics that can be modelled accurately within the least square formulation of the analysis system. Another limitation posed by the analysis system is that at present only uncorrelated control variables can be modelled accurately. To obtain an uncorrelated control variable for humidity, the total variable is divided into an unbalanced/uncorrelated part and a balanced part, with the unbalanced part used as control variable and the balanced part being given by a function of the other variables.

The aim is thus to obtain a gaussian control variable uncorrelated with other control variables. Here we will only consider a control variable for total humidity in order to demonstrate the methodology. Later, when a humidity balance has been defined, statistically and/or analytically, exactly the same methodology can be applied to the unbalanced part of the humidity background error.

In the following sections we will look at the probability distributions of humidity forecast differences and use these distributions to obtain a more gaussian control variable.

2 Analysis of humidity forecast differences

By a gaussian control variable we mean that the background error of the variable, sampled over all gridpoints, has a gaussian distribution. We do not have the background errors available. However, we do have sets of forecast differences which are statistically related to the background errors. Consider two forecasts of the truth x, x_1^b and x_2^b, where

$$x_i^b = x + b^b(x) + \varepsilon_i^b \tag{1}$$

with b^b the bias and ε^b the stochastic error. The difference between forecasts is

$$x_1^b - x_2^b = \varepsilon_1^b - \varepsilon_2^b \tag{2}$$

We know that ε_1^b and ε_2^b are independent stochastic variables with the same pdf $P^b(\varepsilon^b)$. Then for $\delta x = \varepsilon_1^b - \varepsilon_2^b$,

$$P_C^b(\varepsilon_1^b - \varepsilon_2^b) = P_C^b(\delta x) = \int_{-\infty}^{\infty} P^b(\varepsilon_1^b) P^b(-(\delta x - \varepsilon_1^b)) d\varepsilon_1^b \tag{3}$$

So the forecast error differences are a convolution of the background errors with themselves. If $P_C^b(\delta x)$ is gaussian with variance σ^2, it can be shown that $P^b(\varepsilon^b)$ is also a gaussian and with variance $\sigma^2/2$. Therefore

it is of particular interest to find a gaussian description of the forecast differences, since this directly translates into a gaussian description of the background errors.

To study the error distribution of different humidity variables we can create histograms of the forecast differences. The forecast differences are between two forecasts valid at the same time, where each forecast comes from a different assimilation. Each assimilation uses the same set of observations with different perturbations added to each observation (consistent with the observation errors). To begin with, it is enough to consider one set of forecast differences (about 140000 gridpoints per level) to get a good idea about the statistical behaviour of the errors. From studying various candidates for the control variable, two results emerge:

- Forecast differences for q, $\log q$ and RH at a given model level (ca. 850 hPa here) show exponential like, rather than gaussian error distribution (Fig. 1, left). The same was found to be the case for several other humidity variables. For the particular level shown here, the q distribution deviates most from a gaussian.
- The forecast differences for a limited geographical region and/or similar values of the background fields are easier to approximate with a gaussian. For the same 850 hPa model level difference we have now looked at the distribution of differences in a 2.5% interval centered around the median of the background values of each variable (Fig. 1, right). Here q and $\log q$ now still have an exponential distribution, whereas RH is closer to a gaussian. The form of these distributions varies from level to level and for different values of the background. There are instances where for example RH is definitely not gaussian (close to $RH = 0$ and $RH = 1$), and where $\log q$ appears almost gaussian (close to the surface).

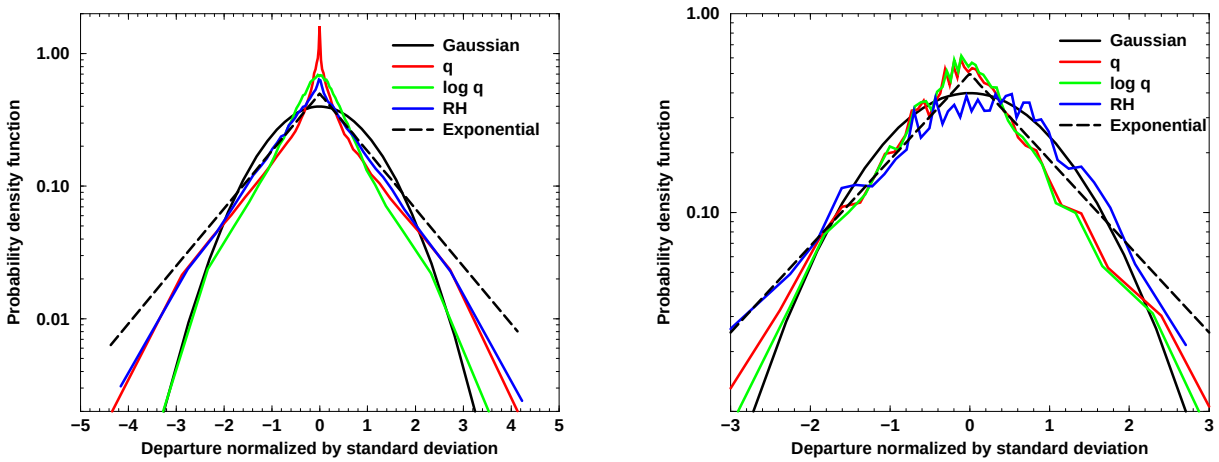

Figure 1: The pdf for a single forecast difference at ca. 850 hPa for q, $\log q$ and RH. The left panel shows all differences, and the right panel shows differences for similar values of the background field (a 2.5% interval centered around the median of the background values of each variable). For comparison, a gaussian and an exponential pdf are shown. The right panel is noisy due to limited number of differences in each interval, but the result remains similar when more fields are added to the statistics.

3 Finding a gaussian control variable

The explanation for the results in the previous section is that the total distribution we see at each level is an integral over different conditions, each with its characteristic errors. The forecast difference $\delta\varphi$ of a given humidity variable $\varphi(q, T, p)$ can be studied as a function of the background, for example as a function of φ

itself. The conditional error distribution $P(\delta\varphi|\varphi)$ varies with φ, and the total distribution is

$$P(\delta\varphi) = \int_{-\infty}^{\infty} P(\delta\varphi|\varphi)d\varphi \tag{4}$$

This function is generally not gaussian. Even if $P(\delta\varphi|\varphi)$ were gaussian for all values of φ (with standard deviation $\sigma(\varphi)$ and bias $b(\varphi)$), $P(\delta\varphi)$ would more probably be exponential due to the variation of σ and b with φ. A special case which would give gaussian $P(\delta\varphi)$ is if σ and b were constants. This gives a clue how to construct a gaussian control variable:

- Find a variable φ whose forecast difference $\delta\varphi$ follows a gaussian conditional error distribution $P(\delta\varphi|\Phi)$ as a function of some variable Φ. This means that the background conditional error $P^b(\varepsilon^b|\Phi)$ will also be gaussian, with variance $\sigma^2/2$. In practice a close to gaussian distribution will be sufficient.
- Determine the bias and standard deviation of the forecast differences as a function of Φ. From technical point of view, we would like the bias $b(\Phi)$ to be negligible. The reason is that if the minimum of the costfunction $J_b(\delta x)$ is not at $\delta x = 0$, then we will have nonzero analysis increments even if there are no observations.
- Normalize forecast differences by the bias and standard deviation,

$$\widetilde{\delta\varphi} = \frac{\delta\varphi - b(\Phi)}{\sigma(\Phi)} \tag{5}$$

  Note that if we have managed to find $\widetilde{\delta\varphi}$ following this procedure, then the forecast differences will be uniformly gaussian for all levels taken together, with zero bias and a standard deviation $\sigma_{\widetilde{\delta\varphi}} = 1$.
- For the analysis this implies a change to a control variable according to Eq. 5, with Φ chosen so that the bias is negligible and with the forecast difference standard deviation $\sigma(\Phi)$ replaced by $\sigma(\Phi)/\sqrt{2}$.

4 Control variable for humidity

4.1 Linear transformation of relative humidity

After experimenting with several formulations of the control variable, it was found that relative humidity RH had reasonably homogeneous statistics, but as shown in Fig. 1 using δRH as a control variable gives exponential error distribution. Including normalization $\widetilde{\delta RH} = \frac{\delta RH - b(RH^b)}{\sigma(RH^b)}$ gives close to gaussian distributions for median values of RH^b, but the distribution is asymmetric for extreme values of RH^b. This is expected since $P(\widetilde{\delta RH}|RH^b)$ is skewed towards negative values for large RH^b and positive values for small RH^b (see Fig. 2, left). An additional problem with this choice of control variable is the non-negligible bias, which does not fit into the formulation of the analysis (Fig. 2, right).

4.2 Symmetrizing transformation of relative humidity

From the study of forecast differences we can see that there is a way to avoid bias and asymmetry. If we have two forecasts RH_a and RH_b, then $P(RH_a - RH_b|RH_a)$ and $P(RH_a - RH_b|RH_b)$ are antisymmetric. This is easily checked by plotting the corresponding graphs (not shown). This antisymmetry can be explained by rewriting $P(RH_a - RH_b|RH_b)$ as $P(-(RH_b - RH_a)|RH_b)$ and noting that since RH_a and RH_b follow the same distribution, they can change place in the calculation of the statistics, so that $P(RH_a - RH_b|RH_b) = P(-(RH_a - RH_b)|RH_a)$. From this follows that if we stratify the statistics of $\delta RH = RH_a - RH_b$ according to the average of the forecasts we get the symmetric distribution $P(RH_a - RH_b|\frac{1}{2}(RH_a + RH_b)) = P(\delta RH|RH_b + \frac{1}{2}\delta RH)$. The result is shown in Fig. 3. The control variable is thus $\widetilde{\delta RH} = \frac{\delta RH}{\sigma(RH^b + \frac{1}{2}\delta RH)}$, where we have been able to eliminate the bias.

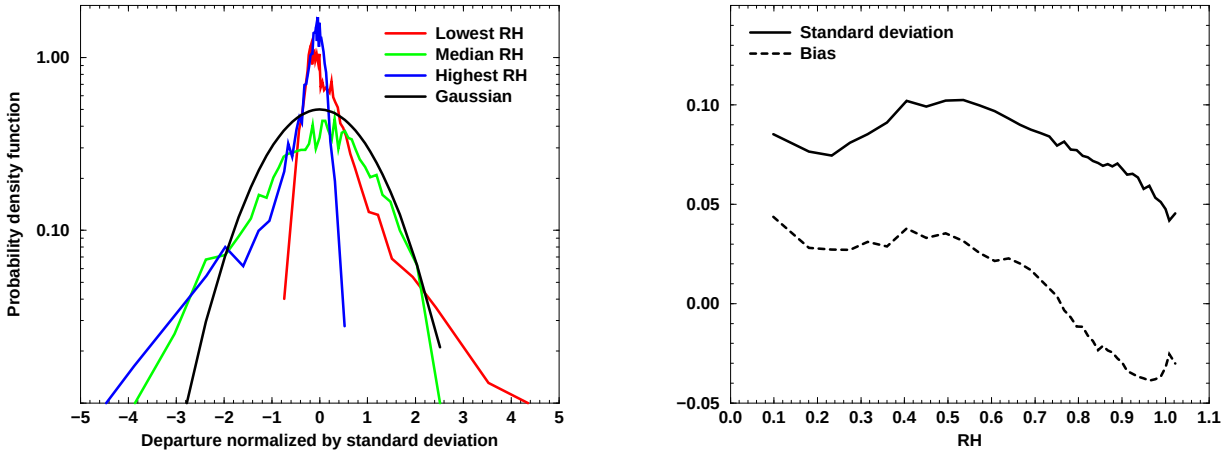

Figure 2: Forecast differences for the 'linear' $\delta RH(RH^b)$ at 850 hPa. The left panel shows the pdf's for lowest, median and highest 2.5% values of RH^b, and the right panel shows the standard deviation and bias as a function of RH^b.

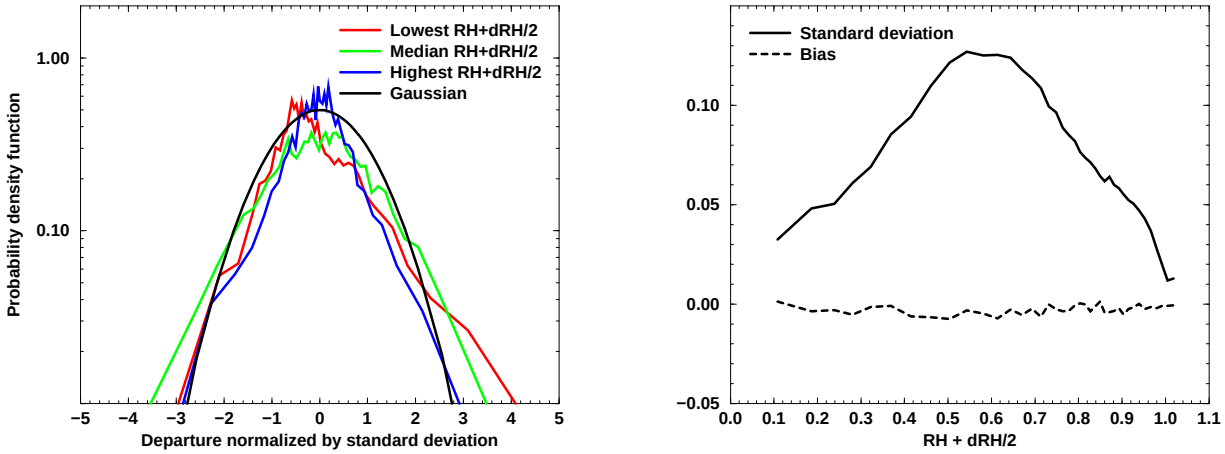

Figure 3: Forecast differences for the 'symmetric' $\delta RH(RH^b + \frac{1}{2}\delta RH)$ at 850 hPa. The left panel shows the pdf's for lowest, median and highest 2.5% values of $RH^b + \frac{1}{2}\delta RH$, and the right panel shows the standard deviation and bias as a function of $RH^b + \frac{1}{2}\delta RH$. All bins can be reasonably modelled by a gaussian. Note that the extreme bins are particularly affected by model and analysis effects of supersaturation clipping and resetting humidity to positive *values.*

4.3 Direct 'gaussianization'

An alternative way to derive a gaussian control variable would be to define a nonlinear transformation $f(\delta\varphi, \Phi)$ directly, based on forecast error differences. As we have seen in previous sections, we need to make the transformation a function of the background conditions if we want to end up with gaussian distributions for all conditions. The way to achieve this is to base the 'gaussianization' transformation on the conditional pdf $P(\delta\varphi|\Phi)$ instead of basing it on $P(\delta\varphi)$. For a given Φ, we just need to find a transformation $f(\delta\varphi, \Phi)$ of the $\delta\varphi$ axis so that the probability that $x \leq \delta\varphi$ equals the probability that $\xi \leq f(\delta\varphi, \Phi)$ for a normal gaussian

distribution,

$$\Pi(\delta\varphi|\Phi) = \int_{-\infty}^{\delta\varphi} P(x|\Phi)dx = \int_{-\infty}^{f(\delta\varphi,\Phi)} \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2} d\xi = \Pi_G(f(\delta\varphi,\Phi)) \tag{6}$$

where Π are the cumulative pdf's. Inverting the gaussian cumulative distribution then gives

$$f(\delta\varphi,\Phi) = \Pi_G^{-1}(\Pi(\delta\varphi|\Phi)) \tag{7}$$

This will always work, but we may end up with a bias term which does not fit into the analysis framework as has been explained above. However, we could apply this transform as a finishing step after the symmetrizing transform and in that way avoid the bias term. This approach makes the search for a gaussian control variable automatic, although it does not guarantee the best control variable. There may be another choice which is less nonlinear for example.

For the normalized relative humidity control variable, we do not need to apply this final 'gaussianization', since the departures are already close to gaussian. But if there is no other way to find a reasonably Gaussian control variable, this approach can always be applied.

4.4 Implementation of the nonlinear control variable

The difficulty with the symmetric control variable is that a nonlinear variable transform is needed to go from the model to the control variable. This nonlinearity is unavoidable when converting from non-gaussian to gaussian control variables. The ECMWF analysis consists of a series of minimizations (inner loops) linearized around a nonlinear reference state (outer loop). In the inner loops we must use a control variable which is linearly related to the model variables, but there is no such restriction in the outer loops. So we can use the nonlinear symmetrizing transform when going between inner and outer loops and use a linearization around the last outer loop in the inner loops.

The new control variable is implemented as a change of variable in the analysis, with the background error at each level approximated by a polynomial $\mathcal{P}(RH^b + \frac{1}{2}\delta RH)$. This makes the relative humidity background error a function of relative humidity and pressure, whereas the earlier formulation was a function of relative humidity and temperature. In additon, there is now a multivariate coupling between q, T, and p through the use of linearized relative humidity.

The background error covariance matrices are now calculated for normalized relative humidity instead of specific humidity. In Fig. 4 we show a comparison of the forecast differences δq and $\widetilde{\delta RH}$ at ca. 850 hPa. These are the actual inputs for the calculation of the covariances. The transformed relative humidity shows a homogeneous difference field which will be much easier to characterize statistically than the specific humidity difference field.

5 Conclusions

We have shown how a detailed study of the probability distributions of forecast differences can help in the design of a gaussian control variable. The main point is to find 'an axis' along which the differences follow gaussian distributions, then determine the bias and standard deviation of the differences along the 'axis', and finally normalize the differences with the bias and standard deviation to obtain a normal gaussian distribution. The background errors are directly related to the forecast differences in that if the forecast differences are gaussian, so are the background errors.

There are a few problems which can arise while finding a gaussian control variable. First, it may be difficult to find an appropriate 'axis', expecially for variables like relative humidity which are bounded from above and below. Asymmetric distributions are a common problem here. To solve this a symmetrizing transform has been

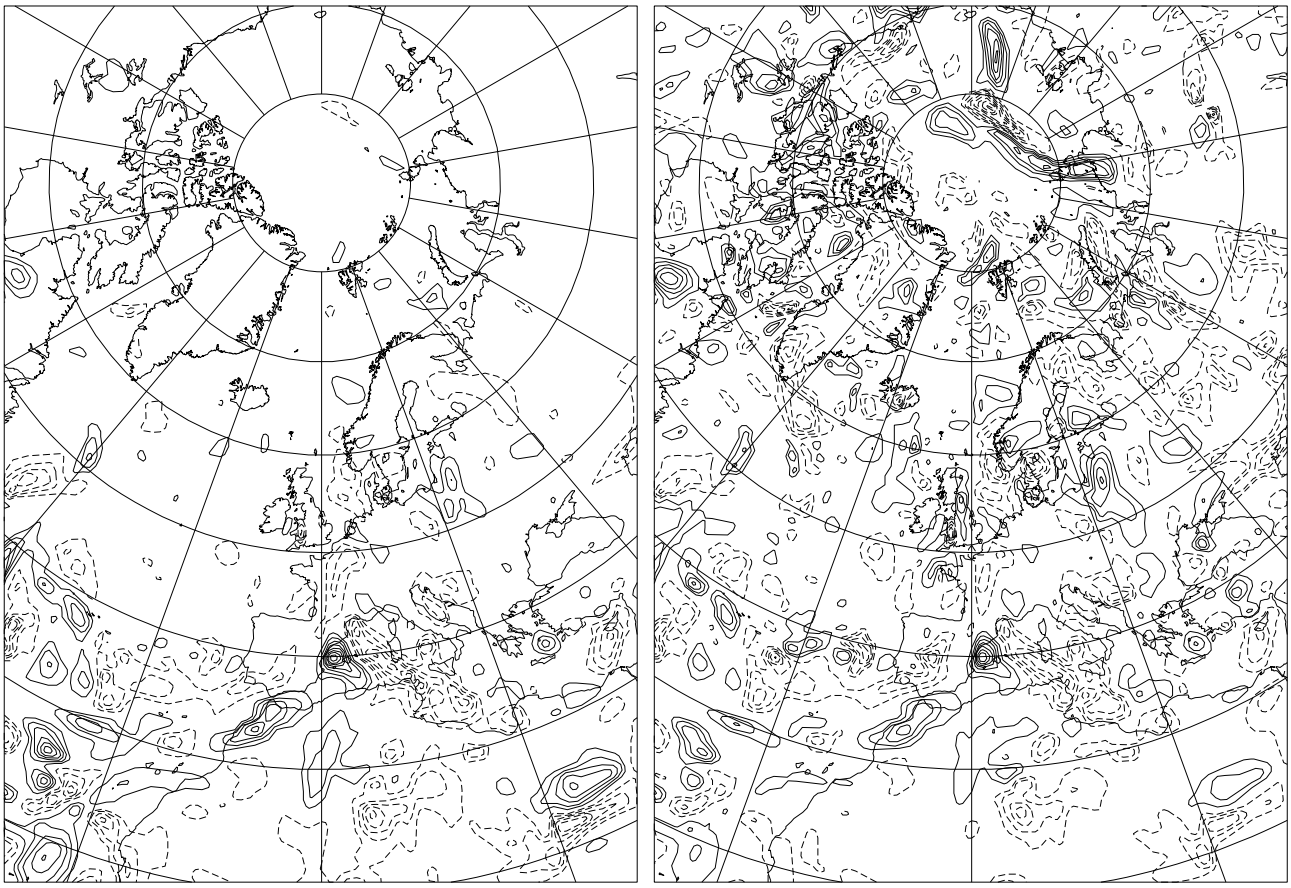

Figure 4: Forecast differences for specific humifity δq (left, isolines 0.005) and normalized relative humidity $\widetilde{\delta RH}$ (right, isolines 1) at a single level (ca. 850 hPa). The two fields show what goes into the statistical determination of covariance matrices for the two variables. Normalized relative humidity is more homogeneous.

defined. Second, from a technical point of view, we would like to be able to neglect any bias term in the change of variable, and this is also automatically solved by the symmetrizing transform. Third, the transform from a non-gaussian to a gaussian variable introduces nonlinearities in the analysis. The solution to this is to apply the full nonlinear transform only when going between inner and outer loops of the analysis, and use a linearized transform in inner loops.

The control variable we found for (total) humidity is a normalized relative humidity. At each model gridpoint relative humidity is divided by a polynomial approximation of the background error, which varies as a function of the background relative humidity in the inner loops, with an additional nonlinear dependency of the increment itself in the outer loops. Since relative humidity is not a model variable, this choice of control variable introduces a multivariate relation between the background errors of specific humidity, temperature, and pressure. We plan to extend the approach here to an unbalanced humidity control variable, as a part of work to further integrate the humidity with other analysis variables.

Acknowledgements

I owe many thanks to my colleagues Mike Fisher, Erik Andersson, and Lars Isaksen for discussions and suggestions during this work.