Bayesian Data assimilation by using Gaussian and non-Gaussian distributed errors

Carlos Pires[1]

[1] Instituto Dom Luiz (IDL), Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal

clpires@fc.ul.pt

Numerical Weather Prediction – Portugal Workshop,
November 26-27, 2018

Instituto Português do Mar e da Atmosfera (IPMA)

Fundamentals of Data Assimilation (DA)

Data assimilation is a mathematical discipline that seeks to optimally combine theoretical (background or prior) estimations or knowledge (X_b) of the state X of a (numerical or qualitative) model (simulating a certain system) with a packet Y of observations, related to X , through a certain (complex or simple, fixed or varying) observation operator $H(X)$.

$$\mathbf{Y} = H(\mathbf{X}) + \boldsymbol{\varepsilon}_o$$

$$\mathbf{X}_b = \mathbf{X} + \boldsymbol{\varepsilon}_b$$

Example: Y= Remote sensing radiative data ; X = 3D, time-varying state vector of a meteorological forecasting physical model; $H(X)$ = Radiative Transfer Model; X_b = Forecast valid now, issued from yesterday;

Fundamentals of Data Assimilation (DA)

The Data Assimilation goal is thus to obtain (from data) an improved (or posterior) estimation (X_a) of the model state (the analysis), optimizing a certain statistical criterium by taking into account the assumed probabilistic distributions $\rho_0(\varepsilon_o), \rho_b(\varepsilon_b)$ of the different errors, coming, both from the prior estimations (due to model errors, truncation, physics etc.) and from observations (instrumental calibration, spatio-temporal representation, spatial colocation, clock errors). Furthermore, the assumed probability distribution errors can eventually be imperfect due to: 1)wrong pdf choice; 2)false assumption of errors independence, 3)error biases, 4)under (or over) quantification of error variances and of error extremes, 5)probability of observation rejection badly assessed.

Data: Observation + background

$$\mathbf{Y} = H(\mathbf{X}) + \boldsymbol{\varepsilon}_o$$
$$\mathbf{X}_b = \mathbf{X} + \boldsymbol{\varepsilon}_b$$

+

pdfs of observation and background errors

$$\rho_o(\boldsymbol{\varepsilon}_o)\ ,\ \rho_b(\boldsymbol{\varepsilon}_b)$$

→

Posterior State and its pdf

$$\mathbf{X}_a = \mathbf{X} + \boldsymbol{\varepsilon}_a$$
$$\rho_{post}(\mathbf{X})$$

The posterior Bayesian-based probability density function (pdf) is a pdf conditioned to data

$$\rho_{post}(\mathbf{X}) = \rho(\mathbf{X} \mid \mathbf{X}_b, \mathbf{Y}) = \frac{\rho(\mathbf{X}_b, \mathbf{Y} \mid \mathbf{X})\rho_{cli}(\mathbf{X})}{\rho(\mathbf{X}_b, \mathbf{Y})}$$

$$= \hat{C}\rho_{cli}(\mathbf{X})\rho_{ob}[\mathbf{X}_b - \mathbf{X}, \mathbf{Y} - H(\mathbf{X})]$$

When errors are statistically independent and climatic variance is much larger than data error's variances, the posterior pdf comes as:

$$\rho_{post}(\mathbf{X}) \approx C\rho_b(\mathbf{X}_b - \mathbf{X})\rho_o[\mathbf{Y} - H(\mathbf{X})]$$

The posterior state according to different criteria

The Most-likely (ML) state: $\mathbf{X}_{ML} = \arg\max_{\mathbf{U}}[\rho_{post}(\mathbf{U})]$

The Median of the posterior pdf (MED): $\mathbf{X}_{MED} = Q_{\rho_{post}}(0.5)$

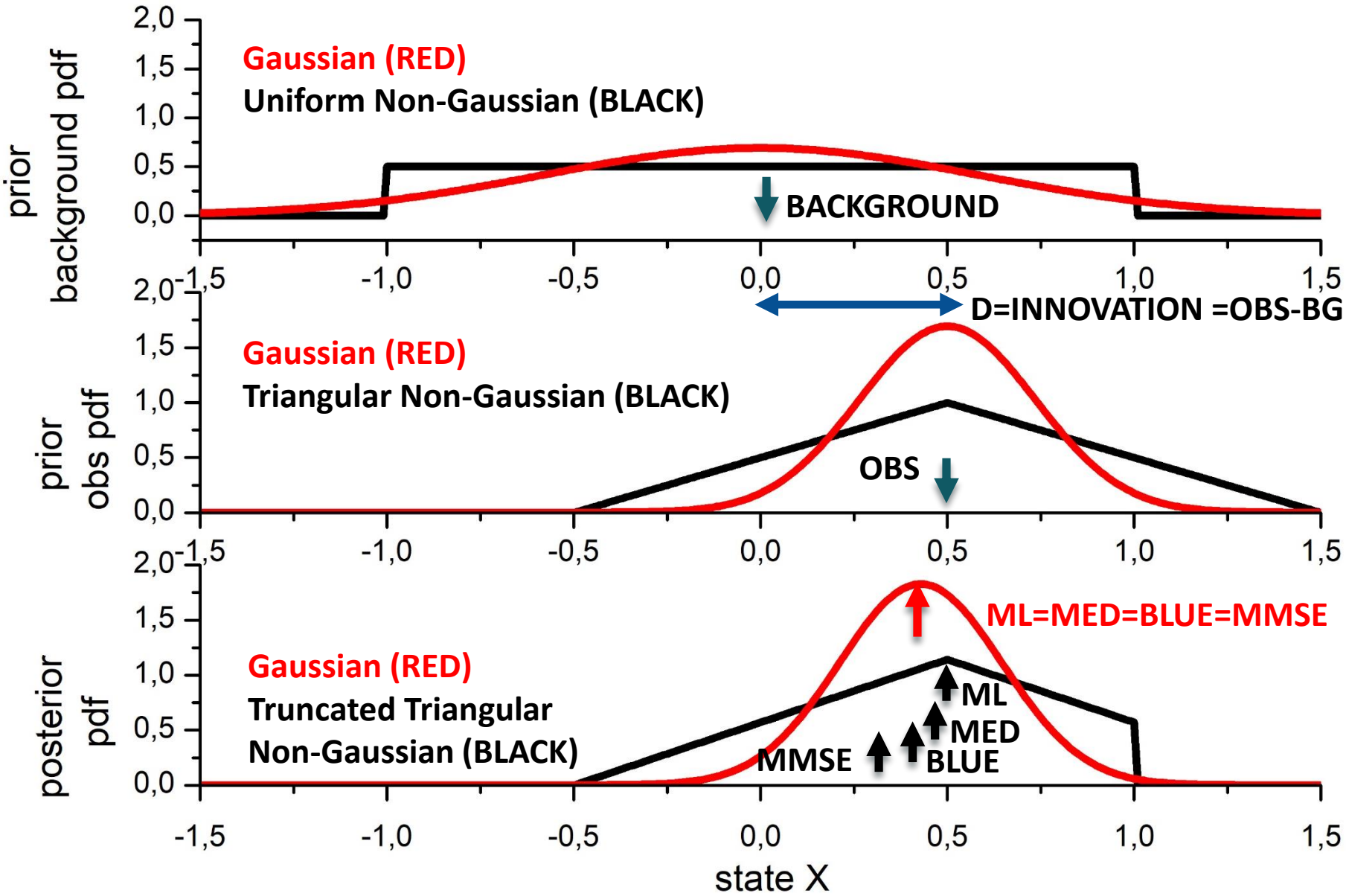
The Best Linear Unbiased Estimator (BLUE):

$$\mathbf{X}_{BLUE} = \mathbf{X}_b + \mathbf{K}^*[\mathbf{Y} - H(\mathbf{X}_b)]$$

$$\mathbf{K}^* = \mathbf{\Sigma_b} \mathbf{H}^T (\mathbf{\Sigma}_o + \mathbf{H}\mathbf{\Sigma_b}\mathbf{H}^T)^{-1} \;;\; \mathbf{H} = \partial H / \partial \mathbf{X}_{\mathbf{X}_b}$$

The Minimum Mean Square Error estimator (MMSE):

$$\mathbf{X}_{MMSE} = E_{\rho_{post}}[\mathbf{X}] = \int \mathbf{U}\rho_{post}(\mathbf{U}) \,|\, d\mathbf{U} \,|$$



$$MMSE = BG + .81\mathbf{D} \qquad MED = BG + .94\,\mathbf{D} \qquad OBS = BG + \mathbf{D}$$

$$BLUE = BG + .86\,\mathbf{D} \qquad ML = BG + \mathbf{D}$$

The choice of errors' pdfs make a difference in the analysis

Therefore,

How may we decide if the chosen pdf of errors is correct or not?

Answer: By collecting statistics of the innovations (differences between the observations and backgrouds

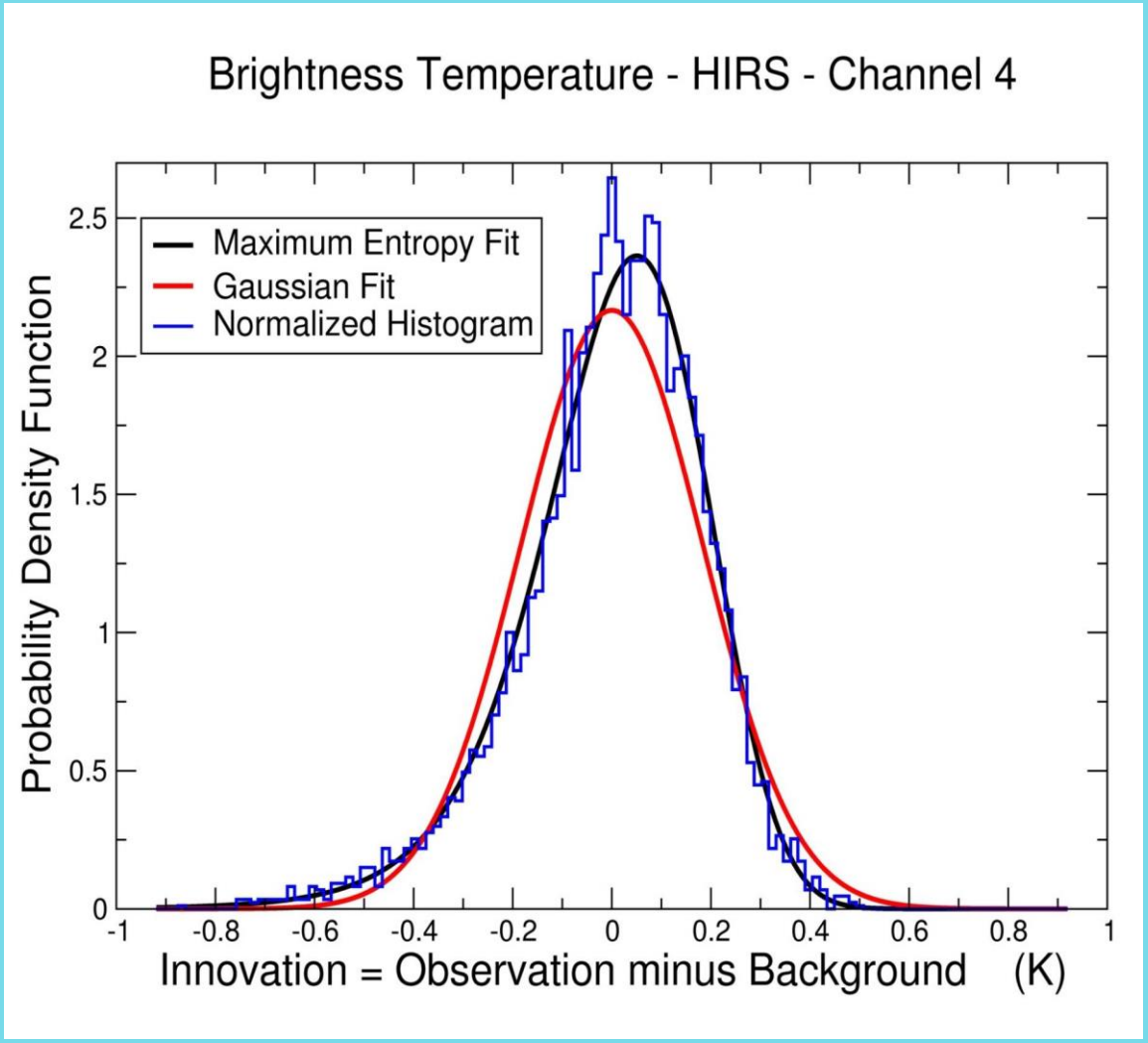
The innovation is expressed as a difference between errors. When independent, Its variance is the sum of error variances.

$$D = OBS - BG = \varepsilon_o - \varepsilon_b \quad ; \quad \sigma_d^2 = \sigma_o^2 + \sigma_b^2$$

The Gaussianity of errors may be tested by computing the skewness (s) and kurtosis excess (k') of innovations D.
Under Gaussian pdfs , s=k'=0.
s and k' of D are shared among the skewnesses and kurtoses of OBS and BG errors as:

$$s_d = \overline{\left(\frac{d'}{\sigma_d}\right)^3} = s_o \left(\frac{\sigma_o}{\sigma_d}\right)^{3/2} - s_b \left(\frac{\sigma_b}{\sigma_d}\right)^{3/2} \quad (Skewness: s)$$

$$k_d' = \overline{\left(\frac{d'}{\sigma_d}\right)^4} - 3 = k_o' \left(\frac{\sigma_o}{\sigma_d}\right)^2 + k_b' \left(\frac{\sigma_b}{\sigma_d}\right)^2 \quad (Kurtosis\ Excess:\ k'=k-3)$$



Example of non-Gaussian innovations

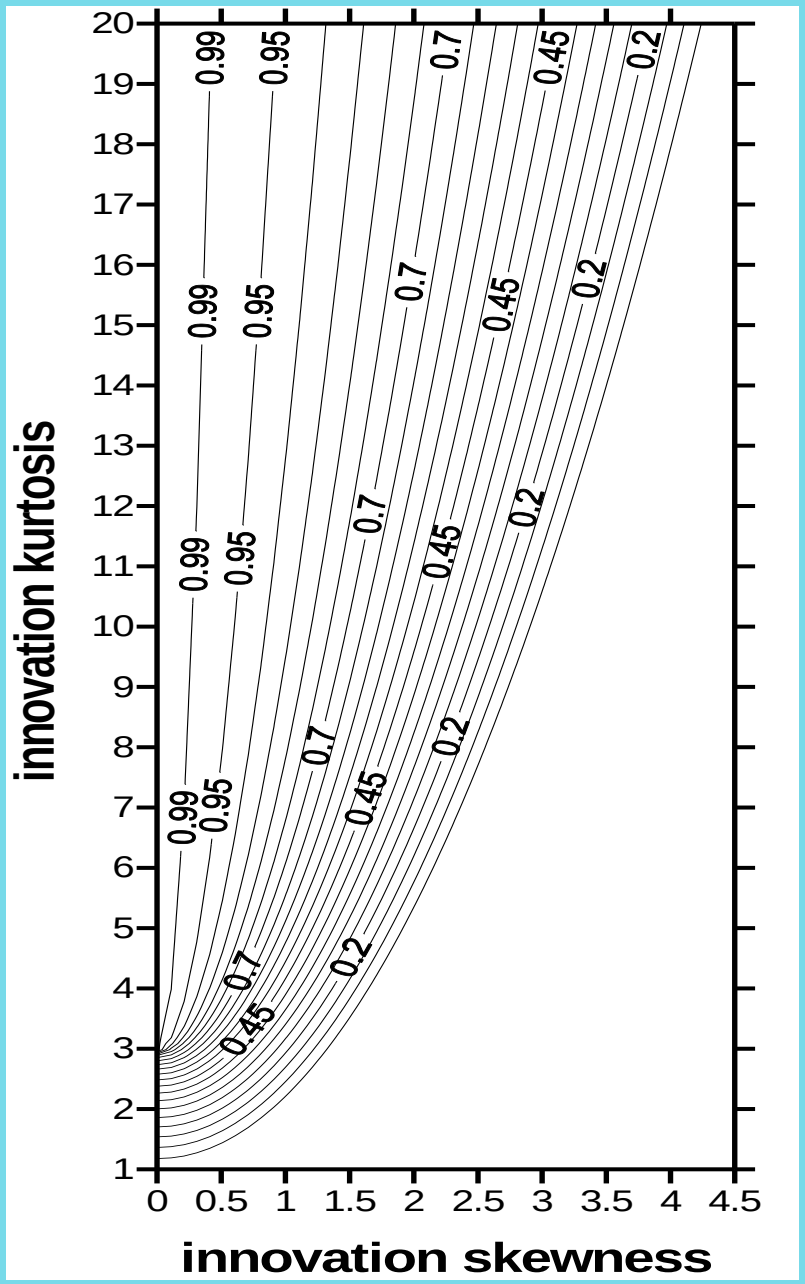
$$\sigma_d = 0.184\,K$$

$$s_d = -0.70$$

$$k'_d = 1.02$$

$$N \sim 5709$$

Histogram and pdf fit of the Innovations of Brightness Temperature. Bg taken from ECMWF forecasts. Note the negatively skewed, leptokurtic distribution of innovations.

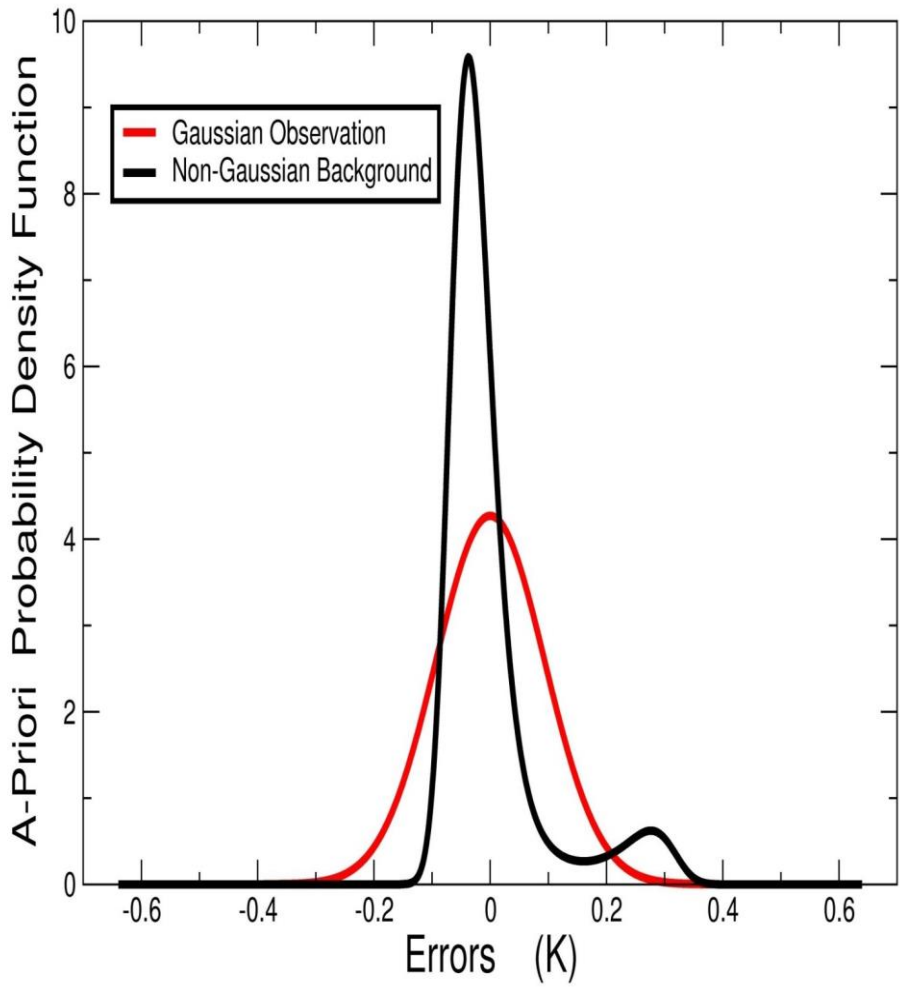


Innovations are non-Gaussian and that Non-Gaussianity may come from one or from both errors. Taking one of the errors to be Gaussian, its variance fraction admits an upper bound (in Fig.) for each value of the skewness s_d and kurtosis k_d of innovations. Therefore non-Gaussianity must be accomodated by a minimum of error variance which is a constraint for the origin of the Non-Gaussianity.

After deciding the NGty origin, and computing the 4 leading central error moments: μ=0, σ, s and k, consistent with innovation statistics, we must build the correspondent Maximum Entropy (ME) pdf. A non-linear minimization problem on Lagrange multipliers is then solved

$$\left(\hat{\lambda}_1,\hat{\lambda}_2,\hat{\lambda}_3,\hat{\lambda}_4\right) = arg\,min\left[\begin{array}{l} ln\int exp\left(\lambda_1 u+\lambda_2 u^2+\lambda_3 u^3+\lambda_4 u^4\right)du \\ -\left(\lambda_2\sigma^2+\lambda_3 s\sigma^3+\lambda_4 k\sigma^4\right)\end{array}\right]$$

$$pdf\ of\ error:\quad \rho(\varepsilon)=\frac{exp\left(\hat{\lambda}_1\varepsilon+\hat{\lambda}_2\varepsilon^2+\hat{\lambda}_3\varepsilon^3+\hat{\lambda}_4\varepsilon^4\right)}{\int exp\left(\hat{\lambda}_1 u+\hat{\lambda}_2 u^2+\hat{\lambda}_3 u^3+\hat{\lambda}_4 u^4\right)du}$$



$$\sigma_b = 0.13 \; ; \; \sigma_o = 0.13$$

$$s_b = 2.2 \quad ; s_b = 0.0$$

$$k_b = 7.7 \quad ; k_o = 3.0$$

ME pdfs of observation Gaussian errors and background non-Gaussian errors obtained by prescribing consistent error statistics σ, s, k. Note the positive skewness of background errors.

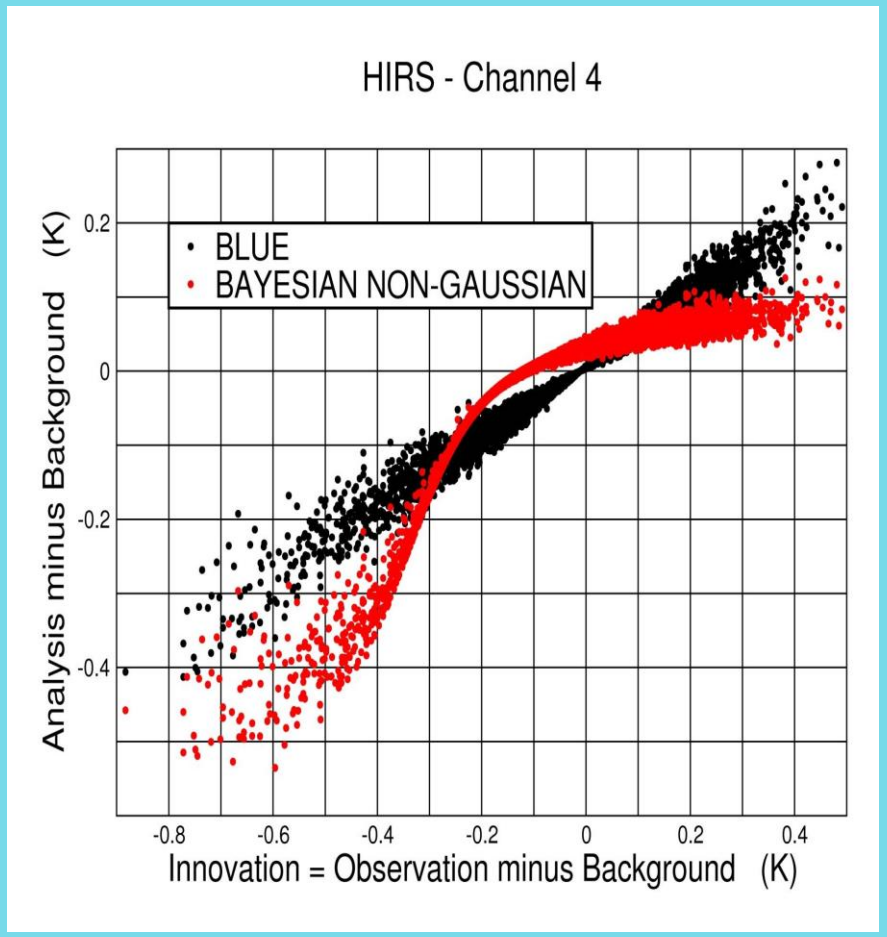
The BLUE and the non-Gaussian MMSE Bayesian estimator for a prescribed bg and innovation d and error pdfs ρ_o, ρ_b writes as:

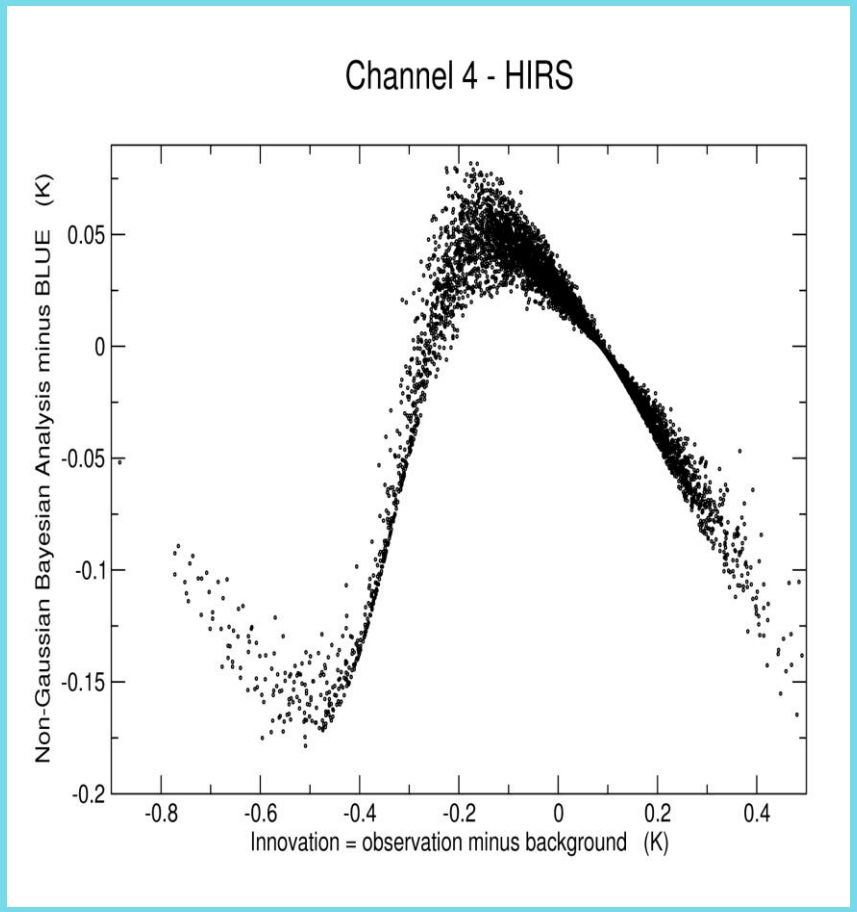
$$x_{BLUE} = x_b + \frac{\sigma_b^2}{\sigma_b^2 + \sigma_o^2} d \quad ; \quad x_{MMSE} = x_b - \frac{\int \varepsilon \rho_b(\varepsilon) \rho_o(d-\varepsilon) d\varepsilon}{\int \rho_b(\varepsilon) \rho_o(d-\varepsilon) d\varepsilon}$$

The impact of non-Gaussian errors is measured by the mean square difference between MMSE and BLUE, normalized by the a-posteriori BLUE's error variance in terms of the:

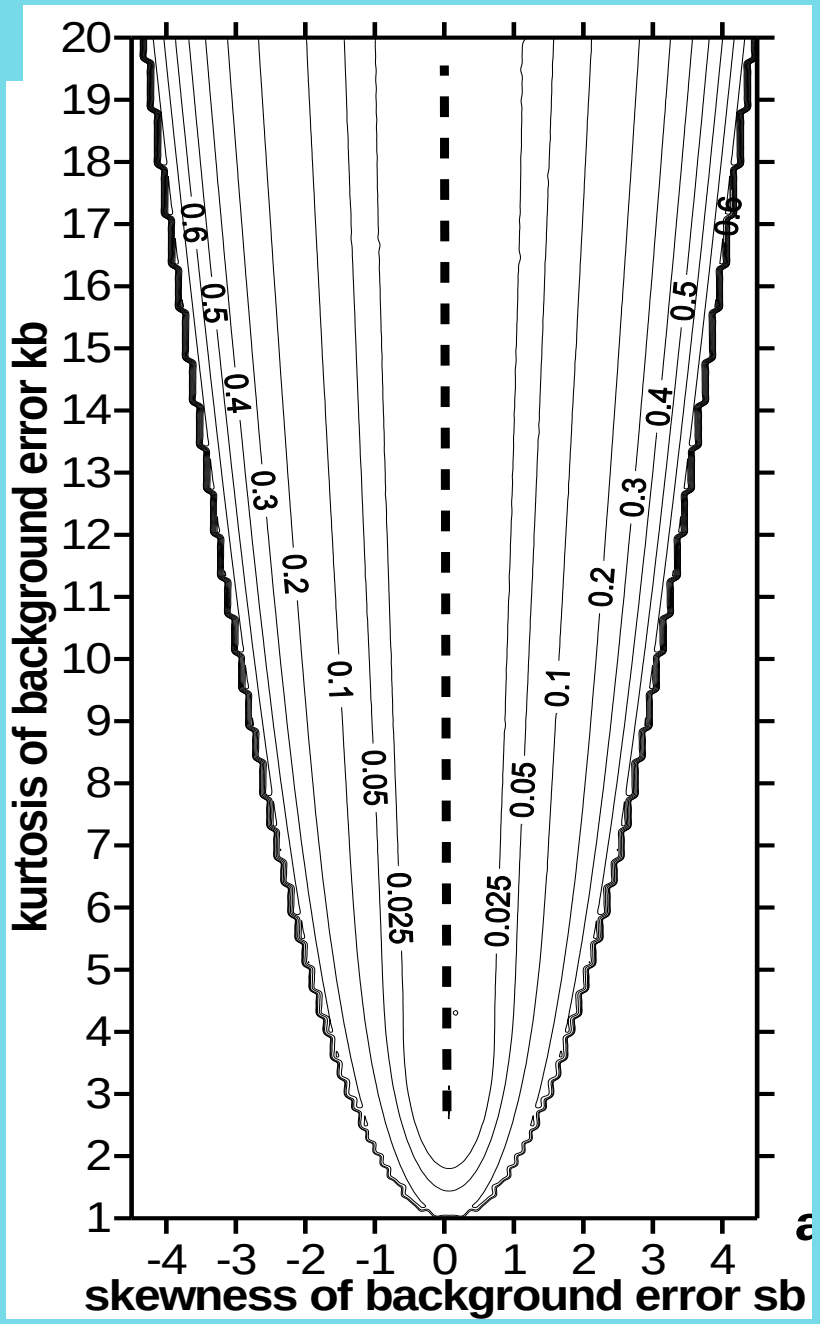
$$\text{SCORE} = \frac{\overline{\left(x_{MMSE} - x_{BLUE}\right)^2}}{\left(\sigma_o^{-2} + \sigma_b^{-2}\right)^{-1}} =$$

$$= \textit{relative potential decrease of the analysis MSE}$$





Scattered plot of the differences: BLUE-BG (black, left), MMSE-BG (red,left) and MMSE-BLUE (black,right) as a function of the innovation (5709 data). The SCORE=0.27 represents a 27% potential reduction of the mean square analysis error, specially when BG>>OBS (highly negative innovations)



What makes a large SCORE?

For observation and background errors of equal variance and innovations' non-Gaussianity coming only from one the errors, the SCORE (in Fig.) tends to mostly increase with assymmetry (skewness) of the assumed non-Gaussian error.

Physica D 239 (2010) 1701–1717

Contents lists available at ScienceDirect

Physica D

journal homepage: www.elsevier.com/locate/physd

Diagnosis and impacts of non-Gaussianity of innovations in data assimilation

Carlos A. Pires [a,*], Olivier Talagrand [b], Marc Bocquet [c,d]

[a] *Instituto Dom Luiz, University of Lisbon, Portugal*

[b] *Laboratoire de Météorologie Dynamique (LMD), École Normale Supérieure, Paris, France*

[c] *Université Paris-Est, CEREA, Joint Laboratory École des Ponts ParisTech and EDF R&D, Champs-sur-Marne, France*

[d] *INRIA, Paris-Rocquencourt Research Center, France*

Conclusions:

1 - Data assimilation executed with non-Gaussian distributed errors may lead to different analysis (posterior sates) as a function of the chosen optimality criterium (ML, MED, BLUE,MMSE)

2 -The assumption of error's Gaussianity may be tested by computing high-order statistics of the innovations

3 - The correction of the prescribed pdf of errors in the observation space may be executed by attributing consistent values of skewness and kurtosis to errors which are consistent to those of the innovations. The least committing pdf is that obtained by the Maximum Entropy method constrained by the imposed moments.

4 -The potential reduction of the analysis mean square error comes mainly from the skewness of errors and may be large for some observables and extreme innovations