# Issues of nonlinearity and non-gaussianity

## A brief tour in non-Gaussian data assimilation with a view to large geophysical systems

Marc Bocquet
(bocquet@cerea.enpc.fr)

CEREA, École des Ponts ParisTech / EDF R&D
Université Paris-Est and INRIA

Thanks to Lin Wu for his valuable suggestions.

# Outline

# Nonlinear statistical estimation: discrete approach

## Dynamics, observation and statistics

$$\mathbf{x}_{k+1} = M_k(\mathbf{x}_k) + \mathbf{w}_k \quad \text{and} \quad \mathbf{y}_k = H_k(\mathbf{x}_k) + \mathbf{v}_k$$

$$\underbrace{p(\mathbf{x}_{k+1}|\mathbf{x}_k) = p_W\left(\mathbf{x}_k - M_k(\mathbf{x}_k)\right)}_{\text{transition kernel}} \;,\; \underbrace{p(\mathbf{y}_k|\mathbf{x}_k) = p_V\left(\mathbf{y}_k - H_k(\mathbf{x}_k)\right)}_{\text{likelihood}} \quad \text{are known}$$

Smoothing approach: Given $\mathbf{X}_k = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ and $\mathbf{Y}_k = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N\}$, recursive application of Bayes and transition rules lead to

$$p(\mathbf{X}_k|\mathbf{Y}_k) = \prod_{k=1}^{K} \left[p_V(\mathbf{y}_k - H_k(\mathbf{x}_k)) p_W(\mathbf{x}_{k+1} - M_k(\mathbf{x}_k))\right] p(\mathbf{x}_0)$$

Maximum a posteriori of $\ln\left(p(\mathbf{X}_k|\mathbf{Y}_k)\right)$ defines the variational cost function.

Sequential approach (filtering problem):

- Forecast (Chapman-Kolmogorov): $p(\mathbf{x}_{k+1}|\mathbf{Y}_k) = \int \mathrm{d}\mathbf{x}_k p_W(\mathbf{x}_{k+1} - M_k(\mathbf{x}_k)) p(\mathbf{x}_k|\mathbf{Y}_k)$.
- Analysis (Bayes): $p(\mathbf{x}_k|\mathbf{Y}_k) = \dfrac{p_{\mathbf{v}}(\mathbf{y}_k - \mathbf{H}(\mathbf{x}_k)) p(\mathbf{x}_k|\mathbf{Y}_{k-1})}{\int \mathrm{d}\mathbf{x}_k p_V(\mathbf{y}_k - \mathbf{H}(\mathbf{x}_k)) p(\mathbf{x}_k|\mathbf{Y}_{k-1})}$.

# Nonlinear statistical estimation: Fokker-Plank and Zakai equations

Continuous time, state space discretized as $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}^\dagger$. Model equation

$$\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)\mathrm{d}t + \mathbf{g}(\mathbf{x}_t, t) \cdot \mathrm{d}\mathbf{w}_t \,.$$

Fokker-Planck equation for the relative probability density function ($\mathbf{Q} = \mathbf{g}_t \mathbf{g}_t^\dagger$)

$$\frac{\partial p_t}{\partial t} = -\nabla \cdot (\mathbf{f}(\mathbf{x}, t) p_t) + \frac{1}{2} \sum_{i,j} \frac{\partial^2}{\partial x_i \, \partial x_j} \left( [\mathbf{Q}]_{ij} \, p_t \right) = \mathscr{L}_{\mathrm{FP}}(p_t) \,.$$

Adding the observation equation $\mathrm{d}\mathbf{y}_t = \mathbf{h}(\mathbf{x}_t, t)\mathrm{d}t + \sqrt{\mathbf{R}}\,\mathrm{d}\mathbf{v}_t$ ,
leads to the Zakai (or normalized Kushner) equation:

$$\mathrm{d}p_t = \mathscr{L}_{\mathrm{FP}}(p_t)\,\mathrm{d}t + p_t \mathbf{h}_t^\dagger \mathbf{R}_t^{-1} \mathrm{d}\mathbf{y}_t \,.$$

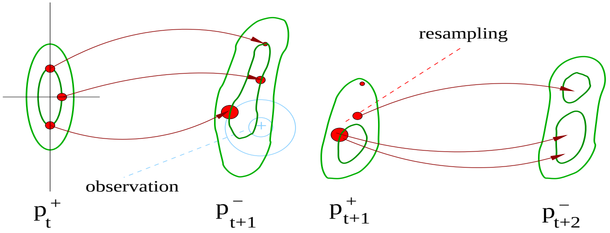
From $\mathbb{R}^N$ to $\mathscr{P}(\mathbb{R})^{\otimes^N}$, the maths exist **but the complexity is too high !**

Similar to the passage from classical to quantum physics ...
[Miller et al. 1999]

# Numerics: particle filter

Monte Carlo approaches to solve these nonlinear filtering equations are called particle filters. Most intuitive one: bootstrap filter

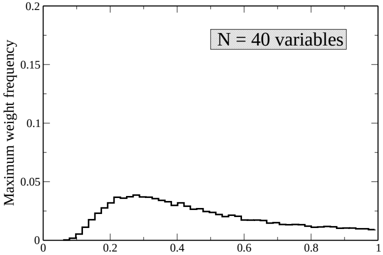
- Particles $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_I\}$ sample the pdf $p_t(\mathbf{x})$: $p_t(\mathbf{x}) \simeq \sum_{i=1}^{I} w_i \delta(\mathbf{x} - \mathbf{x}_t^i)$.
- Propagation of the particles trough the model: $p_{t+1}(\mathbf{x}) \simeq \sum_{i=1}^{I} w_i \delta(\mathbf{x} - \mathbf{x}_{t+1}^i)$.
- Analysis (weights altered by likelihood): $w_{t+1}^i \propto w_t^i p(\mathbf{y}_{t+1} | \mathbf{x}_{t+1}^i)$.
- When necessary, resampling of the ensemble, using the unbalanced weights $w_i$.



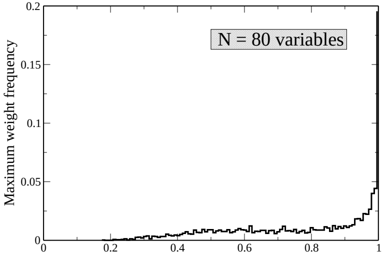

[Handschin et al. 1969, Gordon et al. 1993, Van Leeuwen 2002, Zhou 1996]

# Curse of dimensionality

Problem: particle filters work fine up to $N \sim 4-8$. When the state space and/or the observation space get bigger, degeneracy/collapse of the weights: only one particle remains likely. This implies a failure of the filter as a modal estimator.
Resampling helps but is not solving the issue.




Balanced Lorenz-96 particle filter

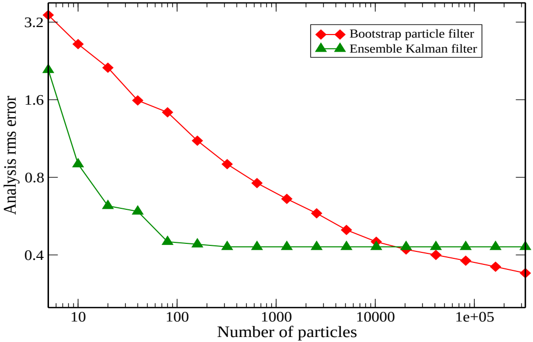



Degenerate Lorenz-96 particle filter

The required ensemble size scales exponentially with the state space size $N$, the observation space size or the innovation variance.
[Snyder, Bengtsson et al. 2007-2008]

# Gaussian as *maskeshift* ?

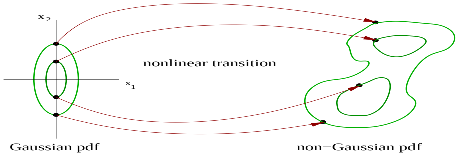
Lorenz-96 ($N = 10$ variables, $F = 8$) experiment




- Still, Gaussian estimation leads to a complexity of $\mathscr{P}(\mathbb{R}) \otimes \mathscr{P}(\mathbb{R})$.
- Mathematically tractable (if the full covariance matrix is not made explicit).
- Supported by central limit theorem.
- Least committed distribution when only first and second moments are known.

# Outline

# Sources of non-gaussianity in a geophysical context




## Nonlinearities in models generate non-Gaussian pdfs

- Nonlinearity of Navier-Stokes leading to chaos, thresholds (cloud, rain), chemistry, increase in resolution (precipitation at convective scale), etc.
- Observation operator model.

## Non-Gaussian priors sometimes more adequate description

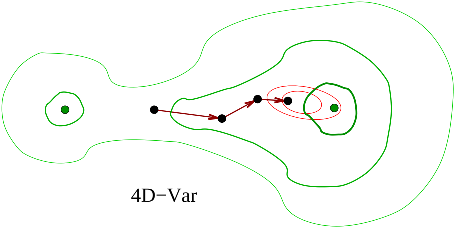
- Background information in state/control space: humidity (Gaussian anamorphosis though), emission inventories in atmospheric chemistry.
- Observation error prior: Huber norm, combination of $l_1$ and $l_2$ (Gaussian + account for gross errors), $l_\infty$ norm, log-normal and multiplicative errors, etc.
- Advanced model error prior.

# Dealing with nonlinearity in a Gaussian framework

The priors can be assumed Gaussian, but the models remain nonlinear [Gauthier 1992, Stensrud et al. 1992, Miller et al. 1994, Pires et al. 1996], and it must be dealt with . . .

## 4D-Var solutions to deal with nonlinearity

- Risk: Gaussian-based Bayesian estimation may rigorously lead to multimodal distribution, whenever nonlinear operators involved.
- Fixes: Outer loop to enforce the full (high-res) nonlinear model. Inner loop to warranty fast optimization (conjugate gradient), and (local) uniqueness of minimum.

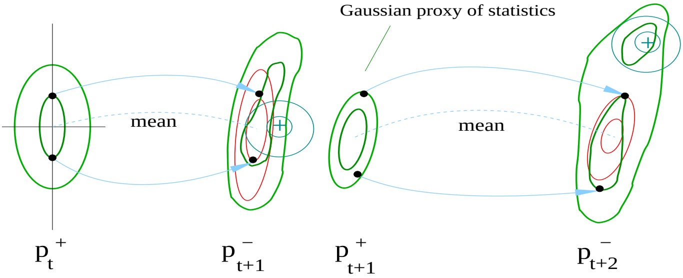


# Dealing with nonlinearity in a Gaussian framework

The priors can be assumed Gaussian, but the models remain nonlinear . . .

## EnKF solutions to deal with nonlinearities

- Ensemble encodes all statistics. Ensemble propagated by model without proxy.
- Fixes: Ensemble statistics assumed Gaussian (a priori and a posteriori, even though they may not be) so has to keep ensemble coherence.




Gaussian ensemble filter

# Measuring non-Gaussianity: how much do we loose being Gaussian ?

## Relative entropy

- Fundamental measure of the discrepancy between two pdfs: relative entropy

$$\mathscr{K}(p,q) = \int \mathrm{d}p \ln \frac{p}{q}\,.$$

Geophysical applications in predictibility [Kleeman 2002], in statistics of geophysical dynamical systems [Majda], in inverse modeling [Bocquet 2005], in modeling of prior pdfs [Eyink et al., Pires et al., 2004-2008].

- Difficult to handle in high-dimensional systems.
- $p =$ prediction or analysis uncertainty pdf.
- $q =$ Gaussian proxy of the pdf with the same first and second-order moments.

## Gram-Charlier/Edgeworth expansions of $\mathscr{K}$

Gram-Charlier/Edgeworth expansion of $p/q$, leads to (skewness and kurtosis order)

$$\mathscr{K}(p,q) \underset{\text{Gra.}}{\simeq} \frac{1}{12}\sum_{i,j,k}(\kappa_{i,j,k})^2 + \frac{1}{48}\sum_{i,j,k,l}(\kappa_{i,j,k,l})^2 \qquad \mathscr{K}(p,q) \underset{\text{Edg.}}{\simeq} \frac{1}{12}\sum_{i,j,k}(\kappa_{i,j,k})^2 + O\left(\frac{1}{I^{3/2}}\right)$$

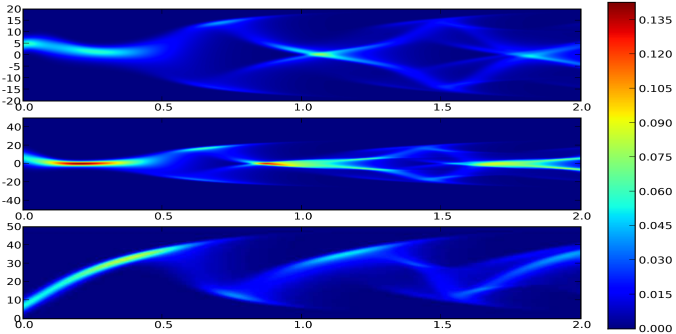
where $\kappa_{i_1,i_2,\ldots,i_n}$ are the *standardized* cumulants of $p$ of order $n$.

# Measuring non-Gaussianity: how much do we loose being Gaussian ?

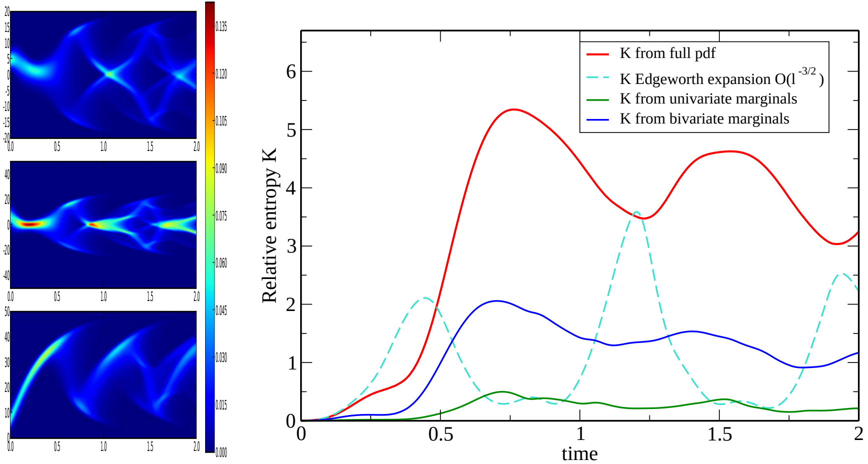
## Multivariate test of normality

- Numerous various test of normality (univariate): Kolmogorov-Smirnov, Anderson-Darling, Shapiro-Wilk test.
- Multivariate case: a few test, difficult to handle for large sample size and large number degrees of freedom.
- Necessary but insufficient test: comparing the Mahalanobis norm of members to a $\chi^2$ law, using an univariate null-hypothesis test, marginals of pdf, ...

Lorenz63
Free run
Initially: Gaussian ensemble with $\sigma = 0.1$



# Measuring non-Gaussianity: how much do we loose being Gaussian ?



# Reducing nonlinearity impact: divide and conquer

With finer discretizations, nature becomes Gaussian (as long as it becomes linear) . . .

## Adaptive data assimilation

- Assimilation could adapt to the varying instability of the flow. For instance, the efficient variational assimilation window length of $\tau_{\mathrm{eff}}(\mathbf{x}) \propto \lambda^{-1}(\mathbf{x})$ [Pires et al. 1996], where $\lambda(\mathbf{x})$ is the typical local Lyapunov exponent: smaller delay between analyses required.
- Identify low dimensional manifold to deploy particle filters [Berliner & Wickle 2007].

## Localizing strategies for particle filters

- A smaller number of particles for smaller areas.
- But contrary to localized EnKF, not trivial glueing of the subsequent local estimates from the analysis, . . . [van Leeuven, 2004-2008]

## Gaussian mixtures

- Many components mixture: ultimately as difficult as part. filters [Bengtsson et al., 2003].
- Can be used to estimate with a finite number of components a non-Gaussian pdf with analytically tractable estimation equations.

# Outline

# Gaussian on non-Gaussian grounds: deviation from climatology

## Maximum entropy filter [Eyink & Kim 2006]

▶ The pdf of an ensemble should be, given its mean and variance, the closest to the climatology pdf $q$. *Distance* measured by the relative entropy:

$$\mathscr{K}(p,q) = \int \mathrm{d}p \ln(p/q)\,.$$

Ensemble second-order statistics:

$$\widehat{\mathbf{y}} = \frac{1}{I}\sum_{i=1}^{I} \mathbf{H}\mathbf{x}_i \quad \text{and} \quad \widehat{\mathbf{Y}} = \frac{1}{I}\sum_{i=1}^{I} \mathbf{H}\mathbf{x}_i (\mathbf{H}\mathbf{x}_i)^{\dagger}.$$

pdf generic form: $p(\mathbf{x},\lambda,\mathbf{\Lambda}) \propto q(\mathbf{x}) \exp\left(\lambda^{\dagger}\mathbf{H}\mathbf{x} - \frac{1}{2}\mathbf{x}^{\dagger}\mathbf{H}^{\dagger}\mathbf{\Lambda}\mathbf{H}\mathbf{x}\right).$

Dual parameter estimations:

$$\lambda, \mathbf{\Lambda} = \mathrm{argmin}\left(\ln(Z(\lambda,\mathbf{\Lambda})) - \lambda^{\dagger}\widehat{\mathbf{y}} + \frac{1}{2}\mathrm{Tr}\left(\widehat{\mathbf{Y}}\mathbf{\Lambda}\right)\right)$$

▶ Assuming linear observation operator $\mathbf{H}$ and Gaussian error (observation $\tilde{\mathbf{y}}$ with error statistics $\mathbf{R}$), the pdf is updated using Bayes rule, within a dual framework.

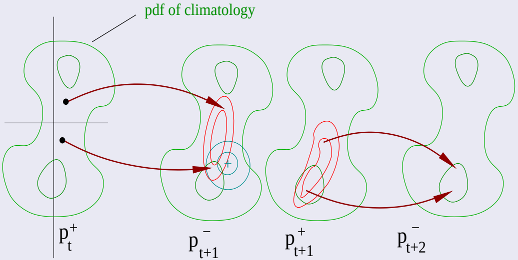
Dual parameters update:

$$\lambda^{+} = \lambda^{-} + \mathbf{R}^{-1}\tilde{\mathbf{y}} \quad \text{and} \quad \mathbf{\Lambda}^{+} = \mathbf{\Lambda}^{-} + \mathbf{R}^{-1}.$$

# Gaussian on non-Gaussian grounds: deviation from climatology

## Maximum entropy filter [Eyink & Kim 2006]

- Resampling (like in deterministic analysis filter)
- In essence: it is a dual ensemble Kalman filter upon a reference pdf given by the climatology. It is efficient on Lorenz-63.




Maximum entropy filter

Lorenz-63 analysis error r.m.s.

| $\Delta t$ | EnKF | MEF |
|---|---|---|
| 1/6 | 1.0457 | 1.7846 |
| 1/3 | 1.5034 | 1.6041 |
| 2/3 | 1.1548 | 1.0200 |
| 4/3 | 0.7212 | 0.6529 |

Eyink and Kim, J.Stat.Phys., 2006

- Degraded version (first moments only): mean-field filter.

# Non-Gaussian on Gaussian grounds: importance filtering

## Main ideas of importance sampling

- Empirical representation with a mix of particles trajectories and weights:

$$p_t(\mathbf{X}_t|\mathbf{Y}_t) \simeq \sum_{i=1}^{N} w_t^i \delta(\mathbf{X}_t - \mathbf{X}_t^i) .$$

where the particles trajectories are drawn from a known *proposal* pdf $q$.
This is possible if the weights are of the form

$$w_t^i \propto \frac{p_t(\mathbf{Y}_t|\mathbf{X}_t^i)p(\mathbf{X}_t^i)}{q_t(\mathbf{X}_t^i|\mathbf{Y}_t)} .$$

- Sequential filtering version:

$$w_t^i \propto w_{t-1}^i \frac{p_t(\mathbf{y}_t|\mathbf{x}_t^i)p_t(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i)}{q_t(\mathbf{x}_t^i|\mathbf{X}_{t-1}^i,\mathbf{Y}_t)} .$$

- If proposal $q_t(\mathbf{x}_t^i|\mathbf{X}_{t-1}^i,\mathbf{Y}_t) \triangleq p_t(\mathbf{x}_t|\mathbf{x}_{t-1})$ then this is a bootstrap filter !
- To avoid too unlikely trajectories, particles should be drawn from a proposal making use of $\mathbf{y}_t$, but this is not considered easy, unless one practices ensemble-based Kalman filters . . .

# Non-Gaussian on Gaussian grounds: importance sampling

## Observation-dependent proposal: Gaussian filters [van der Merwe et al. 2000, Papadakis 2007]

If $\overline{x}_t^i$ and $\mathbf{P}_t^i$ are mean and covariance of an ensemble-based Gaussian filter: EKF, UKF, EnKF, ETKF, etc, then

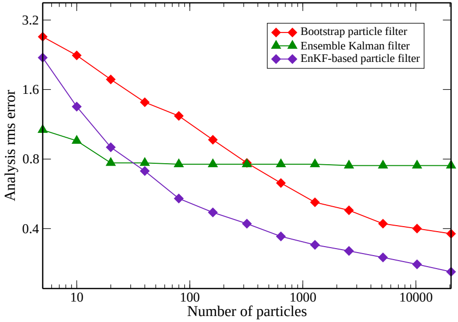
$$q(\mathbf{x}_t^i|\mathbf{X}_{t-1}^i, \mathbf{Y}_t) \triangleq N(\overline{x}_t^i, \mathbf{P}_t^i)$$
$$q(\mathbf{x}_t^i|\mathbf{X}_{t-1}^i, \mathbf{Y}_t) \triangleq N(\overline{x}_t, \mathbf{P}_t)\,.$$

Second one: kind of weighted EnKF but it's a particle filter !

Lorenz-96 experiment
$N = 5$ variables, $F = 8$

$$w_t^i \propto w_{t-1}^i \frac{p_t(\mathbf{y}_t|\mathbf{x}_t^i)p_t(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i)}{N(\overline{x}_t, \mathbf{P}_t)}\,.$$

Less particles are wasted !
It has the good asymptotics !



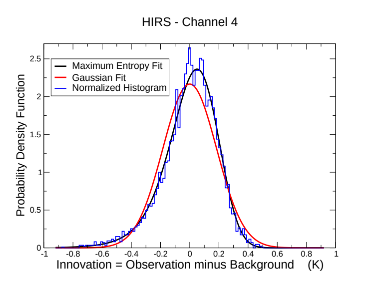
# Non-Gaussian prior construction

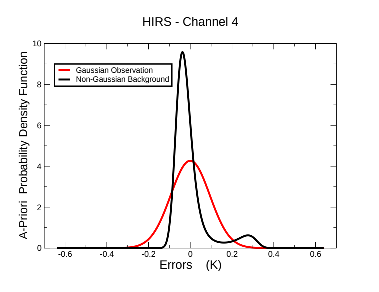
## Measuring innovation non-Gaussianity [Pires & Talagrand 2004-2008]

- Compute the deviation from gaussianity of the innovation $q = y - \mathbf{H}(x_b)$

Statistics: skewness $s = \dfrac{\mathrm{E}\left[(q-\overline{q})^3\right]}{\mathrm{E}\left[(q-\overline{q})^2\right]^{3/2}}$ and kurtosis $k = \dfrac{\mathrm{E}\left[(q-\overline{q})^4\right]}{\mathrm{E}\left[(q-\overline{q})^2\right]^2} - 3$.

- Deviation from Gaussianity estimated by a Gram-Charlier expansion (1d case)
- Compute the least committed pdf consistent with skewness and kurtosis of innovations to construct a joint prior $\nu(\varepsilon_o, \varepsilon_b)$, using the maximum entropy principle.




Innovation pdf fit [Pires & Talagrand, 2008] Priors pdf from ME

# Linear models acting on non-Gaussian priors

## Linear models

- System driven by forcing field or initial condition $\mathbf{x} \in \mathbb{R}^N$ with model/observation error $\mathbf{e} \in \mathbb{R}^p$:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{e},$$

with $\mathbf{H}$ the (up to 4D) model/observation Jacobian.

- Statistical modelling: prior pdf on controls and errors: $\nu(\mathbf{x},\mathbf{e})$, posterior pdf : $p(\mathbf{x},\mathbf{e})$.

## Bayesian inference + maximum a posteriori [Bocquet 2007]

- Primal cost function ($\equiv$ 4D-Var in Gaussian context):

$$\mathscr{L}(\mathbf{x}) = -\ln \nu(\mathbf{x}, \mathbf{y} - \mathbf{H}\mathbf{x}).$$

- If convexity proven, dual cost function ($\equiv$ PSAS in Gaussian context):

$$\widehat{\mathscr{L}}(\lambda) = (-\ln \nu)^* \left(\mathbf{H}^\dagger \lambda, \lambda\right) - \mathbf{y}^\dagger \lambda,$$

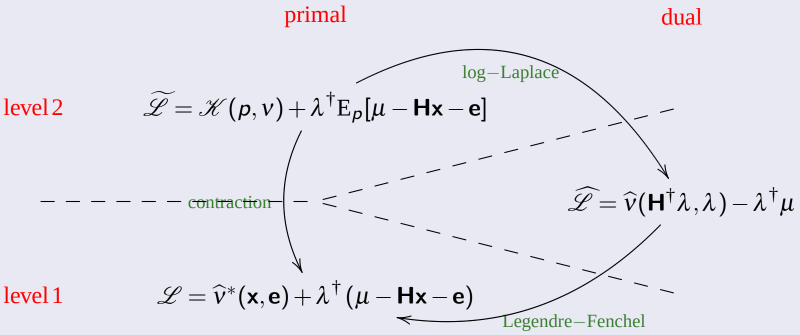
where $\nu^*$ is the Legendre-Fenchel conjugate of $\nu$.

# Linear models acting on non-Gaussian priors

Thanks to nonlinear convex analysis . . .

## Maximum entropy on the mean [Bocquet 2005-2008]

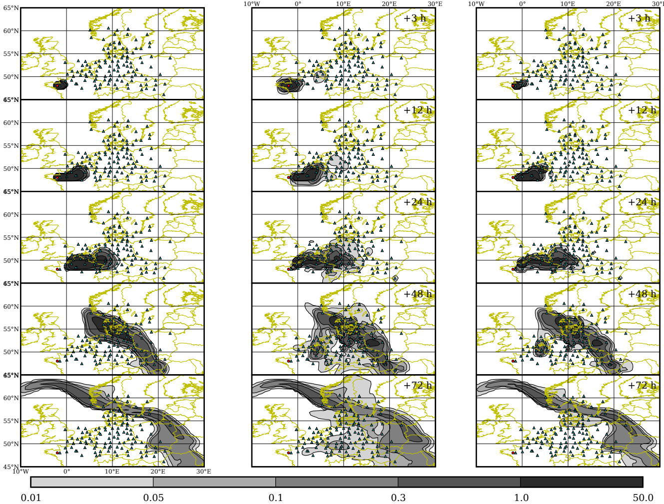
- Fully non-Gaussian generalization of 4D-Var / PSAS when models are linear




- Equivalence of all cost functions thanks to convexity.

# Linear models acting on non-Gaussian priors

Example of forecast of the ETEX-I plume ($10^3$ obs. used, $2 \times 10^5$ control variables).




Reference knowing the release / Gaussian assimilation / non-Gaussian assimilation.

# Outline

# Summary

- Fully non-Gaussian numerical solutions of estimation still not affordable.
- Mathematical tools exist that can objectively measure the departure from Gaussianity.
- Expansion (more or less affordable) around Gaussian filtering is possible
- In specific cases, and sometimes in high dimensions, non-perturbative methods are possible.

# Comments

So do we need non-Gaussian modelling after all ?

- Nonlinearity of models: nothing that will be ultimately be dealt with local in space and/or time ?
- Non-Gaussian approaches: just refinements (deviations from Gaussianity) ?
- Still need to model non-Gaussian priors (that may result from the nonlinearity of models).
- How do we measure the deviations from Gaussianity: criteria based on the flow (singular vectors, breeding modes) or uncertainty based (relative entropy, statistical tests, validation) ?
- So far, very orientated towards getting the best estimator.
  What about really getting the pdf (or higher order moments) ?
  May become a strong issue when passing from best estimate obtained from data assimilation to *best ensemble estimate* obtained from data assimilation ($\simeq$ calibration of ensemble by data assimilation).

Thank you !