Model error in geophysical data assimilation
Some (older and new) ideas

Alberto Carrassi

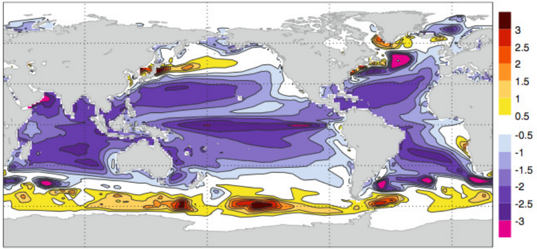
Nansen Environmental and Remote Sensing Center, Norway
Geophysical Institute, University of Bergen, Norway

UNIVERSITETET I BERGEN
Geofysisk institutt

With:
P. Ailliot (Un. Brest, FR), **M. Bocquet** (ENPC-CEREA, FR), **C. Grudzien** (UNV-Reno, USA), **M. Lucini** (Un. Reading, UK), **L. Mitchell** (Un. Adelaide, AUS), **T. Miyoshi** (RIKEN, JP), **M. Pulido** (Un. Reading, UK), **P. Raanes** (NORCE, NO), **P. Tandeo** (IMT, FR), **S. Vannitsem** (RMI, BE), **Y. Zhen** (IMT, FR).

The impact of model error

- ▶ For years model error impacts on NWP predictions was considered small compared to the (growth of) i.c. error, and thus often neglected in DA.
- ▶ The amelioration of the i.c. & the increase of the forecast horizons (seasonal-to-interannual) led to a larger impact of the model error on prediction skill.
- ▶ In DA it often manifests as underestimation of the estimate state error co-variance ⇒ **Inflation**.
- ▶ Particularly on long timescales, model error becomes evident through the emergence of biases.



- ECMWF IFS model coupled with NEMO ocean model.
- Sea surface forecast bias (Years 14–23).
- Figure from Magnusson *et al.*, 2012

Posing of the problem: Nonlinear Gaussian state-space model

It is usually assumed an HMM such as:

$$\mathbf{x}_k = \mathcal{M}_{k:k-1}(\mathbf{x}_{k-1}, \boldsymbol{\lambda}) + \boldsymbol{\eta}_k, \qquad \mathbf{y}_k = \mathcal{H}_k(\mathbf{x}_k) + \boldsymbol{\epsilon}_k. \tag{1}$$

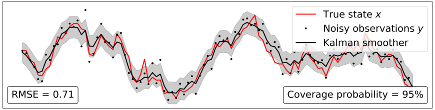
- $\mathbf{x}_k \in \mathbb{R}^m$ and $\boldsymbol{\lambda} \in \mathbb{R}^p$ are the model state and parameter vectors respectively.
- $\mathbf{y}_k \in \mathbb{R}^d$ are noisy observations related to the system's state via the, generally nonlinear, *observation operator*, $\mathcal{H} : \mathbb{R}^m \to \mathbb{R}^d$
- $\mathcal{M}_{k:k-1} : \mathbb{R}^m \to \mathbb{R}^m$ is usually a nonlinear, possibly *chaotic*, function from time t_{k-1} to t_k.
- The model and the observational errors, $\boldsymbol{\eta}_k$ and $\boldsymbol{\epsilon}_k$, are usually assumed to be uncorrelated in time, mutually independent, and Gaussian distributed: $\boldsymbol{\eta}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_k)$ and $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_k)$

- Given the multiple sources of model error a stochastic approach is generally used.
- An accurate estimate of the model error covariance, $\mathbf{Q}_k$, is necessary.

The importance of a good **Q** - 1D illustration

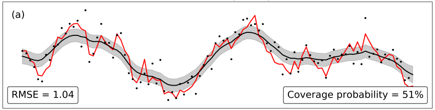
Perfect Q

Kalman smoother with $Q = Q^t$ & $R = R^t$



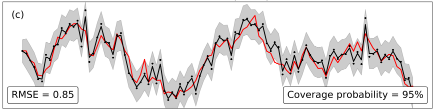
Under-estimated Q

Kalman smoother with $Q = 0.1Q^t$ & $R = R^t$



Over-estimated Q

Kalman smoother with $Q = 10Q^t$ & $R = R^t$



Univariate, linear case.

$$x_k = 0.95 x_{k-1} + \eta_k \quad (2)$$

$$y_k = x_k + \epsilon_k \quad (3)$$

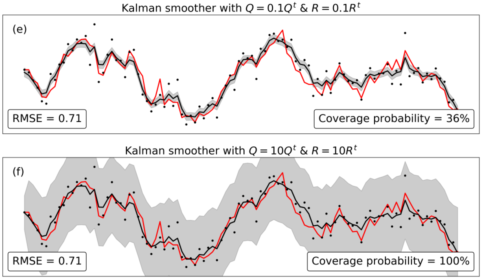
with $\eta_k \sim \mathcal{N}(0, Q^t)$ and $\epsilon_k \sim \mathcal{N}(0, R^t)$

- Promote the use of inflation.

Tandeo *et al*, 2019 - Under review

The importance of a good $||\mathbf{Q}/\mathbf{R}||$ ratio - 1D illustration

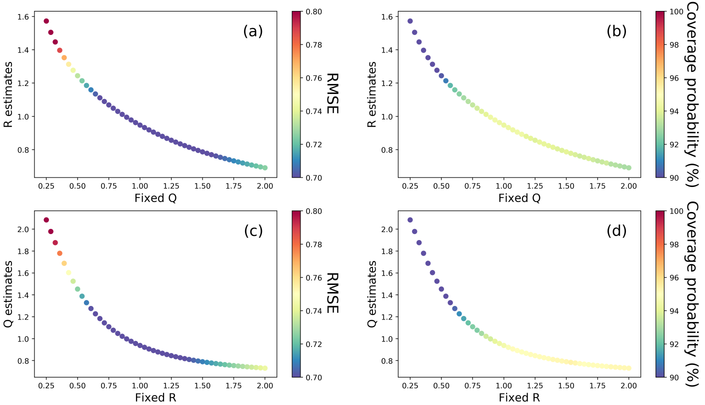
- It is the ratio Q/R that matters for the accuracy of the state estimate.



Tandeo *et al*, 2019 - Under review

- Good Q/R (no matter the individual estimates of Q and R) suffices to get good RMSE
- However it impacts differently the uncertainty quantification (*i.e.* coverage probability).

The importance of simultaneously estimating **Q** and **R** - 1D illustration



- Estimate **Q** or **R** with the Expectation Maximization (EM) (Shumway and Stoffer, 1982)
- Figure from Tandeo *et al*, 2019 - Under Review

It is not possible to fully compensate for the misrepresentation of **Q**/**R** by optimizing **R**/**Q**
$\Rightarrow$ The best is to estimate **Q** and **R** simultaneously.

Estimating $\mathbf{Q}$: key obstacles and objectives

- Large variety of possible error sources (incorrect parametrizations of physical processes, numerical discretizations, unresolved scales, etc..)
- The amount of available data insufficient to realistically describe the model error statistics, *i.e.* $\dim(\mathbf{y}) = d \ll \dim(\mathbf{x}) = m$.
- Lack of a general framework for model error dynamics (as opposed to the dynamics of the i.c. error).

What this talk is about:

1. Is the white-noise assumption always a good one?
2. Can we efficiently estimate $\mathbf{Q}_k$ along with the system state?
3. On **one** mechanism behind the need for the ultimate therapy: Inflation.

Time-correlated model error - Formulation

Let assume to have the model:

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}, \boldsymbol{\lambda})$$

used to describe the true process:

$$\frac{\mathrm{d}\hat{\mathbf{x}}(t)}{\mathbf{d}t} = \hat{\mathbf{f}}(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \boldsymbol{\lambda}^{'}) \qquad \frac{\mathrm{d}\hat{\mathbf{y}}(t)}{\mathrm{d}t} = \hat{\mathbf{h}}(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \boldsymbol{\lambda}^{'})$$

- $\hat{h}(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \boldsymbol{\lambda}^{'})$: unresolved scale; $\Delta\boldsymbol{\lambda} = \boldsymbol{\lambda}^{'} - \boldsymbol{\lambda}$ parametric error.

The evolution of the error covariance in the resolved scale:

$$\mathbf{P}(t) = < \delta\mathbf{x}_0 \delta\mathbf{x}_0^{\mathrm{T}} > + \int_{t_0}^{t} \mathrm{d}\tau \int_{t_0}^{t} \mathrm{d}\tau^{'} < [\mathbf{f}(\mathbf{x}, \boldsymbol{\lambda}) - \hat{\mathbf{f}}(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \boldsymbol{\lambda}')][\mathbf{f}(\mathbf{x}, \boldsymbol{\lambda}) - \hat{\mathbf{f}}(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \boldsymbol{\lambda}')] >^{\mathrm{T}} \tag{4}$$

- The important factor controlling the evolution is the difference between the velocity fields, the *tendencies* $\mathbf{f}(\mathbf{x}, \boldsymbol{\lambda}) - \hat{\mathbf{f}}(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \boldsymbol{\lambda})$

Time-correlated model error - Formulation

- The evolution equation for the model error covariance cannot be implemented in high dimension.
- A suitable approximation can be obtained for short-time (*e.g.* the assimilation window).

$$\mathbf{Q}(t_1,t_2) \lessapprox [\mathbf{f}(\mathbf{x},\boldsymbol{\lambda}) - \hat{\mathbf{f}}(\hat{\mathbf{x}},\hat{\mathbf{y}},\boldsymbol{\lambda}^{'})][\mathbf{f}(\mathbf{x},\boldsymbol{\lambda}) - \hat{\mathbf{f}}(\hat{\mathbf{x}},\hat{\mathbf{y}},\boldsymbol{\lambda}^{'})]^{\mathrm{T}}(t_1-t_2)^2 + O(3) \tag{5}$$

- The difference between the model and the nature tendencies, $\mathbf{f}(\mathbf{x},\boldsymbol{\lambda}) - \hat{\mathbf{f}}(\hat{\mathbf{x}},\hat{\mathbf{y}},\boldsymbol{\lambda}^{'})$ is treated as being correlated in time.
- The white-noise case would correspond to the terms $\mathbf{f}(\mathbf{x},\boldsymbol{\lambda}) - \hat{\mathbf{f}}(\hat{\mathbf{x}},\hat{\mathbf{y}},\boldsymbol{\lambda}^{'})$ being delta-correlated and the short-time evolution would be bound to be linear.

How to estimate the model-to-nature tendencies difference

Making use of the reanalysis

$\Rightarrow \mathbf{Q}_t \approx < (\mathbf{f} - \hat{\mathbf{f}})(\mathbf{f} - \hat{\mathbf{f}})^{\mathrm{T}} > t^2$

- Needs to estimate the statistics of the velocity fields discrepancy.
- Use of the **analysis increments from a reanalysis data-set** assumed to be the "truth":

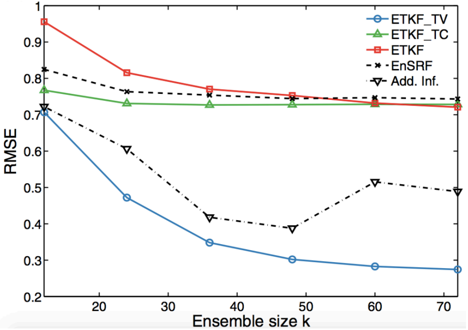
$$\mathbf{f} - \hat{\mathbf{f}} = \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} - \frac{\mathrm{d}\hat{\mathbf{x}}}{\mathrm{d}t} \approx \frac{\mathbf{x}_r^{\mathrm{f}}(t+\tau_r) - \mathbf{x}_r^{\mathrm{a}}(t)}{\tau_r} - \frac{\mathbf{x}_r^{\mathrm{a}}(t+\tau_r) - \mathbf{x}_r^{\mathrm{a}}(t)}{\tau_r} = \frac{\delta\mathbf{x}_r^{\mathrm{a}}}{\tau_r} \Rightarrow$$

$$\mathbf{Q}(t) \approx < \delta\mathbf{x}_r^{\mathrm{a}} \delta\mathbf{x}_r^{\mathrm{a}\,\mathrm{T}} > \frac{\tau^2}{\tau_r^2}$$

with τ_r reanalysis assimilation interval and τ current assimilation interval.

EnKF with short-time correlated model error

- L96 two scales. Neglect the fast scales in the model and observe 12/36 points on the coarse scale.
- ETKF (Bishop *et al*, 2001) with "best tuned" multiplicative inflation and localization (red line).
- ETKF with model error matrix $\mathbf{Q}$ estimated using the short-time approximation and the re-analysis (ETKF-TC, green line).
- ETKF with time-varying model error, randomly sampled from the reanalysis-increment statistics (ETKF-TV blue line) such that $\mathbf{x}_i^f = \mathcal{M}(\mathbf{x}_i^a) + \boldsymbol{\eta}_i \frac{\tau}{\tau_r} \qquad \boldsymbol{\eta}_k \sim \mathcal{N}(\delta\bar{\mathbf{x}}_r^a, \mathbf{Q}) \qquad i = 1, ..., N$



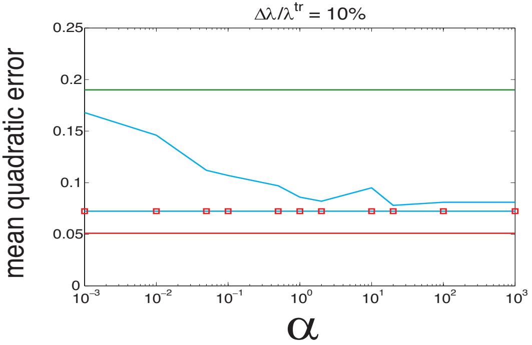
Mitchell and Carrassi, 2015

4DVar with short-time correlated model error

- Minimize the cost-function:

$$2J = \int_0^\tau \int_0^\tau (\delta \mathbf{x}_{t_1})^{\mathrm{T}} \mathbf{Q}_{t_1 t_2}^{-1} (\delta \mathbf{x}_{t_2}) \mathrm{d}t_1 \mathrm{d}t_2 + \ldots$$

- Model *Lorenz 3-variables.*
- **Strong-constraint** - Assume perfect model.
- **Weak constraint 4DVar with uncorrelated model error:** $\mathbf{Q}_t = \alpha \mathbf{B}$ (blue) or $\mathbf{Q}_t = \mathbf{Q}(t)^2$ (red marks)
- **Short-time weak constraint 4DVar with correlated model error** - $\mathbf{Q}(t_1, t_2) \approx \mathbf{Q}_0(t_1)(t_2)$



Carrassi and Vannitsem, 2010

Time-batch estimated model error covariance

- The idea (Pulido *et al*, 2018) is to maximize the log-likelihood of the data (*model evidence*) as a function of the parameter $\boldsymbol{\theta}$

$$l(\boldsymbol{\theta}) = \ln \int p(\mathbf{x}_{K:0}, \mathbf{y}_{K:1} | \boldsymbol{\theta}) \mathrm{d}\mathbf{x}_{K:0}$$

where $\boldsymbol{\theta}$ can be $\boldsymbol{\lambda}$, $\mathbf{R}$ or $\mathbf{Q}$.

- Inserting an arbitrary PDF $q(\mathbf{x}_{K:0})$ and using the Jensen inequality we have

$$l(\boldsymbol{\theta}) \geq \int q(\mathbf{x}_{K:0}) \ln \left(\frac{p(\mathbf{x}_{K:0}, \mathbf{y}_{K:1} | \boldsymbol{\theta})}{q(\mathbf{x}_{K:0})} \right) \mathrm{d}\mathbf{x}_{K:0} \equiv \mathcal{Q}(q, \boldsymbol{\theta})$$

and the equality holds when $q(\mathbf{x}_{K:0}) = p(\mathbf{x}_{K:0} | \mathbf{y}_{K:1}, \boldsymbol{\theta})$ that is the PDF maximizing $\mathcal{Q}(q, \boldsymbol{\theta})$ and a lower bound for $l(\boldsymbol{\theta})$.

- $p(\mathbf{x}_{K:0} | \mathbf{y}_{K:1}, \boldsymbol{\theta})$ can be obtained as the outcome of a DA procedure (*e.g.* EnKF, EnKS ...)

Time-batch estimated model error covariance $\mathbf{Q}$

- This suggests a two-steps algorithms:

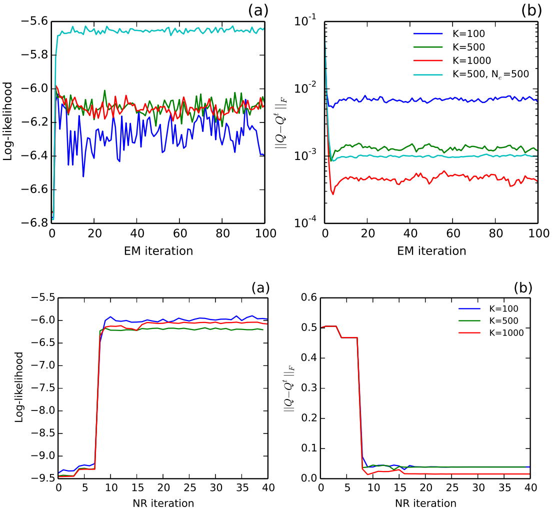
1. **Expectation**: Determine the distribution q that maximizes $\mathcal{Q}$. This is given by $q^* = p(\mathbf{x}_{K:0}|\mathbf{y}_{K:1}, \boldsymbol{\theta}^{'})$. Note that $p(\mathbf{x}_{K:0}|\mathbf{y}_{K:1}, \boldsymbol{\theta}^{'})$ is the outcome (the posterior) of a data assimilation algorithm for the HMM, evaluated at $\boldsymbol{\theta}^{'}$
2. **Maximization**: Determine the likelihood parameter $\boldsymbol{\theta}^*$ that maximizes $\mathcal{Q}(q^*, \boldsymbol{\theta})$ over $\boldsymbol{\theta}$.

We have used the EnKF to estimate $p(\mathbf{x}_{K:0}|\mathbf{y}_{K:1}, \boldsymbol{\theta}^{'})$ in combination with:

- the **expectation–maximization**, **EnKF-EM**
- the **Newton–Raphson**, **EnKF-NR**

to maximize the likelihood associated to the parameters to be estimated.

Numeric with L96 model



- The **EnKF-EM** requires the optimal value in the maximization step to be computed analytically which limits the range of its applications ⇒ Ok in a Gaussian framework, an iterative minimization in nonlinear cases.
- In the **EnKF-NR** one makes use of approximate formulae for the model evidence.
- Convergence of the NR and EM maximization as a function of the iterations for different evidencing window lengths ($K = 100, 500, 1000$).
- **(a)** Log-likelihood function.
- **(b)** Frobenius norm of the model noise estimation error.
- In about 10 iterations, they converge to a good estimation.

Pulido *et al*, 2018

However always use inflation... (better if) adaptively

- Even with a good $\mathbf{Q}$, you "always" need inflation due to sampling error and non-linearity/non-Gaussianity.
- Can avoid tuning by *adaptive inflation*; *e.g.* **EAKF-adaptive** by Anderson, 2007 or **ETKF-adaptive** by Miyoshi, 2011.
- A survey of existing methods in Raanes *et al*, 2019.
- Raanes *et al*, 2019 hybridized the "finite-size" **EnKF-*N*** (Bocquet, 2011) and the ETKF-adaptive ⇒ **EnKF-*N*-hybrid** targets explicit both sampling and model error.
- EnKF-N-hybrid yields best filter accuracy, but only by slight margin.
- See Patrick Raanes's talk tomorrow (10.35 – 11.20)

Rank-deficient filters: the *upwelling effect* and the need for inflation

- Consider a reduced-rank KF (*aka* an EnKF with $n < m$ members).

- Write the model propagator in the basis of the backward Lyapunov vectors (BLVs) using the QR decomposition

$$\mathbf{M}_k = \mathbf{E}_k \mathbf{U}_k \mathbf{E}_k^{\mathrm{T}}, \quad \mathbf{E}_k = (\mathbf{E}_k^{\mathrm{f}}\ \mathbf{E}_k^{\mathrm{u}}) \text{ with } \mathbf{U}_k = \begin{pmatrix} \mathbf{U}_k^{\mathrm{ff}} & \mathbf{U}_k^{\mathrm{fu}} \\ 0 & \mathbf{U}_k^{\mathrm{uu}} \end{pmatrix}$$

and partition the error into **filtered**/**unfiltered** variables $\boldsymbol{\epsilon}_k = \mathbf{E}_k^{\mathrm{f}} \boldsymbol{\epsilon}_k^{\mathrm{f}} + \mathbf{E}_k^{\mathrm{u}} \boldsymbol{\epsilon}_k^{\mathrm{u}}$

- The error in the filtered space ("seen" by DA) is given recursively by

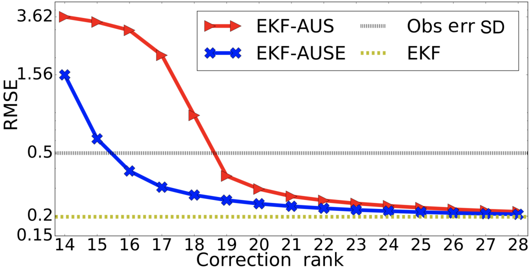
$$\boldsymbol{\epsilon}_{k+1}^{\mathrm{f}} = (\mathbf{U}_{k+1}^{\mathrm{ff}} - \mathbf{U}_{k+1}^{\mathrm{ff}} \mathbf{K}_k \mathbf{H}_k \mathbf{E}_k^{\mathrm{f}}) \boldsymbol{\epsilon}_k^{\mathrm{f}} - \mathbf{U}_{k+1}^{\mathrm{ff}} \mathbf{K}_k \boldsymbol{\epsilon}_k^{\mathrm{obs}} + \boldsymbol{\eta}_k^{\mathrm{f}} + (\mathbf{U}_{k+1}^{\mathrm{fu}} - \mathbf{U}_{k+1}^{\mathrm{ff}} \mathbf{K}_k \mathbf{H}_k \mathbf{E}_k^{\mathrm{u}}) \boldsymbol{\epsilon}_k^{\mathrm{u}}$$

- **The terms in black** correspond to the usual KF-like recursion.

- **The terms in red** disappear when the filtered subspace is the entire state space ($n = m$).

Model error and chaos: the *upwelling effect* and the need for inflation

- When $n < m$, they represent the **dynamical upwelling** of the unfiltered error into the filtered variables [Grudzien *et al* 2018].
- It moves uncertainty from unfiltered to filtered subspace, *i.e.* from the stabler to the unstable subspace.
- This phenomenon **occurs whenever** $n < m$, but is **exacerbated by model error**.
- Leads to underestimating the error in the (En)KF $\Rightarrow$ Need for **inflation** to prevent divergence.



- L96 one-scale, $m = 40$, $n_0 = 14$.
- **EKF** solves the *full-rank* recursion.
- **EKF-AUS** solves the *low-rank* $(n = n_0)$ recursion **without upwelling** (black terms only).
- **EKF-AUSE** solves the *low-rank* recursion **with upwelling** (black+red terms).

Conclusion

- Treating model error as stochastic noise is convenient and coherent with the Bayesian formulation.
- But in many real problems (*e.g.* climate science) it is actually time-correlated and its impact grows with the prediction horizon.
- A time-correlated (deterministic) model error approach has been introduced [Carrassi and Vannitsem, 2016].

- *On-the-fly* estimating the model error covariance matrix $\mathbf{Q}$ is extremely difficult in high-dimension.
- State-augmentation does not work well because the model error component of the error covariance is bound to monotonically decrease with time.
- A new method, based on the computation *the model evidence* is introduced [Pulido *et al*, 2018].
- The method requires the computation of the posterior that can be obtained (under Gaussian hypothesis) using EnKF, EnKS.

- Inflation is always needed to cope with non-Gaussianity and sampling error, but also for not-optimal $\mathbf{Q}$.
- We have demonstrated how in reduced rank filters model error is upwelled from unfiltered to filtered subspace causing error under-estimation and motivating the use of inflation [Grudzien *et al*, 2018].
- An extension of the EnKF-N originally devised for sampling error has been introduced to simultaneously deal with sampling and model error [Raanes *et al*, 2019].

Bibliography

- Carrassi, A. and S. Vannitsem, 2010. Accounting for model error in variational data assimilation. A Deterministic Formulation. *Mon. Weather. Rev.*, **138**, 3369-3386
- Carrassi, A. and S. Vannitsem, 2016: Deterministic treatment of model error in geophysical data assimilation. *Book Chapter in the book "Mathematical Paradigms of Climate Science"*, Springer. INdAM Series 15.
- Grudzien, C., A. Carrassi and M. Bocquet, 2018. Chaotic dynamics and the role of covariance inflation for reduced rank Kalman filters with model error. *Nonlin. Proc. Geophys.*, **25**, 633-648.
- Mitchell, L. and A. Carrassi, 2015. Accounting for model error due to unresolved scales within ensemble Kalman filtering. *Q. J. Roy. Meteor. Soc.*, **141**, 1417–1428
- Pulido, M., P. Tandeo, M. Bocquet, A. Carrassi and M. Lucini, 2018. Stochastic parametrization identification using ensemble Kalman filtering combined with expectation-minimization and Newton-Raphson maximum likelihood methods. *Tellus*, **70**, 1442099
- Raanes, P., M. Bocquet, and A. Carrassi, 2019. Adaptive covariance inflation in the ensemble Kalman filter by Gaussian scale mixtures. *Q. J. R. Meteorol Soc.*, **145**, 53–75.
- Tandeo, P., P. Ailliot, M. Bocquet, A. Carrassi, T. Miyoshi, M. Pulido and Y. Zhen, 2019. Joint Estimation of Model and Observation Error Covariance Matrices in Data Assimilation: a Review. Submitted. Available at https://arxiv.org/pdf/1807.11221.pdf