Parameter estimation beyond the augmented approach

Expectation-Maximization algorithm

Manuel Pulido

University of Reading, UK

On leave from UNNE, Argentina

thanks to: J. Ruiz, G. Scheffler (UBA, Argentina), P. Tandeo (IMT, France), M. Bocquet (CEREA, France), A. Carrasi (NERSC, Norway), PJ van Leeuwen (UReading, UK)

Data assimilation has been rather succesfull in estimating the state of the system.

DA may play a role in model development and process-oriented understanding which is still largely unexplored and underdeveloped.

The talk is biased toward Ensemble Kalman filtering

Potential applications

1. Parameter estimation (Annan et al 2003; Aksoy et al 2006; Tong and Xue, 2008; Ruiz et al 2013, among others)
2. Model error treatment (Ruiz and Pulido, 2015)
3. Model or parameterization development (Lang and van Leeuwen, 2016; Pulido et al 2016)
4. Process-oriented understanding based on model-observations (Scheffler and Pulido 2017)
5. Model and observational error covariance estimation (Ueno et al. 2015, Dreano et al 2017). Localization, inflation estimation.

Augmented state approach

Offline estimation with pseudo-observations

Maximum likelihood methods

Definition - Scope of the term

What do I refer to as “parameters”?

- Deterministic parameters of the dynamical model, e.g. those related to a physical parameterization.
- Stochastic parameters of physical parameterizations.
- Statistical parameters related to observational or model error covariances. Inflation factors. Localization parameters.
- Initial apriori density or background parameters (e.g. mean and/or covariance) $p_0(x)$, $\mathbf{x}^b$, $\mathbf{B}$.

Augmented state approach

$$\mathbf{x}_{ag} = [\mathbf{x}, \mathbf{x}_p], \qquad \mathcal{H}_{ag} = [\mathcal{H}(\cdot), \mathbf{0}]$$

In general, they are assumed constant during model evolution, $\mathbf{x}_{pk} = \mathbf{x}_{p(k-1)}$.

$$\mathcal{M}_{ag} = \begin{bmatrix} \mathcal{M}(\cdot) & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

Each ensemble member has a different set of parameter values.

Do not require any changes in the assimilation code. Any EnKF can be used.

$$\mathbf{x}_p^a = \mathbf{x}_p^f + \mathbf{K}_p(\mathbf{y} - \mathcal{H}(\mathbf{x}^f))$$

$$\mathbf{K}_p = \mathbf{P}_{px}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{P}_{xx}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{\mathrm{T}}$$

$\mathbf{P}_{px}$ propagates observational information to the parameters.

Why is more challenging parameter than state estimation?

- They are non-observed.

  We can only infer them through state-parameter covariances.

- They are static.

  Their impact diminishes with time.

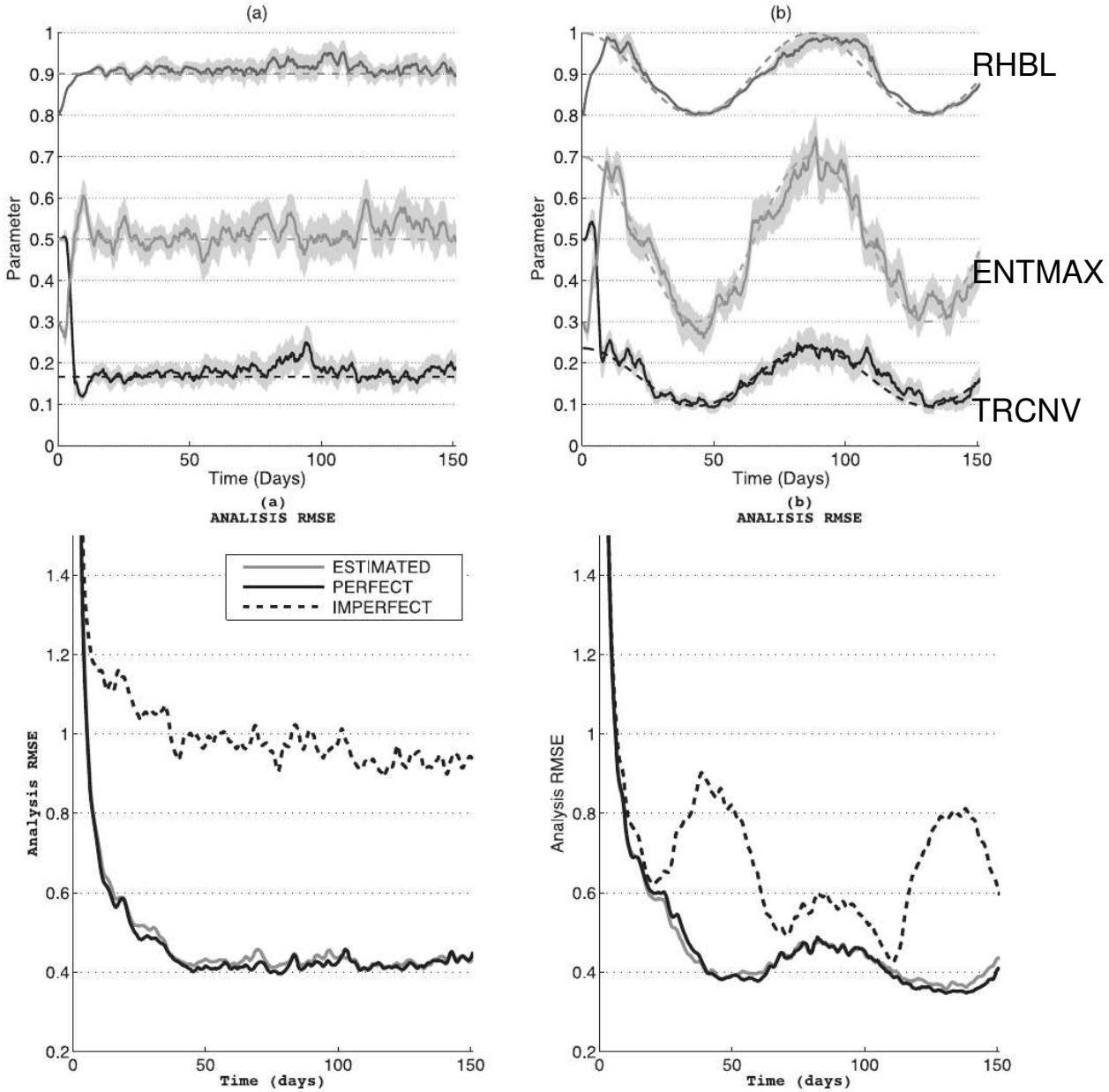
- Parameter posterior distribution are highly nongaussian.

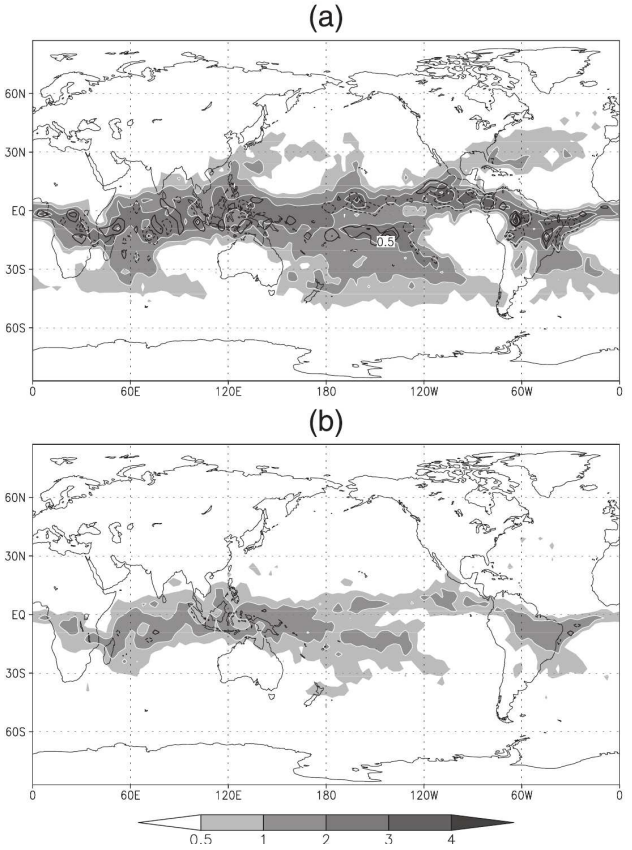
  They usually have a highly nonlinear response (in state variables).

  They are constrained (e.g. only positive parameters make physical sense).

- Global parameters are incompatible with localization.

Convective parameters in a simple GCM



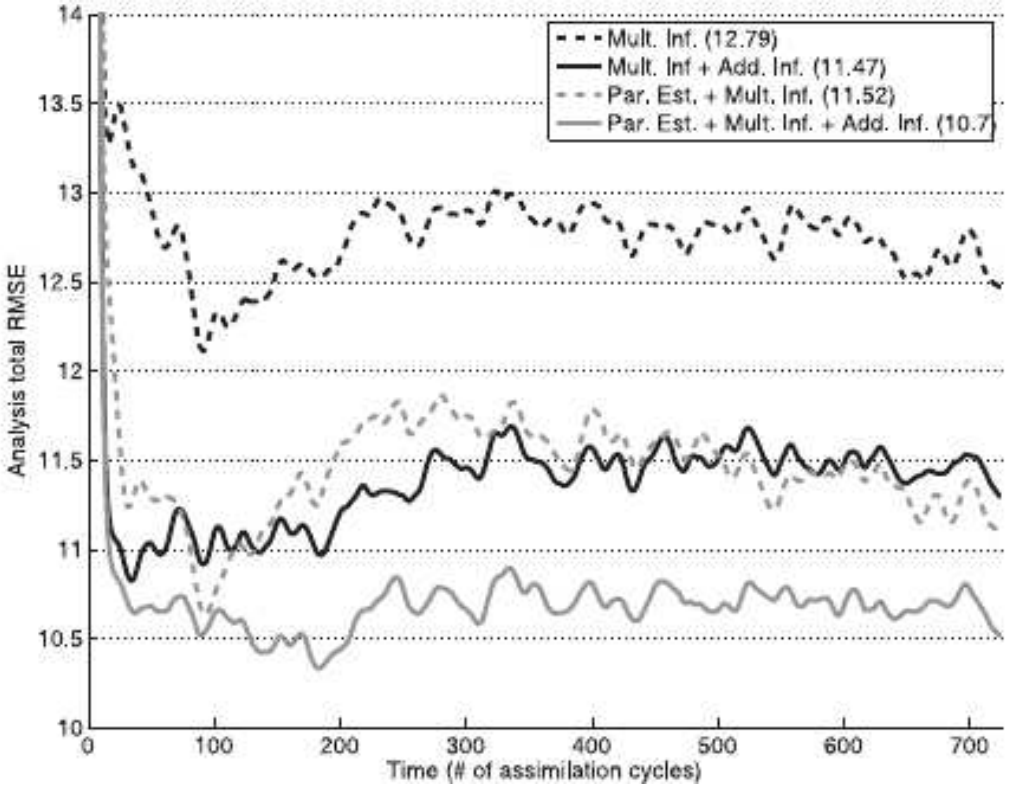


Twin experiments using LETKF and SPEEDY model.

From Ruiz et al (2013)

No need of localization for global parameters. Independent parameter inflation treatment.

Model error treatment with convective parameters



New unstable subspaces could be explored with augmented approach.

From Ruiz and Pulido (2015)

Two caveats: Convective parameters are fixing other sources of model error.

(Apriori) parameter constraints to realistic values are essential.

Offline approach

$$\begin{aligned}
\mathbf{x}_t - \mathcal{M}(\mathbf{x}_{t-1}) &= I_{SL}(\mathbf{x}_{t-1}, \mathbf{x}^S_{t-1}) && \rightarrow \text{interaction term. } \mathbf{x}^S_{t-1} \text{ not resolved.} \\
&= \mathcal{P}(\mathbf{x}_{t-1}, \mathbf{x}_p) && \rightarrow \text{representation by a parameterization} \\
&= \mathbf{F}(t_k) && \rightarrow \text{representation in the assimilation}
\end{aligned}$$

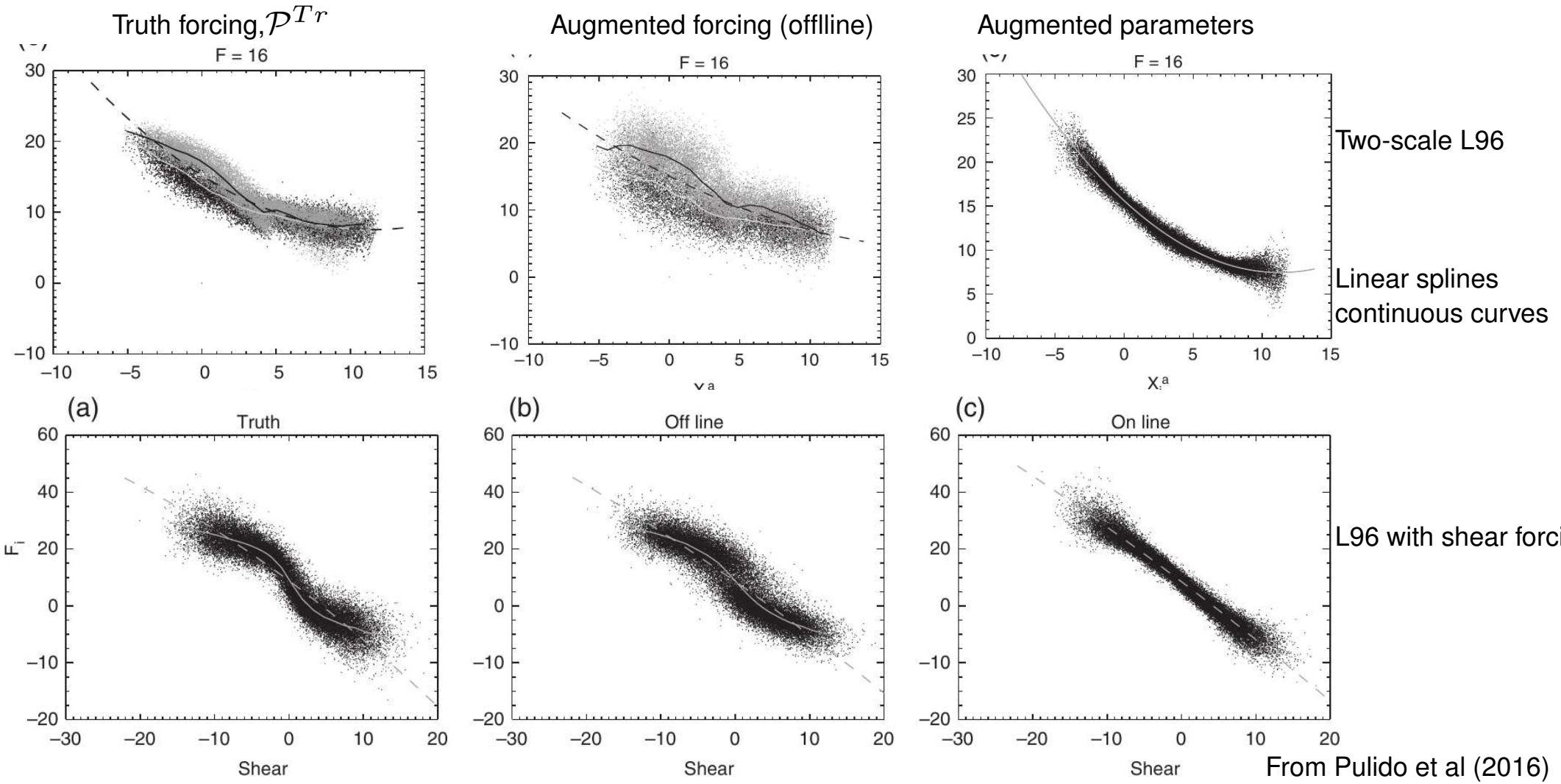
$\mathbf{F}$ would be representing sources of missing momentum or heat. In the aumented state approach, they will be estimated by $\mathbf{x}_{ag} = [\mathbf{x}, \mathbf{F}]$.

Once $\mathbf{F}^a_k$ are estimated, they could be interpreted as pseudo-observations. Suitable for Bayesian or deterministic inverse techniques:

$$J(\mathbf{x}_p) = \sum_k \left\| \mathbf{F}^a_k - \mathcal{P}(\mathbf{x}^a_k, \mathbf{x}_p) \right\|^2_{\mathbf{A}} + \left\| \mathbf{x}_p - \mathbf{x}^b_p \right\|^2_{\mathbf{B}_p}$$

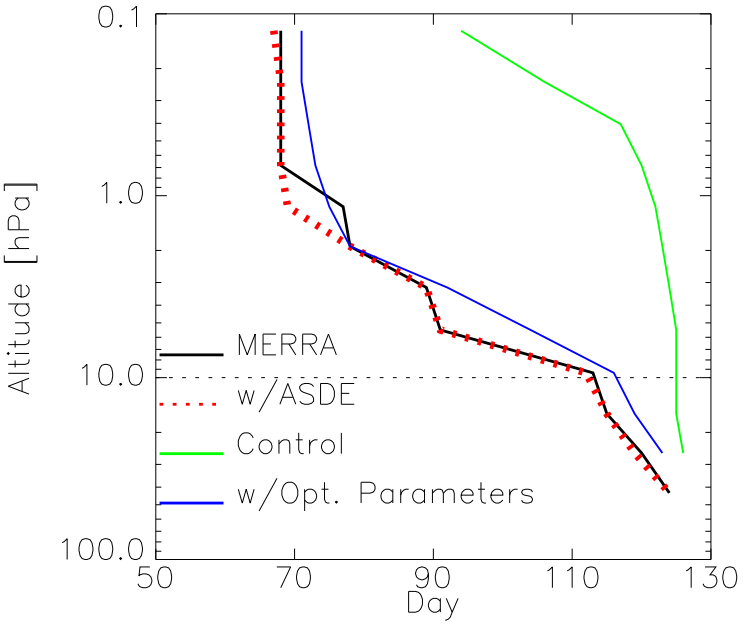
Central assumption: Short assimilation cycles.

Offline approach: proof-of-concept



From Pulido et al (2016)

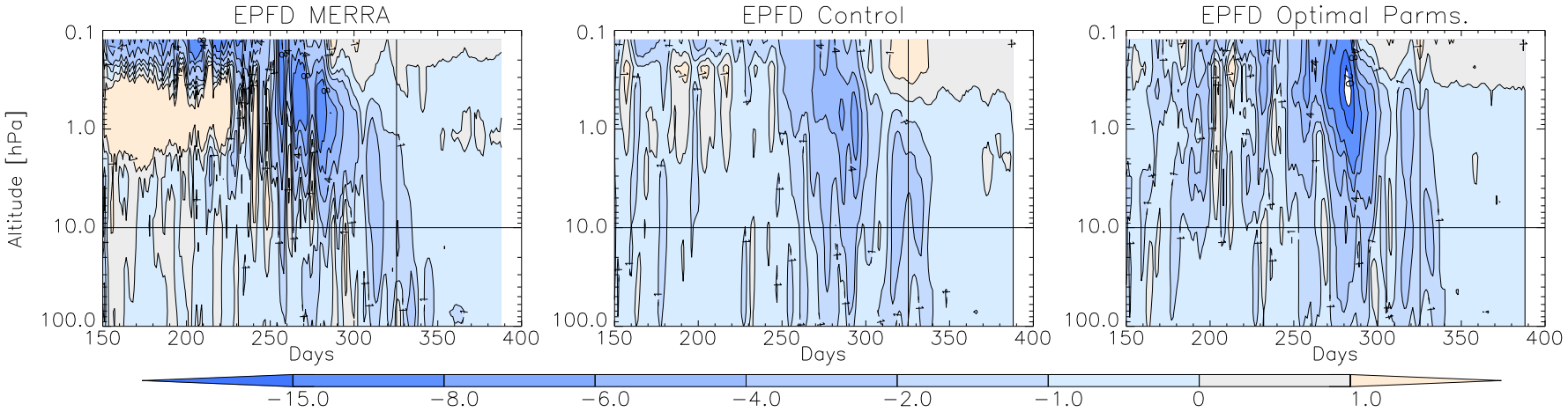
Offline approach in the middle atmosphere



Estimation of gravity wave drag parameters from missing momentum. 4Dvar + Genetic algorithm.

Optimal parameters improve substantially the delay in the vortex breakdown.

The approach can capture the complex resolved-parameterization (unresolved wave) interactions.



From Scheffler and Pulido (2017)

A proposed method to estimate covariances:

Maximum likelihood estimation in hidden Markov models given a batch of observations.

Estimation of model error covariances

Can we estimate with the augmented state approach model error covariances?

Delsole and Yang, 2010:

"The lack of dependence between mean forecast and parameter implies that the mean forecast cannot be used to infer the value of an additive stochastic parameter in a linear model."

In a first order AR model,

$$x_k = \phi x_{k-1} + \sigma\eta, \qquad \eta \sim \mathcal{N}(0,1)$$

$$cov(x_k^f, \sigma_k^f) = \phi\, cov(x_{k-1}^a, \sigma_{k-1}^a)$$

Since $\phi < 1$ then $cov(x_k^f, \sigma_k^f) \to 0$.

Along the same lines, Law and Stuart (2012) show that EnKF does not give a good measure of the uncertainty.

Problem statement

Given a Hidden Markov model,

$$\mathbf{x}_k = \mathcal{M}_{\mathbf{\Omega}}(\mathbf{x}_{k-1}) + \boldsymbol{\eta}_k,$$
$$\mathbf{y}_k = \mathcal{H}(\mathbf{x}_k) + \boldsymbol{\epsilon}_k,$$

where we assume errors are Gaussian $\boldsymbol{\eta}_k = \mathcal{N}(0, \mathbf{Q}_k)$ and $\boldsymbol{\epsilon}_k = \mathcal{N}(0, \mathbf{R}_k)$. The additive assumption is not essential.

Given a set of observations distributed in time, we want to estimate the initial prior distribution $p(\mathbf{x}_0)$, and/or the observation error covariance $\mathbf{R}_k$, the model error covariance $\mathbf{Q}_k$, and/or deterministic and stochastic physical parameters $\mathbf{\Omega}$ of $\mathcal{M}$

Maximum likelihood estimation

We want to find a set of parameters of the densities involved in the HMM that maximizes the observation likelihood under the presence of a hidden state

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

where $p(\mathbf{x}_{0:K}, \mathbf{y}_{1:K}; \boldsymbol{\theta})$ is the joint density.

We seek for the parameter values that give the most likely density.

Note that we are thinking in a batch of observations from t_1 to t_K.

Any (efficient) optimization method could be used. One of the most used in latent variable models is the algorithm of Expectation-Maximization (Dempster, et al 1977).

Expectation-Maximization principles

Suppose a proposal density $q(\mathbf{x}_{0:K})$

$$
\begin{aligned}
l(\theta) &= \ln \int q(\mathbf{x}_{0:K}) \frac{p(\mathbf{x}_{0:K}, \mathbf{y}_{1:K}; \boldsymbol{\theta})}{q(\mathbf{x}_{0:K})} \mathrm{d}\mathbf{x}_{0:K} \\
&\geq \int q(\mathbf{x}_{0:K}) \ln \left(\frac{p(\mathbf{x}_{0:K}, \mathbf{y}_{1:K}; \boldsymbol{\theta})}{q(\mathbf{x}_{0:K})} \right) \mathrm{d}\mathbf{x}_{0:K} \equiv \mathcal{P}(q, \theta).
\end{aligned}
$$

where $\mathcal{P}(q, \theta)$ is the intermidiate function (Neal and Hinton, 1999). It depends on the proposal density and on the parameters.

Expectation Step

Given that $p(\mathbf{x}_{0:K}, \mathbf{y}_{1:K}; \boldsymbol{\theta}) = p(\mathbf{x}_{0:K}|\mathbf{y}_{1:K}; \boldsymbol{\theta})p(\mathbf{y}_{1:K}; \boldsymbol{\theta})$,

$$\mathcal{P}(q, \theta) = -D_{KL}(q|p(\mathbf{x}_{0:K}|\mathbf{y}_{1:K}; \boldsymbol{\theta})) + l(\boldsymbol{\theta})$$

where $D_{KL}(q|p)$ is the Kullback-Leibler divergence between the densities.

Since $D_{KL}(q|p) = 0$ iff $q = p$, then $q = p(\mathbf{x}_{0:K}|\mathbf{y}_{1:K}; \boldsymbol{\theta})$ gives an upper bound of $\mathcal{P}(q, \boldsymbol{\theta})$.

The expectation step in practice involves finding the posterior density $p(\mathbf{x}_{0:K}|\mathbf{y}_{1:K}; \boldsymbol{\theta}')$ assuming the set of parameters are known.

The conditioning is over the whole set of observations $\rightarrow$ a smoother is required!

Maximization Step

Now we want to maximize the intermediate function as a function of $\boldsymbol{\theta}$ for a fixed $\hat{q}$, $\mathcal{P}(\hat{q}, \theta)$,

$$\int p(\mathbf{x}_{0:K}|\mathbf{y}_{1:K};\boldsymbol{\theta}')\ln\left(p(\mathbf{x}_{0:K},\mathbf{y}_{1:K};\boldsymbol{\theta})\right)\mathrm{d}\mathbf{x}_{0:K} \equiv \mathcal{E}\left[\ln\left(p(\mathbf{x}_{0:K},\mathbf{y}_{1:K};\boldsymbol{\theta})\right)|\mathbf{y}_{1:K}\right],$$

Instead of maximizing the joint distribution integrated in the full hidden state, the EM algorithm requires maximizing the expectation of $\ln p$ given the observations.

The maximum is found at

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\mathcal{E}\left[\ln\left(p(\mathbf{x}_{0:K},\mathbf{y}_{1:K};\boldsymbol{\theta})\right)|\mathbf{y}_{1:K}\right] &= 0\\ \mathcal{E}\left[\nabla_{\boldsymbol{\theta}}\ln\left(p(\mathbf{x}_{0:K},\mathbf{y}_{1:K};\boldsymbol{\theta})\right)|\mathbf{y}_{1:K}\right] &= 0\end{aligned}$$

Gaussian HMM

For an HMM, the joint density is factorable,

$$p(\mathbf{x}_{0:K}, \mathbf{y}_{1:K}) = p(\mathbf{x}_0) \prod_{k=1}^{K} p(\mathbf{x}_k|\mathbf{x}_{k-1}) p(\mathbf{y}_k|\mathbf{x}_k).$$

considering errors are Gaussian,

$$\ln(p(\mathbf{x}_{0:K}, \mathbf{y}_{1:K})) = -\frac{1}{2}\ln|\mathbf{P}_0| - \frac{1}{2}\left\|\mathbf{x}_0 - \overline{\mathbf{x}}_0\right\|^2_{\mathbf{P}_0} - \frac{K}{2}\ln|\mathbf{Q}|$$
$$-\frac{1}{2}\sum_{k=1}^{K}\left\|\mathbf{x}_k - \mathcal{M}(\mathbf{x}_{k-1})\right\|^2_{\mathbf{Q}} - \frac{K}{2}\ln|\mathbf{R}| - \frac{1}{2}\sum_{k=1}^{K}\left\|\mathbf{y}_k - \mathcal{H}(\mathbf{x}_k))\right\|^2_{\mathbf{R}}$$

The root of the grad is available analytically:

$$\mathbf{Q} = \frac{1}{K}\sum_{k=1}^{K} \mathcal{E}\left(\left[\mathbf{x}_k - \mathcal{M}(\mathbf{x}_{k-1})\right] \left[\mathbf{x}_k - \mathcal{M}(\mathbf{x}_{k-1})\right]^{\mathrm{T}} \middle| \mathbf{y}_{1:k}, \boldsymbol{\theta}' \right).$$

This expresion could be used with an ensemble Kalman smoother or with a particle smoother.

Algorithm

Input: Given $\mathbf{X}_{0,1} = \{\mathbf{x}_{0,1}^{(j)}\}_{j=1}^{N_p}, \mathbf{y}_{1:K}, \mathbf{R}_0, \mathbf{Q}_0$

for $i = 1, i_{max}$ **do** ▷ or a convergence criterion

$\mathbf{X}_{1:K}^a, \mathbf{X}_{1:K}^f \leftarrow EnKF(\mathbf{X}_{0,i}, \mathbf{y}_{1:K}, \mathbf{Q}_i, \mathbf{R}_i)$ ▷ Filter

$\mathbf{X}_{0:K}^s, \mathbf{X}_{1:K}^f \leftarrow RTS(\mathbf{X}_{1:K}^a, \mathbf{X}_{1:K}^f)$ ▷ Smoother, $\mathbf{X}_k^f = \mathcal{M}(\mathbf{X}_{k-1}^s)$

$\mathbf{Q}_{i+1} \leftarrow \frac{1}{K}\sum_k (\mathbf{X}_k^s - \mathbf{X}_k^f)(\mathbf{X}_k^s - \mathbf{X}_k^f)^{\mathrm{T}}$

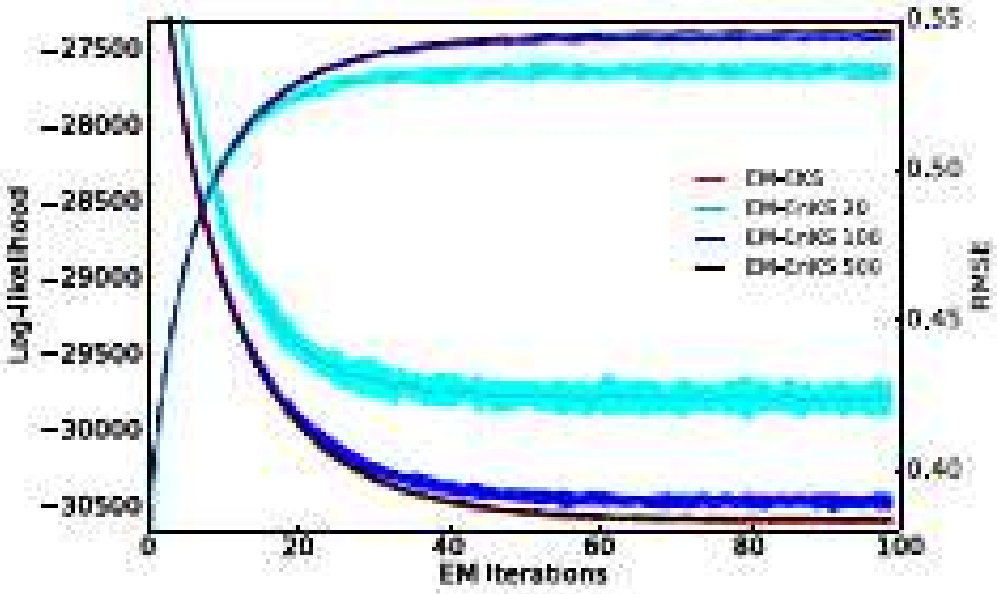
$\mathbf{R}_{i+1} \leftarrow \frac{1}{K}\sum_k (\mathbf{y}_k - \mathcal{H}(\mathbf{X}_k^s))(\mathbf{y}_k - \mathcal{H}(\mathbf{X}_k^s))^{\mathrm{T}}$

$\mathbf{X}_{0,i+1} \leftarrow \mathbf{X}_0^s$ ▷ New initial density

end for

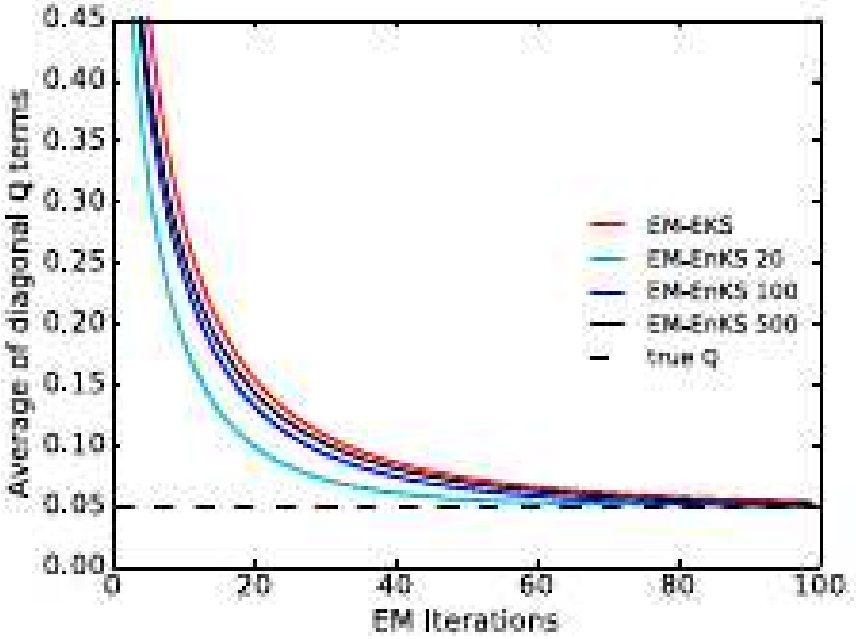
Output: $\{\mathbf{X}_{0,i_{imax}}, \mathbf{R}_{i_{max}+1}, \mathbf{Q}_{i_{max}+1}\}$

Estimating model error covariance in Lorenz-63



Averaged log-lik for 20 experiments.

From Dreano et al. (2017).



Averaged estimated model error covariance for 20 experiments.

Twin experiments with $\mathbf{R} = 2\mathbf{I}$ and $\mathbf{Q}^t = 0.05\mathbf{I}$. 100 cycles.

Stochastic parameterizations. Non-aditive model error

Most parameterizations have non-additive stochastic parameters. The augmented state approach allows to rewrite these systems to one with additive model error.

Any HMM with nonadditive model error

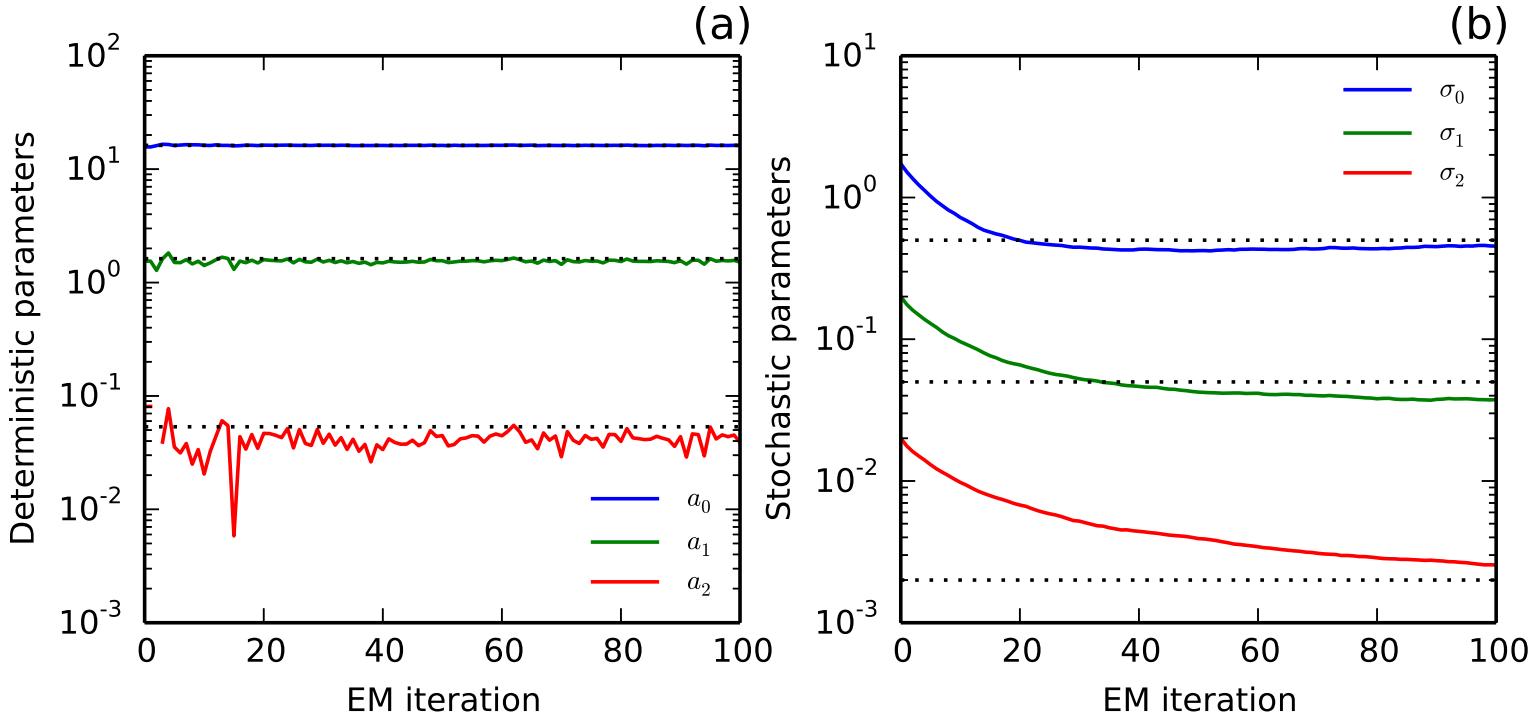
$$\mathbf{x}_k = \mathcal{M}(\mathbf{x}_{k-1}, \boldsymbol{\eta}_k) + \boldsymbol{\nu}_k$$

can be converted in an HMM with additive model error by augmenting the state space,

$$\mathbf{x}_k = \mathcal{M}(\mathbf{x}_{k-1}, \mathbf{z}_{k-1}) + \boldsymbol{\nu}_k$$
$$\mathbf{z}_k = \boldsymbol{\eta}_k$$

where time indeces for $\boldsymbol{\eta}_k$ have been shifted one time backward.

Stochastic parameterizations. Non-aditive model error



Twin experiment with Lorenz-96 and a quadratic stochastic parameterization.

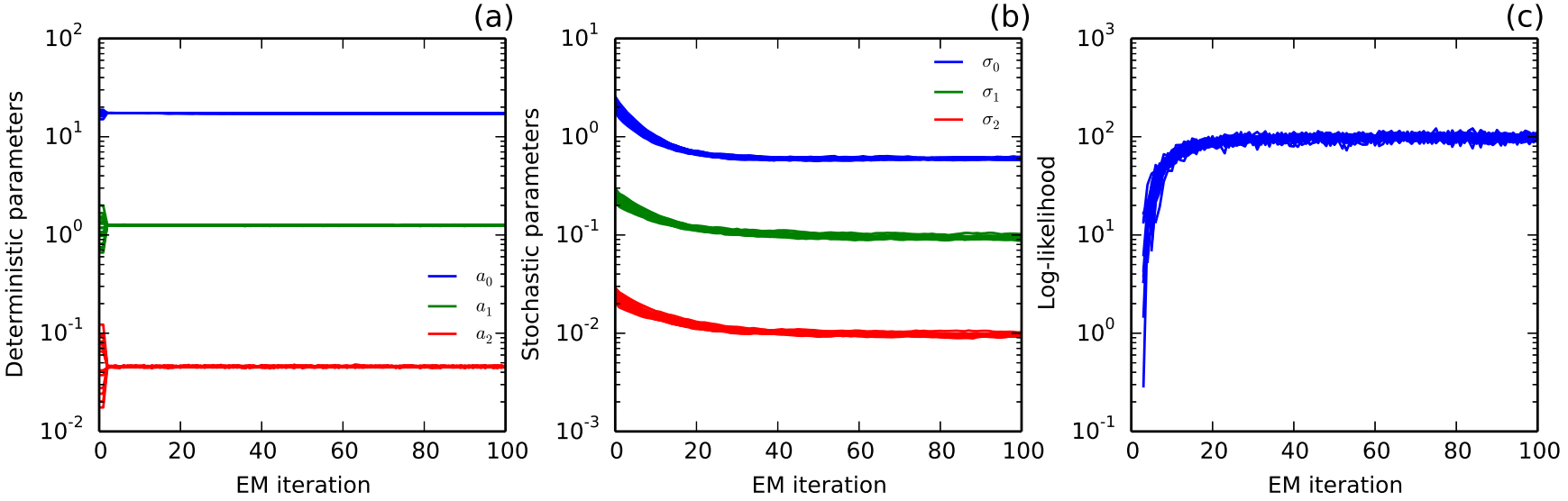
$\mathcal{P}_n(\mathbf{x}, \mathbf{a}) = \alpha_2 x_n^2 + \alpha_1 x_n + \alpha_0$ each parameter a random variable with

$\alpha_i = \mathcal{N}(a_i, \sigma_i)$.

From Pulido et al. in press

Two-scale Lorenz-96

Imperfect model. Suppose we want to find a parameterization of the 1scl-Lorenz-96 of the form $\mathcal{P}(\mathbf{x}, \mathbf{a}) = \sum_i \alpha_i x^i$ to mimic two-scale Lorenz-96.

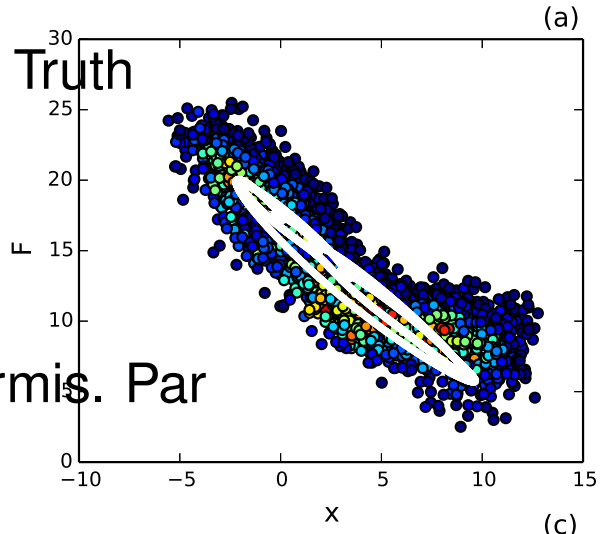


Optimal values are unknown in this case. 20 experiments with different initial values.

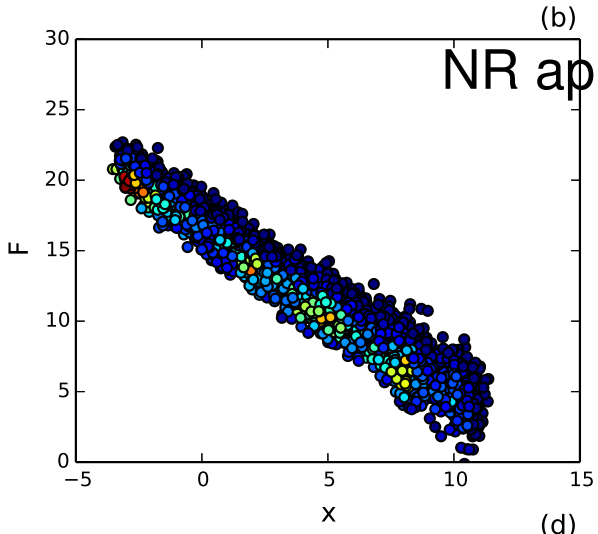
Different structural parameterizations can be evaluated. The larger maximal likelihood the better structure. Second degree is the optimal polynomial param.

Representation of the identified stochastic parameterizations.

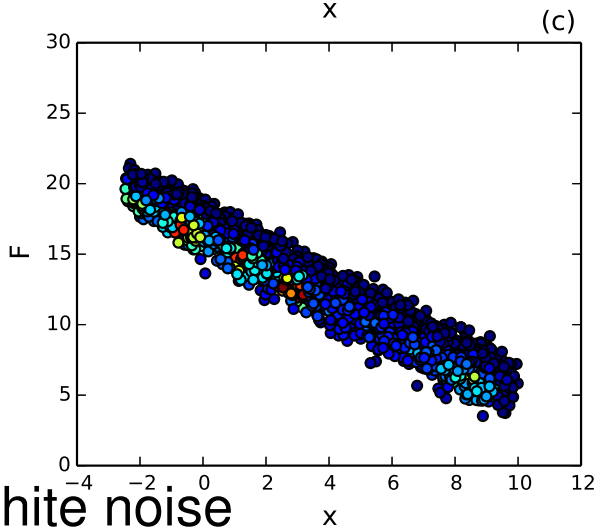
F=18 Truth



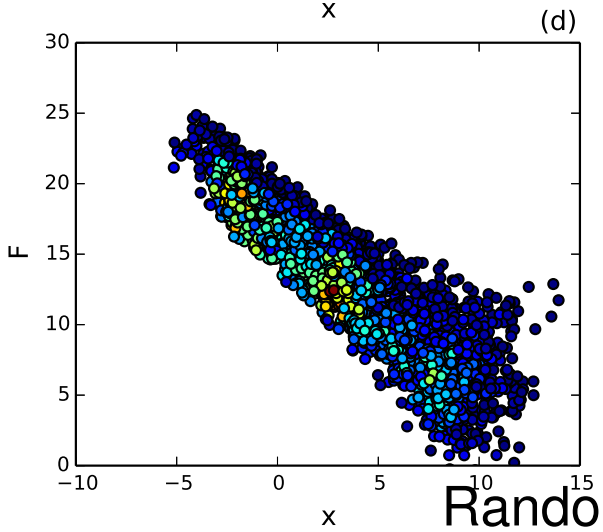
NR approach



Determis. Par



White noise



Random walk

All the knowledge of the parameterization is coming from observations of (only) large-scale variables.

A much better representation of the small-scale effects on large scale variables is obtained with the stochastic parameterization.

Conclusions

- Deterministic parameters are well estimated with augmented EnKF approach. No need of localization for global parameters.
- For model structure identification or parameterization development, the offline approach gives promising results. Further testing are required.
- The EM algorithm is a quite straighforward method to estimate optimal statistical parameters in a data assimilation system.
- EM gives highly robust estimations of stochastic parameters.