Data Assimilation for Weather Forecasting: Reducing the Curse of Dimensionality

Philippe Toint
(with S. Gratton, S. Gürol, M. Rincon-Camacho, E. Simon and J. Tshimanga)

University of Namur, Belgium

Leverhulme Lecture IV, Oxford, December 2015

Thanks

- Leverhulme Trust, UK
- Balliol College, Oxford
- Belgian Fund for Scientific Research (FNRS)
- University of Namur, Belgium

Talk outline

Data assimilation in earth sciences

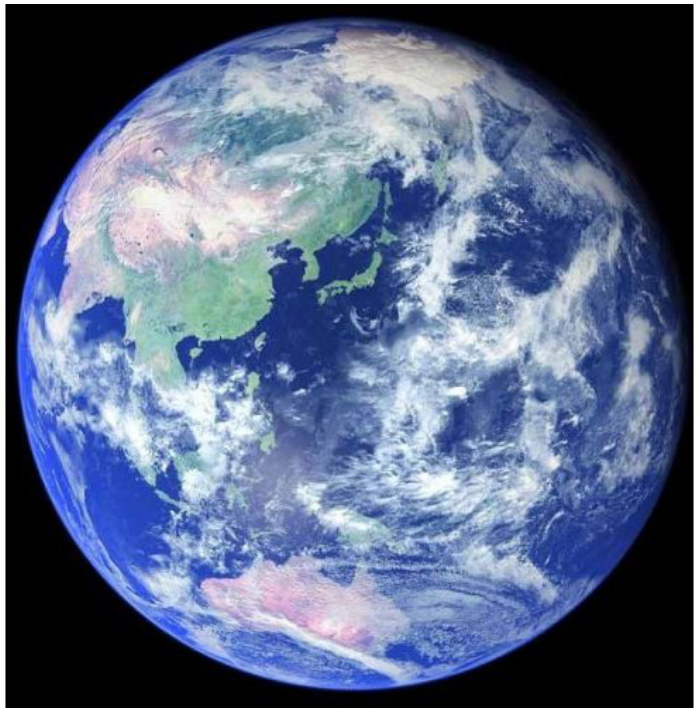
The state of the atmosphere or the ocean (the system) is characterized by state variables that are classically designated as fields:

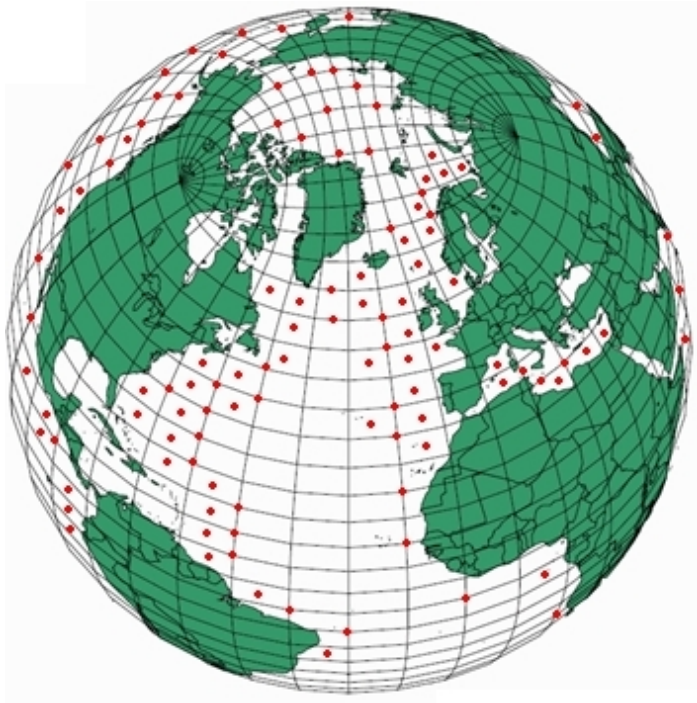
- velocity components
- pressure
- density
- temperature
- salinity

A dynamical model predicts the state of the system at a time given the state of the ocean at a earlier time. We address here this estimation problem. Applications are found in climate, meteorology, ocean, neutronic, hydrology, seismic,... (forecasting) problems. Involving large computers and nearly real-time computations.

Data assimilation in weather forecasting (2)

Data: température, wind, pression, . . . everywhere and at all times !

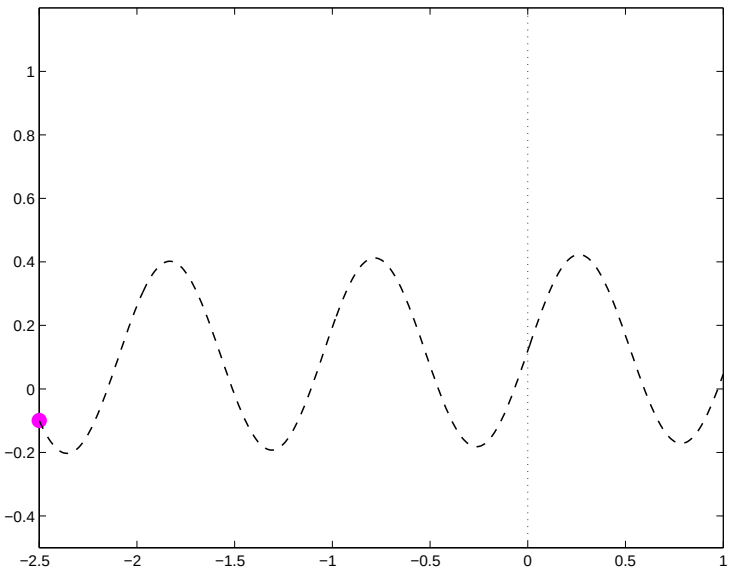




May involve more than 1.000.000.000 variables!

Data assimilation in weather forecasting (3)

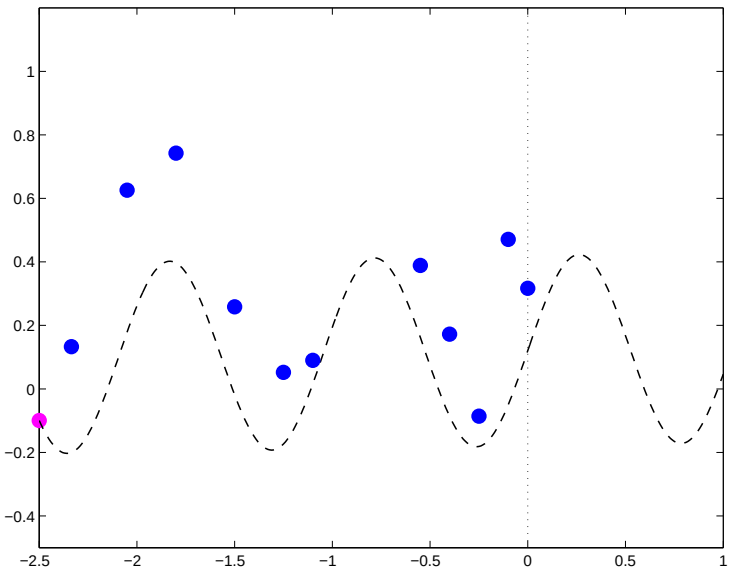
The principle:



- Situation 2.5 days ago and "background" prediction

Data assimilation in weather forecasting (3)

The principle:

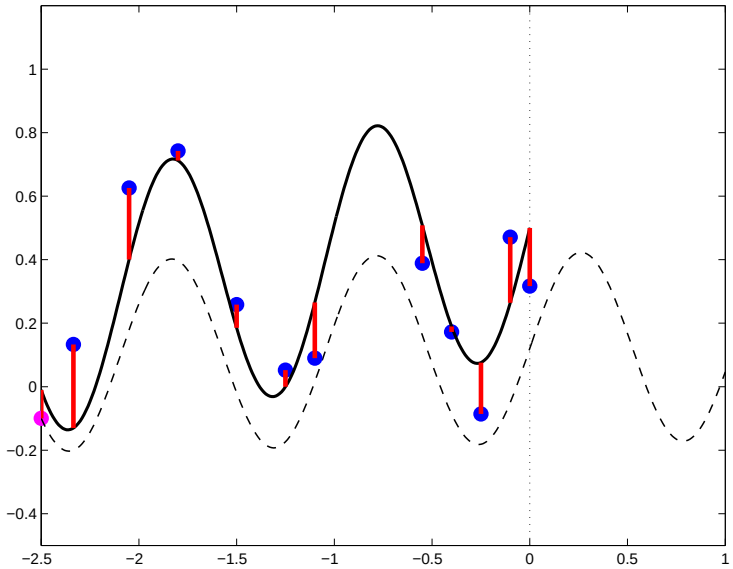


- Situation 2.5 days ago and "background" prediction
- Température for the last 2.5 days

Data assimilation in weather forecasting (3)

The principle:

Minimize the error between model and past observations



temperature/days

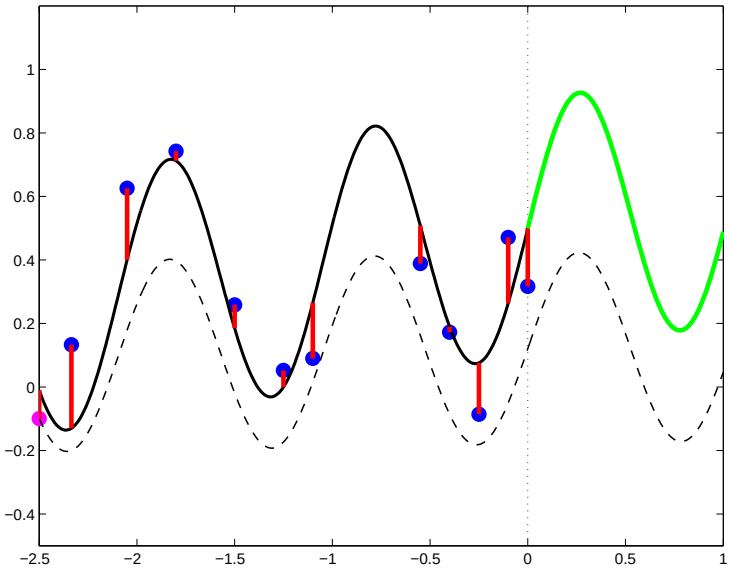
- Situation 2.5 days ago and "background" prediction
- Température for the last 2.5 days
- Run the model to minimize the gap between I model and observations

$$\min_{x_0} \frac{1}{2}\|x_0 - x_b\|^2_{B^{-1}} + \frac{1}{2}\sum_{i=0}^{N} \|\mathcal{H}\mathcal{M}(t_i, x_0) - b_i\|^2_{R_i^{-1}}.$$

Data assimilation in weather forecasting (3)

The principle:

Minimize the error between model and past observations



temperature/days

- Situation 2.5 days ago
  and "background" prediction
- Température for the last 2.5 days
- Run the model to minimize
  the gap between I model and observations
- Predict tomorrow's temperature

Four-Dimensional Variational (4D-Var) formulation

→ Very large-scale nonlinear weighted least-squares problem:

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{2}||x - x_b||^2_{B^{-1}} + \frac{1}{2}\sum_{j=0}^{N} ||\mathcal{H}_j(\mathcal{M}_j(x)) - y_j||^2_{R_j^{-1}}$$

where:

- Size of real (operational) problems: $x, x_b \in \mathbb{R}^{10^9}$, $y_j \in \mathbb{R}^{10^5}$
- The observations y_j and the background x_b are noisy
- $\mathcal{M}_j$ are model operators (nonlinear)
- $\mathcal{H}_j$ are observation operators (nonlinear)
- B is the covariance background error matrix
- R_j are covariance observation error matrices

Incremental 4D-Var

Rewrite the problem as:

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{2}||\rho(x)||_2^2$$

Incremental 4D-Var = inexact/truncated Gauss-Newton algorithm

- Linearize ρ around the current iterate $\tilde{x}$ and solve

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|\rho(\tilde{x}) + J(\tilde{x})(x - \tilde{x})\|_2^2$$

  where $J(\tilde{x})$ is the Jacobian of $\rho(x)$ at $\tilde{x}$

- Solve a sequence of linear systems (normal equations)

$$J^T(\tilde{x})J(\tilde{x})(x - \tilde{x}) = -J^T(\tilde{x})\rho(\tilde{x})$$

  where the matrix is symmetric positive definite and varies along the iterations

Inner minimization

Ignoring superscripts, minimizing

$$J(\delta x_0) = \frac{1}{2}\|\delta x_0 - [x_b - x_0]\|^2_{B^{-1}} + \frac{1}{2}\|H\delta x_0 - d\|^2_{R^{-1}}$$

amounts to iteratively solve

$$(B^{-1} + H^T R^{-1} H)\delta x_0 = B^{-1}(x_b - x_0) + H^T R^{-1} d.$$

whose exact solution is

$$x_b - x_0 + \left(B^{-1} + H^T R^{-1} H\right)^{-1} H^T R^{-1} \left(d - H(x_b - x_0)\right),$$

or equivalently (using the Sherman-Morrison-Woodbury formula)

$$x_b - x_0 + BH^T \left(R + HBH^T\right)^{-1} \left(d - H(x_b - x_0)\right).$$

Solving the 4D-VAR subproblem

That is:

$$(I + BH^T R^{-1} H)x = BH^T R^{-1} d$$

In pratice:

- use Conjugate Gradients
  (or other Krylov space solver – more later on this)
- for a (very) limited number of level-2 iterations
- (with preconditioning – more later on this)

$\Rightarrow$ need products of the type

$$(I + BH^T R^{-1} H)v \quad \text{for a number of vectors } v$$

Focus now on how to compute Bv (B large)

Modelling covariance

A widely used approach (Derber + Rosati, 1989, Weaver + Courtier, 2001):

Spatial background correlation $\approx$ diffusion process

i.e.

Computing Bv
$\approx$
integrating a diffusion equation starting from the state v.

use p steps of an implicit integration scheme

(level-3 iteration, each involving a solve with B!!!)

The integration iteration

Define

$$\Theta_h = I + \frac{\mathcal{L}}{2p}\Delta_h$$

(Δ_h is the discrete Laplacian, $\mathcal{L}$ is the correlation length).
For each integration (z and p given)

1. $u_0 = \left(\mathrm{diag}(\Theta_h^{-p})\right)^{-1/2} z$ (diagonal scaling)
2. $\boxed{u_\ell = \Theta_h^{-1} u_{\ell-1}}$ $(\ell = 1, \ldots, p)$
3. $Bz = \left(\mathrm{diag}(\Theta_h^{-p})\right)^{-1/2} u_p$ (diagonal scaling)

Our question: how to solve $\Theta_h u_\ell = u_{\ell-1}$?

The integration iteration

An (pratically important) question: how to solve $\Theta_h u_\ell = u_{\ell-1}$?

Carrier + Ngodock (2010):

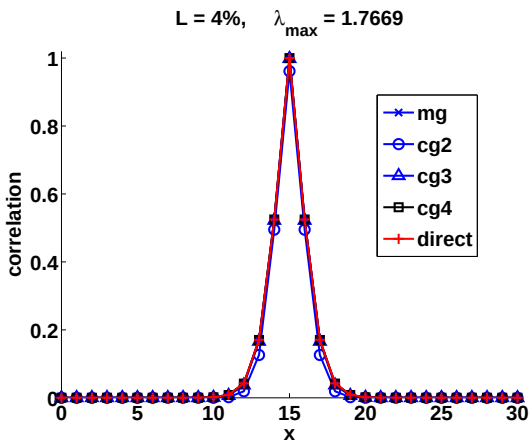
Implicit integration + CG is $\approx$ 5 times faster than explicit integration!

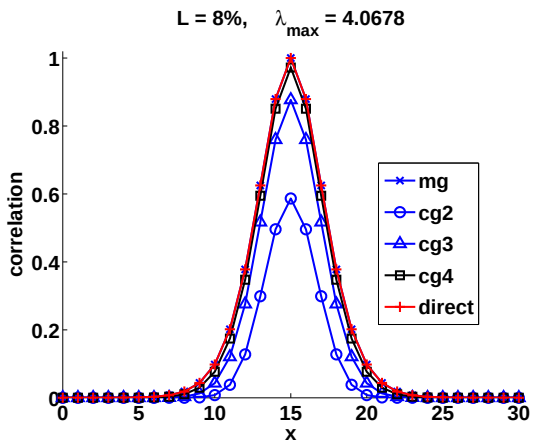
But:

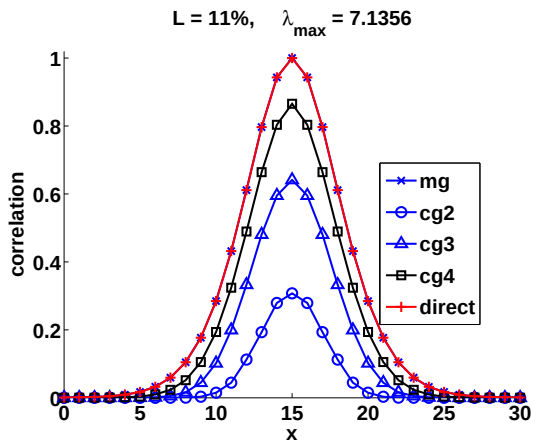
- What about multigrid ??
- Is an approximate solution of the system (CG or MG) altering the spatial properties of the correlation?
- Inexact solves ?

Approximately diffusing a Dirac pulse

Compare the diffusion of a Dirac pulse using approximate linear solvers and exact factorization, as a function of correlation length:



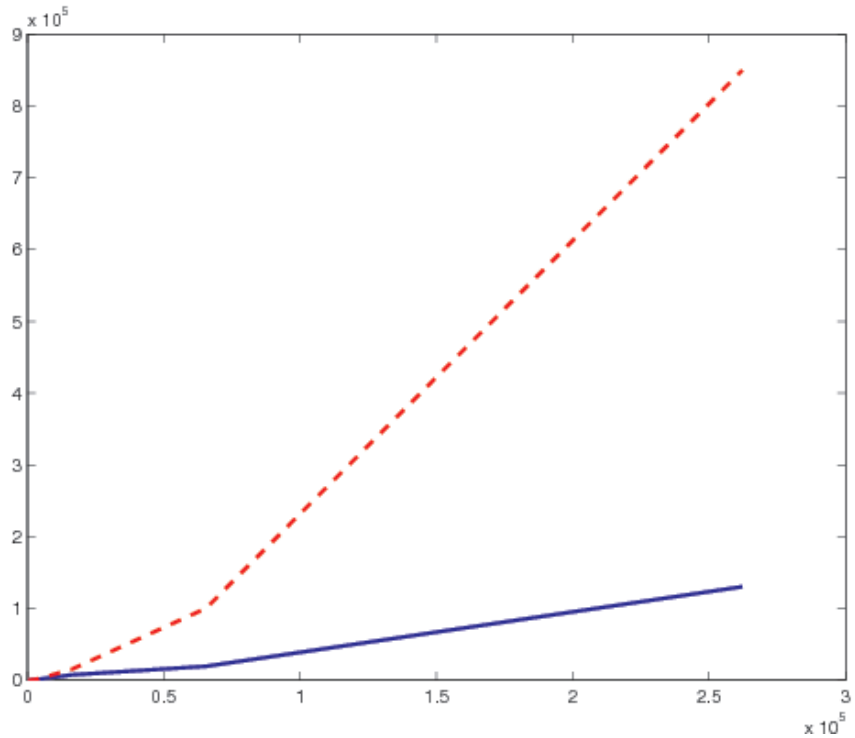




Note: cost(1 MG V-cycle) $\approx$ cost(4 CG iterations)

Comparing the computational costs of CG vs MG

Consider a complete data assimilation exercize on a 2D shallow-water system (3 level-1 iterations, 15 level-2 iterations, $p = 6$, tol $= 10^{-4}$)



Number of "normalized" matrix-vectors products as a function of problem size

MG is a interesting alternative to CG in the integration loop

Solving the Gauss-Newton model: PSAS

System matrix (after level 1 preconditioning) =
low (???) rank perturbation of the identity

1. Very popular when few observations compared to model variables. Stimulated a lots of discussion e.g. in the Ocean and Atmosphere communities (cfr P. Gauthier)
2. Relies on

$$x_b - x_0 + BH^T \left(R + HBH^T\right)^{-1} (d - H(x_b - x_0))$$

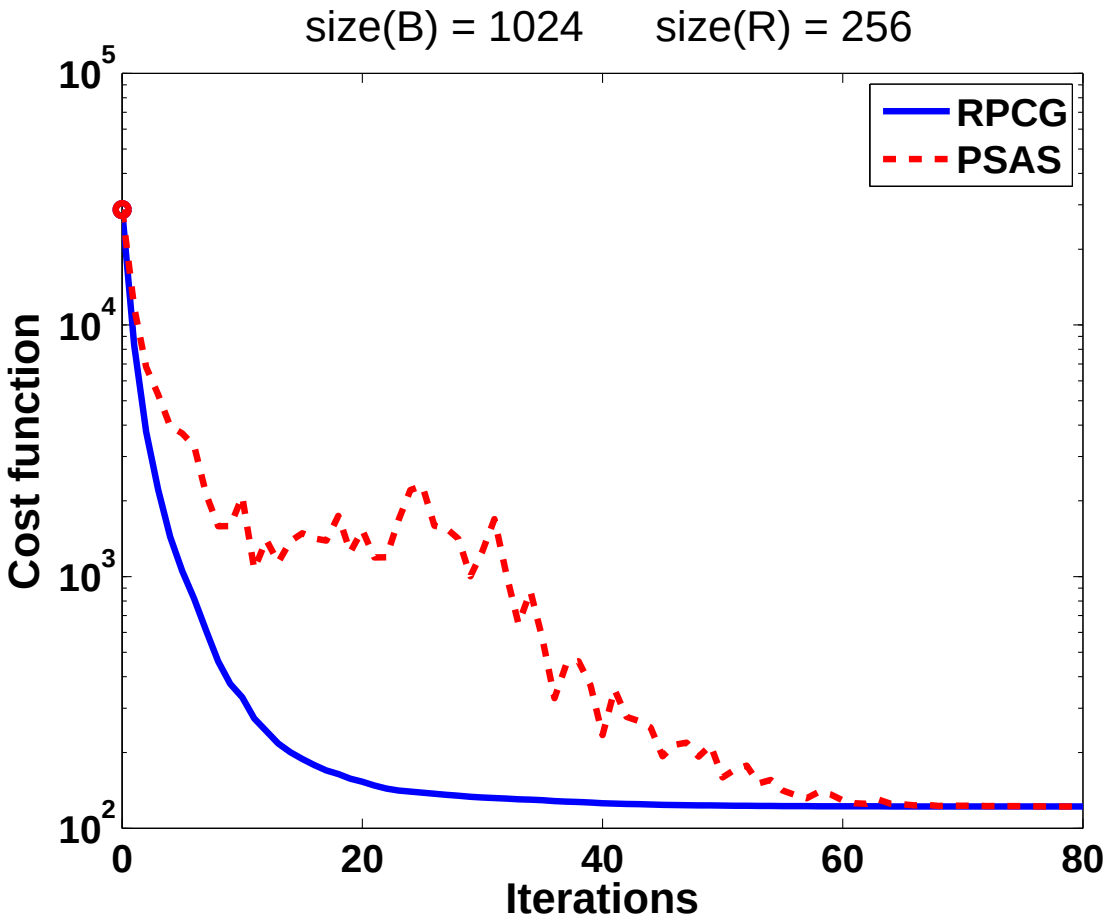
3. Iteratively solve

$$\left(I + R^{-1}HBH^T\right) w = R^{-1}(d - H(x_b - x_0)) \quad \text{for } w$$

4. Set

$$\delta x_0 = x_b - x_0 + BH^T w$$

Experiments



Motivation : PSAS and CG-like algorithm

1. CG minimizes the Incremental 4D-Var function during its iterations. It minimizes a quadratic approximation of the non quadratic function using Gauss-Newton in the model space.
2. PSAS does not minimize the Incremental 4D-Var function during its iterations but works in the observation space.

Our goal : combine the advantages of both approaches:

- the variational property: enforce sufficient descent even when iterations are truncated.
- computational efficiency: work in the (dual) observation space whenever the number of observations is significantly smaller than the size of the state vector

Preserve global convergence in the observation space !

Range space reformulation

- Use conjugate-gradients (CG) to solve the step computation

$$\min_{x_0} J(s) \stackrel{\text{def}}{=} \tfrac{1}{2}(x_s + s - x_b)^T B^{-1}(x_s + s - x_b) + \tfrac{1}{2}(Hs - d)^T R^{-1}(Hs - d)$$

- Reformulate as a constrained problem:

$$\min_{x_0} J(s) \stackrel{\text{def}}{=} \tfrac{1}{2}(x_s + s - x_b)^T B^{-1}(x_s + s - x_b) + \tfrac{1}{2}a^T R^{-1} a$$
$$\text{such that} \qquad a = Hs - d$$

- Write KKT conditions of this problem (for $x_s = x_b$, wlog)

$$(R + HBH^T)\lambda = d, \quad s = BH^T\lambda, \quad a = -R\lambda$$

- Precondition (1rst level) by R, forget a:

$$(I + R^{-1}HBH^T)\lambda = R^{-1}d, \quad s = BH^T\lambda,$$

Solve system using (preconditioned) CG in the H^TBH inner product

$\Rightarrow$ RPCG (Gratton, Tshimanga, 2009)

RPCG and preconditoning

Features of RPCG:

- algorithm form comparable to PCG (additional products)
- only uses vector of size = number of observations
  $\Rightarrow$ suitable for reorthogonalization (if useful)
- sequence of iterates identical to that generated by PCG on primal
  $\Rightarrow$ good descent properties on $J(s)$!
- numerically stable for range-space perturbations
- any (2nd level) preconditioner F for the primal PCG translates to a preconditioner G for the range-space formulation iff

$$FH^T = BH^T G$$

Note: works for limited memory preconditioners

Numerical performance on ocean data assimilation ?

1. NEMOVAR: Global 3D-Var system
   $\longrightarrow$ seasonal forecasting $\longrightarrow$ ocean reanalysis
2. ROMS California Current Ocean Modeling: Regional 4D-Var system
   $\longrightarrow$ coastal applications

Implementation details:

- $m \approx 10^5$ and $n \approx 10^6$
- The model variables: temperature, height, salinity and velocity.
- 1 outer loop of Gauss-Newton ($k = 1$), 40 inner loops

Gürol, Weaver, Moore, Piacentini, Arango, Gratton, 2013. *Quarterly Journal of the Royal Meteorological Society*. In press.

- good numerical performance
- reorthogonalization sometimes necessary

In the nonlinear setting

When solving the general 4DVAR problem...

Several (outer) Gauss-Newton iterations
using a (range-space) trust-region framework

Question: Can one reuse the range-space preconditioner from previous iteration?

$$??? \quad F_{k-1} H_k^T = B H_k^T G_{k-1} \quad ???$$

In general: No!

Some fixes

The old preconditioner is no longer symmetric in the new metric!

How to get around this problem:

1. avoid (2nd level) preconditioning ???
   (last resort decision)
2. recompute the preconditioner in the new metric ??
   (possible with limited memory preconditioners, but costly)
3. ignore the problem and take one's chances ?
   (only reasonable if convergence can still be guaranteed)

Taking measured risks...

A simple proposal for computing the step:

1. compute the Cauchy step (1rst step of RPCG)
2. if negative curvature, recompute a complete step without 2nd level preconditioner
3. otherwise, use equivalence with primal to check simple decrease of f at the Cauchy point
4. if no decrease, then unsuccessful TR outer iteration
5. otherwise, continue RPCG with old preconditioner
6. if unsuccessful, go back to Cauchy point

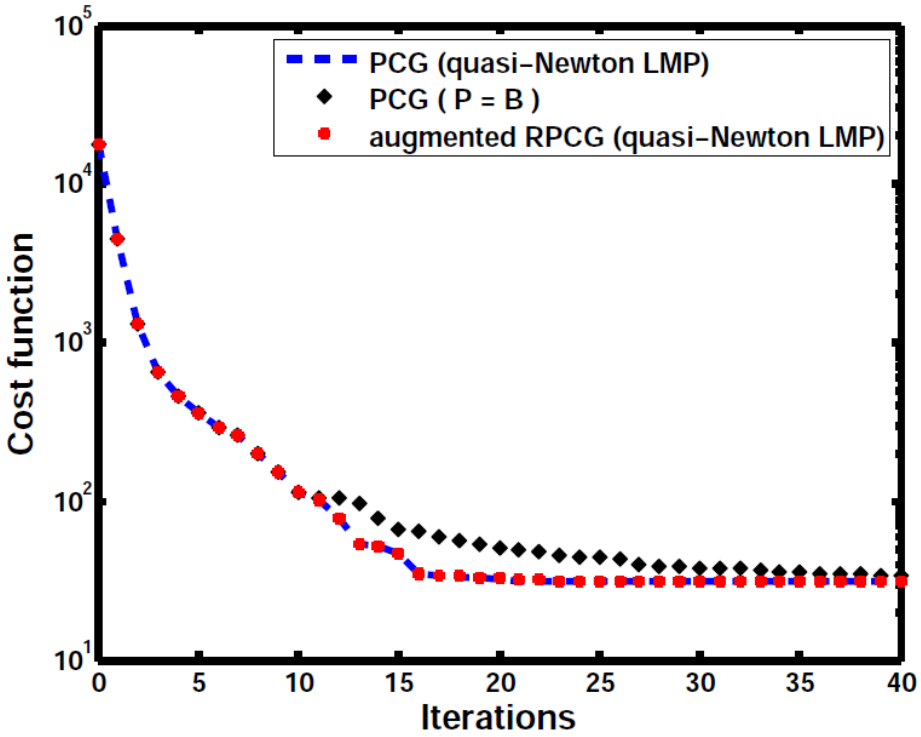
Can be interpreted as a TR "magical step"

S. Gratton, S. Gürol and Ph. L. Toint, 2013. Preconditioning and globalizing conjugate-gradients in dual space for quadratically penalized nonlinear-least squares problems. *Computational Optimization and Applications* 54: 1-25
S. Gürol, PhD thesis, 2013

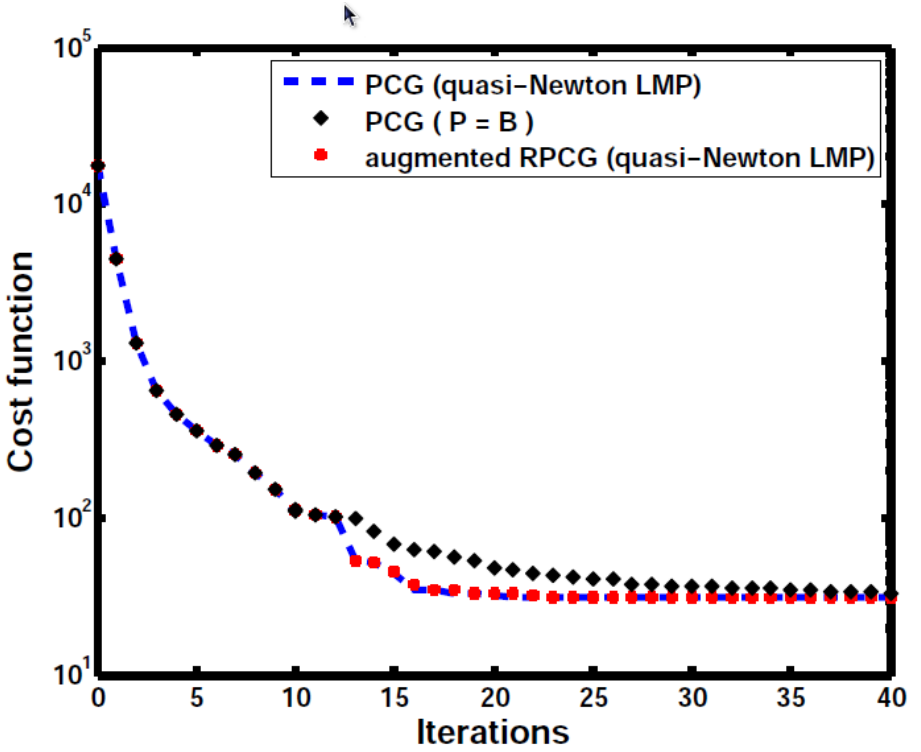
Does it work? (1)

Numerical experiment with a nonlinear heat equation

$f(x) = exp[1x]$

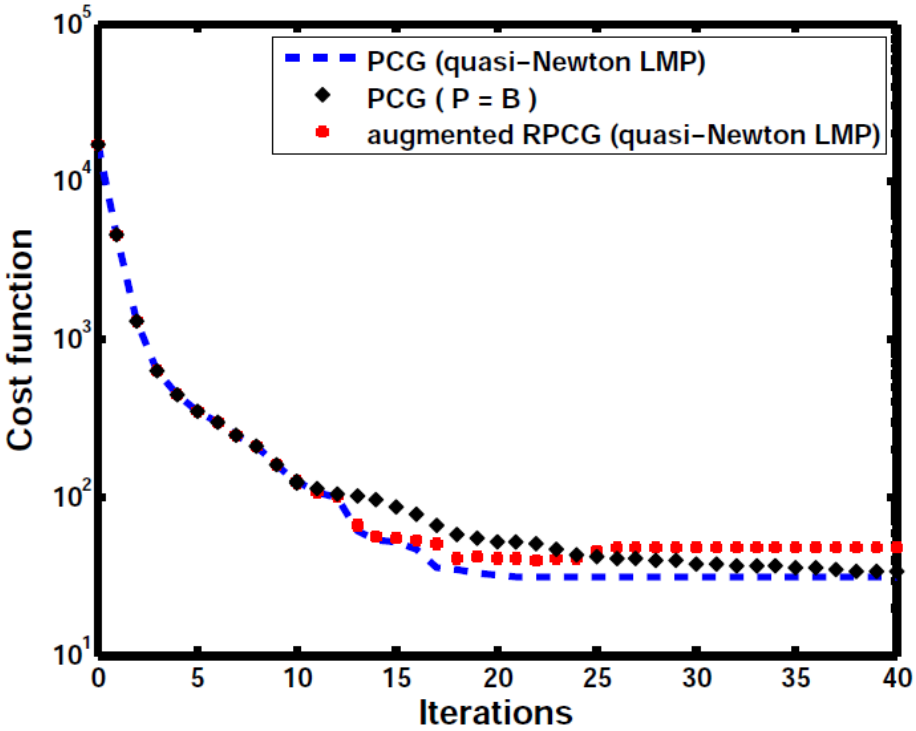


$f(x) = exp[2x]$

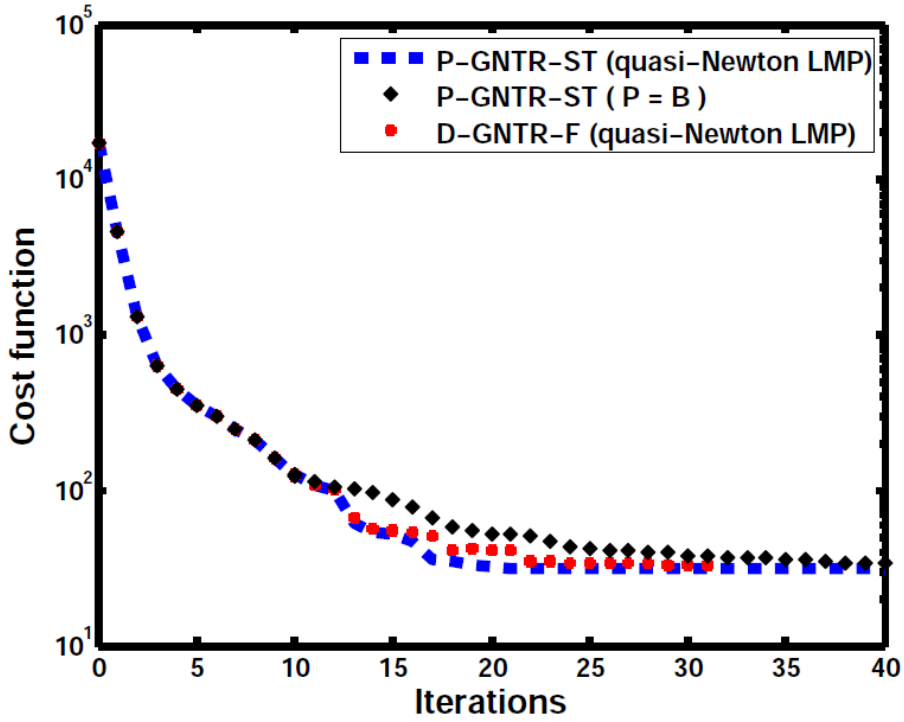


Does it work? (2)

$f(x) = exp[3x]$



$f(x) = exp[3x]$



$\longrightarrow$ The global convergence is ensured by using the **"risky" trust-region algorithm.**
$\longrightarrow$ This algorithm **requires an additional function evaluation** at the Cauchy point for each outer loop.

Conclusions for Section 1

Multigrid approach useful for gandling background covariance matrix

Range-space approach efficient in some data assimilation problems

Suitable 2nd level preconditioners can be built

Potential symmetry problem solved without compromising convergence

Already in use in real operational systems

Observation thinning

In many applications,

- too many observations in some parts of the domain!
- observations can be considered into a nested hierarchy $\{\mathcal{O}_i\}_{i=0}^r$ with

$$\mathcal{O}_i \subset \mathcal{O}_{i+1} \quad i = 0, \dots r-1.$$

(coarse vs fine)

Can we exploit this for reducing computations?

The coarse and fine subproblems

The fine (sub)problem:

$$\min_s \tfrac{1}{2} \| x + s_f - x_b \|_{B^{-1}} + \tfrac{1}{2} \| H_f s - d_f \|^2_{R_f^{-1}}$$

The coarse (sub)problem:

$$\min_s \tfrac{1}{2} \| x + s_c - x_b \|_{B^{-1}} + \tfrac{1}{2} \| \Gamma_f (H_f s - d_f) \|^2_{R_c^{-1}}$$

where Γ_f is the restriction from fine to coarse observations.
Moreover

- fine problem formulation $\Longrightarrow$ fine multiplier λ_f
- coarse problem formulation $\Longrightarrow$ coarse multiplier λ_c

A useful error bound

Question: what is the difference between fine and coarse multipliers ?

If $\Pi_c = \sigma\Gamma_f$ is the prolongation from the coarse observations to the fine ones, then

$$\| \lambda_f - \Pi_c\lambda_c \|_{R_f + H_f B H_f^T} \leq \| d_f - H_f s_c - R_f \Pi_c \lambda_c \|_{R_f^{-1}}$$

(proof somewhat technical. . .)

- Uses d_f but no comptuted quantity at the fine level
- Observation i useful if the i-th component of $\lambda_f - \Pi_c\lambda_c$ is large

How to exploit this?

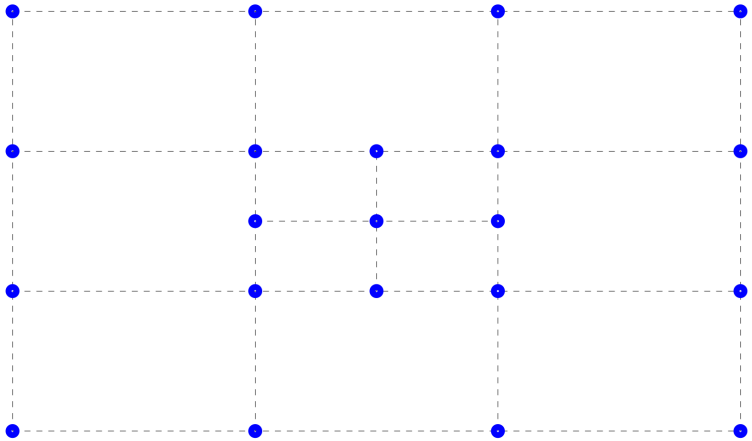
Idea:

Starting from the coarsest observation set, and until the finest observation set is used:

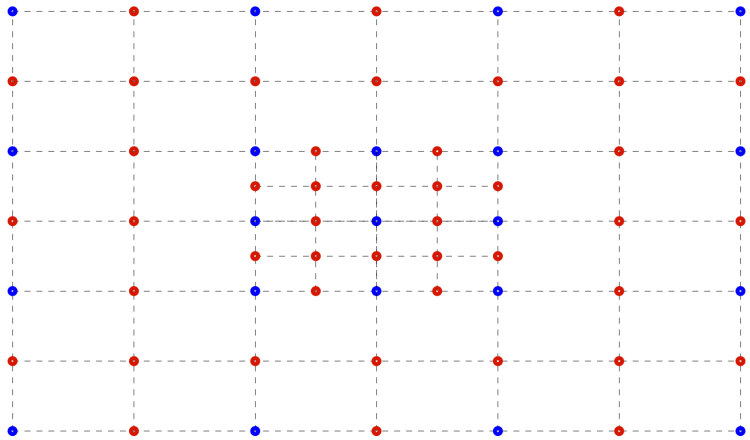
1. solve the coarse problem for (s_c, λ_c)
2. define a finer auxiliary problem by moving up in the hierarchy of observation sets (i.e. consider finer auxiliary observations)
3. use theorem to estimate distances from λ_c to $\lambda_{\text{aux}} = \Pi_c \lambda_c$
4. using this, select a subset of the auxiliary observations whose impacts represents the impacts of these observations well enough (thinning)
5. redefine this selection as the next coarse observation set and loop

(needs: a more formal definition of the observations hierarchy
+ selection procedure)

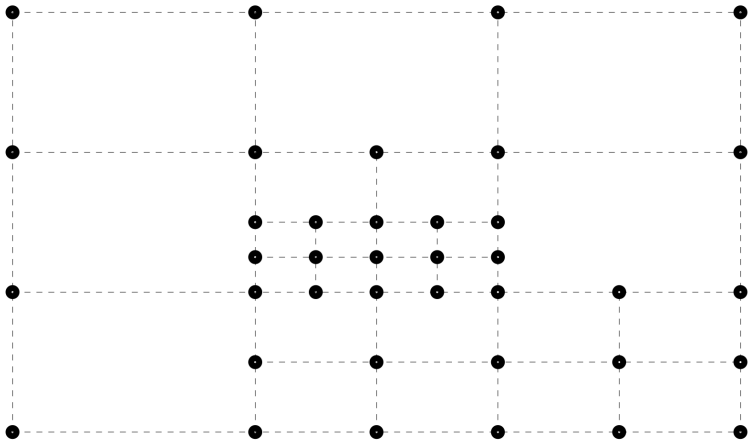
An example of observation sets



Coarse set



Auxilary set



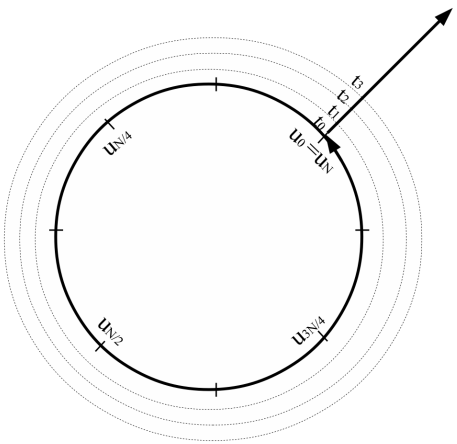
Selected fine set

Example 1: The Lorenz96 chaotic system (1)

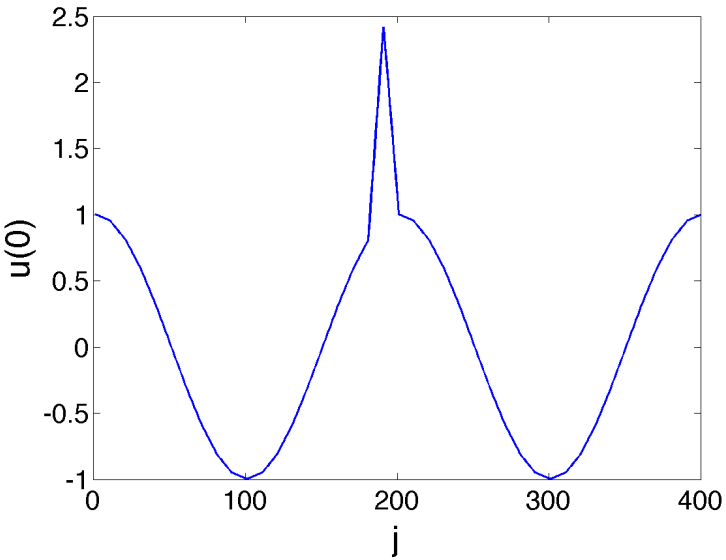
Find u_0, where $\bar{u}$ is an N-equally spaced entries around a circle, obeying

$$\frac{du_{j+\theta}}{dt} = \frac{1}{\kappa}(-u_{j+\theta-2}u_{j+\theta-1} + u_{j+\theta-1}u_{j+\theta+1} - u_{j+\theta} + F),$$

$(j = 1, \ldots 400, \quad \theta = 1, \ldots, 120)$

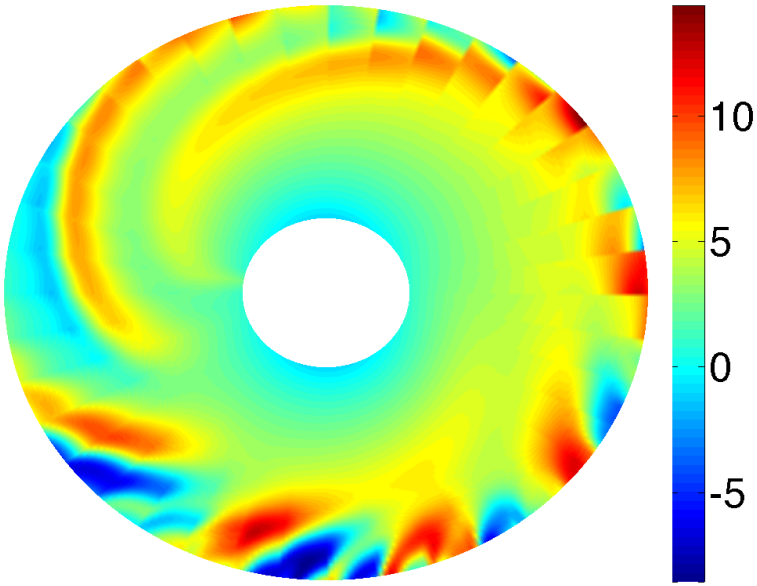


Coordinate system

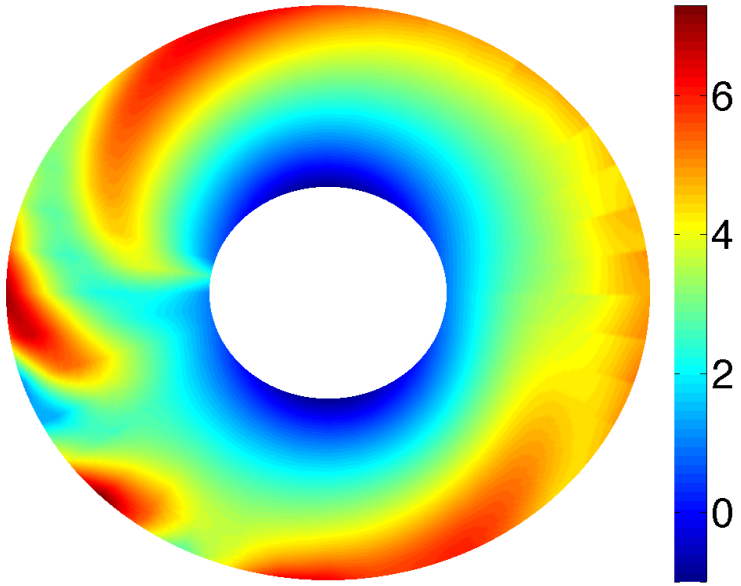


Initial $(u_1(0), u_2(0), \ldots, u_N(0))$

Example 1: The Lorenz96 chaotic system (2)

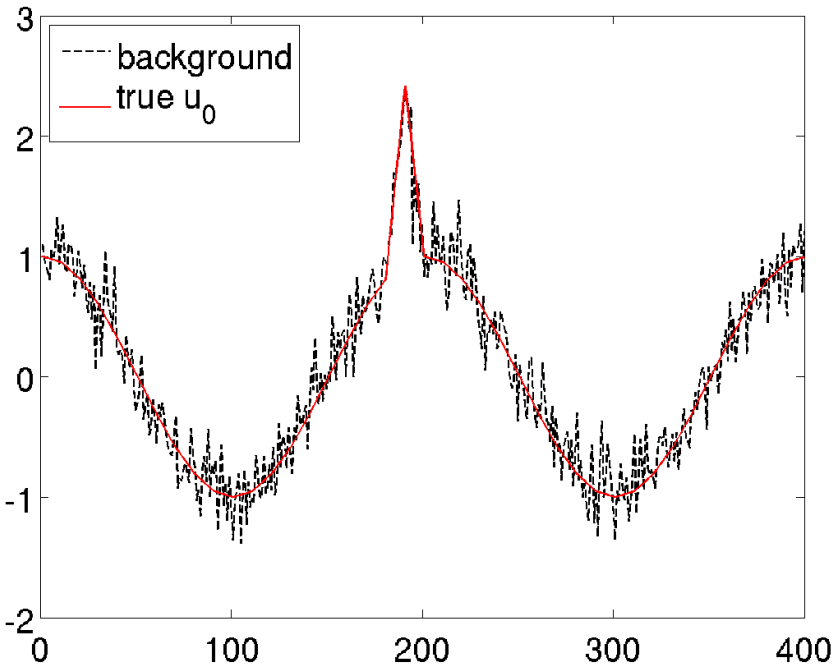


System over space and time

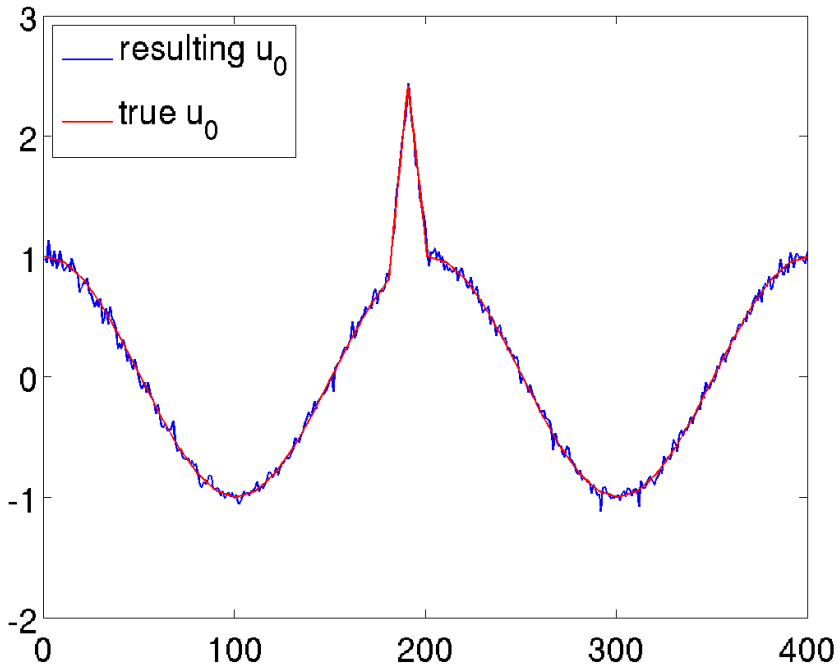


Window of assimilation

The Lorenz96 chaotic system (3)



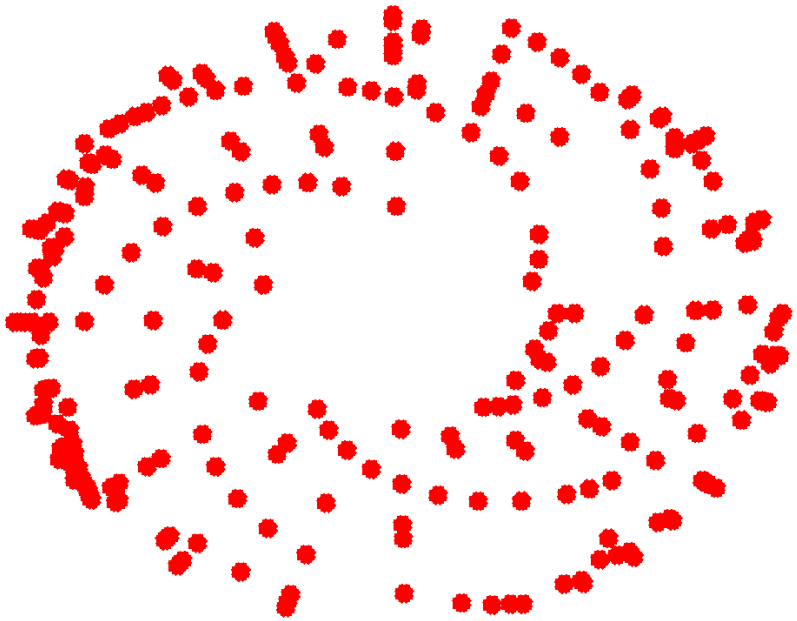
Background and true initial $u(0)$



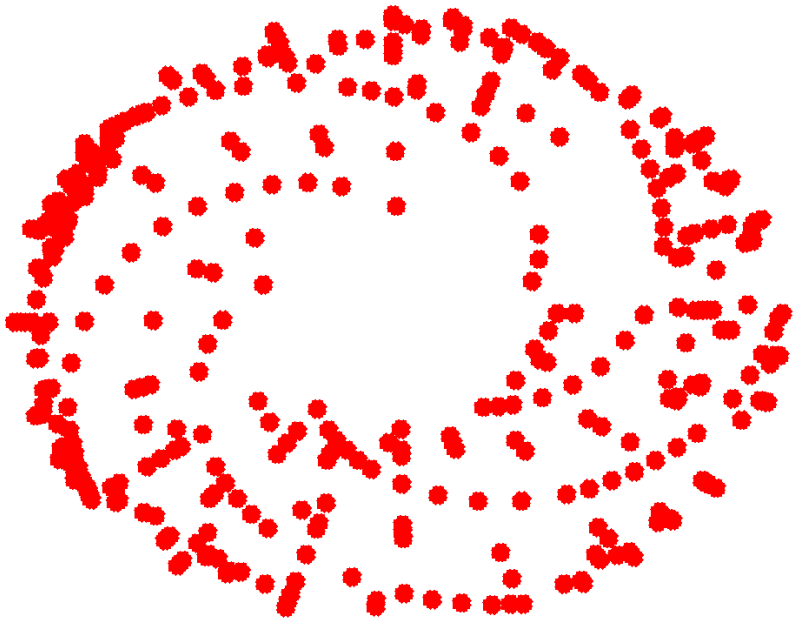
True and computed $u(0)$

Example 1: The Lorenz96 chaotic system (4)

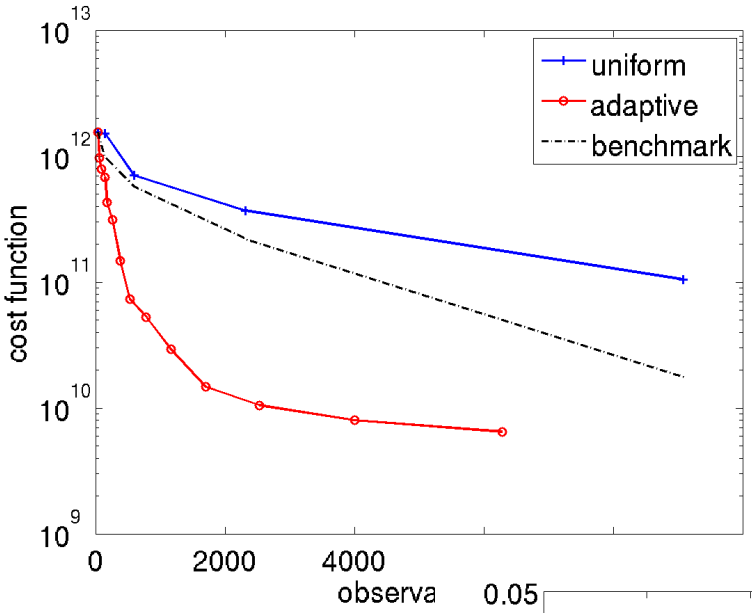
An example of transition from coarse to fine observations sets :



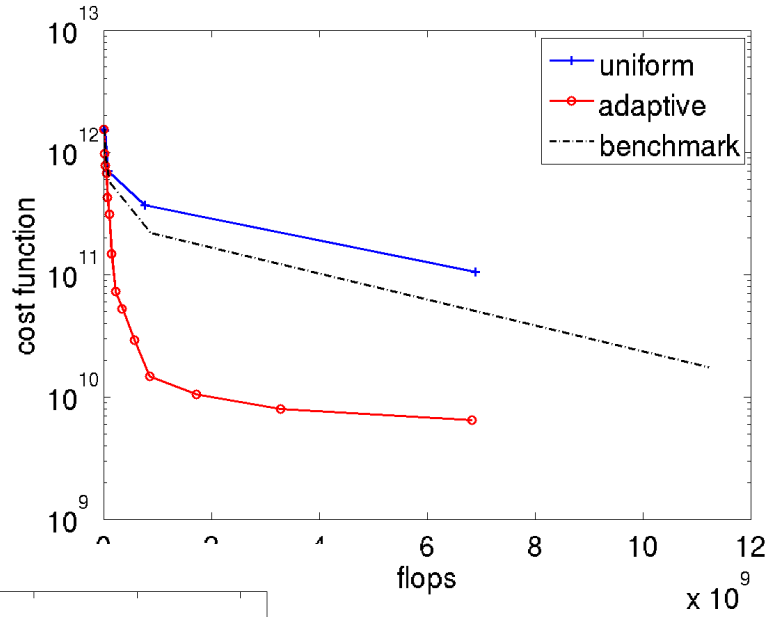
$\mathcal{O}_i$ $\longrightarrow$ $\mathcal{O}_{i+1}$



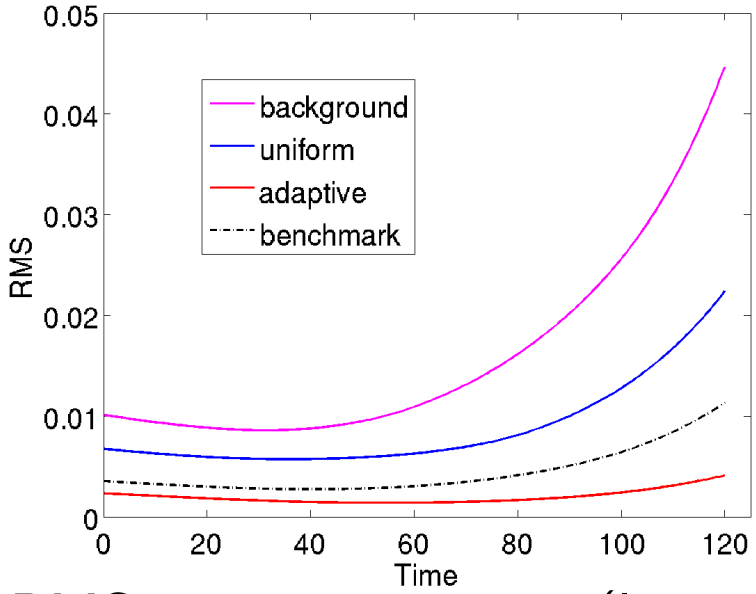
Example 1: The Lorenz96 chaotic system (5)



Cost vs obs



Cost vs flops



RMS error versus time (last iteration)

Example 2: 1D wave system with a shock (1)

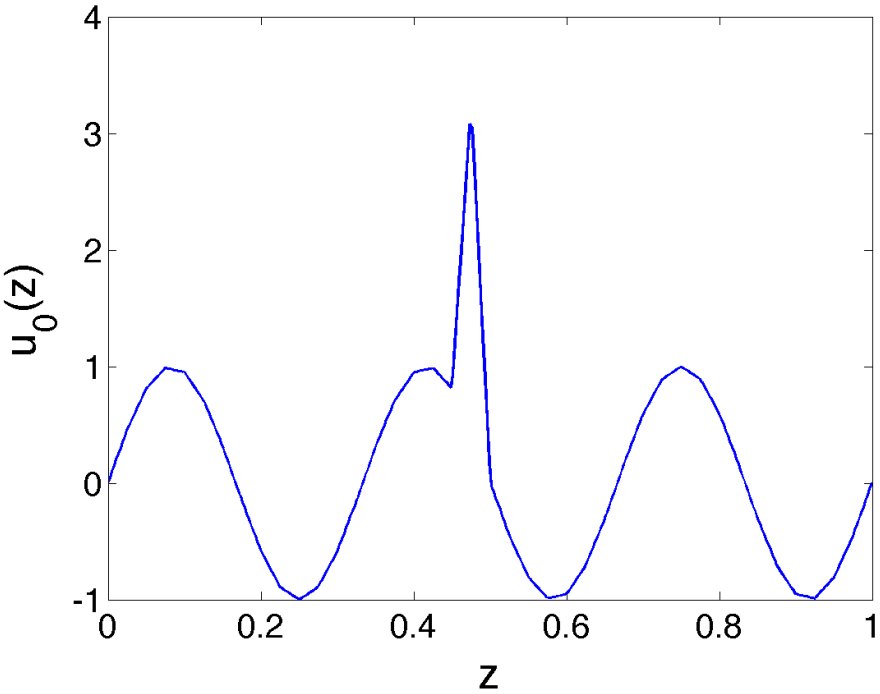
Find $u_0(z)$ in

$$\frac{\partial^2}{\partial t^2}u(z,t) - \frac{\partial^2}{\partial z^2}u(z,t) + f(u) = 0,$$
$$u(0,t) = u(1,t) = 0,$$
$$u(z,0) = u_0(z),\ \frac{\partial}{\partial t}u(z,0) = 0,$$
$$0 \leq t \leq T,\ \ 0 \leq z \leq 1,$$

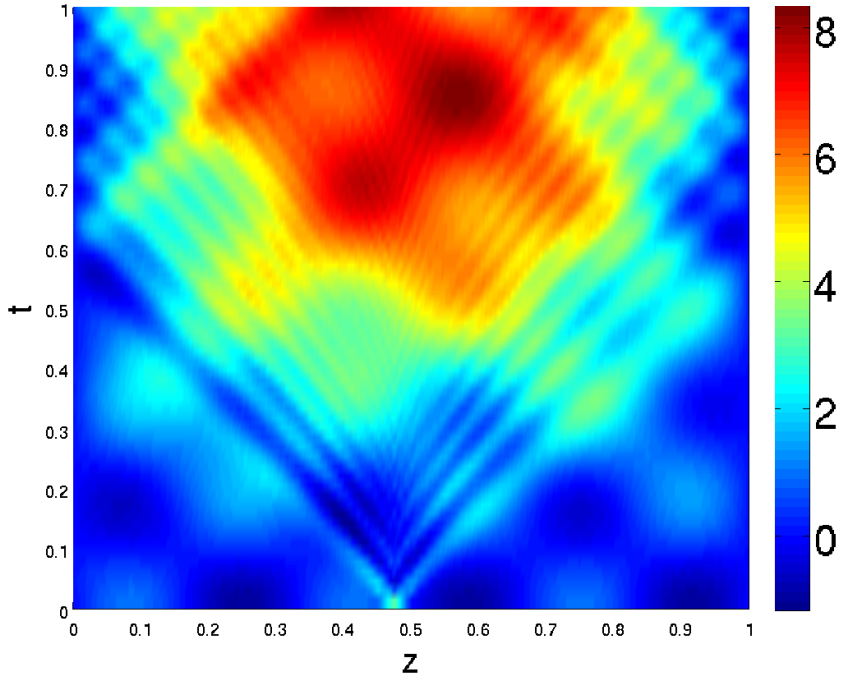
where $f(u) = \mu e^{\eta u}$

(360 grid points, $\Delta x \approx 2.8 \cdot 10^{-3}$, $T = 1$ and $\Delta t = \frac{1}{64}$).

Example 2: 1D wave system with a shock (2)

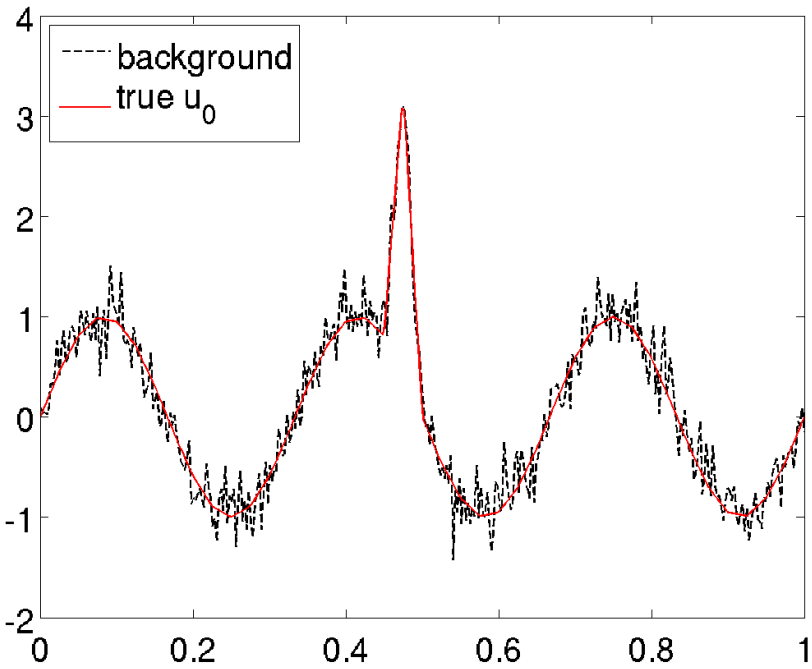


Initial $x_0 = u_0(z)$

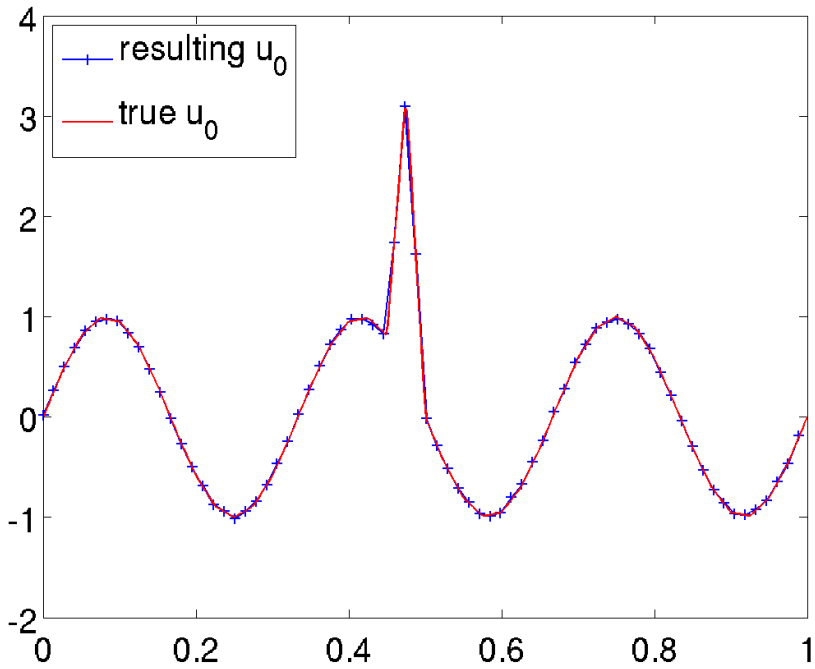


System over space and time

Example 2: 1D wave system with a shock (3)



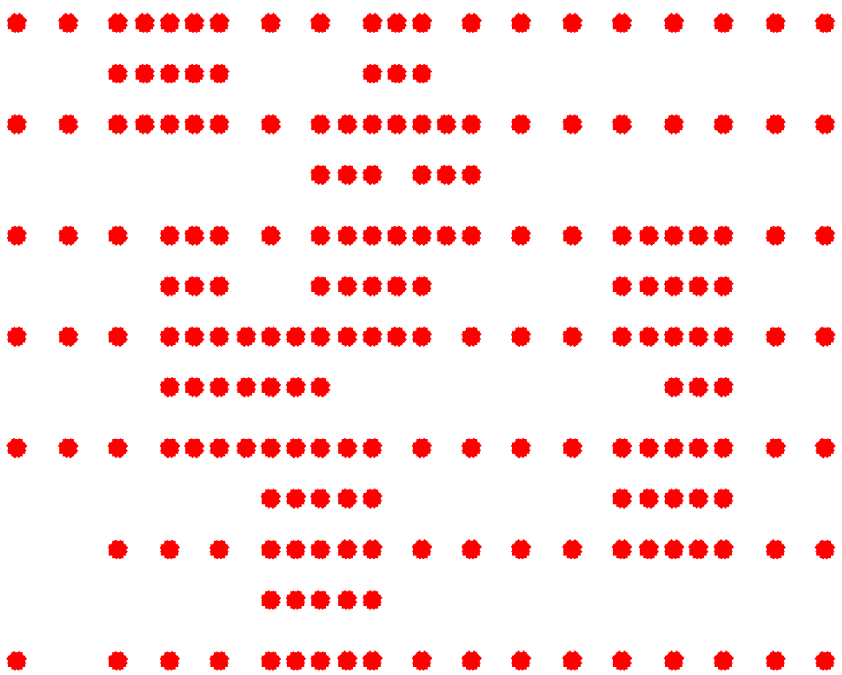
Background and true initial $u(0)$



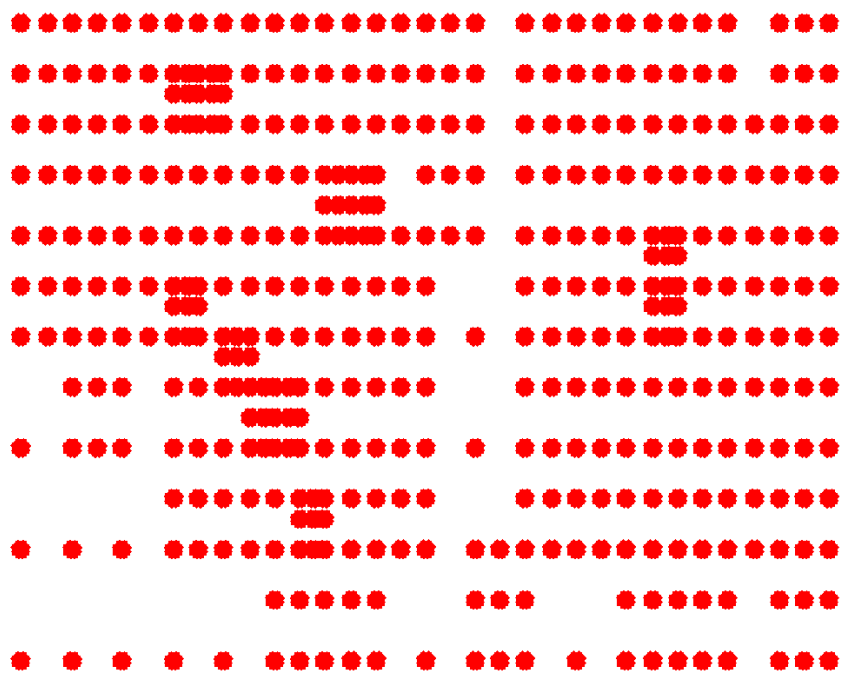
True and computed $u(0)$

Example 2: 1D wave system with a shock (4)

An example of transition from coarse to fine observations sets :

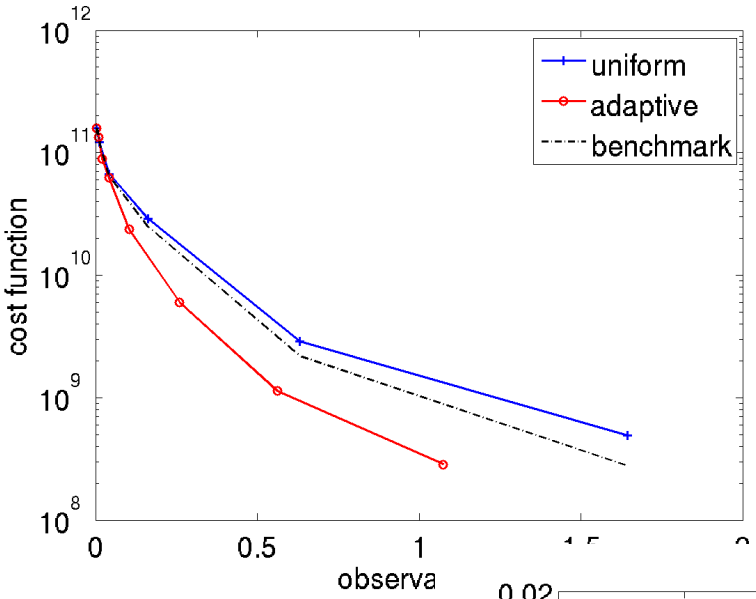


$\mathcal{O}_i$

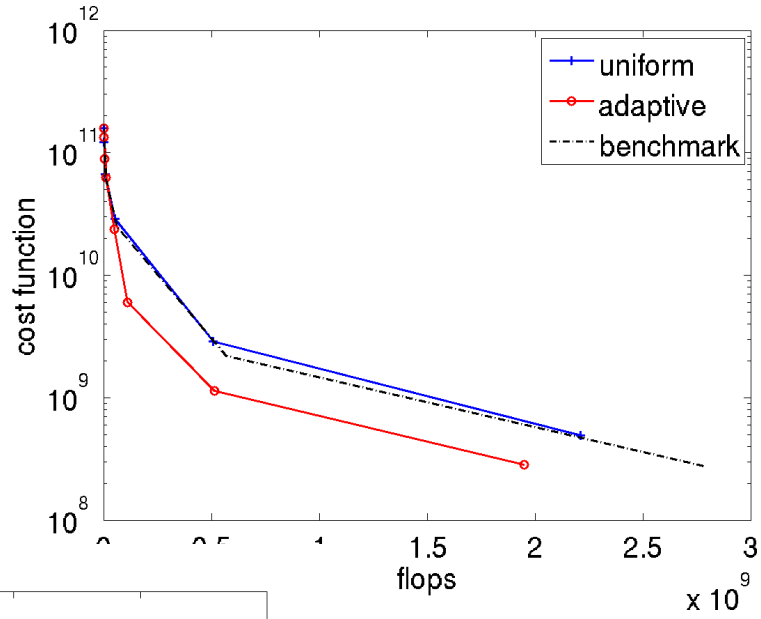


$\longrightarrow$ $\mathcal{O}_{i+1}$

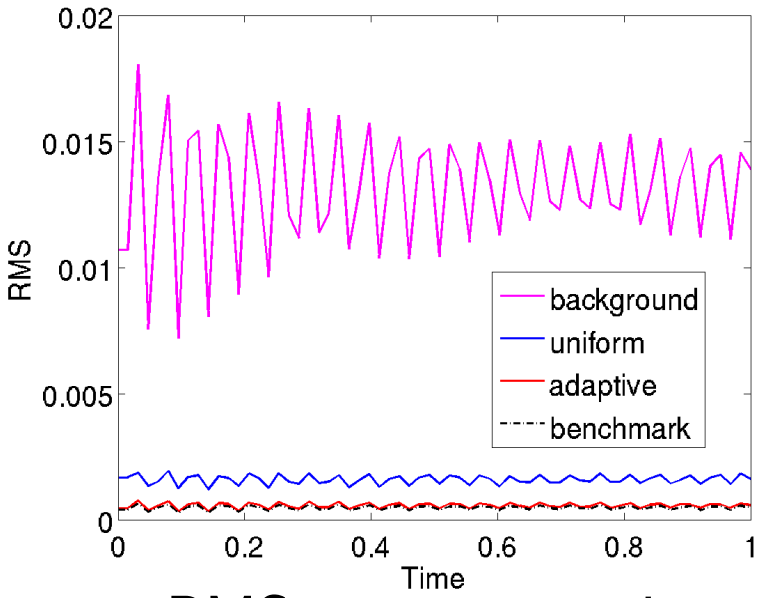
Example 1: 1D wave system with a shock (5)



Cost vs obs



Cost vs flops



RMS error versus time

Conclusions for Section 2

Combining the advantages of the dual approach with adaptive observation thinning is possible

Observation thinning can produce faster solutions

Observation thinning can produce more accurate solutions

Reuse of selected data sets along the nonlinear optimization?

Use this idea for the design of observations campaigns?

Final comments

- large number of algorithmic challenges in 4DVar data assimilation (also true for other approaches such as ensemble methods)
- working in the (possibly thinned) dual space can be very advantageous
- some numerical analysis expertise truly useful
- dialog with practioners not always easy (requires use of real models)
- from-scratch rexamination of the computational chain necessary?

Thank you for your attention!

Reading

- S. Gratton, Ph. L. Toint et J. Tshimanga, "Conjugate-gradients versus multigrid solvers for diffusion-based correlation models in data assimilation", Quarterly Journal of the Royal Meteorological Society, vol. 139, pp. 1481-1487, 2013.
- S. Gratton, S. Gürol, Ph. L. Toint, "Preconditioning and Globalizing Conjugate Gradients in Dual Space for Quadratically Penalized Nonlinear-Least Squares Problems", Computational Optimization and Applications, vol. 54(1), pp. 1-25, 2013..
- S. Gratton, Ph. L. Toint, J. Tshimanga, "Range-space variants and inexact matrix-vector products in Krylov solvers for linear systems arising from inverse problems", SIAM Journal on Matrix Analysis and Applications, vol. 32(3), pp. 969-986, 2011.
- S. Gratton, M. Rincon-Camacho, E. Simon, Ph. L. Toint, "Observations Thinning In Data Assimilation", EURO Journal on Computational Optimization, vol.3, 31-51, 2015.