Introduction to data assimilation. Application to the calibration of a model parameter on the 1D diffusion equation.

B.Delmotte, S.Ricci, M.Rochoux, O.Thual

March 4, 2011

Abstract

This document introduces the basic concepts of data assimilation and presents a simple application of this methodology. The purpose is to calibrate the diffusion parameter for a mono-dimensional diffusion equation using the Best Linear Unbaised Estimator algorithm (further refered to as BLUE). The study is carried out in the context of an Observing System Experiment (OSE) where the observations are generated with the numerical model and represent measurements of the variable state at fixed locations on the computational domain. The data assimilation experiment shows that the estimation of the diffusion parameter using observations of the variable state is feasible. It is also shown in this example that the estimation provided by the BLUE algorithm is not the optimal solution of the real problem since the observation operator that relates the parameter space to the observation space is non linear as it implies the integration of the numerical model. An iterative process of the BLUE is presented here as a valuable solution to overcome some of the limitations of the BLUE in the treatment of the non-linearities problem. This is highlighted by the comparison of the cost function corresponding to the real problem to the cost function associated to the linear problem solved with the BLUE algorithm. The observation operator and its linear approximation are also represented.

Contents

Introduction

Both parameter calibration and physical field description can be formulated as inverse problems [10]. Data assimilation optimally combines all sources of information on a system to produce a more realistic image of its control parameters than if either were taken separately [1, 3, 4]. In this study, data assimilation is used to couple numerical simulations and observations in order to calibrate a specific parameter of a physical problem.

The benefit of data assimilation has already been greatly demonstrated in meteorology [5, 6] and oceanography [2] over the past decades, especially for providing initial conditions for numerical forecast. Data assimilation is now being applied with increasing frequency to other application fields (such as nuclear physics or hydrology). It provides therefore an efficient framework to reduce the uncertainty on the dynamics of a system.

This document proposes a simple application of the Best Linear Unbaised Estimator algorithm to estimate the diffusion parameter for a 1D diffusion equation discretized as a numerical model. It aims at describing the use of data assimilation to correct the estimation of the diffusion coefficient when observation of the state variable are available.

1 Basic concepts in data assimilation

Data assimilation combines numerical and observational information on a system in order to provide a better description of it. While measures are most of the time imperfect and sparse in space and time, the model provides an extensive data of the approximated solution of the physical problem, but its parameters are not precisely known and time and spatial discretizations induce consistency errors. So both observations and model are subject to uncertainties.

So the main objective of data assimilation is to obtain the optimal value of the "real" state of the model, called **analysis**, thanks to the **observations** (measures) and the model parameters estimation (**background**), considering their inaccuracy.

A brief description of data assimilation concepts is presented [9].

1.1 The data assimilation variables

1.1.1 Control vector

The state vector is formed by the finite number of data which represents the discrete state of the model, in other words the simulated field(s) defined at the grid points of the computational domain. The analysis problem is in general not solved for all components of the model state as the dimension of the system can be beyond computer capacities (about 10^6 variables in meteorology for instance). Additionally, it does not only take into account the state variables but also model parameters if they are subject to critical uncertainties. So the idea is to reduce the number of variables to calibrate and to focus the assimilation on those which have the highest uncertainty and to which the system is highly sensitive. These variables are gathered to form the control vector $\mathbf{X}$. This vector is therefore not systematically in the same space as the state vector.

In the context of parameter calibration, the control vector is reduced to a set of n parameters, defining the control space. It is an approximation of the **true control vector** $\mathbf{x^t}$, which represents the true values of these n parameters which are, in reality, completely unknown. Data assimilation aims at determining an optimal value of the control vector, called the **analysis** and denoted by $\mathbf{X}^a$ which is closer to $\mathbf{x^t}$ than the background $\mathbf{X}^b$, the a priori value of the control vector.

1.1.2 Observation vector

Even though they may be sparse in time and space, observations of the system give additional information on the dynamics of the system. Gathered in one vector $\mathbf{Y}^0$ of length p, they represent, in the framework of Observing System Experiment (OSE), measurements of the simulated fields (the state variables) at some fixed spatial locations over the computational domain. They may help to integrate more physical features of the true solution, which are not modeled without data assimilation.

1.1.3 Background information

The information that can be used to produce the analysis is a collection of observed values provided by observations of the true state. In the case of parameter calibration (this tutorial), the model parameters should not reach any aberrant value just to stick to the observations. In order to make the analysis problem more "physical", it is necessary to rely on some background information $\mathbf{X}^b$, an a priori estimate of the model parameters before application of data assimilation ; it is similar to a "first guess" of the model parameters.

When the assimilated value is the state model itself, if the observations are more tightened in space and time than the numerical model solutions, the model state is overdetermined by the observations, then the analysis reduces to an interpolation problem. In most cases the analysis problem is underdetermined (less observations than model solutions) because data is sparse and only indirectly related to the model variables. In this case, to make it a well-posed problem, the background information $\mathbf{X}^b$ is an a priori estimate of the model state.

1.1.4 Observation operator

As the background and observational information need to be combined in the course of data assimilation, an operator H mapping a parameter from the control space onto the observation space is required. This observation operator is generally non-linear as a composition of a model integration M (from model parameters to state variables) and of an interpolation process (from grid points to observational points). In this study, the interpolation process is reduced to a selection process S as the observational points are located at some grid nodes. So H is defined by equation (1).

$$H(\mathbf{X}) = S\left(M(\mathbf{X})\right) \tag{1}$$

In practice, measurements are not perfect and the determination of H implies some assumptions. They are therefore subject to uncertainties, denoted by ϵ^0.

The formulation of the BLUE algorithm requires the linearization of H, denoted by $\mathbf{H}$, which is identified as the Jacobian matrix in the Taylor expansion of H in the vicinity of a reference value of the control vector $\mathbf{X}^g$ (usually the background $\mathbf{X}^b$):

$$H(\mathbf{X}^g + \delta\mathbf{X}^g) = H(\mathbf{X}^g) + \left.\frac{\partial H}{\partial \mathbf{X}}\right|_{\mathbf{X}^g} \delta\mathbf{X}^g + O\left((\delta\mathbf{X}^g)^2\right) \tag{2}$$

with $\lim_{\delta\mathbf{X}^g \to 0} O\left((\delta\mathbf{X}^g)^2\right)\ (\delta\mathbf{X}^g)^{-2} = 0$.

Hence,

$$\mathbf{H} = \left.\frac{\partial H}{\partial \mathbf{X}}\right|_{\mathbf{X}^g} \tag{3}$$

$\mathbf{H}$ is called the linear tangent of H.

1.1.5 Innovation vector

The innovation vector measures the discrepancies between the observation vector and the background projection in the observation space .

$$\mathbf{d}^{ob} = \mathbf{Y}^0 - H(\mathbf{X}^b) \tag{4}$$

1.2 The modeling of the errors

As the true control vector $\mathbf{x}^t$ is unknown, the error on the background $\mathbf{X}^b$, denoted by ϵ^b, and on the observations $\mathbf{Y}^0$, denoted by ϵ^0, are also unknown. As a consequence, the background and observation error covariance matrices, denoted by $\mathbf{B}$ and $\mathbf{R}$ respectively, can only be estimated using an error covariance model.

The modeling of the errors is performed using statistics. If a large number of experiments were undergone under exactly the same conditions, the associated errors would be different each time, but statistics such as expectation value and variance could be established. These two first statistical moments would then converge to values which depend only on the physical processes responsible for the errors, and no longer on any particular realization of these errors. So a statistical model may appear as the reasonable process to approach the errors present in the system.

The best information about the distribution of these errors is given by the Probability Density Function (PDF) function. It describes the relative likelihood for these errors to occur at a given point in the control or observation space, and gives information on their expectation value and variance. This PDF is modeled by a Gaussian function.

1.2.1 Background errors

The background error ϵ^b is defined as the difference between the background $\mathbf{X}^b$ and the true control vector $\mathbf{x}^t$ with $\epsilon^b = \mathbf{X}^b - \mathbf{x}^t$. The statistics of ϵ^b are described in a square symmetric, positive definite matrix $\mathbf{B}$ of size $n \times n$, such that:

$$\mathbf{B} = \mathbb{E}[\epsilon^b \cdot (\epsilon^b)^T] \tag{5}$$

where the diagonal elements of $\mathbf{B}$ represent the error variance for each control parameter, while the off-diagonal terms stand for the covariances between the errors. This means that if $n = 1$, $\mathbf{B}$ is a scalar and is fully described by the variance associated to the single control parameter considered.

$\mathbf{x}^b$ is the realization of the random variable $\mathbf{X}^b$ that is assumed to follow a Gaussian distribution $\mathcal{N}(\mathbf{x}^t, \mathbf{B})$. As the background errors are assumed to be unbiased,

$$\epsilon^b \sim \mathcal{N}(0, \mathbf{B}) \tag{6}$$

1.2.2 Observation errors

As to the background, the observations are represented by a random variable $\mathbf{Y}^0$ that is assumed to follow a Gaussian distribution $\mathcal{N}(H(\mathbf{x}^t), \mathbf{R})$ (provided that $H(\mathbf{x^t})$ is a non-random variable) with the error covariance matrix

$$\mathbf{R} = \mathbb{E}[\epsilon^0 \cdot (\epsilon^0)^T] \tag{7}$$

As the observation error is defined as the difference between the observations and the true control vector projected onto the observation space such that $\mathbf{Y}^0 = H(\mathbf{x}^t) + \epsilon^0$, assuming it is unbiased and uncorrelated, ϵ^0 follows a Gaussian distribution

$$\epsilon^0 \sim \mathcal{N}(0, \mathbf{R}) \tag{8}$$

Furthermore, observed quantities are not necessarily the same as the control variables. The transformations required to map one variable from the control space onto the observation space do not change the analysis problem, only its representation. However, they insert some unvavoidable errors due to the conversion between the observed values and their equivalents in the model state (independently from the observation and background errors), called the representativeness errors. They will not be considered in this case study.

1.3 Principles

Data assimilation provides more realistic physical fields by integrating observations and the physical expertise through a numerical model.

In this context, errors are limited to observations and background. Both follow a Gaussian distribution with a zero average and a specified covariance matrix. The resolution of the data assimilation problem requires therefore four components:

- observations of the physical system $\mathbf{Y}^0$ and their associated errors (ϵ^0, $\mathbf{R}$),
- a background estimation of the control vector $\mathbf{X}^b$ and its associated errors (ϵ^b, $\mathbf{B}$),
- a model M describing the system dynamics,
- an observation operator H with its associated linear tangent operator $\mathbf{H}$.

This defines an **inverse problem** : given $\mathbf{Y}^0$ and $\mathbf{X}^b$, the goal is to approximate as best as possible the true control vector $\mathbf{x}^t$ satisfying

$$\begin{array}{rcl} \mathbf{Y}^0 & = & H(\mathbf{x}^t) + \epsilon^0 \\ \mathbf{X}^b & = & \mathbf{x}^t + \epsilon^b \end{array} \tag{9}$$

2 The Best Linear Unbiased Estimator

By estimating uncertainties on the background and the observations, data assimilation gives an unbiased estimation of the analysis $\mathbf{X}^a$ and minimizes the distance with the true control vector $\mathbf{x}^t$.The BLUE algorithm formulates the analysis as a linear combination of the background $\mathbf{X}^b$ and of the observations $\mathbf{Y}^0$ such that

$$\mathbf{X}^a = \mathbf{L}\mathbf{X}^b + \mathbf{K}\mathbf{Y}^0 \tag{10}$$

with $\mathbf{K}$ the gain matrix and $\mathbf{L}$ a linear operator.

This way, the analysis $\mathbf{X}^a$ approximates the true state $\mathbf{x}^t$.

2.1 The BLUE formulation as a background correction

The BLUE algorithm relies on two major assumptions:

1. The background and observation errors are unbiased.
2. The observation operator is linear : $H = \mathbf{H}$.

Using these assumptions, the linear combination (10) can be written as a background correction, as it will be shown in the following.
From section 1 the data assimilation variables read:

$$\begin{array}{rcl} \mathbf{X}^b & = & \mathbf{x}^t + \epsilon^b \\ \mathbf{Y}^t & = & \mathbf{H}\mathbf{x}^t \\ \mathbf{Y}^0 & = & \mathbf{Y}^t + \epsilon^0 \\ \mathbf{X}^a & = & \mathbf{x}^t + \epsilon^a \end{array} \tag{11}$$

with ϵ^b, ϵ^0 and ϵ^a respectively the background, observation and analysis errors, and $\mathbf{Y}^t$ the true trajectory projected onto the observation space. The background and observation errors are assumed to be unbiased, i.e. $\mathbb{E}[\epsilon^0] = 0$ and $\mathbb{E}[\epsilon^b] = 0$.

By injecting equations (11) in equation (10) it leads to [7] :

$$\begin{array}{rl} \mathbf{x}^t + \epsilon^a & = \mathbf{L}\mathbf{x}^t + \mathbf{L}\epsilon^b + \mathbf{K}\mathbf{Y}^t + \mathbf{K}\epsilon^0 \\ & = \mathbf{L}\mathbf{x}^t + \mathbf{L}\epsilon^b + \mathbf{K}\mathbf{H}\mathbf{x}^t + \mathbf{K}\epsilon^0 \end{array} \tag{12}$$

The BLUE estimator $\mathbf{X}^a$ is also unbiased ($\mathbb{E}[\epsilon^a] = 0$), so taking the expectation value of $\mathbf{x}^t + \epsilon^a$ in Equation (12) leads to:

$$\mathbb{E}[\mathbf{x}^t] = \mathbf{L}\mathbb{E}[\mathbf{x}^t] + \mathbf{K}\mathbf{H}\mathbb{E}[\mathbf{x}^t] \tag{13}$$

hence,

$$\mathbf{L} = \mathbf{I} - \mathbf{K}\mathbf{H} \tag{14}$$

and Equation (10) reads therefore

$$\mathbf{X}^a = \mathbf{X}^b + \mathbf{K}\left(\mathbf{Y^0} - \mathbf{H}\mathbf{X}^b\right) \tag{15}$$

So, the analysis can be considered as a correction of the background with an analysis increment $\delta\mathbf{X}^a = \mathbf{K}(\mathbf{Y}^0 - \mathbf{H}\mathbf{X}^b) = \mathbf{K}\mathbf{d}^{ob}$, with

- $\mathbf{K}$: the gain matrix (see [7] for its calculation),

$$\mathbf{K} = \left(\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{R}^{-1} = \mathbf{B}\mathbf{H}^T\left(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R}\right)^{-1} \tag{16}$$

- $\mathbf{d}^{ob} = \mathbf{Y^0} - \mathbf{H}\mathbf{X}^b$: the innovation vector.

A perfect confidence in the backgound $\mathbf{X}^b$ leads to a zero matrix $\mathbf{B}$ and therefore a zero matrix $\mathbf{K}$ by equation (16). In this context the assimilation correction is zero and $\mathbf{X}^a = \mathbf{X}^b$. Conversely, if the confidence is total in the observations, $\mathbf{R}$ is a zero matrix and then $\mathbf{K} = \mathbf{H}^{-1}$ by equation (16). Thus Equation (10) becomes

$$\mathbf{X}^a = \left(\mathbf{I} - \mathbf{H^{-1}H}\right)\mathbf{X}^b + \mathbf{H}^{-1}\mathbf{Y^0} = \mathbf{H}^{-1}\mathbf{Y}^0$$

So in this case,$\mathbf{X}^a$ is directly the solution of the inverse problem $\mathbf{H}\mathbf{X} = \mathbf{Y}^0$.

.

2.2 The analysis residual

The gain matrix $\mathbf{K}$ provides a posteriori the analysis error covariance matrix $\mathbf{A}$, assuming that both background and observations errors are non-correlated. $\mathbf{A}$ reads

$$\mathbf{A} = \mathbb{E}[\epsilon^a \cdot (\epsilon^a)^T] = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{B} \tag{17}$$

As equation (17) shows, if $\mathbf{H}$ is of full rank (e.g., $\text{rank}(\mathbf{H}) = \min(n, p)$), then $\mathbf{A} \leq \mathbf{B}$, meaning that the BLUE algorithm reduces the error variance of the control variables.

The BLUE formulation can also be tackled with the minimization problem resulting from a data assimilation process [1]

2.3 The 3DVAR formulation

The resolution of most data assimilation algorithms relies on the minimization of a cost function J_{3DVAR} (variational approach referred to as $3DVAR$) defined in equation (18). This quadratic cost function measures the statistically weighted square difference between the background and the control vector on the one hand, the observations and the equivalent of the control vector in the observation space on the other hand :

$$J_{3DVAR}(\mathbf{X}) = \frac{1}{2}\left(\mathbf{X} - \mathbf{X}^b\right)^T\mathbf{B}^{-1}\left(\mathbf{X} - \mathbf{X}^b\right) + \frac{1}{2}\left(\mathbf{Y^0} - H(\mathbf{X})\right)^T\mathbf{R}^{-1}\left(\mathbf{Y^0} - H(\mathbf{X})\right) \tag{18}$$

with H the non-linear observation operator, whose linearization in the vicinity of $\mathbf{X}^b$ is denoted $\mathbf{H}$ so that $\mathbf{H} = \frac{\partial H}{\partial \mathbf{X}}\big|_{\mathbf{X}^b}$.

The minimization of this cost function consists in finding the optimal "compromise" between the background and the observations depending on their precision ($\mathbf{B}^{-1}$ and $\mathbf{R}^{-1}$ respectively).
The gradient J_{3DVAR} reads:

$$\nabla J_{3DVAR}(\mathbf{X}) = \mathbf{B}^{-1}\left(\mathbf{X} - \mathbf{X}^b\right) + \mathbf{H}^T\mathbf{R}^{-1}\left(\mathbf{Y^0} - H(\mathbf{X})\right)$$

The objective of data assimilation is to find $\mathbf{X}^a = argmin\left(J_{3DVAR}(\mathbf{X})\right)$ or equivalently :

$$\nabla J_{3DVAR}(\mathbf{X}^a) = 0 \tag{19}$$

$$\Rightarrow \mathbf{B}^{-1}\left(\mathbf{X}^a - \mathbf{X}^b\right) + \mathbf{H}^T\mathbf{R}^{-1}\left(\mathbf{Y^0} - H(\mathbf{X}^a)\right) = 0 \tag{20}$$

The 3DVAR approach implies the use of a minimizor that requires the evaluation of the cost function and its gradient at each iteration, whereas the BLUE algorithm aims at calculating an analytic solution to this problem using a linearity assumption.

2.4 Equivalence between BLUE and 3DVAR under the linearity assumption

In order to find an analytic solution to the minimization problem (19) the BLUE formulation requires a linearization of the observation operator H (see section 1.1.4).
If H is considered linear, then $H = \mathbf{H}$.
Thus, Equation (20) becomes :

$$\begin{aligned} 0 &= \mathbf{B}^{-1}\left(\mathbf{X}^a - \mathbf{X}^b\right) + \mathbf{H}^T\mathbf{R}^{-1}\left(\mathbf{Y^0} - \mathbf{H}\mathbf{X}^a\right) \\ \Rightarrow \mathbf{X}^a &= \mathbf{X}^b + \left(\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{R}^{-1}\left(\mathbf{Y^0} - \mathbf{H}\mathbf{X}^b\right) \\ \Rightarrow \mathbf{X}^a &= \mathbf{X}^b + \mathbf{K}\left(\mathbf{Y^0} - \mathbf{H}\mathbf{X}^b\right) \end{aligned} \tag{21}$$

$$\Rightarrow \mathbf{X}^a = \mathbf{X}^b + \mathbf{K}\mathbf{d}^{ob} \tag{22}$$

So, if H is linear, equation (19) leads directly to equation (22) which is exactly the same as equation (15).

Consequently, **the 3DVAR and BLUE approaches are equivalent if H is linear** [1].

2.5 Which of 3DVAR and BLUE is the most efficient method ?

In the case of data assimilation when calibrating only a few parameter, the BLUE approach is more efficient than 3DVAR. Calculating an analytic solution using a linear approximation is indeed less costly **[7]** than the evaluation of the cost function and its gradient at each iteration.

However, when the number of control variables increases and/or the observation operator H is highly non-linear, 3DVAR direct minimization may be less costly and/or more precise in the definition of the minimum of J_{3DVAR}.[1]

2.6 External loops : analysis as an iterative process

When using the BLUE approach to calibrate the parameter of a model, the solution can be distorted depending on the level of non-linearity of the observation operator. In this case, it could be interesting to calculate the analysis as an iterative process to be more precise in the minimization problem.

In this configuration, the analysis results from a succession of corrections on the background. For each iteration $\mathbf{H}$ is updated at a new reference control parameter which is the analysis of the previous iteration, and a new increment $\delta\mathbf{X}^a$ is calculated. The process is fully described in the following algorithm.

Algorithm :

1. Determination of the true trajectory.
2. Modeling of the observation error covariance matrix $\mathbf{R}$.
3. Modeling of the background error covariance matrice $\mathbf{B}$.
4. Construction of the observations $\mathbf{Y}^0$, directly extracted from the true state.

[1] To evaluate the efficiency of theBLUE method, it is interesting to plot the cost function J_{3DVAR} and its linearized form J_{BLUE} minimized with the BLUE algorithm, according to the control parameters (if $n \leq 2$). This allows to visually compare $J_{3DVAR}(\mathbf{X}^a)$ and $\min_{\mathbf{X}} J_{3DVAR}(\mathbf{X})$ (see section 4).

5. Determination of the background trajectory $M(\mathbf{X}^b)$ and projection onto the observation space: $H(\mathbf{X}^b)$.
6. Start of external loops:
   (a) Determination of the reference control vector $\mathbf{X}^g$:
      i. -at iteration 1, $\mathbf{X}^g = \mathbf{X}^b$.
      ii. -for all other iterations, $\mathbf{X}^g = \mathbf{X}^a$.
   (b) Determination of the reference trajectory $M(\mathbf{X}^g)$ and projection onto the observation space: $H(\mathbf{X}^g)$.
   (c) Determination of the new "innovation vector" : $\mathbf{d}^{og} = \mathbf{Y}^0 - H(\mathbf{X}^g)$
   (d) Construction of the tangent linear of the observation operator $\mathbf{H} = \frac{\partial H}{\partial \mathbf{X}}\big|_{\mathbf{X}^g}$ in the vicinity of the reference control parameter $\mathbf{X}^g$.
   (e) BLUE application.
   (f) Integration of the analysed trajectory $M(\mathbf{X}^a)$.
7. End of external loops.

Attention must be paid in **Point 6.(e)** [11]:

By taking $\mathbf{X}^g = \mathbf{X}^a$, the formulation of the BLUE differs from the "classical" one. It can be demonstrated by starting with the incremental cost function formulation :

$$J_{inc}(\mathbf{X}, \mathbf{X}^g) = \frac{1}{2}\left(\mathbf{X} - \mathbf{X}^b\right)^T \mathbf{B}^{-1}\left(\mathbf{X} - \mathbf{X}^b\right) + \frac{1}{2}\left(\mathbf{Y}^0 - H(\mathbf{X}^g) - \mathbf{H}(\mathbf{X} - \mathbf{X}^g)\right)^T \mathbf{R}^{-1}\left(\mathbf{Y}^0 - H(\mathbf{X}^g) - \mathbf{H}(\mathbf{X} - \mathbf{X}^g)\right) \tag{23}$$

Given that $\mathbf{X}^a = \arg\min\left(J_{inc}\right)$, it follows

$$\frac{\partial J_{inc}}{\partial \mathbf{X}}(\mathbf{X}^a, \mathbf{X}^g) = 0 \tag{24}$$

$$\Rightarrow \mathbf{B}^{-1}\left(\mathbf{X}^a - \mathbf{X}^b\right) - \mathbf{H}^T\mathbf{R}^{-1}\left(\mathbf{Y}^0 - H(\mathbf{X}^g) - \mathbf{H}(\mathbf{X}^a - \mathbf{X}^g)\right) = 0$$

Using the same calculation as in section 2.4 the new analysis reads:

$$\Rightarrow \mathbf{X}^a = \mathbf{X}^b + \mathbf{K}\left(\mathbf{d}^{og} + \mathbf{H}(\mathbf{X}^g - \mathbf{X}^b)\right) \tag{25}$$

with $\mathbf{d}^{og} = \mathbf{Y}^0 - H(\mathbf{X}^g)$ the difference between the observation vector and the projection of the reference control vector onto the observation space.

3 Parameter Calibration for a 1D Sample Model

On the following, the BLUE algorithm is applied to the calibration of the diffusion coefficient of a 1D diffusion equation.

This study is carried out in the context of OSE, each step of data assimilation is described.

3.1 Model configuration

3.1.1 Diffusion equation

The 1D diffusion equation of a scalar quantity $U(x,t)$ (e.g. the temperature) is defined by the following PDE (e.g. the heat equation) :

$$\begin{cases} \frac{\partial U}{\partial t}(x,t) = K\,\frac{\partial^2 U}{\partial x^2}(x,t) & , t \in [0,T_f],\ x \in [0,L] \\ U(x,0) = U_0(x) = U_{amp} e^{-\left(\frac{(x-0.5L)^2}{2} s\right)} & , x \in [0,L]\,,\ \text{Initial condition} \\ U(0,t) = 0 & , t \in [0,T_f]\,,\ \text{Dirichlet boundary condition} \\ \frac{\partial U}{\partial n}(x,t) = 0 & , t \in [0,T_f]\,,\ \text{Neumann boundary condition} \end{cases} \tag{26}$$

with

- K : the diffusion coefficient.
- T_f : the final time of simulation.
- $L = 5 \cdot 10^5$: the geometry length.
- $U_0(x)$: the initial condition characterized by
  - $U_{amp} = 1$: the maximal amplitude of the initial temperature distribution.
  - $e^{-\left(\frac{(x-0.5L)^2}{2} s\right)}$: the gaussian distribution centered on $x = \frac{L}{2}$ (see figure 1).

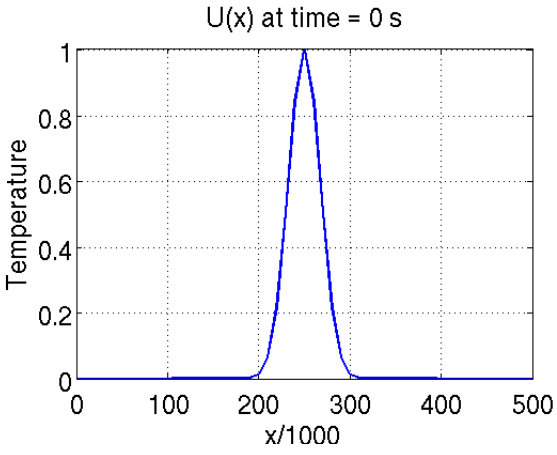


Figure 1: Initial condition $U_0(x)$

3.1.2 Choice of the numerical configuration

For this simple prototype, the computational configuration is chosen so as to keep a low computational cost and an immediate visualization. For this purpose, the size of the matrices and vectors involved in the assimilation process shall keep a relatively small size.

- So all the data assimilation tests are carried out for a mesh of $N_x = 51$ nodes : $x_1 = 0$; ... ; $x_{N_x} = L$, with $dx = \frac{L}{N_x - 1} = 10^4$.
- The integration time step is also chosen so as to reduce calculation costs : $dt = 100$, with $N_t = \left\lfloor \frac{T_f}{dt} \right\rfloor$ the number of time iterations.

Problem (26) is solved using a second order finite difference scheme in space and an Explicit Euler scheme in time. Its discretization gives :

$$\begin{cases} \frac{U_i^{n+1} - U_i^n}{dt} = K \frac{dt}{dx^2} \left(U_{i+1}^n - 2U_i^n + U_{i-1}^n \right) & , \ i \in \{2, ..., N_x - 1\}, \ n \in \{1, ..., N_t\} \\ U_i^1 = U_0(x_i) & , \ i \in \{1, ..., N_x\} \\ U_1^n = 0 & , \ n \in \{1, ..., N_t\} \\ U_{N_{x-1}}^n = U_{N_x}^n & , \ n \in \{1, ..., N_t\} \end{cases} \tag{27}$$

The stability condition for the diffusion equation using a second-order finite difference scheme and an Explicit Euler method is

$$K \frac{dt}{dx^2} < \frac{1}{2} \tag{28}$$

Considering the values of dt and dx, the stability condition becomes :

$$K < 5 \cdot 10^5 \tag{29}$$

This stability condition will always be satisfied in the following example.

3.2 The twin experiments, a validation framework

In the context of Observing System Experiment (OSE), observations result from measurements of the simulated fields (the state variables) at some fixed spatial locations over the computational domain. More precisely, in the framework of twin experiment, the true value $\mathbf{x}^t$ of the control parameters is known and the analysis is obtained thanks to the following process (see figure (2)) :

1. From the true state $\mathbf{x}^t$, an artificial observation vector is calculated :
   (a) The model is integrated with the true value $\mathbf{x}^t$. All the values of the "true" solution in space and time are gathered into the vector $U^t(x,t) = M(\mathbf{x}^t)$ called the "true trajectory".
   (b) Selection of the observation nodes: $H(\mathbf{x}^t) = S\left(M(\mathbf{x}^t)\right)$.
   (c) To be more representative of a real case, the observation vector $\mathbf{Y}^0 = H(\mathbf{x}^t) + \epsilon_0$ is then generated by adding an artificial noise ϵ_0 to $H(\mathbf{x}^t)$ whose probability function is a centered Gaussian distribution of variance σ_0^2 (see section 1.2.2). The value of σ_0 is arbitrarily chosen.
2. The background $\mathbf{X}^b$ is calculated as a perturbation of the true control vector and its standard deviation σ_b is also known (see section 3.4). $H(\mathbf{X}^b)$ and its linear tangent $\mathbf{H}$ can therefore be calculated.
3. Once the required components ($\mathbf{Y}^0, \mathbf{R}, \mathbf{X}^b, \mathbf{B}, H(\mathbf{X}^b)$ and $\mathbf{H}$) are calculated, the BLUE algorithm can be applied. Then, the objective is to compare the resulting analysis to the true state with various diagnostics (see section 3.3.5) in order to evaluate the consistency and optimality of the solution.

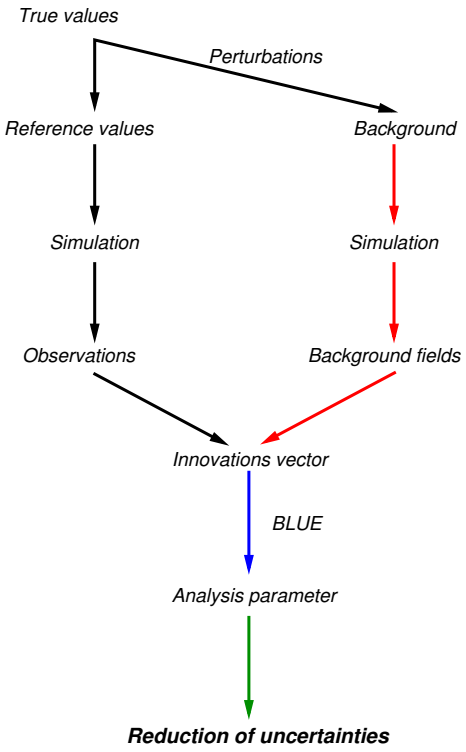


Figure 2: Schematic representation of a twin experiment

So, by simulating the observations, the background state and their respective errors from the true state, the twin experiment is useful to evaluate the robustness of the correction algorithm (BLUE) with calibrated perturbations of the observations and background uncertainties. That is why twin experiments can be considered as a validation framework for this sample model. This assimilation context is deliberate as the goal of this tutorial is to focus on the correction algorithm.

3.3 Steps of the data assimilation process

3.3.1 Integration of the true and background trajectories

The goal of this tutorial is to implement the calibration of the diffusion parameter K present in Equation (26). So the control vector $\mathbf{X}$ is reduced to a scalar ($n = 1$), with K as the single component.
In the context of a twin experiment, the true value of K is supposed to be known and is denoted by K_t. The Equation (26) is integrated for $K = K_t$ to obtain the true trajectory $U^t(x,t) = M(K_t)$.
From this true trajectory are directly extracted p artificial observations with a fixed frequency in space and time :

- There is a single observation located in the middle of the domain at $x_{obs} = \frac{L}{2}$ at a given time.
- The frequency in time is $f_{obs} = \frac{1}{T_{obs}}$, with T_{obs} specified by the user. So, in this case

$$p = 1 \times \left\lfloor \frac{T_f}{T_{obs}} \right\rfloor = 1 \times N_{obs}$$

with N_{obs} the number of observations in time.

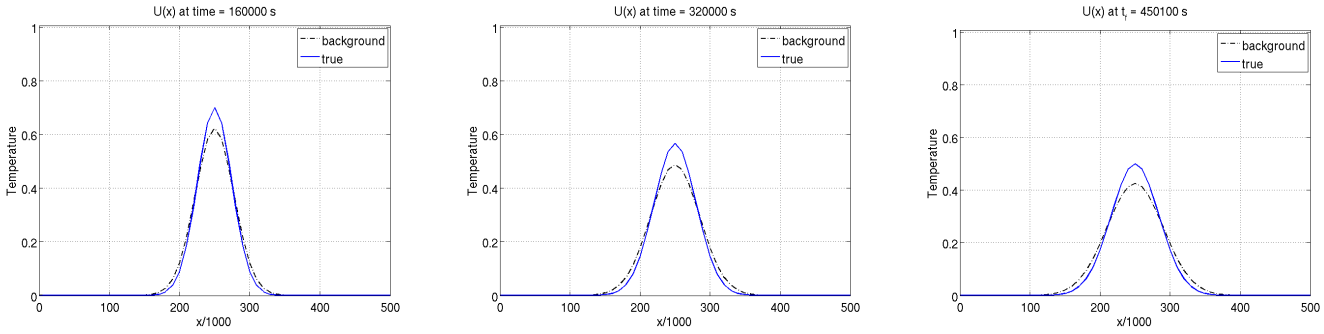
- Then, the observation vector $\mathbf{Y}^0$ is calculated adding artificial gaussian noise:

$$\mathbf{Y}^o = H(K_t) + \epsilon_0 = \left[\underbrace{U^t(x_{obs},\,(k-1)T_{obs},K_t)}_{H(K_t)_k} + \underbrace{\sigma_0 \times G(0,1)_k}_{(\epsilon_0)_k} \right]_{k=1,..,\,N_{obs}}$$

with $G(0,1)$: a vector containing pseudo-random values drawn from a normal distribution with mean 0 and standard deviation 1.

The background vector $\mathbf{X}^b$ has for single component the scalar K_b, which is specified as a perturbation of K_t and which is associated to the background trajectory $U^b(x,t) = M(K_b)$.
After the selection operation and the calculation of $\mathbf{H}$, the BLUE algorithm corrects K_b via the analysis increment $\delta K_a = \mathbf{K}(\mathbf{Y}^0 - \mathbf{H}K_b)$.

In Figure (3), $U^b(x,t)$ is represented for $K_b = 1500$. The diffusion is therefore more intense for the background than for the true parameter ($K_t = 1000$), that is why the magnitude of the background trajectory is lower than the magnitude of the true trajectory at each time step.



(a) Temperature distribution at $t = 16000$ (b) Temperature distribution at $t = 32000$ (c) Temperature distribution at $t = T_f$

Figure 3: Temperature distribution for $K_t = 1000$ and $K_b = 1500$

3.3.2 Formulation of the non-linear observation operator H

K_b and $\mathbf{Y}^0$ do not evolve in the same space, so an observation operator H is required to map the control parameters onto the observation space. In this context, $H(K_b)$ gives the value of the background trajectory at each observation point :

$$H(K_b) = [U(x = x_{obs}, t = (k-1)T_{obs}, K_b)]_{k=1,\ldots,\,N_{obs}} \tag{30}$$

Thus, H is a composition of a model integration M and of a selection matrix S. Only the second step is linear as the solution describing the diffusion process depends non-linearly on the diffusion coefficient K in equation (26). H is therefore non-linear.

3.3.3 Determination of the tangent linear of H

The assimilation algorithm requires a linearization $\mathbf{H}$ of the observation operator H so as to define the optimal matrix gain $\mathbf{K}$. $\mathbf{H}$ results from the Taylor expansion of H in the vicinity of a reference diffusion parameter, denoted K_g

$$H(K_g + \delta K_g) = H(K_g) + \left.\frac{\partial H}{\partial K}\right|_{K_g} \delta K_g + O\left((\partial K_g)^2\right) \tag{31}$$

with $\lim_{\delta K_g \to 0} O\left((\partial K_g)^2\right) (\delta K_g)^{-2} = 0$.

By definition, the tangent linear of H at a reference parameter K_g, also called the linearized observation operator $\mathbf{H}$, is the Jacobian matrix in the Taylor expansion (31) and is approximated using a non-centered finite-difference scheme. Hence,

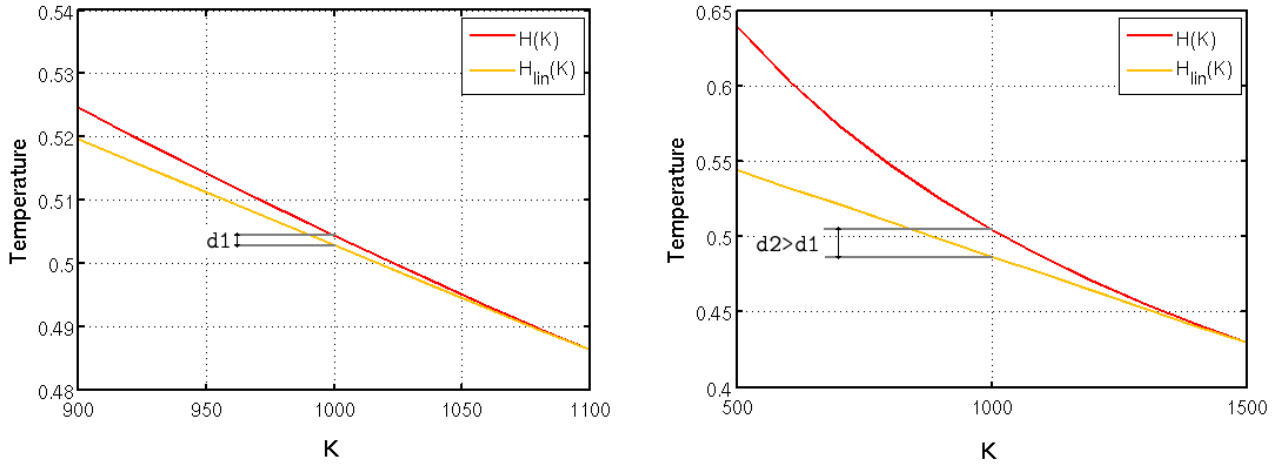
$$\mathbf{H} = \left.\frac{\partial H}{\partial K}\right|_{K_g} \approx \frac{H(K_g + \delta K_g) - H(K_g)}{\delta K_g} \tag{32}$$

In practice, $\mathbf{H}$ results from the Taylor expansion in the vicinity of the background K_b. The computation of $\mathbf{H}$ requires the integration of an additional trajectory corresponding to a perturbation of K_b, denoted by $K_b + \delta K_b$. Then, both trajectories ($M(K_b)$ and $M(K_b + \delta K_b)$) are projected onto the observation space through the selection operator S. The tangent linear $\mathbf{H}$ can be therefore computed for a prescribed perturbation δK_b using equation (32).[2]

As H is non-linear, the quality of the linear approximation is sensitive to the perturbation δK_b introduced to compute $\mathbf{H}$. By construction, equation (32) is valid when the analysis is in the interval $K_b; K_b + \delta K_b]$, but its validity is not ensured outside of this interval. Consequently, the sensitivity of $\mathbf{H}$ shall be carefully studied as a function of the control parameter so as to guarantee the optimality of the solution when $|K_a - K_b| > \delta K_b$ (see Figure 4). In this example, the perturbation δK_b is arbitrarily taken equal to $0.05 K_b$.

Indeed, it is important to notice that the analysis coefficient K_a shall be in the interval $[K_t; K_b]$: if the confidence in the background is high, the correction δK_a will be low and K_a will be close to K_b, whereas K_a will be close to K_t if the observations are granted a high confidence. This result can only be guaranteed if the linear tangent of the observation operator $\mathbf{H}$ is valid along the interval $[K_t; K_b]$. It might not be the case if the distance between K_b and K_t is too large as H is non-linear (figure 4).

[2] It is more precise to calculate $\mathbf{H}$ with a centered finite-difference scheme $\frac{H(K_g+\delta K_g)-H(K_g-\delta K_g)}{2\delta K_g}$ but it implies calculating two additional trajectories : $M(K_g + \delta K_g)$ and $M(K_g - \delta K_g)$ instead of one for the non-centered scheme ($M(K_g + \delta K_g)$). [8]



(a) $H(K)$ and its linear tangent in the vicinity of$K_b = 1100$ (b) $H(K)$ and its linear tangent in the vicinity of $K_b = 1500$

Figure 4: H and its linear tangent $\mathbf{H}$ for $[K_t, K_b] = [1000, 1100]$ and for $[K_t, K_b] = [1000, 1500]$. The discrepancies between H and its linear tangent for a given time at the observation point ($x_{obs} = \frac{L}{2}$) are higher when $[K_t, K_b]$ is larger because the impact of non-linearities is stronger. For example, for $K_t = 1000$, $d2 \simeq 10d1$.

3.3.4 Modeling of the error covariance matrices

For one parameter calibration, $\mathbf{B}$ is a scalar, namely the error variance σ_b^2 associated to the diffusion coefficient. As to the observation error covariance matrix $\mathbf{R}$, its size is $p \times p$ with p the dimension of the observation vector $\mathbf{Y}^o$. In this study, it is assumed that observation errors are not correlated so that $\mathbf{R}$ is reduced to a diagonal matrix (which is a common assumption in data assimilation). In this study observations are prescribed the same standard deviation σ_0 (see section 3.3.1).

3.3.5 Diagnostics for validation

The BLUE algorithm ensures that the analysis error variance is smaller than the background error variance (see section 2.2). Provided that the observation operator is linear, the analysis is in the interval $[\mathbf{X}^b; \mathbf{x}^t]$ for an OSE experiment and $\mathbf{X}^a$ satisfies $Tr(\mathbf{A}) = \sigma_a^2 < Tr(\mathbf{B}) = \sigma_b^2$. In this case the assimilation is only computed once, so this diagnostic can be formulated easily :

$$\text{RMS(TMA)} = \left|\mathbf{x}^t - \mathbf{X}^a\right| \leq \text{RMS(TMB)} = \left|\mathbf{x}^t - \mathbf{X}^b\right| \tag{33}$$

Another common diagnostic in data assimilation is to verify that the analysis error variance is also reduced in the observation space. This implies that the variance of the difference between the analysis trajectory and the observation is smaller than the variance of the difference between the background trajectory and the observation [8] :

$$d^{oa} = \text{RMS(OMA)} = \text{RMS}\left(\mathbf{Y}^0 - H(\mathbf{X}^a)\right) \leq d^{ob} = \text{RMS(OMB)} = \text{RMS}\left(\mathbf{Y}^0 - H(\mathbf{X}^b)\right) \tag{34}$$

3.4 Implementation of data assimilation

The aim of this study is to compare the following two assimilation methods :

1. Twin experiments with two possible variants :

(a) σ_b is consistent with the perturbation of K_t such that $\sigma_b = K_b - K_t$. This is a rigourous twin experiment.

(b) σ_b is chosen arbitrarily. So it is not consistent with the prescribed perturbation of $K_b - K_t$. However, this is close to a real case wher $\mathbf{x}^t$ is completely unknown and σ_b has to be modeled.

2. External loops : as explained in section 2.6, the assimilation is implemented as an iterative process. The analysis results from a succession of corrections on the background, and for each correction $\mathbf{H}$ is updated at a new reference diffusion coefficient (e.g., the analysis of the previous iteration). The two variants are also available for this assimilation method.

4 Results

Various configurations are tested to show some sensitive aspects of the BLUE algorithm. It is then proved that these sensitivity problem can be overcome thanks to external loops.

4.1 Optimality of the BLUE solution in the context rigourous twin experiments[3]

A series of tests is carried out for $K_t = 1000$ in the context of rigourous twin experiments. The observation error variance is set to $\sigma_0^2 = 0.01^2$, $\sigma_0^2 = 0.1^2$ and $\sigma_0^2 = 1$. The difference between the background and the true diffusion coefficient is set to 100 or 500 (this difference sets the background error variance : $\sigma_b = |K_b - K_t|$).
The temperature at the point $x_{obs} = \frac{L}{2}$ is measured each two iteration : $f_{obs} = \frac{1}{T_{obs}} = \frac{1}{200} = \frac{1}{2dt}$.
The simulation time T_f is chosen such as

$$U^t\left(t = T_f, x = \frac{L}{2}\right) = 0.5U_0\left(x = \frac{L}{2}\right)$$

The results of the BLUE analysis are given in Table 1, and the BLUE and 3DVAR cost funtions are presented on figure 5.

	E1	E2	E3	E4	E5	E6
σ_0	0.01	0.1	1	0.01	0.1	1
K_b	1100	1100	1100	1500	1500	1500
K^a_{BLUE}	**996.7**	**1055**	**1099**	**883**	**963**	**1379**
$J_{3DVAR}(K_t)$	8.90	11.92	13.93	21	9.31	8.55
$J_{3DVAR}(K_b)$	38.9	11.83	13.44	597.2	14.61	8.33
$J_{3DVAR}(K^a_{BLUE})$	9.01	11.67	13.44	73.91	9.42	8.29
RMS(AMO)	0.0086	0.1003	1.081	0.0252	0.0877	0.8477
RMS(BMO)	0.0184	0.1014	1.081	0.0721	0.1127	0.8510

Table 1: Description of the BLUE analysis experiments for $K_t = 1000$

The red color indicates that the analysis is not the optimal value of the real cost function J_{3DVAR} : $J_{K_a} \in]J_{K_t}; J_{K_b}]$. It means that the minimum of J_{BLUE} is too far from the minimum of the non-linear J_{3DVAR} cost function.

The blue color indicates that the analysis is close to the optimal value of J_{3DVAR}:

$$J_{3DVAR}(K^a_{BLUE}) \leq \min\left(J_{3DVAR}(K_t)\,;\, J_{3DVAR}(K_b)\right)$$

[3]The results are similar with the inconsistent twin experiment.

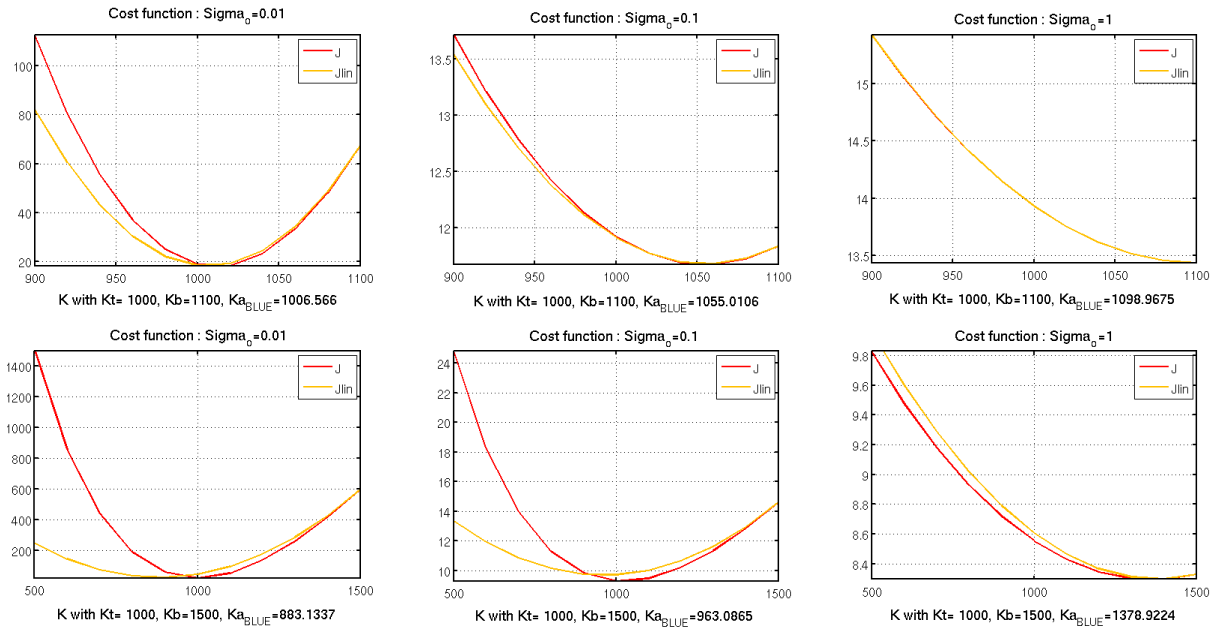


Figure 5: Cost functions J_{3DVAR} (red curves) and J_{BLUE} (orange curves) for different background and observation error variances for experiments E1 to E6.

It can be noted that :

1. The diagnostic RMS(AMO) < RMS(BMO) is always satisfied : the analysis trajectory is in average closer to the observations than the background trajectory.
2. If the uncertainty on the observations is high:
   (a) The analysis stays close to the model than to the observations which are not truth worthy.
   (b) Therefore the analysis is very close to the background and there is no effect of the non-linearities. The optimal value of J_{3DVAR} and J_{BLUE} are close. The BLUE solution is a good approximation of the solution of the real non linear problem given by the 3DVAR (E3, E6).
3. If the uncertainty on the observation is low:
   (a) The analysis is close to the true diffusion coefficient.
   (b) If the background value is far from the true value, the linear approximation for H computed at the background is not valid for the analysis. J_{3DVAR} and J_{BLUE} significantly differ. As a consequence the background is "over-corrected" ($K_a \notin [K_t; K_b]$) and the BLUE optimal solution is different from the 3DVAR optimal solution. The minimum of the orange and red curves are significantly different. The BLUE solution is not the optimal solution of the real diffusion problem (E4,E5).
   (c) To a lesser extent, even if the background is close to the true value, the analysis is potentially closer to K_t. Again, the linear hypothesis for H is not valid for the analysis and the BLUE solution is not the optimal solution for the non-linear problem (E1).

When the analysis is potentially far from the background (E1, E4 and E5), the linear hypothesis is not valid. To curb the impact of these non-linearities, the analysis can be calculated iteratively by updating the linear tangent of **H** at a new reference diffusion coefficient (e.g. the analysis of the previous iteration) (see section 2.6).

4.2 Improvement of the calibration with the use of external loops for sensitive cases.

A four-loop iterative process is launched for the sensitive configurations E1, E4 and E5. The results are given in Table 2, and the successive incremental and 3DVAR cost funtions are presented on Figure 6.

	E1	E4	E5
σ_0	0.01	0.01	0.1
K_b	1100	1500	1500
K^a_{BLUE}	**996.7**	**883**	**963**
K^a_{LOOPS}	**1000**	**1001**	**1024**
$J_{3DVAR}(K_t)$	8.90	21	9.31
$J_{3DVAR}(K_b)$	38.9	597.2	14.61
$J_{3DVAR}(K^a_{BLUE})$	9.01	73.91	9.42
$J_{3DVAR}(K^a_{LOOPS})$	8.87	21	9.29
RMS(AMO)$_{BLUE}$	0.0086	0.0252	0.0877
RMS(AMO)$_{LOOPS}$	0.0085	0.0134	0.0877
RMS(BMO)	0.0184	0.0721	0.1127

Table 2: Comparison between external loops and BLUE. K^a_{LOOPS} stands for the analysis at the end of the iteration process.

The improvements are obvious :

1. The analysis K^a_{LOOPS} is always in the interval $[K_t; K_b]$.
2. For each configuration K^a_{LOOPS} is close to the optimal value of J_{3DVAR}:

$$J_{3DVAR}(K^a_{LOOPS}) \leq \min\left(J_{3DVAR}(K_t)\,;\, J_{3DVAR}(K_b)\right)$$

All the orange curves are close to the red curve in the vicinity of the optimal value of J_{3DVAR}.

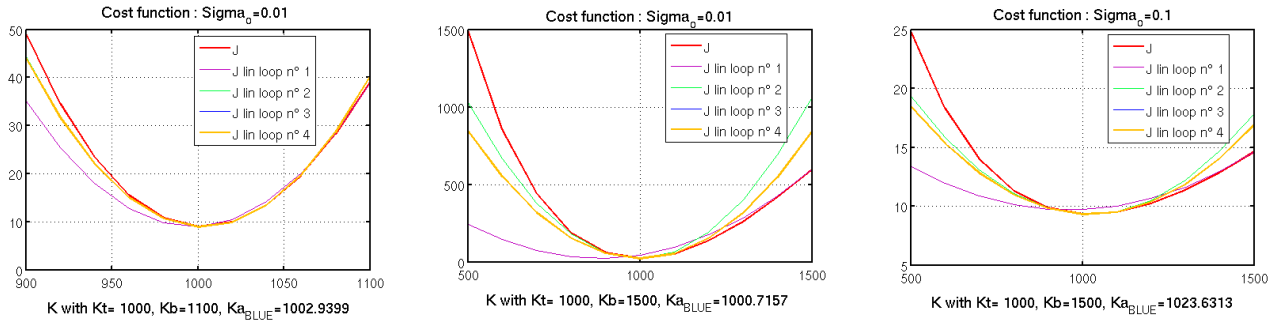


Figure 6: Cost functions J_{3DVAR} (red curves) , $J^{final}_{inc} = J^{(4)}_{inc}$ (orange curves) and incremental $J^{(i)}_{inc}$, $i = \{1, ..., 3\}$ (other colors) for different background and observation error variances for experiments E1, E4 and E5. The incremental cost functions J^i_{inc} get closer to J_{3DVAR} around the optimal value. The violet curve corresponds to the previous J_{BLUE}.

This algorithm reduces step by step the distance between the reference point K_g and the optimal value of J_{3DVAR}. The closer K_g to the optimal value is, the better the linear approximation in the interval $[K_g; K_g+\delta K_g]$ is (see Figure 7).

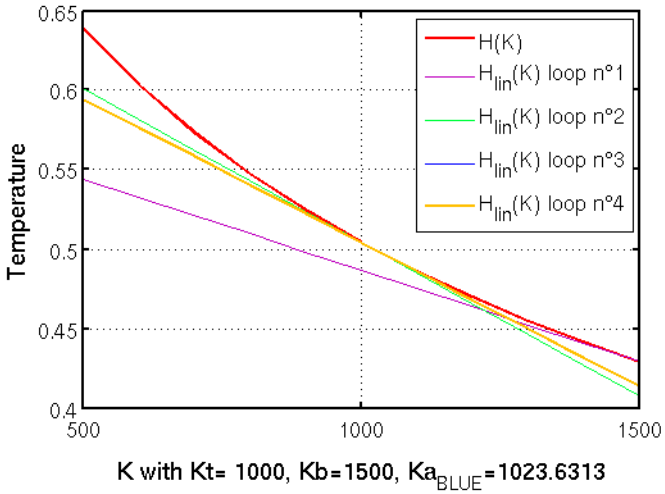


Figure 7: H (red curve) and its successive linear tangents $\mathbf{H}$ in the vicinity of K_g (other colors) for experiment E5. The violet curve is the first one : it corresponds to the linear tangent calculated in the vicinity of K_b. The orange curve approximates H in the vicinity of the optimal value of J_{3DVAR} : the linear hypothesis is therefore valid.

Conclusion

The different experiments show that the BLUE solution is particularly sensitive to the choice of the background K_b compared to K_t, its uncertainty and the uncertainty of the observations. The errors introduced in the analysis by the construction of $\mathbf{H}$ with respect to K_b are due to the magnitude of the non-linearities of H. This magnitude depends on the distance between the true and background coefficients. It increases usually when K_b is further away from K_t. When the sensitivity is too high, these non-linearities can be handled thanks to external loops. Though this process is more costly, it provides a solution in better agreement with the optimal solution (minimum of the red curves) than the classical BLUE algorithm for the sensitive cases identified in section 4. So the external loops approach may appear as a valuable solution to calibrate a few model parameters even though the model is subject to some non-linearities.

Glossary

Random variables are usually denoted in capital letters ($\mathbf{X}$, $\mathbf{Y}$,...), whereas one of their realizations is denoted in minuscule letters ($\mathbf{x}$, $\mathbf{y}$,...). Vector or matrix variables are denoted in bold (for instance, $\mathbf{H}$ or $\mathbf{A}$).

Dimension

- n : dimension of the control vector $\mathbf{X}$.
- p : number of assimilated observations, e.g. the length of the observation vector $\mathbf{Y^0}$

Variables

- $\mathbf{X}$: the control random variable, called the *control vector*, with $\mathbf{x}$ a realization.
- $\mathbf{x}^t$: a non-random variable which represents the true value of the control parameters, called the *true control vector* (denoted by "t" for "true").
- $\mathbf{X}^b$: the background random variable which defines an a priori knowledge of the model state, called the *background vecto*r, with $\mathbf{x}^b$ a realization (denoted by "b" for "background").
- $\mathbf{X}^a$: the analysis random variable which defines the optimal parameters resulting from data assimilation, called the *analysis*, with $\mathbf{x}^a$ a realization (denoted by "a" for "analysis").
- $\mathbf{X}^g$: the reference random variable, in the vicinity of which the operator H and/or M is linearized.
- $\mathbf{Y}^0$: the observation random variable, which contains the experimental measures of the system and which is called the *observation vector*, with $\mathbf{y}^0$ a realization (denoted by "o" for "observation").
- $\mathbf{Y}^t$: the true trajectory projected onto the observation space.

Operators

- H : the observation operator, with $\mathbf{H}$ its tangent linear operator.
- M : the model operator which defines the integration of the equations over time, with $\mathbf{M}$ its linear approximation.
- S : the selection operator which defines how observations are extracted from the simulated fields with a prescribed spatial and time frequency, containing only 0 and 1 values.

Errors/Covariance matrices

- $\epsilon^b = \mathbf{X}^b - \mathbf{x}^t$: the random variable of background errors.
- $\epsilon^0 = \mathbf{Y}^0 - \mathbf{H}\mathbf{x}^t$: the random variable of observation errors.
- $\mathbf{B}$: the covariance matrix for background errors.
- $\mathbf{R}$: the covariance matrix for observation errors.
- $\mathbf{A}$: the covariance matrix for analysed errors.
- $\mathbf{I}$: the identity matrix.
- $\mathbf{K}$: the gain matrix.
- $\mathbf{d}^{ob} = \mathbf{Y}^0 - H(\mathbf{X}^b)$: the innovation vector, measuring the discrepancies between the observation vector and the background in the observation space .

- $\mathbf{d}^{oa} = \mathbf{Y}^0 - H(\mathbf{X}^a)$: the analysis residual in the observation space (also called AMO or OMA).
- $\mathbf{d}^{ab} = H(\mathbf{X}^a) - H(\mathbf{X}^b)$: the residual between the analysis and the background in the observation space.

Acronyms

- OSE : Observation System Experiment.
- AMB or BMA: Analysis Minus Background.
- BMO or OMB: Observation Minus Background.
- TMA : True Minus Analysis
- TMB : True Minus Background.
- BLUE: Best Linear Unbiased Estimator.
- RMS: Root Mean Square.
- PDF : Probability Density Function.

References

[1] François Bouttier and Philippe Courtier. Data assimilation concepts and methods, March 1999.

[2] P. Brasseur and J. Verron. *The SEEK fillter method for data assimilation in oceanography: a synthesis.* Springer Berlin / Heidelberg, 2006.

[3] K. Ide, P. Courtier, M. Ghil, and A.C Lorenc. Unified notation for data assimilation: operational, sequential and variational. *Journal of the Meteorological Society of Japan*, 75(1B):181–189, 1997.

[4] Eugenia Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability.* 2003.

[5] D.F Parrish and J.C Derber. The national meteorological center's spectral statistical interpolation analysis system. *Monthly Weather Review*, (120):1747–1763, 1992.

[6] F. Rabier, H. Jarvinen, E. Kilnder, J.F Mahfouf, and A. Simmons and. The ecmwf operational implementation of four-dimensional variational assimilation. part i: Experimental results with simplified physics. *Quarterly Journal of The Royal Meteorological Society*, (126):1143–1170, 2000.

[7] Sophie Ricci. *Assimilation variationnelle océanique : modélisation multivariée de la matrice de covariance d'erreur d'ébauche.* PhD thesis, Université Paul Sabatier - Toulouse III, March 2004.

[8] Sophie Ricci. *Calage de paramètre sur un toy model*, February 2010.

[9] Mélanie Rochoux. Preliminary investigation of data assimilation methodologies for forest fire propagation. Technical report, Cerfacs - University of Maryland, November 2010.

[10] A. Tarantola. Inverse problem theory and methods for model parameter estimation. *SIAM: Society for Industrial and Applied Mathematics*, 1987.

[11] Olivier Thual. Boucle externe sans oscillations. In *EPI-ALEZ, Échange de Projets et Idées "Assimilation de données pour le Lez"*, September 2010.