Theoretical Issues in the Application of Bayesian Data Assimilation in Context of Meteorology. Part 2: Stability and convergence

Jean-Pierre Issartel, X Busch, Maithili Sharan

Theoretical Issues in the Application of Bayesian Data Assimilation in Context of Meteorology. Part 2: Stability and convergence

By Jean-Pierre Issartel[1], X. Busch[2], Maithili Sharan[3]

[1]*Laboratoire de Mecanique et d'Energetique d'Evry, Universite d'Evry Val d'Essonne, 40 rue du Pelvoux, CE 1455 Courcouronnes, 91020 Evry Cedex,*
[2]*DGA Maitrise NRBC, 5, rue Lavoisier, BP 3, 91710 Vert le Petit cedex, France,*
[3]*Centre for Atmospheric Sciences, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110 016, India*

The elicitation of the background error covariance matrix (becm) is a major difficulty of Bayesian assimilation used in atmospheric and oceanic sciences based on Kalman filter in the form of 3Dvar or 4Dvar. In a preliminary companion paper, the definition of a becm and link to earlier information were seen poorly practicable. Here, more decisive arguments are obtained by reviewing mathematical studies, many in optimal control, a field dual to filtering.

When the system is discretized, and in the simplest steady and linear case, controllability and detectability conditions are generally fulfilled implying convergence in theory, not necessarily in practice as the limit may be very large. The non steady or nonlinear case may be stabilized by covariance inflation. However, mathematical works about inflation disprove the common belief that the converged result would be the becm of the system.

Considering the complete infinite dimensional system is even more disappointing. The becm is replaced by a trace class operator solution, in the steady linear case, to an infinite-dimensional Riccati equation. This equation has never been studied corresponding to the specific features of geophysical assimilation of data. These correspond to limit cases where the existence of solutions is no longer guaranteed. There is no ground, accordingly, for presuming filter convergence in terms of the true infinite-dimensional system. Counter-examples are proposed. This conclusion jeopardizes the physical meaning of any finite-dimensional becm that could be computed based on discretization.

Keywords: Assimilation of data, meteorology, Bayesian statistics

1. Introduction

The concept of background error covariance matrix (becm) is pivotal in the Bayesian framework of data assimilation used in atmospheric and oceanic sciences, hereafter termed jointly as geophysical sciences, to readjust numerical models to real observations. The matrix is theoretically related to earlier information by 3Dvar or 4Dvar iterations based on Kalman filter. However, the huge system of equations defies any complete numerical resolution and the simplified versions eventually diverge. Meteorologists commonly analyse this *catastrophic filter divergence* in view of the simplifications and approximations. A preliminary companion article discussed the definition of the becm showing the link to earlier information is poorly practicable. In addition to the well known numerical difficulties, the

becm is so sensitive to environmental conditions and sensor arrangement that any change makes it unusable.

Emerging opinion suggests catastrophic filter divergence is a rigorous property of the filter regardless of the simplifications (Kelly et al., 2015). This is the opinion supported here based on an extensive review of the mathematical literature. A large part is from optimal control as this field is equivalent by duality to filtering. The concepts of controllability, observability rooted in optimal control, are widely used in the theoretical works about Kalman filter. This is reminded and explained in appendices.

This article successively addresses finite and infinite dimensional systems, the first corresponding in practice to a discretized model of the second. The reason for this division is as follows. The becm is defined for finite-dimensional systems most often arising from discretization. Thus, the convergence of the discrete filter must be verified. However, and this is regularly neglected, it is also necessary to verify the significance of any discrete result with respect to a true property of the non dicretized system.

After a section 2 of statements and definitions, section 3 is devoted to finite-dimensional systems. The simplest ones are linear and steady: evolution and measurement operators are linear and repeated same at each time step. The steady linear systems are generally controllable and detectable: the filter converges and so does the becm. In practice, owing to the relatively small number of observations compared to a reasonable discretization, the theoretical becm is probably very large and numerically out of reach. Divergence occurs whenever a system is not controllable or not detectable. Convergence cannot be presumed. Covariance inflation is commonly used in geosciences to stabilize the filter, especially for time-varying or nonlinear systems. Surprisingly, this procedure is also well described in mathematical studies about Kalman filter. Its stabilizing efficiency was first understood by Xiong et al. (2009) and Dymirkovsky (2012). However, mathematical works disprove the belief, advocated in geosciences, that the converged result is the becm of the system.

Infinite-dimensional systems are considered in section 4. The becm must be replaced by a background error covariance operator (beco) of trace class, i.e. having a finite trace. The mathematics involved is much more complex than in finite-dimension, and the results are not so comprehensive. Owing to this limitation, the study of infinite-dimensional systems is restricted here for the steady linear case. This allows to reach significant conclusions by focussing at the Riccati equation fulfilled by the beco corresponding to the filter when time goes to infinity. This equation has been much studied in optimal control, unfortunately not corresponding to the specific features of geophysical assimilation of data. The trace class requirement is always ignored. The existence of solutions is proven generally under the condition that the evolution operator be power stable. It is shown here that this assumption is not acceptable in geophysics. The literature suggests that divergence is common among systems that are not power stable; counter-examples are obtained. There is no ground, accordingly, for presuming filter convergence in terms of the true infinite-dimensional system. This conclusion jeopardizes the physical meaning of any finite-dimensional becm that could be computed based on discretization.

2. Preliminaries and statement of the problem

(*a*) *Description of the system*

A system is described using a state function s belonging to a Hilbert space $\mathcal{S}$ of functions with sufficient regularity. The evolution and observation of the system at successive

time steps $t_1, t_2, ..., t_k$,... is described by the following noisy equations:

$$s_{k+1} = \mathcal{E}_k s_k + e_k \tag{2.1a}$$

$$\vec{\mu}_k = \mathcal{H}_k s_k + \vec{r}_k \tag{2.1b}$$

in which the label k refers to time t_k, s_k denotes the state at t_k, $\mathcal{E}_k$ is the evolution operator between t_k and t_{k+1}, e_k the dynamical error in the evolution model, $\vec{\mu}_k \in \mathbb{R}^{n_k}$ is the observation vector consisting of a finite number n_k of measurements, $\mathcal{H}_k$ is the observation operator, $\vec{r}_k$ the error in the observation model. In a 3Dvar approach, $s_k \in \mathcal{S}$ describes the instantaneous state of the system at t_k and $\vec{\mu}_k$ corresponds to observations at this time exactly. In a 4Dvar approach, $s_k \in \mathcal{S}$ describes the evolution of the system during the time interval $[t_{k-1}, t_k]$ and $\vec{\mu}_k$ corresponds to observations during this interval.

(*b*) *Filtering in finite dimension*

If the operators $\mathcal{E}_k$, $\mathcal{H}_k$ are linear and the state space is finite dimensional, $\mathcal{S} = \mathbb{R}^N$, the equations 2.1 may be rewritten in matrix form as:

$$s_{k+1} = \mathbf{E}_k s_k + \mathbf{M}_k \epsilon_k \tag{2.2a}$$

$$\vec{\mu}_k = \mathbf{H}_k s_k + \mathbf{N}_k \epsilon_k \tag{2.2b}$$

in which $\mathbf{E}_k$ (size $N \times N$), $\mathbf{H}_k$ (size $n_k \times N$) are the dynamical and measurement matrices, ϵ_k is a Gaussian noise with the identity of dimension $N + n_k$ as covariance matrix, $\mathbf{M}_k$ (size $N \times (N + n_k)$), $\mathbf{N}_k$ (size $n_k \times (N + n_k)$) are related to the covariance matrices $\mathbf{Q}_k$ (size $N \times N$), $\mathbf{R}_k$ (size $n_k \times n_k$) as:

$$\mathbf{M}_k = \left[\sqrt{\mathbf{Q}_k} \mid 0\right], \qquad \mathbf{N}_k = \left[0 \mid \sqrt{\mathbf{R}_k}\right] \tag{2.3}$$

The independent dynamical and measurement noises have covariance matrices described as:

$$\mathbf{Q}_k = \mathbf{M}_k \mathbf{M}_k^\intercal, \qquad \mathbf{R}_k = \mathbf{N}_k \mathbf{N}_k^\intercal, \quad \text{with} \quad \mathbf{M}_k \mathbf{N}_k^\intercal = 0 \tag{2.4}$$

Let s_k^b denote the background state at t_k, i.e. the state estimated before the observations $\vec{\mu}_k$ are taken into account. In terms of successive background states, filter equations are:

$$s_{k+1}^b = \mathbf{E}_k \left[s_k^b + \mathbf{P}_k \mathbf{H}_k^\intercal (\mathbf{H}_k \mathbf{P}_k \mathbf{H}_k^\intercal + \mathbf{R}_k)^{-1} \Delta\vec{\mu}_k\right] \tag{2.5a}$$

$$\mathbf{P}_{k+1} = \mathbf{E}_k (\mathbf{P}_k^{-1} + \mathbf{H}_k^\intercal \mathbf{R}_k^{-1} \mathbf{H}_k)^{-1} \mathbf{E}_k^\intercal + \mathbf{Q}_k \tag{2.5bi}$$

$$= \mathbf{E}_k \left(\mathbf{P}_k - \mathbf{P}_k \mathbf{H}_k^\intercal (\mathbf{H}_k \mathbf{P}_k \mathbf{H}_k^\intercal + \mathbf{R}_k)^{-1} \mathbf{H}_k \mathbf{P}_k\right) \mathbf{E}_k^\intercal + \mathbf{Q}_k \tag{2.5bii}$$

in which $\mathbf{P}_k$ is the becm describing the statistics of the departure of background from true state $s_k^b - s_k$.

When the operators $\mathcal{E}_k$, $\mathcal{H}_k$ are nonlinear, it is still possible to linearize them around s_k^b and put the system in the form 2.2. This strategy defines the *extended* Kalman filter.

3. Convergence and stability of discrete Kalman filter: a short review

Owing to their importance in many technological fields, Kalman filter and discrete time Riccati equation have drawn a considerable attention from mathematicians since the founding work by Kalman (1960). A number of important results, regarding convergence and stability, were already known by Jazwinski (1970), in particular the pivotal role of observability and controllability. Accurate definitions of these concepts are reminded in appendix C in a basic form corresponding to steady linear systems. An updated comprehensive and explanatory review of existing results may be found in the master's thesis by T. Karvonen (2014) about both linear and non-linear extended Kalman filter. This section reminds some results relevant for accepting the divergence and numerical instability of the scheme 2.5 as a theoretical insurmountable constraint in the context of complex geophysical systems.

Observability and controllability are properties of the sequences $(\mathbf{E}_k, \mathbf{H}_k)$, $(\mathbf{E}_k, \mathbf{M}_k)$ respectively. The weaker notions of detectability and stabilizability (appendix C) are often sufficient. In case of time varying systems, when $\mathbf{E}_k$, $\mathbf{M}_k$, $\mathbf{H}_k$, $\mathbf{N}_k$ are not constant, these properties must be uniform, as defined by Anderson & Moore (1979), to produce results in terms of convergence or stability of the filter with respect to initial conditions.

If the positive matrices $\mathbf{Q}_k$ are definite, controllability is obtained in one step as the system is steered from s_k to s_{k+1} by just inverting $\mathbf{Q}_k$. A non invertible $\mathbf{Q}_k$ would imply the existence of some degree of freedom free from dynamical model errors. Therefore, the geophysical systems investigated here are most probably controllable and this controllability will be often uniform corresponding to roughly steady dynamical errors. There is a problem, on the contrary with observability, because the number N of degrees of freedom in the states is huge compared to the number of independent measurements at each time step, i.e. the rank of $\mathbf{H}_k$.

(*a*) *Linear time-invariant systems and filter convergence*

The ergodic assumption that background errors have the same statistics at each time step is essential to link the becm $\mathbf{P}$ with the successive observations. It is optimally fulfilled in steady systems. They are described as:

$$s_{k+1} = \mathbf{E}s_k + \mathbf{M}\epsilon_k \tag{3.1a}$$

$$\vec{\mu}_k = \mathbf{H}s_k + \mathbf{N}\epsilon_k \tag{3.1b}$$

in which the dynamical and measurement noises are independent with covariance matrices:

$$\mathbf{Q} = \mathbf{M}\mathbf{M}^\intercal, \qquad \mathbf{R} = \mathbf{N}\mathbf{N}^\intercal, \quad \text{with} \quad \mathbf{M}\mathbf{N}^\intercal = 0 \tag{3.2}$$

The system is associated with the steady Riccati equation of unknown matrix $\mathbf{P}$ (compare equation 2.5bii):

$$\mathbf{P} = \mathbf{E}\left(\mathbf{P} - \mathbf{P}\mathbf{H}^\intercal(\mathbf{H}\mathbf{P}\mathbf{H}^\intercal + \mathbf{R})^{-1}\mathbf{H}\mathbf{P}\right)\mathbf{E}^\intercal + \mathbf{Q} \tag{3.3}$$

The following result was first proven by Kalman & Bucy (1961) for continuous time filter. The discrete time version may be found in (Anderson & Moore, 1979, section 4.4; Lancaster & Rodman, 1995, theorem 17.5.3; Lewis et al., 2008, theorem 2.3).

Theorem 3.1. *Suppose the steady system* 3.1 *is controllable. Then, it is detectable if and only if the steady Riccati equation* 3.3 *has a unique positive solution* $\mathbf{P}_\infty$*. In addition, this solution is positive definite and is the limiting value* $\lim_{k\to\infty} \mathbf{P}_k = \mathbf{P}_\infty$ *of Kalman filter becm for every choice of non-negative symmetric* $\mathbf{P}_0$*.*

By Hautus-Popov-Belevitch test (appendix A), controllability is a generic property of the eigenvectors of $\mathbf{E}^\intercal$, $\mathbf{M}^\intercal$, detectability is a generic property of the eigenvectors of $\mathbf{E}$, $\mathbf{H}$. Thus, with unlikely exceptions, theorem 3.1 guarantees that $\mathbf{P}_\infty$ exists. It is interesting to notice that, under the same conditions, the computation of $\mathbf{P}_\infty$ by Kalman filter iterations is numerically stable; other stable algorithms exist with similar or lesser computational cost (Assimakis et al., 1997).

Uncontrollable systems exist however and among them, some are unstabilizable corresponding to the strongest form of uncontrollability. In this case, Kučera (1972) shows that equation 3.3 may not have a unique positive definite solution: either zero or at least two. Take care that Kučera works from the point of view of optimal control dual to filtering (appendix C, section b), unstabilizability and undetectability are inverted. Fletcher (2017), in his tutorial book about data assimilation, devotes two chapters to optimal control and clearly relates equation 3.3 to that field. It is surprising that he does not invite to verify the stabilizability of a system and does not propose unstabilizability as a possible cause of filter divergence.

(*b*) *Linear time-varying systems and filter stability*

Let's consider linear time-varying systems of the form 2.2. In view of the linearity, the matrices $\mathbf{E}_k$, $\mathbf{M}_k$, $\mathbf{H}_k$, $\mathbf{N}_k$ are independent of the state. However, they may vary with the time so that the filtering distribution with $\mathbf{P}_k$ is not expected to converge. It is still reasonable to wish Kalman filter will forget the initial inputs. A sufficient condition is given by the following result (Jazwinski, 1970, theorem 7.5; Kamen & Su, 1999, theorem C.4):

Theorem 3.2. *If the system* 2.2 *is uniformly stabilizable and uniformly detectable, the corresponding Kalman filter is stable with respect to the initial conditions.*

Stability with respect to the initial conditions does not mean convergence. It just means that the influence of initial input state and becm is progressively forgotten so that after many time steps, output state and output becm depend only on the observations.

(*c*) *Nonlinear systems*

With rare exceptions, the situations addressed in geosciences involve nonlinearities. If the system 2.2 is obtained from a non-linear system 2.1, the matrices $\mathbf{E}_k$, $\mathbf{M}_k$, $\mathbf{H}_k$, $\mathbf{N}_k$ are updated based on the estimated state. This dependence on the states successively estimated does not even allow the stability of $\mathbf{P}_k$ with respect to initial conditions, at least from a strict mathematical point of view.

From a physical point of view, one may still wish that at each moment t_k, the background state s_k^b departs little from the true state s_k^{tr}, so that $\mathbf{E}_k$, $\mathbf{M}_k$, $\mathbf{H}_k$, $\mathbf{N}_k$ depart little from their values at s_k^{tr}. The conditions of theorem 3.2 from section c would be approximately fulfilled. Hopefully, stability with respect to initial conditions would approximately apply.

This hope may be addressed by two theorems of Reif et al. (1999, theorems 3.1 and 4.1) slightly improved by Rhudy et al. (2012). These are still the core of the mathematical knowledge about the stability of the extended Kalman filter. Theorem 3.1 of Reif et al. states that the estimation error $s_k^b - s_k^{tr}$ remains bounded provided very conservative bounds apply to the non-linearities and to matrices $\mathbf{E}_k$, $\mathbf{M}_k$, $\mathbf{H}_k$, $\mathbf{N}_k$ including $\mathbf{P}_k$. The

boundedness of $\mathbf{P}_k$ is subjected, according to their theorem 4.1, to non-linear observability and controllability conditions. Theorem 4.1 was originally proven for nul dynamical noise but this restriction can be removed provided the initial error $s_0^b - s_0^{tr}$ be sufficiently small (Karvonen, 2014).

The conservative bounds of these theorems can hardly be determined. Fortunately, numerical experiments show they can be considerably relaxed in practice while maintaining $s_k^b - s_k^{tr}$ bounded (Dymirkovsky, 2012). These experiments also show the stabilizing efficiency of artificially increasing the covariance matrices $\mathbf{Q}_k$, $\mathbf{R}_k$ here in the form of $\mathbf{M}_k = \left[\sqrt{\mathbf{Q}_k} \mid 0\right], \mathbf{N}_k = \left[0 \mid \sqrt{\mathbf{R}_k}\right]$ (equation 2.3). The technique is similar to the inflation used in geosciences. It was investigated early by the mathematicians and some theoretical results are available.

(*d*) *Covariance tuning and filter stabilization*

The liberty taken by the mathematicians of replacing the true covariance matrices originates in two points as (i) Kalman filter is used sometimes as an observer for noise-free deterministic systems (Reif et al., 1996) and (ii) the linearised extended Kalman filter, anyway, is not an optimal filter, the analysed outputs s_k^a, $\mathbf{P}'_k$ are at most approximations of the true conditional mean and posterior becm. The stabilizing effect of appropriately chosen instrumental matrices $\mathbf{Q}_k^+$, $\mathbf{R}_k^+$ was observed numerically and utilized before being understood. Corresponding to this usage, the aforementioned stability theorems by Reif et al. (section c) distinguish the true and modified covariance matrices.

The stabilizing effect of enlarged covariance matrices was first understood by Xiong et al. (2009) and Dymirkovsky (2012). Both used a technique introduced by Boutayeb and Aubry (1999) to represent the analysis error $s_k^a - s_k^{tr}$ as a continuous function of filter inputs. An appropriate enlargement or inflation $\mathbf{Q}_k^+ = \mathbf{Q}_k + \Delta\mathbf{Q}_k$, immediately transferred to $\mathbf{P}_{k+1}$ (equation 2.5b), allows some matrix $\Xi_{k+1} = \mathbf{P}_{k+1} - \beta_k$, where β_k is some symmetric positive matrix, to remain positive thus increasing the domain of initial conditions for which the estimation error remains bounded. The extra positive matrix $\Delta\mathbf{Q}_k$ is conveniently taken as diagonal. These results are proven for the unscented Kalman filter.

The unscented Kalman filter is an ensemble based non-linear filter introduced by Julier et al. (1995), widely used in navigation control. The filter relies on the *unscented transform* of background error statistics that is a representation by a set of states called *sigma points*. Non-linear observability and controllability conditions are naturally required.

Many researchers using Bayesian approach for geophysical problems understand the stabilizing efficiency of covariance inflation as revealing a tendency of the undersampled ensemble Kalman filter to underestimate the diagonal terms of the becm. Inflation would offset this underestimation thus avoiding filter divergence. This view leads to argue about a *true* value of the inflation parameters that might be estimated from the observations (Li et al., 2009). Anderson (2007) followed by Miyoshi (2011) consider inflation as an additional state parameter to be filtered with some probability distribution, and this point of view is illustrated by numerical experiments in which the filter does compensate the intentional underestimation of a true covariance matrix.

Such interpretation is not supported by the mathematical studies stated above. These clearly distinguish the true covariance matrices $\mathbf{Q}_k$, $\mathbf{R}_k$ from their enlarged counterparts $\mathbf{Q}_k^+$, $\mathbf{R}_k^+$. However, these studies do not distinguish the true and computed becm. This notational negligence arises from the fact that, as the linearised filter is suboptimal, even without inflation, the output $\mathbf{P}_k$ is not the true becm. In fact, the addition of $\Delta\mathbf{Q}_k$ stabilizes

the filter at the price of a degraded accuracy (Xiong, 2006). Tuning inflation is a matter of balance between stability and accuracy. The causes of covariance underestimation described by the meteorologists are well documented. It is just said here that inflation may not be seen as a correction for that.

(*e*) *Numerical instability of Kalman filter*

A reasonable discretization of the atmospheric or oceanic environment should involve a number N of degrees of freedom very large compared to the number of measurements n_k available at any time step t_k. The number of time steps required to possibly reach observability is too large compared to computational requirements. In practice, the system is unobservable.

To illustrate this, suppose the $N \times N$ matrices $\mathbf{E}_k$, $\mathbf{Q}_k$ and $\mathbf{H}_k^{\mathsf{T}}\mathbf{R}_k^{-1}\mathbf{H}_k$ involved in the second equation 2.5 are diagonal. If in addition $\mathbf{P}_0$ is diagonal, so is $\mathbf{P}_k$ at all times. These matrices may be described in terms of their diagonal coefficients, respectively $a_{(k)i}, q_{(k)i}, g_{(k)i}, p_{(k)i}$, $i = 1, 2, ..., N$. Equation 2.5 for $\mathbf{P}_{k+1}$ becomes:

$$\begin{aligned} p_{(k+1)i} &= a_{(k)i}^2 p_{(k)i} + q_{(k)i} \quad \text{if} \quad g_{(k)i} = 0 \\ p_{(k+1)i} &= a_{(k)i}^2 \frac{p_{(k)i}}{1 + p_{(k)i} g_{(k)i}} + q_{(k)i} \quad \text{if} \quad g_{(k)i} \neq 0 \end{aligned} \tag{3.4}$$

For each $k \geq 0$ there are at least $N - n_k$ coefficients $g_{(k)i} = 0$ corresponding to the rank n_k of $\mathbf{H}_k^{\mathsf{T}}\mathbf{R}_k^{-1}\mathbf{H}_k$. Since N is large compared to $n = \max_{k=0}^{\infty} n_k$, for most labels i, non-zero coefficients $g_{(k)i}$ are rare, the first one $g_{(p)i} > 0$ occurring for a time ν typically of magnitude N/n. The first equation 3.4 then brings:

$$p_{(\nu+1)i} = p_{(0)i} \prod_{l=0}^{\nu} a_{(l)i}^2 + \sum_{k=0}^{\nu} q_{(k)i} \prod_{l=k+1}^{\nu} a_{(l)i}^2 \quad \text{if} \quad q_{(k)i} = 0, \;\; 0 \leq k \leq \nu \tag{3.5}$$

Since the matrices $\mathbf{E}_k$ are not in general contractive, among the many labels i subject to equation 3.5, it is reasonable to find some such that $a_{(k)i} > 1$ for $0 \leq k \leq \nu$. The related coefficients $p_{(k)i}$ of the background error covariance matrix will have an increase more than exponential until time $p \approx N/n$.

Of course, the matrices involved in the real problems are not simultaneously diagonalizable. The example nevertheless confirms that one should not be surprised of numerical instabilities when computing the becm from a filter with insufficient ratio N/n. This may be expected regardless of filter simplifications.

4. Divergence of the infinite-dimensional Kalman filter

The mathematical review in previous section 3 is related to finite-dimensional discretized systems. It highlights the importance, for filter convergence, of observability and controllability conditions. They are fulfilled in general, but not always, for the steady linear systems and convergence is confirmed by theorem 3.1. When the conditions are not fulfilled, divergence occurs. It is accordingly not possible to presume convergence as many geoscientits tacitly do when considering divergence as a numerical artifact (see Part 1 for details). In the time-varying or nonlinear cases, the stability is confirmed theoretically too, but subjected to so strong restrictions that instability is expected in practice. Covariance inflation

is validated as a filter stabilizing procedure, but its ability for identifying the true becm is clearly denied by detailed studies.

These results do not mean, after all, that the becm cannot be determined from earlier information using Kalman filter in the form of 3Dvar or 4Dvar: they just point out difficulties. Paragraph 3e suggests the enormous gap between the discretized dimension and relatively small number of measurements might lead to large background errors, numerically out of reach. The question arises of what happens when the whole complexity of the system is taken into account without discretization. This is the task of the present section, focussed at the steady linear case. It is adressed by reviewing the literature, essentially from optimal control, about the infinite dimensional Riccati equation. To begin with, it is necessary to see how the filter adapts for the infinite dimension.

(*a*) *Notations*

The Hilbert space $\mathcal{S}$ of system states is now supposed infinite-dimensional. This section examines how the previous developments should be modified. Let's denote

- $(.,.)_{\mathcal{S}}$ and $\| \ . \ \|_{\mathcal{S}}$ the scalar product and related norm in Hilbert space $\mathcal{S}$;
- $\mathcal{S}^*$ the dual space consisting of the continuous linear forms from $\mathcal{S}$ into $\mathbb{R}$;
- $s^* \in \mathcal{S}^*$ the state adjoint to a state s such that $s^*(\sigma) = (s, \sigma)_{\mathcal{S}}$ for all $\sigma \in \mathcal{S}$;
- $\mathcal{X}^*$ the adjoint to an operator $\mathcal{X}$;
- $\mathscr{L}(\mathcal{S})$ the space of bounded (i.e. continuous) linear operators on $\mathcal{S}$.

The linear operators between Hilbert spaces that are continuous are exactly those bounded on the unit sphere and thus, they are rather called the bounded operators.

(*b*) *Operator Riccati equation*

When the infinite dimension of a complex system is taken into account, the matrix description 2.2 is replaced by an operator description. The operators considered hereafter, are linear and, unless otherwise specified, they are also bounded:

$$s_{k+1} = \mathcal{E}_k s_k + \mathcal{M}_k \epsilon_k \tag{4.1a}$$

$$\vec{\mu}_k = \mathcal{H}_k s_k + \mathcal{N}_k \epsilon_k \tag{4.1b}$$

In this system, randomness is accounted for by ϵ_k now drawn from an infinite-dimensional Hilbert space. This variable may be taken as Gaussian but its covariance operator $\mathcal{V}$ must be trace class (appendix B) excluding identity utilized in finite-dimension. The expressions for the covariance operator $\mathcal{Q}_k$ of dynamical noise and the covariance operator $\mathbf{R}_k$ of measurement noise, this one still possible to describe as a $n_k \times n_k$ matrix, are modified as:

$$\mathcal{Q}_k = \mathcal{B}_k \mathcal{V} \mathcal{B}_k^*, \qquad \mathbf{R}_k = \mathcal{D}_k \mathcal{V} \mathcal{D}_k^*, \quad \text{with} \quad \mathcal{B}_k \mathcal{V} \mathcal{D}_k^* = 0 \tag{4.2}$$

Kalman filter 2.5 should be rewritten in operator form with successive becm replaced by background error covariance operators (beco) $\mathcal{P}_k$ that must be trace class (appendix B). If

the system 4.1 is steady, a converged beco $\mathcal{P}$ should be a self-adjoint, non-negative, trace class solution of the equation:

$$\mathcal{P} = \mathcal{E}\left(\mathcal{P} - \mathcal{P}\mathcal{H}^*(\mathcal{H}\mathcal{P}\mathcal{H}^* + \mathbf{R})^{-1}\mathcal{H}\mathcal{P}\right)\mathcal{E}^* + \mathcal{Q} \tag{4.3}$$

The point is now to determine whether such solution exists. The following paragraphs show this cannot be presumed. As far as geophysical systems are concerned, the equation 4.3 does not fulfill the conditions usually retained and examples are easily obtained where no solution exists.

(*c*) *Example of systems leading to filter divergence*

Two examples are given hereafter of systems for which equation 4.3 does not admit any solution $\mathcal{P}$. These are infinite dimensional steady linear systems corresponding to classical geophysical assimilation problems. The evolution operators are either bounded or unbounded.

(i) *Bounded evolution operator: source of a trace species*

The authors have utilized data assimilation most often to retrieve the source $s(x,t)$ of an atmospheric tracer based on concentration measurements. Such source is commonly determined by complex processes and human behaviours. No clear logics can be identified and the evolution of the source may be modelled as purely random. This corresponds to a null evolution model in continuous time: $\frac{\partial s}{\partial t} = 0s + \varepsilon$, to an identical evolution model in discrete time: $s_{k+1} = \mathcal{I}s_k + \varepsilon_k$. This evolution operator $\mathcal{I}$ is bounded, i.e. continuous. It has 1 as an eigenvalue of infinite multiplicity. The following result indicates that, for this system, Kalman filter may not converge.

Theorem 4.1. *If (i) the measurement operator $\mathcal{H}$ has finite rank n, (ii) the evolution operator $\mathcal{E}$ has an eigenvalue λ, $|\lambda| \geq 1$, of multiplicity* $\dim\ker(\mathcal{P} - \lambda\mathcal{I}) > n$ *and (iii) $\mathcal{Q} > 0$, there does not exist any symmetric positive operator $\mathcal{P}$ solution of equation 4.3.*

Proof : Suppose $\mathcal{P}$ is a positive solution of equation 4.3. It is easily seen that the operator $\mathcal{Q} - |\lambda|^2\mathcal{P}\mathcal{C}^*(\mathcal{C}\mathcal{P}\mathcal{C}^* + \mathbf{R})^{-1}\mathcal{C}\mathcal{P}$ must be non-positive through $\ker(\mathcal{P} - \lambda\mathcal{I})$. But this is impossible. The rank of $|\lambda|^2\mathcal{P}\mathcal{C}^*(\mathcal{C}\mathcal{P}\mathcal{C}^* + \mathbf{R})^{-1}\mathcal{C}\mathcal{P}$, less than n, does not allow to compensate for the positiveness of $\mathcal{Q}$ over all of $\ker(\mathcal{P} - \lambda\mathcal{I})$ having too large a dimension.

□

(ii) *Unbounded evolution operator: shallow water equations*

This section provides a relatively simple example of an evolution operator making the Riccati equation 4.3 impossible to solve. It corresponds to the very common problem of shallow water equations. The evolution operator is unbounded, which is a situation poorly studied in optimal control.

Let's consider the propagation in the x-direction of an incompressible flow perturbation in a basin with constant depth H. The perturbation is characterized by a horizontal velocity $u(x,t)$ and vertical displacement $h(x,t)$. This is modelled using one-dimensional shallow water equations:

$$\begin{bmatrix} u \\ h \end{bmatrix} = \Gamma \begin{bmatrix} u \\ h \end{bmatrix} \quad \text{with} \quad \Gamma = \begin{bmatrix} -\gamma & -g\frac{\partial}{\partial x} \\ -H\frac{\partial}{\partial x} & 0 \end{bmatrix} \tag{4.4}$$

$g = 9.81 m.s^{-1}$ is acceleration due to gravity, γ is a constant viscosity. The evolution between two assimilation steps at an interval τ is described as:

$$\begin{bmatrix} u_{k+1} \\ h_{k+1} \end{bmatrix} = \mathcal{E} \begin{bmatrix} u_k \\ h_k \end{bmatrix} \quad \text{with} \quad \mathcal{E} = e^{\tau\Gamma} \tag{4.5}$$

where the evolution operator is the time propagator associated with Γ, formally denoted as $e^{\tau\Gamma}$. The measurement operator $\mathcal{C}$ might consist in observing at some detector locations one or both state variables, u or h. This is consistent with the claim of Bannister (2008): *Background errors at different locations can be correlated and appear in the* $\mathbf{B}$*-matrix* [the becm] *as off-diagonal elements. These cause information to be transferred between these locations during assimilation.* He also writes later in the text: *The* $\mathbf{B}$*-matrix spreads information to other variables and imposes balance*.

For any $\lambda \in \mathbb{C}$, the differential equation $\Gamma \begin{bmatrix} u \\ h \end{bmatrix} = \lambda \begin{bmatrix} u \\ h \end{bmatrix}$ or $\mathcal{E} \begin{bmatrix} u \\ h \end{bmatrix} = e^{\lambda} \begin{bmatrix} u \\ h \end{bmatrix}$ is ordinary with respect to x, linear, with constant coefficients. Thus, non-trivial solutions exist. In other words, the eigenvalues of $\mathcal{E}$ correspond to the whole complex field but zero. The following result indicates that, for this system, Kalman filter may not converge.

Theorem 4.2. *If (i) the measurement operator $\mathcal{H}$ has finite rank, (ii) the eigenvalues of the evolution operator $\mathcal{E}$ are unbounded and (iii) $\mathcal{Q} > 0$, there does not exist any symmetric positive operator $\mathcal{P}$ solution of equation 4.3.*

Proof : A positive solution of equation 4.3 is necessarily definite because $\mathcal{Q} > 0$ and $\mathcal{P} > \mathcal{Q}$. The positive operator $\mathcal{P}\mathcal{H}^*(\mathcal{H}\mathcal{P}\mathcal{H}^* + \mathbf{R})^{-1}\mathcal{H}\mathcal{P} \geq 0$ is less than $\mathcal{P}$ and its rank is finite, bounded by that of $\mathcal{H}$, Thus, a constant α, $0 < \alpha < 1$, may be found such that:

$$\mathcal{P} \geq (1-\alpha)\mathcal{E}\mathcal{P}\mathcal{E}^* + \mathcal{Q} \tag{4.6}$$

Now suppose $\zeta \in \mathcal{S}$ is eigenvector of $\mathcal{E}$ with eigenvalue λ, $\mathcal{E}\zeta = \lambda\zeta$, $\| \zeta \|_{\mathcal{S}} = 1$. From equation 4.6 one deduces:

$$\zeta^*\mathcal{P}\zeta \geq |\lambda|^2(1-\alpha)\zeta^*\mathcal{P}\zeta + \zeta^*\mathcal{Q}\zeta \tag{4.7}$$

Since $\zeta^*\mathcal{Q}\zeta > 0$, this implies $|\lambda|^2 < \frac{1}{1-\alpha}$ in contradiction with the assumption that the eigenvalues of $\mathcal{E}$ are unbounded.

□

(*d*) *Literature about the operator Riccati equation*

The resolution of Riccati equation is an active topic of control theory. The focus is on control rather than filtering, and equation 4.3, instead of being regarded as a filter Riccati equation, is regarded as the control Riccati equation of the adjoint system (appendix C, sections b, c). This does not change the relevance of the results but obliges to some translation: the dual properties of observability and controllability should be interchanged and as well their degraded versions of detectability and stabilizability.

Unfortunately, the interest of existing works for the present purpose is limited. On the one hand, most studies are related to finite-dimensional systems such as the discretized systems discussed in section 3. On the other hand, most often continuous time systems are addressed of the form $\frac{ds}{dt} = \mathbf{E}s(t) + \mathbf{M}\epsilon(t), \vec{\mu}(t) = \mathbf{H}s(t) + \mathbf{N}\epsilon(t)$. A different continuous

time Riccati equation applies with different observability and controllability conditions.

The literature about the discrete time Riccati equation 4.3 for infinite-dimensional operators is accordingly limited, produced essentially between 1970 and 2000. Even then, its exploitation is difficult for geophysical assimilation.

First, the evolution operator $\mathcal{E}$ in control studies are always bounded. However, unboundedness is common in geophysical problems: see previous section 4ii.

Second, the trace class requirement for prescribed $\mathcal{Q}$ and sought $\mathcal{P}$ has not drawn the attention of control theoreticians. Regarding $\mathcal{Q}$, a common assumption is that it is coercive: $\alpha\mathcal{Q}$ is greater than identity for some constant $\alpha > 0$. Such $\mathcal{Q}$ cannot be trace class. Coercivity is sometimes assumed by just requiring that $\mathcal{Q}$ be boundedly invertible. The very interesting theorem 4.2 of Zabczyk (1974) about the uniqueness of a possible solution $\mathcal{P}$ to equation 4.3 and geometric convergence of the successive $\mathcal{P}_k$ is deduced assuming $\mathcal{Q}$ is positive definite, denoted $\mathcal{Q} > 0$. However, a careful reading shows that, for this and many authors, positive-definiteness means coercivity and thus, the theorem does not apply in the Bayesian context. Regarding $\mathcal{P}$, trace estimations have been investigated only in finite dimension (e.g. Komaroff & Shahian, 1992; Hua Dai, 2011); these results are not directly applicable to decide whether a solution of equation 4.3 is trace class.

Third, there are many infinite-dimensional generalizations of the dual concepts of controllability/stabilizability, observability/detectability (appendix A, section b). Most studies if not all are rely on power stability/stabilizability. Such stability is excessive for geophysical problems (see remark at the end of appendix A). The theorem 8.3 of Ostveen & Zwart (1996) states the existence and uniqueness of a physically relevant solution (stabilizing solution) to equation 4.3 when $\mathcal{C}$ has finite rank (the number of measurements) and under a spectral condition generically fulfilled. Unfortunately, it is based on power stability.

The lemma 3.6 of Opmeer & Curtain (2007) indicates that a filter Riccati equation has a nonnegative selfadjoint solution, not necessarily unique, if and only if the adjoint system fulfills a *finite cost condition*. However, this result is obtained for an equation slightly different from 4.3 (see remark at the end of appendix C, section a). The difference amounts to modifying the cost function in the adjoint system so that it becomes coercive. As a consequence, the lemma does not apply in the present context neither.

To conlude this disappointing review, as of now, there is no theoretical support for presuming the existence and uniqueness of a selfadjoint, nonnegative, trace class solution to equation 4.3. The point has just remained out of scope.

(*e*) *Literature relating Kalman filter and Riccati equation*

In the past decades some works have been devoted to the infinite-dimensional Kalman filter from the point of view of Riccati equation. These works might suggest that the stability of the filter is well established.

Infinite-dimensional Kalman filter has been studied very early. In his founding work, Falb (1967) generalizes the filter for infinite-dimensional spaces. His work deals with continuous time filtering, but this is not what restricts the support discrete time assimilation can obtain from it. Falb shows with his theorem 7.10 that the Riccati equation associated with the filter has indeed a solution. However, this solution is a function of the time. In terms of discrete-time data assimilation, this amounts to saying that an operator $\mathcal{P}_k$ is well defined at each time step following the operator version of equation 2.5b. It is not even suggested that $\mathcal{P}_k$ might converge to a limit, or that the sequence is bounded.

In a recent publication, Aalto (2018) shows that the infinite dimensional Kalman filter is well approximated by a discretized filter. The state estimate and covariance matrix of estimation errors obtained based on discretization converge, when discretization is refined, to the values that would be obtained from the infinite dimensional filter. However, Aalto requires in his theorem 4.1 that the system be power stable which was seen (section 4d) to be an unacceptable assumption for most geophysical systems.

5. Conclusions

The elicitation of the Bayesian priors in the form of a becm is essential in current data assimilation. Meteorologists and oceanologists usually argue the becm is derived from successive observations by filter iterations. This argument is purely theoretical to support the consistency of the framework. In practice, the numerical models utilized in geosciences contain so many discrete elements that filter equations cannot be handled. The becm is either chosen from empirical assumptions or the equations are strongly simplified eventually leading to *catastrophic filter divergence*.

The companion to this article (part 1) started to question the common claim that *catastrophic filter divergence* occurs due to simplifications. This claim amounts to tacitly presume that the non-simplified filter should converge and to reaffirm the consistency of the theoretical framework. The definition of a becm and the link with earlier information were seen poorly practicable. They unrealistically require no variation in the number and arrangement of the detectors or weather conditions without even guaranteeing any regularity of the estimate. The classical assumption of evenly distributed background errors is not appropriate. The present article (part 2) discusses the question: is it justified to presume that the non-simplified filter should converge?

An extensive literature is available about Kalman filter and nonlinear extended Kalman filter for the finite-dimensional systems. The steady linear systems of finite dimension are generally observable and controllable implying filter convergence, but the result may be numerically out of reach. Unobservable or uncontrollable systems exist: they lead to theoretical divergence. Even in the simplest situations, convergence may not be simply presumed. In the time-varying or nonlinear cases, convergence is subjected to so strong restrictions that instability is expected. Covariance inflation is validated by mathematical studies as a filter stabilizing procedure, but its ability for identifying the true becm is clearly denied.

The infinite-dimensional case was investigated based on the Riccati equation fulfilled by the background error covariance operator of a steady linear system. This equation has been studied exclusively in the field of optimal control that is equivalent by duality to filtering. However, the conditions considered in optimal control never meet exactly the detailed requirements of geophysical assimilation. The trace class requirement is ignored. The existence of solutions to the Ricati equation is established essentially for power stabilizable systems. The evolution operators considered in geosciences are not power stabilizable; this would mean that all initial states are finally dissipatively focussed to finite dimension. So, the geophysical systems correspond to the limit case where convergence is not guaranteed. Diverging counter-examples are given. There is no ground, accordingly, for presuming filter convergence. If a becm has been obtained based on some discretization, there is also no ground for regarding it as an approximation of a true property.

At this stage, it is unavoidable to question the very definition of the becm. Let's first stress the fact that the question is not so big as it looks like. In practice, a correct implementation of the Bayesian framework is not possible and the methods used in reality are always

described in terms of simplifications or approximations. It is not possible to determine how far these simplifications/approximations depart from the theory. After all, meteorological and oceanic centres have been producing reliable forecasts for the past decades without ever computing any converged becm.

A given observation may have many causes. The idea seems therefore reasonable that, based on observations, one might describe the quality of a forecast as a stochastic error/correction distribution. Such distribution may not be obtained from Bayesian statistics. Accordingly, there is a need for another theoretical framework. Let's consider two possibilities for determining the probabilities of head and tails when throwing a coin. Using earlier information would mean throwing the coin hundreds of times before predicting that in next throw, head and tail have the approximately same probability of 50%. In practice, we reach the same result much faster by just considering the symmetry of the coin. The first strategy may be compared to the tentative derivation of a becm using Kalman filter. The relevance of the symmetry-based strategy for geophysical systems is not obvious. It will be explained in a forthcoming article.

References

AAlto, A., Spatial discretization error in Kalman filtering for discrete-time infinite dimensional systems, IMA Journal of Mathematical Control and Information, Volume 35, Issue Supplement 1, April 2018.

Anderson, B. D. O., Moore, J. B., Optimal Filtering. Information and System Sciences Series, Prentice-Hall, 1979.

Anderson, J. L., An adaptive covariance inflation error corection algorithm for ensemble filters, Tellus 59A, pp. 210-224, 2007.

Assimakis, N. D., Lainiotis, D. Q., Katsikas, S. K., Sanida, F. L., A survey of recursive algorithms for the solution of the discrete time Riccati equation, Nonlinear Analysis, Theory, Methods & Applications, Vol. 30, no. 4, pp. 2409-2420. 1997

Bannister, R. N., A review of forecast error covariance statistics in atmospheric variational data assimilation. I: Characteristics and measurements of forecast error covariances, Q. J. R. Meteorol. Soc. 134, PP. 1971-1996, 2008.

Boutayeb, M., Aubry, D., A strong tracking extended Kalman observer for nonlinear discrete-time systems, IEEE Transactions on Automatic Control, 44(8):1550-1556, 1999.

Dymirkovsky, G., New results on stochastic stability of discrete-time unscented Kalman filter, proceedings of the 7th IEEE Conference on Industrial Electronics and Applications held in Singapore, 18-20 July 2012, pp. 1543-1548

Engwerda, J. C., Stabilizability and detectability of discrete-time time-varying systems, Memorandum COSOR; Vol. 8816, Eindhoven, Technische Universiteit Eindhoven, 1988.

Fletcher, S. J., Data Assimilation for the Geosciences: From Theory to Application, Elsevier, 15 march 2017.

Gohberg, I.C., Goldberg, S., Kaashoek, M.A., Classes of linear operators, vol. 1, Birkhäuser Verlag, Basel, 1990.

Hautus, M. L. J., Controllability and observability conditions of linear autonomous systems. Ned. Akad. Wetenschappen, Proc. Ser. A 72, pp 443-448, 1969.

Hua Dai, On Eigenvalue Bounds and Iteration Methods for Algebraic Riccati Equations, Journal of Computational Mathematics, May 2011.

Huang, Z.Y., Yan, J.A., Introduction to Infinite Dimensional Stochastic Analysis, Mathematics and Its Applications (Dordrecht). 502. Dordrecht: Kluwer Academic Publishers. Beijing: Science Press (2000).

Jazwinski, A. H., Stochastic Processes and Filtering Theory, Academic Press, 1970.

Kalman, R. E., A new approach to linear filtering and prediction problems, Transactions of the ASME, Journal of Basic Engineering, 82 (1), pp. 35-45, 1960.

Kalman, R. E., Bucy, R. S., New results in linear filtering and prediction theory, J. Basic Eng 83(1), pp 95-108, March 1961.

Kamen, E. W., Su, J. K., Introduction to Optimal Estimation. Advanced Textbooks in Control and Signal Processing, Springer 1999.

Karvonen, T., Stability of linear and non-linear Kalman filters, Master's thesis, Simo Särkkä advisor, University of Helsinki, December 4, 2014.

Kato, T. Perturbation theory for linear operators, Springer-Verlag, 1980.

Kelly, D., Majda, A. J., Tong, X. T., A concrete ensemble Kalman filter with rigorous catastrophic filter divergence, Proc. Natl. Acad. Sci. USA. 112(34), pp 10589-10594, Aug 25, 2015.

Komaroff, N., Shahian, B., Lower Summation Bounds for the Discrete Riccati and Lyapunov Equations, IEEE Transactions on Automatic Control, vol. 31, 1, July 1992.

Kučera, V., The discrete Riccati equation of optimal control, Kybernetika, vol. 8, issue 5, 1972.

Lancaster, P., Rodman, L., Algebraic Riccati equation, Clarendon Press, 7 sept. 1995.

Lewis, F. L., Lihua Xie, Popa D., Optimal and robust estimation: with an introduction to stochastic control theory (Second Edition), CRC Press Taylor & Francis Group, 2008.

Malinen, J., Discrete Time H^∞ Algebraic Riccati Equations, XXXX, 2000.

Miyoshi, T., The Gaussian approach to adaptive covariance inflation and its implementation with the local ensemble transform Kalman filter, Monthly Weather Review, pp. 1519-1535, may 2011.

Opmeer, M. R., Curtain, R. F., LQG balancing for discrete-time infinite-dimensional linear systems, SIAM Journal on Control and Optimization 46(5):1831-1848, 2007.

Reif, K., Sonnemann, F., Unbehauen, R., Modification of the extended Kalman filter with an additive term, Proceedings of the 35th Conference on Decision and Control, Kobe, Japan, Dec. 1996.

Reif, K., Günther, S., Yaz, E., Unbehauen, R., Stabilität des zeitkontinuierlichen erweiterten Kaiman-Filters, Automatisierungstechnik, 46(12):592-601, 1998, translated in English as: Stochastic stability of the discrete-time extended Kalman filter, IEEE Transactions on Automatic Control, 44(4):714-728, 1999.

Tarantola, A., Inverse Problem Theory, Elsevier, 1987.

Xiong, K., Nonlinear Filter and Its Application in Satellite Attitude, PhD thesis, Beihang University, 2006.

Xiong, K., Liu, L. D., and Zhang, H., Modified unscented Kalman filtering and its application in autonomous satellite navigation. Aerospace Science and Technology, 13(4-5):238-246, 2009.

Zabczyk, J., Remarks on the control of discrete-time distributed parameter systems, SIAM J. Control, vol 12, 4, nov. 1974.

A. Observability, controllability

This appendix is a reminder of the definitions of observability and controllability for the steady linear systems. A nice presentation is also proposed by Fletcher (2017) in the context of meteorological assimilation. Owing to significant differences, the finite and infinite-dimensional cases are treated separately hereafter. The definitions are deep-rooted in optimal control theory. The reader will find in the specialized literature the many adaptations for non-steady or nonlinear systems (e.g. Engwerda, 1988). Filtering and control theories are working with the same systems but the points of view are different. The definitions and theorems below apply to the systems regardless of filter or control point of view. However, the definition of controllability is more easily understood from the control point of view. The duality between both theories is further reminded in appendix C.

(*a*) *Finite-dimensional systems*

In finite dimension, the definitions of observability and controllability are classical and relatively simple. In many applications, they can be slightly relaxed as detectability and stabilizability. These properties can all be characterized with the Hautus-Popov-Belevitch test (HPB test; Hautus, 1969). The following system A.1 in matrix form is analogous to 3.1, but the notations are changed. This is to meet the way of thinking of control theory in which the variable u_k is seen as a command instead of ϵ_k seen as a noise:

$$x_{k+1} = \mathbf{A}x_k + \mathbf{B}u_k \tag{A.1a}$$

$$\vec{y}_k = \mathbf{C}x_k + \mathbf{D}u_k \tag{A.1b}$$

Definition A.1. (*observability*): *The steady system* A.1 *is said to be observable if its initial state x_0 can be deduced from knowledge of finitely many successive variables u_k and observations $\vec{y}_k$.*

The noise u_k is mentionned in this definition to mean that it can be eliminated in the retrieval of the initial state. In other words, observability is a property of the homogeneous (noise free) part of the system depending only on matrices $(\mathbf{A}, \mathbf{C})$. It is characterized by the following classical theorem:

Theorem A.2. : *The following are equivalent:*

- *the pair $(\mathbf{A}, \mathbf{C})$ is observable;*
- *the observability matrix* $\begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \vdots \\ \mathbf{CA}^{n-1} \end{bmatrix}$ *has full-column rank N;*
- *there exists no state $x \neq 0$ such that $\mathbf{A}x = \lambda x$ and $\mathbf{C}x = 0$ (HPB test).*

Definition A.3. (*controllability*): *The steady system* A.1 *is said to be controllable if it can be steered from any initial state x_0 to any final state x_f in finite time by an appropriate choice of successive variables u_k.*

Controllability is a property of matrices $(\mathbf{A}, \mathbf{B})$ only. It is characterized by the following classical theorem:

Theorem A.4. : *The following are equivalent:*

- *the pair $(\mathbf{A}, \mathbf{B})$ is controllable;*
- *the controllability matrix $\left[\mathbf{B}, \mathbf{AB}, ..., \mathbf{A}^{n-1}\mathbf{B}\right]$ has full-row rank N;*
- *there exists no state $x \neq 0$ such that $x^\intercal \mathbf{A} = \lambda x^\intercal$ and $x^\intercal \mathbf{B} = 0$ (HPB test).*

The eigenvalues of $\mathbf{A}$, same from left or right, are called the poles of the system and these may be (i) observable or unobservable, (ii) controllable or uncontrollable.

A pole λ is called unobservable if some $x \neq 0$ satisfies $\mathbf{A}x = \lambda x$ and $\mathbf{C}x = 0$. If the

state of the system is decomposed in terms of eigenvectors of $\mathbf{A}$, and more generally in terms of the right characteristic spaces of $\mathbf{A}$, the state components associated with the unobservable poles remain unknown. The unobservable poles such that $|\lambda| < 1$ are said to be stable because as the discrete time k passes, the norm of the corresponding unobserved component of the state is reduced in proportion to $|\lambda|^k$ (or $k^p|\lambda|^k$ with exponent p depending on the multiplicity of λ as an unobservable pole. Thus, if the input variable u_k remains bounded, so does the stable unobserved state component. This remark has led to the following definition inferior to observability but of similar practical interest:

Definition A.5. (*detectability*): *The steady system* A.1 *is said to be detectable if all its unobservable poles are stable.*

Theorem A.6. (*HPB test for detectability*): *The pair* $(\mathbf{A}, \mathbf{C})$ *is detectable if and only if there exists no state* $x \neq 0$ *such that* $\mathbf{A}x = \lambda x$ *with* $|\lambda| \geq 1$ *and* $\mathbf{C}x = 0$.

Similarly, a pole λ is called uncontrollable if some $x \neq 0$ satisfies $x^\intercal \mathbf{A} = \lambda x^\intercal$ and $x^\intercal \mathbf{B} = 0$; again the pole is stable if $|\lambda| < 1$. The components of the system state corresponding to the right characteristic spaces for some uncontrollable pole cannot be controlled. However, if the poles are stable, these components decrease with time. This leads to the following definition, slightly inferior to controllability:

Definition A.7. (*stabilizability*): *The steady system* A.1 *is said to be stabilizable if all its uncontrollable poles are stable.*

Theorem A.8. (*HPB test for stabilizability*): *The pair* $(\mathbf{A}, \mathbf{B})$ *is stabilizable if and only if there exists no state* $x \neq 0$ *such that* $x^\intercal \mathbf{A} = \lambda x^\intercal$ *with* $|\lambda| \geq 1$ *and* $x^\intercal \mathbf{B} = 0$.

(*b*) *Infinite-dimensional systems*

Let's now address the case of systems defined with states in a Hilbert space $\mathcal{S}$ of infinite dimension. The spectral theory of infinite-dimensional operators is relatively complex and the simple Hautus test becomes inoperative. There are many generalizations of the dual concepts of observability/detectability, controllability/stabilizability. A specific vocabulary is often used in the infinite-dimensional setting:

- observability/detectability become output stability/stabilizability,
- controllability/stabilizability become stability/stabilizability often clarified as input stability/stabilizability.

The system 3.1 in operator form is reproduced hereafter. The linear operators $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$, defined between various Hilbert spaces, should all be bounded i.e. continuous.

$$x_{k+1} = \mathcal{A}x_k + \mathcal{B}u_k \tag{A.2a}$$

$$\vec{y}_k = \mathcal{C}x_k + \mathcal{D}u_k \tag{A.2b}$$

Most studies are based on the following very restrictive generalization, inappropriate for most geophysical systems (cf. remark at the end of the appendix). The operator $\mathcal{A}$ is power stabilizable if its spectral radius verifies $r(\mathcal{A}) < 1$. The system A.2 is power output stabilizable (detectable) if a bounded operator $\mathcal{F}$ exists so that $r(\mathcal{A} + \mathcal{B}\mathcal{F}) < 1$, power input stabilizable if a bounded operator $\mathcal{L}$ exists so that $r(\mathcal{A} + \mathcal{L}\mathcal{C}) < 1$.

The following definitions are more general but difficult to handle in practice. They can be found in (Malinen, 2000, stability) and (Opmeer & Curtain, 2007, stability and stabilizability) with in the latter work a little complication due to the formulation of their cost function (see appendix C section a). The pair $(\mathcal{A}, \mathcal{C})$ is output stable (observable) if $\sum_{k=0}^{+\infty} \| \mathcal{C}\mathcal{A}^k x\|^2 < +\infty$ for all $x \in \mathcal{S}$. The system A.2 is output stabilizable (detectable) if a bounded operator $\mathcal{F}$ exists so that the pair $(\mathcal{A} + \mathcal{B}\mathcal{F}, \mathcal{C} + \mathcal{D}\mathcal{F})$ be output stable.

The pair $(\mathcal{A}, \mathcal{B})$ is input stable if $(\mathcal{A}^*, \mathcal{B}^*)$ is output stable. The system A.2 is input stabilizable (controllable) if a bounded operator $\mathcal{L}$ exists so that $(\mathcal{A} + \mathcal{L}\mathcal{C}, \mathcal{B} + \mathcal{L}\mathcal{D})$ be input stable.

Remark: The concept of power input stabilizability is inappropriate for most geophysical systems. To see this, a few spectral theory is required. Remind that the spectrum $\sigma(\mathcal{A})$ of $\mathcal{A}$ contains a subset $\sigma_{ess}(\mathcal{A})$ called the essential spectrum consisting of the accumulation points and eigenvalues of infinite multiplicity. The essential spectrum is stable by compact perturbation (Kato, 1980, theorem 5.35).

Suppose $(\mathcal{A}, \mathcal{C})$ is power stabilizable, i.e. $r(\mathcal{A} + \mathcal{L}\mathcal{C}) = 1 - \varepsilon$ for some bounded $\mathcal{L}$ and $\varepsilon > 0$. Accordingly, $\sigma(\mathcal{A} + \mathcal{L}\mathcal{C})$ is bounded by $1 - \varepsilon$ and so is its subset $\sigma_{ess}(\mathcal{A} + \mathcal{L}\mathcal{C})$. In data assimilation, the number of measurements described by $\mathcal{C}$ is finite. Thus, the bounded operator $\mathcal{L}\mathcal{C}$ having finite rank is compact so that $\sigma_{ess}(\mathcal{A}) = \sigma_{ess}(\mathcal{A} + \mathcal{L}\mathcal{C})$. Since $\mathcal{A}$ is bounded, so is its spectrum. Thus, $\sigma(\mathcal{A})$ contains at most a finite number of points λ_i, $i = 1, 2, ..., , p$, such that $|\lambda_i| \geq 1 - \frac{\varepsilon}{2}$ since otherwise these would accumulate out of the essential spectrum. These isolated points of the spectrum are necessarily eigenvalues of finite multiplicity. By Riesz spectral decomposition (Gohberg et al., 1990), the operator may be written as $\mathcal{A} = \mathcal{A}' + \mathcal{A}''$. In this, $\mathcal{A}'$, $\mathcal{A}''$ are bounded operators, $\sigma(\mathcal{A}') = \{\lambda_1, \lambda_2, ..., \lambda_p\}$, $\sigma(\mathcal{A}'') = \sigma(\mathcal{A}) - \sigma(\mathcal{A}')$ and $\mathcal{A}'\mathcal{A}'' = \mathcal{A}''\mathcal{A}' = 0$. In particular, $\mathcal{A}^k = \mathcal{A}'^k + \mathcal{A}''^k$. Since $r(\mathcal{A}'') \leq 1 - \frac{\varepsilon}{2}$, for any initial state, $\mathcal{A}''^k x_0$ vanishes for sufficiently large k so that the state becomes modelled as $\mathcal{A}'^k x_0$ in the finite dimensional vector space corresponding to the eigenvalues λ_i. Except in a few dissipative situations, this is not acceptable for describing the geophysical processes. The situation is even worse if $\lambda_i > 1$ for some i.

B. Infinite-dimensional background statistics

Geophysical data assimilation is currently based exclusively on finite-dimensional discretised models considered an approximation of the infinite-dimensional geophysical reality. This appendix is aimed at explaining how the concept of a becm, tightly related to the finite dimension, should be evolved in infinite dimension. It is based on the work of Huang & Yan (2000) describing both operator and probability theory in infinite dimension. In finite dimension, a Gaussian probability is defined with a probability density derived from a covariance matrix. The definition of a Gaussian probability measure γ on $\mathcal{S}$ raises two difficulties when this space is infinite-dimensional.

First, in infinite dimension, there is no notion volume and thus, γ may not be defined in terms of a density.

Second, there is also no notion of matrix. The becm must replaced by a background error covariance operator (beco). Logically, the background state and this beco should be defined as averages from integrals: $s^b = \int_{\mathcal{S}} s\gamma(ds)$, $\mathcal{P} = \int_{\mathcal{S}} (s - s^b)(s - s^b)^* \gamma(ds)$. However, the integrands $s \in \mathcal{S}$ and $(s - s^b)(s - s^b)^* \in \mathscr{L}(\mathcal{S})$ take values in infinite-dimensional spaces so that they cannot be integrated in classical Lebesgue sense.

The difficulties are resolved with the following definition of a Gaussian probability measure. The suitability of this definition for both finite and infinite dimensions is supported by theorem B.2.

Definition B.1. *A probability measure on $\mathcal{S}$ is called Gaussian if for any $s \in \mathcal{S}$, the random variable $(s, \cdot)_{\mathcal{S}}$ has a Gaussian distribution.*

The following result is found as definition 4.8 and theorem 4.11 in (Huang & Yan, 2000):

Theorem B.2. *The mean $s^b \in \mathcal{S}$ and covariance operator $\mathcal{P} \in \mathscr{L}(\mathcal{S})$ of a Gaussian probability measure exist as averages in terms of weak integrals, i.e. for all σ_1, σ_2 in $\mathcal{S}$:*

$$(s^b, \sigma_1)_{\mathcal{S}} = \int_{\mathcal{S}} (s, \sigma_1)_{\mathcal{S}} \gamma(ds) \quad \textit{and} \quad (\mathcal{P}\sigma_1, \sigma_2)_{\mathcal{S}} = \int_{\mathcal{S}} (s - s^b, \sigma_1)_{\mathcal{S}} (s - s^b, \sigma_2)_{\mathcal{S}} \gamma(ds)$$

In addition, $\mathcal{P}$ is trace class, i.e. there exists a number $\mathrm{tr}\mathcal{P}$, called the trace of $\mathcal{P}$, that can be computed in some, then in any, Hilbert base $e_i, i = 0, 1, 2, ...$ of $\mathcal{S}$ as:

$$\mathrm{tr}\mathcal{P} = \sum_{i=0}^{+\infty} (\mathcal{P}e_i, e_i)_{\mathcal{S}}, \qquad 0 \leq \mathrm{tr}\mathcal{P} < +\infty$$

Weak integrals, also known as Pettis integrals, are defined based on Lebesgue integrals. In particular, the integrals in theorem B.2 are classical Lebesgue integrals with a real-valued integrand $(s, \sigma_1)_{\mathcal{S}}$ or $(s - s^b, \sigma_1)_{\mathcal{S}}(s - s^b, \sigma_2)_{\mathcal{S}}$ regardless of the fact that $\gamma(ds)$ is an infinite-dimensional measure. The covariance operator is obviously symmetric (i.e. self-adjoint) and non-negative. Notice that, if $\mathcal{S}$ is infinite-dimensional, an operator proportional to identity is not trace class. This is a further reason to discard the ad-hoc assumption of evenly distributed background errors.

C. Optimal control

For the sake of completeness, the relevant concepts of optimal control and Riccati equation (Lancaster & Rodman, chapter 16, 1995) theory are reminded in this section.

(*a*) *Riccati equation for optimal control*

In the framework of control theory, a time invariant linear dynamical system with state x in a Hilbert space $\mathcal{S}$ is traditionally written as:

$$x_{k+1} = \mathcal{A}x_k + \mathcal{B}u_k \tag{C.1a}$$

$$\vec{y}_k = \mathcal{C}x_k + \mathcal{D}u_k \tag{C.1b}$$

in which $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$ are bounded linear operators between Hilbert spaces. The purpose of optimal control is how to choose the control sequence $\boldsymbol{u} = \{u_k\}_{k=0}^{+\infty}$ to steer the system from an initial state x_0 to a prescribed state x. Owing to linearity, this is equivalent to

steering $x_0 - x$ to 0 so that there is no loss of generality in targetting the state 0. A steering strategy consists of finding an optimal control minimizing a quadratic cost function:

$$\boldsymbol{u}^{\text{opt}} = \arg\min_{\boldsymbol{u}} J(x_0, \boldsymbol{u}) \quad \text{with} \quad J(x_0, \boldsymbol{u}) = \sum_{k=0}^{+\infty} \| \vec{y}_k \|^2 \tag{C.2}$$

in which $\vec{y}_k$ is according to the system and simply $\| \vec{y}_k \|^2 = \vec{y}_k^{\intercal} \vec{y}_k$. Using equation C.1b, the cost function C.2 may be rewritten in terms of a scalar product norm $\| \, ., . \, \|_{\mathcal{G}}$ of joint state-control variables with a bounded operator $\mathcal{G}$:

$$J(x_0, \boldsymbol{u}) = \sum_{k=0}^{+\infty} \| x_k, u_k \|_{\mathcal{G}}^2, \qquad \mathcal{G} = \left[\begin{array}{c|c} \mathcal{C}^*\mathcal{C} & \mathcal{C}^*\mathcal{D} \\ \hline \mathcal{D}^*\mathcal{C} & \mathcal{D}^*\mathcal{D} \end{array} \right] \tag{C.3}$$

In view of equation C.3, the problem of optimal control is sometimes formulated without equation C.1b. The observations and related noise are indirectly defined by the choice of the cost function.

The following result is given by Lancaster & Rodman, (theorem 16.6.4; 1995) for the finite-dimensional systems. These authors call the rank condition on **G** *nondegeneracy* which, for a scalar product matrix, is not the standard meaning that it be invertible.

Theorem C.1. *If the system* C.1 *is finite-dimensional, stabilizable and observable and* rank $\mathcal{G}$ = rank $\mathcal{A}$+rank $\mathcal{D}$*, then the problem* C.2 *has a unique solution for any* $x_0 \in \mathcal{S}$*, and this solution is obtained as:* $u_k^{\text{opt}} = \mathcal{F}(\mathcal{A}+\mathcal{B}\mathcal{F})^k x_0$ *with* $\mathcal{F} = -\left(\mathcal{B}^{\intercal}\mathcal{X}\mathcal{B} + \mathcal{D}^{\intercal}\mathcal{D}\right)^{-1}\left(\mathcal{B}^{\intercal}\mathcal{X}\mathcal{A} + \mathcal{D}^{\intercal}\mathcal{C}\right)$ *in which* $\mathcal{X} \in \mathscr{L}(\mathcal{S})$ *is symmetric, positive definite and is the maximal hermitian solution of the Riccati equation:*

$$\mathcal{X} = \mathcal{A}^{\intercal}\mathcal{X}\mathcal{A} - \left(\mathcal{A}^{\intercal}\mathcal{X}\mathcal{B} + \mathcal{C}^{\intercal}\mathcal{D}\right)\left(\mathcal{B}^{\intercal}\mathcal{X}\mathcal{B} + \mathcal{D}^{\intercal}\mathcal{D}\right)^{-1}\left(\mathcal{B}^{\intercal}\mathcal{X}\mathcal{A} + \mathcal{D}^{\intercal}\mathcal{C}\right) + \mathcal{C}^{\intercal}\mathcal{C} \tag{C.4}$$

This theorem has infinite-dimensional counterparts with, unfortunately, a lot more technicalities (e.g. Malinen, 2000).

Remark: Many theoretical works (e.g. Opmeer & Curtain, 2007) study the following operator Riccati equation slightly different from equation C.4:

$$\mathcal{X} = \mathcal{A}^*\mathcal{X}\mathcal{A} - (\mathcal{A}^*\mathcal{X}\mathcal{B} + \mathcal{C}^*\mathcal{D})(\mathcal{B}^*\mathcal{X}\mathcal{B} + \mathcal{I} + \mathcal{D}^*\mathcal{D})^{-1}(\mathcal{B}^*\mathcal{X}\mathcal{A} + \mathcal{D}^*\mathcal{C}) + \mathcal{C}^*\mathcal{C} \tag{C.5}$$

An additional identity operator $\mathcal{I}$ appears in the term $(\mathcal{B}^*\mathcal{X}\mathcal{B} + \mathcal{I} + \mathcal{D}^*\mathcal{D})^{-1}$. Equation C.5 arises from the following optimal problem:

$$\boldsymbol{u}^{\text{opt}} = \arg\min_{\boldsymbol{u}} J'(x_0, \boldsymbol{u}) \quad \text{with} \quad J'(x_0, \boldsymbol{u}) = \sum_{k=0}^{+\infty} \| u_k \|^2 + \| \vec{y}_k \|^2 \tag{C.6}$$

with the cost function J (equation C.2) replaced by J'. This replacement is equivalent to a modification of the output of the dynamical system C.1; operator $\mathcal{G}$ is modified as $\mathcal{G}'$:

$$\begin{aligned} x_{k+1} &= \mathcal{A}x_k + \mathcal{B}u_k \\ \vec{y}_k &= \left[\begin{array}{c} 0 \\ \mathcal{C} \end{array} \right] s_k + \left[\begin{array}{c} \mathcal{I} \\ \mathcal{D} \end{array} \right] u_k, \end{aligned} \qquad \mathcal{G}' = \left[\begin{array}{c|c} \mathcal{C}^*\mathcal{C} & \mathcal{C}^*\mathcal{D} \\ \hline \mathcal{D}^*\mathcal{C} & \mathcal{I} + \mathcal{D}^*\mathcal{D} \end{array} \right] \tag{C.7}$$

(*b*) *Adjoint dynamical system*

The following dynamical system, defined with state space $\mathcal{S}'$ dual to $\mathcal{S}$, is called adjoint to system 3.1:

$$x_{k+1} = \mathbf{E}^\intercal x_k + \mathbf{H}^\intercal u_k \tag{C.8a}$$

$$\vec{y}_k = \mathbf{M}^\intercal x_k + \mathbf{N}^\intercal u_k \tag{C.8b}$$

According to theorem C.1, the equation for the operator $\mathbf{P} \in \mathscr{L}(\mathcal{S}')$ achieving optimal control in this system is:

$$\mathbf{P} = \mathbf{EPE}^\intercal - (\mathbf{EPH}^\intercal + \mathbf{MN}^\intercal)\,(\mathbf{HPH}^\intercal + \mathbf{NN}^\intercal)^{-1}\,(\mathbf{HPE}^\intercal + \mathbf{NM}^\intercal) + \mathbf{MM}^\intercal \tag{C.9}$$

Equation C.9 is seen to coincide with equation 3.3 when $\mathbf{NM}^\intercal = 0$. The latter constraint is same as in equation 2.4 describing independent dynamical and measurement noises in system 3.1. In other words, the becm associated with system 3.1 from the point of view of filtering, is same as the matrix characterizing optimal control in the adjoint system.

Notice that the adjoint system C.8 is observable/detectable or controllable/stabilizable exactly when the original system C.1 is controllable/stabilizable or observable/detectable, respectively. When dealing with infinite-dimensional systems, various definitions are utilized to generalize these notions, always respecting the same duality.

Remark: The link between a system and its adjoint is not submitted to the constraint $\mathbf{NM}^\intercal = 0$. The case $\mathbf{NM}^\intercal \neq 0$ may account for correlated dynamical and measurement noises as studied by Jazwinski (1970, chapter 7, example 7.6). In fact, Jazwinski studies a time varying system with $\mathbf{N}_k\mathbf{M}_k^\intercal \neq 0$. We hasten to point out that this description of correlations between the dynamical noise $\mathbf{M}_k\epsilon_k$ at t_k and the measurement noise $\mathbf{N}_k\epsilon_k$ at t_k is not the most realistic in geophysical practice. Indeed, the dynamical model $\mathcal{E}_k$ describes phenomena after t_k whereas the measurement model $\mathcal{H}_k$ describes phenomena before t_k. Noise correlations are accordingly sequential in nature, between $\mathbf{M}_k\epsilon_k$ and $\mathbf{N}_l\epsilon_l$ for $k > l$.

(*c*) *Adjoint dynamical system: infinite dimension*

When dealing with such infinite dimensional filtering system as 4.1, the derivation of an adjoint system is subject to an additional difficulty. It is indeed necessary to introduce a trace class covariance operator $\mathcal{V}$ for describing the statistics of the noise ϵ_k. This operator interferes with the formulation of the filter Riccati equation 4.3 through definitions 4.2. Such interference is not addressed in the framework of usual optimal control. One shall simply observe that this equation may as well be seen as the control Riccati equation for the system:

$$x_{k+1} = \mathcal{E}^* x_k + \mathcal{H}^* u_k \tag{C.10a}$$

$$\vec{y}_k = \sqrt{\mathcal{V}}\mathcal{M}^* x_k + \sqrt{\mathcal{V}}\mathcal{N}^* u_k \tag{C.10b}$$

Notice that the square root of a bounded positive self-adjoint operator is well defined (Huang & Yan, 2000, theorem 1.27).