Observation data compression for variational assimilation of dynamical systems

Sibo Cheng, Didier Lucor, Jean-Philippe Argaud

Observation data compression for variational assimilation of dynamical systems

Sibo Cheng[1,2,3], Didier Lucor[2], Jean-Philippe Argaud[3]

[1] *Data Science Institute, Department of Computing, Imperial College London, UK*
[2] *Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, France*
[3] *EDF R&D Saclay, France*

Abstract

Accurate estimation of error covariances (both background and observation) is crucial for efficient observation compression approaches in data assimilation of large-scale dynamical problems. We propose a new combination of a covariance tuning algorithm with existing PCA-type data compression approaches, either observation- or information-based, with the aim of reducing the computational cost of real-time updating at each assimilation step. Relying on a local assumption of flow-independent error covariances, dynamical assimilation residuals are used to adjust the covariance in each assimilation window. The estimated covariances then contribute to better specify the principal components of either the observation dynamics or the state-observation sensitivity. The proposed approaches are first validated on a shallow water twin experiment with correlated and non-homogeneous observation error. Proper selection of flow-independent assimilation windows, together with sampling density for background error estimation, and sensitivity of the approaches to the observations error covariance knowledge, are also discussed and illustrated with various numerical tests and results. The method is then applied to a more challenging industrial hydrological model with real-world data and non-linear transformation operator provided by an operational precipitation-flow simulation software.

Keywords: Data assimilation, Observation compression, Error covariance estimation, Information entropy, Hydrological application

1. Introduction

Data assimilation (DA) is applied in a wide range of industrial problems, such as numerical weather prediction (NWP) [1], hydrology, fire forecasting [2] or nuclear engineering [3]. Recently, DA methods have also been used to COVID-19 pandemic analysis, including predicting disease diffusion and proposing optimal vaccination strategies ([4]). DA algorithms are often used in dynamical systems for continuously updating state estimation/prediction. They have recently made their way to other fields such as biomedical applications [5] or quantitative economics [6]. These methods rely on a weighted combination of different sources of noisy information, including prior numerical estimation (also known as background states) and real-time observations, to improve field reconstruction or parameters calibration. DA methods are often used to deal with problems of large dimensions, especially in NWP [7], [8] (up to 10^9) or in geoscience [9], leading to computational difficulty for real-time updating, if not infeasible. Several strategies for optimizing the computational cost have been developed, including graph-based domain localization [10], observations selection [11], matrix decomposition [12] or reduced-order Kalman Filter [13]. It is also a common practice to combine DA algorithms with classical dynamical system reduction techniques, such as the Proper Orthogonal Decomposition (POD) or the Empirical Interpolation Method (EIM e.g. [14]). Most of these methods rely on either precise knowledge of state variables (e.g. modes in POD) or strong prior assumptions (e.g. cut-off radius in domain localization [15]). Meanwhile, with the increase of available observation precision in DA applications, the observation data compression via low-rank approximation methods has been continuously studied for alleviating the computational cost, especially in a sequential data assimilation chain. These methods, which consist of extracting principal information in observation data, have been widely applied in various branches of engineering, especially for high dimensional problems. An important advantage of observation compression, regarding other methods that directly reduce the state space dimension, is that no extra operation/-knowledge of the state dynamics is required, making the compression error more controllable and estimable. Two classical compression methods are discussed and implemented in this work: the POD-type projection by extracting principal components in the observation dynamic [16] and the information-based compression based on the information entropy analysis [8]. The latter aims to select the most impacting observations to the analyzed state by calculating the prior-posterior information entropy gap. Since the noises are introduced by prior errors in DA systems, the information entropy estimation relies on both background and observation error covariance matrices.

For both observation- and information-based approaches, the data compression is carried out with a noise-normalized dataset [7], [8]. The knowledge of prior error covariances thus becomes crucial for applying these methods. However, the specification of these covariances, especially the background matrix, remains one of the most challenging problems in data assimilation due to the high dimension of the problem and limited prior data [17],[18]. Much attention was given to improving the error covariance specification in dynamical data assimilation models, particularly by the meteorological society. Several methods have been developed to this end, such as the NMC approach [19], the DI01 [20] iterative method and the Desroziers estimation [21]. In this paper, we focus on the latest. Unlike some other methods (e.g. [20], [18]), the Desroziers estimation does not depend on the specific structure of the error covariances, and it provides a non-parametric estimation of full covariances as output of the algorithm. Based on the residual analysis in variational assimilation, this approach has been widely applied in industrial problems, especially in NWP. Recent works of [22] prove its convergence in the ideal case. Another considerable strength of the Desroziers estimation is that dynamic residual data can be used for the covariance estimation. For this reason, a huge ensemble size is not required for high dimensional problems, unlike, for instance, in the NMC method.

In this paper, based on the Desroziers estimation, we have introduced the concept of piecewise covariance estimation for both observation- and information-based compression strategies. We apply the Desroziers method to estimate error covariances in a fixed time range, also known as the flow-independent window where the error covariances are supposed to be time-invariant. Therefore, the choice of the flow-independent window and the residual samplings play an essential role in this algorithm. The window size should be sufficiently long to gather enough time-variant sampling but not too long to consider the error covariances, especially the background matrix, being constant.

The observation- and information-based (with piecewise covariances estimation) data compression are first implemented in a twin experiment framework using 2D shallow water equations with a linear transformation operator. The observation covariance is supposed to be perfectly known *a priori*. The two approaches with different choices of flow-independent windows are compared in this model while changing the truncation parameter. Numerical results show that the observation-based (POD-type) compression is in general over-performed by the information-based approach and that a non-balanced sampling in piecewise covariance estimation results in a less optimal compression. We then apply these methods to a real-world hydrological model to improve river flow prediction/reanalysis by correcting historical daily precipitation measures [23]. Both the precipitation and the river flow data are spatially distributed. The physical simulation is performed using the operating MORDOR-TS software [24], developed by EDF and the study area is around the Tarn river, in the south of France. The precipitation-flow simulation is carried out through conceptual watersheds modeling, which ensures its high computational efficiency. In this hydrological application, both the background and the observation matrices are estimated using the Desroziers method with daily observed flow data for around 10 years (1990 to 2000). Results show that in this industrial application where both $\mathbf{B}$ and $\mathbf{R}$ are not well known, the performance of the information-based strategy is similar to the one of observation-based.

The paper is organized as follows. In section 2, the principle and the notation of data assimilation are briefly introduced. We then introduce the observation- and information-based compression strategies in section 3. The applications of 2D shallow water twin experiments and an industrial hydrological model are shown respectively in section 5 and 6. We finish the paper with a discussion.

2. Variational data assimilation

The objective of data assimilation algorithms is to improve the estimation of some physical fields or parameters $\mathbf{x}$ based on two sources of information: a prior simulation/forecast $\mathbf{x}_b$ and an observation vector $\mathbf{y}$. The theoretical value of the current state is denoted by a vector $\mathbf{x}_{\text{true}}$, also known as the true state. Variational DA algorithms aim to find an optimally weighted compromise between the prior estimation $\mathbf{x}_b$ and the observation $\mathbf{y}$ by minimising the cost function J defined as

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\mathbf{x}_b)^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b) + \frac{1}{2}(\mathbf{y}-\mathcal{H}(\mathbf{x}))^T\mathbf{R}^{-1}(\mathbf{y}-\mathcal{H}(\mathbf{x})) \tag{1}$$

$$= \frac{1}{2}||\mathbf{x}-\mathbf{x}_b||^2_{\mathbf{B}^{-1}} + \frac{1}{2}||\mathbf{y}-\mathcal{H}(\mathbf{x})||^2_{\mathbf{R}^{-1}} \tag{2}$$

where $\mathcal{H}$ denotes the transformation operator from the state space to one of the observations. $\mathbf{B}$ and $\mathbf{R}$ are the associated error covariance matrices, i.e.

$$\mathbf{B} = \text{Cov}(\epsilon_b, \epsilon_b), \quad \mathbf{R} = \text{Cov}(\epsilon_y, \epsilon_y), \tag{3}$$

where

$$\epsilon_b = \mathbf{x}_b - \mathbf{x}_{\text{true}}, \quad \epsilon_y = \mathcal{H}(\mathbf{x}_{\text{true}}) - \mathbf{y}. \tag{4}$$

Thus the inverse of these covariance matrices (i.e. $\mathbf{B}^{-1}, \mathbf{R}^{-1}$) represents the weights of these two information sources in the objective function. Prior errors ϵ_b, ϵ_y are supposed to be centered Gaussian, characterised by the error covariance matrices, i.e.

$$\epsilon_b \sim \mathcal{N}(0, \mathbf{B}), \quad \epsilon_y \sim \mathcal{N}(0, \mathbf{R}). \tag{5}$$

The optimization problem of Eq. 1, so called three-dimensional variational (3D-Var) formulation, is a general representation of variational assimilation while the model error is not considered. The output of Eq. 1 is denoted as $\mathbf{x}_a$, i.e.

$$\mathbf{x}_a = \underset{\mathbf{x}}{\text{argmin}}\Big(J(\mathbf{x})\Big). \tag{6}$$

If $\mathcal{H}$ can be approximated by some linear operator $\mathbf{H}$, Eq. 6 can be solved via BLUE (Best Linearized Unbiased Estimator) formulation,

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}_b) \tag{7}$$

$$\mathbf{A} = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{B} \tag{8}$$

where $\mathbf{A} = \text{Cov}(\mathbf{x}_a - \mathbf{x}_{\text{true}})$ is the analyzed error covariance and the $\mathbf{K}$ matrix, given by

$$\mathbf{K} = \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1} \tag{9}$$

is so called the Kalman gain matrix. In the rest of this paper, we denote $\mathbf{H}$ as the linearized transformation operator. The case when $\mathcal{H}$ is non-linear is more challenging for finding the minimum of Eq. 1, especially for high-dimensional problems. The resolution involves often gradient descent algorithms (relying on algorithms such as "L-BFGS-B" [25] and on adjoint-based [11] numerical techniques.

Variational assimilation algorithms could be applied to dynamical systems through sequential applications using a transition operator $\mathcal{M}_{t^k \to t^{k+1}}$ (from time t^k to t^{k+1}), where

$$\mathbf{x}_{t^{k+1}} = \mathcal{M}_{t^k \to t^{k+1}}(\mathbf{x}_{t^k}). \tag{10}$$

The forecasting thus depends on the knowledge of transition operator $\mathcal{M}_{t^k \to t^{k+1}}$ and the corrected state at the current time $\mathbf{x}_{a,t^k}$. Typically, the current background state is often given by the forecasting from the previous step, i.e.

$$\mathbf{x}_{b,t^k} = \mathcal{M}_{t^{k-1} \to t^k}(\mathbf{x}_{a,t^{k-1}}). \tag{11}$$

Obviously, a more accurate reanalysis $\mathbf{x}_{a,t^{k-1}}$ leads to a more reliable forecasting $\mathbf{x}_{b,t^k}$. It is known that as long as the transformation operator $\mathcal{H}$ and the transition operator $\mathcal{M}$ are linear, the analysis based on the variational method and the Kalman filter results in the same forecasting [9], for dynamical (4D-Var) assimilation problems. Theoretically, the evolution of the $\mathbf{B}$ matrix could also be estimated thanks to the transition operator. However, in practice, the pefect knowledge of $\mathcal{M}$ is often unavailable. Much attention is given to quantify the model error in assimilation, for example, in weak-constraint *4D-VAR* [26]. Recent work of [27] involves deep learning techniques to improve the estimation of $\mathcal{M}_{t^{k-1} \to t^k}$.

3. Observation data compression

DA algorithms are often used to perform real-time corrections of dynamical systems with large dimensions, leading to an essential requirement of computational efficiency. In this work, we are interested in a low-rank approximations of the observation vector which can reduce the cost of real-time updating in DA algorithms.

3.1. Observation-based compression (OC)

The works of [28] and [29] are based on a PCA-type reduction of the observation dynamics. More precisely, a set of n_{obs} observation snapshots is represented by a matrix $\mathbf{Y} \in \mathbb{R}_{[\dim(\mathbf{y}) \times n_{obs}]}$ where each column $\mathbf{Y}[:,.]$ represents an individual observation vector of dimension m at a fixed time t_i, i.e.

$$\mathbf{Y}[:,i] = \mathbf{y}_{t=t_i}. \tag{12}$$

Thus $\mathbf{Y}$ describes the evolution of the observation vector $\mathbf{y}$ including observation error. We work with the error-normalized data $\mathbf{R}^{-1/2}\mathbf{Y}$ [7] whose empirical covariance $\mathbf{C}$ can be written and decomposed as

$$\mathbf{C} = \frac{1}{n_{obs}-1}\mathbf{R}^{-1/2}\mathbf{Y}\mathbf{Y}^T\mathbf{R}^{-1/2} = \tilde{\mathbf{L}}\tilde{\mathbf{D}}\tilde{\mathbf{L}}^T \tag{13}$$

where the columns of $\tilde{\mathbf{L}}$ are the principal components and $\tilde{\mathbf{D}}$ represents the associated eigenvalues in a decreasing order. This decomposition is known as the principal component analysis (PCA) decomposition. We can construct a projection operator $\tilde{\mathbf{L}}_q$ with minimum loss of information (represented by eigenvalues in the covariance matrix) by simply keeping the q first columns in $\tilde{\mathbf{L}}$. q is also known as the truncation parameter. In fact, this projection operator can also be obtained by a singular value decomposition (SVD), without computing the full covariance matrix $\mathbf{C}$, i.e.

$$\mathbf{R}^{-1/2}\mathbf{Y} = \tilde{\mathbf{L}}_q \tilde{\Sigma} \tilde{\mathbf{V}}_q^{\ T} \tag{14}$$

where $\tilde{\mathbf{L}_q}$ and $\tilde{\mathbf{V}}_q$ are orthogonal matrices, i.e. $\tilde{\mathbf{L}_q}^T \tilde{\mathbf{L}_q} = \tilde{\mathbf{V}_q}^T \tilde{\mathbf{V}_q} = \mathbf{I}$ and $\tilde{\Sigma}\tilde{\Sigma}^T = \tilde{\mathbf{D}}$ since all eigenvalues are non negative. The assumption is made for the observation error covariances to be constant (flow-independent), which is a common pratice in data assimilation (e.g [7]). For each DA optimization, instead of updating with the full observation vector $\mathbf{y}$, the correction is made with the reduced observation

$$\tilde{\mathbf{y}}_q = \tilde{\mathbf{L}}_q^T \mathbf{R}^{-1/2}\mathbf{y}. \tag{15}$$

The new observation error covariance $\tilde{\mathbf{R}}$ and the new state-observation transformation operator $\tilde{\mathcal{H}}$ can be written as

$$\tilde{\mathbf{R}}_q = \tilde{\mathbf{L}}_q^T \mathbf{R}^{-1/2}\mathbf{R}\mathbf{R}^{-1/2}\tilde{\mathbf{L}}_q = \mathbf{I}_q, \quad \tilde{\mathcal{H}}_q = \tilde{\mathbf{L}}_q^T \mathbf{R}^{-1/2} \circ \mathcal{H}. \tag{16}$$

The DA algorithm can then be performed on $(\mathbf{x}_b, \tilde{\mathbf{y}}_q, \mathbf{B}, \tilde{\mathbf{R}}_q, \tilde{\mathcal{H}}_q)$ instead of $(\mathbf{x}_b, \mathbf{y}, \mathbf{B}, \mathbf{R}, \mathcal{H})$. This method could be seen as a classical POD approach applied to error-normalised observation data by extracting modes of higher variances against time. It is pointed out by [28] and [7] that performing PCA on noise-normalised observation data can improve the method efficiency and reduce the impact of observation error during the compression procedure.

3.2. Information-based compression (IC)

The observation-based data reduction retains the principal directions of the observation dynamic. However, these directions are not necessarily the most impacting in state correction. A continuous effort has been devoted to quantify and compute the sensitivity of the analysis states to the observations (e.g. [11]), which leads to a more refined observation compression in DA. More precisely, this sensitivity may be expressed by the influence matrix $\mathbf{S}$ [30], defined as

$$\mathbf{S} = \frac{\partial \mathcal{H}(\mathbf{x}_a)}{\partial \mathbf{x}_a} = \mathbf{K}^T\mathbf{H}^T. \tag{17}$$

According to [8], the information given by the influence matrix can be roughly quantified via two indicators, the degree of freedom for signal (DFS) which represents the prior-posterior mutual information and the entropy reduction (ER) which represents the evolution of Shannon information content, respectively defined as

$$\text{DFS} = \mathbb{E}[(\mathbf{x}_a - \mathbf{x}_b)^T \mathbf{B}^{-1}(\mathbf{x}_a - \mathbf{x}_b)] = Tr(\mathbf{S}) \tag{18}$$

$$\text{ER} = H(\mathbf{x}) - H(\mathbf{x}|\mathbf{y}) = -\frac{1}{2}\ln\Big(det(\mathbf{I} - \mathbf{S})\Big) \tag{19}$$

where H is the entropy of a distribution, noted here $H(x)$ for simplicity. Eqs. 18 and 19 are derived for a centred Gaussian vector $\mathbf{x}$. For both measures, we observe that observations associated with the largest eigenvalues of $\mathbf{S}$ have the greatest information content. Using an intermediate matrix $\mathbf{M} = \mathbf{R}^{-1/2}\mathbf{H}\mathbf{B}^{1/2}$, Eqs. 18-19 could be rewritten as

$$\text{DFS} = \text{Tr}(\mathbf{M}\mathbf{M}^T(\mathbf{I} + \mathbf{M}\mathbf{M}^T)^{-1}) \tag{20}$$

$$\text{ER} = \frac{1}{2}\ln\Big(det(\mathbf{I} + \mathbf{M}\mathbf{M}^T)\Big). \tag{21}$$

As stated in the work of [31], the observation projection operator which minimizes the information loss is given by $\hat{\mathbf{L}}_q\mathbf{R}^{-1/2}$, where $\hat{\mathbf{L}}_q$ is the matrix whose columns contain the eigenvectors of $\mathbf{M}\mathbf{M}^T = \mathbf{R}^{-1/2}\mathbf{H}\mathbf{B}\mathbf{H}^T\mathbf{R}^{-1/2}$. DA algorithms could then be performed with

$$\hat{\mathbf{y}}_q = \hat{\mathbf{L}}_q^T \mathbf{R}^{-1/2}\mathbf{y}, \quad \hat{\mathbf{R}}_q = \mathbf{I}_q, \quad \hat{\mathcal{H}}_q = \hat{\mathbf{L}}_{q_q}^T \mathbf{R}^{-1/2} \circ \mathcal{H}. \tag{22}$$

We remind that, from the computational point of view, the only difference between the OC and IC is the way the low-rank projection $\mathbf{L}_q$ is obtained. For both approaches, the specification of error covariance matrices (either background or observation) is crucial to provide an efficient compression. On the other hand, data compression strategies can reduce the computational cost of covariance tuning methods, especially for multidimensional and multivariate problems. Therefore, the precise knowledge of $\mathbf{HBH}^T$ and $\mathbf{R}$ is crucial for this method. However, as pointed out by [8], the condition number of the analysis covariance matrix $\mathbf{A}$ can be higher when using IC approach compared to performing DA with the full observation data set. Therefore, the risk of matrix ill-conditioning is worth monitoring when applying this compression method.

3.3. Optimal truncation parameter for compression methods

The determination of the truncated parameter q, i.e. number of modes kept in the reduced space, is crucial in data compression. The choice of the threshold often depends on available data [32]. Several criteria were considered, such as the information losing rate E_q and the matrix conditioning *a posteriori* μ_q, defined as

$$E_q = \frac{||\Sigma - \phi_q||_\infty}{||\Sigma||_\infty} = 1 - \frac{\sigma_{q-1}}{\sigma_q} \tag{23}$$

$$\mu_q = \frac{\sigma_1}{\sigma_q} \tag{24}$$

where Σ is the diagonal matrices with all eigenvalues of the covariance matrix and $\sigma_{i,i=1..}$ represent the associated real eigenvalues in the decreasing order of absolute value. According to the study of [33], an optimal choice of the truncation parameter can be obtained by combining the two previous indicators, with an objective function f, defined as

$$f(\sigma_q) = E_q + \mu_q = \frac{\sigma_q \sigma_{q-1} + \sigma_1}{\sigma_1 \sigma_q}. \tag{25}$$

Assuming $||\sigma_q - \sigma_1|| >> ||\sigma_q - \sigma_{q-1}||$, one could easily prove that Eq. 25 achieves the minimum when $\sigma_q = \sqrt{\sigma_1}$. With this choice, we manage to both reduce the matrix ill-conditioning and remove less significant modes, as proved in real-world DA application[33]. Another advantage of this criteria is that the computation of the full spectrum of covariances is not required. By applying Lancozs-type methods [34], we can stop the algorithm when the current eigenvalue is inferior to $\sqrt{\sigma_1}$.

4. Piecewise estimation of error covariances

The $\mathbf{R}$ matrix is required for both OC and IC approaches. Furthermore, the construction of $\hat{\mathbf{L}}_q^T$ in IC requires a precise knowledge of the matrix production $\mathbf{HBH}^T$. However, the knowledge of both matrices often remains challenging in data assimilation [17]. Continuous effort was devoted to improve the error covariances specification [35], [18]. A classical approach based on residual analysis, and later a more complete version are respectively given by [36] and [21]. They show that under the assumption of flow-independent error covariance, i.e. $\mathbf{B}$ and $\mathbf{R}$ being invariant against time in a certain period, the following equations hold

$$\mathbf{R} = \mathbb{E}\Big[\Big(\mathbf{y} - \mathcal{H}(\mathbf{x}_a)\Big)\Big(\mathbf{y} - \mathcal{H}(\mathbf{x}_b)\Big)^T\Big] \tag{26}$$

$$\mathbf{HBH}^T = \mathbb{E}\Big[\Big(\mathbf{y} - \mathcal{H}(\mathbf{x}_b)\Big)\Big(\mathbf{y} - \mathcal{H}(\mathbf{x}_b)\Big)^T\Big] - \mathbf{R}. \tag{27}$$

Under these hypothesis, combining Eq. 26 and 27 leads to

$$\mathbf{HBH}^T = \mathbb{E}\Big[\Big(\mathcal{H}(\mathbf{x}_a) - \mathcal{H}(\mathbf{x}_b)\Big)\Big(\mathbf{y} - \mathcal{H}(\mathbf{x}_b)\Big)^T\Big]. \tag{28}$$

In order to somewhat alleviate these strong hypotheses, a simple idea is to take the expectation operators in Eq. 26, 27 and 28 in assimilation windows where the flow-independent assumption stands, resulting in a piecewise estimation of both $\mathbf{B}$ and $\mathbf{R}$. More precisely, a sequence of estimated background matrices $\mathbf{B}_{T_i}$ could be computed via residual covariances, where T_i refer to flow-independent periods of $\mathbf{B}$ in a dynamical system. In other words, $\mathbf{B}$ is considered as invariant between $t = T_i$ and $t = T_{i+1}$. The estimation of $\mathbf{R}_{T_i}$, if required, follows the same principle using Eq. 26. When the knowledge of $\mathbf{R}$ matrix is precise *a priori*, the estimation of Eq. 27 is privileged because of its lower computational cost since no evaluation of the analyzed state $\mathbf{x}_a$ is required. According to [36], when the observation error is dominated by background error (i.e. $\mathrm{Tr}(\mathbf{R}) << \mathrm{Tr}(\mathbf{B})$), $\mathbf{HBH}^T$ can be estimated directly by $\mathbb{E}\Big[\Big(\mathbf{y}-\mathcal{H}(\mathbf{x}_b)\Big)\Big(\mathbf{y}-\mathcal{H}(\mathbf{x}_b)\Big)^T\Big]$.

By definition,

$$\mathbf{HBH}^T = \mathbb{E}\Big[\Big(\mathcal{H}(\mathbf{x}_{\text{true}}) - \mathcal{H}(\mathbf{x}_b)\Big)\Big(\mathcal{H}(\mathbf{x}_{\text{true}}) - \mathcal{H}(\mathbf{x}_b)\Big)^T\Big] \tag{29}$$

represents the background error covariances projected in the observation space.Therefore the information-based observation compression which is based on a PCA-type analysis, can also be interpreted as a projection of $\mathbf{y}$ along the directions where the background errors are most important. Recently, it was also reported in the literature (e.g. [37]) that the convergence (towards the exact observation matrix) of the iterative method can still be ensured when the background and observation error correlation length-scales are similar, which was contrary to what was previously thought [38]. Although this innovation-based covariance estimation approach has been widely applied in DA applications, some drawbacks have also been noticed. For example, the application of this method in real problems often requires post-processing of the $\mathbf{R}$ matrix. It is shown in [23] that the regularized matrix may converge to some other solution rather than the exact observation matrix.

5. Shallow water twin experiments

5.1. Experiments set up

For evaluating the performance of different data compression approaches, we set up a twin experiment framework with a simplified 2D shallow water dynamical model which is frequently used for testing data assimilation algorithms ((e.g [11], [18]). A cylinder of water is positioned in the middle of the study field of size $20mm \times 20mm$ and released at the initial time $t = 0s$ (i.e. with no initial speed), leading to a non-linear wave-propagation. The dynamics of the water level h (in mm), as well as horizontal and vertical velocity (in $0.1m/s$) field (respectively denoted as u and v), is given by the non-conservative shallow water equations

$$\begin{aligned}
\frac{\partial u}{\partial t} &= -g\frac{\partial}{\partial x}(h) - bu \\
\frac{\partial v}{\partial t} &= -g\frac{\partial}{\partial y}(h) - bv \\
\frac{\partial h}{\partial t} &= -\frac{\partial}{\partial x}(uh) - \frac{\partial}{\partial y}(vh) \\
u_{t=0} &= 0 \\
v_{t=0} &= 0
\end{aligned} \tag{30}$$

where $b = 0.1$ is the viscous drag coefficient and the earth gravity constant g is thus scaled to 1. These equations are discretized in a 20×20 regular grid, solved by first-order finite difference method with a time discretization $\delta_t = 10^{-4}s$. This resolution is considered as the reference (i.e. the true state $\mathbf{x}_{\text{true}}$) latter when performing DA algorithms. The state variables in this DA modeling are the combination of the velocity fields $\{u\}_{20\times 20}$ and $\{v\}_{20\times 20}$. The evolution of the reference ($\mathbf{x}_{\text{true},t}$) state is illustrated in Fig. 1. Spatially correlated prior error is then generated artificially for simulating the background state with a standard deviation $\sigma_{b,0} = 0.2$, i.e.

$$\mathbf{x}_{b,t=0} \sim \mathcal{N}(\mathbf{x}_{\text{true},t=0}, \mathbf{B}_{t=0}) \quad \text{where} \quad \mathbf{B}_{t=0} = {\sigma_{b,0}}^2 \text{corr}(\mathbf{B}). \tag{31}$$

The background error correlation matrix corr($\mathbf{B}$) is set to be isotropic (rotational invariant), following the second-order auto-aggressive (SOAR, also known as Balgovind) function,

$$\phi_{\mathbf{B}}(r) = \left(1 + \frac{r}{L_{\mathbf{B}}}\right) \exp(-\frac{r}{L_{\mathbf{B}}}), \tag{32}$$

where r denotes the spatial distance and $L_{\mathbf{B}}$ is the correlation scale length, fixed as $L_{\mathbf{B}} = 4$ in this application. Being part of Matern kernels, the SOAR function is often used in DA for prior error correlation modeling [18],[3] thanks to its smoothness and good conditioning. The simulation of $\mathbf{x}_{b,t} = [u_{b,t}, v_{b,t}]$ via the same discretization of Eq. 30 (except the initial conditions) is used as background states at time t in the DA modeling. For the knowledge of the exact[1] background error covariance $\mathbf{B}_{\text{E},t}$ at different time, 10^3 background trajectories $\{\mathbf{x}_{b,t}^{\gamma=1\ldots1000}\}$ are independently generated via Eq. 31. This exact matrix, hidden for compression approaches, is seldomly used to evaluate the performance of DA algorithms with reduced observation. To simulate an industrial context, only 10 trajectories $\{\mathbf{x}_{b,t}^{\gamma=1\ldots10}\}$ are used in the piecewise estimation of $\mathbf{HBH}^T$ of a flow-independent window, making the ensemble size (10) much smaller than the problem dimension ($20 \times 20 = 400$).

The observations in these twin experiments are generated from the model equivalent based on the true states (i.e $\mathbf{H}(\mathbf{x}_{\text{true}})$), separately for the fields u and v, respectively denoted as $\mathbf{y}_u$ and $\mathbf{y}_v$. For both fields, the observation $\mathbf{y}_t = [\mathbf{y}_{u,t}, \mathbf{y}_{v,t}]$ at time t is the sum of u_t and v_t in a 2×2 cells area with an observation error ϵ_{y_t},

$$\mathbf{y}_{u,i,j,t} = u_{\text{true},2i,2j,t} + u_{\text{true},2i+1,2j,t} + u_{\text{true},2i,2j+1,t} + u_{\text{true},2i+1,2j+1,t} + \epsilon_{y_{u,i,j,t}} \tag{33}$$

and identical for $\mathbf{y}_{v,i,j,t}$. Thus $\mathbf{y}$ represents also the evolution of the velocity field u and v with a "coarser" measure as shown in Fig. 1 [g-h].

In these experiments, we have set a non-homogeneous observation error covariance where the error deviation in the center (of radius 4) of the field is 4 times higher, compared to boundary observations as show in Fig. 3[a]. They are both of the same order of magnitude as $\sigma_{b,0}$, following also the SOAR function with a smaller scale length $L_{\mathbf{R}} = 1$, compared to background error correlation. The full error covariance $\mathbf{R}$ of observations $\mathbf{y}$ (after being converted to a 1D vector by concatenating rows of the original 2D grid model), supposed invariant against time, is illustrated in Fig. 3 [b]. The $\mathbf{R}$ matrix is supposed to be known in this application, thus only 10 observation trajectories $\{\mathbf{y}_t^{\gamma=1\ldots10}\}$ are generated to simulate an ensemble of small size while evaluating $\mathbf{HBH}^T$ through Eq. 27. In this experiment, we make the choice to circumvent the difficulty by setting a temporal correlated $\epsilon_{b,t}$, as the background noises are only added at the beginning of the simulation, and a temporal uncorrelated $\epsilon_{y,t}$. In fact, temporally correlated background errors are difficult to handle for the Desroziers method since it treats the innovation quantities as independent samples for covariance estimation. These assumptions are realistic and widely adopted in DA problems since background simulations are often taken successively while observations are usually discrete. However, it is beneficial to have both time uncorrelated $\epsilon_{b,t}$ and $\epsilon_{y,t}$ for Desroziers-type estimation, as long as the error covariance could still be considered flow-independent [22].

[1]Here, by the term "exact", we refer to the covariance truly corresponding to the prior errors present in the background state, no matter the level of optimality of the chosen assimilation scheme.

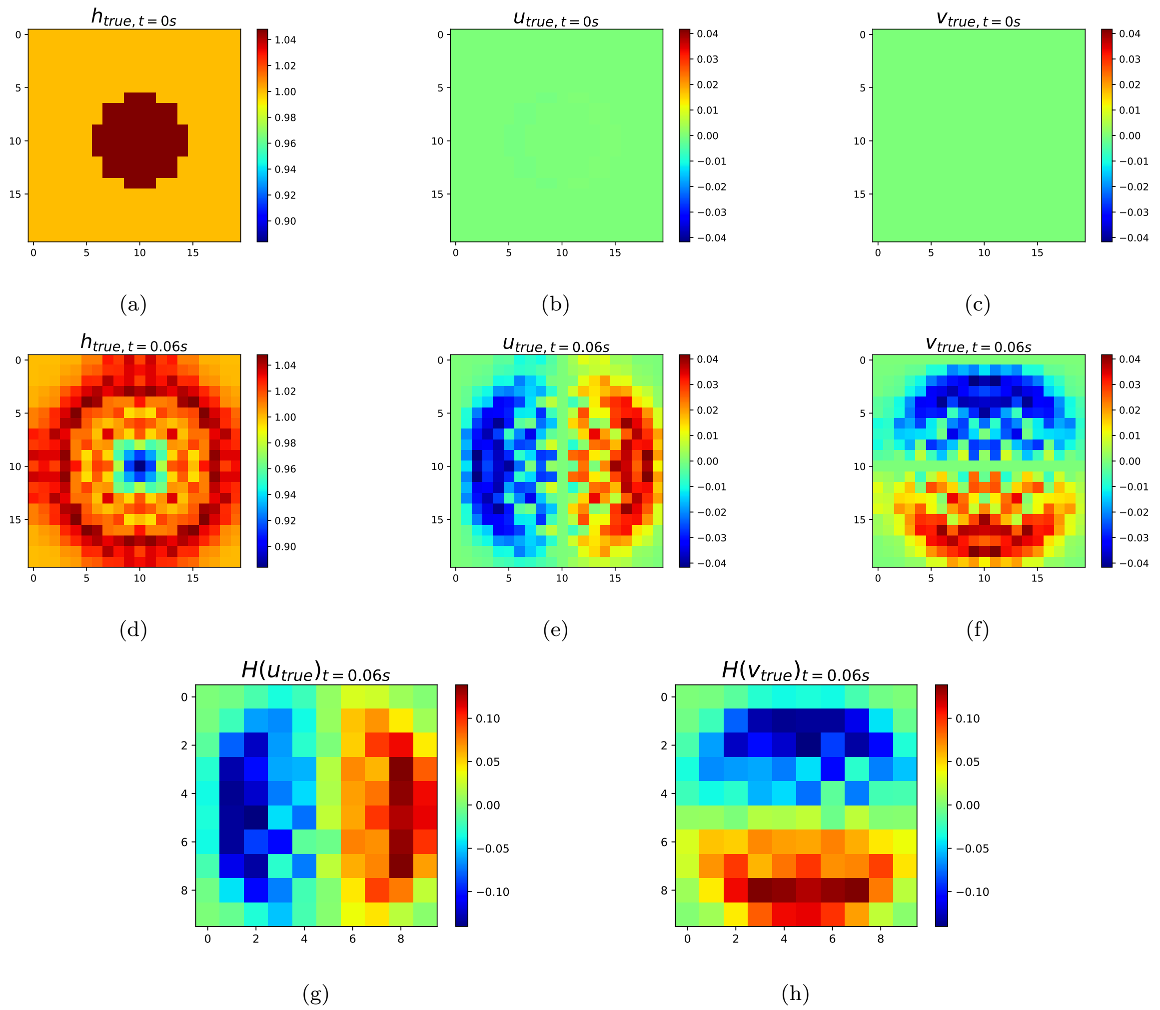


Figure 1: Evolution of the shallow water model of h, u, v (true states) at different time steps (a-f) and the error-free model equivalent $\mathbf{H}(\mathbf{x}_{true})$ for observations (g-h).

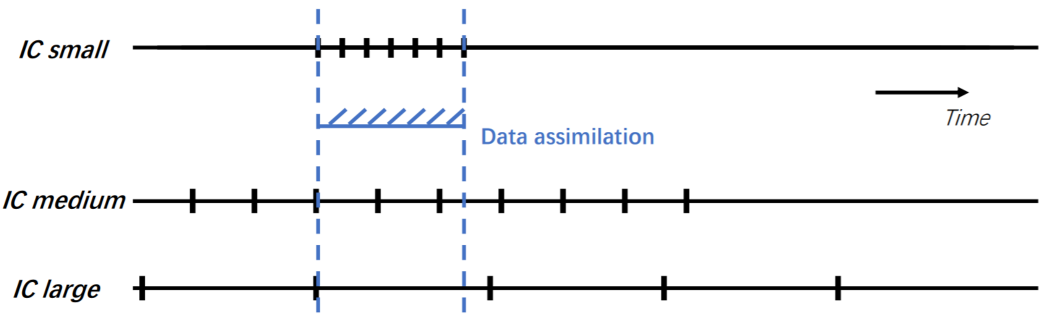
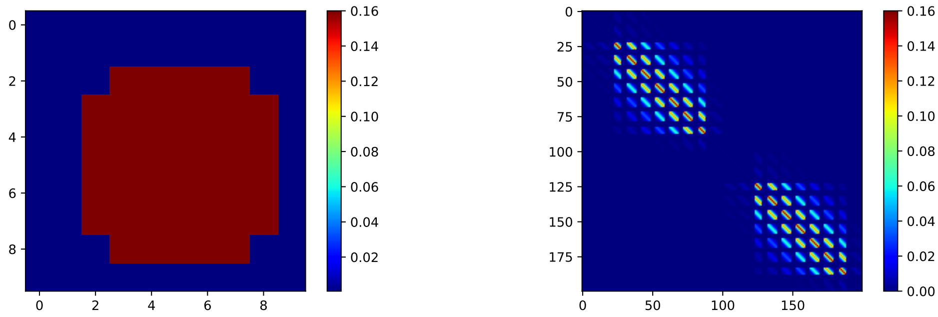


Figure 2: Simple sketch illustrating the three IC sampling strategies. The two vertical blue lines indicate where data assimilation experiments take place (as mentioned in Eq. 34).



(a) (b)

Figure 3: The observation error variance of $\mathbf{y}_u$ and $\mathbf{y}_v$ in the shallow water model[a] and the Balgovind error covariances ($\mathbf{R}$) after the observation vector (originally in a 2D grid) being converted to a 1D vector [b]

5.2. Numerical results for different compression strategies

We then apply different strategies of observation compression and compare the performance of 3D-Var data assimilation using the reduced observation data. For each assimilation, only the current observation $\mathbf{y}_t$ is used to correct the background state $\mathbf{x}_{b,t}$. Thanks to the 1000 background trajectories $\{\mathbf{x}_{b,t}^{\gamma=1...1000}\}$ simulated, the exact $\mathbf{B}_{\mathrm{E},t}$ matrix can be empirically estimated at different time steps, allowing an accurate estimation of analysis error covariance $\mathbf{A}_t$ via Eq. 8 since the $\mathbf{R}$ matrix is supposed to be known. The matrix trace $\mathrm{Tr}(\mathbf{A})$ then represents the sum of marginal analysis error, equivalent to the square of L^2 norm, i.e $\mathbb{E}\left(||\mathbf{x}_a - \mathbf{x}_t||_2^2\right)$, often used as an important indicator of DA schemes [18].

Another objective of this experiment is to inspect the impact on the assimilation error given by different sampling densities, which is critical in information-based compression, as stated in the introduction. We display three sampling strategies for $\mathbf{HBH}^T$ estimation with different assumed flow-independent periods $[T_{\mathrm{s}}, T_{\mathrm{f}}]$,

- *IC small*: Dense sampling in a small period, $\Delta_t = 0.001s$ with $T_{\mathrm{s}} = 0.16s$ and $T_{\mathrm{f}} = 0.18s$
- *IC large*: Sparse sampling in a long period, $\Delta_t = 0.1s$ with $T_{\mathrm{s}} = 0s$ and $T_{\mathrm{f}} = 2s$
- *IC medium*: Between *IC small* and *IC large*, with $\Delta_t = 0.01s$ $T_{\mathrm{s}} = 0.1s$ and $T_{\mathrm{f}} = 0.3s$,

as shown in Fig. 2, where Δ_t is the uniform time discretization between two snapshots. For all these three strategies, the $\mathbf{HBH}^T$ is estimated via 20 time steps (i.e $T_{\mathrm{f}} - T_{\mathrm{s}} = 20\Delta_t$), each with 10 background ($\{\mathbf{x}_{b,t}^{\gamma=1...10}\}$) and observation ($\{\mathbf{y}_t^{\gamma=1...10}\}$) states/residuals. To gain a robust comparison, the posterior error variance $\mathcal{E}_{\text{posterior}}$ is averaged using $\mathrm{Tr}(\mathbf{A}_t)$ at four different time, included in all three assumed flow-independent windows,

$$\mathcal{E}_{\text{posterior}} = \frac{\sum \mathrm{Tr}(\mathbf{A}_t)}{4} \quad \text{for} \quad t \in \{0.16, 0.165, 0.170, 0.175\}. \tag{34}$$

We illustrate in Fig. 4[d], the evolution of $\mathcal{E}_{\text{posterior}}$ against the truncation parameter q, varying from 0 to 200. In fact, when $q = 200$, all methods are equivalent since we work with the full observation data. From Fig. 4[d], we observe that all the information-based strategies with different sampling densities are always more optimal compared to the observation-based method for $q \in (0, 200)$. We apply the stopping criteria as described in section 3.3, by calculating the eigenvalues of $\mathbf{HBH}^T$ for the medium sampling strategy. We obtain the optimal truncation parameter $q_{\text{optimal}} = 29$. The distribution of these eigenvalues are shown by the right vertical axes in Fig. 4[d](numerical log scale is represented by the right vertical axis in matching color). With 29 modes, the assimilation correction is achieved from 53.3% to 69.8%, compared to the background model equivalent $\mathcal{H}(\mathbf{x}_b)$ as shown in table 1, which is compatible to the results obtained in [8] when $L_{\mathbf{B}} > L_{\mathbf{R}}$. Among the three sampling strategies, the one of "IC medium" owns the lowest output error variances, close to the optimal information-based compression where the $\mathbf{HBH}^T$ is computed directly using $\mathbf{B}_{\mathrm{E},t}$. The latter, drawn with blue color in Fig. 4[d], stands for an optimal target for all information-based approaches since we suppose the exact background matrix is out of reach for data compression. As shown in this experiment, the choice of sampling strategy can significantly impact the compression optimality. If the samplings are too close, the residuals might not be uncorrelated, and if the samplings are too sparse, the flow independence of the $\mathbf{B}$ matrix could be threatened. We remind that the stopping criteria for the truncation parameter q varies for the different sampling strategies as shown in table 1. However, in this experiment the values of the optimal truncation parameters obtained do not qualitatively change the results as shown in Fig. 4[d].

In Fig. 4[a,c], we display the evolution of the exact background error variances (i.e. $\mathrm{Tr}(\mathbf{B}_{\mathrm{E},t})$) and error correlation (for fixed distances, $r = 1$ and $r = 2$) against time. The estimation of background error correlation in the 2D space, also based on $\mathbf{B}_{\mathrm{E},t}$, is calibrated using the same method shown in [18]. We observe that the error variances increase continuously for both u and v while the spatial error correlation tends to shrink, both being significantly time-variant between $t = 0s$ and $t = 1.4s$. In order to illustrate the non-linear and turbulent nature of error propagation, we show in Fig. 4[b] the error evolution $||\mathbf{x}_{b,t} - \mathbf{x}_{\mathrm{true},t}||_2$ of a single background trajectory. Obviously, these facts lead to problem of flow independent assumption for the *IC large* approach (between $0s$ and $2s$), conducing a less optimal compression strategy as shown in Fig. 4[d]. From this twin experiment, we notice the advantage of information-based compression by selecting the most impacting observation components. The optimal sampling strategy may strongly depend on the characteristics (e.g chaosity, stability) of the dynamical system.

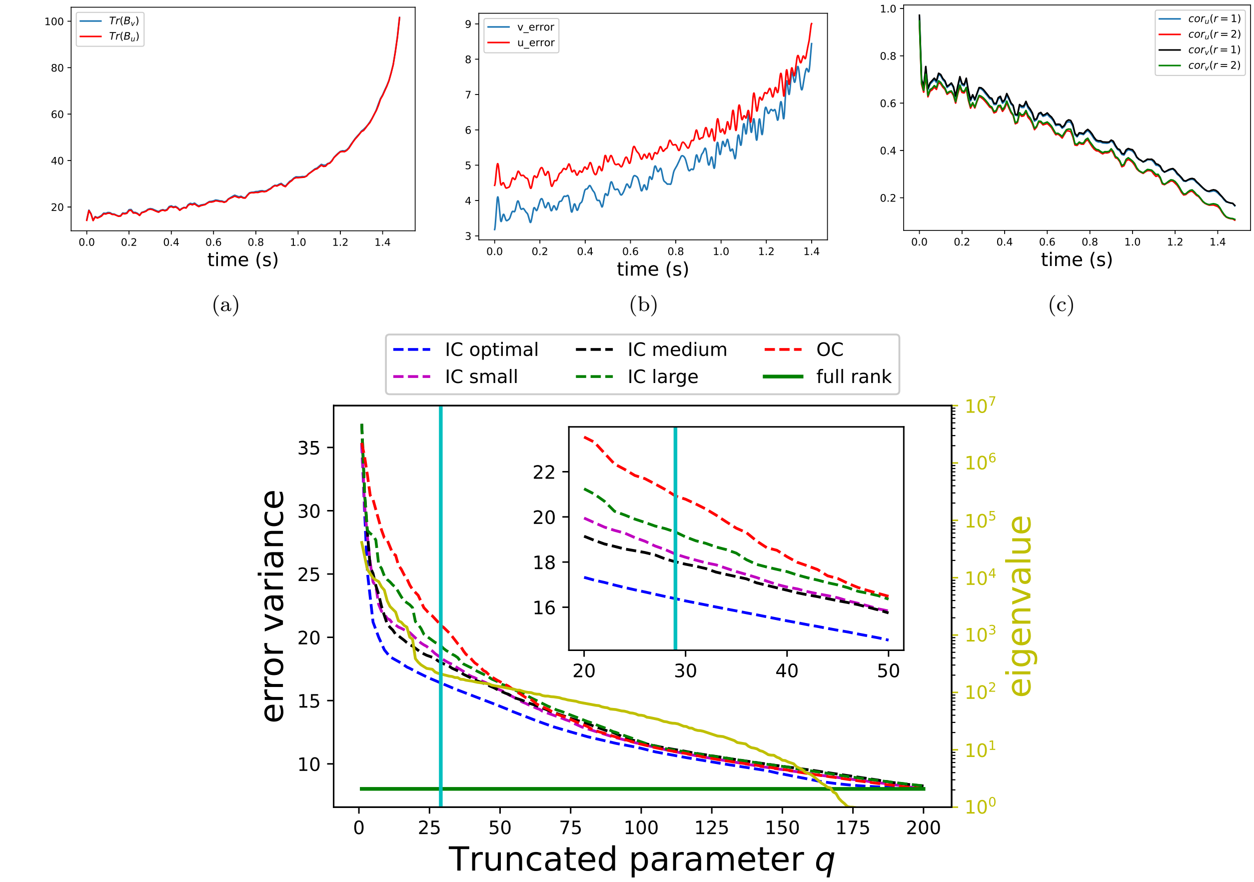
Until now, we have shown that, in the idealised case where the observation matrix is known *a priori* and the transformation operator is time-invariant, the information-based approach exhibits advantageous performance compared to the observation-based approach. However, as pointed out by [8], IC approach can be sensitive to prior errors of covariance estimation. In order to investigate the impact of a potential misknowledge of matrix $\mathbb{R}$, we present here two cases where the difference between the assumed/estimated matrix $\mathbf{R}_{\mathrm{A}}$ and the exact matrix $\mathbf{R}$ is voluntarily large. We explore two cases where the amplitude and the structure are misspecified, respectively:

- (a): $\mathbf{R}_{\mathrm{A}}$ has the same correlation structure as $\mathbf{R}$ with an homogeneous marginal error variance (i.e. $\mathbf{R}_{\mathrm{A},i,i} = 0.04$ which is different to $\mathbf{R}$ (cf. Fig. 3(a)).)
- (b): The correlation scale $L_{R_{\mathrm{A}}}$ is set to be 5 while $L_R = 1$ as explained in section 5.1 with same marginal error variances.

In both cases, the observation compression is implemented using $\mathbf{R}_{\mathrm{A}}$ while the observation matrix in the reduced space is set to be $\mathbf{R}_{\mathrm{A}}^{-1/2}\mathbf{R}\,\mathbf{R}_{\mathrm{A}}^{-1/2}$ instead of the identity matrix in Eq. 16 and Eq. 22. The performance of these compression methods is illustrated in Fig. 5, respectively for case (a) and (b). The optimal IC solutions (same as the blue lines in Fig. 4(d)) are drawn in dashed blue lines for comparison purposes. Both OC and IC approaches exhibit less optimal performance compared to Fig. 4. Furthermore, as shown in Fig. 5(a), the IC method can be more sensitive to the mis-specification of the $\mathbf{R}$ matrix amplitude, leading in this case to larger output error variances while IC behaves better than OC for misspecified R matrix correlation length.

	OC	IC large	IC medium	IC small	IC optimal
q_{optimal}	22	48	29	25	78
Correction for $q = 29$	53.3%	61.5%	65.7%	62.7%	69.8%

Table 1: The ratio of background minus analysis innovation ($||\mathcal{H}(\mathbf{x}_b) - \mathcal{H}(\mathbf{x}_a)||_2$) using compressed observation, relative to the one obtained with full observation



(a) (b) (c)

(d)

Figure 4: [a]: evolution of the exact background variance of u ($\mathrm{Tr}(\mathbf{B}_u)$) and v ($\mathrm{Tr}(\mathbf{B}_v)$) against time; [b]: evolution of $||\mathbf{x}_{b,t} - \mathbf{x}_{\text{true},t}||_2$ of a single background trajectory; [c]: evolution of average error correlation, of fixed distances ($r = 1$ and $r = 2$) in the 2D space; [d]: analysis error variance $\mathcal{E}_{\text{posterior}}$ (left y-axis) and eigenvalues of the estimated background error covariance in observation space ($\mathbf{HBH}^T$) (right y-axis) of the medium sampling strategy as a function of the truncation parameter for $t \in [0.16s, 0.18s]$. The vertical line represents the stopping criteria of $\sigma_q = \sqrt{\sigma_1}$;

6. Application to an operating hydrological model

6.1. DA modelling for flow reanalysis/prediction

The compression strategies introduced in previous sections are applied to a hydrological application using a precipitation-flow simulator MORDOR-TS developed by Électricté de France (EDF, the French electric utility company). This software is widely applied in operating hydraulic/horological problems, e.g. [39], [24], [40]. Based on information on spatially distributed physical parameters, such as precipitation or temperature, it provides a simulation of river flow relying on conceptual watersheds modeling. For more details about MORDOR-TS, interested readers are referred to [24] and [23]. MORDOR-TS is used as a non-linear state-observation transformation operator in data assimilation. We concentrate on a study area in the south of France, around the Tarn river where 9 streamflow gauges positioned at different mesh outlets are available. The Tarn river, being known for its extreme variability of water-level values and high sensitivity to precipitations [23], is an ideal benchmark for comparing different DA strategies. Located downstream, the Tarn river outlet at Millau (hereby denoted as TM) is of particular interest in the hydrological study. As an example, we show in Fig. 6 the simulated and daily observed Tarn river discharges at Millau, for 3 months in 1990 with the averaged precipitation over 28 spatially distributed regions (see [23]). Significant impacts of precipitation on the river flow of TM is observed with a delay of 2 to 5 days. The objective of this DA modeling is to improve the river flow prediction and reanalysis (history matching) by performing corrections on the daily precipitation in the 28 regions. Other physical quantities (e.g temperature) are

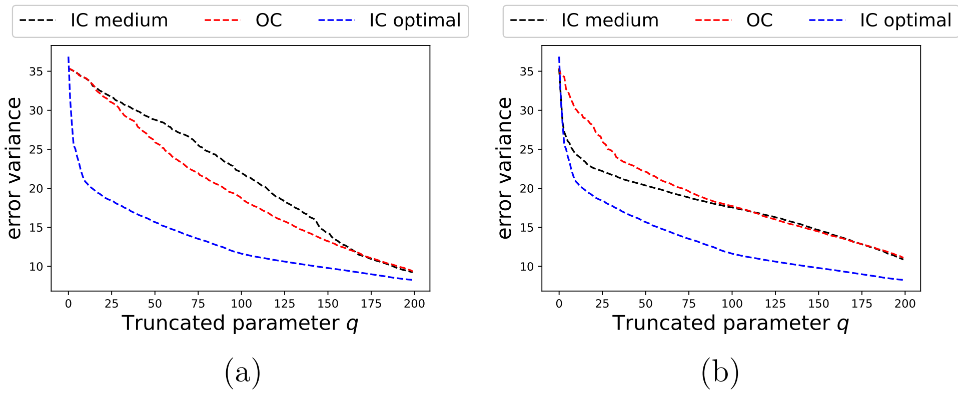


Figure 5: Evolution of error variance when the observation matrix is mis-specified.

considered as invariant parameters in this study. The variational assimilation is performed using the ADAO [41] package of SALOME platform, also developed by EDF.

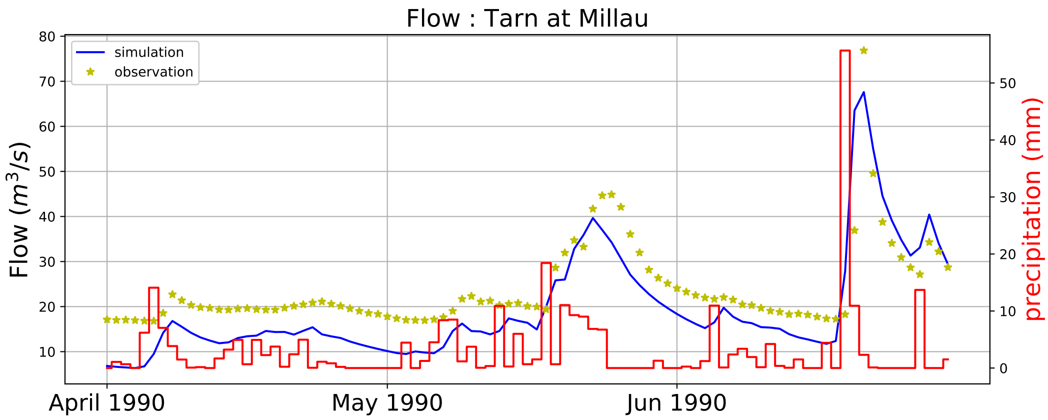


Figure 6: Example of simulation predicted by MORDOR-TS using daily precipitation, and observed Tarn discharges at Millau for three months in 1990. Simultaneous observed precipitations are in red bars (with the scale on the right vertical axis).

As mentioned in [23], performing DA correction on all precipitation inputs (i.e 28 regions) can probably introduce an over-parameterization and thus induces an overfitting, with a high risk to deteriorate flow forecasts. Therefore, we make the choice to proceed with uniform additional increments ξ^p_t for all 28 regions, depending only on time t. Incremental variables $\xi^{r,j}(j = 1..8)$ on the eight parameters which determine the initial (at $t = 0$) reservoir level is also added in the state space to adjust the river flow at the beginning of each assimilation window. These windows are fixed of 30 days, leading to an observation vector of dimension 270 with 9 gauges. Temporal correlation is considered for both background and observation errors. The DA modelling is summarized in Table 2 and a more detailed description can be found in [23] and [40]. The main objective of this application stands for improving short-range flow forecasting by correcting historical precipitation. Since the impact of the precipitation on the river flow is only significant within 3 to 4 days (see [23] for details), we fix the prediction window to 3 days in this study.

DA modelling	state: $\mathbf{x}$	$dim(\mathbf{x})$	Observations: $\mathbf{y}$	$dim(\mathbf{y})$	invariant parameters
Incremental 3DVar	ξ^p_t $\xi^{r,j}$	38	river flow $Q_{q,t}$	270	temperature, etc

Table 2: Details of DA modelling where $t = 0..29$ is the time (days) relative to the beginning of the assimilation window; $q = 1..9$ represents the 9 gauges where $\xi^p_t, \xi^{r,j}$ represent respectively the increments of daily precipitation and initial reservoir level.

6.2. Observation compression

Despite that MORDOR-TS is computationally efficient (it may take only a few CPU seconds to simulate a spatially distributed flow simulation of several years), the application of variational assimilation algorithms could be expensive, due to the non-linearity of the transformation operator. As shown in Table. 2, in this DA modeling, the dimension of the observation vector is much larger compared to the state dimension, promoting the utilization of observation data compression. We then implement DA algorithms in the hydrological model with compressed data using either OC or IC approaches. To make the compression strategies more general, in both cases the principal components are constructed using the daily observed flow data from 1990 to 2000 in the 9 gauges. The objective of this study is to make an efficient use of the observation vector $\mathbf{y}$ with an optimal number of modes selected, which we expect to be much smaller than the full observation dimension ($\dim(\mathbf{y}) = 270$).

A major hurdle of this application is that the *a priori* knowledge of both $\mathbf{B}$ and $\mathbf{R}$ is very limited. As a remedy, we start as described in [23], by considering the background covariance matrix $\mathbf{B}$ of Balgovind-type since we wish to model the existence of temporal correlation in the precipitation data. Moreover, the initial $\mathbf{R}$ matrix is set to be diagonal. The DI01 algorithm [20] is then applied several times to to come up with a reasonable approximation of the ratio between $\mathrm{Tr}(\mathbf{B})$ and $\mathrm{Tr}(\mathbf{R})$ at the first stage. In a second stage, we then perform the estimation of $\mathbf{HBH}^T$ and $\mathbf{R}$, relying on Desroziers formulation (respectively Eq. 26 and 28) using 3400 assimilation windows of 30 days from 1990 to 2000. By then, post-processing is required to ensure the symmetric positive definiteness (SPD) of the $\mathbf{R}$ matrix. More precisely,

$$\mathbf{R} \longleftarrow \frac{1}{2}(1-\mu)(\mathbf{R}+\mathbf{R}^T) + \mu\mathbf{C}, \tag{35}$$

where $\mu = 0.1$ and $\mathbf{C} = \mathrm{Tr}(\mathbf{R}) \times \mathbf{I}$. The Desroziers method is iterated twice, using the same data set, to ensure the stability of the estimated matrices. The algorithm outputs produced after the first and the second iterations are very similar as shown in [23]. We emphasize that the estimated $\mathbf{R}$ matrix is not only used for the observation compression but also in the DA algorithm in the full observation space. The $\mathbf{HBH}^T$ matrix is obtained through Eq. 27, once the $\mathbf{R}$ matrix is specified. As a remark, even if the system considered here is not very large, the computational burden associated with the data assimilation of this nonlinear system (for which prior information is degraded) remains important because of a multi-stage tuning approach which combined several offline and online covariance tuning algorithms can be implemented to improve the reanalysis and the forecasting accuracy of this hydrological application. However, these methods are computationally expensive, especially when iterations are needed (e.g. [18]). With advanced data compression methods, the computational burden can be released, allowing more precise covariance tuning to improve the DA performance.

6.3. DA with compressed data

6.3.1. Averaged performance

Extracting the principal components $\tilde{\mathbf{L}}$ and $\hat{\mathbf{L}}$, respectively based on estimated $\mathbf{HBH}^T$ and $\mathbf{R}$, we then apply the compression methodology described in sect.3. The objective is to compare the assimilation output $\mathbf{x}_{a,\text{compression}}$ and $\mathbf{x}_{a,\text{full}}$, obtained using either the compressed observation $\hat{\mathbf{y}}_q$, $\tilde{\mathbf{y}}_q$ or the full observation vector $\mathbf{y}$. More precisely, we are interested in the observation minus analysis (O-A) innovation quantity for both flow reanalysis and forecast. Varying the truncated parameter q, DA processes are performed respectively with $(\mathbf{x}_b, \tilde{\mathbf{y}}_q, \mathbf{B}, \tilde{\mathbf{R}}_q, \tilde{\mathcal{H}}_q)$ and $(\mathbf{x}_b, \hat{\mathbf{y}}_q, \mathbf{B}, \hat{\mathbf{R}}_q, \hat{\mathcal{H}}_q)$ for 12 assimilation windows in 1993, each of 30 days starting at the first day of every month. We draw the averaged compressed/full O-A innovation ratio ∇, defined as

$$\nabla = \frac{||\mathbf{y} - \mathcal{H}(\mathbf{x}_{a,\text{compression}})||_2}{||\mathbf{y} - \mathcal{H}(\mathbf{x}_{a,\text{full}})||_2}, \tag{36}$$

in Fig. 7 for both reanalysis[a] and prediction[b] at TM. More particularly, $\nabla = 100\%$ means the reanalysis/prediction accuracy of the current solution is equivalent to the one obtained with the full observation vector.

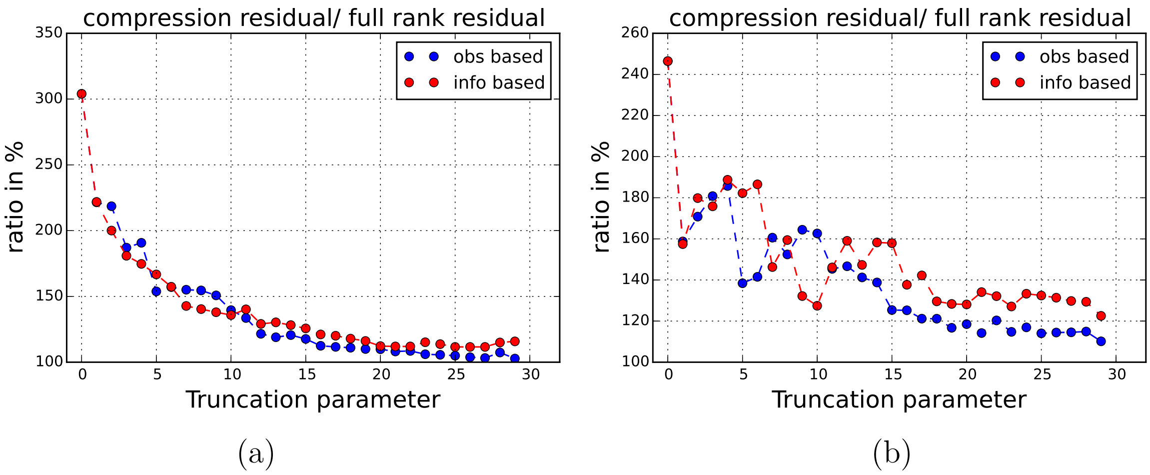


Figure 7: Evolution of ∇, averaged using 12 assimilation windows, against the number of truncation parameter q for reanalysis[a] and prediction[b] at TM.

We observe from Fig. 7 that the performance of these two approaches is similar to the reanalysis while the observation-based method is slightly more optimal on average for flow forecasting. The evolution of the reconstruction error (Fig. 7[a]) is much smoother, compared to the prediction error ((Fig. 7[b])), both against the truncation parameter. In fact, the reconstruction error is estimated using assimilation windows of 30 days while prediction windows are solely of 3 days. Therefore, the estimation of the prediction ratio has significantly more sampling noise. Furthermore, since both $\mathbf{B}$ and $\mathbf{R}$ are not well specified *a priori*, extra noise can be introduced while estimating the information entropy. For both methods, the assimilation results obtained using 15 to 20 modes (around 5% to 7.5% of total observation dimension) are close to the full rank solution in terms of both reanalysis and prediction. Without deteriorating the assimilation result, these compression strategies make the DA algorithm certainly more efficient, allowing more optimization iterations if needed.

6.3.2. Performance in each DA window

We draw the reconstructed river flow (i.e $\mathcal{H}(\mathbf{x}_a)$) at TM of each of those 12 assimilation windows, for both corrections with compressed and full observation data in Fig. 8 where the yellow stars represent the daily observations. Based on the method described in Eq. 25, the optimal truncation parameter reads $q_{\text{optima l}} = 22$. Here, we display the results when $q = 10$ in order to voluntary emphasize the difference between the two approaches as shown in Fig 7. A vertical line in each graph separates the reanalysis (left) and the prediction (right). We notice that the reconstructed curves issued from OC (blue) and IC (red) are similar in most cases, both being adequately close to the full rank assimilation (green), compared to the original simulation. Some exceptions can be found, for example, in the assimilation window of December 1993 where the prediction is covered by a flood period. It seems that the information-based approach provides a better performance, especially for flow forecasting at that moment. In general, as demonstrated in [23], meteorological factors can impact the assimilation precision significantly. DA algorithms often perform better during drought periods (see Jun, July, August in Fig. 8) where the prior observation minus background (O-B) innovations are more consistent (i.e always being under-estimated or over-estimated). Contrarily, in flood periods where O-B innovations are usually more turbulent, more careful attention might be taken when performing compression methods.

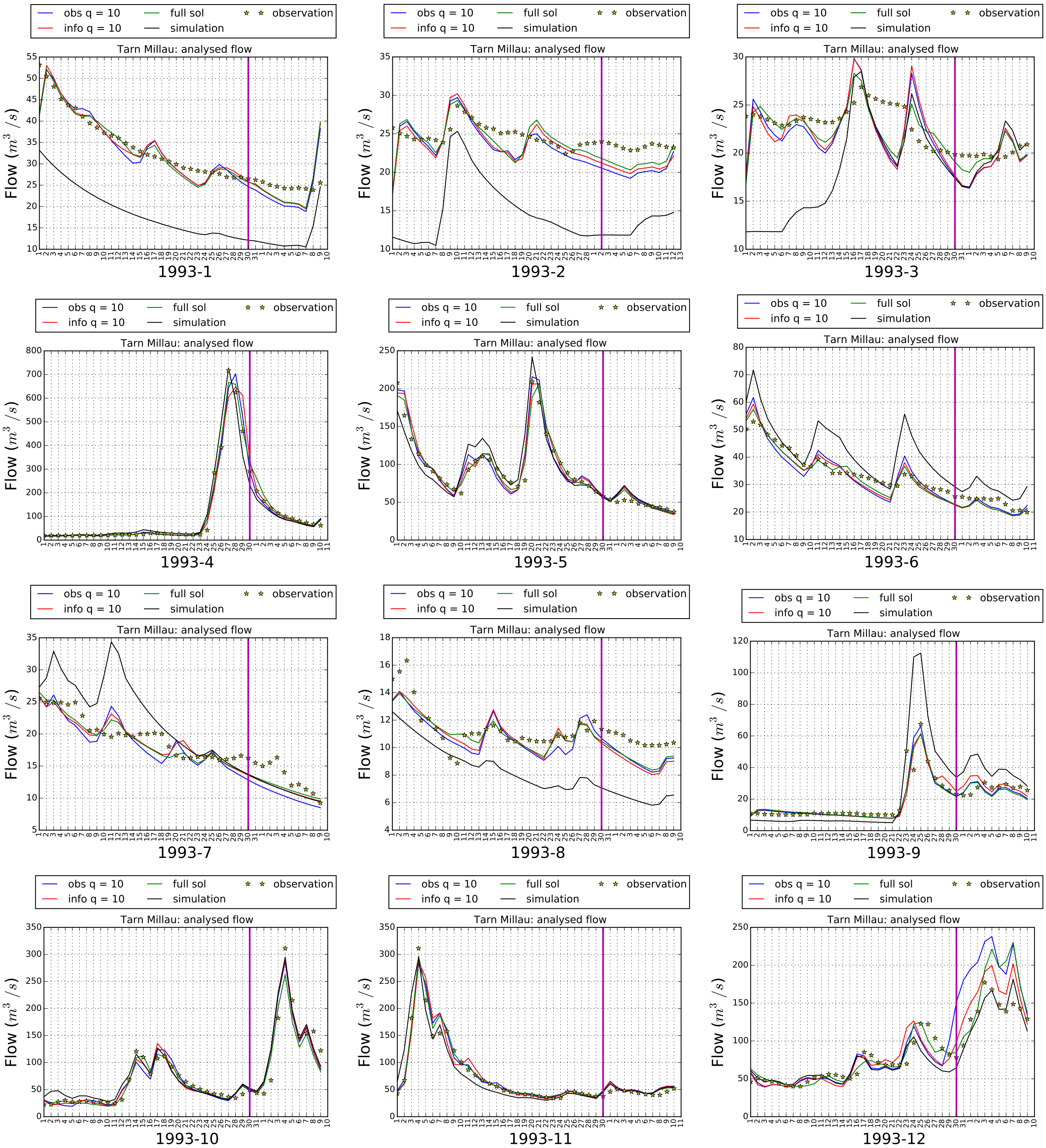


Figure 8: The reconstructed and predicted river flow at TM for OC (obs) ($q = 10$), IC (info) ($q = 10$) and full rank DA solutions of different months in 1993. The left side of the vertical line represents the flow reanalysis while the right side represents the prediction

7. Discussion

Sequential data assimilation algorithms can be computationally challenging, especially for large scale systems such as NWP, remote sensing, or geophysical problems. Data compression techniques commonly used in DA problems have recently received increasing interest in reducing the computational burden. Much effort has been devoted to improving the algorithm efficiency without diminishing the accuracy of assimilation reconstruction and forecasting. Classical compression approaches consist of either extracting the principal vectors of observation dynamics or identifying the directions that contribute the most to the prior-posterior information gap. For both methods, the lack of precise knowledge on prior error covariances stands for an essential obstacle, as mentioned in several previous studies. Furthermore, the limited number of background/observation trajectories often entails a poor empirical estimation. In this paper, we have introduced a concept of observation compression benefiting from existing piecewise covariance estimation, establishing a natural connection between the posterior error covariance diagnosis and data compression techniques. More precisely, we assume that the error covariances (both $\mathbf{B}$ and $\mathbf{R}$) are flow-independent over some specific time periods, which allows an estimation based on time-variant residuals. Therefore, a much smaller number of background/observation trajectories are required for non-parametric covariance estimation. Different estimation formulations are possible depending on the prior knowledge of the $\mathbf{R}$ matrix. The choice of flow-independent windows, as well as the residual sampling densities, is essential in these approaches, especially for the $\mathbf{HBH}^T$ estimation. When the samplings are either too dense or too sparse, the assumptions of covariance estimation approaches might be unsatisfied, leading to a less optimal observation compression. These aspects are numerically analyzed in the twin experiments of a 2D shallow water model with non-linear dynamics with the perfect knowledge of the $\mathbf{R}$. Numerical results show a significant advantage of the information-based compression in terms of assimilation accuracy, compared to the observation-based one. As for the industrial hydrological model, posterior covariance estimation which requires the knowledge of the analyzed states $\mathbf{x}_a$, is needed since the $\mathbf{R}$ matrix is not known *a priori*. In this application, both the OC and IC compression methods rely on the flow-independent estimation of $\mathbf{R}$, showing competitive performance regarding the flow reanalysis and the forecasting accuracy. A meteorological effect is also briefly discussed in this hydrological application, which indicates that different numbers of modes should be chosen in different periods of the year regarding the hydrological properties. Future work can be considered to improve the algorithm efficiency and flexibility under industrial conditions, for example, by using parametric covariance tuning methods or spatial localization techniques. Another important limitation of the current approach stands for the time invariance of the observations error covariance on some time-scale, limiting for instance the use of moving observation sensors. Indeed, if observation positions change, both the observation matrix $\mathbf{R}$ and the transformation operator $\mathcal{H}$ can not be considered flow-independent, leading to difficulties when applying Desroziers-type methods. Future work can be considered to use interpolation approaches to construct a global observation set which includes all time-variant observation positions. Another perspective of this study could be to further examine the optimal choice of the sampling density while estimating the error covariances, for example, with the help of uncertainty quantification methods for dynamical systems.

Acknowledgement

The authors would like to thank Dr. Bertrand Iooss and Dr. Angélique Ponçot for fruitful discussions about the compression methodology and the hydrological application. This work was supported by EDF R&D. This research was partially funded by the Leverhulme Centre for Wildfires, Environment and Society through the Leverhulme Trust, grant number RC-2018-023. The authors are grateful to an anonymous reviewer for the useful remarks on the manuscript.

Bibliography

[1] F. Rabier, Overview of global data assimilation developments in numerical weather-prediction centres, Quarterly Journal of the Royal Meteorological Society 131 (2005) 3215–3233.

[2] M. C. Rochoux, S. Ricci, D. Lucor, B. Cuenot, A. Trouvé, Towards predictive data-driven simulations of wildfire spread–part i: Reduced-cost ensemble kalman filter based on a polynomial chaos surrogate model for parameter estimation, Natural Hazards and Earth System Sciences 14 (2014) 2951–2973.

[3] H. Gong, Y. Yu, Q. Li, C. Quan, An inverse-distance-based fitting term for 3D-Var data assimilation in nuclear core simulation, Annals of Nuclear Energy 141 (2020) 107346.

[4] S. Cheng, R. Arcucci, C. C. Pain, Y.-K. Guo, Optimal vaccination strategies for covid-19 based on dynamical social networks with real-time updating, arXiv preprint arXiv:2103.00485 (2021).

[5] D. Lucor, O. P. Le Maître, Cardiovascular modeling with adapted parametric inference, ESAIM: ProcS 62 (2018) 91–107.

[6] P. Nadler, R. Arcucci, Y. Guo, Data assimilation for parameter estimation in economic modelling, in: 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS) 2019, 2019, pp. 649–656.

[7] A. D. Collard, A. P. McNally, F. I. Hilton, S. B. Healy, N. C. Atkinson, The use of principal component analysis for the assimilation of high-resolution infrared sounder observations for numerical weather prediction, Quarterly Journal of the Royal Meteorological Society 136 (2010) 2038–2050.

[8] A. Fowler, Data compression in the presence of observational error correlations, Tellus A: Dynamic Meteorology and Oceanography 71 (2019) 1634937.

[9] A. Carrassi, M. Bocquet, L. Bertino, G. Evensen, Data assimilation in the geosciences: An overview of methods, issues, and perspectives, Wiley Interdisciplinary Reviews: Climate Change 9 (2018) e535.

[10] S. Cheng, J.-P. Argaud, B. Iooss, A. Ponçot, D. Lucor, A graph clustering approach to localization for adaptive covariance tuning in data assimilation based on state-observation mapping, accepted for publication in *Mathematical Geosciences*, 2021. arXiv:2001.11860.

[11] A. Cioaca, A. Sandu, Low-rank approximations for computing observation impact in 4D-Var data assimilation, Computers & Mathematics with Applications 67 (2014) 2112 – 2126.

[12] E. D. Nino-Ruiz, A. Sandu, X. Deng, A parallel implementation of the ensemble kalman filter based on modified cholesky decomposition, Journal of Computational Science 36 (2019) 100654.

[13] I. Hoteit, D.-T. Pham, J. Blum, A simplified reduced order Kalman filtering and application to altimetric data assimilation in tropical pacific, Journal of Marine Systems 36 (2002).

[14] J.-P. Argaud, B. Bouriquet, F. Caso, H. Gong, Y. Maday, O. Mula, Sensor placement in nuclear reactors based on the generalized empirical interpolation method, Journal of Computational Physics 363 (2018) 354 – 370.

[15] J. A. Waller, S. L. Dance, N. K. Nichols, On diagnosing observation-error statistics with local ensemble data assimilation, Quarterly Journal of the Royal Meteorological Society 143 (2017) 2677–2686.

[16] M. Matricardi, A. P. McNally, The direct assimilation of principal components of IASI spectra in the ECMWF 4D-Var, Quarterly Journal of the Royal Meteorological Society 140 (2014) 573–582.

[17] M. Fisher, Background error covariance modelling, in: Seminar on Recent developments in data assimilation for atmosphere and ocean (Shinfield Park, Reading, 8-12 September), ECMWF, 2003.

[18] S. Cheng, J.-P. Argaud, B. Iooss, D. Lucor, A. Ponçot, Background error covariance iterative updating with invariant observation measures for data assimilation, Stochastic Environmental Research and Risk Assessment 33 (2019) 2033–2051.

[19] D. F. Parrish, J. C. Derber, The National Meteorological Center's spectral statistical-interpolation analysis system, Monthly Weather Review 120 (1992) 1747–1763.

[20] G. Desroziers, S. Ivanov, Diagnosis and adaptive tuning of observation-error parameters in a variational assimilation, Quarterly Journal of the Royal Meteorological Society 127 (2001) 1433 – 1452.

[21] G. Desroziers, L. Berre, B. Chapnik, P. Poli, Diagnosis of observation, background and analysis-error statistics in observation space, Quarterly Journal of the Royal Meteorological Society 131 (2005) 3385 – 3396.

[22] K. Bathmann, Justification for estimating observation-error covariances with the Desroziers diagnostic, Quarterly Journal of the Royal Meteorological Society 144 (2018) 1965–1974.

[23] S. Cheng, J.-P. Argaud, B. Iooss, D. Lucor, A. Ponçot, Error covariance tuning in variational data assimilation: application to an operating hydrological model, Stochastic Environmental Research and Risk Assessment 35 (2021) 1019–1038.

[24] L. Rouhier, M. Le Lay, F. Garavaglia, N. Moine, F. Hendrickx, C. Monteil, P. Ribstein, Impact of mesoscale spatial variability of climatic inputs and parameters on the hydrological response, Journal of Hydrology 553 (2017) 13 – 25.

[25] W. Fulton, Eigenvalues, invariant factors, highest weights, and schubert calculus, Bulletin of The American Mathematical Society 37 (2000) 209–250.

[26] F. Uboldi, M. Kamachi, Time-space weak-constraint data assimilation for nonlinear models, Tellus A 52 (2000) 412–421.

[27] J. Brajard, A. Carrassi, M. Bocquet, L. Bertino, Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the Lorenz 96 model, Journal of Computational Science 44 (2020) 101171.

[28] P. Antonelli, H. E. Revercomb, L. A. Sromovsky, W. L. Smith, R. O. Knuteson, D. C. Tobin, R. K. Garcia, H. B. Howell, H.-L. Huang, F. A. Best, A principal component noise filter for high spectral resolution infrared measurements, Journal of Geophysical Research: Atmospheres 109 (2004).

[29] D. Tobin, P. Antonelli, H. Revercomb, S. Dutcher, D. Turner, J. Taylor, R. Knuteson, K. Vinson, Hyperspectral data noise characterization using principle component analysis: Application to the atmospheric infrared sounder, Journal of Applied Remote Sensing 1 (2006) 013515.

[30] C. Cardinali, S. Pezzulli, E. Andersson, Influence-matrix diagnostic of a data assimilation system, Quarterly Journal of the Royal Meteorological Society 130 (2004) 2767–2786.

[31] S. Migliorini, Information-based data selection for ensemble data assimilation, Quarterly Journal of the Royal Meteorological Society 139 (2013) 2033–2054.

[32] R. Cangelosi, A. Goriely, Component retention in principal component analysis with application to cdna microarray data, Biology Direct 2 (2007) 2.

[33] R. Arcucci, L. Mottet, C. Pain, Y.-K. Guo, Optimal reduced space for variational data assimilation, Journal of Computational Physics 379 (2018) 51–69.

[34] C. Lanczos, An iteration method for the solution of the eigenvalue problem of linear differential and integral operators, Journal of research of the National Bureau of Standards 45 (1950) 255–282.

[35] P. Tandeo, P. Ailliot, M. Bocquet, A. Carrassi, T. Miyoshi, M. Pulido, Y. Zhen, A review of innovation-based methods to jointly estimate model and observation error covariance matrices in ensemble data assimilation, Monthly Weather Review (2020) 1–68.

[36] A. Hollingsworth, P. Lönnberg, The verification of objective analyses: Diagnostics of analysis system performance, Meteorology and Atmospheric Physics 40 (1989) 3–27.

[37] J. A. Waller, S. L. Dance, N. K. Nichols, Theoretical insight into diagnosing observation error correlations using observation-minus-background and observation-minus-analysis statistics, Quarterly Journal of the Royal Meteorological Society 142 (2016) 418–431.

[38] B. Chapnik, G. Desroziers, F. Rabier, O. Talagrand, Property and first application of an error-statistics tuning method in variational assimilation, Quarterly Journal of the Royal Meteorological Society 130 (2004) 2253 – 2275.

[39] R. Garçon, Prévision opérationnelle des apports de la Durance à Serre-Ponçon à l'aide du modèle MORDOR. Bilan de l'année 1994-1995, La Houille Blanche (1996) 71–76.

[40] S. Cheng, Error covariance specification and localization in data assimilation with industrial application, Ph.D. thesis, Paris-Saclay Unversity, France, 2020.

[41] J.-P. Argaud, User documentation, in the SALOME 9.3 platform, of the ADAO module for "Data Assimilation and Optimization", Technical report 6125-1106-2019-01935-EN, EDF / R&D, 2019.