



# On gauss-verifiability of optimal solutions in variational data assimilation problems with nonlinear dynamics



I.Yu. Gejadze<sup>a,\*</sup>, V. Shutyaev<sup>b</sup>

<sup>a</sup> UMR G-EAU, IRSTEA-Montpellier, 361 Rue J.F. Breton, BP 5095, 34196, Montpellier, France

<sup>b</sup> Institute of Numerical Mathematics, Russian Academy of Sciences, Moscow Institute for Physics and Technology, 119333 Gubkina 8, Moscow, Russia

## ARTICLE INFO

### Article history:

Received 12 June 2014

Received in revised form 24 September 2014

Accepted 26 September 2014

Available online 2 October 2014

### Keywords:

Large-scale geophysical flow model

Nonlinear dynamics

Data assimilation

Optimal control

Identifiability

Confidence region

Analysis error covariance

Non-gaussianity

## ABSTRACT

The problem of variational data assimilation for a nonlinear evolution model is formulated as an optimal control problem to find the initial condition. The optimal solution (analysis) error arises due to the errors in the input data (background and observation errors). Under the gaussian assumption the confidence region for the optimal solution error can be constructed using the analysis error covariance. Due to nonlinearity of the model equations the analysis pdf deviates from the gaussian. To a certain extent the gaussian confidence region built on a basis of a non-gaussian analysis pdf remains useful. In this case we say that the optimal solution is “gauss-verifiable”. When the deviation from the gaussian further extends, the optimal solutions may still be partially (locally) gauss-verifiable. The aim of this paper is to develop a diagnostics to check gauss-verifiability of the optimal solution. We introduce a relevant measure and propose a method for computing decomposition of this measure into the sum of components associated to the corresponding elements of the control vector. This approach has the potential for implementation in realistic high-dimensional cases. Numerical experiments for the 1D Burgers equation illustrate and justify the presented theory

© 2014 Elsevier Inc. All rights reserved.

## 1. Introduction

For the distributed parameter systems the notions of controllability, observability and identifiability are of significant importance [38]. The idea of controllability was introduced by Kalman in 1959 [22], and later it was realized that controllability plays a fundamental role in the study of control problems of all types (see Fattorini [8], Kalman et al. [23], Lions [15, 34, 35], Triggiani [45], and others).

Consider the mathematical model of a physical process that is described by a system of time dependent partial differential equations, which contains control functions, denoted by  $v$ . We shall denote by  $\varphi(v)$  the solution of the system for a given control  $v$ . Let us suppose that we have also some partial observations  $y_{obs}$  of the system variables, and we want to find a control  $v$  such that  $C\varphi(v) = y_{obs}$ , where  $C$  is the observation operator mapping the “state space” into the “observation space”. The system is said to be *controllable* (or *exactly controllable*) [34] if for any  $y_{obs}$  there exists  $v$  from the “space of controls” such that  $C\varphi(v) = y_{obs}$ . In other words, if  $v$  runs over all the “space of controls”, then  $C\varphi(v)$  covers

\* Corresponding author. Tel.: +33 467166408; fax: +33 467166440.

E-mail addresses: igor.gejadze@irstea.fr (I.Yu. Gejadze), shutyaev@inm.ras.ru (V. Shutyaev).

all the observation space. The system is *approximately controllable* if  $C\varphi(v)$  belongs to a small neighborhood of  $y_{obs}$ . This definition is an analog of the well-known definition for the final observation  $C\varphi(v) = \varphi|_{t=T}$ .

One can also study controllability in the space dual to the state space. Intuitively, it corresponds to the concept of observability (Kalman [22,23], Markus [37]). A linear evolution system is *observable* if  $v = 0$ ,  $C\varphi(0) = 0$  implies that the initial state equals zero. In other words, with  $v = 0$  for two different initial conditions  $u_1, u_2$ , the values  $C\varphi_1, C\varphi_2$  are different. This means that the initial state can be uniquely determined from the observations [7].

Observability is of central importance in the study of prediction and filtering problems [24]. For linear problems the results on controllability also apply often to observability by simply replacing vectors and linear transformations by their duals [23,37]. A necessary and sufficient conditions for observability were studied by Krasovski [26]. Later this notion was generalized by Kurzanski [30] (see also Kurzanski and Khapalov [31]) who introduced the so-called “informational domain” of initial states of the system consistent with the measurement data and characterized observability as the boundedness of the informational domain [31].

In the theory of optimal control of linear dynamical systems with a quadratic cost function controllability is needed to prove the existence. Uniqueness is often given by observability. To prove stability, one utilizes both controllability and observability, but often it is not enough. The notion of identifiability comes to help for nonlinear systems. A control  $v$  is said to be *identifiable* for the observation operator  $C$  if the mapping  $G$  (control  $\rightarrow$  observation) is injective, i.e.  $G(v) = C\varphi(v)$  has a unique inverse. In other words,  $C\varphi(v_1) \neq C\varphi(v_2)$  if  $v_1 \neq v_2$  (see Chavent [4], Kitamura and Nakagiri [25], Goodson and Polis [16], Kubrusly [28], Kravaris and Seinfeld [27]). For linear systems identifiability is equivalent to observability if the control  $v$  is the initial condition itself. Once the identifiability has been established, it is important to assess whether the control estimate, also referred to as *optimal solution*, is stable with respect to the perturbations of the data.

These three conditions (existence, uniqueness and stability) are known as Hadamard conditions of well-posedness. In the classical optimal control theory an optimal solution  $v$  is said to be stable if the inverse operator  $G^{-1}$  (observation  $\rightarrow$  control) is continuous [27]. This condition is usually presented in the form  $\|v\| \leq \epsilon \|y_{obs}\|$ , which involves a constant  $\epsilon$  dependent on the properties of the operator  $G^{-1}$ . Only extremely crude estimates of  $\epsilon$  are usually available in practice.

Variational Data Assimilation (DA) is a deterministic approach based on the optimal control theory, suitable for high-dimensional large-scale models arising in geophysical applications [32]. In particular, the method called “4D-Var” is the preferred method implemented at some major operational weather and ocean forecasting centers [6]. The main aim of variational DA is to find the optimal solution  $v$  (usually the initial condition), which is an approximation to the true initial state, consistent with observations. In the DA community this optimal solution is called *analysis*. The cost function in variational DA includes the background term (the regularization term), the presence of which usually guarantees that all of three Hadamard conditions are formally met. From this fact, however, very little can be concluded on how close to the truth the optimal solution actually is. That is why variational DA is often considered in a probabilistic context (including the Bayesian context, see e.g. [36,42,43]), where the confidence region for the optimal solution can be constructed on a basis of the analysis error covariance [9]. This implies that the analysis error probability density function (pdf) is reasonably close to the gaussian. We shall say that the optimal solution is *gauss-verifiable* if a reliable covariance-based confidence region for the optimal solution error can be actually constructed.

The fundamental difficulty here is related to the nonlinearity of the model equations (and of the observation operator) [3,12]. The nonlinear least squares estimator is *asymptotically normal* [1,20], however for a finite number of observations this is not the case. On one hand, the nonlinearity may distort its gaussian properties to the extent when the covariance becomes no longer useful for constructing the confidence region (even for the gaussian data errors). On the other hand, this distortion may happen to be localized in certain spatial areas (since it is related to the nonlinearity), whereas outside of these areas the estimator holds its gaussian properties. Assuming the optimal solution is gauss-verifiable, estimating of the analysis error covariance is not an easy task [9,11] in practical terms. The major difficulty can be attributed to the high-dimensionality of the state vector combined with the complexity of the governing equations. This results into extremely high computational costs of a single optimal solution, whereas for computing the sample covariance one needs an ensemble of optimal solutions.

Formally speaking, gauss-verifiability is tied to the approximate gaussianity (normality) of the estimator. An immediate idea would be to use classical test statistics for multivariate normality [18] or the Kullback–Leibler divergence [29] in the form of ‘negentropy’. However, there are a few points why these may not be the best option. First, neither the classical test statistics nor negentropy measure the gauss-verifiability (as we understand it) directly, so it is difficult to make a sensible interpretation of results. Secondly, they measure local properties of the nonlinear estimator, which strongly depend on the point of evaluation, whereas we would rather prefer to know its global properties. Thirdly, the sample (ensemble) on a basis of which these statistics or negentropy are calculated is likely to be extremely small as compared to the size of the control vector. At the same time almost any invariant test statistic is a function of the Mahalanobis distances and angles, which involve the inverse square root of the sample covariance matrix (see e.g. [18]). The difficulty of evaluating the inverse square root of a matrix of a deficient rank is well known.

The aim of this paper is to develop a tool for checking the gauss-verifiability, both total and partial (local). We consider the analysis pdf defined on the “true” state and its approximation defined on the optimal solution. The basic idea is to quantify our ability to recognize the truth among statistically significant events associated to the analysis pdf, defined by the analysis (the mode), and by the analysis error covariance. First, we introduce the *coexistence principle*. Then, in order to quantify its violation (further referred as *coexistence breach*) we define the *coexistence measure* (CM). The decomposition of

the CM into the sum of components, each being associated to the corresponding element of the control vector, is introduced. The distribution of these components in space shows the subsets of the control vector for which the gaussian confidence regions cannot be properly defined. Numerical experiments for the 1D Burgers equation illustrate the developed theory and demonstrate the usefulness of the suggested measure and, especially, of its element-wise decomposition.

The paper is organized as follows. In Section 2, we provide the statement of the variational DA problem to identify the initial condition for a nonlinear evolution model. In Section 3, we introduce the CM and in Section 4 – its element-wise decomposition. A set of approximations accepted to achieve computational feasibility is described in Section 5, the computational approach is presented in Section 6. The confidence intervals for the decomposition components and for the CM itself are introduced in Section 7. Numerical implementation for the nonlinear 1D Burgers equation is briefly described in Section 9. Review of numerical experiments and discussion of numerical results are presented in Section 10. Main findings of this paper are summarized in Conclusions.

## 2. Statement of the problem

Consider the mathematical model of a physical process that is described by the evolution problem:

$$\begin{cases} \frac{\partial \varphi}{\partial t} = F(\varphi) + f, & t \in (0, T) \\ \varphi|_{t=0} = u, \end{cases} \quad (2.1)$$

where  $\varphi = \varphi(t)$  is the unknown function belonging for any  $t \in (0, T)$  to a state space  $X$ ,  $u \in X$ ,  $F$  is a nonlinear operator mapping  $X$  into  $X$ . Let  $Y = L_2(0, T; X)$  be a space of functions  $\varphi(t)$  with values in  $X$ ,  $\|\cdot\|_Y = (\cdot, \cdot)_Y^{1/2}$ ,  $f \in Y$ . Suppose that for a given  $u \in X$ ,  $f \in Y$  there exists a unique solution  $\varphi \in Y$  to (2.1). Further we accept the ‘perfect model’ assumption, i.e.  $f$  is known without error.

Let  $u^t$  be the “true” initial state and  $\varphi^t$  – the solution to the problem (2.1) with  $u = u^t$ , i.e. the “true” state evolution. We define the input data as follows: the background function  $u_b \in X$ ,  $u_b = u^t + \xi_b$  and the observations  $y \in Y_o$ ,  $y = C(\varphi^t) + \xi_o$ , where  $C: Y \rightarrow Y_o$  is a bounded operator (observation operator) and  $Y_o$  is an observation space. The functions  $\xi_b \in X$  and  $\xi_o \in Y_o$  may be regarded as the background and the observation error, respectively. We assume that these errors are normally distributed (Gaussian) with zero mean and the covariance operators  $V_b \cdot = E[(\cdot, \xi_b)_X \xi_b]$  and  $V_o \cdot = E[(\cdot, \xi_o)_{Y_o} \xi_o]$ , i.e.  $\xi_b \sim \mathcal{N}(0, V_b)$ ,  $\xi_o \sim \mathcal{N}(0, V_o)$ , where “ $\sim$ ” is read “is distributed as”. We also assume that  $\xi_o$ ,  $\xi_b$  are mutually uncorrelated and  $V_b$ ,  $V_o$  are positive definite, hence invertible.

Let us formulate the following DA problem (optimal control problem) with the aim to identify the initial condition: for given  $f \in Y$  find  $u \in X$  and  $\varphi \in Y$  such that they satisfy (2.1), and on the set of solutions to (2.1), a cost functional  $J$  takes the minimum value, i.e.

$$J(u, u_b, y) = \inf_{v \in X} J(v, u_b, y), \quad (2.2)$$

where

$$J(u, u_b, y) = \frac{1}{2} (V_b^{-1}(u - u_b), u - u_b)_X + \frac{1}{2} (V_o^{-1}(C\varphi - y), C\varphi - y)_{Y_o}. \quad (2.3)$$

The necessary optimality condition reduces the problem (2.2)–(2.3) to the following system [34]:

$$\begin{cases} \frac{\partial \varphi}{\partial t} = F(\varphi) + f, & t \in (0, T) \\ \varphi|_{t=0} = u, \end{cases} \quad (2.4)$$

$$\begin{cases} -\frac{\partial \varphi^*}{\partial t} - (F'(\varphi))^* \varphi^* = -(C')^* V_o^{-1}(C\varphi - y), & t \in (0, T) \\ \varphi^*|_{t=T} = 0, \end{cases} \quad (2.5)$$

$$V_b^{-1}(u - u_b) - \varphi^*|_{t=0} = 0 \quad (2.6)$$

with the unknowns  $\varphi$ ,  $\varphi^*$ ,  $u$ , where  $(F'(\varphi))^*$  is the adjoint to the Frechet derivative of  $F$ , and  $(C')^*$  is the adjoint to the Frechet derivative of  $C$  defined by  $(C'\varphi, \psi)_{Y_o} = (\varphi, (C')^*\psi)_Y$ ,  $\varphi \in Y$ ,  $\psi \in Y_o$ . Below for simplicity we consider a linear observation operator  $C(\varphi^t) = C\varphi^t$  and, therefore, use  $C^*$  instead of  $(C')^*$ .

The formulas (2.2)–(2.3) define implicitly a data-to-control map (or estimator) in the form

$$u = G^{-1}(y, u_b) = G^{-1}(C\varphi(u^t) + \xi_o, u^t + \xi_b). \quad (2.7)$$

This estimator can be characterized by the analysis pdf (see [14])

$$\begin{aligned} \rho_a(u, u^t) &= c_1(u^t) \cdot \exp\left(-\frac{1}{2} \|V_b^{-1/2}(u - u^t)\|_X^2 - \frac{1}{2} \|V_o^{-1/2}(C\varphi(u) - C\varphi(u^t))\|_{Y_o}^2\right) \\ &= c_1(u^t) \cdot \exp(-J(u, u^t, C\varphi(u^t))), \end{aligned} \quad (2.8)$$

where  $c_1(u^t) > 0$  is a normalization constant.

Below we assume that the estimation bias is very small as compared to deviations, i.e.

$$E(u - u^t) = \int (u - u^t) \rho_a(u, u^t) du \approx 0.$$

### 3. Coexistence measure

Let us consider an independent variable  $w \geq 0$ . For each  $w$ , the solution  $\Gamma(w) \in X$  to the equation

$$J(\Gamma(w), u^t, C\varphi(u^t)) = w$$

represents a manifold (locus) of equal likelihood  $c_1(u^t) \exp(-w)$  in the control space  $X$ . The manifold bounds the domain  $\Omega(w)$  where the cumulative density function  $\beta(w)$  is defined as follows:

$$\beta(w) = \int_{\Omega(w)} \rho_a(u, u^t) du.$$

This function shows the probability that an event  $u \sim \rho(u, u^t)$  falls ‘inside’ the domain  $\Omega(w)$ . Let us note that the manifold may consist of a few disconnected subsurfaces, and the domain – of a few disconnected subdomains. For a given confidence level  $\gamma$  the corresponding value of  $w^*$  satisfying the equation

$$\beta(w^*) = \gamma, \quad (3.1)$$

defines the confidence region  $\Omega(w^*)$ . All events  $u$  falling ‘outside’  $\Omega(w^*)$  are considered as ‘unlikely’ or ‘statistically insignificant’ events to be discarded. It is worth mentioning that for DA with nonlinear models, the confidence region can be topologically very complex, therefore the notions of ‘inside’ and ‘outside’ cannot be trivial.

Let us note that  $\beta(0) = 0$ ,  $\beta(\infty) = 1$  and  $\beta(w)$  is a monotonic increasing function of  $w$ . Therefore,  $\beta(w) < \gamma$  when  $w < w^*$ , and the criteria to test whether or not  $u$  falls into the confidence region  $\Omega(w^*)$  reads

$$J(u, u^t, C\varphi(u^t)) < w^*. \quad (3.2)$$

A particular optimal solution

$$\bar{u} = \arg \min_u J(u, \bar{u}_b, \bar{y}) = G(\bar{y}, \bar{u}_b) = G(C\varphi(u^t) + \bar{\xi}_o, u^t + \bar{\xi}_b)$$

corresponds to the actually observed data  $\bar{y}$  and  $\bar{u}_b$ , defined by the data errors  $\bar{\xi}_o$  and  $\bar{\xi}_b$  which have actually come to pass. Given  $\bar{u}$  as the best available approximation of  $u^t$ , the estimate of the analysis pdf is

$$\begin{aligned} \rho_a(u, \bar{u}) &= c_2(\bar{u}) \cdot \exp\left(-\frac{1}{2} \|V_b^{-1/2}(u - \bar{u})\|_X^2 - \frac{1}{2} \|V_o^{-1/2}(C\varphi(u) - C\varphi(\bar{u}))\|_{Y_o}^2\right) \\ &= c_2(\bar{u}) \cdot \exp(-J(u, \bar{u}, C\varphi(\bar{u}))), \quad c_2(\bar{u}) = \text{const} > 0. \end{aligned} \quad (3.3)$$

Let us denote  $\bullet|_{u^t}$  an “object associated to  $\rho_a(u, u^t)$ ”, and  $\bullet|_{\bar{u}}$  an “object associated to  $\rho_a(u, \bar{u})$ ”. We shall say that  $\bar{u}$  and  $u^t$  coexist if, simultaneously,  $\bar{u}$  is a statistically significant event in the distribution  $\rho_a(u, u^t)$ , i.e.  $\bar{u} \in \Omega(w^*)|_{u^t}$ , and  $u^t$  is a statistically significant event in the distribution  $\rho_a(u, \bar{u})$ , i.e.  $u^t \in \Omega(w^*)|_{\bar{u}}$ . Since both conditions have probability  $\gamma$  and they are assumed to be statistically independent, the “coexistence” is a random event with probability  $\gamma^2$  and has to be quantified as a random variable. In terms of the testing criteria (3.2) this reads as follows:

$$P[J(\bar{u}, u^t, C\varphi(u^t)) < w^*|_{u^t}, J(u^t, \bar{u}, C\varphi(\bar{u})) < w^*|_{\bar{u}}] = \gamma^2. \quad (3.4)$$

The *coexistence principle* simply means that we always expect with high probability that the truth falls within the confidence region defined for the analysis pdf  $\rho_a(u, \bar{u})$ . The *coexistence breach* may occur if the original non-gaussian analysis pdf is approximated by the gaussian pdf. In this case we shall say that the optimal solution is not *gauss-verifiable* on the whole.

In the finite-dimensional case ( $X = \mathbf{R}^n$ ) the gaussian approximation of the analysis density  $\rho_a(u, \bar{u})$  is given by

$$\bar{\rho}_a(u, \bar{u}) = c(\bar{u}) \exp\left(-\frac{1}{2} \|V^{-1/2}(\bar{u})(u - \bar{u})\|_X^2\right) \equiv c(\bar{u}) \exp(-\tilde{J}(u, \bar{u})), \quad (3.5)$$

where  $c(\bar{u}) = (2\pi)^{-n/2} (\det V(\bar{u}))^{-1/2}$ ,  $n$  is the dimension of the state space  $X$ ,  $V(\bar{u})$  is the covariance computed from the pdf (3.3), and the function  $\tilde{J}$  is called the *origin* [13]. For  $\bar{\rho}_a(u, \bar{u})$  the cumulative density function  $\beta(w)$  is known to be the  $\chi_n^2$ -cumulative density function  $F(w, n)$ , and  $w^*|_{\bar{u}} = \chi_n^2(\gamma)$  is the critical point of  $\chi_n^2$  distribution. Taking into account the definition of  $J$  in (2.3) we notice that

$$J(u^t, \bar{u}, C\varphi(\bar{u})) = J(\bar{u}, u^t, C\varphi(u^t)). \quad (3.6)$$

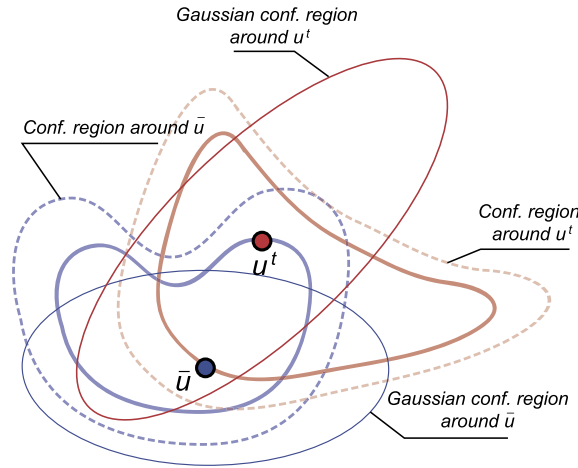


Fig. 1. Confidence regions.

This is an important condition for the coexistence principle to hold. However, as a result of using the gaussian  $\tilde{\rho}(u, \bar{u})$  instead of  $\rho(u, \bar{u})$  this condition may no longer be valid. In addition, the critical point  $w^*|_{\bar{u}}$  becomes  $\chi_n^2(\gamma)$ . Therefore, the coexistence is affected by the two differences:

$$\tilde{J}(u^t, \bar{u}) - J(\bar{u}, u^t, C\varphi(u^t)) \quad (3.7)$$

and

$$\chi_n^2(\gamma) - w^*|_{u^t}.$$

Let us note that  $w^*|_{u^t}$  is an integral quantity of the corresponding pdf and may differ from  $\chi_n^2(\gamma)$  not too significantly (unless the specified confidence level  $\gamma$  is too close to 1). In this case the difference (3.7) can be considered as a major cause of the coexistence breach. Therefore, the only condition for our approach to be valid is as follows:

$$\|\tilde{J}(u^t, \bar{u}) - J(\bar{u}, u^t, C\varphi(u^t))\| \gg \|\chi_n^2(\gamma) - w^*|_{u^t}\|. \quad (3.8)$$

Because the truth is not known, instead of (3.7) we can consider the difference

$$\tilde{J}(\bar{u}, u^t) - J(u^t, \bar{u}, C\varphi(\bar{u})) = \frac{1}{2} \|V^{-1/2}(u^t)(u^t - \bar{u})\|_X^2 - J(u^t, \bar{u}, C\varphi(\bar{u})). \quad (3.9)$$

In the above formulation the truth  $u^t$  becomes a random variable which falls inside a neighborhood of  $\bar{u}$  consistent with  $\rho_a(u, \bar{u})$ , i.e.  $u^t = \bar{u} + v$ , where  $v$  is the optimal solution (analysis) error. Thus, we are interested in evaluating the difference

$$\theta(v, \bar{u}) = \frac{1}{2} \|V^{-1/2}(\bar{u} + v)v\|_X^2 - J(\bar{u} + v, \bar{u}, C\varphi(\bar{u})) \quad (3.10)$$

averaged over  $\rho_a(u, \bar{u})$ :

$$E[\theta(v, \bar{u})] = \int \theta(v, \bar{u}) \rho_a(\bar{u} + v, \bar{u}) dv. \quad (3.11)$$

We shall call  $E[\theta(v, \bar{u})]$  the *coexistence measure*. It can be used to quantify gauss-verifiability of optimal solutions. From other perspective, it is also a global measure of deviation of  $\rho_a(u, \bar{u})$  from normality.

The above idea is illustrated in Fig. 1 for the 2D case. Here we present the exact confidence regions associated to  $u^t$  and  $\bar{u}$  in dashed lines, the gaussian confidence regions – in solid thin lines and the contours of equal likelihood – in solid thick lines. Due to the nonlinearity of the model equations  $J$  is not quadratic and the analysis pdf (2.8) and (3.3) are not gaussian, therefore the shape of the confidence regions significantly differs from ellipsoidal. However, we can see, that  $\bar{u}$  lies on the certain likelihood locus associated to  $u^t$ , whereas  $u^t$  lies on the same likelihood locus associated to  $\bar{u}$ . However,  $u^t$  falls outside the Gaussian confidence region associated to  $\bar{u}$ .

**Remark 1.** In the Bayesian approach the posterior pdf is given by Bayes' rule (e.g. [42]):

$$\begin{aligned} \rho(u) &= c_3(\bar{u}_b, \bar{y}) \cdot \exp\left(-\frac{1}{2} \|V_b^{-1/2}(u - \bar{u}_b)\|_X^2 - \frac{1}{2} \|V_o^{-1/2}(C\varphi(u) - \bar{y})\|_{Y_o}^2\right) \\ &= c_3(\bar{u}_b, \bar{y}) \cdot \exp(-J(u, \bar{u}_b, \bar{y})), \end{aligned} \quad (3.12)$$

where  $c_3(\bar{u}_b, \bar{y})$  is the normalization constant. The maximizer of  $\rho(u)$  is obtained by minimizing  $J(u, \bar{u}_b, \bar{y})$ . Thus, the mode of  $\rho(u)$  is the particular optimal solution. However, the pdf  $\rho_a(u, \bar{u})$  and  $\rho(u)$  are different, which explains the difference between the analysis error covariance and the Bayesian posterior covariance. Whereas the density  $\rho_a(u, u^t)$  has been artificially derived to characterize the estimator (2.7), one can, in turn, derive an estimator characterized by  $\rho(u)$ . It is in the form [14]

$$u = G^{-1}(y, u_b) = G^{-1}(\bar{y} + \xi_o, \bar{u}_b + \xi_b), \quad (3.13)$$

where  $G$ , as before, is defined implicitly by (2.2)–(2.3). Since the formulas (2.8) and (3.12) are rarely useful in practical computations with high-dimensional systems, the estimators (2.7) and (3.13) can be used for evaluating the moments of  $\rho_a(u, u^t)$  and  $\rho(u)$ , correspondingly. In the variational DA context this method is used in [11–14], being referred to as the “fully nonlinear ensemble method”. All results of this paper will remain valid if the Bayesian estimator (3.13) is considered instead of (2.7).

#### 4. Coexistence measure decomposition

It is important to present  $E[\theta(v, \bar{u})]$  as a sum of contributions associated to perturbations  $v_i$  in the elements of the control vector  $\bar{u}$ . Those could be obtained as an outcome of a global sensitivity analysis applied to  $E[\theta(v, \bar{u})]$ , but such analysis is hardly feasible for the models in mind. One possible way to achieve the mentioned decomposition is to use the relationship

$$E[\|V^{-1/2}(\bar{u} + v)v\|_X^2] = \text{tr}\{E[V^{-1}(\bar{u} + v)vv^T]\}.$$

There is no guarantee that the elements of the trace are non-negative values, so we consider a modification of  $E[\theta(v, \bar{u})]$  allowing us to mitigate this difficulty.

Since the expectation  $E[V^{-1}(\bar{u} + v)vv^T]$  is the integral with respect to  $v$ , under the conditions of the mean value theorem, there exist  $v_0$  such that

$$E[V^{-1}(\bar{u} + v)vv^T] = V^{-1}(\bar{u} + v_0)E[vv^T] = V^{-1}(\bar{u} + v_0)V, \quad (4.1)$$

where  $V$  is the covariance of the distribution  $\rho_a(\bar{u} + v, \bar{u})$ :

$$V = E[vv^T] = \int vv^T \rho_a(\bar{u} + v, \bar{u}) dv. \quad (4.2)$$

Since  $v_0$  is not known, instead of  $V^{-1}(\bar{u} + v_0)$  we will consider in (4.1) its expectation  $E[V^{-1}(\bar{u} + v)]$ , then  $E[V^{-1}(\bar{u} + v)]V = E[V^{-1}(\bar{u} + v)V]$ , and instead of  $E[\theta(v, \bar{u})]$  we introduce

$$\mathcal{D} = \frac{1}{2} \text{tr}\{E[V^{-1}(\bar{u} + v)V]\} - C_1, \quad (4.3)$$

where

$$C_1 = E[J(\bar{u} + v, \bar{u}, C\varphi(\bar{u}))]. \quad (4.4)$$

Consider the square-root decomposition of  $V$  in the form

$$V = Q Q^T, \quad (4.5)$$

such that  $Q : X \rightarrow X$ . Then

$$\text{tr}\{V^{-1}(\bar{u} + v)V\} = \text{tr}\{V^{-1}(\bar{u} + v)Q Q^T\} = \text{tr}\{Q^T V^{-1}(\bar{u} + v)Q\}$$

and

$$\mathcal{D} = \text{tr}\left\{\frac{1}{2}E[Q^T V^{-1}(\bar{u} + v)Q] - \frac{1}{n}C_1 I_n\right\}. \quad (4.6)$$

The last formula implies

$$\mathcal{D} = \sum_{i=1}^n d_i, \quad (4.7)$$

where

$$d_i = \frac{1}{2}(Ae_i, e_i)_X - \frac{C_1}{n}, \quad A = E[Q^T V^{-1}(\bar{u} + v)Q].$$

The operator  $A$  acts from  $X$  to  $X$ , therefore formula (4.7) helps to evaluate an individual contribution of each state variable into the integral value  $\mathcal{D}$ .

**Remark 2.** Since  $V^{-1}(\bar{u} + v)$  is positive definite for each  $v$ , it is easily seen that  $(Ae_i, e_i)_X \geq 0$ , because

$$(Q^T V^{-1}(\bar{u} + v) Q e_i, e_i)_X = (V^{-1}(\bar{u} + v) Q e_i, Q e_i)_X \geq \epsilon \|Q e_i\|^2 \geq 0, \quad \epsilon = \text{const} > 0.$$

Let us note that  $d_i = (Ae_i, e_i)_X/2 - C_1/n$  may not always be positive in theory. In practice, the coexistence breach mainly occurs when  $(Ae_i, e_i)_X/2 \gg C_1/n$ .

## 5. Coexistence measure approximations

Let  $V(\bar{u})$  be the covariance of the distribution  $\rho_a(u, \bar{u})$ . It is easy to show that

$$\frac{1}{2} \int \|V^{-1/2}(\bar{u})v\|_X^2 \rho_a(\bar{u} + v, \bar{u}) dv = \frac{n}{2}. \quad (5.1)$$

If the functional  $J(\bar{u} + v, \bar{u}, C\varphi(\bar{u}))$  is quadratic ( $F$  is linear) it can be exactly represented through the covariance  $V(\bar{u})$ , otherwise one may assume

$$J(\bar{u} + v, \bar{u}, C\varphi(\bar{u})) \approx \|V^{-1/2}(\bar{u})v\|_X^2, \quad \bar{u} + v \sim \rho_a(u, \bar{u}). \quad (5.2)$$

It follows from (5.1) and (5.2) that

$$E[J(\bar{u} + v, \bar{u}, C\varphi(\bar{u}))] \approx \frac{n}{2} \quad (5.3)$$

and, therefore,  $C_1$  can be taken as  $n/2$  in (4.6) and it can be re-written as

$$\mathcal{D} = \text{tr} \left\{ \frac{1}{2} E[Q^T V^{-1}(\bar{u} + v) Q] - \frac{I_n}{2} \right\}. \quad (5.4)$$

Assuming that the estimation biases are much smaller than particular estimation deviations one may expect that the accuracy of the approximation (5.3) for expectation is far more accurate than the approximation (5.2) for a particular event  $v$ . This is easy to check numerically.

For practical computations in (4.6) one needs to define the inverse (or pseudo-inverse) of the covariance  $V(\bar{u} + v)$  for each integration point  $w$ . Computing invertible  $V(\bar{u} + v)$  by the Monte Carlo method requires an ensemble of optimal solutions of a size greater than  $n$ , which would be an enormous computational task for large  $n$ . However, the covariance  $V(\cdot)$  can be approximated by the inverse Hessian  $H(\cdot)$  of the auxiliary (linearized) DA problem (see e.g. [11,44]):

$$V^{-1}(\bar{u} + v) \approx H(\bar{u} + v). \quad (5.5)$$

Taking into account (4.3) we note that instead of (5.5) we are actually trying to approximate the expectation of the inverse covariance by the expectation of the Hessian:

$$E(V^{-1}(\bar{u} + v)) \approx E(H(\bar{u} + v)). \quad (5.6)$$

As before with  $C_1$ , the accuracy of the approximation (5.6) is generally far better than the accuracy of (5.5) for a particular event  $v$ .

Using the approximations (5.3) and (5.6), the CM defined by (4.3) can be represented as follows:

$$\mathcal{D} \approx D = \text{tr} \left\{ \frac{1}{2} E[H(\bar{u} + v) V] - \frac{I_n}{2} \right\}.$$

If the decomposition (4.5) is valid, the above formula implies that

$$D = \text{tr} \left\{ \frac{1}{2} E[Q^T H(\bar{u} + v) Q] - \frac{I_n}{2} \right\}. \quad (5.7)$$

In this case formula (4.7) holds and we can finally write

$$D = \sum_{i=1}^n d_i, \quad d_i = \frac{1}{2} (Ae_i, e_i)_X - \frac{1}{2}, \quad A = E[Q^T H(\bar{u} + v) Q]. \quad (5.8)$$

**Remark 3.** The Hessian  $H(\cdot)$  is defined as follows [11]:

$$\begin{cases} \frac{\partial \psi}{\partial t} - F'(\varphi(\cdot))\psi = 0, & t \in (0, T), \\ \psi|_{t=0} = v, \end{cases} \quad (5.9)$$

$$\begin{cases} -\frac{\partial \psi^*}{\partial t} - (F'(\varphi(\cdot)))^* \psi^* = -C^* V_o^{-1} C \psi, & t \in (0, T) \\ \psi^*|_{t=T} = 0, \end{cases} \quad (5.10)$$

$$H(\cdot)v = V_b^{-1}v - \psi^*|_{t=0}. \quad (5.11)$$



## 6. Computational procedure

To compute  $D$  and  $d_i$  by (5.8) it is necessary to know the square-root  $Q : X \rightarrow X$  from (4.5). Let us assume that we know the sample covariance  $\hat{V}(\bar{u})$  of full rank  $n$ . In this case we use  $V = \hat{V}(\bar{u})$  and the SVD decomposition is valid:

$$\hat{V} = USU^T,$$

where  $S = \text{diag}\{s_1, \dots, s_n\}$  are singular values of  $\hat{V}$ , and the columns of the matrix  $U$  are the singular vectors of  $\hat{V}$ . Then

$$\hat{V} = US^{1/2}S^{1/2}U^T = US^{1/2}U^TUS^{1/2}U^T = QQ^T, \quad (6.1)$$

where  $Q = US^{1/2}U^T : X \rightarrow X$ .

In practice the full-rank sample covariance  $\hat{V}$  is unlikely to be constructed, thus instead of  $\hat{V}(\bar{u})$  one can use the inverse Hessian, i.e.  $V = H^{-1}(\bar{u})$ . In this case we have the SVD decomposition for the Hessian:

$$H(\bar{u}) = USU^T,$$

and

$$Q = US^{-1/2}U^T.$$

When it is important to distinguish between  $Q$  obtained using  $\hat{V}$  or  $H^{-1}$  we will denote them  $Q|_{\hat{V}}$  or  $Q|_H$ , respectively.

An efficient way of computing  $D$  defined by (5.7) is by performing the eigenvalue analysis of the matrix

$$\tilde{H}(\bar{u} + v) = Q^T H(\bar{u} + v) Q. \quad (6.2)$$

If  $\lambda_j$ ,  $j = 1, \dots, n$  are the eigenvalues of  $\tilde{H}(\bar{u} + v)$ , then

$$\text{tr}\{\tilde{H}(\bar{u} + v)\} = \sum_{j=1}^n \lambda_j$$

and

$$D = \frac{1}{2} E \left[ \sum_{j=1}^n \lambda_j - 1 \right]. \quad (6.3)$$

It follows from (6.3) that a reasonable approximation of  $D$  can be achieved by using a subset of eigenvalues  $\{\lambda_j\}$  most distinct from 1. Such a subset can be computed by means of iterative methods (Lanczos or Arnoldi), which require only the matrix-vector product involving  $\tilde{H}(\bar{u} + v)$ . In this case we use the limited-memory representation

$$\tilde{H}(\bar{u} + v) = I^{(n)} + W(\Lambda - I^{(m)})W^T, \quad (6.4)$$

where  $\Lambda = \text{diag}\{\lambda_1, \dots, \lambda_m\}$ ,  $W$  is an  $n \times m$ -matrix containing  $m$  eigenvectors  $W_i$  as its columns, and  $I^{(n)}$ ,  $I^{(m)}$  are the identity matrices of sizes  $n$  and  $m$ . From (6.4) it follows that

$$(\tilde{H}(\bar{u} + v))_{ii} = (\tilde{H}(\bar{u} + v)e_i, e_i)_X = 1 + \sum_{j=1}^m (\lambda_j - 1) W_{ij}^2.$$

Therefore,

$$D = \frac{1}{2} E \left[ \sum_{j=1}^m \lambda_j - 1 \right] \quad (6.5)$$

and

$$D = \sum_{i=1}^n d_i, \quad d_i = \frac{1}{2} E \left[ \left( \sum_{j=1}^m \lambda_j - 1 \right) W_{ij}^2 \right]. \quad (6.6)$$

The expectation is taken over the analysis pdf as defined in (3.11). In actual computations  $E[\cdot]$  is approximated by the sample mean, i.e.

$$D = \frac{1}{2K} \sum_{k=1}^K \left[ \sum_{j=1}^m \lambda_j|_{v^{(k)}} - 1 \right] \quad (6.7)$$



and

$$D = \sum_{i=1}^n d_i, \quad d_i = \frac{d_i^*}{2K}, \quad d_i^* = \sum_{k=1}^K \left[ \left( \sum_{j=1}^m \lambda_j|_{v^{(k)}} - 1 \right) W_{ij}^2|_{v^{(k)}} \right]. \quad (6.8)$$

Above  $v^{(k)} = u^{(k)} - \bar{u}$ ,  $u^{(k)} \sim \rho(u, \bar{u})$  and  $\bullet|_{v^{(k)}}$  means “object associated to  $v^{(k)}$ ”, which is used in (6.2). An ensemble (sample) of optimal solutions  $\{u^{(k)}\}$ ,  $k = 1, \dots, K$  is generated by solving  $K$  times the DA problem (2.2)–(2.3) with perturbed data  $u_b = \bar{u} + \xi_b^{(k)}$ ,  $\xi_b \sim \mathcal{N}(0, V_b)$  and  $y = C\varphi(\bar{u}) + \xi_o^{(k)}$ ,  $\xi_o \sim \mathcal{N}(0, V_o)$ , see [11].

Finally, the computational procedure can be presented in the form of

#### Algorithm 1.

1. Compute ensemble (sample) of optimal solutions  $\{u^{(k)}\}$ ,  $k = 1, \dots, K$
2. Compute the sample covariance  $\hat{V}(\bar{u})$  (optional)
3. Compute SVD of  $H(\bar{u})$  or  $\hat{V}(\bar{u})$  (optional) to define  $Q|_H$  or  $Q|_{\hat{V}}$  (optional)
4. Start loop on  $k$ 
  - 4.1. Compute  $m$  eigenpairs  $\{\lambda_j, W_j\}$  of  $\tilde{H}(u^{(k)}) = Q^T H(u^{(k)}) Q$  using the Lanczos method, whereas the Hessian-vector product  $H(u^{(k)})v$  is defined by (5.9)–(5.11)
  - 4.2. Compute  $d_i^*$  as defined in (6.8)
5. End loop on  $k$
6. Compute  $d_i$  and  $D$  as defined (6.8).

**Remark 4.** In practice, for the purpose of finding  $D$  and  $d_i$  only a very small ensemble ( $K \ll n$ ) is likely to be generated. However, for the research purpose we can get ensembles of significant size, including those with  $K \gg n$ . Then, the sample mean and the sample covariance are defined as follows:

$$\hat{u} = \frac{1}{K} \sum_{k=1}^K u^{(k)}, \quad \hat{V}(\bar{u}) = \frac{1}{K-1} \sum_{k=1}^K (u^{(k)} - \hat{u})(u^{(k)} - \hat{u})^T. \quad (6.9)$$

The full-rank sample covariance is required to get the square-root factor  $Q|_{\hat{V}}$ . As discussed above, in real applications the covariance  $\hat{V}(\bar{u})$  has to be approximated by the inverse Hessian and  $Q|_H$  will be used instead of  $Q|_{\hat{V}}$ . Thus,  $Q|_{\hat{V}}$  is needed to investigate the error related to this approximation. Also,  $\hat{V}(\bar{u})$  will be used in numerical experiments for computing a test statistic for multivariate normality.

**Remark 5.** To avoid overloading this paper with technical details, the question of using the limited-memory form of  $H(\bar{u})$ , as well as some other computational techniques which seem useful for computing the CM, will be discussed separately.

## 7. Confidence interval for coexistence measure

Let us consider the representation (5.4) in the form:

$$D = \sum_{i=1}^n E[b_i], \quad (7.1)$$

where

$$b_i = \frac{1}{2} (e_i^T Q^T V^{-1}(\bar{u} + v) Q e_i - 1) = \frac{1}{2} (Q_i^T V^{-1}(\bar{u} + v) Q_i - 1).$$

We rewrite  $b_i$  as follows:

$$b_i = \frac{1}{2n} (\sqrt{n} Q_i^T V^{-1}(\bar{u} + v) \sqrt{n} Q_i - n).$$

This quantity defines the difference between the likelihood of a sigma-point  $\sqrt{n} Q_i$  of  $V(\bar{u})$  (see [21]) in the distribution  $\mathcal{N}(0, V(\bar{u} + v))$  and in its parent distribution  $\mathcal{N}(0, V(\bar{u}))$ . In order to access the bounds for  $b_i$  let us consider instead of  $\sqrt{n} Q_i$  a random vector  $\xi \sim \mathcal{N}(0, V(\bar{u} + v))$  in its parent distribution. Given  $V(\bar{u} + v) = \bar{Q} \bar{Q}^T$ ,  $\xi = \bar{Q} r$ , where  $r \sim \mathcal{N}(0, I)$ , we obtain

$$\bar{b} = \frac{1}{2n} (\xi^T H(\bar{u} + v) \xi - n) = \frac{1}{2n} (r^T \bar{Q}^T \bar{Q}^{-T} \bar{Q}^{-1} \bar{Q} r - n) = \frac{1}{2n} (r^T r - n).$$

It is well known that  $r^T r - n$  has the centered  $\chi_n^2$  distribution with  $n$  degrees of freedom [5], which can be well approximated by the gaussian for large  $n$ . Therefore, for the moments of  $\bar{b}$  we have

**Table 1**Summary of asymptotic values of integral measures,  $K = 2500$ .

Case	$E[\theta(v, \bar{u})]$	$D _{\hat{\varphi}}$	$D _{H^{-1}}$	$d^*/D$	$(S_{JB} - n)/2$	$D^*$
A1	507.41	548.17	1158.23	1.3E–4	216.60	30.00
A2	53.00	45.26	46.21	3.25E–3	110.34	30.00
B1	268.53	263.20	240.37	6.24E–4	575.10	30.00
B2	23.19	20.14	18.38	8.16E–3	2.14	30.00
C1	315.30	348.59	252.32	5.94E–4	560.36	30.00
C2	11.64	15.37	14.65	1.02E–2	452.22	30.00

$$E[\bar{b}] = \frac{1}{2n} E[r^T r - n] = 0,$$

$$E[\bar{b}^2] = \frac{1}{4n^2} E[(r^T r - n)(r^T r - n)] = \frac{2n}{4n^2} = \frac{1}{2n}.$$

The confidence interval for  $\bar{b}$  can now be defined as follows:

$$\bar{b} < d^* = \frac{\alpha}{\sqrt{2n}}, \quad (7.2)$$

where  $\alpha > 0$  is a real number. For example,  $\alpha = 2$  approximately corresponds to  $\chi_n^2(0.05)$ , and  $\alpha = 3$  – to  $\chi_n^2(0.001)$ . If condition (7.2) is not satisfied,  $\xi$  is considered as an event from a different distribution. Since  $D = n\bar{b}$ , then  $E[D^2] = n^2 E[\bar{b}^2] = n/2$  and the confidence interval for  $D$  is

$$D < D^* = \alpha \sqrt{\frac{n}{2}}. \quad (7.3)$$

## 8. General description of numerical experiments

The purpose of numerical experiments is to calculate the approximated coexistence measure  $D$  and its element-wise decomposition  $d_i$ ,  $i = 1, \dots, n$  using Algorithm 1.

First, a large ensemble (sample) of optimal solutions  $\{u^{(k)}\}$ ,  $k = 1, \dots, K$  is generated,  $K = 2500$ . On a basis of this ensemble the “asymptotic” values of  $D$  and  $d_i$  are computed. The values of  $D$  are summarized in Table 1, whereas the scaled decompositions  $d_i/D$  as functions of  $x = (i-1)h_x$ ,  $h_x = 1/(n-1)$ , are presented in Figs. 4–6, mid panels. In practice, only very small ensembles are likely to be generated. Thus, in order to assess the sampling error we consider disjoint subsets (sub-ensembles) of the large ensemble, each of size  $K_1 \ll K$ ,  $K_1 = 50$ . On a basis of each sub-ensemble the corresponding  $d_i^{(k_1)}$  is computed, together they form the ensemble

$$\{d_i^{(k_1)}\}, \quad k_1 = 1, \dots, \text{int}(K/K_1).$$

This ensemble is used for constructing the envelope for  $d_i/D$ , which includes about 70% of all  $d^{(k_1)}/D$ . The envelopes are presented in Figs. 4–6, lower panels.

The values of  $D$  have to be compared with the value of the test statistic

$$(S_{JB} - n)/2, \quad (8.1)$$

where

$$S_{JB} = nK \left( \frac{b_{1,n}^2}{6} + \frac{(b_{2,n} - 3)^2}{24} \right) \quad (8.2)$$

is the Jarque–Bera test statistic for multivariate normality;  $b_{1,n}$  and  $b_{2,n}$  are Srivastava’s multivariate sample skewness and kurtosis, respectively (see [19] and [41]). The Jarque–Bera statistic has  $\chi_{n+1}^2$  distribution, which for large  $n$  is well approximated by the Gaussian, with  $E[S_{JB}] = n$ ,  $E[(S_{JB} - n)^2] = 2n$ . Therefore  $E[(S_{JB} - n)/2] = 0$ ,  $E[((S_{JB} - n)/2)^2] = n/2$  and the confidence interval for  $(S_{JB} - n)/2$  coincides with the interval for  $D$  given in (7.3). This comparison is useful to see both the similarity and the difference between the classical test statistics for multivariate normality and the CM. Let us remind that the CM should be considered as an indirect measure of deviation from the gaussian.

The major attention is paid to the scaled decomposition  $d_i/D$ , showing the partial contributions associated to the elements of the control vector  $u_i$ . It reveals the subsets of the control vector which contribute most significantly into  $D$ . We expect that the estimated gaussian characterization of these subsets is not reliable. For example, the estimated univariate gaussian pdf for an element from this subset may not approximate its actual marginal pdf. Since we consider a distributed control, these subsets can be interpreted as localized spatial areas where gauss-verifiability does not hold.

The scaled decomposition  $d_i/D$  has to be compared to the results of the sensitivity analysis of  $E[\theta(v, \bar{u})]$ . There is no obvious mathematical connection between the sensitivity analysis and the suggested decomposition approach, however, intuitively, these two approaches should provide similar results. Let us rewrite (3.10)–(3.11) in the form

$$E[\theta(v, \bar{u})] \approx \frac{1}{2K} \sum_{k=1}^K [\|V^{-1/2}(\bar{u} + v^{(k)})v^{(k)}\|_X^2 - 2J(\bar{u} + v^{(k)}, \bar{u}, C\varphi(\bar{u}))].$$

The contribution of the  $i$ -th element of the control vector into  $E[\theta(v, \bar{u})]$  can be estimated as

$$(E[\theta(v, \bar{u})])_i = E[\theta(v, \bar{u})]_{|v_i^{(k)}=0}.$$

Then, the sensitivity is defined as follows:

$$\bar{z}_i = z_i/c, \quad (8.3)$$

where

$$z_i = \frac{(E[\theta(v, \bar{u})])_i}{E[\theta(v, \bar{u})]} - 1, \quad c = \sum_{i=1}^n |z_i|. \quad (8.4)$$

Following the above sensitivity definition we expect  $\bar{z}_i < 0$ . This is certainly true for linear problems, but may not hold for nonlinear problems. Unfortunately, more sophisticated approaches such as ANOVA [2], which guarantee a definite sign of the sensitivities, are significantly more computationally expensive.

## 9. Numerical implementation

### 9.1. Numerical model

As a model we use the 1D Burgers equation with a nonlinear viscous term

$$\frac{\partial \varphi}{\partial t} + \frac{1}{2} \frac{\partial (\varphi^2)}{\partial x} = \frac{\partial}{\partial x} \left( \mu(\varphi) \frac{\partial \varphi}{\partial x} \right), \quad \varphi = \varphi(x, t), \quad t \in (0, T), \quad x \in (0, 1), \quad (9.1)$$

with the Neumann boundary conditions

$$(d\varphi/dx)|_{x=0} = (d\varphi/dx)|_{x=1} = 0 \quad (9.2)$$

and the viscosity coefficient

$$\mu(\varphi) = \mu_0 + \mu_1(d\varphi/dx)^2, \quad \mu_0, \mu_1 = \text{const} > 0. \quad (9.3)$$

The nonlinear diffusion term with  $\mu(\varphi)$  dependent on  $\partial\varphi/\partial x$  is introduced to mimic eddy viscosity (turbulence), which depends on the field gradients (pressure, temperature), rather than on the field value itself. This type of  $\mu(\varphi)$  also allows us to formally qualify the problem (9.1)–(9.3) as strongly nonlinear [10]. Burgers equations are sometimes considered in the DA context as a simple model describing elements of atmospheric flow motion.

We use the implicit time discretization as follows

$$\frac{\varphi^i - \varphi^{i-1}}{h_t} + \frac{\partial}{\partial x} \left( \frac{1}{2} w(\varphi^i) \varphi^i - \mu(\varphi^i) \frac{\partial \varphi^i}{\partial x} \right) = 0, \quad i = 1, \dots, N, \quad x \in (0, 1), \quad (9.4)$$

where  $i$  is the time integration index,  $h_t = T/N$  is a time step. The spatial differential operator is discretized on a uniform grid ( $h_x$  is the spatial discretization step,  $j = 1, \dots, M$  is the node number,  $M$  is the total number of grid nodes) using the ‘power law’ first-order scheme as described in [40], which yields quite a stable discretization (this scheme allows  $\mu(\varphi)$  as small as  $10^{-5}$  for  $M = 201$  without noticeable oscillations). For each time step we perform nonlinear iterations on coefficients  $w(\varphi) = \varphi$  and  $\mu(\varphi)$ , assuming initially that  $\mu(\varphi^i) = \mu(\varphi^{i-1})$  and  $w(\varphi^i) = \varphi^{i-1}$ , and keep iterating until (9.4) is satisfied (i.e. the norm of the left-hand side in (9.4) becomes smaller than a threshold  $\epsilon_1 = 10^{-12}\sqrt{M}$ ).

In all computations presented in this paper we use the following parameters: observation period  $T = 0.32$ , discretization steps  $h_t = 0.004$ ,  $h_x = 0.005$ , state vector dimension  $n = M = 201$ , and parameters in (9.3)  $\mu_0 = 10^{-4}$ ,  $\mu_1 = 10^{-6}$ .

A general property of Burgers solutions is that a smooth initial state evolves into a state characterized by areas of severe gradients (or even shocks in the inviscid case). These are precisely the areas of a strong nonlinearity where the coexistence breach is likely to occur. This behavior can be seen in Fig. 2, cases A and B. In case C, a more complex behavior is simulated. Here we observe positive and negative sub-domains of the state variable moving towards the center of the domain, where they eventually collide.

The observation scheme for each case consists of a set of  $L$  stationary sensors located at: case A –  $x_k = (0.35, 0.45, 0.55, 0.65)$ , case B –  $x_k = (0.3, 0.4, 0.5, 0.6, 0.7)$  and case C –  $x_k = (0.35, 0.4, 0.5, 0.6, 0.65)$ . Observations are assimilated each time step at every sensor. Observation errors are uncorrelated with the error variance  $\sigma_o^2 = 0.001$ . The observation error covariance matrix  $V_o$  is therefore diagonal of size  $(N + 1) \times L$ .

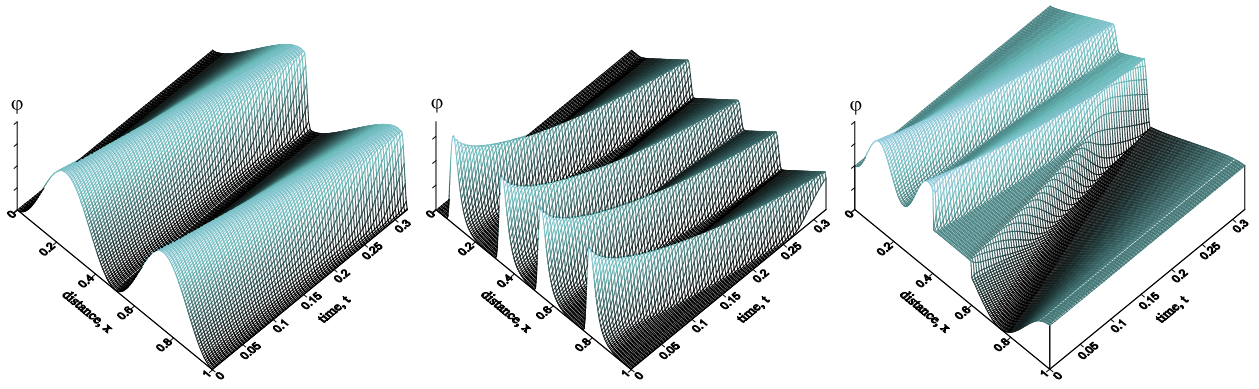


Fig. 2. Field evolution for different initial conditions: left/center/right – cases A/B/C, correspondingly.

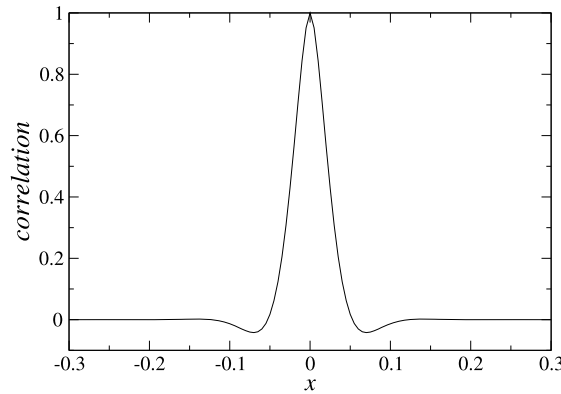


Fig. 3. Correlation function.

## 9.2. A priori information

The background error covariance matrix  $V_b$  in (2.3) is defined under the assumption that the background error belongs to the Sobolev space  $W_2^2[0, 1]$  (see [12], Section 5.1). The resulting background correlation function is presented in Fig. 3. For each initial state two different background variance functions are considered. The background  $3\sigma$ -envelope positive margin, i.e.  $+3\sigma_b(x)$ , can be seen in Figs. 4–6, upper panels.

## 9.3. Miscellaneous

- (a) In all numerical experiments  $\mathcal{D}$  and  $d_i$  are computed based on the assigned true state, i.e.  $\bar{u} = u^t$ .
- (b) Number of eigenpairs of  $\tilde{V}(\bar{u} + v^{(k)})$  in (6.2) computed by the Lanczos method is  $m = 20$ .
- (c) Truncated eigenvalue decomposition of symmetric operators (Hessian, covariance matrix) is computed by means of the Implicitly Restarted Arnoldi Method implemented in the ARPACK software [33]. This is referred to as ‘Lanczos’ throughout.
- (d) Consistent tangent linear and adjoint models have been generated from the original forward solver implementing (9.4) by the Automatic Differentiation tool TAPENADE [17] and checked using the standard gradient test.
- (e) The random series are produced by a pseudo-random generator, which uses the subroutines ‘gasdev’ and ‘run2’ provided in [39].

# 10. Numerical results

## 10.1. Summary of results

For numerical experiments we consider three cases (A, B and C) which correspond to three different initial states; the corresponding state evolution fields  $\varphi(x, t)$  are presented in Fig. 2. For each initial state we consider two different background variance functions; thus each case splits in two sub-cases (for example: A1 and A2). In sub-cases ‘2’ the background variance is 10-times smaller than in sub-case ‘1’. Consequently, the confidence region  $\Omega(w^*)$  around  $\bar{u}$  (based on  $\rho_a(u, \bar{u})$ ) is smaller as compared to sub-case ‘1’ and a significant reduction in  $D$  has to be expected.

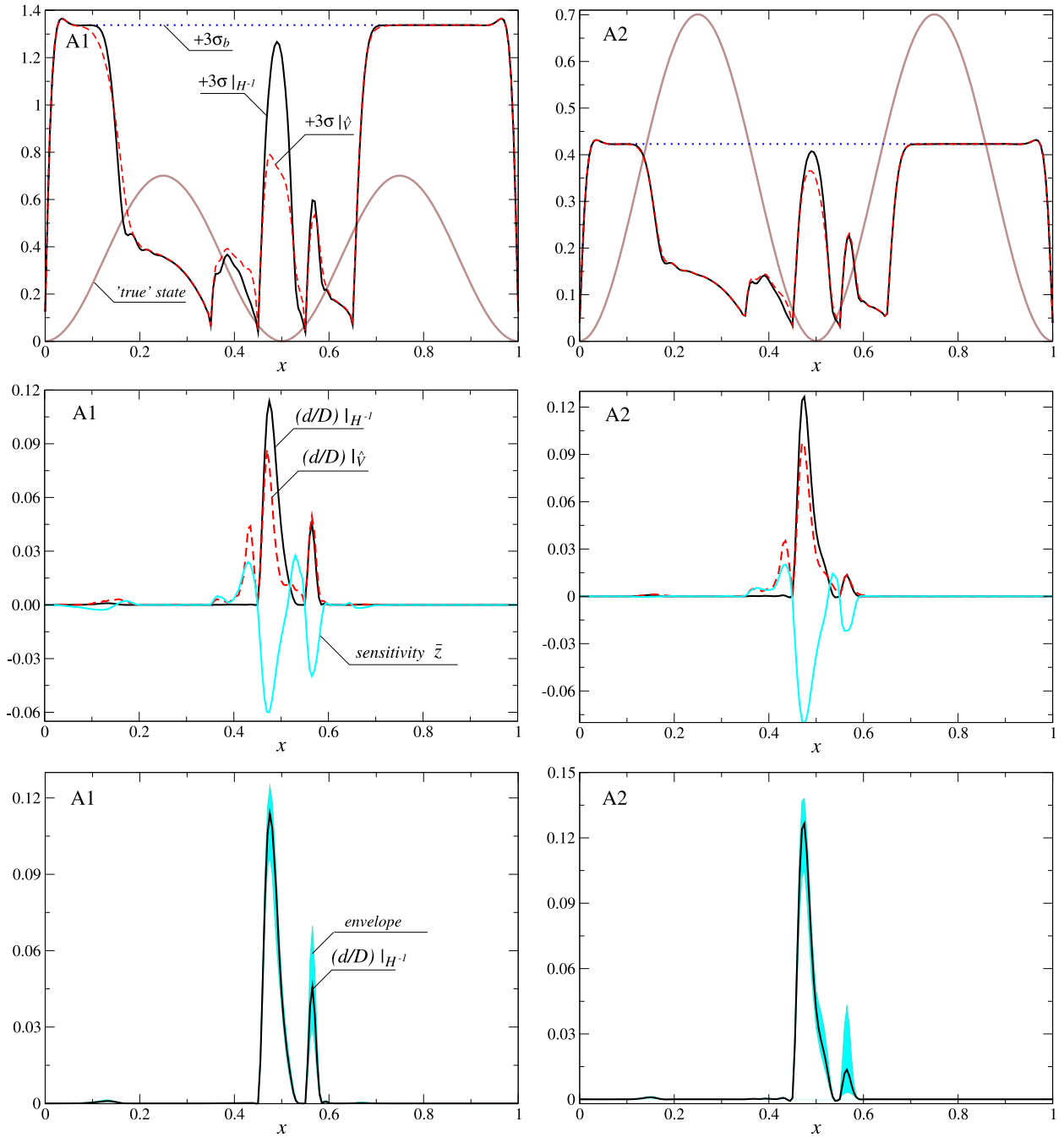


Fig. 4. Cases A1 and A2, see Section 10.1 for detail.

The results on  $D$  and related integral quantities are summarized in Table 1. It contains:

- $E[\theta(v, \bar{u})]$ , the original CM defined by (3.11);
- $D|_{\hat{\psi}}$ , the approximated CM computed at  $Q|_{\hat{\psi}(u^t)}$  in (6.2);
- $D|_{H^{-1}}$ , the approximated CM computed at  $Q|_{H^{-1}(u^t)}$  in (6.2);
- $d^*/D = d^*/D|_{H^{-1}}$ , the scaled confidence threshold computed at  $Q|_{H^{-1}(u^t)}$  in (6.2);
- $(S_{JB} - n)/2$ , the modified Jarque–Bera test statistic for multivariate normality (8.1)–(8.2);
- $D^*$ , the critical value (7.3) for  $D$  and  $(S_{JB} - n)/2$ .

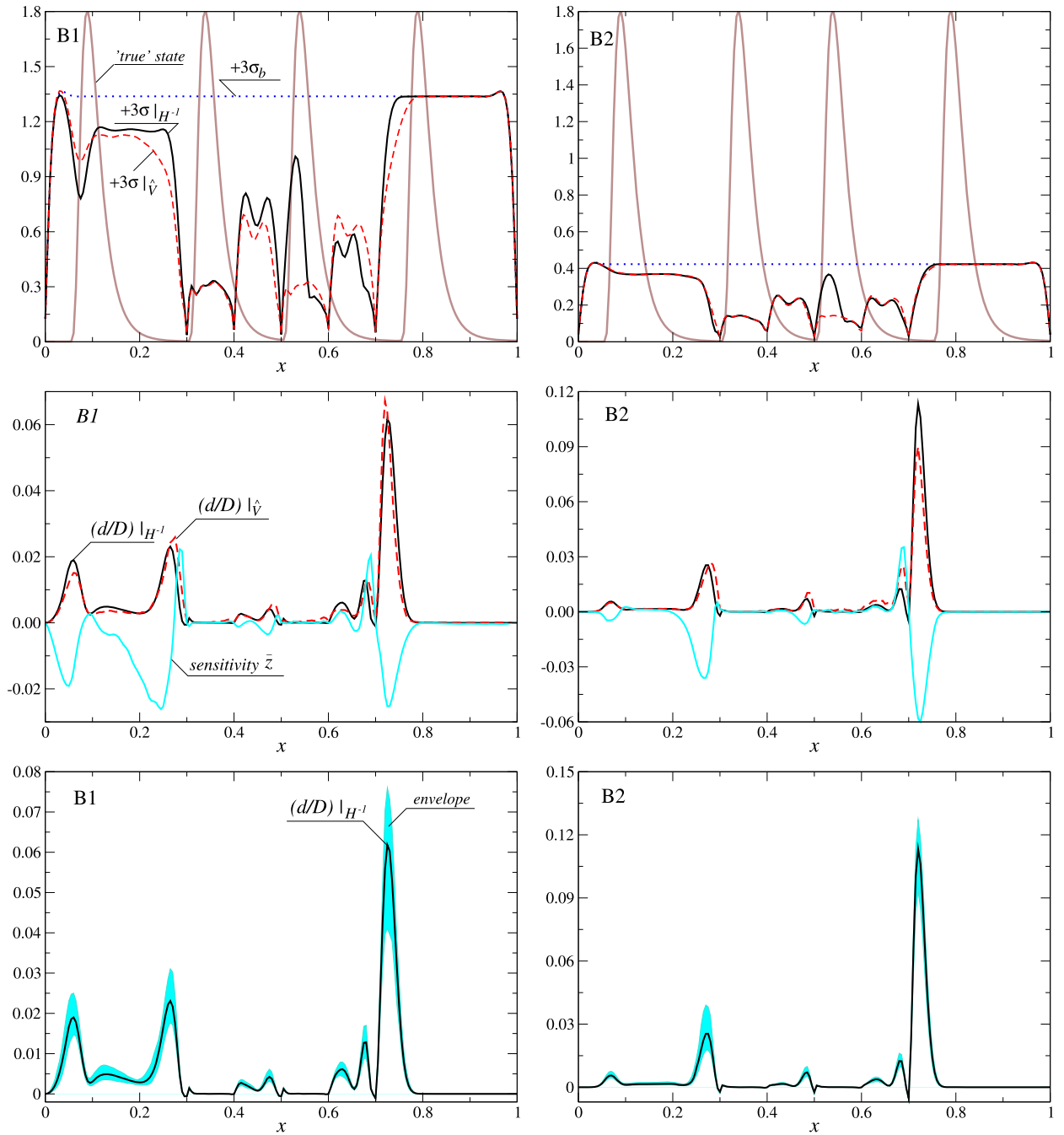


Fig. 5. Cases B1 and B2, see Section 10.1 for detail.

The following presentation pattern is accepted in Figs. 4–6:

(a) Upper panel shows:

- ‘true’ state  $u^t$ , the initial condition  $u = \varphi|_{t=0}$ , which is the unknown control in data assimilation problem;
- $+3\sigma_b$ , where  $\sigma_b^2 = \text{diag}(V_b)$ , i.e.  $3\sigma$ -envelope positive margin defined by the background error covariance  $V_b$ ;
- $+3\sigma|_{H^{-1}}$ , where  $\sigma^2|_{H^{-1}} = \text{diag}(H^{-1}(u^t))$ , i.e.  $3\sigma$ -envelope positive margin defined by the analysis error covariance approximated by the inverse Hessian  $H^{-1}(u^t)$ ;
- $+3\sigma|_{\hat{\psi}}$ , where  $\sigma^2|_{\hat{\psi}} = \text{diag}(\hat{V}(u^t))$ , i.e.  $3\sigma$ -envelope positive margin defined by the analysis error covariance approximated by the sample covariance  $\hat{V}(u^t)$ .

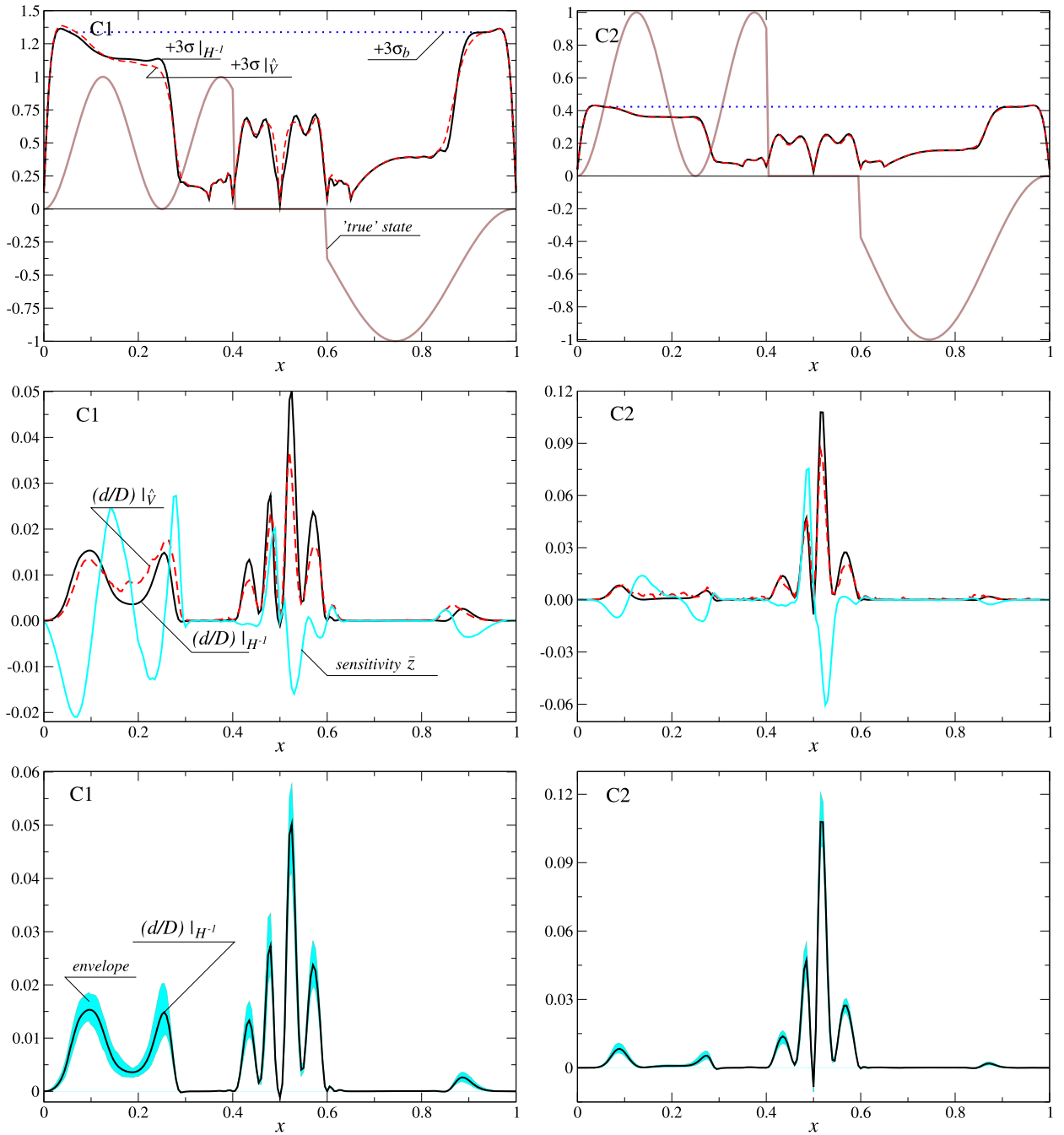


Fig. 6. Cases C1 and C2, see Section 10.1 for detail.

(b) Mid panel shows for  $i = 1, \dots, n$ :

- $(d_i/D)|_{\hat{v}}$ , scaled decomposition computed at  $Q|_{\hat{v}(u^t)}$  in (6.2), in solid line;
- $(d_i/D)|_{H^{-1}}$ , scaled decomposition computed at  $Q|_{H^{-1}(u^t)}$  in (6.2), in dashed line;
- $\bar{z}_i$ , sensitivity defined by (8.3)–(8.4), in pale solid line.

(c) Lower panel shows for  $i = 1, \dots, n$ :

- $(d_i/D)|_{H^{-1}}$ , scaled decomposition computed at  $Q|_{H^{-1}(u^t)}$  in (6.2);
- $\pm 1\sigma$ -envelope of  $(d_i/D)|_{H^{-1}}$ .



## 10.2. Discussion

Let us look at Table 1. First of all, we notice that  $E[\theta(v, \bar{u})]$  is significantly larger than the critical value  $D^*$  in cases A1, B1 and C1. This indicates that the optimal solutions are not gauss-verifiable on the whole in these cases. However, in cases A2, B2 and C2,  $E[\theta(v, \bar{u})]$  is less than  $D^*$  (or not too far from it in case A2), which means that the solutions are gauss-verifiable on the whole. Let us recall that the difference between sub-cases '1' and '2' is due to a different magnitude of the background variance: in sub-cases '2' it is ten times smaller than in sub-cases '1'. Subsequently, in sub-cases '2' the confidence region  $\Omega(w^*)$  associated to the analysis pdf  $\rho_a(u, u^t)$  is much narrower. For the gaussian approximation of the analysis pdf the confidence region is represented by the  $n$ -dimensional ellipsoid with the  $i$ -axis half-length equal to  $\alpha\sigma_i$ . The  $3\sigma$ -envelopes for sub-cases '1' and '2' can be compared in Figs. 4–6, upper panels. Since in sub-cases '2' the solution neighborhood around  $\bar{u}$  is smaller, the errors around  $\bar{u}$  are smaller and their evolution is much better approximated by the tangent linear model. As a result, the gaussian properties of the estimator are not critically distorted.

In order to see the relationship between the CM and the classical test statistic for multivariate normality we compare  $E[\theta(v, \bar{u})]$  to  $(S_{JB} - n)/2$ , and the latter to  $D^*$ . We notice that  $(S_{JB} - n)/2$  is significantly larger than the critical value  $D^*$  in cases A1, B1 and C1. This indicates that the gaussianity of the estimator is seriously distorted in these cases. We also notice that  $(S_{JB} - n)/2$  in cases A2, B2 and C2 is smaller than in cases A1, B1 and C1, but only in case B2 one can assume that the estimator is gaussian. In cases A2 and C2 the value  $(S_{JB} - n)/2$  is still large enough to conclude that the estimator is not gaussian, yet the solutions are gauss-verifiable, according to  $E[\theta(v, \bar{u})]$ . This is a clear sign that the gauss-verifiability, though related to the gaussianity of the estimator, is a different property.

Comparing  $E[\theta(v, \bar{u})]$  to  $D|_{\hat{\gamma}}$  one can conclude that the latter approximates the former reasonably well. Let us remind that  $D|_{\hat{\gamma}}$  incorporates two "inevitable" approximations described in Section 5. Next, using  $D|_{H^{-1}}$  instead of  $D|_{\hat{\gamma}}$  seems to be working well in most cases. The largest approximation error is observed in case A1, however it cannot alter our previous conclusions on this case.

Let us look now at Figs. 4–6, mid panels. The most important fact to be noticed is an existence of localized subsets/areas of the control vector  $u$  contributing the most weight into the total of  $D$ . It has been previously concluded (by considering  $D$ ) that the optimal solutions are not gauss-verifiable on the whole in cases A1, B1 and C1. However, the look of  $d_i/D$  suggests that this is only true for some parts of the control vector, whereas other parts can be locally gauss-verifiable in the sense that  $d_i/D < d^*/D$ . This is clearly demonstrated in case A1, where the non-verifiable subsets are located at  $x \in [0.46, 0.5]$  and  $x \in [0.56, 0.58]$ . On the other hand, in cases A2, B2 and C2 the conclusion has been made that the solutions are gauss-verifiable on the whole. Yet, in case A2 a subset at  $x \in [0.46, 0.50]$  may not be gauss-verifiable because, locally,  $d_i/D \gg d^*/D$  ( $d^*$  is given by (7.2), for  $\alpha = 3$ ,  $n = 201$  and  $D|_{H^{-1}} = 46.21$ , thus we obtain  $d^*/D = 3.25 \times 10^{-3}$ , see Table 1). These examples demonstrate the importance of computing  $d_i$ , alongside with the integral measure  $D$ . Looking at Fig. 4, mid panel, we notice that using  $D|_{H^{-1}}$  instead of  $D|_{\hat{\gamma}}$  may cause a significant structural error in  $d_i/D$ . In particular, as a result of this approximation the pick around  $x \approx 0.43$  is completely lost, both in cases A1 and A2.

In Figs. 4–6, mid panels, the sensitivities  $\bar{z}_i$  are presented for comparison. We observe that the non-verifiable subsets/areas of the control vector generally match the areas of a large sensitivity magnitude. The most significant mismatch can be found in case C1 (Fig. 6, mid panel), however, in case C2, i.e. for smaller errors, the match is reasonably good again. The latter points out the connection between the sensitivities  $\bar{z}_i$  and the decomposition  $d_i/D$ , though no obvious mathematical relationship has been established so far. This is an interesting issue to be further investigated.

It has been already mentioned that the numbers in Table 1 and all graphs in Figs. 4–6, mid panels, represent the asymptotic values of the corresponding quantities, i.e. the values obtained with the large ensemble ( $K = 2500$ ). In practice, the decomposition  $(d_i/D)|_{H^{-1}}$  is likely to be computed on a basis of a very small ensemble. Thus, in Figs. 4–6, lower panels, we present both the asymptotic  $(d_i/D)|_{H^{-1}}$  and its envelope which includes 70% of all  $d_i/D$  computed with small ensembles ( $K = 50$ ). The main conclusion to be made from these graphs is that  $(d_i/D)|_{H^{-1}}$  is fairly stable to the ensemble size and, therefore, its estimates obtained on a basis of small ensembles are reliable. Let us note that the pdf of  $d_i/D$  is far from the gaussian, so the envelope should not be interpreted as  $1\sigma$ -envelope of the Gaussian pdf.

## 11. Conclusions

The optimal solution in variational DA is an approximation of the true state of nature at a given time instant. It is an important practical task to access the accuracy of this approximation, however the deterministic error estimates are usually very crude. More useful error estimates are defined in the probabilistic framework. Particularly in the gaussian context, errors can be characterized by the confidence regions based on the analysis error covariance matrix.

Technical difficulties in computing this covariance are related to the high dimensionality of the state vector, whereas a fundamental difficulty is related to the nonlinearity of the model equations and of the observation operator. Nonlinearity distorts the gaussian properties of the variational DA estimator and, as a result, the constructed gaussian confidence regions may render a totally wrong error characterization. In this case we say that the optimal solution is not gauss-verifiable. Since the distortion of the gaussian properties is due to nonlinearity, one may expect the loss of gauss-verifiability taking place *locally*, in and around the spatial areas where the nonlinear phenomena are particularly strong (such as shock waves, vortices, etc.). In other words, there may exist non-verifiable localized subsets of the state vector, while for the rest of the

state vector the gaussian description of the error remains useful. This paper develops a measure of this usefulness on a basis of which the non-verifiable subsets can be detected.

At a glance, to assess deviation from the gaussianity one should use the classical test statistics for multivariate normality. However, these statistics do not measure gauss-verifiability directly and usually require large ensembles/samples of optimal solutions for practical implementation, which is not feasible for the models in mind. In this paper we introduce a new statistic called the coexistence measure (CM). This statistic can also be considered as a 'global' test statistic for multivariate normality. We suggest a method for its decomposition into the sum of (predominantly) positive components associated to the elements of the control vector. The subsets contributing the most weight into the total value of the CM can be considered as non-verifiable. The CM and its decomposition is feasible to compute because: (a) it can be evaluated on a basis of a very small ensemble of optimal solutions; (b) most computations and storage are implemented in a matrix-free form. This is achieved by approximating the sample covariance, which has deficient rank, by the inverse of the Hessian. The latter is defined by the successive solution of the tangent linear and adjoint models.

In numerical experiments the CM and its decomposition have been tested. We notice that the integral measures not always provide a sufficiently clear insight on the nature and extent of non-gaussianity in variational DA systems. On the contrary, the CM decomposition offers a delicate tool for analysis of gauss-verifiability. Moreover, computing CM is fairly reliable with sample-deficient ensembles because of exploiting information provided by the Hessians. For example, such an ensemble could be a natural outcome of the 'ensemble 4D-Var'. Let us finally mention that even though our focus here is on DA for geophysical applications, very similar problems (inverse problems for distributed dynamical systems) arise in many other areas of science and engineering. For these problems our results are directly applicable.

## Acknowledgements

The first author acknowledges the funding through the UK Natural Environment Research Council (NERC grant NE/J018201/1), the co-author acknowledges the support by the Russian Science Foundation (project 14-11-00609).

## References

- [1] T. Amemiya, Non-linear regression models, in: Z. Griliches, M.D. Intriligator (Eds.), *Handbook of Econometrics*, vol. 1, North-Holland Publishing Company, Amsterdam, 1983, pp. 333–389, Chapter 6.
- [2] G.E.B. Archer, A. Saltelli, I.M. Sobol', Sensitivity measures, ANOVA-like techniques and the use of bootstrap, *J. Stat. Comput. Simul.* 58 (1997) 99–120.
- [3] M. Bocquet, C.A. Pires, L. Wu, Beyond Gaussian statistical modeling in geophysical data assimilation, *Mon. Weather Rev.* (2010) 2997–3023.
- [4] G. Chavent, Identification of distributed parameter systems: about the output least square method, its implementation, and identifiability, in: *Proc. of 5-th IFAC Symp. on Identification and System Parameter Estimation*, Pergamon Press, Oxford, 1979, pp. 85–97.
- [5] V. Chew, Confidence, prediction, and tolerance regions for the multivariate normal distribution, *J. Am. Stat. Assoc.* 61 (315) (1966) 605–617.
- [6] P. Courtier, J.N. Thépaut, A. Hollingsworth, A strategy for operational implementation of 4D-Var, using an incremental approach, *Q. J. R. Meteorol. Soc.* 120 (1994) 1367–1388.
- [7] M.C. Delfour, S.K. Mitter, Controllability and observability for infinite-dimensional systems, *SIAM J. Control* 10 (1972) 329–333.
- [8] H.O. Fattorini, On complete controllability of linear systems, *J. Differ. Equ.* 3 (1967) 391–402.
- [9] M. Fisher, P. Courtier, Estimating the covariance matrices of analysis and forecast error in variational data assimilation, ECMWF Research Department, 1995, Techn. Memo. 220.
- [10] S. Fučík, A. Kufner, *Nonlinear Differential Equations*, Elsevier, Amsterdam, 1980.
- [11] I. Gejadze, F.-X. Le Dimet, V. Shutyaev, On analysis error covariances in variational data assimilation, *SIAM J. Sci. Comput.* 30 (4) (2008) 1847–1874.
- [12] I. Gejadze, F.-X. Le Dimet, V. Shutyaev, On optimal solution error covariances in variational data assimilation problems, *J. Comput. Phys.* 229 (2010) 2159–2178.
- [13] I. Gejadze, F.-X. Le Dimet, V. Shutyaev, Computation of the optimal solution error covariance in variational data assimilation problems with nonlinear dynamics, *J. Comput. Phys.* 230 (2011) 7923–7943.
- [14] I. Gejadze, V. Shutyaev, F.-X. Le Dimet, Analysis error covariance versus posterior covariance in variational data assimilation, *Q. J. R. Meteorol. Soc.* 138 (2012) 1–16.
- [15] R. Glowinski, J.L. Lions, Exact and approximate controllability for distributed parameter systems, *Acta Numer.* (1994) 269–378.
- [16] R.E. Goodson, M.P. Polis, Parameter identification in distributed systems: a synthesizing overview, *Proc. IEEE* 64 (1) (1976).
- [17] L. Hascoët, V. Pascual, TAPENADE 2.1 user's guide, INRIA Technical Report, 2004, No. 0300, 78 pp.
- [18] N. Henze, Invariant tests for multivariate normality: a critical review, *Stat. Pap.* 43 (2002) 467–506.
- [19] C.M. Jarque, A.K. Bera, A test for normality of observations and regression residuals, *Int. Stat. Rev.* 55 (2) (1987) 163–172.
- [20] R.I. Jennrich, Asymptotic properties of nonlinear least square estimation, *Ann. Math. Stat.* 40 (1969) 633–643.
- [21] S.J. Julier, J.K. Uhlmann, A new extension of the Kalman filter to nonlinear systems, in: *Proceedings of AeroSense: The 11th International Symposium on Aerospace/Defence Sensing, Simulation and Controls*, 1997.
- [22] R.E. Kalman, Contributions to the theory of optimal control, in: *Proceedings of the Mexico City Conference on Ordinary Differential Equations*, 1959, in: *Bol. Soc. Mat. Mexicana*, 1960, p. 102.
- [23] R.E. Kalman, Y.C. Ho, K.S. Narendra, Controllability of linear dynamical systems, *Contrib. Differ. Equ.* 1 (2) (1963) 189–213.
- [24] R.E. Kalman, R.S. Bucy, New results in linear filtering and prediction theory, *Trans. ASME, Ser. D, J. Basic Eng.* 83 (1961) 95.
- [25] S. Kitamura, S. Nakagiri, Identifiability of spatially-varying and constant parameters in distributed systems of parabolic type, *SIAM J. Control Optim.* 15 (5) (1977) 217–241.
- [26] N.N. Krasovski, *Theory of Control of Motion*, Nauka, Moscow, 1968.
- [27] C. Kravaris, H. Seinfeld, Identification of parameters in distributed parameter systems by regularization, *SIAM J. Control Optim.* 23 (2) (1985) 785–802.
- [28] C.S. Kubrusly, Distributed parameter system identification, a survey, *Int. J. Control* 26 (4) (1977) 509–535.
- [29] S. Kullback, *Information Theory and Statistics*, John Wiley and Sons, New York, 1959.
- [30] A.B. Kurzhanski, *Control and Observation Under Uncertainty Conditions*, Nauka, Moscow, 1977.
- [31] A.B. Kurzhanski, A.Yu. Khapalov, On state estimation problem for distributed systems, in: *Lect. Notes Control Inf. Sci.*, vol. 83, 1983, pp. 102–113.

- [32] F.X. Le Dimet, O. Talagrand, Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects, *Tellus* 38A (1986) 97–110.
- [33] R.B. Lehoucq, D.C. Sorensen, C. Yang, *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, SIAM, New York, 1998.
- [34] J.L. Lions, *Contrôle optimal des systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.
- [35] J.L. Lions, On controllability of distributed systems, *Proc. Natl. Acad. Sci. USA* 94 (1997) 4828–4835.
- [36] A.C. Lorenc, Analysis methods for numerical weather prediction, *Q. J. R. Meteorol. Soc.* 112 (1986) 1177–1194.
- [37] L. Markus, Controllability and observability, in: E. Caianiello (Ed.), *Functional Analysis and Optimization*, Acad. Press, New York, 1966, pp. 133–143.
- [38] I.M. Navon, Practical and theoretical aspects of adjoint parameter estimation and identifiability in meteorology and oceanography, *Dyn. Atmos. Ocean.* (1997) 55–79.
- [39] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, *Numerical Recipes in Fortran 77: The Art of Scientific Computing*, second edition, Cambridge Press, 1992.
- [40] S.V. Patankar, *Numerical Heat Transfer and Fluid Flow*, Hemisphere Publishing Corporation, New York, 1980.
- [41] M.S. Srivastava, A measure of skewness and kurtosis and a graphical method for assessing multivariate normality, *Stat. Probab. Lett.* 2 (1984) 263–267.
- [42] A.M. Stuart, Inverse problems: a Bayesian perspective, *Acta Numer.* 19 (2010) 451–559.
- [43] A. Tarantola, *Inverse Problems Theory: Methods for Data Fitting and Model Parameter Estimation*, Elsevier, New York, 1987.
- [44] W.C. Thacker, The role of the Hessian matrix in fitting models to measurements, *J. Geophys. Res.* 94 (C5) (1989) 6177–6196.
- [45] R. Triggiani, Controllability and observability in Banach space with bounded operators, *SIAM J. Control* 13 (2) (1975) 462–491.