Tellus
SERIES A
DYANAMIC
METEOROLOGY
AND OCEANOGRAPHY
PUBLISHED BY THE INTERNATIONAL METEOROLOGICAL INSTITUTE IN STOCKHOLM

Stochastic parameterization identification using ensemble Kalman filtering combined with maximum likelihood methods

By MANUEL PULIDO[1,2*], PIERRE TANDEO[3], MARC BOCQUET[4], ALBERTO CARRASSI[5] and MAGDALENA LUCINI[6], [1]*Department of Physics, FaCENA, Universidad Nacional del Nordeste, Corrientes, Argentina;* [2]*IMIT, IFAECI, CNRS-CONICET, Buenos Aires, Argentina;* [3]*IMT Atlantique, Lab-STICC, UBL, Brest, France;* [4]*CEREA, Joint Laboratory École des Ponts ParisTech and EDF R&D, Université Paris-Est, Champs-sur-Marne, France;* [5]*Nansen Environmental and Remote Sensing Center, Bergen, Norway;* [6]*Department of Mathematics, FaCENA, Universidad Nacional del Nordeste and CONICET, Corrientes, Argentina*

(Manuscript received 21 September 2017; in final form 9 February 2018)

ABSTRACT

For modelling geophysical systems, large-scale processes are described through a set of coarse-grained dynamical equations while small-scale processes are represented via parameterizations. This work proposes a method for identifying the best possible stochastic parameterization from noisy data. State-of-the-art sequential estimation methods such as Kalman and particle filters do not achieve this goal successfully because both suffer from the collapse of the posterior distribution of the parameters. To overcome this intrinsic limitation, we propose two statistical learning methods. They are based on the combination of the ensemble Kalman filter (EnKF) with either the expectation–maximization (EM) or the Newton–Raphson (NR) used to maximize a likelihood associated to the parameters to be estimated. The EM and NR are applied primarily in the statistics and machine learning communities and are brought here in the context of data assimilation for the geosciences. The methods are derived using a Bayesian approach for a hidden Markov model and they are applied to infer deterministic and stochastic physical parameters from noisy observations in coarse-grained dynamical models. Numerical experiments are conducted using the Lorenz-96 dynamical system with one and two scales as a proof of concept. The imperfect coarse-grained model is modelled through a one-scale Lorenz-96 system in which a stochastic parameterization is incorporated to represent the small-scale dynamics. The algorithms are able to identify the optimal stochastic parameterization with good accuracy under moderate observational noise. The proposed EnKF-EM and EnKF-NR are promising efficient statistical learning methods for developing stochastic parameterizations in high-dimensional geophysical models.

Keywords: parameter estimation, model error estimation, stochastic parameterization, expectation–maximization algorithm

1. Introduction

The statistical combination of observations of a dynamical model with a priori information of physical laws allows the estimation of the full state of the model even when it is only partially observed. This is the main aim of data assimilation (Kalnay, 2002). One common challenge of evolving multi-scale systems in applications ranging from meteorology, oceanography, hydrology and space physics to biochemistry and biological systems is the presence of parameters that do not rely on known physical constants so that their values are unknown and unconstrained. Data assimilation techniques can also be formulated to estimate these model parameters from observations (Jazwinski et al., 1970; Wikle and Berliner, 2007).

There are several multi-scale physical systems which are modelled through coarse-grained equations. The most paradigmatic cases being climate models (Stensrud, 2009), large-eddy simulations of turbulent flows (Mason and Thomson, 1992) and electron fluxes in the radiation belts (Kondrashov et al., 2011). These imperfect models need to include subgrid-scale effects through physical parameterizations (Nicolis, 2004). In the last years, stochastic physical parameterizations have been incorporated in weather forecast and climate models (Palmer, 2001; Christensen et al., 2015; Shutts, 2015). They are called stochastic

*Corresponding author. e-mail: pulido@unne.edu.ar

Citation: Tellus A: 2018, **70**, 1442099, https://doi.org/10.1080/16000870.2018.1442099

parameterizations because they represent stochastically a process that is not explicitly resolved in the model, even when the unresolved process may not be itself stochastic. The forecast skill of ensemble forecast systems has been shown to improve with these stochastic parameterizations (Palmer, 2001; Christensen et al., 2015; Shutts, 2015). Deterministic integrations with models that include these parameterizations have also been shown to improve climate features (see e.g. Lott et al., 2012). In general, stochastic parameterizations are expected to improve coarse-grained models of multi-scale physical systems (Katsoulakis et al., 2003; Majda and Gershgorin, 2011). However, the functional form of the schemes and their parameters, which represents small-scale effects, are unknown and must be inferred from observations. The development of automatic statistical learning techniques to identify an optimal stochastic parameterization and estimate its parameters is, therefore, highly desirable.

One standard methodology to estimate physical model parameters from observations in data assimilation techniques, such as the traditional Kalman filter, is to augment the state space with the parameters (Jazwinski et al., 1970). This methodology has also been implemented in the ensemble-based Kalman filter (see e.g. Anderson, 2001). The parameters are constrained through their correlations with the observed variables. However, three challenges are posed for parameter estimation in EnKFs. Firstly, parameter probability distributions are in general non-Gaussian, even though Kalman-based filters rely on the Gaussian assumption. Secondly, the estimation of global parameters is theoretically incompatible with the use of domain localization (Bellsky et al., 2014), which is very often employed to implement the EnKF in high-dimensional systems. Thirdly, the parameters are usually assumed to be governed by persistence so that their impact on the augmented error covariance matrix diminishes with time (Ruiz et al., 2013a).

The collapse of the parameter posterior distribution found in both ensemble Kalman filters (Delsole and Yang, 2010; Ruiz et al., 2013a, 2013b; Santitissadeekorn and Jones, 2015) and particle filters (West and Liu, 2001) is a major contention point when one is interested in estimating stochastic parameters of non-linear dynamical models. Hereafter, we refer as stochastic parameters to those that define the covariance of a Gaussian stochastic process (Delsole and Yang, 2010). In other words, the sequential filters are, in principle, able to estimate deterministic physical parameters, the mean of the parameter posterior distribution, through the augmented state-space procedure, but they are unable to estimate stochastic parameters of the model, because of the collapse of the corresponding posterior distribution. Using the Kalman filter with the augmentation method, Delsole and Yang (2010) proved analytically the collapse of the parameter covariance in a first-order autoregressive model. They proposed a generalized maximum likelihood estimation using an approximate sequential method to estimate stochastic parameters. Carrassi and Vannitsem (2011) derived the evolution of the augmented error covariance in the extended Kalman filter using a quadratic in time approximation that mitigates the collapse of the parameter error covariance. Santitissadeekorn and Jones (2015) proposed a particle filter blended with an ensemble Kalman filter and use a random walk model for the parameters. This technique was able to estimate stochastic parameters in the first-order autoregressive model, but a tunable parameter in the random walk model needs to be introduced.

The expectation–maximization (EM) algorithm (Dempster et al., 1977; Bishop, 2006) is a widely used method to maximize the likelihood function in a broad spectrum of applications. One of the advantages of the EM algorithm is that its implementation is rather straigthforward. Wu (1983) showed that if the likelihood is smooth and unimodal, the EM algorithm converges to the unique maximum likelihood estimate. Accelerations of the EM algorithm have been proposed for its use in machine learning (Neal and Hinton, 1999). Recently, it was used in an application with a highly non-linear observation operator (Tandeo et al., 2015). The EM algorithm was able to estimate subgrid-scale parameters with good accuracy while standard ensemble Kalman filter techniques failed. It has also been applied to the Lorenz-63 system to estimate model error covariance (Dreano et al., 2017).

In this work, we combine the ensemble Kalman filter (Evensen, 1994; Evensen, 2003) with maximum likelihood estimators for stochastic parameterization identification. Two maximum likelihood estimators are evaluated: the EM (Dempster et al., 1977; Bishop, 2006) and the Newton-Raphson algorithm (Cappé et al., 2005). The derivation of the techniques is explained in detail and simple terms so that readers that are not from those communities can understand the basis of the methodologies, how they can be combined, and hopefully foresee potential applications in other geophysical systems. The learning statistical techniques are suitable to infer the functional form and the parameter values of stochastic parameterizations in chaotic spatio-temporal dynamical systems. They are evaluated here on a two-scale spatially extended chaotic dynamical system (Lorenz, 1996) to estimate deterministic physical parameters, together with additive and multiplicative stochastic parameters. Pulido et al. (2016) evaluated methods based on the EnKF alone to estimate subgrid-scale parameters in a two-scale system: they showed that an offline estimation method is able to recover the functional form of the subgrid-scale parameterization, but none of the methods was able to estimate the stochastic component of the subgrid-scale effects. In the present work, the results show that the NR and EM techniques are able to uncover the functional form of the subgrid-scale parameterization while successfully determining the stochastic parameters of the representation of subgrid-scale effects.

This work is organized as follows. Section 2 briefly introduces the EM algorithm and derives the marginal likelihood of the data using a Bayesian perspective. The implementation of the EM and NR likelihood maximization algorithms in the context of data assimilation using the ensemble Kalman filter is also discussed. Section 3 describes the experiments which are based on the

one- and two-scale Lorenz-96 systems. The former includes simple deterministic and stochastic parameterizations to represent the effects of the smaller scale to mimic the two-scale Lorenz-96 system. Section 4 focuses on the results: Section 4.1 discusses the experiments for the estimation of model noise. Section 4.2 shows the results of the estimation of deterministic and stochastic parameters in a perfect-model scenario. Section 4.3 shows the estimation experiments for an imperfect model. The conclusions are drawn in Section 5.

2. Methodology

2.1. Hidden Markov model

A hidden Markov model is defined by a stochastic non-linear dynamical model $\mathcal{M}$ that evolves in time the hidden variables $\mathbf{x}_{k-1} \in \mathbb{R}^N$, according to,

$$\mathbf{x}_k = \mathcal{M}_{\boldsymbol{\Omega}}(\mathbf{x}_{k-1}) + \boldsymbol{\eta}_k, \tag{1}$$

where k stands for the time index. The dynamical model $\mathcal{M}$ depends on a set of deterministic and stochastic physical parameters denoted by $\boldsymbol{\Omega}$. We assume an additive random model error, $\boldsymbol{\eta}_k$, with covariance matrix $\mathbf{Q}_k = \mathcal{E}\left(\boldsymbol{\eta}_k \boldsymbol{\eta}_k^{\mathrm{T}}\right)$. The notation $\mathcal{E}\,()$ stands for the expectation operator, $\mathcal{E}\,[f(x)] \equiv \int f(x)p(x)\mathrm{d}x$ with p being the probability density function of the underlying process X.

The observations at time k, $\mathbf{y}_k \in \mathbb{R}^M$, are related to the hidden variables through the observational operator $\mathcal{H}$,

$$\mathbf{y}_k = \mathcal{H}(\mathbf{x}_k) + \boldsymbol{\epsilon}_k, \tag{2}$$

where $\boldsymbol{\epsilon}_k$ is an additive random observation error with observation error covariance matrix $\mathbf{R}_k = \mathcal{E}\left(\boldsymbol{\epsilon}_k \boldsymbol{\epsilon}_k^{\mathrm{T}}\right)$.

Our estimation problem: *Given a set of observation vectors distributed in time,* $\{\mathbf{y}_k, k = 1, \ldots, K\}$, *a nonlinear stochastic dynamical model,* $\mathcal{M}$, *and a nonlinear observation operator,* $\mathcal{H}$, *we want to estimate the initial prior distribution* $p(\mathbf{x}_0)$, *the observation error covariance* $\mathbf{R}_k$, *the model error covariance* $\mathbf{Q}_k$, *and deterministic and stochastic physical parameters* $\boldsymbol{\Omega}$ *of* $\mathcal{M}$.

In the EM literature, the term 'parameters' is used for all the parameters of the likelihood function including the moments of the statistical distributions. Here, the parameters of the likelihood function are referred more specifically to as *likelihood parameters*. The likelihood parameters may include the deterministic and stochastic physical parameters, the observation error and the model error covariances and the first two moments of the initial prior distribution.

The estimation method we derived is based on maximum likelihood estimation. Given a set of independent and identically distributed (iid) observations from a probability density function represented by $p(\mathbf{y}_{1:K}|\boldsymbol{\theta})$, we seek to maximize the likelihood function $L(\mathbf{y}_{1:K}; \boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$. We denote $\{\mathbf{y}_1, \ldots, \mathbf{y}_K\}$ by $\mathbf{y}_{1:K}$ and the set of likelihood parameters to be estimated by $\boldsymbol{\theta}$: the deterministic and stochastic physical parameters $\boldsymbol{\Omega}$ of the dynamical model $\mathcal{M}$ as well as observation error covariances $\mathbf{R}_k$, model error covariances $\mathbf{Q}_k$ and the mean $\overline{\mathbf{x}}_0$ and covariance $\mathbf{P}_0$ of the initial prior distribution $p(\mathbf{x}_0)$. In practical applications, the statistical moments $\mathbf{R}_k$, $\mathbf{Q}_k$ and $\mathbf{P}_0$ are usually poorly constrained. It may thus be convenient to estimate them jointly with the physical parameters. The dynamical model is assumed to be non-linear and to include stochastic processes represented by some of the physical parameters in $\boldsymbol{\Omega}$.

The estimation technique used in this work is a batch method: a set of observations taken along a time interval is used to estimate the model state trajectory that is closest to them, considering measurement and model errors with a least-square criterion to be established below. The simultaneous use of observations distributed in time is essential to capture the interplay of the several likelihood parameters included in the estimation problem. The required minimal length K for the observation window is evaluated in the numerical experiments. The estimation technique may be applied in successive K-windows. For stochastic parameterizations in which the parameters are sensitive to processes of different time scales, a batch method may also be required to capture the sensitivity to slow processes.

2.2. Expectation-maximization algorithm

The EM algorithm maximizes the log-likelihood of observations as a function of the likelihood parameters $\boldsymbol{\theta}$ in the presence of a hidden state $\mathbf{x}_{0:K}$,[1]

$$l(\boldsymbol{\theta}) = \ln L(\mathbf{y}_{1:K}; \boldsymbol{\theta}) = \ln \int p(\mathbf{x}_{0:K}, \mathbf{y}_{1:K}; \boldsymbol{\theta}) \mathrm{d}\mathbf{x}_{0:K}. \tag{3}$$

An analytical form for the log-likelihood function, (3), can be obtained only in a few ideal cases. Furthermore, the numerical evaluation of (3) may involve high-dimensional integration of the complete likelihood (integrand of (3)). Given an initial guess of the likelihood parameters $\boldsymbol{\theta}$, the EM algorithm maximizes the log-likelihood of observations as a function of the likelihood parameters in successive iterations without the need to evaluate the complete likelihood.

2.2.1. The principles. Let us introduce in the integral (3) an arbitrary probability density function of the hidden state, $q(\mathbf{x}_{0:K})$,

$$l(\boldsymbol{\theta}) = \ln \int q(\mathbf{x}_{0:K}) \frac{p(\mathbf{x}_{0:K}, \mathbf{y}_{1:K}; \boldsymbol{\theta})}{q(\mathbf{x}_{0:K})} \mathrm{d}\mathbf{x}_{0:K}. \tag{4}$$

We assume that the support of $q(\mathbf{x}_{0:K})$ contains that of $p(\mathbf{x}_{0:K}, \mathbf{y}_{1:K}; \boldsymbol{\theta})$. The density, $q(\mathbf{x}_{0:K})$ may be thought, in particular, as a function of a set of fixed likelihood parameters $\boldsymbol{\theta}'$, $q(\mathbf{x}_{0:K}; \boldsymbol{\theta}')$. Using Jensen inequality, a lower bound for the log-likelihood is obtained,

$$l(\boldsymbol{\theta}) \geq \int q(\mathbf{x}_{0:K}) \ln\left(\frac{p(\mathbf{x}_{0:K}, \mathbf{y}_{1:K}; \boldsymbol{\theta})}{q(\mathbf{x}_{0:K})}\right) \mathrm{d}\mathbf{x}_{0:K} \equiv \mathcal{Q}(q, \theta). \tag{5}$$

If we choose $q(\mathbf{x}_{0:K}) = p(\mathbf{x}_{0:K}|\mathbf{y}_{1:K}; \boldsymbol{\theta}')$, the equality is satisfied in (5), therefore $p(\mathbf{x}_{0:K}|\mathbf{y}_{1:K}; \boldsymbol{\theta}')$ is an upper bound to $\mathcal{Q}$ and so it is the q function that maximises $\mathcal{Q}(q, \boldsymbol{\theta})$. The intermediate function $\mathcal{Q}(q, \boldsymbol{\theta})$ may be interpreted physically as the free energy of the system, so that $\mathbf{x}$ are interpreted as the physical states and the energy is the joint density (Neal and Hinton, 1999). Rewriting the joint density in (5) as a function of the conditional density, $p(\mathbf{x}_{0:K}, \mathbf{y}_{1:K}; \boldsymbol{\theta}) = p(\mathbf{x}_{0:K}|\mathbf{y}_{1:K}; \boldsymbol{\theta}) L(\mathbf{y}_{1:K}; \boldsymbol{\theta})$, the intermediate function may be related to the Kullback–Leibler divergence,

$$\mathcal{Q}(q, \theta) = -D_{KL}(q|p(\mathbf{x}_{0:K}|\mathbf{y}_{1:K}; \boldsymbol{\theta})) + l(\theta), \tag{6}$$

where $D_{KL}(q|p) \equiv \int q \ln\left(\frac{q}{p}\right) \mathrm{d}x$ is a positive definite function and $D_{KL}(q|p) = 0$ iff $q = p$. From (6), using the properties of the Kullback–Leibler divergence, it is clear that the upper bound of $\mathcal{Q}$ is obtained for $q = p(\mathbf{x}_{0:K}|\mathbf{y}_{1:K}; \boldsymbol{\theta})$.

From (5), we see that if we maximize $\mathcal{Q}(q, \boldsymbol{\theta})$ over $\boldsymbol{\theta}$, we find a lower bound for $l(\boldsymbol{\theta})$. The idea of the EM algorithm is to first find the probability density function q that maximizes $\mathcal{Q}$, the conditional probability of the hidden state given the observations, and then to determine the parameter $\boldsymbol{\theta}$ that maximizes $\mathcal{Q}$. Hence, the EM algorithm encompasses the following steps:

Expectation: Determine the distribution q that maximizes $\mathcal{Q}$. This function is easily shown to be $q^* = p(\mathbf{x}_{0:K}|\mathbf{y}_{1:K}; \boldsymbol{\theta}')$ (see (5); Neal and Hinton 1999). The function q^* is the conditional probability of the hidden state given the observations. In practice, this is obtained by evaluating the conditional probability at $\boldsymbol{\theta}'$.

Maximization: Determine the likelihood parameters $\boldsymbol{\theta}^*$ that maximize $\mathcal{Q}(q^*, \boldsymbol{\theta})$ over $\boldsymbol{\theta}$. The new estimation of the likelihood parameters is denoted by $\boldsymbol{\theta}^*$ while the (fixed) previous estimation by $\boldsymbol{\theta}'$. The expectation step is a function of these old likelihood parameters $\boldsymbol{\theta}'$. The part of function $\mathcal{Q}$ to maximize is given by:

$$\int p(\mathbf{x}_{0:K}|\mathbf{y}_{1:K}; \boldsymbol{\theta}') \ln\left(p(\mathbf{x}_{0:K}, \mathbf{y}_{1:K}; \boldsymbol{\theta})\right) \mathrm{d}\mathbf{x}_{0:K} \equiv \mathcal{E}\left[\ln\left(p(\mathbf{x}_{0:K}, \mathbf{y}_{1:K}; \boldsymbol{\theta})\right)|\mathbf{y}_{1:K}\right], \tag{7}$$

where we use the notation $\mathcal{E}(f(x)|y) \equiv \int f(x) p(x|y) \mathrm{d}x$ (Jazwinski et al., 1970). While the function that we want to maximize is the log-likelihood, the intermediate function (7) to maximize in the EM algorithm is the expectation of the joint distribution *conditioned to the observations*.

2.2.2. Expectation-maximization for a hidden Markov model. The joint distribution of a hidden Markov model using the definition of the conditional probability distribution reads

$$p(\mathbf{x}_{0:K}, \mathbf{y}_{1:K}) = p(\mathbf{y}_{1:K}|\mathbf{x}_{0:K}) p(\mathbf{x}_{0:K}). \tag{8}$$

The model state probability density function can be expressed as a product of the transition density from t_k to t_{k+1} using the definition of the conditional probability distribution and the Markov property,

$$p(\mathbf{x}_{0:K}) = p(\mathbf{x}_0) \prod_{k=1}^{K} p(\mathbf{x}_k|\mathbf{x}_{k-1}). \tag{9}$$

The observations are mutually independent and are conditioned on the current state (see (2)) so that

$$p(\mathbf{y}_{1:K}|\mathbf{x}_{0:K}) = \prod_{k=1}^{K} p(\mathbf{y}_k|\mathbf{x}_k). \tag{10}$$

Then, replacing (9) and (10) in (8) yields

$$p(\mathbf{x}_{0:K}, \mathbf{y}_{1:K}) = p(\mathbf{x}_0) \prod_{k=1}^{K} p(\mathbf{x}_k|\mathbf{x}_{k-1}) p(\mathbf{y}_k|\mathbf{x}_k). \tag{11}$$

If we now assume a Gaussian hidden Markov model, and that the covariances $\mathbf{R}_k$ and $\mathbf{Q}_k$ are constant in time, the logarithm of the joint distribution (11) is then given by:

$$\begin{aligned}
&\ln(p(\mathbf{x}_{0:K}, \mathbf{y}_{1:K})) \\
&\quad = -\frac{(M+N)}{2} \ln(2\pi) - \frac{1}{2} \ln|\mathbf{P}_0| - \frac{1}{2} (\mathbf{x}_0 - \bar{\mathbf{x}}_0)^{\mathrm{T}} \mathbf{P}_0^{-1} \\
&\qquad \times (\mathbf{x}_0 - \bar{\mathbf{x}}_0) - \frac{K}{2} \ln|\mathbf{Q}| - \frac{1}{2} \sum_{k=1}^{K} (\mathbf{x}_k - \mathcal{M}(\mathbf{x}_{k-1}))^{\mathrm{T}} \\
&\qquad \times \mathbf{Q}^{-1} (\mathbf{x}_k - \mathcal{M}(\mathbf{x}_{k-1})) - \frac{K}{2} \ln|\mathbf{R}| \\
&\qquad - \frac{1}{2} \sum_{k=1}^{K} (\mathbf{y}_k - \mathcal{H}(\mathbf{x}_k))^{\mathrm{T}} \mathbf{R}^{-1} (\mathbf{y}_k - \mathcal{H}(\mathbf{x}_k)).
\end{aligned} \tag{12}$$

The Markov hypothesis implies that model error is not correlated in time. Otherwise, we would have cross terms in the model error summation of (12). The assumption of a Gaussian hidden Markov model is central to derive a closed analytical form for the

likelihood parameters that maximize the intermediate function. However, the dynamical model and observation operator may have non-linear dependencies so that the Gaussian assumption is not strictly held. We therefore consider an iterative method in which each step is an approximation. In general, the method will converge through successive approximations. For severe non-linear dependencies in the dynamical model, the existence of a single maximum in the log-likelihood is not guaranteed. In that case, the EM algorithm may converge to a local maximum. As suggested by Wu (1983), one way to avoid that the EM algorithm be trapped in a local maximum of the likelihood function is to apply the algorithm for different starting parameters. Then, the EM simulation with the highest likelihood is chosen and the corresponding estimated parameters. In practice, the stochastic nature of the likelihood function may contribute to avoid the EM algorithm gets stuck in a local maximum (as in stochastic optimization).

We consider (12) as a function of the likelihood parameters $\boldsymbol{\theta}$ in this Gaussian state-space model. In this way, given the known values of the observations the log-likelihood function in (3), is a function of the likelihood parameters, namely: $\bar{\mathbf{x}}_0$, $\mathbf{P}_0$, $\mathbf{Q}$, $\mathbf{R}$, and $\boldsymbol{\Omega}$, the physical parameters from $\mathcal{M}$.

In this Gaussian state-space model, the maximum of the intermediate function in the EM algorithm, (7), may be determined analytically from

$$\begin{aligned} 0 &= \nabla_{\boldsymbol{\theta}} \mathcal{E}\left[\ln\left(p(\mathbf{x}_{0:K}, \mathbf{y}_{1:K}; \boldsymbol{\theta})\right) | \mathbf{y}_{1:K}\right] \\ &= \int p(\mathbf{x}_{0:K}|\mathbf{y}_{1:K}; \boldsymbol{\theta}') \nabla_{\boldsymbol{\theta}} \ln(p(\mathbf{x}_{0:K}, \mathbf{y}_{1:K}; \boldsymbol{\theta}))\, \mathrm{d}\mathbf{x}_{0:K} \\ &= \mathcal{E}\left[\nabla_{\boldsymbol{\theta}} \ln\left(p(\mathbf{x}_{0:K}, \mathbf{y}_{1:K}; \boldsymbol{\theta})\right) | \mathbf{y}_{1:K}\right] \end{aligned} \tag{13}$$

Note that $\boldsymbol{\theta}'$ is fixed in (13). We only need to find the critical values of the likelihood parameters $\mathbf{Q}$ and $\mathbf{R}$. The physical parameters are appended to the state, so that their model error is included in $\mathbf{Q}$. The $\bar{\mathbf{x}}_0$, $\mathbf{P}_0$ are at the initial time so that they are obtained as an output of the smoother which gives a Gaussian approximation of $p(\mathbf{x}_k|\mathbf{y}_{1:K})$ with $k = 0, \ldots, K$. The smoother equations are shown in the Appendix 1.

Differentiating (12) with respect to $\mathbf{Q}$ and $\mathbf{R}$ and applying the expectation conditioned to the observations, we can determine the root of the condition, (13), which gives the maximum of the intermediate function. The value of the model error covariance, solution of (13), is

$$\mathbf{Q} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{E}\left(\left[\mathbf{x}_k - \mathcal{M}\left(\mathbf{x}_{k-1}\right)\right]\left[\mathbf{x}_k - \mathcal{M}\left(\mathbf{x}_{k-1}\right)\right]^{\mathrm{T}} | \mathbf{y}_{1:K}\right). \tag{14}$$

In the case of the observation error covariance, the solution is:

$$\mathbf{R} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{E}\left(\left[\mathbf{y}_k - \mathcal{H}\left(\mathbf{x}_k\right)\right]\left[\mathbf{y}_k - \mathcal{H}\left(\mathbf{x}_k\right)\right]^{\mathrm{T}} | \mathbf{y}_{1:K}\right). \tag{15}$$

Therefore, we can summarize the EM algorithm for a hidden Markov model as:

Expectation: The required set of expectations given the observations must be evaluated at $\boldsymbol{\theta}_i$, i being the iteration number, specifically, $\mathcal{E}\left(\mathbf{x}_k|\mathbf{y}_{1:K}\right)$, $\mathcal{E}\left(\mathbf{x}_k\mathbf{x}_k^{\mathrm{T}}|\mathbf{y}_{1:K}\right)$, etc. The outputs of a classical smoother are indeed $\mathcal{E}\left(\mathbf{x}_k|\mathbf{y}_{1:K}\right)$, $\mathcal{E}\left(\left(\mathbf{x}_k - \mathcal{E}\left(\mathbf{x}_k|\mathbf{y}_{1:K}\right)\right)\left(\mathbf{x}_k - \mathcal{E}\left(\mathbf{x}_k|\mathbf{y}_{1:K}\right)\right)^{\mathrm{T}}|\mathbf{y}_{1:K}\right)$ which fully characterize $p(\mathbf{x}_k|\mathbf{y}_{1:K})$ in the Gaussian case. Hence, this expectation step involves the application of a foward filter and a backward smoother.

Maximization: Since we assume Gaussian distributions, the optimal value of $\boldsymbol{\theta}_{i+1}$ can be determined analytically, which in our model are $\mathbf{Q}$ and $\mathbf{R}$, as derived in (14) and (15). These equations are evaluated using the expectations determined in the Expectation step.

The basic steps of this EM algorithm are depicted in Fig. 1a. In this work, we use an ensemble-based Gaussian filter, the ensemble transform Kalman filter (Hunt et al., 2007) and the Rauch–Tung–Striebel (RTS) smoother (Cosme et al., 2012; Raanes, 2016).[2] A short description of these methods is given in the Appendix. The empirical expectations are determined using the smoothed ensemble member states at t_k, $\mathbf{x}_m^s(t_k)$. For instance,

$$\mathcal{E}\left(\mathbf{x}_k\mathbf{x}_k^{\mathrm{T}}|\mathbf{y}_{1:K}\right) = \frac{1}{N_e} \sum_{m=1}^{N_e} \mathbf{x}_m^s(t_k)\mathbf{x}_m^s(t_k)^{\mathrm{T}}, \tag{16}$$

where N_e is the number of ensemble members. Then, using these empiral expectations $\mathbf{R}$ and/or $\mathbf{Q}$ are computed from (14) and/or (15).

The EM algorithm applied to a linear Gaussian state-space model using the Kalman filter was first proposed by Shumway and Stoffer (1982). Its approximation using an ensemble of draws (Monte Carlo EM) was proposed in Wei and Tanner (1990). It was later generalized with the extended Kalman filter and Gaussian kernels by Ghahramani and Roweis (1999). The use of the EnKF and the ensemble Kalman smoother permits the extension of the EM algorithm to non-linear high-dimensional dynamical models and non-linear observation operators.

2.3. Maximum likelihood estimation via Newton–Raphson

The EM algorithm is highly versatile and can be readily implemented. However, it requires the optimal value in the maximization step to be computed analytically which limits the range of its applications. If physical deterministic parameters of a non-linear model need to be estimated, an analytical expression for the optimal likelihood parameter values may not be available. Another approach to find an estimate of the likelihood parameters consists

in maximizing an approximation of the likelihood function $l(\boldsymbol{\theta})$ with respect to the parameters, (3). This maximization may be conducted using standard optimization methods (Cappé et al., 2005).

Following Carrassi et al. (2017), the observation probability density function can be decomposed into the product

$$p(\mathbf{y}_{1:K};\boldsymbol{\theta}) = \prod_{k=1}^{K} p(\mathbf{y}_k|\mathbf{y}_{1:k-1};\boldsymbol{\theta}), \tag{17}$$

with the convention $\mathbf{y}_{1:0} = \varnothing$. In the case of sequential application of NR maximization in successive K-windows, the a priori probability distribution $p(\mathbf{x}_0)$ can be taken from the previous estimation. For such a case, we leave implicit the conditioning in (17) on all the past observations, $p(\mathbf{y}_{1:K};\boldsymbol{\theta}) = p(\mathbf{y}_{1:K}|\mathbf{y}_{:0};\boldsymbol{\theta})$, $\mathbf{y}_{:0} = \{\mathbf{y}_0, \mathbf{y}_{-1}, \mathbf{y}_{-2}, \ldots\}$ which is called contextual evidence in Carrassi et al. (2017). The times of the evidencing window, $1:K$, required for the estimation are the only ones that are kept explicit in (17).

Replacing (17) in (3) yields

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{k=1}^{K} \ln p(\mathbf{y}_k|\mathbf{y}_{1:k-1};\boldsymbol{\theta}) \\ &= \sum_{k=1}^{K} \ln\left(\int p(\mathbf{y}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{y}_{1:k-1};\boldsymbol{\theta})\mathrm{d}\mathbf{x}_k\right). \end{aligned} \tag{18}$$

If we assume Gaussian distributions and linear dynamical and observational models, the integrand in (18) is exactly the analysis distribution given by a Kalman filter (Carrassi et al., 2017). The likelihood of the observations conditioned on the state at each time is then given by:

$$\begin{aligned} p(\mathbf{y}_k|\mathbf{x}_k) = [(2\pi)^{M/2}|\mathbf{R}|^{1/2}]^{-1} \\ \times \exp\left[-\frac{1}{2}(\mathbf{y}_k - \mathcal{H}(\mathbf{x}_k))^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}_k - \mathcal{H}(\mathbf{x}_k))\right], \end{aligned} \tag{19}$$

and the prior forecast distribution,

$$\begin{aligned} p(\mathbf{x}_k|\mathbf{y}_{1:k-1};\boldsymbol{\theta}) = [(2\pi)^{N/2}|\mathbf{P}_k^f|^{1/2}]^{-1} \\ \times \exp\left[-\frac{1}{2}(\mathbf{x}_k - \mathbf{x}_k^f)^{\mathrm{T}}(\mathbf{P}_k^f)^{-1}(\mathbf{x}_k - \mathbf{x}_k^f)\right], \end{aligned} \tag{20}$$

where $\mathbf{x}_k^f = \mathcal{M}(\mathbf{x}_{k-1}^a) + \boldsymbol{\eta}_k$ is the forecast with $\boldsymbol{\eta}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_k)$, $\mathbf{x}_{k-1}^a$ is the analysis state – filter mean state estimate – at time $k-1$ and $\mathbf{P}_k^f$ is the forecast covariance matrix of the filter.

The resulting approximation of the observation likelihood function which is obtained replacing (19) and (20) in (18), is

$$\begin{aligned} l(\boldsymbol{\theta}) \approx -\frac{1}{2}\sum_{k=1}^{K}\Big[(\mathbf{y}_k - \mathbf{H}\mathbf{x}_k^f)^{\mathrm{T}}(\mathbf{H}\mathbf{P}_k^f\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1} \\ \times (\mathbf{y}_k - \mathbf{H}\mathbf{x}_k^f) + \ln(|\mathbf{H}\mathbf{P}_k^f\mathbf{H}^{\mathrm{T}} + \mathbf{R}|)\Big] + C \end{aligned} \tag{21}$$

where C stands for the constants independent of $\boldsymbol{\theta}$ and the observational operator is assumed linear, $\mathcal{H} = \mathbf{H}$. Equation (21) is exact for linear models $\mathcal{M} = \mathbf{M}$, but just an approximation for non-linear ones. As in EM, the point we made is that we expect that *the likelihood in the iterative method can converge through successive approximations*.

The evaluation of the model evidence (21) does not require the smoother. The forecasts $\mathbf{x}_k^f$ in (21) are started from the analysis – filter state estimates. In this case, the initial likelihood parameters $\mathbf{x}_0$ and $\mathbf{P}_0$ need to be good approximations (e.g. an estimation from the previous evidencing window) or they need to be estimated jointly to the other potentially unknown parameters $\boldsymbol{\Omega}$, $\mathbf{R}$, and $\mathbf{Q}$. Note that (21) does not depend explicitly on $\mathbf{Q}$ because the forecasts $\mathbf{x}_k^f$ already include the model error. The steps of the NR method are sketched in Fig. 1b.

For all the cases in which we can find an analytical expression for the maximization step of the EM algorithm, we can also derive a gradient of the likelihood function (Cappé et al., 2005). However, for the application of the NR maximization in both cases; when the EM maximization step can be derived analytically but also when it cannot, we have implemented an NR maximization based on a so-called derivative-free optimization method, i.e. a method that does not require the likelihood gradient, to be described in the next section.

3. Design of the numerical experiments

A first set of numerical experiments consists of twin experiments in which we first generate a set of noisy observations using the model with known parameters. Then, the maximum likelihood estimators are computed using the same model with the synthetic observations. Since we know the true parameters, we can evaluate the error in the estimation and the performance of the proposed algorithms. A second set of experiments applies the method for model identification. The (imperfect) model represents the multi-scale system through a set of coarse-grained dynamical equations and an unknown stochastic physical parameterization. The model-identification experiments are imperfect model experiments in which we seek to determine the stochastic physical parameterization of the small-scale variables from observations. In particular, the 'nature' or true model is the two-scale Lorenz-96 model and it is used to generate the synthetic observations, while the imperfect model is the one-scale Lorenz-96 model forced by a physical parameterization which has to be

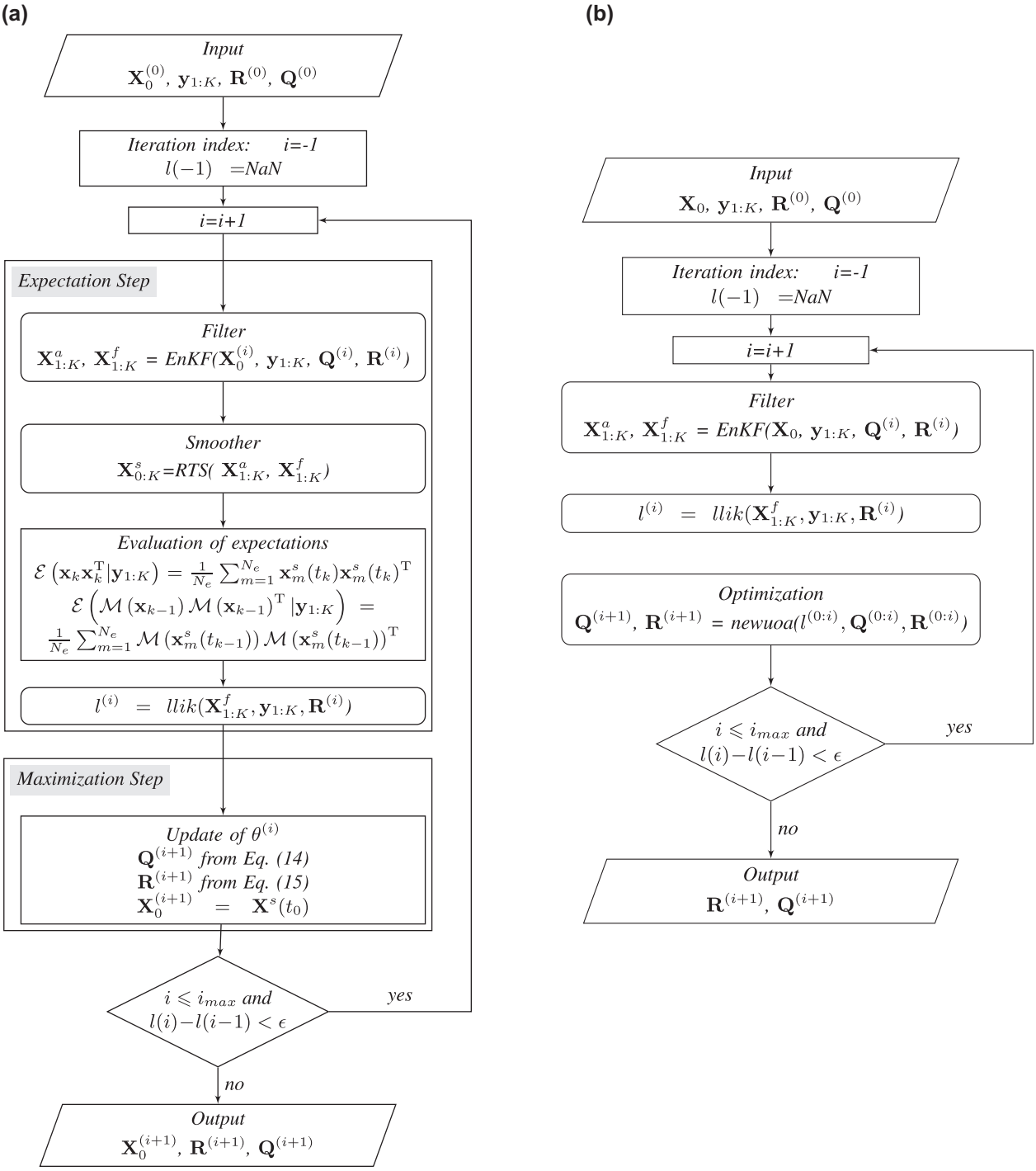

Fig. 1. (a) Flowchart of the EM algorithm (left panel). (b) NR flowchart (right panel). Each column of the matrix $\mathbf{X}_k$ is an ensemble member state $\mathbf{X}_k \equiv \mathbf{x}_{1:N_e}(t_k)$ at time k. Subscript (i) means ith iteration. A final application of the filter may be required to obtain the updated analysis state at $i+1$. The function $llik$ is the log-likelihood calculation from (21). The newuoa function in the optimization step refers to the 'new' unconstrained optimization algorithm (Powell, 2006).

identified. This parameterization should represent the effects of small-scale variables on the large-scale variables. In this way, the coarse-grained one-scale model with a physical parameterization with tunable deterministic and stochastic parameters is adjusted to account for the (noisy) observed data. We evaluate whether the EM algorithm and the NR method are able to determine the set of optimal parameters, assuming they exist.

The synthetic observations are taken from the known nature integration by, see (2),

$$\mathbf{y}_k \;=\; \mathbf{H}\mathbf{x}_k + \boldsymbol{\epsilon}_k \tag{22}$$

with $\mathbf{H} = \mathbf{I}$, i.e. all the state is observed. Futhermore, we assume non-correlated observations $\mathbf{R}_k = \mathcal{E}\left(\boldsymbol{\epsilon}_k \boldsymbol{\epsilon}_k^{\mathrm{T}}\right) = \alpha_R \mathbf{I}$.

3.1. Twin experiments

In the twin experiments, we use the one-scale Lorenz-96 system and a physical parameterization that represents subgrid-scale effects. The nature integration is conducted with this model and a set of 'true' physical parameter values. These parameters characterize both deterministic and stochastic processes. By virtue of the perfect model assumption, the model used in the estimation experiments is exactly the same as the one used in the nature integration except that the physical parameter values are assumed to be unknown. Although for simplicity we call this 'twin experiment', this experiment could be thought as a model selection experiment with parametric model error in which we know the 'perfect functional form of the dynamical equations' but the model parameters are completely unknown and they need to be selected from noisy observations.

The equations of the one-scale Lorenz-96 model are:

$$\frac{\mathrm{d}X_n}{\mathrm{d}t} + X_{n-1}(X_{n-2} - X_{n+1}) + X_n \;=\; G_n(X_n, a_0, \ldots, a_J)\,, \tag{23}$$

where $n = 1, \ldots, N$. The domain is assumed periodic, $X_{-1} \equiv X_{N-1}$, $X_0 \equiv X_N$, and $X_{N+1} \equiv X_1$.

We have included in the one-scale Lorenz-96 model a *physical parameterization* which is taken to be,

$$G_n(X_n,\; a_0, \ldots, a_2) \;=\; \sum_{j=0}^{2} (a_j \;+\; \eta_j(t)) \cdot (X_n)^j\,, \tag{24}$$

where a noise term, $\eta_j(t)$, of the form,

$$\eta_j(t) \;= \eta_j\;(t - \Delta t) \;+\; \sigma_j\; \nu_j(t), \tag{25}$$

has been added to *each* deterministic parameter. Equation (25) represents a random walk with standard deviation of the process σ_j, the stochastic parameters, and $\nu_j(t)$ is a realization of a Gaussian distribution with zero mean and unit variance. The standard deviation in the Runge–Kutta scheme is taken proportional to the square root of the time step $\sqrt{\Delta t}$ (Hansen and Penland, 2006). The parameterization (24) is assumed to represent subgrid-scale effects, i.e. effects produced by the small-scale variables to the large-scale variables (Wilks, 2005).

3.2. Model-identification experiments

In the model-identification experiments, the nature integration is conducted with the two-scale Lorenz-96 model (Lorenz, 1996). The state of this integration is taken as the true state evolution. The equations of the two-scale Lorenz-96 model, 'true' model, are given by N equations of large-scale variables X_n,

$$\begin{aligned}\frac{\mathrm{d}X_n}{\mathrm{d}t} + X_{n-1}(X_{n-2} - X_{n+1}) + X_n \;=\\ = F - \frac{h\,c}{b} \sum_{j=N_S/N(n-1)+1}^{nN_S/N} Y_j\;;\end{aligned} \tag{26}$$

with $n = 1, \ldots, N$; and N_S equations of small-scale variables Y_m, given by:

$$\begin{aligned}\frac{\mathrm{d}Y_m}{\mathrm{d}t} + c\;b\;Y_{m+1}(Y_{m+2} - Y_{m-1}) + c\;Y_m\\ = \frac{h\,c}{b}\;X_{\mathrm{int}[(m-1)/N_S/N]+1}\;,\end{aligned} \tag{27}$$

where $m = 1, \ldots, N_S$. The two set of equations, (26) and (27), are assumed to be defined on a periodic domain, $X_{-1} \equiv X_{N-1}$, $X_0 \equiv X_N$, $X_{N+1} \equiv X_1$, and $Y_0 \equiv Y_{N_S}$, $Y_{N_S+1} \equiv Y_1$, $Y_{N_S+2} \equiv Y_2$.

The imperfect model used in the model-identification experiments is the one-scale Lorenz-96 model (23) with a parameterization (24) meant to represent small-scale effects (right-hand side of (26)).

3.3. Numerical experiment details

As used in previous works (see e.g. Wilks 2005; Pulido et al. 2016), we set $N = 8$ and $N_S = 256$ for the large- and small-scale variables, respectively. The constants are set to the standard values $b = 10$, $c = 10$ and $h = 1$. The external forcing for the model-identification experiments is taken to be $F = 18$. The ordinary differential equations (26)–(27) are solved by a fourth-order Runge–Kutta algorithm. The time step is set to $dt = 0.001$ for integrating (26) and (27).

For the model-identification experiments, we aim to mimic the dynamics of the large-scale equations of the two-scale Lorenz-96 system with the one-scale Lorenz-96 system (23) forced by a

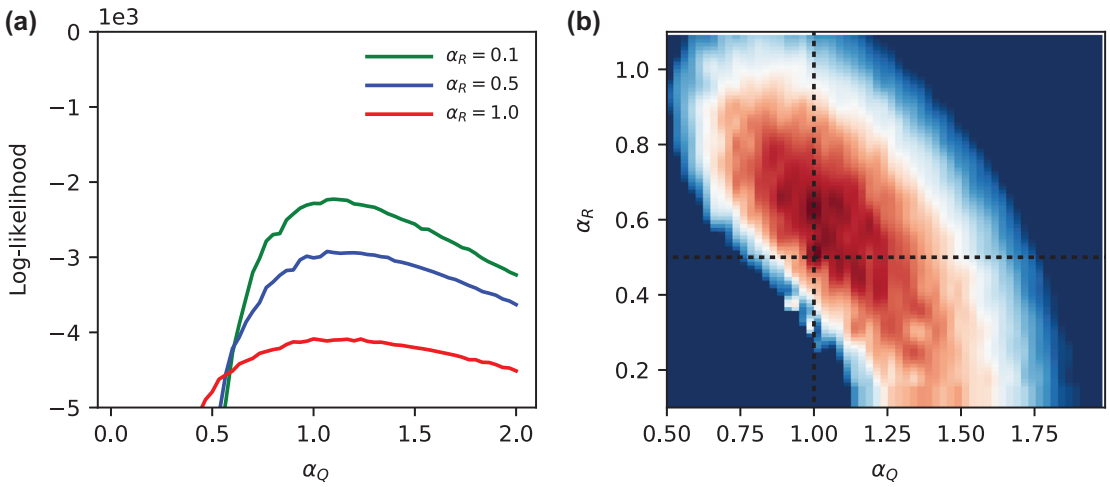

Fig. 2. Log-likelihood function as a function of (a) model noise for three true observational noise values, $\alpha_R^t = 0.1,\ 0.5,\ 1.0$; and as a function of (b) model noise (α_Q) and observational noise (α_R) for a case with $\alpha_Q^t = 1.0$ and $\alpha_R^t = 0.5$. Darker red shading represents larger log-likelihood.

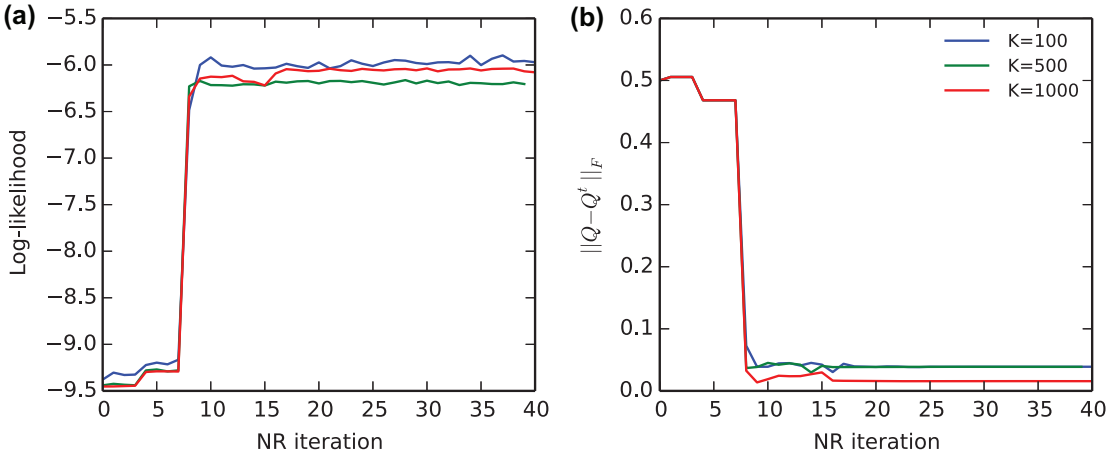

Fig. 3. Convergence of the NR maximization as a function of the iteration of the outer loop (inner loops are composed of $2N_C + 1$ function evaluations, where N_C is the control space dimension) for different evidencing window lengths ($K = 100,\ 500,\ 1000$). (a) Log-likelihood function. (b) Frobenius norm of the model noise estimation error.

physical parameterization (24). In other words, our nature is the two-scale model, while our imperfect coarse-grained model is the forced one-scale model. For this reason, we take 8 variables for the one-scale Lorenz-96 model for the twin experiments (as the number of large-scale variables in the model-identification experiments). Equations (23) are also solved by a fourth-order Runge–Kutta algorithm. The time step in all the experiments is also set to $dt = 0.001$.

The EnKF implementation we use is the ensemble transform Kalman filter (Hunt et al., 2007) without localization. A short description of the ensemble transform Kalman filter is given in the Appendix. The time interval between observations (cycle) is 0.05 (an elapsed time of 0.2 represents about 1 day in the real atmosphere considering the error growth rates; Lorenz, 1996). The number of ensemble members is set to $N_e = 50$. The number of assimilation cycles (observation times) is $K = 500$. This is the 'evidencing window' (Carrassi et al., 2017) in which we seek for the optimal likelihood parameters. The measurement variance error is set to $\alpha_R = 0.5$ except otherwise stated. We do not use any inflation factor, since the model error covariance matrix is estimated.

The optimization method used in the NR maximization is 'newuoa' (Powell, 2006). This is an unconstrained minimization algorithm which does not require derivatives. It is suitable

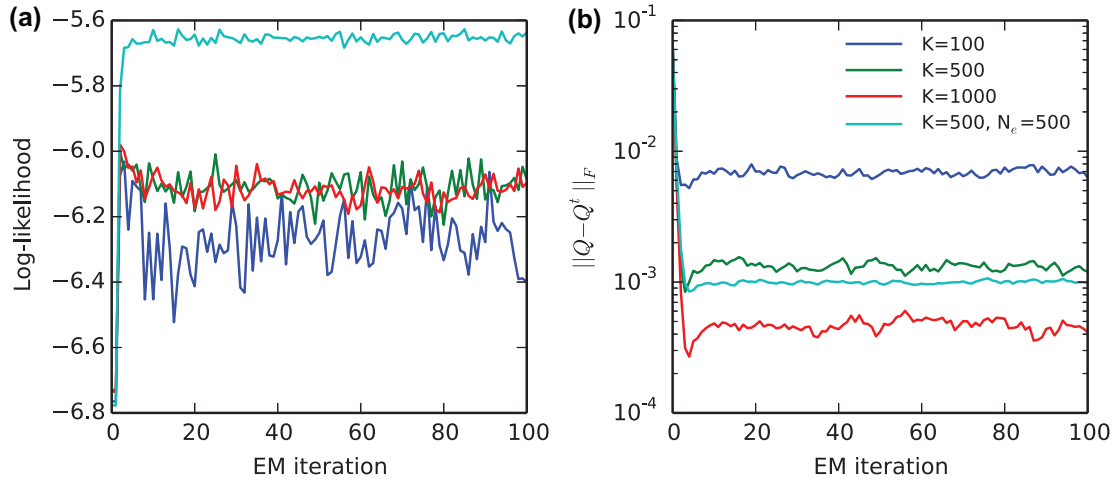

Fig. 4. Convergence of the EM algorithm as a function of the iteration for different observation time lengths (evidencing window). An experiment with $N_e = 500$ ensemble members and $K = 500$ is also shown. (a) Log-likelihood function. (b) The Frobenius norm of the model noise estimation error.

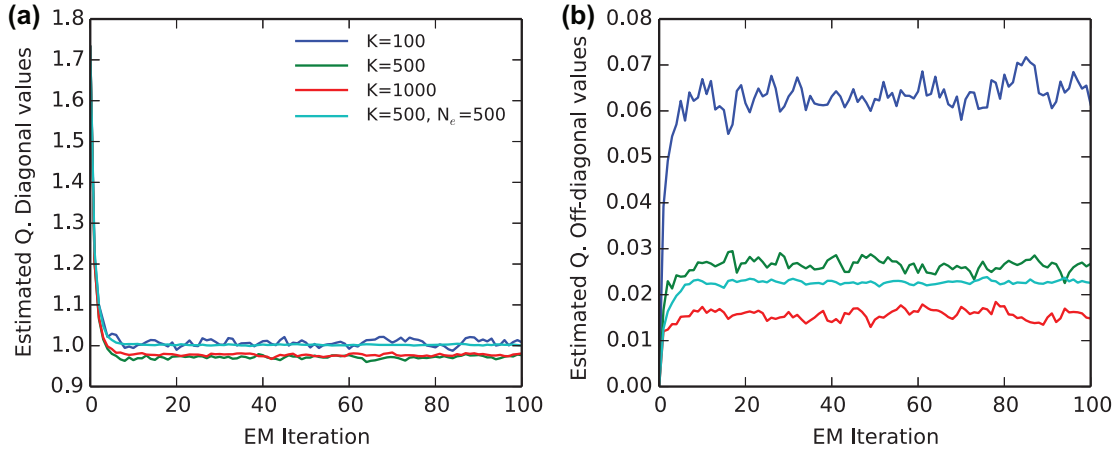

Fig. 5. Estimated model noise as a function of the iteration in the EM algorithm. (a) Mean diagonal model noise (true value is 1.0). (b) Mean off-diagonal absolute model noise value (true value is 0.0).

for control spaces of about a few hundred dimensions. This derivative-free method could eventually permit to extend the NR maximization method to cases in which the state evolution (1) incorporates a non-additive model error.

4. Results

4.1. Twin experiments: Estimation of model noise parameters

The nature integration is obtained from the one-scale Lorenz-96 model (23) with a constant forcing of $a_0 = 17$ without higher orders in the parameterization; in other words a one-scale Lorenz-96 model with an external forcing of $F = 17$. Information quantifiers show that for an external forcing of $F = 17$, the Lorenz-96 model is in a chaotic regime with maximal statistical complexity (Pulido and Rosso, 2017). The true model is represented by (1) with model noise following a normal density, $\boldsymbol{\eta}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}^t)$. The true model noise covariance is defined by $\mathbf{Q}^t = \alpha_Q^t \mathbf{I}$ with $\alpha_Q^t = 1.0$ (true parameter values are denoted by a t superscript). The observations are taken from the nature integration and perturbed using (22).

A first experiment examines the log-likelihood (21) as a function of α_Q for different true measurement errors, $\alpha_R^t = 0.1, 0.5, 1.0$ (Fig. 2a). A relatively smooth function is found with a well-defined maximum. The function is better conditioned for

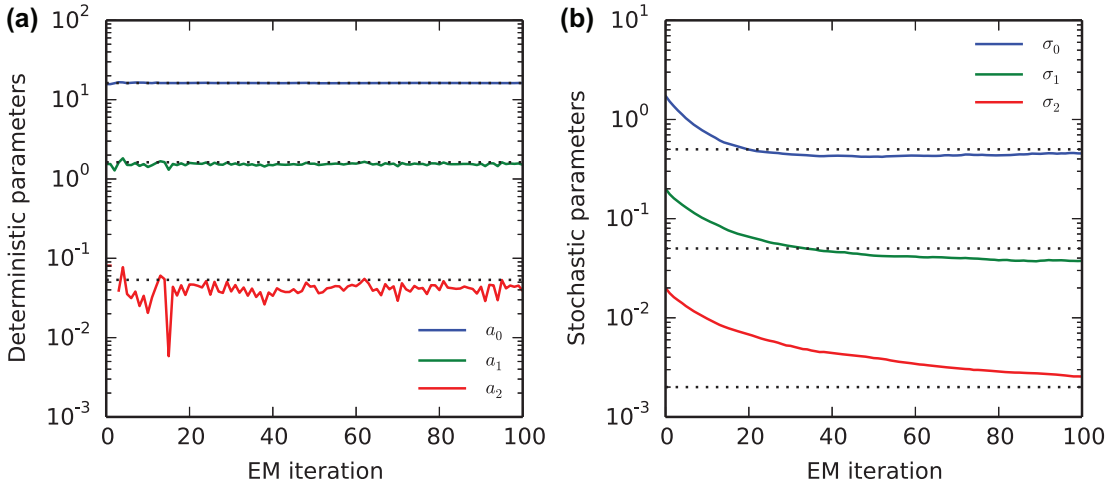

Fig. 6. (a) Estimated mean deterministic parameters, a_i, as a function of the EM iterations for the twin parameter experiment. (b) Estimated stochastic parameters, σ_i.

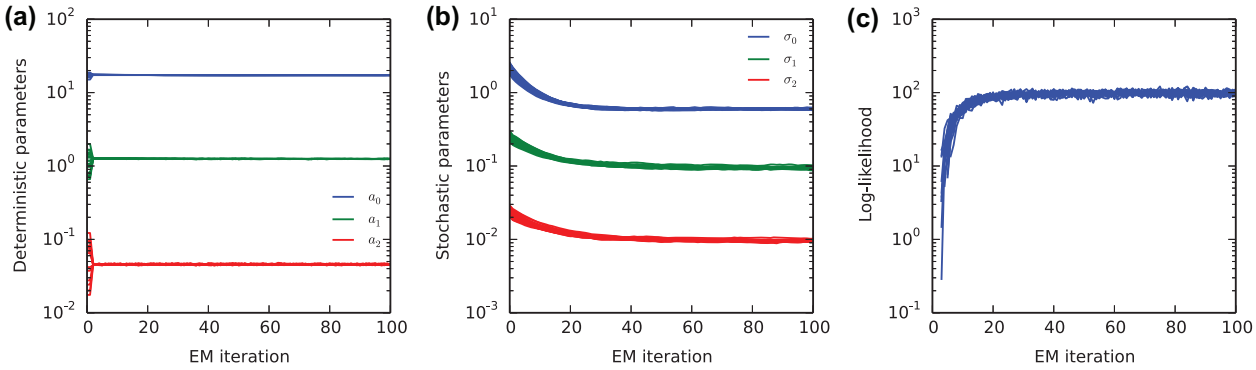

Fig. 7. (a) Estimated deterministic parameters as a function of the EM iterations for the model-identification experiment. Twenty experiments with random initial deterministic and stochastic parameters are shown. (b) Estimated stochastic parameters. (c) Log-likelihood function.

the experiments with smaller observational noise, α_R. Figure 2b shows the log-likelihood as a function of α_Q and α_R. The darkest shading is around $(\alpha_Q, \alpha_R) \approx (1.0, 0.5)$. However, note that because of the asymmetric shape of the log-likelihood function (Fig. 2a), the darker red region is shifted toward higher α_Q and α_R values. The up-left bottom-right orientation of the likelihood pattern in the plane α_Q and α_R reveals a correlation between them: the larger α_Q, the smaller α_R for the local maximal likelihood.

We conducted a second experiment using the same observations but the estimation of model noise covariance matrix is performed through the NR method. The control space is of $8\times8=64$ dimensions, i.e. the full $\mathbf{Q}$ model error covariance matrix is estimated (note that $N = 8$ is the model state dimension). Figure 3a depicts the convergence of the log-likelihood function in three experiments with evidencing window $K = 100$, 500 and 1000. The Frobenius norm of the error in the estimated model noise covariance matrix, i.e. $\|\mathbf{Q} - \mathbf{Q}^t\|_F = \sqrt{\sum_{ij}\left(Q_{ij} - Q^t_{ij}\right)^2}$, is shown in Fig. 3b. As the number of cycles used in a single batch process increases, the estimation error diminishes.

The convergence of the EM algorithm applied for the estimation of model noise covariance matrix only ($8\times8=64$ dimensions) is shown in Fig. 4. This work is focused on the estimation of model parameters so that the observation error covariance matrix is assumed to be known. The method would allow to estimate it jointly through (15), however, this is beyond the main aim of this work. This is similar to the previous experiment, using the EM instead of the NR method. In 10 iterations, the EM algorithm achieves a reasonable estimation, which is not further improved for larger number of iterations. The obtained log-likelihood value is rather similar to the NR method. The noise in the log-likelihood function diminishes with longer evidencing windows. The amplitude of the log-likelihood function noise for $K = 100$ is about 3%. These fluctuations are caused by sampling

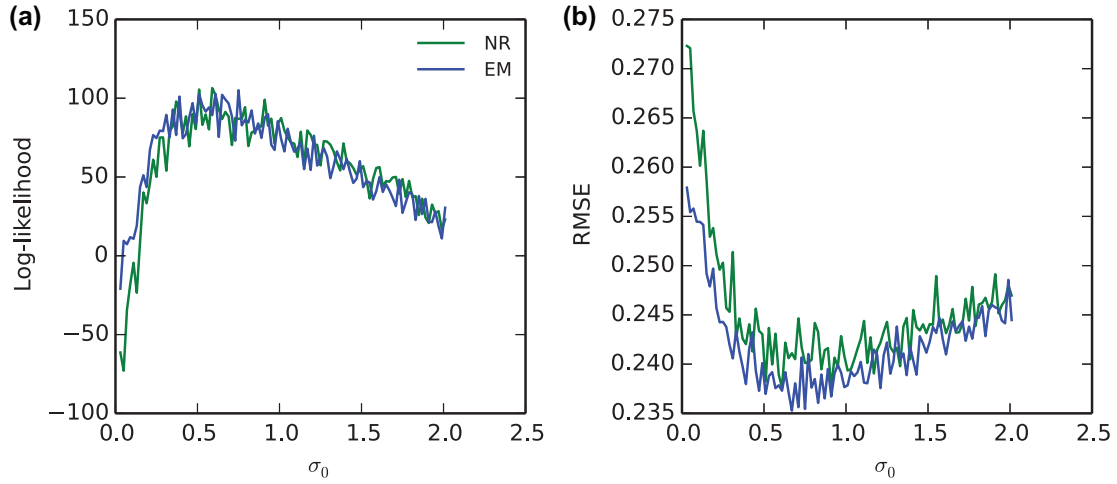

Fig. 8. (a) Log-likelihood as a function of the σ_0 parameter at the σ_1 and σ_2 optimal values for the NR estimation (green curve) and with the optimal values for the EM estimation (blue curve) for the imperfect-model experiment. (b) Analysis RMSE as a function of the σ_0 parameter.

noise. Note that the number of likelihood parameters is 64 and the evidencing window $K = 100$ in this case. For larger K, the log-likelihood noise is diminished $< 1\%$. As mentioned above a certain amount of noise may be beneficial for the convergence of the algorithm.

Comparing the standard $N_e = 50$ experiments with $N_e = 500$ in Fig. 4a, the noise also diminishes by increasing the number of ensemble members. Increasing the number of members does not appear to impact on the estimation of off-diagonal values, but it does so on the diagonal stochastic parameter values (Fig. 5a and b). The error in the estimates is about 7% in both diagonal and off-diagonal terms of the model noise covariance matrix for $K = 100$, and lower than 2% for the $K = 1000$ cycles case (Fig. 5).

4.2. Twin experiments: estimation of deterministic and stochastic parameters

A second set of twin experiments evaluates the estimation of deterministic and stochastic parameters from a physical parameterization. The model used to generate the synthetic observations is (23) with the physical parameterization (24). The length of the assimilation cycle is set to its standard value, 0.05. The deterministic parameters to conduct the nature integration are fixed to $a_0^t = 17.0$, $a_1^t = -1.15$ and $a_2^t = 0.04$ and the model error variance in each parameter is set to $\sigma_0^t = 0.5$, $\sigma_1^t = 0.05$, and $\sigma_2^t = 0.002$, respectively. The true parameters are governed by a stochastic process (25). This set of deterministic parameters is a representative physical quadratic polynomial parameterization, which closely resembles the dynamical regime of a two-scale Lorenz-96 model with $F = 18$ (Pulido and Rosso, 2017). The observational noise is set to $\alpha_R = 0.5$. An augmented state space of 11 dimensions is used, which is composed by appending to the 8 model variables the 3 physical parameters (a_0, a_1, a_2). The evolution of the augmented state is represented by (1) for the state vector component and a random walk for the parameters. The EM algorithm is then used to estimate the additive augmented state model error $\mathbf{Q}$ which is an 11×11 covariance matrix. Therefore, the smoother recursion gives an estimate of both the state variables and deterministic parameters. The recursion formula for the model error covariance matrix (and the parameter covariance submatrix) is given by (14).

Figure 6a shows the estimation of the mean deterministic parameters as a function of the EM iterations. The estimation of the deterministic parameters is rather accurate; a_2 has a small true value and it presents the lowest sensitivity. The estimation of the stochastic parameters by the EM algorithm converges rather precisely to the true stochastic parameters (Fig. 6b). The convergence requires of about 80 iterations. The estimated model error for the state variables is in the order of 5×10^{-2}. This represents the additive inflation needed by the filter for an optimal convergence. It establishes a lower threshold for the estimation of additive stochastic parameters.

A similar experiment was conducted with NR maximization for the same synthetic observations. A scaling of $S_\sigma = (1, 10, 100)$ was included in the optimization to increase the condition number. A good convergence was obtained with the optimization algorithm. The estimated optimal parameter values are $\sigma_0 = 0.38$ $\sigma_1 = 0.060$ $\sigma_2 = 0.0025$ for which the log-likelihood is $l = -491$. The estimation is reasonable with a relative error of about 25%.

4.3. Model-identification experiment: estimation of the deterministic and stochastic parameters

As a proof-of-concept model-identification experiment, we now use synthetic observations with an additive observational noise of $\alpha_R = 0.5$ taken from the nature integration of the two-

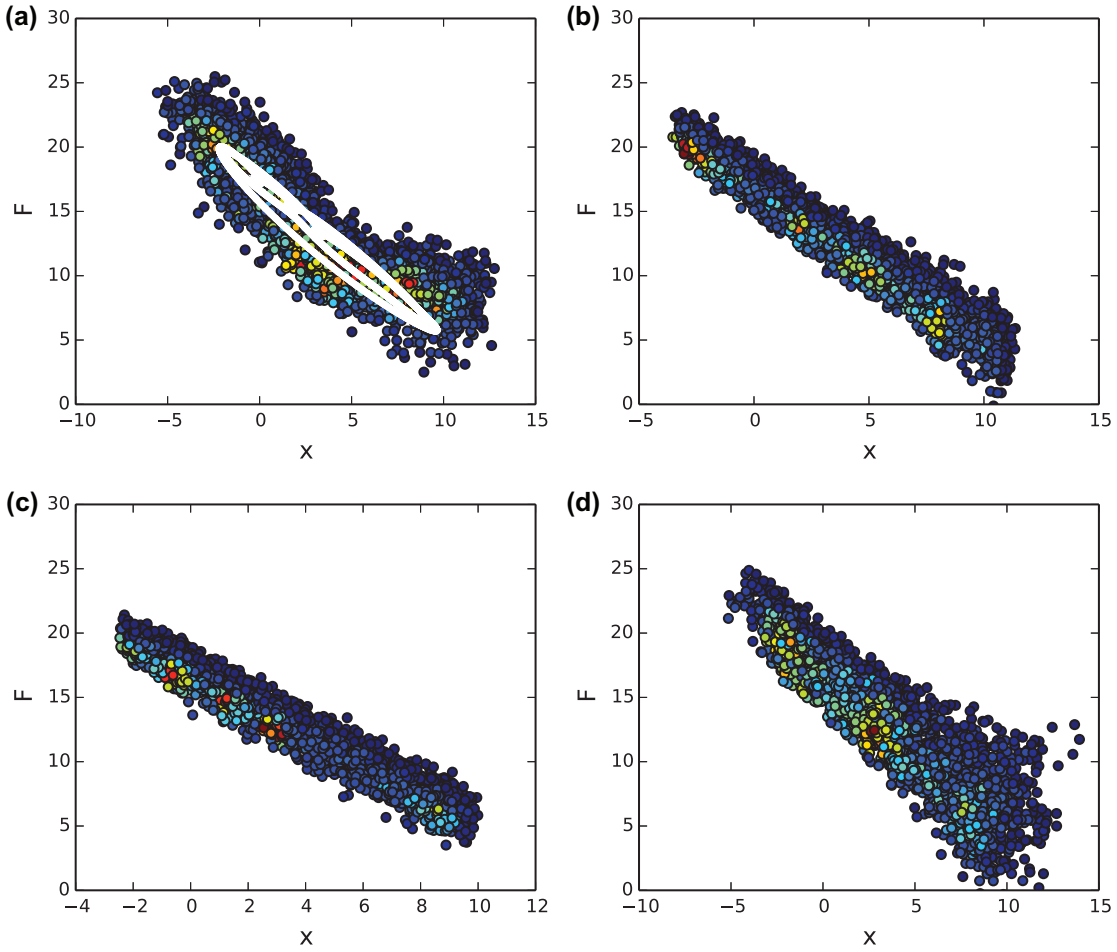

Fig. 9. (a) Scatterplot of the true small-scale effects in the two-scale Lorenz-96 model as a function of a large-scale variable (coloured dots) and scatterplot of the deterministic parameterization with optimal parameters (white dots). (b) Scatterplot from the stochastic paramerization with optimal parameters obtained with the EM algorithm and (c) with the NR method. (d) Scatterplot given by a constrained random walk with optimal EM parameters.

scale Lorenz-96 model with $F = 18$. On the other hand, the one-scale Lorenz-96 model is used in the ensemble Kalman filter with a physical parameterization that includes the quadratic polynomial function, (24), and the stochastic process (25). The deterministic parameters are estimated through an augmented state space while the stochastic parameters are optimized via the algorithm for the maximization of the log-likelihood function. The model error covariance estimation is constrained for these experiments to the three stochastic parameters alone. Figure 7a shows the estimated deterministic parameters as a function of the EM iterations. Twenty experiments with different initial deterministic parameters and initial stochastic parameter values were conducted. The deterministic parameter estimation does not manifest a significant sensitivity to the stochastic parameter values. The mean estimated values are $a_0 = 17.3$, $a_1 = -1.25$ and $a_3 = 0.0046$. Note that the deterministic parameter values estimated with information quantifiers in Pulido and Rosso (2017) for the two-scale Lorenz-96 with $F = 18$ are $(a_0, a_1, a_2) = (17.27, -1.15, 0.037)$. Figure 7b depicts the convergence of the stochastic parameters. The mean of the optimal stochastic parameter values are $\sigma_0 = 0.60$, $\sigma_1 = 0.094$ and $\sigma_2 = 0.0096$ with the log-likelihood value being 98.8 (single realization). The convergence of the log-likelihood is shown in Fig. 7c.

NR maximization is applied to the same set of synthetic observations. The mean estimated deterministic and stochastic parameters are $(a_0, a_1, a_2) = (17.2, -1.24, 0.0047)$ and $(\sigma_0, \sigma_1, \sigma_2) = (0.59, 0.053, 0.0064)$ from 20 optimizations. As in the EM experiment, only the three stochastic parameters were estimated as likelihood parameters. Preliminary experiments with the full augmented model error covariance gave smaller estimated σ_0 values and nonnegligible model error variance (not shown). The log-likelihood function (Fig. 8a) and the analysis root-mean-square error (RMSE, Fig. 8b) are shown as a

function of σ_0 at the σ_1 and σ_2 optimal values given by the Newton–Rapshon method (green curve) and at the σ_1 and σ_2 optimal values given by the EM algorithm (blue curve). The log-likelihood values are indistinguishable. A slightly smaller analysis RMSE is obtained for the EM algorithm (Fig. 8b), which is likely related to the improvement with the iterations of the initial prior distribution in the EM algorithm, while this distribution is fixed in the NR method.

Long integrations (10^6 time cycles) of the nature model and the identified coarse-grained models were conducted to evaluate the parameterizations. The true effects of the small-scale variables on a large-scale variable from the two-scale Lorenz-96 model are shown in Fig. 9 as a function of the large-scale variable. This true scatterplot is obtained by evaluating the right-hand side of (26). The deterministic quadratic parameterization with the optimal parameters from the EnKF is also represented in Fig. 9a. A poor representation of the functional form and variability is obtained. Figure 9b shows the scatterplot with a stochastic parameterization which stochastic parameters are the ones estimated with EM algorithm, while Fig. 9c shows it for the stochastic parameters estimated with the NR method. The two methods, NR and EM, give scatterplots of the parameterization which are almost indistinguishable and improve the small-scale representation with respect to the deterministic parameterization. Figure 9d shows the scatterplot resulting from the quadratic parameterization using a random walk for the parameters set to the estimated values with the EM algorithm. The values of the parameters are limited to the $a_i \pm 4\sigma_i$ range. The parameter values need to be constrained, because for these long free simulations, some parameter values given by the random walk produce numerical instabilities in the Lorenz-96 model (Pulido et al., 2016). The stochastic parameterization which was identified by the statistical learning technique improves substantially the functional form of the effects of the small-scale variables. Using a constrained random walk appears to give the best simulation.

5. Conclusions

Two novel methods to estimate and charactize physical parametrizations in stochastic multi-scale dynamical systems have been introduced, the expectation–maximization algorithm (EnKF-EM) and Newton–Raphson likelihood maximization (EnKF-NR) combined with the ensemble Kalman filter. These new methods are suitable for the estimation of both stochastic and deterministic parameters, based on sparse and noisy observations. Both methods determine the maximum of the observation likelihood , also known as model evidence, given a set of spatio-temporally distributed observations, using the ensemble Kalman filter to combine observations with model predictions. The methods are first evaluated in a controlled model experiment in which the true parameters are known and then, in the two-scale Lorenz-96 dynamics which is represented with a stochastic coarse-grained model. The performance of the methods is excellent, even in the presence of moderate observational noise. The methods do not require neither inflation factors nor any other tunable parameters , because the methodology includes an additive model noise term or stochastic parameters, which compensate for the underestimation of the forecast error covariance. The level of model noise to be added is not arbitrarily chosen but the one that gives the maximal observation likelihood.

The estimation based on the expectation–maximization algorithm gives very promising results in these medium-sized experiments (≈100 parameters). About 50 iterations are needed to achieve an estimation error lower than 10% using 100 observation times. Using a longer observation time interval, the accuracy is improved. The estimation of stochastic parameters included the case of additive, i.e. a_0, and multiplicative parameters, i.e. $a_1 X_n$ and $a_2 X_n^2$. The number of ensemble members has a strong impact on the stochastic parameter variance, while the length of the observation time interval appears to have a stronger impact on the stochastic parameter correlations.

The computational cost of the algorithm is directly related to the number of iterations needed for convergence. Each iteration requires the application of an ensemble Kalman filter and a smoother (which needs an extra inversion through singular value decomposition). In the model-identification experiments, 50 EM iterations were chosen as a secure option, with a minimal iteration number of 20 for coarse convergence. In an operational high-dimensional data assimilation system, the application of 20–50 ensemble Kalman filters would be prohibitive. On the other hand, these experiments would be computationally feasible for model identification, during the model development phase, even for high-dimensional systems or for tuning the data assimilation scheme.

The estimation based on the NR method also presents good convergence for the twin experiment with an additive stochastic parameter. For the more realistic model-identification experiments, the model evidence presents some noise which may affect the convergence. The free-derivative optimization requires about 10 iterations of $2N_C + 1$ evaluations where N_C is the control space dimension (number of parameters to be estimated). For higher dimensional problems and large number of parameters, optimization algorithms that use the gradient of the likelihood to the likelihood parameters need to be implemented. Moreover, the use of simulated annealing or other stochastic gradient optimization techniques suitable for noisy cost functions would be required.

The EM algorithm assumes a Gaussian additive model error term, which leads to an analytical expression for the maximization step. Besides, the derivation of the likelihood function in the NR method also assumes Gaussian additive model and observation errors. The methods could be extended for non-Gaussian statistics, in which case the maximization step in the EM algorithm can be conducted through an optimization itera-

tive method. For cases with multimodal statistics, the application of a particle filter (vanLeeuwen, 2009) and smoother (Briers et al., 2010) instead of the Kalman filter and RTS smoother would be required.

Both estimation methods can be applied to a set of different dynamical models to address which one is more reliable given a set of noisy observations; the so called 'model selection' problem. A comparison of the likelihood from the different models with the optimal parameters gives a measure of the model fidelity to the observations. Majda and Gershgorin (2011) seeked to improve imperfect models by adding stochastic forcing and used a measure from information theory that gives the closest model distribution to the observed probability distribution. The model-identification experiments in the current work can be viewed as pursuing a similar objective, stochastic processes are added to the physical parameterization to improve the model representation of the unresolved processes. Different structural parameterizations can be compared through their maximal observation likelihood, the one that gets the larger maximal observation likelihood for the optimal likelihood parameters using the same set of observations is the parameterization that best suits the data.

Hannart et al. (2016) proposed to apply the observation likelihood function, model evidence, that results from assimilating a set of observations, for the detection and attribution of climate change. They suggest to evaluate the likelihood in two possible model configurations, one with the current anthropogenic forcing scenario (factual world) and one with the preindustrial forcing scenario (contrafactual world). If the evidencing window where the observations are located includes, for instance, an extreme event then one could determine the fraction of attributable risk as the fraction of the change in the observation likelihood of the extreme event which is attributable to the anthropogenic forcing.

The increase of data availability in many areas has fostered the number of applications of the ensemble Kalman filter. In particular, it has been used for influenza forecasting (Shaman et al., 2013) and for determining a neural network structure (Hamilton et al., 2013). The increase in spatial and temporal resolution of data offers great opportunities for understanding multi-scale strongly-coupled systems such as atmospheric and oceanic dynamics. This has lead to the proposal of purely data-driven modelling which uses past observations to reconstruct the dynamics through the ensemble Kalman filter without a dynamical model (Hamilton et al., 2016; Lguensat et al., 2017). The use of automatic statistical learning techniques that can use measurements for improvement of multi-scale models is also a promising venue. Following this recent stream of research, in this work, we propose the coupling of the EM algorithm and NR method with the ensemble Kalman filter which may be applicable to a wide range of multi-scale systems to improve the representation of the complex interactions between different scales.

Acknowledgements

The authors wish to acknowledge the members of the DADA CNRS team for insightful discussions, in particular Alexis Hannart, Michael Ghil and Juan Ruiz.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Nordic Center of Excellence EmblA of the Nordic Countries Research Council, NordForsk, by the project REDDA of the Norwegian Research Council; and by ANPCyT [grant number PICT2015-2368]. Cerea is a member of Institut Pierre-Simon Laplace (IPSL).

Notes

1. We use the notation ';', $p(\mathbf{y}_{1:K};\boldsymbol{\theta})$ instead of conditioning '|' to emphasize that $\boldsymbol{\theta}$ is not a random variable but a parameter. NR maximization and EM are point estimation methods so that $\boldsymbol{\theta}$ is indeed assumed to be a parameter (Cappé et al., 2005).
2. In principle what is required in (7) is $p(\mathbf{x}_{0:K}|\mathbf{y}_{1:K})$ so that a fixed-interval smoother needs to be applied. However, it has been shown by Raanes (2016) that the Rauch–Tung–Striebel smoother and the ensemble Kalman smoother, a fixed-interval smoother, are equivalent even in the non-linear, non-Gaussian case.

References

Anderson, J. 2001. An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev.* **142**, 2884–2903.

Bellsky, T., Berwald, J. and Mitchell, L. 2014. Nonglobal parameter estimation using local ensemble Kalman filtering. *Mon. Wea. Rev.* **142**, 2150–2164.

Bishop, C. 2006. *Pattern Recognition and Machine Learning* Springer.

Briers, M., Doucet, A. and Maskell, S. 2010. Smoothing algorithms for state-spacemodels. *Ann. Inst. Stat. Math.* **62**, 61–89.

Cappé, O., Moulines, E. and Rydén, T. 2005. *Inference in Hidden Markov Models*. Springer, New York, NY.

Carrassi, A., Bocquet, M., Hannart, A. and Ghil, M. 2017. Estimating model evidence using data assimilation. *Q. J. R. Meteorol. Soc.* **143**, 866–880.

Carrassi, A. and Vannitsem, S. 2011. State and parameter estimation with the extended Kalman filter: an alternative formulation of the model error dynamics. *Q. J. R. Meteorol. Soc.* **137**, 435–451.

Christensen, H., Moroz, I. M. and Palmer, T. N. 2015. Stochastic and perturbed parameter representations of model uncertainty in convection parameterization. *J. Atmos. Sci.* **72**, 2525–2544.

Cosme, E., Verron, J., Brasseur, P., Blum, J. and Auroux, D. 2012. Smoothing problems in a Bayesian framework and their linear gaussian solutions. *Mon. Weather Rev.* **140**, 683–695.

Delsole, T. and Yang, X. 2010. State and parameter estimation in stochastic dynamical models. *Physica D* **239**, 1781–1788.

Dempster, A., Laird, N. and Rubin, D. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **9**, 1–38.

Dreano, D., Tandeo, P., Pulido, M., Ait-El-Fquih, B., Chonavel, T. and co-authors. 2017. Estimation of error covariances in nonlinear state-space models using the expectation maximization algorithm. *Q. J. R. Meteorol. Soc.* **142**, 1877–1885.

Evensen, G. 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *J. Geophys. Res.* **99**, 10143–10162.

Evensen, G. 2003. The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dyn.* **53**, 343–367.

Ghahramani, Z. and Roweis, S. 1999. Learning nonlinear dynamical systems using an EM algorithm. In: *Advances in Neural Information Processing Systems*, MIT Press, pp. 431–437.

Hamilton, F., Berry, T., Peixoto, N. and Sauer, T. 2013. Real-time tracking of neuronal network structure using data assimilation. *Phys. Rev. E* **88**, 052715.

Hamilton, F., Berry, T. and Sauer, T. 2016. Ensemble Kalman filtering without a model. *Phys. Rev. X* **6**, 011021.

Hannart, A., Carrassi, A., Bocquet, M., Ghil, M., Naveau, P. and co-authors. 2016. Dada: data assimilation for the detection and attribution of weather and climate-related events. *Clim. Change* **136**, 155–174.

Hansen, J. and Penland, C. 2006. Efficient approximate techniques for integrating stochastic differential equations. *Mon. Wea. Rev.* **134**, 3006–3014.

Hunt, B., Kostelich, E. J. and Szunyogh, I. 2007. Efficient data assimilation for spatio-temporal chaos: a local ensemble transform Kalman filter. *Physica D* **77**, 437–471.

Jazwinski, A. H. 1970. *Stochastic and Filtering Theory*. Mathematics in Sciences and Engineering Series, Vol. 64. Academic Press, London and New York, p. 376.

Kalnay, E. 2002. *Atmospheric Modeling, Data Assimilation, and Predictability* Cambridge University Press, Cambridge.

Katsoulakis, M., Majda, A. and Vlachos, D. 2003. Coarse-grained stochastic processes for microscopic lattice systems. *Proc. Nat. Acad. Sci.* **100**, 782–787.

Kondrashov, D., Ghil, M. and Shprits, Y. 2011. Lognormal Kalman filter for assimilating phase space density data in the radiation belts. *Space Weather* **9**, 11.

Lguensat, R., Tandeo, P., Fablet, R., Pulido, M. and Ailliot, P. 2017. The analog ensemble-based data assimilation. *Mon. Wea. Rev.* **145**, 4093–4107.

Lorenz, E. (1996). *Predictability–A Problem Partly Solved*. Reading: ECMWF. (pp. 1–18)

Lott, F., Guez, L. and Maury, P. 2012. A stochastic parameterization of nonorographic gravity waves: formalism and impact on the equatorial stratosphere. *Geophys. Res. Lett.* **39**, L06807.

Majda, A. and Gershgorin, B. 2011. Improving model fidelity and sensitivity for complex systems through empirical information theory. *Proc. Nat. Acad. Sci.* **100**, 10044–10049.

Mason, P. and Thomson, D. 1992. Stochastic backscatter in large-eddy simulations of boundary layers. *J. Fluid Mech.* **242**, 51–78.

Neal, R. and Hinton, G. 1999. *A View of the EM Algorithm that Justifies Incremental, Sparse and other Variants*. Springer, Dordrecht.

Nicolis, N. 2004. Dynamics of model error: the role of unresolved scales revisited. *J. Atmos. Sci.* **61**, 1740–1753.

Palmer, T. 2001. A nonlinear dynamical perspective on model error: a proposal for non-local stochastic-dynamic parameterization in weather and climate prediction models. *Q. J. R. Meteorol. Soc.* **127**, 279–304.

Powell, M. 2006. The NEWUOA software for unconstrained optimization without derivatives. In: *Large-Scale Nonlinear Optimization*. Springer, Boston, MA, pp. 255–297.

Pulido, M. and Rosso, O. 2017. Model selection: using information measures from ordinal symbolic analysis to select model sub-grid scale parameterizations. *J. Atmos. Sci.* **74**, 3253–3269.

Pulido, M., Scheffler, G., Ruiz, J., Lucini, M. and Tandeo, P. 2016. Estimation of the functional form of subgrid-scale schemes using ensemble-based data assimilation: a simple model experiment. *Q. J. R. Meteorol. Soc.* **142**, 2974–2984.

Raanes, P. 2016. On the ensemble Rauch-Tung-Striebel smoother and its equivalence to the ensemble Kalman smoother. *Q. J. R. Meteorol. Soc.* **142**, 1259–1264.

Ruiz, J., Pulido, M. and Miyoshi, T. 2013a. Estimating parameters with ensemble-based data assimilation a review. *J. Meteorol. Soc. Jpn.* **91**, 79–99.

Ruiz, J., Pulido, M. and Miyoshi, T. 2013b. Estimating parameters with ensemble-based data assimilation parameter covariance treatment. *J. Meteorol. Soc. Jpn.* **91**, 453–469.

Santitissadeekorn, N. and Jones, C. 2015. Two-stage filtering for joint state-parameter estimation. *Mon. Wea. Rev.* **143**, 2028–2042.

Shaman, J., Karspeck, A., Yang, W., Tamerius, J. and Lipsitch, M. 2013. Real-time influenza forecasts during the 2012–2013 season. *Nat. Commun.* **4**, 2837.

Shumway, R. and Stoffer, D. 1982. An approach to time series smoothing and forecasting using the EM algorithm. *J. Time Ser. Anal.* **3**, 253–264.

Shutts, G. 2015. A stochastic convective backscatter scheme for use in ensemble prediction systems. *Q. J. R. Meteorol. Soc.* **141**, 2602–2616.

Stensrud, D. 2009. *Parameterization Schemes: Keys to Understanding Numerical Weather Prediction Models* Cambridge University Press, Cambridge.

Tandeo, P., Pulido, M. and Lott, F. 2015. Offline estimation of subgrid-scale orographic parameters using EnKF and maximum likelihood error covariance estimates. *Q. J. R. Meteorol. Soc.* **141**, 383–395.

van Leeuwen, P. J. 2009. Particle filtering in geophysical systems. *Mon. Wea. Rev.* **407**, 4089–4114.

Wei, G. and Tanner, M. A. 1990. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Stat. Assoc.* **85**, 699–704.

West, M. and Liu, J. 2001. Combined parameter and state estimation in simulation-based filtering. In: *Sequential Monte Carlo Methods in Practice*. Springer, New York, pp. 197–223.

Wikle, C. and Berliner, L. 2007. A Bayesian tutorial for data assimilation. *Physica D* **230**, 1–16.

Wilks, D. S. 2005. Effects of stochastic parametrizations in the Lorenz 96 system. *Q. J. R. Meteorol. Soc.* **131**, 389–407.

Wu, C. 1983. On the convergence properties of the EM algorithm. *Ann. Stat.* **11**, 95–103.

Appendix 1. Ensemble Kalman filter and smoother

The ensemble Kalman filter determines the probability density function of a dynamical model conditioned to a set of past observations, i.e. $p(\mathbf{x}_k|\mathbf{y}_{1:k})$, based on the Gaussian assumption. The mean and covariances are represented by a set of possible states, called *ensemble members*. Let us assume that the a priori ensemble members at time k are $\mathbf{x}^f_{1:N_e}(t_k)$, so that the empirical mean and covariance of the a priori hidden state are:

$$\overline{\mathbf{x}}^f(t_k) = \frac{1}{N_e}\sum_{m=1}^{N_e}\mathbf{x}^f_m(t_k),$$
$$\mathbf{P}^f(t_k) = \frac{1}{N_e-1}\mathbf{X}^f(t_k)[\mathbf{X}^f(t_k)]^{\mathrm{T}}, \tag{A1}$$

where $\mathbf{X}^f(t_k)$ is a matrix with the ensemble member perturbations, $\mathbf{x}^f_m(t_k) - \overline{\mathbf{x}}^f(t_k)$, as the m-th column.

To obtain the estimated hidden state, called *analysis state*, the observations are combined statistically with the a priori model state using the Kalman filter equations. In the case of the ensemble transformed Kalman filter (Hunt et al., 2007), the analysis state is a linear combination of the N_e ensemble member perturbations,

$$\overline{\mathbf{x}}^a = \overline{\mathbf{x}}^f + \mathbf{X}^f\overline{\mathbf{w}}^a, \quad \mathbf{P}^a = \mathbf{X}^f\tilde{\mathbf{P}}^a(\mathbf{X}^f)^{\mathrm{T}}. \tag{A2}$$

The optimal ensemble member weights $\overline{\mathbf{w}}^a$ are obtained considering the distance between the projection of member states to the observational space, $\mathbf{y}^f_m \equiv \mathcal{H}(\mathbf{x}^f_m)$, and observations $\mathbf{y}$. These weights and the analysis covariance matrix in the perturbation space are:

$$\overline{\mathbf{w}}^a = \tilde{\mathbf{P}}^a(\mathbf{Y}^f)^{\mathrm{T}}\mathbf{R}^{-1}[\mathbf{y} - \overline{\mathbf{y}}^f],$$
$$\tilde{\mathbf{P}}^a = [(N_e-1)\mathbf{I} + (\mathbf{Y}^f)^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y}^f]^{-1}. \tag{A3}$$

All the quantities in (A2) and (A3) are at time t_k so that the time dependence is omitted for clarity. A detailed derivation of (A2) and (A3) and a thorough description of the ensemble transformed Kalman filter and its numerical implementation can be found in Hunt et al. (2007).

To determine each ensemble member of the analysis state, the ensemble transformed Kalman filter uses the square root of the analysis covariance matrix, thus it belongs to the so-called square-root filters,

$$\mathbf{x}^a_m = \overline{\mathbf{x}}^f + \mathbf{X}^f\mathbf{w}^a_m \tag{A4}$$

where the perturbations of $\mathbf{w}^a_m$ are the columns of $\mathbf{W}^a = [(N_e - 1)\tilde{\mathbf{P}}^a]^{1/2}$.

The analysis state is evolved to the time of the next available observation t_{k+1} through the dynamical model equations which given the a priori or *forecasted* state,

$$\mathbf{x}^f_m(t_{k+1}) = \mathcal{M}(\mathbf{x}^a_m(t_k)). \tag{A5}$$

The smoother determines the probability density function of a dynamical model conditioned to a set of past and future observations, i.e. $p(\mathbf{x}_k|\mathbf{y}_{1:K})$, based on the Gaussian assumption. Applying the Rauch–Tung–Striebel smoother retrospective formula to each ensemble member (Cosme et al., 2012),

$$\mathbf{x}^s_m(t_k) = \mathbf{x}^a_m(t_k) + \mathbf{K}^s(t_k)[\mathbf{x}^s_m(t_{k+1}) - \mathbf{x}^f_m(t_{k+1})], \tag{A6}$$

where $\mathbf{K}^s(t_k) = \mathbf{P}^a(t_k)\mathbf{M}^{\mathrm{T}}_{k\to k+1}[\mathbf{P}^f(t_{k+1})]^{-1}$, and $\mathbf{M}_{k\to k+1}$ being the linear tangent model. For the application of the smoother in conjunction with the ensemble transformed Kalman filter, the smoother gain is re-expressed as:

$$\mathbf{K}^s(t_k) = \mathbf{X}^f(t_k)\mathbf{W}^a[\mathbf{X}^f(t_{k+1})]^{\dagger}. \tag{A7}$$

In practice, the pseudo inversion of the forecast state perturbation matrix $\mathbf{X}^f$ required in (A7) is conducted through singular value decomposition.