Nonlin. Processes Geophys., 19, 127–143, 2012
www.nonlin-processes-geophys.net/19/127/2012/
doi:10.5194/npg-19-127-2012

Nonlinear Processes in Geophysics

On closure parameter estimation in chaotic systems

J. Hakkarainen[1,2,*], A. Ilin[3,*], A. Solonen[1,2,*], M. Laine[1], H. Haario[2], J. Tamminen[1], E. Oja[3], and H. Järvinen[1]

[1]Finnish Meteorological Institute, Helsinki, Finland
[2]Lappeenranta University of Technology, Lappeenranta, Finland
[3]Aalto University School of Science, Espoo, Finland
[*]These authors contributed equally to this work.

Correspondence to: J. Hakkarainen (janne.hakkarainen@fmi.fi)

Received: 21 November 2011 – Revised: 3 February 2012 – Accepted: 7 February 2012 – Published: 15 February 2012

Abstract. Many dynamical models, such as numerical weather prediction and climate models, contain so called closure parameters. These parameters usually appear in physical parameterizations of sub-grid scale processes, and they act as "tuning handles" of the models. Currently, the values of these parameters are specified mostly manually, but the increasing complexity of the models calls for more algorithmic ways to perform the tuning. Traditionally, parameters of dynamical systems are estimated by directly comparing the model simulations to observed data using, for instance, a least squares approach. However, if the models are chaotic, the classical approach can be ineffective, since small errors in the initial conditions can lead to large, unpredictable deviations from the observations. In this paper, we study numerical methods available for estimating closure parameters in chaotic models. We discuss three techniques: off-line likelihood calculations using filtering methods, the state augmentation method, and the approach that utilizes summary statistics from long model simulations. The properties of the methods are studied using a modified version of the Lorenz 95 system, where the effect of fast variables are described using a simple parameterization.

1 Introduction

Many dynamical models in atmospheric sciences contain so called closure parameters. These parameters are usually connected to processes that occur on smaller and faster scales than the model discretization allows. For instance, the computational grid used in modern climate and numerical weather prediction (NWP) models is too coarse to directly model cloud micro-physics and many cloud-related phenomena are therefore represented by parameterization schemes. For example, consider the cloud shortwave optical properties, which are related to the cloud liquid water amount, and affect the shortwave radiation fluxes in atmospheric models. These properties can be specified with parameters, such as the mean effective radius of cloud water droplets (Martin et al., 1994).

The closure parameters act as "tuning handles" of the models. Parameter tuning is particularly necessary whenever new and improved parameterized processes are implemented into the models. Currently, the parameters are usually predefined by experts using a relatively small number of model simulations. This tuning procedure is somewhat subjective and therefore open for criticism.

In this paper, we discuss different algorithmic ways to estimate the tuning parameters, that is, how to find the optimal closure parameters by fitting the model to available observations. While this problem has not been studied much, there are a few recent papers that address the problem in the context of climate modeling (Jackson et al., 2008; Villagran et al., 2008; Järvinen et al., 2010; Sexton et al., 2011). Numerical weather prediction (NWP) is considered to be more of an initial value problem than a parameter estimation problem (Annan and Hargreaves, 2007) and tuning of closure parameters is in general done manually by using samples of test forecasts. In a recently proposed approach, NWP parameter estimation is embedded into ensemble prediction systems (Järvinen et al., 2011; Laine et al., 2011).

While the motivation behind this work is the closure parameter estimation problem in atmospheric models, we note that similar parameterizations appear in many multi-scale models in computational fluid dynamics. The parameter estimation is complicated by the fact that these models are often chaotic, which means that a small change in the initial conditions can lead to a completely different simulated trajectory. Therefore, the classical parameter estimation approaches that are based on directly comparing model simulations and observations using, for instance, a least squares approach, may be inefficient.

To fix the notation, let us assume for simplicity that a dynamical model can be described by a discrete state space model

$$\boldsymbol{x}_k = \mathcal{M}(\boldsymbol{x}_{k-1}, \boldsymbol{\theta}) \quad (1)$$

$$\boldsymbol{z}_k = \mathcal{K}(\boldsymbol{x}_k), \quad (2)$$

where $\boldsymbol{x}$ denotes the state of the system, the model operator $\mathcal{M}$ solves the equations that describe the dynamics of the system, k is the index of the time, $\boldsymbol{z}$ are the variables that can be observed, $\mathcal{K}$ is the observation operator and $\boldsymbol{\theta}$ denotes the (closure) parameters. The model operator $\mathcal{M}$ is assumed to contain everything that is needed to simulate the system, including also as external forcing terms and boundary conditions. In the real-world setting, we would like to tune parameters $\boldsymbol{\theta}$ of the model in Eqs. (1)–(2) using a set of available observations $\boldsymbol{y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ taken at some time instances $\{t_1, \ldots, t_n\}$. Note that $\boldsymbol{y}$ are measured values, while $\boldsymbol{z}$ are simulated values of the same variables.

In parameter estimation, we follow the Bayesian methodology, in which the knowledge about the unknown parameters is inferred from the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta}), \quad (3)$$

which is evaluated using the prior $p(\boldsymbol{\theta})$ and the likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$. The likelihood function specifies how plausible the observed data are given model parameter values. Therefore, defining a proper likelihood function is the central problem in parameter estimation. The prior contains the information that we have about the parameters based on the accumulated information from the past. For an introduction to Bayesian estimation, see, for example, the book by Gelman et al. (2003).

Traditionally, parameters of dynamical systems are estimated by comparing model simulations to observed data using a measure such as a sum of squared differences between $\boldsymbol{z}$ and $\boldsymbol{y}$. This corresponds to the assumption that the observations are noisy realizations of the model values. The problem in applying these techniques directly to chaotic systems is that the dynamically changing model state $\boldsymbol{x}$ is not known exactly, and small errors in the state estimates can grow in an unpredictable manner, making direct comparisons of model simulations and observations meaningless over long time periods.

In this paper, we consider three ways to estimate the closure parameters of chaotic models. In the first approach, observations and model simulations are summarized in the form of statistics, which are typically some temporal and spatial averages of the data. The likelihood model is constructed in terms of the summary statistics such that model parameters producing statistics that are closer to the observed statistics would have higher likelihood. This kind of an approach is employed in climate model parameter estimation in several recent studies (Jackson et al., 2008; Järvinen et al., 2010; Sexton et al., 2011). In the summary statistics approach, the problem of chaotic behavior can be alleviated, since the statistics computed from long simulations are less dependent on the initial conditions than the specific values of the state variables.

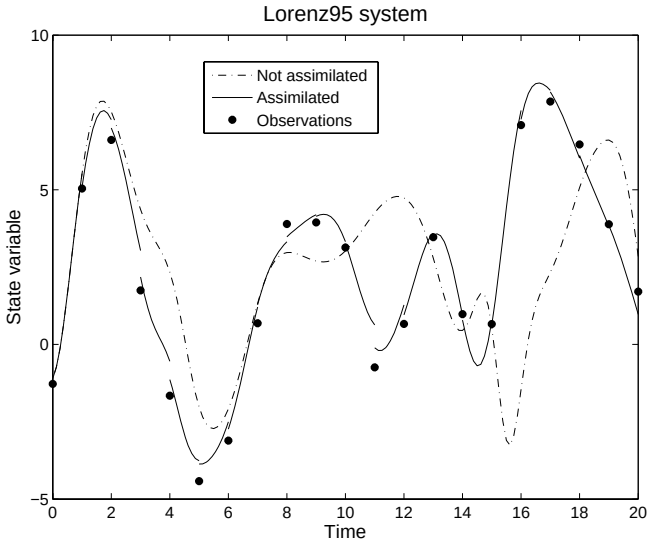


Fig. 1. An illustration of sequential data assimilation in a chaotic system. After some time the control run, even with optimal parameter values, gets "off track" due to chaos. Data assimilation keeps the model in the same trajectory with the data.

The other two approaches are based on embedding the parameter estimation techniques into dynamical state estimation (data assimilation) methods that constantly update the model state as new observations become available. Thus, the model is kept in the vicinity of the data, and the problems caused by chaotic behavior can be alleviated. This is illustrated in Fig. 1 by running the Lorenz system – that is used for experimentation in Sect. 5 – two times from the same initial values, with and without data assimilation. One can see that the model run without assimilation eventually deviates from the trajectory of the observations.

We consider two ways to implement parameter estimation within a data assimilation system. In the *state augmentation* approach (see Sect. 4), the model parameters are treated as artificial states and assimilated together with the actual model state (see, e.g. Kitagawa, 1998; Ionides et al., 2006; Dowd, 2011). In the *likelihood* approach, detailed in Sect. 3, the likelihood of a parameter value is evaluated by running a state estimation method over a chosen data set, keeping the parameter value fixed. The likelihood is constructed using the filter residuals (the squared differences between the observations and the short-range forecasts), see Fig. 1. This resembles classical parameter estimation, but the uncertainty in the model state is "integrated out" using a state estimation technique. The problem of chaoticity is circumvented by computing the likelihood components from short simulations, where chaotic behavior does not yet appear. The likelihood approach is a standard technique in

parameter estimation of stochastic models (see, e.g. Kivman, 2003; Singer, 2002; Dowd, 2011), but less studied in connection with deterministic, chaotic systems.

The paper is organized as follows. In Sect. 2, we present the methodology for the summary statistics approach. The likelihood approach is discussed in Sect. 3, and the sequential parameter estimation via state augmentation is presented in Sect. 4. In Sect. 5, the case study setup and numerical results are presented. In Sect. 6 we discuss some specific issues related to the properties of the methods. Section 7 concludes the paper.

2 Likelihood based on summary statistics

Several recent studies (Jackson et al., 2008; Järvinen et al., 2010; Sexton et al., 2011) on parameter estimation in climate models formulated the likelihood in terms of summary statistics. The advantage of this approach is that it is computationally feasible and rather straightforward to implement. It avoids the problem of chaotic behavior as in sufficiently long simulations the effect of the initial values diminishes.

In this approach, the observations are transformed to a set of summary statistics $\boldsymbol{s} = \boldsymbol{s}(\boldsymbol{y}_{1:n})$, where $\boldsymbol{y}_{1:n}$ denotes all observations for n steps of the simulation model Eqs. (1)–(2). The posterior distribution of model parameters $\boldsymbol{\theta}$ is evaluated as

$$p(\boldsymbol{\theta}|\boldsymbol{s}) \propto p(\boldsymbol{\theta})p(\boldsymbol{s}|\boldsymbol{\theta}). \tag{4}$$

The likelihood $p(\boldsymbol{s}|\boldsymbol{\theta})$ is constructed so that models $\boldsymbol{\theta}$ producing summary statistics $\boldsymbol{s}_{\boldsymbol{\theta}}$ which are close to the observed values $\boldsymbol{s}$ get higher probability. Here, $\boldsymbol{s}_{\boldsymbol{\theta}} = \boldsymbol{s}(\boldsymbol{z}_{\boldsymbol{\theta},1:n})$ denotes the summary statistics computed from data $\boldsymbol{z}_{\boldsymbol{\theta},1:n}$ simulated for n steps with model parameters $\boldsymbol{\theta}$. The approach is related to approximate Bayesian computation (ABC, see, e.g. Cornuet et al., 2008), where summary statistics are used to do statistical inference in situations where the exact likelihood is intractable.

2.1 Matching observed and simulated statistics

When the simulated and observed summary statistics are directly matched, the likelihood can be formulated, for instance, as

$$p(\boldsymbol{s}|\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}C(\boldsymbol{s},\boldsymbol{s}_{\boldsymbol{\theta}})\right), \tag{5}$$

where $C(\boldsymbol{s},\boldsymbol{s}_{\boldsymbol{\theta}})$ is a cost function penalizing the misfit between $\boldsymbol{s}$ and $\boldsymbol{s}_{\boldsymbol{\theta}}$. For example, one could use the Gaussian assumption yielding the form

$$C(\boldsymbol{s},\boldsymbol{s}_{\boldsymbol{\theta}}) = (\boldsymbol{s}-\boldsymbol{s}_{\boldsymbol{\theta}})^T \Sigma^{-1}(\boldsymbol{s}-\boldsymbol{s}_{\boldsymbol{\theta}}), \tag{6}$$

where the covariance matrix Σ takes into account possible correlations between the summary statistics. When some of the correlations are ignored (Eq. 6) becomes a sum of multiple terms. For example, in Järvinen et al. (2010) the cost function was similar to

$$C(\boldsymbol{s},\boldsymbol{s}_{\boldsymbol{\theta}}) = (s^{\mathrm{g}} - s^{\mathrm{g}}_{\boldsymbol{\theta}})^2/\sigma^2_{\mathrm{g}} + \sum_{t=1}^{12}\sum_{i,k}\left(\boldsymbol{s}^{ikt} - \boldsymbol{s}^{ikt}_{\boldsymbol{\theta}}\right)^T \Sigma^{-1}_{ikt}\left(\boldsymbol{s}^{ikt} - \boldsymbol{s}^{ikt}_{\boldsymbol{\theta}}\right), \tag{7}$$

where s^{g} is the annual global mean of the net radiative flux at the top of the atmosphere (TOA) and $\boldsymbol{s}^{ikt}$ are zonal and monthly averages of the k-th variable computed for latitude i and month t. The first term in Eq. (7) penalizes unstable models which have an unrealistic balance of the global-mean TOA radiation, whereas the second term ensures a realistic annual cycle for the radiation. The same statistics computed from simulated data are denoted by $s^{\mathrm{g}}_{\boldsymbol{\theta}}$ and $\boldsymbol{s}^{ikt}_{\boldsymbol{\theta}}$.

The goal of the studies by Järvinen et al. (2010) was to explore the uncertainty of the parameters which have a large effect on the radiative balance and therefore only the net radiative flux at TOA was used to compute the zonal and monthly averages $\mathbf{s}^{ikt}$ in Eq. (7). In Jackson et al. (2008), several variables were included in the cost function and the covariance matrix Σ^{ikt} was formulated in terms of a few leading empirical orthogonal functions (EOFs).

One problem of the direct matching of observed and simulated statistics is that the resulting likelihood Eq. (5) may not be a smooth function of the parameters, as will be seen in the experimental results (e.g. Fig. 2). A possible reason for this are the random effects caused by the finite length of the simulations. The noise in the objective function may complicate the parameter estimation procedure. Another problem is that the matching approach is not based on a well justified statistical model for the summary statistics: it is not easy to define what values for the summary statistics are "good" in the statistical sense. For example, it is not straightforward to select the scaling parameter Σ_{ikt} in Eq. (7).

2.2 Fitting a probabilistic model for summary statistics

The problems mentioned above can be partly overcome by building a probabilistic model for the summary statistics. The summary statistics are treated as random variables which are systematically affected by varying the model parameters $\boldsymbol{\theta}$:

$$\boldsymbol{s}_{\boldsymbol{\theta}} = \mathbf{f}(\boldsymbol{\theta}) + \boldsymbol{\epsilon}, \tag{8}$$

where function $\mathbf{f}$ is often called an emulator or a surrogate model (Rougier, 2008; Sexton et al., 2011) and the noise term $\boldsymbol{\epsilon}$ can be assumed Gaussian with zero mean and covariance matrix Σ. The emulator and the noise model can be estimated from training samples which are pairs $\{\boldsymbol{s}_{\boldsymbol{\theta}_i}, \boldsymbol{\theta}_i\}$ of simulated statistics $\boldsymbol{s}_{\boldsymbol{\theta}_i}$ and parameter values $\boldsymbol{\theta}_i$ used in the simulations. This is a nonlinear regression problem which have been studied intensively in statistics and machine learning (see, e.g. Bishop, 2006). Examples of parametric models

for $\mathbf{f}$ include polynomials, radial basis function networks and multi-layer perceptrons.

Thus, prior to constructing the likelihood, the model has to be simulated many times with different parameter values selected over a suitable range. This is computationally expensive, but a necessary step. The likelihood is then constructed as

$$\boldsymbol{s}|\boldsymbol{\theta} \sim N(\mathbf{f}(\boldsymbol{\theta}), \Sigma). \tag{9}$$

One difficulty of learning Eq. (8) is that the set of all possible statistics in $\boldsymbol{s}_{\boldsymbol{\theta}}$ is highly multidimensional while the number of examples to train the emulator Eq. (8) is potentially limited because of the high computational costs of model simulations. This problem can be solved by reducing the number of summary statistics somehow before training the model Eq. (8). The simplest way is to consider only a linear combination of the summary statistics, which means neglecting the variability of summary statistics outside a selected subspace. Thus, Eq. (8) is replaced by

$$\mathbf{W}\boldsymbol{s} = \mathbf{f}_*(\boldsymbol{\theta}) + \boldsymbol{\epsilon}_*, \tag{10}$$

where $\mathbf{W}$ is a properly chosen matrix and the likelihood is formulated in terms of the projected data:

$$\mathbf{W}\boldsymbol{s}|\boldsymbol{\theta} \sim N(\mathbf{f}_*(\boldsymbol{\theta}), \Sigma_*). \tag{11}$$

Sexton et al. (2011), computed $\mathbf{W}$ using principal component analysis (PCA) and the dimensionality of the summary statistics was reduced from 175 000 to only six. Thus, the criterion for dimensionality reduction used there was the maximum amount of variance retained in the projected statistics. Another possible approach is to find projections $\mathbf{W}\boldsymbol{s}$ which are most informative about closure parameters $\boldsymbol{\theta}$. For example, canonical correlation analysis is mentioned as a more appropriate dimensionality method by Rougier (2008). In the experiments presented in Sect. 5.2.1, we find $\mathbf{W}$ by fitting a linear regression model $\boldsymbol{\theta} \approx \mathbf{W}\boldsymbol{s}_{\boldsymbol{\theta}}$ to the training data $\{z_{\boldsymbol{\theta}_i}, \boldsymbol{\theta}_i\}$. The emulator (10) is then estimated for the projected statistics. Note that one can analyze the elements of $\mathbf{W}$ computed to maximize correlations between $\boldsymbol{\theta}$ and $\boldsymbol{s}_{\boldsymbol{\theta}}$ in order to have a clue on which summary statistics are affected by varying the closure parameters.

3 Likelihood with filtering methods

Traditionally, the likelihood for a parameter in non-chaotic dynamical models is calculated by comparing the model to data using a goodness-of-fit measure, such as the sum of squared differences between the model and the observations. In many cases, however, the model state is not known accurately and it has to be estimated together with the parameters. This is especially important with chaotic models, where small errors in the model state can grow quickly when the model is integrated in time.

State estimation in dynamical models can be carried out using filtering methods, where the distribution of the model state is evolved with the dynamical model and sequentially updated as new observations become available. When static parameters are estimated, filtering can be used to "keep the model on track" with the measurements. In this section, we present how the likelihood in chaotic models can be computed using filtering methods. First, we present the general filtering formulas and then consider the special case of Gaussian filters.

3.1 General formulas

Let us consider the following discrete state space model at time step k with unknown parameters $\boldsymbol{\theta}$:

$$\boldsymbol{x}_k \sim p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}, \boldsymbol{\theta}) \tag{12}$$

$$\boldsymbol{y}_k \sim p(\boldsymbol{y}_k|\boldsymbol{x}_k) \tag{13}$$

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta}). \tag{14}$$

Thus, in addition to the unknown, dynamically changing model state $\boldsymbol{x}_k$, we have static parameters $\boldsymbol{\theta}$, from which we have some prior information $p(\boldsymbol{\theta})$. As mentioned in the introduction, the goal in parameter estimation, in Bayesian terms, is to find the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y}_{1:n})$ of the parameters, given a fixed data set $\boldsymbol{y}_{1:n}$. Here, the notation $\boldsymbol{y}_{1:n} = \{\boldsymbol{y}_1, ..., \boldsymbol{y}_n\}$ means all observations for n time steps.

Filtering methods (particle filter, Kalman filters, etc.) estimate the dynamically changing model state sequentially. They give the marginal distribution of the state given the measurements obtained until the current time k. Thus, for a given value for $\boldsymbol{\theta}$, filtering methods estimate $p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k}, \boldsymbol{\theta})$.

Filters work by iterating two steps: prediction and update. In the prediction step, the current distribution of the state is evolved with the dynamical model to the next time step:

$$p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k-1}, \boldsymbol{\theta}) = \int p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}, \boldsymbol{\theta}) \times p(\boldsymbol{x}_{k-1}|\boldsymbol{y}_{1:k-1}, \boldsymbol{\theta}) d\boldsymbol{x}_{k-1}. \tag{15}$$

When the new observation $\boldsymbol{y}_k$ is obtained, the model state is updated using the Bayes' rule with the predictive distribution $p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k-1}, \boldsymbol{\theta})$ as the prior:

$$p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k}, \boldsymbol{\theta}) \propto p(\boldsymbol{y}_k|\boldsymbol{x}_k, \boldsymbol{\theta}) p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k-1}, \boldsymbol{\theta}). \tag{16}$$

This posterior is used inside the integral (15) to obtain the prior for the next time step.

Using the state posteriors obtained in the filtering method, it is also possible to compute the predictive distribution of the next observation. For observation $\boldsymbol{y}_k$, the predictive distribution, given all previous observations, can be written as

$$p(\boldsymbol{y}_k|\boldsymbol{y}_{1:k-1}, \boldsymbol{\theta}) = \int p(\boldsymbol{y}_k|\boldsymbol{x}_k, \boldsymbol{\theta}) p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k-1}, \boldsymbol{\theta}) d\boldsymbol{x}_k. \tag{17}$$

The latter term $p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k-1}, \boldsymbol{\theta})$ in the integral is the predictive distribution given by Eq. (15).

Let us now proceed to the original task of estimating static parameters $\boldsymbol{\theta}$ from observations $\boldsymbol{y}_{1:n}$, i.e., computing the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y}_{1:n})$. Applying the Bayes' formula and the chain rule for joint probability, we obtain

$$p(\boldsymbol{\theta}|\boldsymbol{y}_{1:n}) \propto p(\boldsymbol{y}_{1:n}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \quad (18)$$

$$= p(\boldsymbol{y}_n|\boldsymbol{y}_{1:n-1},\boldsymbol{\theta})p(\boldsymbol{y}_{n-1}|\boldsymbol{y}_{1:n-2},\boldsymbol{\theta})\ldots \times p(\boldsymbol{y}_2|\boldsymbol{y}_1,\boldsymbol{\theta})p(\boldsymbol{y}_1|\boldsymbol{\theta})p(\boldsymbol{\theta}). \quad (19)$$

In the filtering context, the predictive distributions $p(\boldsymbol{y}_i|\boldsymbol{y}_{1:i-1},\boldsymbol{\theta})$ are calculated based on the marginal posterior of the model states, see Eq. (17).

Thus, the likelihood of the whole data $\boldsymbol{y}_{1:n}$ can be calculated as the product of the predictive distributions of the individual observations. That is, to check how well the model with parameter vector $\boldsymbol{\theta}$ fits the observations, one can check how individual predictions made from the current posterior fit the next observations. The only difference to traditional model fitting is that the state distribution is updated after each measurement.

Note that the above analysis only tells how the parameter likelihood is related to filtering methods. We have not yet discussed how the parameter estimation can be implemented in practice. In order to obtain the parameter estimates, two steps are required: (a) a filtering method to compute the posterior density for a given parameter value and (b) a parameter estimation algorithm to obtain the estimates. In this paper, we use variants of the Kalman filter for task (a), but other filtering methods, such as particle filters (see, e.g. Cappe et al., 2007), could be applied as well. For task (b) there are several optimization and Monte Carlo approaches available, and the method of choice depends on the case. In our examples, we use three different methods to estimate the posterior distribution: a maximum a posteriori (MAP) optimization approach with a Gaussian approximation of the posterior, a Markov chain Monte Carlo (MCMC) algorithm, and an importance sampling approach. For the sake of completeness, these algorithms are briefly reviewed in Appendix A.

Next, we will present how the parameter estimation is performed in the more familiar case, where the distributions are assumed to be Gaussian, and the extended Kalman filter (EKF) is used as the filtering method.

3.2 EKF likelihood

The extended Kalman filter is one of the most popular methods for state estimation in nonlinear dynamical models. EKF is an extension to the Kalman filter (KF, Kalman, 1960), where the model is assumed to be linear.

Let us now write the state space model as follows:

$$\boldsymbol{x}_k = \mathcal{M}(\boldsymbol{x}_{k-1},\boldsymbol{\theta}) + \boldsymbol{E}_k \quad (20)$$

$$\boldsymbol{y}_k = \mathcal{K}(\boldsymbol{x}_k) + \boldsymbol{e}_k. \quad (21)$$

Unlike in the standard EKF, the model $\mathcal{M}$ now depends on parameters θ. The model and observation errors are assumed to be zero mean Gaussians: $\boldsymbol{E}_k \sim N(\boldsymbol{0}, \mathbf{C}_k^{\mathrm{E}})$ and $\boldsymbol{e}_k \sim N(\boldsymbol{0}, \mathbf{C}_k^{\mathrm{e}})$.

In KF, the prediction and update steps can be written down analytically, since everything is linear and Gaussian. EKF uses the KF formulas, but the model matrix in KF is replaced by a linearization of the nonlinear model. In EKF, the predictive distribution for the state at time k is

$$\boldsymbol{x}_k|\boldsymbol{y}_{1:k-1},\boldsymbol{\theta} \sim N(\boldsymbol{x}_k^p, \mathbf{C}_k^p), \quad (22)$$

where $\boldsymbol{x}_k^p = \mathcal{M}(\boldsymbol{x}_{k-1}^{\mathrm{est}},\boldsymbol{\theta})$ is the posterior mean $\boldsymbol{x}_{k-1}^{\mathrm{est}}$ from the previous time evolved with the model. The prior covariance $\mathbf{C}_k^p = \mathbf{M}_k^{\boldsymbol{\theta}}\mathbf{C}_{k-1}^{\mathrm{est}}\mathbf{M}_k^{\boldsymbol{\theta}^T} + \mathbf{C}_k^{\mathrm{E}}$ is the covariance of the state estimate evolved with the linearized model $\mathbf{M}_k^{\boldsymbol{\theta}} = \partial\mathcal{M}(\boldsymbol{x}_{k-1}^{\mathrm{est}},\boldsymbol{\theta})/\partial\boldsymbol{x}_{k-1}^{\mathrm{est}}$.

In the update step, the prior in Eq. (22) is updated with the new observation $\boldsymbol{y}_k$. In the EKF formulation, the posterior is

$$\boldsymbol{x}_k|\boldsymbol{y}_{1:k},\boldsymbol{\theta} \sim N(\boldsymbol{x}_k^{\mathrm{est}}, \mathbf{C}_k^{\mathrm{est}}), \quad (23)$$

where $\boldsymbol{x}_k^{\mathrm{est}}$ and $\mathbf{C}_k^{\mathrm{est}}$ are given by the Kalman filter formulas:

$$\boldsymbol{x}_k^{\mathrm{est}} = \boldsymbol{x}_k^p + \mathbf{G}_k(\boldsymbol{y}_k - \mathcal{K}(\boldsymbol{x}_k^p)) \quad (24)$$

$$\mathbf{C}_k^{\mathrm{est}} = \mathbf{C}_k^p - \mathbf{G}_k\mathbf{K}_k\mathbf{C}_k^p. \quad (25)$$

Here $\mathbf{G}_k = \mathbf{C}_k^p\mathbf{K}_k^T(\mathbf{K}_k\mathbf{C}_k^p\mathbf{K}_k^T + \mathbf{C}_k^{\mathrm{e}})^{-1}$ is the Kalman gain matrix and $\mathbf{K}_k = \partial\mathcal{K}(\boldsymbol{x}_k^p)/\partial\boldsymbol{x}_k^p$ is the linearized observation operator. The predictive distribution of measurement $\boldsymbol{y}_k$, needed in the likelihood evaluation, is given by

$$\boldsymbol{y}_k|\boldsymbol{y}_{1:k-1},\boldsymbol{\theta} \sim N(\mathcal{K}(\boldsymbol{x}_k^p), \mathbf{C}_k^y), \quad (26)$$

where $\mathbf{C}_k^y = \mathbf{K}\mathbf{C}_k^p\mathbf{K}^T + \mathbf{C}_k^{\mathrm{e}}$. Now, applying the general formula (19), the total likelihood of observing $\boldsymbol{y}_{1:n}$, given parameters $\boldsymbol{\theta}$, can be written as

$$p(\boldsymbol{y}_{1:n}|\boldsymbol{\theta}) = p(\boldsymbol{y}_1|\boldsymbol{\theta})\prod_{k=2}^{n} p(\boldsymbol{y}_k|\boldsymbol{y}_{1:k-1},\boldsymbol{\theta})$$

$$= \prod_{k=1}^{n}\exp\left(-\frac{1}{2}\boldsymbol{r}_k^T(\mathbf{C}_k^y)^{-1}\boldsymbol{r}_k\right)(2\pi)^{-d/2}|\mathbf{C}_k^y|^{-1/2}$$

$$\propto \exp\left(-\frac{1}{2}\sum_{k=1}^{n}\left[\boldsymbol{r}_k^T(\mathbf{C}_k^y)^{-1}\boldsymbol{r}_k + \log|\mathbf{C}_k^y|\right]\right), \quad (27)$$

where $\boldsymbol{r}_k = \boldsymbol{y}_k - \mathcal{K}(\boldsymbol{x}_k^p)$ and $|\cdot|$ denotes the determinant. Note that the normalization constants of the likelihood terms depend on θ implicitly through the covariances $\mathbf{C}_k^p$, and the term $\log|\mathbf{C}_k^y| = \log|\mathbf{K}_k\mathbf{C}_k^p\mathbf{K}_k^T + \mathbf{C}_k^{\mathrm{e}}|$ therefore needs to be included.

The above likelihood resembles the traditional "least squares" type of Gaussian likelihoods. The difference is that the model state is allowed to change between time steps, and the residuals are weighted by the model prediction uncertainty term $\mathbf{K}_i\mathbf{C}_i^p\mathbf{K}_i^T$ in addition to the measurement error covariance $\mathbf{C}_i^{\mathrm{e}}$. In fact, removing the model uncertainty terms $\mathbf{K}_i\mathbf{C}_i^p\mathbf{K}_i^T$ reduces the likelihood to the classical Gaussian

likelihood often used in parameter estimation. Adding the prediction error covariances to the sum of squares terms essentially gives more weight to situations that are predictable, and down-weights the terms where the model prediction is uncertain, due to, e.g. chaotic behavior of the system.

If, in addition to the actual model parameters $\boldsymbol{\theta}$, the model error covariance $\mathbf{C}_{\mathrm{E}}$ is unknown, we can parameterize it and estimate its parameters from the measurements together with the model parameters. As discussed later in more detail, the ability to estimate the variance parameters is one of the advantages of the likelihood approach, compared to the state augmentation and summary statistics methods.

Unfortunately, as the dimension of the model increases, EKF soon becomes practically infeasible. An approximation to EKF that can be implemented in large-scale systems is the ensemble Kalman filter (EnKF). In EnKF and its numerous variants (see, e.g. Evensen, 2007; Ott et al., 2004; Whitaker and Hamill, 2002; Zupanski, 2005), the computational issues in EKF are circumvented by using sample statistics in EKF formulas, computed from a relatively small number of ensembles. Hence, when EnKF is used, the likelihood (27) can be computed simply by defining $\boldsymbol{x}_i^p$ and $\mathbf{C}_i^p$ as sample mean and covariance matrix estimated from the ensemble.

Most implementations of the EnKF involve random perturbations of the model states and observations. This introduces randomness in the likelihood function (27): two evaluations with the same parameter value give different likelihood values. As noted by Dowd (2011), this complicates the parameter inference, and one has to resort to stochastic optimization methods that can handle noise in the target function (see Shapiro et al., 2009, for an introduction). Note that some recent variants of EnKF, such as many of the so called ensemble square root filters (Tippett et al., 2003) do not involve random components, and they might be more suitable for parameter estimation purposes. We test a variant called Local Ensemble Transform Kalman Filter (LETKF, Hunt et al., 2007) in the experiments of Sect. 5.

4 Parameter estimation with state augmentation

In the previous section, the parameter estimation was carried out off-line by repeatedly sweeping through the data using a filter. In state augmentation (SA), the parameters are added to the state vector and estimated on-line in the filter. In practice this means that model parameters are updated together with the state, whenever new observations become available. Next, we will present how SA can be implemented with EKF.

Let us consider the following state space model, where the parameter vector is modeled as an additional dynamical variable:

$$\boldsymbol{x}_{k+1} = \mathcal{M}(\boldsymbol{x}_k, \boldsymbol{\theta}_k) + \boldsymbol{E}_{\boldsymbol{x}} \tag{28}$$

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \boldsymbol{E}_{\boldsymbol{\theta}} \tag{29}$$

$$\boldsymbol{y}_{k+1} = \mathcal{K}(\mathbf{x}_{k+1}) + \boldsymbol{e}. \tag{30}$$

For notational convenience, we have dropped the time index k from the error terms.

In SA, we treat the combined vector $\boldsymbol{s}_k = [\boldsymbol{x}_k, \boldsymbol{\theta}_k]^T$ as the state vector that is updated at each time step k. The model for the combined vector can be written as

$$\boldsymbol{s}_{k+1} = \tilde{\mathcal{M}}(\boldsymbol{s}_k) + \boldsymbol{E}_{\boldsymbol{x},\boldsymbol{\theta}}, \tag{31}$$

where $\tilde{\mathcal{M}}(\boldsymbol{s}_k) = [\mathcal{M}(\boldsymbol{x}_k, \boldsymbol{\theta}_k), \boldsymbol{\theta}_k]^T$ and $\boldsymbol{E}_{\boldsymbol{x},\boldsymbol{\theta}}$ is the error of the augmented model $\tilde{\mathcal{M}}$, here assumed to be zero mean Gaussian with covariance matrix $\mathbf{C}_{\boldsymbol{x},\boldsymbol{\theta}}$.

In EKF, we now need to linearize $\tilde{\mathcal{M}}(\boldsymbol{s}_k)$ with respect to $\boldsymbol{s}_k$, which results in the following Jacobian matrix:

$$\mathbf{M}_k = \frac{\partial \tilde{\mathcal{M}}(\boldsymbol{s}_k)}{\partial \boldsymbol{s}_k} = \begin{bmatrix} \partial \mathcal{M}(\boldsymbol{s}_k)/\partial \boldsymbol{x}_k & \partial \mathcal{M}(\boldsymbol{s}_k)/\partial \boldsymbol{\theta}_k \\ \partial \boldsymbol{\theta}_k/\partial \boldsymbol{x}_k & \partial \boldsymbol{\theta}_k/\partial \boldsymbol{\theta}_k \end{bmatrix} \tag{32}$$

$$= \begin{bmatrix} \partial \mathcal{M}(\boldsymbol{s}_k)/\partial \boldsymbol{x}_k & \partial \mathcal{M}(\boldsymbol{s}_k)/\partial \boldsymbol{\theta}_k \\ \mathbf{0} & \mathbf{I} \end{bmatrix}. \tag{33}$$

Now, this matrix $\mathbf{M}_k$ can be used in the EKF formulas. Note that the top left term in the matrix is the linearization with respect to the actual states, which is needed in the standard states-only EKF as well. In addition, the derivative with respect to the parameters is needed (the top right term).

In EKF, we also need to define the model error covariance matrix $\mathbf{C}_{\boldsymbol{x},\boldsymbol{\theta}}$. In SA, this must be defined in the joint space of the state and the parameters. The errors in the state and parameters are hardly uncorrelated, but for simplicity we model them here as independent random variables, which yields a block diagonal error covariance matrix

$$\mathbf{C}_{\boldsymbol{x},\boldsymbol{\theta}} = \begin{bmatrix} \mathbf{C}_{\boldsymbol{x}} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{\boldsymbol{\theta}} \end{bmatrix}. \tag{34}$$

The model error in the state, $\mathbf{C}_{\boldsymbol{x}}$, has a clear interpretation: it represents the statistical properties of the error that the model makes in a filter time step. However, the parameter error covariance matrix $\mathbf{C}_{\boldsymbol{\theta}}$ lacks such an interpretation. We consider $\mathbf{C}_{\boldsymbol{\theta}}$ as an additional tuning parameter of the SA approach. Roughly speaking, increasing $\mathbf{C}_{\boldsymbol{\theta}}$ allows more sudden changes from $\boldsymbol{\theta}_k$ to $\boldsymbol{\theta}_{k+1}$. Note that, unlike in the full likelihood approach, the model error covariance $\mathbf{C}_{\boldsymbol{x}}$ cannot be estimated from data using SA. A simple example illustrating this problem is shown by DelSole and Yang (2010). The effect of (and the sensitivity to) $\mathbf{C}_{\boldsymbol{x},\boldsymbol{\theta}}$ is studied in more detail in the experimental section. In Appendix B, we give some theoretical discussion of the effect of the selected $\mathbf{C}_{\boldsymbol{x},\boldsymbol{\theta}}$.

As in the likelihood approach, the SA parameter estimation method can be implemented using other filtering methods besides EKF. For instance, replacing EKF with EnKF is straightforward: the ensembles now contain perturbations of both the model states and the parameters. The EnKF SA approach has been implemented, for instance, for a marine biochemistry model by Dowd (2011) and for an atmospheric model by Annan et al. (2005).

Conceptually, SA is straightforward, although, like noted by Järvinen et al. (2010), it implicitly assumes the static parameters as dynamical quantities and the parameter estimates therefore change at every update step of the filter. In some applications, such as numerical weather prediction (NWP), this may be critical. Operational NWP systems perform under strict requirements of timely product delivery to end-users. The "drifting" model parameters have to be therefore carefully considered from the system reliability point-of-view.

5 Case study: parametrized Lorenz 95

In this section, we will demonstrate the discussed parameter estimation approaches using a modified Lorenz system. We start by describing the model and the experiments, and then present the results for the three different methods.

5.1 Description of the experiment

To demonstrate and compare the parameter estimation approaches, we use a modified version of the Lorenz 95 ODE system, detailed by Wilks (2005). The chaotic Lorenz model (Lorenz, 1995) is commonly used as a low order test model to study estimation algorithms. The system used here is similar to the original system, but the state variables x_i are affected by forcing due to fast variables y_j, too. The full system is written as

$$\frac{dx_k}{dt} = -x_{k-1}(x_{k-2} - x_{k+1}) - x_k + F - \frac{hc}{b}\sum_{j=J(k-1)+1}^{Jk} y_j \tag{35}$$

$$\frac{dy_j}{dt} = -cby_{j+1}\left(y_{j+2} - y_{j-1}\right) - cy_j + \frac{c}{b}F_y + \frac{hc}{b}x_{1+\lfloor \frac{j-1}{J} \rfloor} \tag{36}$$

where $k = 1,\ldots,K$ and $j = 1,\ldots,JK$. That is, each of the "slow" state variables x_i are forced by a sum of the additional fast variables y_j. The fast variables have dynamics similar to the slow variables, but they are also coupled with the slow variables. We use values $K = 40$, $J = 8$, $F = F_y = 10$, $h = 1$ and $c = b = 10$, adopted from (Leutbecher, 2010).

The system (35)–(36) is considered as the "truth" and used for generating synthetic data. As a forecast model, we use a version where the net effect of the fast variables is described using a deterministic parameterization. The forecast model reads as

$$\frac{dx_k}{dt} = -x_{k-1}(x_{k-2} - x_{k+1}) - x_k + F - g(x_k, \boldsymbol{\theta}), \tag{37}$$

where $g(x_k, \boldsymbol{\theta})$ is the parameterization in which the missing fast variables y_j are modeled using the "resolved" variables. Here, we use a polynomial parameterization,

$$g(x_k, \boldsymbol{\theta}) = \sum_{i=0}^{d} \theta_i x_k^{(i)}, \tag{38}$$

with $d = 1$. The goal is to "tune" the parameters $\boldsymbol{\theta}$ so that the model fits the observations as well as possible. The parameter estimation resembles the closure parameter estimation problem in atmospheric models: the forecast model is solved with a time step $\Delta t = 0.025$, which is too crude for modeling the fast variables that operate on a finer time scale.

The observations for parameter estimation are generated as follows. The model is solved with dense time stepping ($\Delta t = 0.0025$) for altogether 2500 days (in the Lorenz model, one day corresponds to 0.2 time units). Then Gaussian noise is added to the model output with zero mean and covariance $(0.1\sigma_{\text{clim}})^2\mathbf{I}$, where $\sigma_{\text{clim}} = 3.5$ (standard deviation from long simulations). When the parameters are estimated using the filtering approaches, only 24 out of the 40 slow variables are assumed to be observed each day. The observation operator, used also in previous data assimilation studies (Auvinen et al., 2009, 2010), picks the last three state variables from every set of five states and we thus observe states $3, 4, 5, 8, 9, 10, \ldots, 38, 39, 40$. Partial observations were assumed to emulate a realistic data assimilation setting. In the experiments with the summary statistics approach, all the 40 states are assumed to be observed because hiding some of the states would introduce problems in the computation of the statistics.

Note that with this set-up, it is possible to use the values of the fast variables y_j simulated in the full system (35)–(36) to estimate parameters θ of the forcing model

$$g(x_k, \boldsymbol{\theta}) \approx \frac{hc}{b}\sum_{j=J(k-1)+1}^{Jk} y_j\,.$$

We will use the term "reference parameter values" for $\boldsymbol{\theta}$ which minimize the errors of this forcing model in the least squares sense. Naturally, such fitting cannot be performed in real applications since the actual sub-grid scale forcing is not known.

5.2 Results

In this section, we will present the results using the summary statistics, likelihood and state augmentation approaches. Our emphasis is on comparing the accuracy and the properties of these different approaches.

5.2.1 Summary statistics

In a synthetic example like Lorenz 95, summarizing the data in the form of a few statistics is not a trivial task. For example, if one wants to repeat the parameter estimation procedure similarly to Järvinen et al. (2010); Sexton et al. (2011), it is not clear what would be a suitable counterpart for the zonal and monthly averages in Eq. (7).

Matching observed and simulated statistics

As mentioned in Sect. 2.1, the simulated statistics may not be a smooth function of the parameters. One method that is rather insensitive to that kind of behavior in the likelihood is importance sampling (see Appendix A3 for details). We perform importance sampling for the two parameters $[\theta_0, \theta_1]$ of the model Eqs. (37)–(38). First, we draw 1000 candidate values uniformly and independently for the two parameters from the intervals $\theta_0 \in [1.4, 2.2]$ and $\theta_1 \in [0, 0.12]$. Then, the system defined by Eqs. (37)–(38) is simulated for each candidate value, the summary statistics are computed and the likelihood is calculated. The parameter ranges were chosen so that the shape of the posterior distribution is clearly visible.

In the first experiment, the cost function was constructed around a set of summary statistics which were selected arbitrarily. We used six statistics: mean, variance, autocovariance with time lag 1, covariance of a node with its neighbor and cross-covariance of a node with its two neighbors for time lag 1. Since the model is symmetric with respect to the nodes, we averaged these statistics across different nodes. The cost function was

$$C(\hat{\boldsymbol{s}}, \boldsymbol{s}_{\boldsymbol{\theta}}) = \sum_{i=1}^{6} (\hat{s}^i - s^i_{\boldsymbol{\theta}})^2 / \hat{\sigma}_i^2, \tag{39}$$

where $s^i_{\boldsymbol{\theta}}$ is one of the six statistics computed from data simulated with parameters $\boldsymbol{\theta}$, and $\hat{s}^i$, $\hat{\sigma}_i^2$ are the mean and variance of the same statistics computed from a relatively long simulation of the full-system (35)–(36) similarly to (Järvinen et al., 2010) . All the 40 variables x_k were assumed to be observed and the observation period was taken to be 200 days.

Figure 2 shows the importance sampling weights obtained for the 1000 candidate values. The results show that the cost function (39) does not restrict the model parameters much and the posterior distribution is rather broad. The parameter estimates are also clearly biased: the values obtained using the knowledge of the simulated fast variables are outside the obtained distribution. Note also the "spiky" behavior of the cost function: the weights do not vary smoothly as a function of the parameters.

Likelihood based on an emulator

In the next experiment, the likelihood was computed using an emulator trained on the same set of six summary statistics. We again performed importance sampling of parameters θ_0 and θ_1 of the model (37)–(38) using the same 1000 candidate values drawn from $\theta_0 \in [1.4, 2.2]$ and $\theta_1 \in [0, 0.12]$. The likelihood Eq. (11) was constructed using an emulator, as explained in Sect. 2.2.

Figure 3a presents the results for the likelihood Eq. (11) in which the dimensionality reduction was performed using PCA with only two principal components retained. Figure 3b

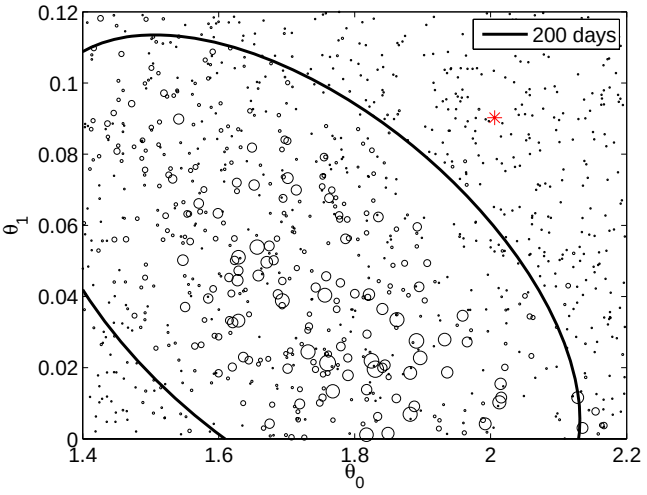


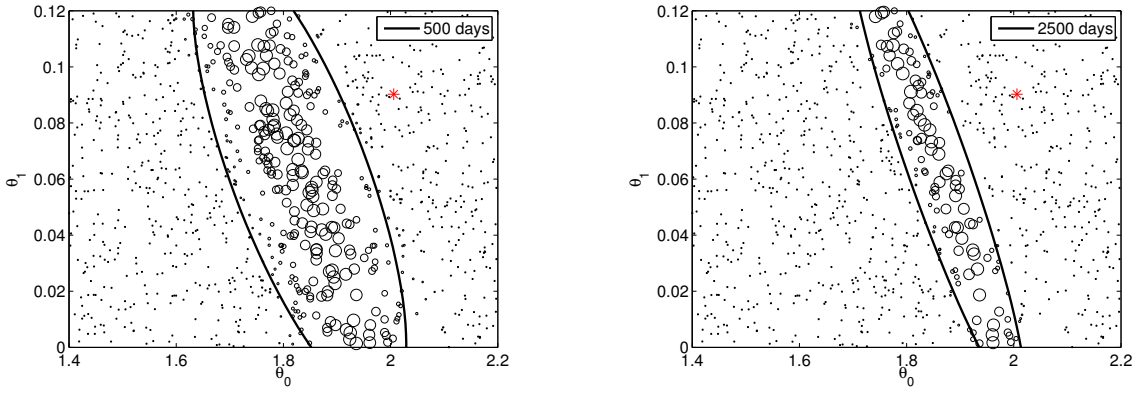
Fig. 2. The scatter plots of the parameter value candidates and their likelihood (represented by the size of the markers) obtained with cost function (39). The marker size reflects the value of the weights for importance sampling. The black ellipse represents the first two moments estimated with importance sampling. The red star represents the parameter values estimated by fitting the forcing model to the simulated fast variables (see Sect. 5 for details).

presents similar results for the case when the dimensionality of the features was reduced from six to two by fitting a linear regression model $\boldsymbol{\theta} \approx \mathbf{W}\boldsymbol{s}_{\boldsymbol{\theta}}$ and by using a feed-forward neural network (see, e.g. Bishop, 2006) to build an emulator.

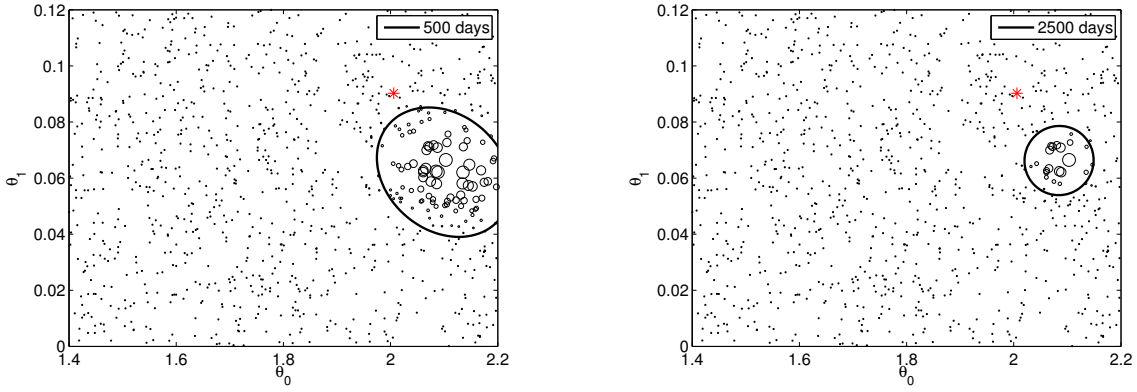
There are a few remarks that we can make based on the obtained results. Using the emulator results in a posterior density which is a smooth function of the parameters, thus the problem of the spiky behavior is solved. The parameters found with this approach are close to the reference values but there is a bias which is affected by the choice of the summary statistics. This effect is related to the known result from the field of Approximate Bayesian Computation that only using *sufficient* summary statistics yields the same posterior distribution as when the full data set is used (Marjoram and Tavaré, 2006). A longer observational period results in a more narrow posterior distribution but the bias problem remains.

The results are generally sensitive to the dimensionality reduction approach and to the number of components retained in the model. In this simple example, using more or less components leads to qualitatively similar results (biased estimates). In more complex cases, a cross-validation procedure (e.g. similar to Sexton et al., 2011) might be needed to estimate the right number of the required components.

Another problem is that the observed values of some of the summary statistics cannot be obtained by varying the parameters of the surrogate model. This situation can easily occur in real model tuning and it may result in over-fitting of the parameters to such problematic statistics. In the results shown in Fig. 3a, this problem is concealed by the fact that



(a) Dimensionality reduction by PCA



(b) Dimensionality reduction by fitting a linear regression model $\theta \approx \mathbf{W}\mathbf{s}_\theta$

Fig. 3. The scatter plots of the parameter value candidates and their likelihood (represented by the size of the markers) obtained with cost function (11). The simulation length is 500 days (left) and 2500 days (right). The marker size reflects the value of the weights for importance sampling. The black ellipses represent the first two moments estimated with importance sampling. The red start represents the parameter values estimated by fitting the forcing model to the simulated fast variables (see Sect. 5 for details).

only two principal components are retained in the analysis and those principal components can be simulated well by the surrogate model.

The summary statistics approach has a few other potential problems. The choice of the summary statistics has a critical impact on the estimated parameter values. Some arbitrarily selected statistics may not be affected by varying the model parameters and the idea behind the most informative projections is to diminish this problem. In the tuning of climate models, summary statistics often include only monthly and regional averages of some state variables, which means the focus is on how well climate models reproduce the seasonal cycle. It may be that some model parameters have little effect on the seasonal cycle but they can be important for the overall quality of a climate model.

Thus, the summary statistics approach has a few practical problems and can result in biased estimates. We think that the essential problem is the averaging procedure in which a lot of important information is lost. We argue that the sequential methods provide a more appropriate way to determine the likelihood for the parameters.

5.2.2 Likelihood calculations using filtering methods

In this section, we estimate the parameters of the forecast model (37) for the Lorenz 95 system using the filtering methodology presented in Sect. 3.

Filtering with EKF

Implementation of EKF requires linearization of the model (37), which is rather straightforward in this synthetic example. As mentioned in Sect. 3.2, the EKF filtering procedure also requires the model error covariance matrix $\mathbf{C}_\mathrm{E}$. We use a simple parameterization:

$$\mathbf{C}_\mathrm{E} = \sigma^2 \mathbf{I}, \tag{40}$$

where σ^2 is a parameter which needs to be estimated together with parameters $\boldsymbol{\theta}$ of the forecast model (37). In practice, we estimate the logarithm of σ^2, which guarantees the positivity of the variance and yields a posterior distribution whose shape is closer to Gaussian. We perform parameter estimation using delayed rejection adaptive Metropolis (DRAM) MCMC (Haario et al., 2006, Appendix A).

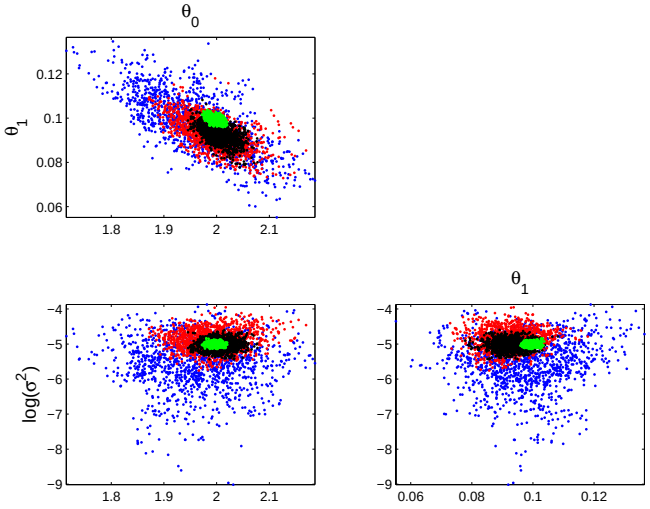


Fig. 4. Scattering plot of parameter pairs from MCMC runs using 10 (blue), 20 (red), 50 (black) and 500 (green) day simulations. Clear tightening of posterior can be observed as the simulation length increases. The third parameter is related to the model error covariance.

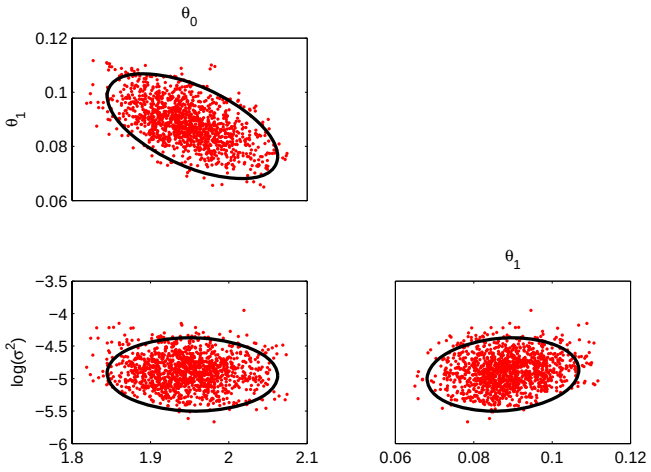


Fig. 5. Posterior distribution using 20 day simulations with MCMC (red dots) and Gaussian approximation (black ellipse).

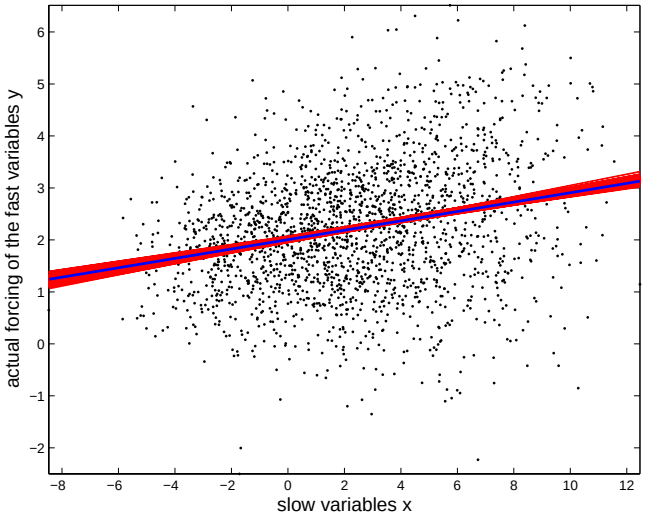


Fig. 6. Actual forcing of fast variables (black cloud). Red lines indicate forcing from MCMC runs. Red lines in the figure represent results from the 50 day MCMC simulations. Blue line is gotten by formally fitting the parameter values in the cloud.

In Fig. 4, the pairwise marginal distributions for the parameters are illustrated using 10, 20, 50 and 500 day simulations. As expected, the distribution becomes tighter as the simulation length increases. Note that the posterior distribution is much tighter compared to the summary statistics approach (see, e.g. Fig. 3) even though almost half of the states are not observed and the filtering procedure is applied to a relatively short observation sequence. The parameter estimates are closer to the reference values obtained using the knowledge of the simulated fast variables, compared to the estimates obtained via the summary statistics approach. We also observe that the parameter distribution is approximately Gaussian when sufficiently long simulations are used, as shown in Fig. 5.

In Fig. 6, we plot the true forcing in the simulated full system (35)–(36) against the slow variables. The red lines in the figure represent the parameter values from the 50 day MCMC simulation. The blue line represent the parameter values obtained by fitting a line to the true forcing in the least squares sense. We observe good agreement with our results and the fitted line.

The estimates obtained by the likelihood approach are close to the reference values obtained using the knowledge of the fast variables. However, there is no reason to think that the reference values are optimal, for instance in the sense of forecast accuracy. Therefore, to further demonstrate that likelihood approach produces good parameter estimates, we study the effect of the parameters on the forecast skill of the model. We grid the 2-dimensional parameter space and calculate the average 6 day forecast skill for different parameter values. The average forecast skill is computed by making a 6 day forecast starting every 24 h for 100 days. The averaged forecast skill can be written as

$$f(\boldsymbol{\theta}) = \frac{1}{NK\sigma_{\text{clim}}^2} \sum_{i=1}^{N} \|\mathcal{M}_6(\boldsymbol{x}_i^{\text{true}}, \boldsymbol{\theta}) - \boldsymbol{x}_{i+6}^{\text{true}}\|_2^2,$$

where $N = 100$, $K = 40$ and $\sigma_{\text{clim}} = 3.5$. The notation $\mathcal{M}_6(\boldsymbol{x}_i^{\text{true}}, \boldsymbol{\theta})$ means a 6 day prediction launched from the true state $\boldsymbol{x}_i^{\text{true}}$ with parameter values $\boldsymbol{\theta}$. The contour lines of the average forecast skill and the parameter values obtained by 10, 50 and 500 day MCMC runs are shown in Fig. 7. Again, we observe a good agreement: parameters tuned with the likelihood approach yield a good average forecast skill provided that the simulation length is sufficient.

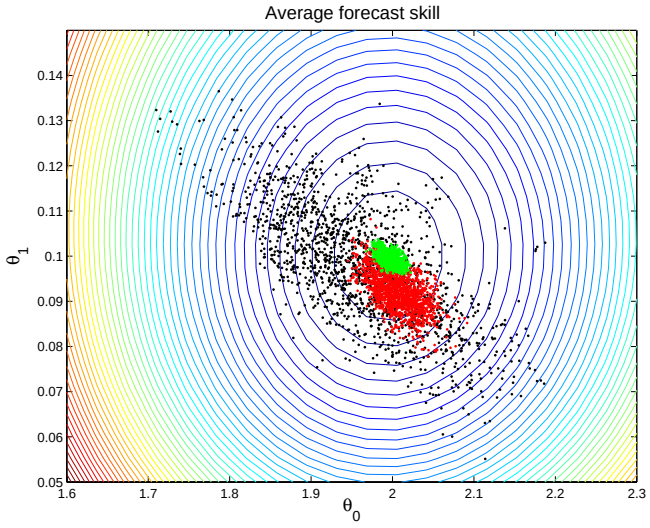


Fig. 7. An illustration of the average forecast skill. Black, red and green dots indicate the results from 10, 50 and 500 day MCMC simulations, respectively. Blue contour colors indicate high forecast skill.

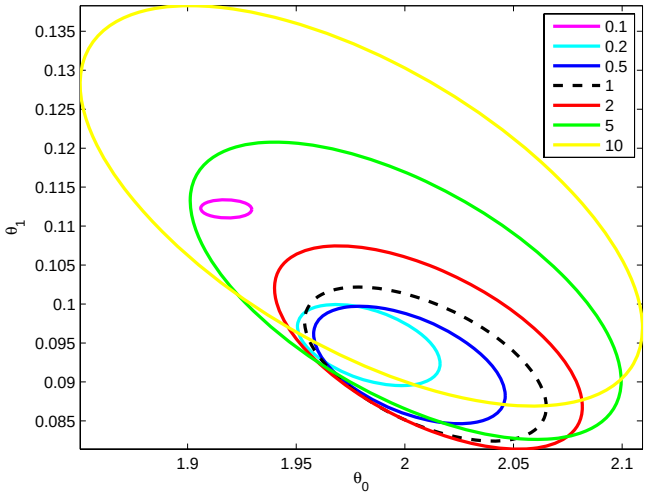


Fig. 8. Gaussian posterior approximations with 10, 5 and 2 times too small and too large σ^2.

Sensitivity to model error covariance

The possibility to estimate the model error covariance $\mathbf{C}_E$ from data is an advantage of the parameter estimation based on filtering. In data assimilation, $\mathbf{C}_E$ is often considered as a tuning parameter which has to be selected somehow. In large scale models like NWP, the model error covariance is usually estimated in a separate procedure, and finding an effective parametrization of $\mathbf{C}_E$ is an open question (see, e.g. Bonavita et al., 2008).

Therefore, in the following experiment, we test how specifying non-optimal values for the model error covariance affects the quality of parameter estimation. We use the same parameterization (40) and vary the variance parameter σ^2 so that it is two, five and 10 times smaller or greater than the optimal value obtained in the EKF exercise with the 500 day-long simulations (see Fig. 4). We run the likelihood approach with only 50 days of data. Since the posterior distribution is approximately Gaussian, we do not perform the computationally heavy MCMC runs, but compute the MAP estimate using an optimizer and calculate the Gaussian approximation of the posterior at the MAP estimate (see Appendix A).

The results are shown in Fig. 8. We observe that specifying the model error covariance wrongly can lead to biased parameter estimates: too small values of σ^2 lead to an underestimated posterior covariance of the parameters and vice versa. In this specific example, we change the correct $\mathbf{C}_E$ only by one order of magnitude and still obtain reasonable parameter estimates. For larger errors in $\mathbf{C}_E$, parameter estimates can be severely biased.

Filtering with EnKF

As discussed in Sect. 3.2, replacing the deterministic EKF with a stochastic filter, such as the EnKF, leads to a noisy likelihood function. We performed importance sampling for the two parameters similarly as in the summary statistics experiment in Section 5.2.1. The EnKF likelihood was evaluated by fixing the model error variance σ^2 to the optimum found in the EKF exercise, and setting the number of ensemble members to 100.

In our experiment, the noise in the likelihood function dominated the statistical analysis: most of the importance weights were assigned to only a few candidate values. That is, statistical inference could not be performed properly, and the noise in the likelihood seems to be a real issue in the likelihood approach computed with EnKF. Here, we settle for plotting the negative log-likelihood surface, as was done with EnKF parameter estimation by Dowd (2011). From Fig. 9 we observe that the general pattern is good: low values are found from the correct region, and a reasonable MAP estimate might be found using stochastic optimization techniques. However, statistical inference is complicated by the noisy likelihood. Smoothing methods could be used to alleviate the noise in the likelihood, but this question is not pursued further here.

In addition, we also tested the likelihood set-up with LETKF (Hunt et al., 2007) which falls into the category of ensemble square-root filters with no random perturbations. In this method, the model error term is neglected, but a covariance inflation term is used to inflate the posterior covariance and to account for the missing model error term. The covariance inflation parameter (see Hunt et al., 2007, for details) can be estimated together with the parameters like the model error term in EKF. In Fig. 10 we show MCMC runs

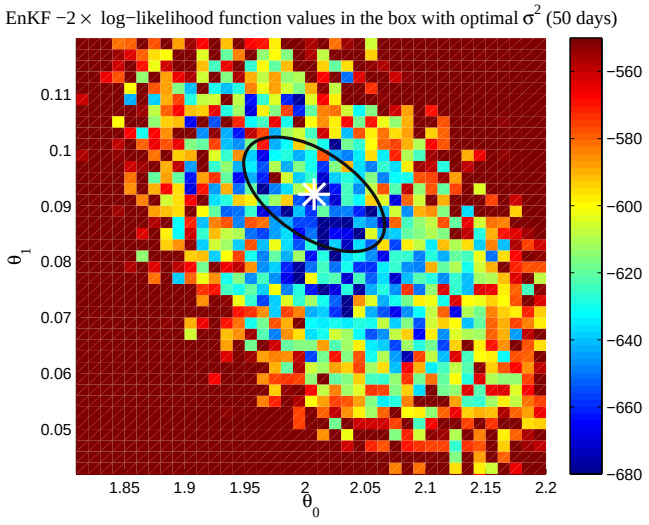


Fig. 9. The use of stochastic filtering method yields a noisy likelihood function. The results of 50 day run using EnKF based likelihood function is illustrated in the Figure. The values indicate negative log-likelihood times two values. White star and black ellipse is acquired from correspondent EKF likelihood calculations.

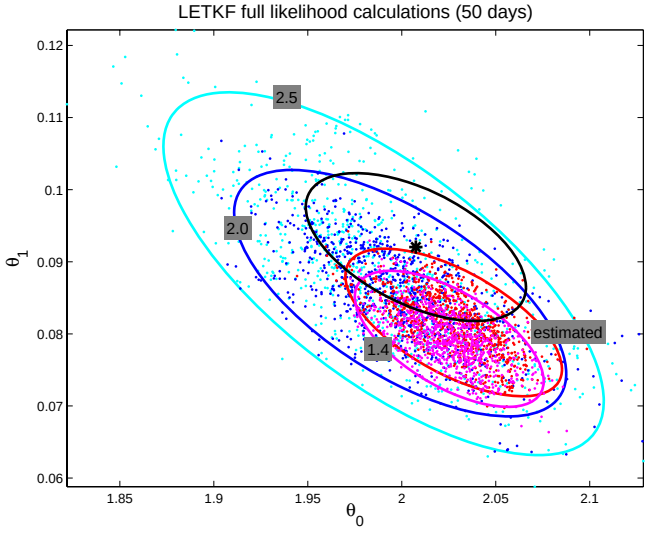


Fig. 10. Full likelihood calculations using LETKF data assimilation method. Different MCMC runs with different covariance inflation factor together with 95 % trust ellipses. Red color indicates a run, where the factor is also estimated (mean value is 1.52). Magenta, blue and cyan colors indicate a run where covariance inflation factor is fixed to 1.4, 2.0 and 2.5, respectively. Black star and ellipse is acquired from correspondent EKF likelihood calculations.

using different values for the covariance inflation parameter and a run where the inflation parameter is estimated together with the model parameters. Although there is a small bias, depending on the value of the covariance inflation parameter, the agreement with EKF calculations is rather good. Thus, deterministic ensemble filters seem to be more suitable for parameter estimation purposes.

5.2.3 State augmentation

As discussed in Sect. 4, in SA the model parameters are modeled as dynamical quantities, and the estimates do not converge to any fixed value as more observations are added. The rate at which parameter values can change from one filter time step to another is controlled by the extra tuning parameter, the model error covariance for the parameters, $\mathbf{C}_{\boldsymbol{\theta}}$. Here, we study how the SA method performs in parameter estimation and specifically how the tuning of $\mathbf{C}_{\boldsymbol{\theta}}$ affects the results.

Tuning of the parameter error covariance

In our experiments, we use a diagonal matrix as $\mathbf{C}_{\boldsymbol{\theta}}$ and keep it fixed during the runs. The model error for the state vector was fixed to its "optimal value", obtained from the likelihood experiments. In Fig. 11 we show four different runs using the EKF version of the SA method. The runs are made so that the diagonal elements of $\mathbf{C}_{\boldsymbol{\theta}}$ are taken to be 0.1 %, 1 %,10 % and 100 % of the optimal parameter values acquired from the likelihood experiment. In all cases, the SA method converged quickly to values near the optimum. The effect of the size of $\mathbf{C}_{\boldsymbol{\theta}}$ was as expected. When $\mathbf{C}_{\boldsymbol{\theta}}$ is set to be small, the method reacts slowly on new data and the parameter values take small steps. On the other hand, if $\mathbf{C}_{\boldsymbol{\theta}}$ is set large, the method allows larger deviations, but can yield unrealistic values for the parameters. Some theoretical discussion about the effect of $\mathbf{C}_{\boldsymbol{\theta}}$ is given in Appendix B.

In this example, we do not observe any systematic temporal variations in the parameter value. However, it is worth pointing out that the SA method could be useful in checking if such variations exist. Since the parameter trajectories are stationary, one could use the mean value of the trajectories as the final parameter estimate. In the current example, the mean is a good estimate, and it is also rather insensitive to the tuning of $\mathbf{C}_{\boldsymbol{\theta}}$. In general, however, the parameter trajectories cannot be interpreted in the statistical sense, since the parameter values and their variation depend entirely on the tuning of $\mathbf{C}_{\boldsymbol{\theta}}$. Thus, the SA method cannot be used for statistical inference of the parameters in the same sense as the likelihood approach.

Sensitivity to model error covariance

If, on the other hand, we keep the "parameter model error covariance" $\mathbf{C}_{\boldsymbol{\theta}}$ fixed and vary the model error covariance $\mathbf{C}_{\boldsymbol{x}}$, the effects are somewhat different than in the likelihood approach. Too small σ^2 values can cause the filter to diverge leading to unrealistic parameter estimates. Examples of runs with too large model error covariance are illustrated in Fig. 12. We observe that too large σ^2 values can cause bias to the estimates. In addition, $\mathbf{C}_{\boldsymbol{x}}$ affects the rate of change of

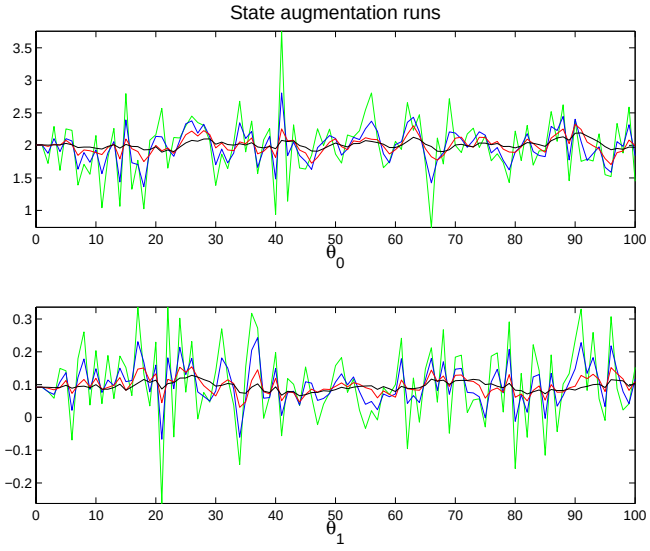


Fig. 11. Runs using the EKF version of the SA method, when the diagonal elements of $\mathbf{C}_{\boldsymbol{\theta}}$ are taken to be 0.1 % (black), 1 % (red),10 % (blue) and 100 % (green) of the optimal initial values. The effect of the size of $\mathbf{C}_{\boldsymbol{\theta}}$ was as expected. When $\mathbf{C}_{\boldsymbol{\theta}}$ is set to be small, the method reacts slowly on new data and the parameter values take small steps.

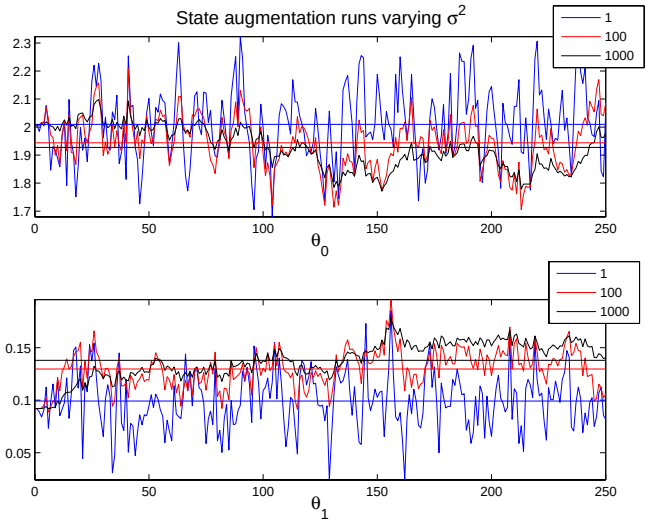


Fig. 12. Examples of too large model error: the model error variance is multiplied with 100 and 1000. Too large model error in state augmentation will cause bias in the parameter values. Note that the rate of change of the parameters becomes smaller as the variance grows, although $\mathbf{C}_{\boldsymbol{\theta}}$ is kept fixed in all runs. The straight lines represent the means of the parameter trajectories.

the parameters: the higher σ^2 is, the smaller are the parameter changes. The latter effect can be theoretically justified (see Appendix B for details).

State augmentation with EnKF

The use of ensemble based filtering methods is possible in state augmentation system. In our tests, with a large enough ensemble size, the results were similar to the EKF results. In the Lorenz system the minimum required ensemble size is roughly 50. Smaller ensemble size leads to underestimation of the covariances and can cause the filter to diverge. We note that state augmentation does not have the problem with the stochasticity of EnKF, which was encountered in the likelihood approach.

6 Remarks and discussion

6.1 Applicability to large scale systems

The discussed parameter estimation methods differ in their applicability to large scale systems like NWP and climate models. The summary statistics based approaches are computationally expensive although straightforward to implement: one only needs to simulate the model, and compare the selected summary statistics of the simulations and the observations. The difficulty lies in selecting the appropriate summary statistics that enable the identification of the parameters.

The state augmentation and the likelihood approaches depend on a data assimilation system, which is often available for NWP systems, but not commonly for climate models. The state augmentation method requires modifications to the assimilation method. In deterministic assimilation systems, such as the variational approximations to EKF that are often used in operational NWP systems (Rabier et al., 2000; Gauthier et al., 2007), one needs to add derivative computations with respect to the parameters. If an ensemble data assimilation method is used, such as a variant of the EnKF (see Houtekamer et al., 2005) parameter perturbations need to be added. Computationally, state augmentation is economical, since it requires only one assimilation sweep over the selected data.

The filter likelihood approach is computationally much more challenging than the state augmentation approach, since it involves many repeated filter sweeps, the number of which depends on the parameter estimation technique used. The likelihood approach requires, in addition to a data assimilation system, a method to estimate the forecast error covariance. In ensemble data assimilation systems, the covariance can be estimated from the ensemble. Variational data assimilation methods do not contain error propagation mechanisms, and it is not immediately clear how the likelihood approach can be implemented in these systems. A potential way is to obtain the covariance from ensemble prediction systems (EPS), that are built for approximating the prediction errors (Molteni et al., 1996; Palmer et al., 2005). Our preliminary tests with the low order system suggest that such EPS information could be used to approximate the likelihood approach, but verifying this is a topic of on-going research.

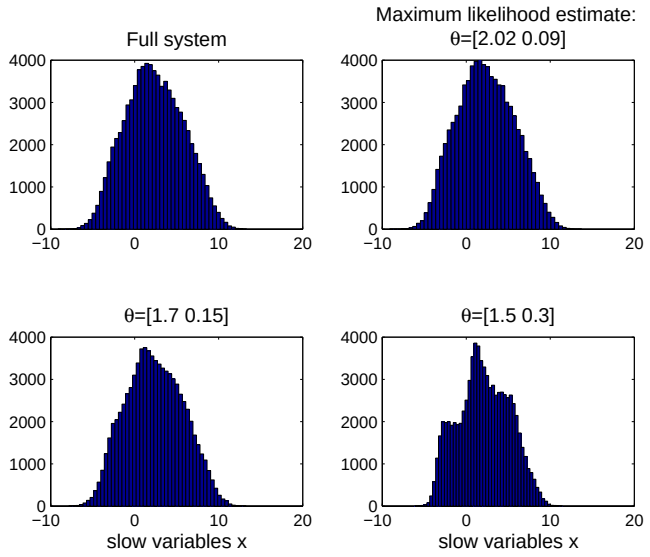


Fig. 13. Distribution of the state variables in the full Lorenz system (top left) and in the parameterized system with maximum likelihood estimate (top left) and two arbitrarily chosen "poor" parameter values (bottom row).

6.2 Climatologies with tuned parameters

In our likelihood experiment, the optimal parameter values led to improved short-range average forecast skills, as expected. Another question is related to the effect of parameter tuning on the quality of long model simulations (or "climatologies"): the tuned parameters should, in addition to improving short-range forecasts, improve climatologies, too. In Fig. 13, we compare the histograms of the state variables in the full Lorenz system and in the forecast model with different parameter values. We compare the statistics of the full system to the statistics produced by the forecast model with maximum likelihood parameter estimate and two arbitrarily chosen "poor" parameter values. We observe that, in this case, the parameters tuned with the likelihood approach produce also the correct climatologies. We also note that the overall statistics of the system can be quite good even with rather poor parameter values. This highlights the difficulties in choosing the correct likelihood for the summary statistics approach.

7 Conclusions

In this paper, we review three methods for parameter estimation in chaotic systems. In the summary statistics approach, the selected statistics computed from model simulations are compared to the same statistics calculated from observations. In the state augmentation method, unknown parameters are added to the state vector and estimated "on-line" together with the model state in a data assimilation system. In the likelihood approach, the likelihood for a parameter value is computed by running a data assimilation method "off-line" over a selected data set. All methods were studied using a modified version of the Lorenz 95 model.

Our results indicate that the summary statistics approach, albeit relatively easy to implement and compute, can have problems in properly identifying the parameters, and may lead to biased estimates. This result is supported by the previous climate model parameter estimation experiments (Järvinen et al., 2010) where simple summary statistics were not enough to uniquely identify all selected model parameters.

The state augmentation approach can work well and converge fast, if properly tuned. State augmentation contains additional tuning parameters, to which the performance of the method is somewhat sensitive: one must correctly specify the model error covariance both for the actual model states and for the parameters. The state augmentation approach is computationally feasible, since parameters are estimated "on-line" instead of repeatedly comparing model simulations to observations. The implementation of the method requires a modification to the data assimilation system. A downside of the approach is that the "static" model parameters are modeled as dynamical quantities, and one needs to accept the fact that the parameter estimates change at every time step and do not converge to a fixed value. Moreover, the method does not support statistical inference of the model parameters, since the obtained parameter values depend directly on the tuning of the model error covariance.

The likelihood approach performed well in our tests. The performance of the method was somewhat sensitive to the tuning of the model error covariance, like in the state augmentation approach. The likelihood approach assumes that the parameter values are static, and allows for statistical inference of the model parameters. The method requires a data assimilation system, and a method to propagate model error statistics. This may be restrictive in large-scale systems. The computational burden is much higher than in the state augmentation approach, and may be a bottleneck when scaling up to large scale NWP and climate models. The likelihood can be implemented with ensemble data assimilation methods, but the statistical analysis may be complicated by the stochasticity introduced into the likelihood function, if random perturbations are used in the ensemble method.

Appendix A

Parameter estimation algorithms

A1 MAP estimation and Gaussian approximation

The Maximum a Posteriori (MAP) estimate can be found by maximizing the posterior density $p(\boldsymbol{\theta}|\boldsymbol{y})$ with respect to $\boldsymbol{\theta}$, or, equivalently, minimizing the negative logarithm of the posterior

$$L(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta}|\boldsymbol{y}) = -\log p(\boldsymbol{y}|\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}).$$

The maximization can be done by different numerical methods (see, e.g. Nocedal and Wright, 1999). Once the estimate $\hat{\boldsymbol{\theta}} = \operatorname{argmin} L(\boldsymbol{\theta})$ has been obtained, one can construct a multivariate Gaussian approximation of $p(\boldsymbol{\theta}|\boldsymbol{y})$ around the point $\hat{\boldsymbol{\theta}}$. It is well known (see, e.g. Gelman et al., 2003) that the covariance matrix at $\hat{\boldsymbol{\theta}}$ can be approximated by the inverse Hessian of the negative logarithm of the posterior:

$$\mathrm{Cov}(\boldsymbol{\theta}) \approx \mathbf{H}(\hat{\boldsymbol{\theta}})^{-1},$$

where the Hessian matrix $\mathbf{H}(\hat{\boldsymbol{\theta}})$ contains the second derivatives of the negative logarithm of the likelihood, evaluated at $\hat{\boldsymbol{\theta}}$:

$$\mathbf{H}_{ij}(\hat{\boldsymbol{\theta}}) = \left.\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

The Hessian can be calculated analytically or numerically. In our examples, we have used a standard finite difference approximation applying the central difference formula (Nocedal and Wright, 1999).

A2 MCMC sampling

In principle, the Bayes formula, see Eq. (3), solves the parameter estimation problem in a fully probabilistic sense. However, the problem of calculating the integral of the normalizing constant is faced. This integration is often a formidable task, even for only moderately high number of parameters in a nonlinear model, and direct application of the Bayes formula is intractable for all but trivial nonlinear cases. The MCMC methods provide a tool to handle this problem. They generate a sequence of parameter values $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots \boldsymbol{\theta}_N$, whose empirical distribution approximates the true posterior distribution for large enough sample size N.

In many MCMC methods, instead of sampling directly from the true distribution, one samples from an artificial *proposal* distribution. Combining the sampling with a simple accept/reject procedure, the posterior can be correctly approximated. The simplest MCMC method is the *Metropolis algorithm* (Metropolis et al., 1953):

1. Initialize by choosing a starting point $\boldsymbol{\theta}_1$.
2. Choose a new candidate $\hat{\boldsymbol{\theta}}$ from a suitable proposal distribution $q(.|\boldsymbol{\theta}_n)$ that may depend on the previous point of the chain.
3. *Accept* the candidate with probability

$$\alpha(\boldsymbol{\theta}_n, \hat{\boldsymbol{\theta}}) = \min\left(1, \frac{\pi(\hat{\boldsymbol{\theta}})}{\pi(\boldsymbol{\theta}_n)}\right).$$

   If rejected, repeat the previous point in the chain. Go back to step 2.

So, points with $\pi(\hat{\boldsymbol{\theta}}) > \pi(\boldsymbol{\theta}_n)$, i.e., steps "uphill", are always accepted. But also points with $\pi(\hat{\boldsymbol{\theta}}) < \pi(\boldsymbol{\theta}_n)$, i.e. steps "downhill", may be accepted, with probability that is given by the *ratio* of the π values. In practice, this is done by generating a uniformly distributed random number $u \in [0,1]$ and accepting $\hat{\boldsymbol{\theta}}$ if $u \leq \pi(\hat{\boldsymbol{\theta}})/\pi(\boldsymbol{\theta}_i)$. Note that only the ratios of π at consecutive points are needed, so the main problem of calculation the normalizing constant is circumvented, since the constant cancels out.

However, the choice of the proposal distribution may still pose a problem. It should be chosen so that the "sizes" of the proposal q and target distributions suitably match. This may be difficult to achieve, and an unsuitable proposal can leadt to inefficient sampling. For simple cases, the proposal might be relatively easy to find by some hand-tuning. However, the "size" of the proposal distribution is not a sufficient specification. Especially In higher dimensions, the shape and orientation of the proposal are crucial. The most typical proposal is a multi–dimensional Gaussian (Normal) distribution. In the *random walk* version, the center point of the Gaussian proposal is chosen to be the current point of the chain. The task then is to find a covariance matrix that produces efficient sampling.

Several efficient adaptive methods have been recently proposed, for example, the adaptive Metropolis (AM) algorithm (Haario et al., 2001). In adaptive MCMC, one uses the sample history to automatically tune the proposal distribution "on-line" as the sampling proceeds. In AM, the empirical covariance from the samples obtained so far is used as the covariance of a Gaussian proposal. In this paper, a variant of AM called the delayed rejection adaptive Metropolis (DRAM, Haario et al., 2006) is used for all sampling tasks.

A3 Importance sampling

Some methods considered here use a likelihood which depends on initial conditions, random seeds and other settings, and the estimated likelihood is therefore random. For such methods, we used the following importance sampling procedure for estimating the parameters. The likelihood was computed for a set of candidate parameter values $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ which were drawn from an *importance function* $g(\boldsymbol{\theta})$. The posterior distribution of the parameters was evaluated by weighting each sample according to their likelihood values with respect to the importance function:

$$w_i = p(z \mid \boldsymbol{\theta}_i)/g(\boldsymbol{\theta}_i).$$

One can now compute the required statistics using samples $\boldsymbol{\theta}_i$ with weights w_i. Here, we evaluated the weighted posterior mean

$$\bar{\boldsymbol{\theta}} = \frac{1}{W}\sum_{i=1}^{N} w_i \boldsymbol{\theta}_i$$

with $W = \sum_{i=1}^{N} w_i$ and the weighted covariance matrix

$$\mathbf{C}_{\boldsymbol{\theta}} = \frac{1}{W}\sum_{i=1}^{N} w_i (\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})^T (\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})$$

to approximate the confidence intervals for the closure parameters.

Appendix B

The effect of the model error covariance matrix in the state augmentation method

Here we study the effect of the block diagonal error covariance matrix

$$\mathbf{C}_{\boldsymbol{x},\boldsymbol{\theta}} = \begin{bmatrix} \mathbf{C}_{\boldsymbol{x}} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{\boldsymbol{\theta}} \end{bmatrix}$$

in the state augmentation set up, that is when we assume that the errors in the state and the parameters are uncorrelated. For notational convenience, we do not use the time index k. Naturally, we do not observe the model parameters and the observation operator is

$$\tilde{\mathbf{K}} = \begin{bmatrix} \mathbf{K} & \mathbf{0} \end{bmatrix},$$

where $\mathbf{K}$ is the original observation operator.

The Kalman gain matrix $\mathbf{G}$, that defines how much the prior state and covariance are changed by an observation, can be written as

$$\begin{aligned} \mathbf{G} &= \mathbf{C}^p \tilde{\mathbf{K}}^T (\tilde{\mathbf{K}} \mathbf{C}^p \tilde{\mathbf{K}}^T + \mathbf{C}^{\mathrm{e}})^{-1} \\ &= \mathbf{M}^{\boldsymbol{\theta}} \mathbf{C}^{\mathrm{est}}_{k-1} \mathbf{M}^{\boldsymbol{\theta}^T} \tilde{\mathbf{K}}^T (\tilde{\mathbf{K}} \mathbf{M}^{\boldsymbol{\theta}} \mathbf{C}^{\mathrm{est}}_{k-1} \mathbf{M}^{\boldsymbol{\theta}^T} \tilde{\mathbf{K}}^T \\ &\quad + \tilde{\mathbf{K}} \mathbf{C}_{\boldsymbol{x},\boldsymbol{\theta}} \tilde{\mathbf{K}}^T + \mathbf{C}^{\mathrm{e}})^{-1} \\ &\quad + \mathbf{C}_{\boldsymbol{x},\theta} \tilde{\mathbf{K}}^T (\tilde{\mathbf{K}} \mathbf{M}^{\theta} \mathbf{C}^{\mathrm{est}}_{k-1} \mathbf{M}^{\boldsymbol{\theta}^T} \tilde{\mathbf{K}}^T \\ &\quad + \tilde{\mathbf{K}} \mathbf{C}_{\boldsymbol{x},\boldsymbol{\theta}} \tilde{\mathbf{K}}^T + \mathbf{C}^{\mathrm{e}})^{-1}. \end{aligned} \tag{B1}$$

Our augmented model error covariance matrix $\mathbf{C}_{\boldsymbol{x},\boldsymbol{\theta}}$ appears in the gain matrix only as $\mathbf{C}_{\boldsymbol{x},\boldsymbol{\theta}} \tilde{\mathbf{K}}^T$. Now,

$$\mathbf{C}_{\boldsymbol{x},\boldsymbol{\theta}} \tilde{\mathbf{K}}^T = \begin{bmatrix} \mathbf{C}_{\boldsymbol{x}} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{\boldsymbol{\theta}} \end{bmatrix} \begin{bmatrix} \mathbf{K}^T \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{C}_{\boldsymbol{x}} \mathbf{K}^T \\ \mathbf{0} \end{bmatrix}.$$

This means that the parameter part $\mathbf{C}_{\boldsymbol{\theta}}$ of the model error covariance matrix has no (direct) effect on the gain matrix. Hence, it would be the same as it would be directly inserted to its place in the posterior error covariance matrix. Especially this can be noted in the first round of the state augmentation: the selected matrix $\mathbf{C}_{\boldsymbol{\theta}}$ has no effect on parameters or state.

From the expansion (B1) of $\mathbf{G}$ we can note that only the first term $\mathbf{M}^{\boldsymbol{\theta}} \mathbf{C}^{\mathrm{est}}_{k-1} \mathbf{M}^{\boldsymbol{\theta}^T} \tilde{\mathbf{K}}^T (\tilde{\mathbf{K}} \mathbf{M}^{\boldsymbol{\theta}} \mathbf{C}^{\mathrm{est}}_{k-1} \mathbf{M}^{\boldsymbol{\theta}^T} \tilde{\mathbf{K}}^T + \tilde{\mathbf{K}} \mathbf{C}_{\boldsymbol{x},\boldsymbol{\theta}} \tilde{\mathbf{K}}^T + \mathbf{C}^{\mathrm{e}})^{-1}$ affects the parameter part of the gain matrix, since the second term has a multiplier $\mathbf{C}_{\boldsymbol{x},\boldsymbol{\theta}} \tilde{\mathbf{K}}^T$. In that term the model error term $\mathbf{C}_{\boldsymbol{x},\boldsymbol{\theta}}$ appears only in the inverse, so if we will increase $\mathbf{C}_{\boldsymbol{x}}$ in the experiments, it will cause a smaller rate of change to the parameters. In addition to the inverse, the previous posterior covariance matrix $\mathbf{C}^{\mathrm{est}}_{k-1}$ appears also in the "numerator". Hence, the effect of increasing $\mathbf{C}_{\boldsymbol{\theta}}$ will saturate at some point.

Acknowledgements. Martin Leutbecher from ECMWF is gratefully acknowledged for his support and help in designing the numerical experiments carried out with the stochastic Lorenz-95 system. The research has been supported by the Academy of Finland (project numbers 127210, 132808, 133142 and 134999) and the Centre of Excellence in Inverse Problems Research.

Edited by: W. Hsieh
Reviewed by: two anonymous referees

References

Annan, J. and Hargreaves, J.: Efficient estimation and ensemble generation in climate modelling, Phil. Trans. R. Soc. A, 365, 2077–2088, doi:10.1098/rsta.2007.2067, 2007.

Annan, J. D., Lunt, D. J., Hargreaves, J. C., and Valdes, P. J.: Parameter estimation in an atmospheric GCM using the Ensemble Kalman Filter, Nonlin. Processes Geophys., 12, 363–371, doi:10.5194/npg-12-363-2005, 2005.

Auvinen, H., Bardsley, J. M., Haario, H., and Kauranne, T.: Large-Scale Kalman Filtering Using the Limited Memory BFGS Method, Electron. T. Numer. Ana., 35, 217–233, 2009.

Auvinen, H., Bardsley, J., Haario, H., and Kauranne, T.: The variational Kalman filter and an efficient implementation using limited memory BFGS, Int. J. Numer. Meth. Fl., 64, 314–335, 2010.

Bishop, C. M.: Pattern Recognition and Machine Learning, Information Science and Statistics, Springer, New York, 2nd Edn., 2006.

Bonavita, M., Torrisi, L., and Marcucci, F.: The ensemble Kalman filter in an operational regional NWP system: preliminary results with real observations, Q. J. Roy. Meteor. Soc., 134, 1733–1744, doi:10.1002/qj.313, 2008.

Cappe, O., Godsill, S., and Moulines, E.: An overview of existing methods and recent advances in sequential Monte Carlo, Proceedings of IEEE, 95, 899–924, doi:10.1109/JPROC.2007.893250, 2007.

Cornuet, J.-M., Santos, F., Beaumont, M. A., Robert, C. P., Marin, J.-M., Balding, D. J., Guillemaud, T., and Estoup, A.: Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation, Bioinformatics, 24, 2713–2719, 2008.

DelSole, T. and Yang, X.: State and Parameter Estimation in Stochastic Dynamical Models, Physica D, 239, 1781–1788, 2010.

Dowd, M.: Estimating parameters for a stochastic dynamic marine ecological system, Environmetrics, 22, 501–515, doi:10.1002/env.1083, 2011.

Evensen, G.: Data assimilation: The ensemble Kalman filter, Springer, 2007.

Gauthier, P., Tanguay, M., Laroche, S., Pellerin, S., and Morneau, J.: Extension of 3DVAR to 4DVAR: Implementation of 4DVAR at the Meteorological Service of Canada, Mon. Weather Rev., 135, 2339–2354, 2007.

Gelman, A., Carlin, J., Stern, H., and Rubin, D.: Bayesian Data Analysis, Chapman & Hall, 2nd Edn., 2003.

Haario, H., Saksman, E., and Tamminen, J.: An adaptive Metropolis algorithm, Bernoulli, 7, 223–242, 2001.

Haario, H., Laine, M., Mira, A., and Saksman, E.: DRAM: Efficient adaptive MCMC, Stat. Comput., 16, 339–354, doi:10.1007/s11222-006-9438-0, 2006.

Houtekamer, P., Herschel, L., Mitchell, G., Buehner, M., Charron, M., Spacek, L., and Hansen, B.: Atmospheric Data Assimilation with an Ensemble Kalman Filter: Results with Real Observations, Mon. Weather Rev., 133, 604–620, 2005.

Hunt, B. R., Kostelich, E. J., and Szunyogh, I.: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter, Physica D, 230, 112–126, 2007.

Ionides, E., Breto, C., and King, A.: Inference for nonlinear dynamical systems, Proc. Nat. Aca. Sci., 103, 18438–18443, 2006.

Jackson, C. S., Sen, M. K., Huerta, G., Deng, Y., and Bowman, K. P.: Error Reduction and Convergence in Climate Prediction, J. Climate, 21, 6698–6709, doi:10.1175/2008JCLI2112.1, 2008.

Järvinen, H., Räisänen, P., Laine, M., Tamminen, J., Ilin, A., Oja, E., Solonen, A., and Haario, H.: Estimation of ECHAM5 climate model closure parameters with adaptive MCMC, Atmos. Chem. Phys., 10, 9993–10002, doi:10.5194/acp-10-9993-2010, 2010.

Järvinen, H., Laine, M., Solonen, A., and Haario, H.: Ensemble prediction and parameter estimation system: the concept, Q. J. Roy. Meteor. Soc., doi:10.1002/qj.923, published online, 2011.

Kalman, R.: A New Approach to Linear Filtering and Prediction Problems, Transactions of the ASME – Journal of Basic Engineering, Series D, 82, 35–42, 1960.

Kitagawa, G.: A self-organizing state space model, J. Am. Stat. Assoc., 93, 1203–1215, 1998.

Kivman, G. A.: Sequential parameter estimation for stochastic systems, Nonlin. Processes Geophys., 10, 253–259, doi:10.5194/npg-10-253-2003, 2003.

Laine, M., Solonen, A., Haario, H., and Järvinen, H.: Ensemble prediction and parameter estimation system: the method, Q. J. Roy. Meteor. Soc., published online, doi:10.1002/qj.922, 2011.

Leutbecher, M.: Predictability and Ensemble Forecasting with Lorenz-95 systems, Lecture notes, ECMWF meteorological training course on Predictability, diagnostics and seasonal forecasting, available at: http://www.ecmwf.int/newsevents/training/meteorological_presentations/pdf/PR/Practice_L95.pdf, 2010.

Lorenz, E.: Predictability: A problem partly solved, Proceedings of the Seminar on Predicability, European Center on Medium Range Weather Forecasting, 1, 1–18, 1995.

Marjoram, P. and Tavaré, S.: Modern computational approaches for analysing molecular genetic variation data, Nat. Rev. Genet., 7, 759–770, 2006.

Martin, G. M., Johnson, D. W., and Spice, A.: The measurement and parameterization of effective radius of droplets in warm stratocumulus, J. Atmos. Sci., 51, 1823–1842, 1994.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E.: Equations of State Calculations by Fast Computing Machines, J. Chem. Phys., 21, 1087–1092, 1953.

Molteni, F., Buizza, R., Palmer, T., and Petroliagis, T.: The ECMWF Ensemble Prediction System: Methodology and validation, Q. J. Roy. Meteor. Soc., 122, 73–119, 1996.

Nocedal, G. and Wright, S.: Numerical Optimization, Springer, Berlin, 1999.

Ott, E., Hunt, B., Szunyogh, I., Zimin, A., Kostelich, E., Corazza, M., Kalnay, E., Pati, D., and Yorke, J.: A local ensemble Kalman filter for atmospheric data assimilation, Tellus A, 56, 415–428, 2004.

Palmer, T., Shutts, G., Hagedorn, R., Doblas-Reyes, F., Jung, T., and Leutbecher, M.: Representing Model Uncertainty in Weather and Climate Prediction, Ann. Rev. Earth Planet. Sci., 33, 163–193, 2005.

Rabier, F., Järvinen, H., Klinker, E., Mahfouf, J.-F., and Simmons, A.: The ECMWF implementation of four dimensional variational assimilation, Part I: Experimental results with simplified physics, Q. J. Roy. Meteor. Soc., 126, 1143–1170, 2000.

Rougier, J. C.: Efficient Emulators for Multivariate Deterministic Functions, J. Comput. Graphic. Stat., 17, 827–843, 2008.

Sexton, D., Murphy, J., Collins, M., and Webb, M.: Multivariate Prediction Using Imperfect Climate Models Part I: Outline of Methodology, Clim. Dynam., 1–30, doi:10.1007/s00382-011-1208-9, 2011.

Shapiro, A., Dentcheva, D., and Ruszczynski, A.: Lectures on Stochastic Programming: Modeling and Theory, MOS-SIAM Series on Optimization 9, Soc. Ind. Appl. Math., 2009.

Singer, H.: Parameter estimation of nonlinear stochastic differential equations: Simulated maximum likelihood versus extended Kalman filter and Ito-Taylor expansion, J. Comput. Graph. Stat., 11, 972–995, 2002.

Tippett, M., Anderson, J., Bishop, G., Hamill, T., and Whitaker, J.: Ensemble Square Root Filters*, Mon. Weather Rev., 131, 1485–1490, doi:10.1175/1520-0493(2003)131<1485:ESRF>2.0.CO;2, 2003.

Villagran, A., Huerta, G., Jackson, C. S., and Sen, M. K.: Computational Methods for Parameter Estimation in Climate Models, Bayesian Analysis, 3, 823–850, doi:10.1214/08-BA331, 2008.

Whitaker, J. and Hamill, T.: Ensemble data assimilation without perturbed observations, Mon. Weather Rev., 130, 1913–1924, 2002.

Wilks, D.: Effects of stochastic parametrizations in the Lorenz '96 system, Q. J. Roy. Meteor. Soc., 131, 389–407, doi:10.1256/qj.04.03, 2005.

Zupanski, M.: Maximum Likelihood Ensemble Filter: Theoretical Aspects, Mon. Weather Rev., 133, 1710–1726, 2005.