# A review of innovation-based methods to jointly estimate model and observation error covariance matrices in ensemble data assimilation

Article

Accepted Version

# www.reading.ac.uk/centaur

**CentAUR**

Central Archive at the University of Reading

Reading's research outputs online

# A Review of Innovation-Based Methods to Jointly Estimate Model and Observation Error Covariance Matrices in Ensemble Data Assimilation

Pierre Tandeo[*]

*IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238, France & RIKEN Center for Computational Science, Kobe, Japan*

Pierre Ailliot

*LMBA, UMR CNRS 6205, University of Brest, France*

Marc Bocquet

*CEREA Joint Laboratory École des Ponts ParisTech and EDF R&D, Université Paris-Est, Champs-sur-Marne, France*

Alberto Carrassi

*Dept. of Meteorology and National Centre for Earth Observation, University of Reading, UK & Mathematical Institute, University of Utrecht, Netherlands*

Takemasa Miyoshi

*RIKEN Center for Computational Science, Kobe, Japan*

Manuel Pulido

Generated using v4.3.2 of the AMS LATEX template

1

*Universidad Nacional del Nordeste and CONICET, Corrientes, Argentina & Department of*

*Meteorology, University of Reading, UK*

Yicun Zhen

*IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238, France*

\*_Corresponding author address:_ Dept.  Signal & Communications, IMT Atlantique, 655 Avenue

du Technopôle, 29200 Plouzané, France

E-mail: pierre.tandeo@imt-atlantique.fr

2

ABSTRACT

Data assimilation combines forecasts from a numerical model with observations. Most of the current data assimilation algorithms consider the model and observation error terms as additive Gaussian noise, specified by their covariance matrices Q and R, respectively. These error covariances, and specifically their respective amplitudes, determine the weights given to the background (i.e., the model forecasts) and to the observations in the solution of data assimilation algorithms (i.e., the analysis). Consequently, Q and R matrices significantly impact the accuracy of the analysis. This review aims to present and to discuss, with a unified framework, different methods to jointly estimate the Q and R matrices using ensemble-based data assimilation techniques. Most of the methodologies developed to date use the innovations, defined as differences between the observations and the projection of the forecasts onto the observation space. These methodologies are based on two main statistical criteria: (i) the method of moments, in which the theoretical and empirical moments of the innovations are assumed to be equal, and (ii) methods that use the likelihood of the observations, themselves contained in the innovations. The reviewed methods assume that innovations are Gaussian random variables, although extension to other distributions is possible for likelihood-based methods. The methods also show some differences in terms of levels of complexity and applicability to high-dimensional systems. The conclusion of the review discusses the key challenges to further develop estimation methods for Q and R. These challenges include taking into account time-varying error covariances, using limited observational coverage, estimating additional deterministic error terms, or accounting for correlated noise.

4

## 1. Introduction

In meteorology and other environmental sciences, an important challenge is to estimate the state of the system as accurately as possible. In meteorology, this state includes pressure, humidity, temperature and wind at different locations and elevations in the atmosphere. Data assimilation (hereinafter DA) refers to mathematical methods that use both model predictions (also called background information) and partial observations to retrieve the current state vector with its associated error. An accurate estimate of the current state is crucial to get good forecasts, and it is particularly so whenever the system dynamics is chaotic, such as it is the case for the atmosphere.

The performance of a DA system to estimate the state depends on the accuracy of the model predictions, the observations, and their associated error terms. A simple, popular and mathematically justifiable way of modeling these errors is to assume them to be independent and unbiased Gaussian white noise, with covariance matrices $\mathbf{Q}$ for the model and $\mathbf{R}$ for the observations. Given the aforementioned importance of $\mathbf{Q}$ and $\mathbf{R}$ in estimating the analysis state and error, a number of studies dealing with this problem has arisen in the last decades. This review work presents and summarizes the different techniques used to estimate simultaneously the $\mathbf{Q}$ and $\mathbf{R}$ covariances. Before discussing the methods to achieve this goal, the mathematical formulation of DA is briefly introduced.

### a. Problem statement

Hereinafter, the unified DA notation proposed in Ide et al. (1997) is used[1]. DA algorithms are used to estimate the state of a system, $\mathbf{x}$, conditionally on observations, $\mathbf{y}$. A classic strategy is to use sequential and ensemble DA frameworks, as illustrated in Fig. 1, and to combine two sources of information: model forecasts (in green) and observations (in blue). The ensemble framework

---

[1]Other notations are also used in practice

5

70 uses different realizations, also called members, to track the state of the system at each assimilation

71 time step.

72 The forecasts of the state are based on the usually incomplete and approximate knowledge of the

73 system dynamics. The evolution of the state from time $k-1$ to $k$ is given by the model equation:

$$\mathbf{x}(k) = \mathscr{M}_k\left(\mathbf{x}(k-1)\right) + \boldsymbol{\eta}(k), \tag{1}$$

74 where the model error $\boldsymbol{\eta}$ implies that the dynamic model operator $\mathscr{M}_k$ is not perfectly known.

75 Model error is usually assumed to follow a Gaussian distribution with zero mean (i.e., the model

76 is unbiased) and covariance $\mathbf{Q}$. The dynamic model operator $\mathscr{M}_k$ in Eq. (1) has also an explicit

77 dependence on $k$, because it may depend on time-dependent external forcing terms. At time $k$,

78 the forecasted state is characterized by the mean of the forecasted states, $\mathbf{x}^f$, and its uncertainty

79 matrix, namely $\mathbf{P}^f$, which is also called the background error covariance matrix, and noted $\mathbf{B}$ in

80 DA.

81 The forecast covariance $\mathbf{P}^f$ is determined by two processes. The first is the uncertainty propa-

82 gated from $k-1$ to $k$ by the model $\mathscr{M}_k$ (the green shade within the dashed ellipse in Fig. 1, and

83 denoted by $\mathbf{P}^m$). The second process is the model error covariance $\mathbf{Q}$ accounted by the noise term

84 at time $k$ in Eq. (1). Given that model error is largely unknown and originated by various and

85 diverse sources, the matrix $\mathbf{Q}$ is also poorly known. Model error sources encompass the model $\mathscr{M}$

86 deficiencies to represent the underlying physics, including deficiencies in the numerical schemes,

87 the cumulative effects of errors in the parameters, and the lack of knowledge of the unresolved

88 scales. Its estimation is a challenge in general, but it is particularly so in geosciences because we

89 usually have far fewer observations than those needed to estimate the entries of $\mathbf{Q}$ (Daley 1992;

90 Dee 1995). The sum of the two covariances $\mathbf{P}^m$ and $\mathbf{Q}$ gives the forecast covariance matrix, $\mathbf{P}^f$

91 (full green ellipse in Fig. 1). In the illustration given here, a large contribution of the forecast co-

6

92 variance $\mathbf{P}^f$ is due to $\mathbf{Q}$. This situation reflects what is common in ensemble DA, where $\mathbf{P}^m$ can be

93 too small, as a consequence of the ensemble undersampling of the initial condition error (i.e., the

94 covariance estimated at the previous analysis). In that case, inflating $\mathbf{Q}$ could partially compensate

95 for the bad specification of $\mathbf{P}^m$.

96 DA uses a second source of information, the observations $\mathbf{y}$, which are assumed to be linked to

97 the true state $\mathbf{x}$ through the time-dependent operator $\mathscr{H}_k$. This step in DA algorithms is formalized

98 by the observation equation:

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

99 where the observation error $\boldsymbol{\epsilon}$ describes the discrepancy between what is observed and the truth.

100 In practice, it is important to remove as much as possible the large-scale bias in the observation

101 before DA. Then, it is common to state that the remaining error $\boldsymbol{\epsilon}$ follows a Gaussian and unbiased

102 distribution with a covariance $\mathbf{R}$ (the blue ellipse in Fig. 1). This covariance takes into account er-

103 rors in the observation operator $\mathscr{H}$, the instrumental noise and the representation error associated

104 with the observation, typically measuring a higher resolution state than the model represents. Op-

105 erationally, a correct estimation of $\mathbf{R}$ that takes into account all these effects is often challenging

106 (Janjić et al. 2018).

107 DA algorithms combine forecasts with observations, based on the model and observation equa-

108 tions, respectively given in Eq. (1) and Eq. (2). The corresponding system of equations is a non-

109 linear state-space model. As illustrated in Fig. 1, this Gaussian DA process produces a posterior

110 Gaussian distribution with mean $\mathbf{x}^a$ and covariance $\mathbf{P}^a$ (red ellipse). The system given in Eqs. (1)

111 and (2) is representative of a broad range of DA problems, as described in seminal papers such

112 as Ghil and Malanotte-Rizzoli (1991), and still relevant today as referenced by Houtekamer and

113 Zhang (2016) and Carrassi et al. (2018). The assumptions made in Eqs. (1) and (2) about model

114 and observation errors (additive, Gaussian, unbiased, and mutually independent) are strong, yet

7

115   convenient from the mathematical and computational point of view. Nevertheless, these assump-

116   tions are not always realistic in real DA problems. For instance, in operational applications, sys-

117   tematic biases in the model and in the observations are recurring problems. Indeed, biases affect

118   significantly the DA estimations and a specific treatment is required; see Dee (2005) for more

119   details.

120      From Eqs. (1) and (2), noting that $\mathscr{M}$, $\mathscr{H}$ and $\mathbf{y}$ are given, the only parameters that influence the

121   estimation of $\mathbf{x}$ are the covariance matrices $\mathbf{Q}$ and $\mathbf{R}$. These covariances play an important role in

122   DA algorithms. Their importance was early put forward in Hollingsworth and Lönnberg (1986),

123   in section 4.1 of Ghil and Malanotte-Rizzoli (1991) and Daley (1991) in section 4.9. The results

124   of DA algorithms highly depend on the two error covariance matrices $\mathbf{Q}$ and $\mathbf{R}$, which have to be

125   specified by the users. But these covariances are not easy to tune. Indeed, their impact is hard to

126   grasp in real DA problems with high-dimensionality and nonlinear dynamics. We thus illustrate

127   the problem with a simple example first.

128   *b. Illustrative example*

129      In either variational or ensemble-based DA methods, the quality of the reconstructed state (or

130   hidden) vector $\mathbf{x}$ largely depends on the relative amplitudes between the assumed observation and

131   model errors (Desroziers and Ivanov 2001). In Kalman filter based methods, the signal-to-noise

132   ratio $\left\| \mathbf{P}^f \right\| / \left\| \mathbf{R} \right\|$, where $\mathbf{P}^f$ depends on $\mathbf{Q}$, impacts the Kalman gain, which gives the relative

133   weights of the observations against the model forecasts. Here, the $\|.\|$ operator represents a matrix

134   norm. For instance, Berry and Sauer (2013) used the Frobenius norm to study the effect of this

135   ratio in the reconstruction of the state in toy models.

136      The importance of $\mathbf{Q}$, $\mathbf{R}$ and $\left\| \mathbf{P}^f \right\| / \left\| \mathbf{R} \right\|$ is illustrated with the aid of a toy example, using

137   a scalar state $x$ and simple linear dynamics. This simplified setup avoids several issues typical

8

138 of realistic DA applications: the large dimension of the state, the strong nonlinearities and the

139 chaotic behavior. In this example, the dynamic model in Eq. (1) is a first-order autoregressive

140 model, denoted by AR(1) and defined by

$$x(k) = 0.95x(k-1) + \eta(k), \tag{3}$$

141 with $\eta \sim \mathcal{N}(0, Q^t)$ where the superscript $t$ means "true" and $Q^t = 1$. Furthermore, observations $y$

142 of the state are contaminated with an independent additive zero-mean and unit-variance Gaussian

143 noise, such that $R^t = 1$ in Eq. (2) with $\mathcal{H}(x) = x$. The goal is to reconstruct $x$ from the noisy ob-

144 servations $y$ at each time step. The AR(1) dynamic model defined by Eq. (3) has an autoregressive

145 coefficient close to one, representing a process which evolves slowly over time, and a stochastic

146 noise term $\eta$ with variance $Q^t$. Although the knowledge of these two sources of noise is crucial

147 for the estimation problem, identifying them is not an easy task. Given that the dynamic model is

148 linear and the error terms are additive and Gaussian in this simple example, the Kalman smoother

149 provides the best estimation of the state (see section 2 for more details). To evaluate the effect

150 of badly specified $Q$ and $R$ errors on the reconstructed state with the Kalman smoother, different

151 experiments were conducted with values of $\{0.1, 1, 10\}$ for the ratio $Q/R$ (in this toy example, we

152 use $Q/R$ instead of $\left\| \mathbf{P}^f \right\| / \left\| \mathbf{R} \right\|$ for simplicity).

153 Figure 2 shows, as a function of time, the true state (red line) and the smoothing Gaussian

154 distributions represented by the 95% confidence intervals (gray shaded) and their means (black

155 lines). We also report the Root Mean Squared Error (RMSE) of the reconstruction and the so-

156 called "coverage probability", or percentage of $x$ that falls in the 95% confidence intervals (defined

157 as the mean $\pm 1.96$ the standard deviation in the Gaussian case). In this synthetic experiment, the

158 best RMSE and coverage probability obtained, applying the Kalman smoother with true $Q^t =$

159 $R^t = 1$, are 0.71 and 95%, respectively. Using a small model error variance $Q = 0.1Q^t$ in Fig. 2(a),

9

the filter gives a large weight to the forecasts given by the quasi-persistent autoregressive dynamic

model. On the other hand, with a small observation error variance $R = 0.1R^t$ in Fig. 2(b), excessive

weight is given to the observation and the reconstructed state is close to the noisy measurements.

These results show the negative impact of independently badly scaled $Q$ and $R$ error variances. In

the case of overestimated model error variance as in Fig. 2(c), the mean reconstructed state vector

and thus its RMSE are identical to Fig. 2(b). In the same way, overestimated observation error

variance like in Fig. 2(d) gives similar mean reconstruction, as in Fig. 2(a). These last two results

are due to the fact that in both cases, the ratio $Q/R$ are equal, respectively, to 10 and 0.1. Now,

we consider in Fig. 2(e) and Fig. 2(f) the case where the $Q/R$ ratio is equal to 1, but, respectively,

using the simultaneous underestimation and overestimation of model and observation errors. In

both cases, the mean reconstructed state is equal to that obtained with the true error variances (i.e.,

RMSE=0.71). The main difference is the gray confidence interval, which is supposed to contain

95% of the true trajectory: the spread is clearly underestimated in Fig. 2(e) and overestimated in

Fig. 2(f), with respective coverage probability of 36% and 100%.

We used a simple synthetic example, but for large dimensional and highly nonlinear dynamics,

such an underestimation or overestimation of uncertainty may have a strong effect and may cause

filters to collapse. The main issue in ensemble-based DA is an underdispersive spread, as in

Fig. 2(e). In that case, the initial condition spread is too narrow, and model forecasts (starting

from these conditions) would be similar and potentially out of the range of the observations. In

the case of an overdispersive spread, as in Fig. 2(f), the risk is that only a small portion of model

forecasts would be accurate enough to produce useful information on the true state of the system.

This illustrative example shows how important is the joint tuning of model and observation errors

in DA. Since the 1990s, a substantial number of studies have dealt with this topic.

*c. Seminal work in the data assimilation community*

In a seminal paper, Dee (1995) proposed an estimation method for parametric versions of $\mathbf{Q}$ and $\mathbf{R}$ matrices. The method, based on maximizing the likelihood of the observations, yields an estimator which is a function of the innovation defined by $\mathbf{y} - \mathscr{H}(\mathbf{x}^f)$. Maximization is performed at each assimilation step, with the current innovation computed from the available observations. This technique was later extended to estimate the mean of the innovation, which depends on the biases in the forecast and in the observations (Dee et al. 1999a). The methodology was then applied to realistic cases in Dee et al. (1999b), making the maximization of innovation likelihood a promising technique for the estimation of errors in operational forecasts.

Following a distinct path, Desroziers and Ivanov (2001) proposed using the observation-minus-analysis diagnostic. It is defined by $\mathbf{y} - \mathscr{H}(\mathbf{x}^a)$ with $\mathbf{x}^a$ the analysis (i.e., the output of DA algorithms). The authors proposed an iterative optimization technique to estimate a scaling factor for the background $\mathbf{B} = \mathbf{P}^f$ and observation $\mathbf{R}$ matrices. The procedure was shown to converge to a proper fixed-point. As in Dee's work, the fixed-point method presented in Desroziers and Ivanov (2001) is applied at each assimilation step, with the available observations at the current step.

Later, Chapnik et al. (2004) showed that the maximization of the innovation likelihood proposed by Dee (1995) makes the observation-minus-analysis diagnostic of Desroziers and Ivanov (2001) optimal. Moreover, the techniques of Dee (1995) and Desroziers and Ivanov (2001) have been further connected to the generalized cross-validation method previously developed by statisticians (Wahba and Wendelberger 1980).

These initial studies clearly nurtured the discussion of the estimation of observation $\mathbf{R}$, model $\mathbf{Q}$, or background $\mathbf{B} = \mathbf{P}^f$ error covariance matrices in the modern DA literature. For demonstration purposes, the algorithms proposed in Dee (1995) and Desroziers and Ivanov (2001) were tested on

11

realistic DA problems, using a shallow-water model on a plane with a simplified Kalman filter, and using the French ARPEGE three-dimensional variational framework, respectively. In both cases, although good performances have been obtained with a small number of iterations, the proposed algorithms have shown some limits, in particular with regard to the simultaneous estimation of the two sources of errors: observation and model (or background). In this context, Todling (2015) pointed out that using only the current innovation is not enough to distinguish the impact of $\mathbf{Q}$ and $\mathbf{R}$, which still makes their simultaneous estimation challenging. Given that our preliminary focus here is to review methods for the joint estimate of $\mathbf{Q}$ and $\mathbf{R}$, the work Dee (1995) and Desroziers and Ivanov (2001) are not further detailed hereafter. After these two seminal studies, various alternatives were proposed. They are based on the use of several types of innovations and are discussed in this review.

### d. Methods presented in this review

The main topic of this review is the "joint estimation of $\mathbf{Q}$ and $\mathbf{R}$". Thus, only methods based on this specific goal are presented in detail. A history of what have been, in our opinion, the most relevant contributions and the key milestones for $\mathbf{Q}$ and $\mathbf{R}$ covariance estimation in DA is sketched in Fig. 3. The highlighted papers are discussed in this review, with a summary of the different methodologies, given in Table 1. We distinguish four methods and we can classify them into two categories: those which rely on moment-based methods, and those using likelihood-based methods. Both methods make use of the innovations. The main concepts of the techniques are briefly introduced below.

On the one hand, moment-based methods assume equality between theoretical and empirical statistical moments. A first approach is to study different type of innovations in the observation space (i.e., working in the space of the observations instead of the space of the state). It has

12

been initiated in DA by Rutherford (1972) and Hollingsworth and Lönnberg (1986). A second

approach extracts information from the correlation between lag innovations, namely innovations

between consecutive times. On the other hand, likelihood-based methods aim to maximize likeli-

hood functions with statistical algorithms. One option is to use a Bayesian framework, assuming

prior distributions for the parameters of $\mathbf{Q}$ and $\mathbf{R}$ covariance matrices. Another option is to use the

iterative expectation–maximization algorithm to maximize a likelihood function.

The four methodologies listed in Fig. 3 will be examined in this paper. Before doing that, it is

worth mentioning existing review work that have attempted to summarize the methodologies in

DA context and beyond.

*e. Other review papers*

Other review papers on parameter estimation (including $\mathbf{Q}$ and $\mathbf{R}$ matrices) in state-space models

have appeared in the statistical and signal processing communities. The first one (Mehra 1972)

introduces moment- and likelihood-based methods in the linear and Gaussian case (i.e., when $\boldsymbol{\eta}$

and $\epsilon$ are Gaussians and $\mathscr{M}$ is a linear operator in Eqs. (1) and (2)). Many extensions to nonlinear

state-space models have been proposed since the seminal work of Mehra, and these studies are

summarized in the recent review by Duník et al. (2017), with a focus on moment-based methods

and the extended Kalman filter (Jazwinski 1970). The book chapter by Buehner (2010) presents

another review of moment-based methods, with a focus on the modeling and estimation of spatial

covariance structures $\mathbf{Q}$ and $\mathbf{R}$ in DA with the ensemble Kalman filter algorithm (Evensen 2009).

In the statistical community, the recent development of powerful simulation techniques, known

as sequential Monte-Carlo algorithms or particle filters, has led to an extensive literature on the

statistical inference in nonlinear state-space models relying on likelihood-based approaches. A

recent and detailed presentation of this literature can be found in Kantas et al. (2015). However,

13

252 these methods typically require a large number of particles, which make them impractical for

253 geophysical DA applications.

254 The review presented here focuses on methods proposed in DA, especially the moment- and

255 likelihood-based techniques which are suitable for geophysical systems (i.e., with high dimen-

256 sionality and strong nonlinearities).

*f. Structure of this review*

258 The paper is organized as follows. Section 2 briefly presents the filtering and smoothing DA

259 algorithms used in this work. The main families of methods used in the literature to jointly

260 estimate error covariance matrices $\mathbf{Q}$ and $\mathbf{R}$ are then described. First, moment-based methods

261 are introduced in section 3. Then, we describe in section 4 the likelihood-based methods. We

262 also mention other alternatives in section 5, along with methods used in the past but not exactly

263 matching the scope of this review, and diagnostic tools to check the accuracy of $\mathbf{Q}$ and $\mathbf{R}$. Finally,

264 in section 6, we provide a summary and discussion on what we consider to be the forthcoming

265 challenges in this area.

266

## 2. Filtering and smoothing algorithms

268 This review paper focuses on the estimation of $\mathbf{Q}$ and $\mathbf{R}$ in the context of ensemble-based DA

269 methods. For the overall discussion of the methods and to set the notation, a short description of

270 the ensemble version of the Kalman recursions is presented in this section: the ensemble Kalman

271 filter (EnKF) and ensemble Kalman smoother (EnKS).

272 The EnKF and EnKS estimate various state vectors $\mathbf{x}^f(k)$, $\mathbf{x}^a(k)$, $\mathbf{x}^s(k)$ and covariance matrices

273 $\mathbf{P}^f(k)$, $\mathbf{P}^a(k)$, $\mathbf{P}^s(k)$, at each time step $1 \leq k \leq K$, where $K$ represents the total number of assimila-

14

274   tion steps. Kalman-based algorithms assume a Gaussian prior distribution $p\left(\mathbf{x}(k)|\mathbf{y}(1:k-1)\right) \sim$

275   $\mathcal{N}\left(\mathbf{x}^f(k), \mathbf{P}^f(k)\right)$. Then, filtering and smoothing estimates correspond to the Gaussian posterior

276   distributions $p\left(\mathbf{x}(k)|\mathbf{y}(1:k)\right) \sim \mathcal{N}\left(\mathbf{x}^a(k), \mathbf{P}^a(k)\right)$ and $p\left(\mathbf{x}(k)|\mathbf{y}(1:K)\right) \sim \mathcal{N}\left(\mathbf{x}^s(k), \mathbf{P}^s(k)\right)$ of the

277   state conditionally to past/present observations and past/present/future observations respectively.

278       The basic idea of the EnKF and EnKS is to use an ensemble $\mathbf{x}_1, \ldots, \mathbf{x}_{N_e}$ of size $N_e$ to track

279   Gaussian distributions over time with the empirical mean vector $\bar{\mathbf{x}} = {}^1\!/_{N_e} \sum_{i=1}^{N_e} \mathbf{x}_i$ and the empirical

280   error covariance matrix ${}^1\!/_{(N_e-1)} \sum_{i=1}^{N_e} \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^{\mathrm{T}}$.

$_{281}$      The EnKF/EnKS equations are divided into three main steps, $\forall i = 1, \ldots, N_e$ and $\forall k = 1, \ldots, K$:

<div align="center">Forecast step (forward in time):</div>

$$\mathbf{x}_i^f(k) = \mathcal{M}_k\left(\mathbf{x}_i^a(k-1)\right) + \eta_i(k) \tag{4a}$$

<div align="center">Analysis step (forward in time):</div>

$$\mathbf{d}_i(k) = \mathbf{y}(k) - \mathcal{H}_k\left(\mathbf{x}_i^f(k)\right) + \varepsilon_i(k) \tag{4b}$$

$$\mathbf{K}^f(k) = \mathbf{P}^f(k)\mathcal{H}_k^{\mathrm{T}}\left(\mathcal{H}_k\mathbf{P}^f(k)\mathcal{H}_k^{\mathrm{T}} + \mathbf{R}(k)\right)^{-1} \tag{4c}$$

$$\mathbf{x}_i^a(k) = \mathbf{x}_i^f(k) + \mathbf{K}^f(k)\mathbf{d}_i(k) \tag{4d}$$

<div align="center">Reanalysis step (backward in time):</div>

$$\mathbf{K}^s(k) = \mathbf{P}^a(k)\mathcal{M}_k^{\mathrm{T}}\left(\mathbf{P}^f(k+1)\right)^{-1} \tag{4e}$$

$$\mathbf{x}_i^s(k) = \mathbf{x}_i^a(k) + \mathbf{K}^s(k)\left(\mathbf{x}_i^s(k+1) - \mathbf{x}_i^f(k+1)\right) \tag{4f}$$

$_{282}$ with $\mathbf{K}^f(k)$ and $\mathbf{K}^s(k)$ the filter and smoother Kalman gains, respectively. Here, $\mathbf{P}^f(k)$ and

$_{283}$ $\mathcal{H}_k\mathbf{P}^f(k)\mathcal{H}_k^{\mathrm{T}}$ denote the empirical covariance matrices of $\mathbf{x}_i^f(k)$ and $\mathcal{H}_k(\mathbf{x}_i^f(k))$, respectively.

$_{284}$ Then, $\mathbf{P}^f(k)\mathcal{H}_k^{\mathrm{T}}$ and $\mathbf{P}^a(k)\mathcal{M}_k^{\mathrm{T}}$ denote the empirical cross-covariance matrices between $\mathbf{x}_i^f(k)$

$_{285}$ and $\mathcal{H}_k(\mathbf{x}_i^f(k))$ and between $\mathbf{x}_i^a(k)$ and $\mathcal{M}_k(\mathbf{x}_i^a(k))$, respectively. These quantities are estimated

$_{286}$ using $N_e$ ensemble members.

$_{287}$      In some of the methods presented in this review, the ensembles are also used to approximate $\mathcal{M}_k$

$_{288}$ and $\mathcal{H}_k$ by linear operators $\mathbf{M}_k$ and $\mathbf{H}_k$ such as

$$\mathbf{M}_k = \mathbf{E}_k^{\mathcal{M}(a)}(\mathbf{E}_{k-1}^a)^{\dagger} \tag{5a}$$

$$\mathbf{H}_k = \mathbf{E}_k^{\mathcal{H}(f)}(\mathbf{E}_k^f)^{\dagger} \tag{5b}$$

<div align="center">16</div>

with † the pseudo-inverse, $\mathbf{E}_k^{\mathscr{M}(a)}$, $\mathbf{E}_{k-1}^a$, $\mathbf{E}_k^{\mathscr{H}(f)}$ and $\mathbf{E}_k^f$ the matrices containing along their columns the ensemble perturbation vectors (the centered ensemble vectors) of $\mathscr{M}_k(\mathbf{x}_i^a(k-1))$, $\mathbf{x}_i^a(k-1)$, $\mathscr{H}_k(\mathbf{x}_i^f(k))$ and $\mathbf{x}_i^f(k)$, respectively.

In Eq. (4b), the innovation is denoted as $\mathbf{d}$ and tracked by $\mathbf{d}_1(k), \ldots, \mathbf{d}_{N_e}(k)$. The innovation is the key ingredient of the methods presented in sections 3 and 4.

## 3. Moment-based methods

In order to constrain the model and observational errors in DA systems, initial efforts were focused on the statistics of relevant variables which could contain information on covariances. The innovation, given in Eq. (4b), corresponds to the difference between the observations and the forecast in the observation space. This variable implicitly takes into account the $\mathbf{Q}$ and $\mathbf{R}$ covariances. Unfortunately, as explained in Blanchet et al. (1997), by using only current observations, their individual contributions cannot be easily disentangled. Thus, the techniques with only the classic innovation $\mathbf{y}(k) - \mathscr{H}_k(\mathbf{x}^f(k))$ are not discussed further in this review.

Two main approaches have been proposed in the literature to address this issue. They are based on the idea of producing multiple equations involving $\mathbf{Q}$ and $\mathbf{R}$. The first approach uses different type of innovation statistics (i.e., not only the classic one). The second approach is based on lag innovations, or differences between consecutive innovations. From a statistical point of view, they refer to the "methods of moments", where we construct a system of equations that links various moments of the innovations with the parameters and then replace the theoretical moments by the empirical ones in these equations.

17

*a. Innovation statistics in the observation space*

This first approach, based on the Desroziers diagnostic (Desroziers et al. 2005), is historical and now popular in the DA community. It does not exactly fit the topic of this review paper (i.e., estimating the model error $\mathbf{Q}$), since it is based on the inflation of the background covariance matrix $\mathbf{P}^f$. However, this forecast error covariance is defined by $\mathbf{P}^f(k) = \mathbf{M}_k \mathbf{P}^a(k-1)\mathbf{M}_k^{\mathrm{T}} + \mathbf{Q}$ in the Kalman filter, considering a linear model operator $\mathbf{M}_k$. Thus, even if DA systems do not use an explicit model error perturbation controlled by $\mathbf{Q}$, the inflation of the background covariance matrix $\mathbf{P}^f$ has similar effects, compensating for the lack of an explicit model uncertainty.

Desroziers et al. (2005) proposed examining various innovation statistics in the observation space. It is based on different type of innovation statistics between observations, forecasts and analysis, with all of them defined in the observation space: namely, $\mathbf{d}^{o-f}(k) = \mathbf{y}(k) - \mathscr{H}_k\left(\mathbf{x}^f(k)\right)$ as in Eq. (4b) and $\mathbf{d}^{o-a}(k) = \mathbf{y}(k) - \mathscr{H}_k\left(\mathbf{x}^a(k)\right)$. In theory, in the linear and Gaussian case, for unbiased forecast and observation, and when $\mathbf{P}^f(k)$ and $\mathbf{R}(k)$ are correctly specified, the Desroziers innovation statistics should verify the equalities:

$$
\begin{cases}
\mathrm{E}\left[\mathbf{d}^{o-f}(k)\mathbf{d}^{o-f}(k)^{\mathrm{T}}\right] = \mathbf{H}_k\mathbf{P}^f(k)\mathbf{H}_k^{\mathrm{T}} + \mathbf{R}(k) & \text{(6a)} \\
\mathrm{E}\left[\mathbf{d}^{o-a}(k)\mathbf{d}^{o-f}(k)^{\mathrm{T}}\right] = \mathbf{R}(k) & \text{(6b)}
\end{cases}
$$

with E the expectation operator. Equation (6a) is given by using Eq. (4b):

$$
\begin{aligned}
\mathbf{d}^{o-f}(k)\mathbf{d}^{o-f}(k)^{\mathrm{T}} = {}& -\mathbf{y}(k)\mathbf{x}^f(k)^{\mathrm{T}}\mathbf{H}_k^{\mathrm{T}} \\
& -\mathbf{H}_k\mathbf{x}^f(k)\mathbf{y}(k)^{\mathrm{T}} \\
& +\mathbf{H}_k\mathbf{x}^f(k)\mathbf{x}^f(k)^{\mathrm{T}}\mathbf{H}_k^{\mathrm{T}} \\
& +\mathbf{y}(k)\mathbf{y}(k)^{\mathrm{T}},
\end{aligned}
\tag{7}
$$

then applying the expectation operator and using the definition of $\mathbf{P}^f$ and $\mathbf{R}$. The observation-minus-forecast innovation statistics in Eq. (6a) is not useful to constrain model error $\mathbf{Q}$. Indeed,

18

₃₂₀ $\mathbf{d}^{o-f}$ does not depend explicitly on $\mathbf{Q}$, but rather on the forecast error covariance matrix $\mathbf{P}^f$. Thus,

₃₂₁ the combination of Eq. (6a) and Eq. (6b) can be used as a diagnosis of the forecast and obser-

₃₂₂ vational error covariances in the system. A mismatch between the Desroziers statistics and the

₃₂₃ actual covariances, namely the left- and right-hand side terms in Eq. (6a) and Eq. (6b), indicates

₃₂₄ inappropriate estimated covariances $\mathbf{P}^f(k)$ and $\mathbf{R}(k)$.

₃₂₅ The forecast covariance $\mathbf{P}^f$ is sometimes badly estimated in ensemble-based assimilation sys-

₃₂₆ tems. The limitations may be attributed to a number of causes. The limited number of ensemble

₃₂₇ members produces an over- or, most of the time, underestimation of the forecast variance. An-

₃₂₈ other limitation is the inaccuracies in methods used to sample initial condition or model error. The

₃₂₉ underestimation of the forecast covariance produces negative feedback, and the estimated analysis

₃₃₀ covariance $\mathbf{P}^a$ is thus underestimated, which in turn produces a further underestimation of the fore-

₃₃₁ cast covariance in the next cycle. This feedback process leads to filter divergence, as was pointed

₃₃₂ out by Pham et al. (1998), Anderson and Anderson (1999) or Anderson (2007). To avoid this

₃₃₃ filter divergence, inflating the forecast covariance $\mathbf{P}^f$ has been proposed. This covariance inflation

₃₃₄ accounts for both sampling errors and the lack of representation of model errors, like a too small

₃₃₅ amplitude for $\mathbf{Q}$ or the fact that a bias is omitted in $\boldsymbol{\eta}$ and $\boldsymbol{\epsilon}$, Eqs. (1) and (2). In this context, the

₃₃₆ diagnostics given by the Desroziers innovation statistics have been proposed as a tool to constrain

₃₃₇ the required covariance inflation in the system.

₃₃₈ We distinguish three inflation methods: multiplicative, additive and relaxation-to-prior. In the

₃₃₉ multiplicative case, the forecast error covariance matrix $\mathbf{P}^f$ is usually multiplied by a scalar coeffi-

₃₄₀ cient greater than 1 (Anderson and Anderson 1999). Using innovation statistics in the observation

₃₄₁ space, adaptive procedures to estimate this coefficient have been proposed by Wang and Bishop

₃₄₂ (2003), Anderson (2007), Anderson (2009) conditionally to the spatial location, Li et al. (2009),

₃₄₃ Miyoshi (2011), Bocquet (2011), Bocquet and Sakov (2012), Miyoshi et al. (2013), Bocquet et al.

344 (2015), El Gharamti (2018) and Raanes et al. (2019). In order to prevent excessive inflation or de-

345 flation, some authors have proposed assuming a priori distribution for the multiplicative inflation

346 factor. The most usual a priori distributions used by the authors are Gaussian in Anderson (2009),

347 inverse-gamma in El Gharamti (2018) or inverse chi-square in Raanes et al. (2019).

348 In practice, multiplicative inflation tends to excessively inflate in the data-sparse regions and

349 inflate too little in the densely observed regions. As a result, the spread looks like exaggeration of

350 data density (i.e., too much spread in sparsely observed regions, and vice versa). Additive inflation

351 solves this problem, but requires a lot of samples for additive noise; these drawbacks and benefits

352 are discussed in Miyoshi et al. (2010). In the additive inflation case, the diagonal terms of the

353 forecast and analysis empirical covariance matrices is increased (Mitchell and Houtekamer 2000;

354 Corazza et al. 2003; Whitaker et al. 2008; Houtekamer et al. 2009). This regularization also avoids

355 the problems corresponding to the inversion of the covariance matrices.

356 The last alternative is the relaxation-to-prior method. In application, this technique is more effi-

357 cient than both additive and multiplicative inflations because it maintains a reasonable spread struc-

358 ture. The idea is to relax the reduction of the spread at analysis. We distinguish the method pro-

359 posed in Zhang et al. (2004), where the forecast and analysis ensemble perturbations are blended,

360 from the one given in Whitaker and Hamill (2012), which multiplies the analysis ensemble with-

361 out blending perturbations. This last method is thus a multiplicative inflation, but applied after the

362 analysis, not the forecast. Finally, Ying and Zhang (2015) and Kotsuki et al. (2017b) proposed

363 methods to adaptively estimate the relaxation parameters using innovation statistics. Their con-

364 clusions are that adaptive procedures for relaxation-to-prior methods are robust to sudden changes

365 in the observing networks and observation error settings.

366 Closely connected to multiplicative inflation estimation is statistical modeling of the error vari-

367 ance terms proposed by Bishop and Satterfield (2013) and Bishop et al. (2013). From numerical

20

368 evidence based on the 10-dimensional Lorenz-96 model, the authors assume an inverse-gamma

369 prior distribution for these variances. This distribution allows for an analytic Bayesian update of

370 the variances using the innovations. Building on Bocquet (2011); Bocquet et al. (2015); Ménétrier

371 and Auligné (2015), this technique was extended in Satterfield et al. (2018) to adaptively tune a

372 mixing ratio between the true and sample variances.

373 Adaptive covariance inflations are estimation methods directly attached to a traditional filtering

374 method (such as the EnKF used here), with almost negligible overhead computational cost. In

375 practice, the use of this technique does not necessarily imply an additive error term $\boldsymbol{\eta}$ in Eq. (1).

376 Thus, it is not a direct estimation of $\mathbf{Q}$ but rather an inflation applied to $\mathbf{P}^f$ in order to compensate

377 for model uncertainties and sampling errors in the EnKFs, as explained in Raanes et al. (2019,

378 their section 4 and appendix C). Several DA systems work with an inflation method and use it for

379 its simplicity, low cost, and efficiency. As an example of inflation techniques, the most straight-

380 forward inflation estimation is a multiplicative factor $\lambda$ of the incorrectly scaled $\tilde{\mathbf{P}}^f(k)$, so that the

381 corrected forecast covariance is given by $\mathbf{P}^f(k) = \lambda(k)\tilde{\mathbf{P}}^f(k)$. The estimate of the inflation factor

382 is given by taking the trace of Eq. (6a):

$$\tilde{\lambda}(k) = \frac{\mathbf{d}^{o-f}(k)^{\mathrm{T}}\mathbf{d}^{o-f}(k) - \mathrm{Tr}\left(\mathbf{R}(k)\right)}{\mathrm{Tr}\left(\mathbf{H}_k\tilde{\mathbf{P}}^f(k)\mathbf{H}_k^{\mathrm{T}}\right)}. \tag{8}$$

383 The estimated inflation parameter $\tilde{\lambda}$ computed at each time $k$ can be noisy. The use of temporal

384 smoothing of the form $\lambda(k+1) = \rho\tilde{\lambda}(k) + (1-\rho)\lambda(k)$ is crucial in operational procedures. Al-

385 ternatively, Miyoshi (2011) proposed calculating the estimated variance of $\lambda(k)$, denoted as $\sigma_{\lambda(k)}^2$,

386 using the central limit theorem. Then, $\lambda(k+1)$ is updated using the previous estimate $\lambda(k)$ and

387 the Gaussian distribution with mean $\tilde{\lambda}(k)$ and variance $\sigma_{\lambda(k)}^2$. From the Desroziers diagnostics,

388 at each time step $k$ and when sufficient observations are available, an estimate of $\mathbf{R}(k)$ is possible

389 using Eq. (6b). For instance, Li et al. (2009) proposed estimating each component of a diagonal

21

$_{390}$ and averaged $\mathbf{R}$ matrix. However, the diagonal terms cannot take into account spatial correlated

$_{391}$ error terms, and constant values for observation errors are not realistic. Then, Miyoshi et al. (2013)

$_{392}$ proposed additionally estimating the off-diagonal components of the time-dependent matrix $\mathbf{R}(k)$.

$_{393}$ The Miyoshi et al. (2013) implementation is summarized in the appendix, Algorithm 1.

$_{394}$ The Desroziers diagnostic method has been applied widely to estimate the real observation error

$_{395}$ covariance matrix $\mathbf{R}$ in Numerical Weather Prediction (NWP). The observations are coming from

$_{396}$ different sources. In the case of satellite radiances, Bormann et al. (2010) applied three meth-

$_{397}$ ods, including the Desroziers diagnostic and the method detailed in Hollingsworth and Lönnberg

$_{398}$ (1986) to estimate a constant diagonal term of $\mathbf{R}$ using the innovation $\mathbf{d}^{o-f}$ and its correlations

$_{399}$ in space, assuming that horizontal correlations in $\mathbf{d}^{o-f}$ samples are purely due to $\mathbf{P}^f$. Weston

$_{400}$ et al. (2014) and Campbell et al. (2017) then included the inter-channel observation error correla-

$_{401}$ tions of satellite radiances in DA and obtained improved results compared with the case using a

$_{402}$ diagonal $\mathbf{R}$. For spatial error correlations in $\mathbf{R}$, Kotsuki et al. (2017a) estimated the horizontal ob-

$_{403}$ servation error correlations of satellite-derived precipitation data. Including horizontal observation

$_{404}$ error correlations in DA for densely-observed data from satellites and radars is more challenging

$_{405}$ than including inter-channel error correlations in DA. Indeed, the number of horizontally error-

$_{406}$ correlated observations is much larger, and some recent studies have been tackling this issue (e.g.,

$_{407}$ Guillet et al. (2019)).

$_{408}$ To conclude, the Desroziers diagnostic is a consistency check and makes it possible to detect if

$_{409}$ the error covariances $\mathbf{P}^f$ and $\mathbf{R}$ are incorrect. When and how this method can result in accurate

$_{410}$ or inaccurate estimates, and convergence properties, have been studied in depth by Waller et al.

$_{411}$ (2016) and Ménard (2016). The Desroziers diagnostic is also useful to estimate off-diagonal terms

$_{412}$ of $\mathbf{R}$, for instance taking into account the spatial error correlations. However, covariance localiza-

413 tion used in the ensemble Kalman filter might induce erroneous estimates of spatial correlations

414 (Waller et al. 2017).

## *b. Lag innovation between consecutive times*

416     Another way to estimate error covariances is to use multiple equations involving $\mathbf{Q}$ and $\mathbf{R}$,

417 exploiting cross-correlations between lag innovations. More precisely, it involves the current in-

418 novation $\mathbf{d}(k) = \mathbf{d}^{o-f}(k)$ defined in Eq. (4b) and past innovations $\mathbf{d}(k-1)$, ..., $\mathbf{d}(k-l)$. Lag

419 innovations were introduced by Mehra (1970) to recover $\mathbf{Q}$ and $\mathbf{R}$ simultaneously for Gaussian,

420 linear and stationary dynamic systems. In such a case, $\{\mathbf{d}(k)\}_{k \geq 1}$ is completely characterized by

421 the lagged covariance matrix $\mathbf{C}_l = \mathrm{Cov}(\mathbf{d}(k), \mathbf{d}(k-l))$, which is independent of $k$. In other words,

422 the information encoded in $\{\mathbf{d}(k)\}_{k \geq 1}$ is completely equivalent to the information provided by

423 $\{\mathbf{C}_l\}_{l \geq 0}$. Moreover, for linear systems in a steady state, analytic relations exist between $\mathbf{Q}$, $\mathbf{R}$ and

424 $\mathrm{E}\left[\mathbf{d}(k)\mathbf{d}(k-l)^{\mathrm{T}}\right]$. However, these linear relations can be dependent and redundant for different

425 lags $l$. Therefore, as stated in Mehra (1970), only a limited number of $\mathbf{Q}$ components can be

426 recovered.

427     Bélanger (1974) extended these results to the case of time-varying linear stochastic processes,

428 taking $\mathbf{d}(k)\mathbf{d}(k-l)^{\mathrm{T}}$ as "observations" of $\mathbf{Q}$ and $\mathbf{R}$ and using a secondary Kalman filter to update

429 them iteratively. On the one hand, considering the time-varying case may increase the number of

430 components in $\mathbf{Q}$ that can be estimated. On the other hand, as pointed out in Bélanger (1974),

431 this method would no longer be analytically exact if $\mathbf{Q}$ and $\mathbf{R}$ were updated adaptively at each

432 time step. One numerical difficulty of Bélanger's method is that it needs to invert a matrix of size

433 $m^2 \times m^2$, where $m$ refers to the dimension of the observation vector. However, this difficulty has

434 been largely overcome by Dee et al. (1985) in which the matrix inversion is reduced to $\mathscr{O}(m^3)$, by

435 taking the advantage of the fact that the big matrix comes from some tensor product.

23

<sup>436</sup> More recent work have focused on high-dimensional and nonlinear systems using the extended

<sup>437</sup> or ensemble Kalman filters. Berry and Sauer (2013) proposed a fast and adaptive algorithm in-

<sup>438</sup> spired by the use of lag innovations proposed by Mehra. Harlim et al. (2014) applied the original

<sup>439</sup> Bélanger algorithm empirically to a nonlinear system with sparse observations. Zhen and Harlim

<sup>440</sup> (2015) proposed a modified version of Bélanger's method, by removing the secondary filter and

<sup>441</sup> alternatively solving $\mathbf{Q}$ and $\mathbf{R}$ in a least-squares sense based on the averaged linear relation over a

<sup>442</sup> long term.

Here, we briefly describe the algorithm of Berry and Sauer (2013), considering the lag-zero and

lag-one innovations. The following equations are satisfied in the linear and Gaussian case, for

unbiased forecast and observation when $\mathbf{P}^f(k)$ and $\mathbf{R}(k)$ are correctly specified:

$$
\begin{cases}
\mathrm{E}\left[\mathbf{d}(k)\mathbf{d}(k)^{\mathrm{T}}\right] = \mathbf{H}_k\mathbf{P}^f(k)\mathbf{H}_k^{\mathrm{T}} + \mathbf{R}(k) = \Sigma(k) & \text{(9a)} \\[2mm]
\mathrm{E}\left[\mathbf{d}(k)\mathbf{d}(k-1)^{\mathrm{T}}\right] = \mathbf{H}_k\mathbf{M}_k\mathbf{P}^f(k-1)\mathbf{H}_{k-1}^{\mathrm{T}} & \\[2mm]
-\mathbf{H}_k\mathbf{M}_k\mathbf{K}^f(k-1)\Sigma(k-1). & \text{(9b)}
\end{cases}
$$

<sup>443</sup> Equation (9a) is equivalent to Eq. (6a). Moreover, Eq. (9b) results from the fact that developing

<sup>444</sup> the expression of $\mathbf{d}(k)$ using consecutively Eqs. (2), (1), (4a), and (4d), the innovation can be

<sup>445</sup> written as

$$
\begin{aligned}
\mathbf{d}(k) =& \mathbf{y}(k) - \mathbf{H}_k\mathbf{x}^f(k) \\
=& \mathbf{H}_k\left(\mathbf{x}(k) - \mathbf{x}^f(k)\right) + \boldsymbol{\epsilon}(k) \\
=& \mathbf{H}_k\left(\mathbf{M}_k\mathbf{x}(k-1) - \mathbf{x}^f(k) + \boldsymbol{\eta}(k)\right) + \boldsymbol{\epsilon}(k) \\
=& \mathbf{H}_k\left(\mathbf{M}_k\left(\mathbf{x}(k-1) - \mathbf{x}^a(k-1)\right) + \boldsymbol{\eta}(k)\right) + \boldsymbol{\epsilon}(k) \\
=& \mathbf{H}_k\mathbf{M}_k\left(\mathbf{x}(k-1) - \mathbf{x}^f(k-1) - \mathbf{K}^f(k-1)\mathbf{d}(k-1)\right) \\
& + \mathbf{H}_k\boldsymbol{\eta}(k) + \boldsymbol{\epsilon}(k). && \text{(10)}
\end{aligned}
$$

24

₄₄₆ Hence, the innovation product $\mathbf{d}(k)\mathbf{d}(k-1)^{\mathrm{T}}$ between two consecutive times is given by

$$\mathbf{H}_k\mathbf{M}_k\left(\mathbf{x}(k-1)-\mathbf{x}^f(k-1)\right)\mathbf{d}(k-1)^{\mathrm{T}}$$

$$-\mathbf{H}_k\mathbf{M}_k\left(\mathbf{K}^f(k-1)\mathbf{d}(k-1)\right)\mathbf{d}(k-1)^{\mathrm{T}}$$

$$+\mathbf{H}_k\boldsymbol{\eta}(k)\mathbf{d}(k-1)^{\mathrm{T}}+\boldsymbol{\epsilon}(k)\mathbf{d}(k-1)^{\mathrm{T}}, \tag{11}$$

₄₄₇ and assuming that the model $\boldsymbol{\eta}$ and observation $\boldsymbol{\epsilon}$ error noises are white and mutually uncorrelated,

₄₄₈ then $\mathrm{E}\left[\boldsymbol{\eta}(k)\mathbf{d}(k-1)^{\mathrm{T}}\right]=0$ and $\mathrm{E}\left[\boldsymbol{\epsilon}(k)\mathbf{d}(k-1)^{\mathrm{T}}\right]=0$. Finally, developing $\mathrm{E}\left[\mathbf{d}(k)\mathbf{d}(k-1)^{\mathrm{T}}\right]$,

₄₄₉ Eq. (9b) is satisfied.

₄₅₀ The algorithm in Berry and Sauer (2013) is summarized in the appendix, Algorithm 2. It is

₄₅₁ based on an adaptive estimation of $\mathbf{Q}(k)$ and $\mathbf{R}(k)$, which satisfies the following relations in the

₄₅₂ linear and Gaussian case:

$$\tilde{\mathbf{P}}(k)=(\mathbf{H}_k\mathbf{M}_k)^{-1}\mathbf{d}(k)\mathbf{d}(k-1)^{\mathrm{T}}\mathbf{H}_{k-1}^{-\mathrm{T}},$$

$$+\mathbf{K}^f(k-1)\mathbf{d}(k-1)\mathbf{d}(k-1)^{\mathrm{T}}\mathbf{H}_{k-1}^{-\mathrm{T}} \tag{12a}$$

$$\tilde{\mathbf{Q}}(k)=\tilde{\mathbf{P}}(k)-\mathbf{M}_{k-1}\mathbf{P}^a(k-2)\mathbf{M}_{k-1}^{\mathrm{T}}, \tag{12b}$$

$$\tilde{\mathbf{R}}(k)=\mathbf{d}(k)\mathbf{d}(k)^{\mathrm{T}}-\mathbf{H}_k\mathbf{P}^f(k)\mathbf{H}_k^{\mathrm{T}}. \tag{12c}$$

₄₅₃ In operational applications, when the number of observations is not equal to the number of

₄₅₄ components in state $\mathbf{x}$, $\mathbf{H}$ is not a square matrix and Eq. (12a) is ill-defined. To avoid the inversion

₄₅₅ of $\mathbf{H}$, Berry and Sauer (2013) proposed considering parametric models for $\mathbf{Q}$ and then solving a

₄₅₆ linear system associated with Eqs. (12a) and (12b). It is written as a least-squares problem such

25

that

$$
\begin{aligned}
\tilde{\mathbf{Q}}(k) = \arg\min_{\mathbf{Q}} || & \mathbf{d}(k)\mathbf{d}(k-1)^{\mathrm{T}} \\
& + \mathbf{H}_k\mathbf{M}_k\mathbf{K}^f(k-1)\mathbf{d}(k-1)\mathbf{d}(k-1)^{\mathrm{T}} \\
& - \mathbf{H}_k\mathbf{M}_k\mathbf{M}_{k-1}\mathbf{P}^a(k-2)\mathbf{M}_{k-1}^{\mathrm{T}}\mathbf{H}_{k-1}^{\mathrm{T}} \\
& - \mathbf{H}_k\mathbf{M}_k\mathbf{Q}\mathbf{H}_{k-1}^{\mathrm{T}}||.
\end{aligned}
\tag{13}
$$

In this adaptive procedure, joint estimations of $\tilde{\mathbf{Q}}(k)$ and $\tilde{\mathbf{R}}(k)$ can abruptly vary over time. Thus, the temporal smoothing of the covariances being estimated becomes crucial. As suggested by Berry and Sauer (2013), such temporal smoothing between current and past estimates is a reasonable choice:

$$
\mathbf{Q}(k+1) = \rho\tilde{\mathbf{Q}}(k) + (1-\rho)\mathbf{Q}(k),
\tag{14a}
$$

$$
\mathbf{R}(k+1) = \rho\tilde{\mathbf{R}}(k) + (1-\rho)\mathbf{R}(k)
\tag{14b}
$$

with $\mathbf{Q}(1)$ and $\mathbf{R}(1)$ the initial conditions and $\rho$ the smoothing parameter. When $\rho$ is large (close to 1), weight is given to the current estimates $\tilde{\mathbf{Q}}$ and $\tilde{\mathbf{R}}$, and when $\rho$ is small (close to 0) it gives smoother $\mathbf{Q}$ and $\mathbf{R}$ sequences. The value of $\rho$ is arbitrary and may depend on the system and how it is observed. For instance, in the case where the number of observations equals the size of the system, Berry and Sauer (2013) uses $\rho = 5 \times 10^{-5}$ in order to estimate the full matrix $\mathbf{Q}$ for the Lorenz-96 model.

The algorithm in Berry and Sauer (2013) only considers lag-zero and lag-one innovations. By incorporating more lags, Zhen and Harlim (2015) and Harlim (2018) showed that it makes it possible to deal with the case in which some components of Q are not identifiable from the method in Berry and Sauer (2013). For instance, let us consider the two-dimensional system with any stationary operator $\mathbf{M}$ and $\mathbf{H} = [1,0]$, meaning that only the first component of the system is

26

observed. This is a linear, Gaussian, stationary system, and Mehra's theory implies that two parameters of $\mathbf{Q}$ are identifiable. However, using only lag-one innovations as in Berry and Sauer (2013), Eq. (13) becomes a scalar equation and only one parameter of $\mathbf{Q}$ can be determined. The idea of considering more lag innovations to estimate more components of $\mathbf{Q}$ was tested in Zhen and Harlim (2015). Numerical results show that considering more than one lag can improve the estimates of $\mathbf{Q}$ and $\mathbf{R}$. For instance, Zhen and Harlim (2015) focused on the Lorenz-96 model. Results show that when $\mathbf{Q}$ is stationary, the trace of $\mathbf{Q}$ and $\mathbf{R}$ are equal, and when observations are taken at twenty fixed equally spaced grid points for every five integration time steps, the optimal RMSE of the estimates of $\mathbf{Q}$ and $\mathbf{R}$ is achieved when four time lags are considered. But with more lags, the performance is degraded.

To summarize, methods based on lag innovation between consecutive times have been studied for a long time in the signal processing community. The original methods (Mehra 1970; Bélanger 1974) were analytically established for linear systems with Gaussian noises. Inspired by these foundational ideas, empirical methods have been established for nonlinear systems in DA (Berry and Sauer 2013; Harlim et al. 2014; Zhen and Harlim 2015). Although these methods have not been tested in any operational experiment, the idea of using lagged innovations seems to have significant potential.

**4. Likelihood-based methods**

This section focuses on methods based on the likelihood of the observations, given a set of statistical parameters. The conceptual idea behind what we refer to as likelihood-based methods is to determine the optimal statistical parameters (i.e., $\mathbf{Q}$ and $\mathbf{R}$) that maximize the likelihood function for a given set of observations which may be distributed over time. In this way, the aim

27

$_{495}$ is to derive estimation methods that use the observations to find the most suitable, or most likely

$_{496}$ parameters.

$_{497}$ Early studies in Dee (1995), Blanchet et al. (1997), Mitchell and Houtekamer (2000) and Liang

$_{498}$ et al. (2012) proposed finding the optimal $\mathbf{Q}$ and $\mathbf{R}$ that maximize the current innovation likelihood

$_{499}$ at time $k$. Unfortunately, if only the current observations are used, the joint estimation of $\mathbf{Q}$ and $\mathbf{R}$

$_{500}$ is not well constrained (Todling 2015). To tackle this issue, several solutions have been recently

$_{501}$ proposed where the likelihood function considers observations distributed in time over several

$_{502}$ assimilation cycles.

$_{503}$ The likelihood-based methods are broadly divided into two categories. One approach uses a

$_{504}$ Bayesian framework. It assumes a priori knowledge about the parameters and estimate jointly the

$_{505}$ posterior distribution of $\mathbf{Q}$ and $\mathbf{R}$ together with the state of the system, or alternatively to estimate

$_{506}$ them in a two-stage process[2]. The second one is based on the frequentist viewpoint and attempts

$_{507}$ a point estimate of the parameters by maximizing a total likelihood function.

$_{508}$ *a. Bayesian inference*

$_{509}$ In the Bayesian framework, the elements of the covariance matrices $\mathbf{Q}$ and $\mathbf{R}$ are assumed to

$_{510}$ have a priori distributions which are controlled by hyperparameters. In practice, it is difficult to

$_{511}$ have prior distributions for each element of $\mathbf{Q}$ and $\mathbf{R}$, especially for large DA systems. Instead,

$_{512}$ parametric forms are used for the matrices, typically describing the shape and level noise. We

$_{513}$ denote the corresponding parameters as $\boldsymbol{\theta}$.

---

[2]Some of the methods presented in section 3 also use the Bayesian philosophy; for instance they assume a priori distribution for the multiplicative

inflation parameter $\lambda$ (Anderson 2009; El Gharamti 2018).

514   The inference in the Bayesian framework aims to determine the posterior density $p\left(\boldsymbol{\theta}|\mathbf{y}(1:k)\right)$.

515   Two techniques have appeared, the first based on a state augmentation and the second based on a

516   rigorous Bayesian update of the posterior distribution.

517   1) STATE AUGMENTATION

518   In the Bayesian framework, $\boldsymbol{\theta}$ is a random variable such that the state is augmented with these

519   parameters by defining $\mathbf{z}(k) = (\mathbf{x}(k), \boldsymbol{\theta})$. To define an augmented state-space model, one has to

520   define an evolution equation for the parameters. This leads to a new state-space model of the form

521   of Eqs. (1) and (2) with $\mathbf{x}$ replaced by $\mathbf{z}$. Therefore, the state and the parameters are estimated

522   jointly using the DA algorithms.

523   State augmentation was first proposed in Schmidt (1966) and is known as the Schmidt–Kalman

524   filter. This technique was mainly used to estimate both the state of the system and additional pa-

525   rameters, including bias, forcing terms and physical parameters. These kinds of parameters are

526   strongly related to the state of the system (Ruiz et al. 2013a). Therefore, they are identifiable

527   and suitable for an augmented state approach. However, Stroud and Bengtsson (2007) and later

528   Delsole and Yang (2010) formally demonstrated that augmentation methods fail for variance pa-

529   rameters like $\mathbf{Q}$ and $\mathbf{R}$. The explanation is that in the EnKF, the empirical forecast covariance $\mathbf{P}^f$

530   is computed using all the ensemble members, each one with a different realization of the random

531   variable $\boldsymbol{\theta}$. Thus, $\mathbf{P}^f$ and consequently the Kalman gain $\mathbf{K}^f$, are mixing the effects of $\mathbf{Q}$ and $\mathbf{R}$

532   parameters contained in $\boldsymbol{\theta}$. Therefore, after applying Eq. (4d), the update of $\mathbf{z}$ corresponding to

533   the $\boldsymbol{\theta}$ parameters is the same for all the parameters. To capture the impact of a single variance

534   parameter on the prediction covariance and circumvent the limitation of the state augmentation,

535   Scheffler et al. (2019) proposed to use an ensemble of states integrated with the same variance

536   parameter. The choice of an ensemble of states for each variance parameter leads to two nested

29

537 ensemble Kalman filters. The technique performs successfully under different model error covari-

538 ance structures but has an important computational cost.

539 Another critical aspect of state augmentation is that one needs to define an evolution model for

540 the augmented state $\mathbf{z}(k) = (\mathbf{x}(k), \boldsymbol{\theta}(k))$. If persistence is assumed in the parameters such that they

541 are constant in time, this leads to filter degeneracy, since the estimated variance of the error in $\boldsymbol{\theta}$

542 is bound to decrease in time. To prevent or at least mitigate this issue, it was suggested to use an

543 independent inflation factor on the parameters (Ruiz et al. 2013b) or to impose artificial stochastic

544 dynamics for $\boldsymbol{\theta}$, typically a random walk or AR(1) model, as introduced in Eq. (3) and proposed

545 in Liu and West (2001). The tuning of the parameters introduced in these artificial dynamics may

546 be difficult, and this introduces bias into the procedure, which is hard to quantify.

547 2) BAYESIAN UPDATE OF THE POSTERIOR DISTRIBUTION

548 Instead of the inference of the joint posterior density using a state augmentation strategy, the

549 state $\mathbf{x}(k)$ and parameters $\boldsymbol{\theta}$ can be divided into a two-step inference procedure using the following

550 formula:

$$p(\mathbf{x}(k), \boldsymbol{\theta}|\mathbf{y}(1:k)) =$$
$$p(\mathbf{x}(k)|\mathbf{y}(1:k), \boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathbf{y}(1:k)), \tag{15}$$

551 which is a direct consequence of the conditional density definition. In Eq. (15), $p(\mathbf{x}(k)|\mathbf{y}(1:k), \boldsymbol{\theta})$

552 represents the posterior distribution of the state, given the observations and the parameter $\boldsymbol{\theta}$. It can

553 be computed using a filtering DA algorithm. The second term on the right-hand side of Eq. (15)

554 corresponds to the posterior distribution of the parameters, given the observations up to time $k$.

555   The latter can be updated sequentially using the following Bayesian hierarchy:

$$p\left(\boldsymbol{\theta}|\mathbf{y}(1:k)\right) \propto$$

$$p\left(\mathbf{y}(k)|\mathbf{y}(1:k-1),\boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}|\mathbf{y}(1:k-1)\right), \tag{16}$$

556   where $p\left(\mathbf{y}(k)|\mathbf{y}(1:k-1),\boldsymbol{\theta}\right)$ is the likelihood of the innovations.

557   Different approximations have been used for $p\left(\boldsymbol{\theta}|\mathbf{y}(1:k)\right)$ in Eq. (16); these include parametric

558   models based on Gaussian (Stroud et al. 2018), inverse-gamma (Stroud and Bengtsson 2007) or

559   Wishart distributions (Ueno and Nakamura 2016), particle-based approximations (Frei and Künsch

560   2012; Stroud et al. 2018) and grid-based approximation (Stroud et al. 2018).

561   The methods proposed in the literature also differ by the approximation used for the likelihood

562   of the innovations. We emphasize that $p\left(\mathbf{y}(k)|\mathbf{y}(1:k-1),\boldsymbol{\theta}\right)$ needs to be evaluated for different

563   values of $\boldsymbol{\theta}$ at each time step, and that this requires applying the filter from the initial time with

564   a single value of $\boldsymbol{\theta}$, which is computationally impossible for applications in high dimensions. To

565   reduce computational time, it is generally assumed that $\mathbf{x}^f$ and $\mathbf{P}^f$ are independent of $\boldsymbol{\theta}$, and only

566   observations $\mathbf{y}(k-l:k-1)$ in a small time window from the current observation are used when

567   computing the likelihood of the innovations (see Ueno and Nakamura (2016); Stroud et al. (2018)

568   for a more detailed discussion). A summary of the Bayesian method from Stroud et al. (2018) is

569   given in the appendix, Algorithm 3. It was implemented within the EnKF framework and is one

570   of the most recent studies based on the Bayesian approach.

571   Applications of the Bayesian methodology in the DA context are now discussed. It has mainly

572   been used to estimate shape and noise parameters of $\mathbf{Q}$ and $\mathbf{R}$ error covariance matrices. For

573   instance, Purser and Parrish (2003) and Solonen et al. (2014) estimated statistical parameters con-

574   trolling the magnitude of the variance and the spatial dependencies in the model error $\mathbf{Q}$, assuming

575   that $\mathbf{R}$ is known. There are also applications aimed at estimating parameters governing the shape

31

576 of the observation error covariance matrix $\mathbf{R}$ only: Frei and Künsch (2012) and Stroud et al. (2018)

577 in the Lorenz-96 system, Winiarek et al. (2012, 2014) for the inversion of the source term of air-

578 borne radionuclides using a regional atmospheric model, and Ueno and Nakamura (2016) using a

579 shallow-water model to assimilate satellite altimetry.

580 As pointed out in Stroud and Bengtsson (2007), Bayesian update algorithms work best when the

581 number of unknown parameters in $\boldsymbol{\theta}$ is small. This limitation may explain why the joint estimation

582 of parameters controlling both model and observation error covariances is not systematically ad-

583 dressed. For instance, Stroud and Bengtsson (2007) used the EnKF with the Lorenz-96 model for

584 the estimation of a common multiplicative scalar parameter for predefined matrices $\mathbf{Q}$ and $\mathbf{R}$. Al-

585 ternatively, Stroud et al. (2018) tested the Bayesian method on different spatio-temporal systems

586 to estimate the signal-to-noise ratio between $\mathbf{Q}$ and $\mathbf{R}$. Nevertheless, based on the experiments

587 about the importance of the signal-to-noise ratio $\left\|\mathbf{P}^f\right\| / \|\mathbf{R}\|$ presented in Fig. 2, we know that this

588 estimation of the ratio is not optimal.

589 Widely used in the statistical community, the Bayesian framework is useful incorporating phys-

590 ical knowledge about error covariance matrices and constraining their estimation process. In the

591 DA literature, authors have used a priori distributions for the shape and noise parameters of $\mathbf{Q}$

592 or $\mathbf{R}$, but rarely both. Operationally, only a limited number of parameters can be estimated. To

593 address this issue, Stroud and Bengtsson (2007) suggested combining Bayesian algorithms with

594 other techniques.

595 *b. Maximization of the total likelihood.*

596 The innovation likelihood at time $k$, $p(\mathbf{y}(k)|\mathbf{y}(1:k-1),\boldsymbol{\theta})$ in Eq. (16), can be maximized to

597 find the optimal $\boldsymbol{\theta}$ (i.e., $\mathbf{Q}$ and $\mathbf{R}$ matrices or parameterizations of them). In practice, when this

598 maximization is done at each time step, two issues arise. Firstly, the innovation covariance matrix

32

₅₉₉ $\boldsymbol{\Sigma}(k) = \mathbf{H}_k \mathbf{P}^f(k) \mathbf{H}_k^{\mathrm{T}} + \mathbf{R}(k)$ combines the information about $\mathbf{R}$ and $\mathbf{Q}$, the latter being contained

₆₀₀ in $\mathbf{P}^f$. When using only time $k$, it is difficult to disentangle the model and observation error

₆₀₁ covariances; in application, the aforementioned studies only estimated one of them. Secondly,

₆₀₂ the number of observations at each time step is in general limited and, as pointed out by Dee

₆₀₃ (1995), available observations should exceed "the number of tunable parameters by two or three

₆₀₄ orders of magnitude". To overcome these limitations, a reasonable alternative is to use a batch of

₆₀₅ observations within a time window and to assume $\boldsymbol{\theta}$ to be constant in time. The resulting total

₆₀₆ likelihood expressed sequentially through conditioning is given by

$$p\left(\mathbf{y}(1:K)|\boldsymbol{\theta}\right) = \prod_{k=1}^{K} p\left(\mathbf{y}(k)|\mathbf{y}(1:k-1),\boldsymbol{\theta}\right). \tag{17}$$

₆₀₇ Because it is an integration of innovation likelihoods over a long period of time from $k = 1$ to $k =$

₆₀₈ $K$, Eq. (17) provides more observational information to estimate $\mathbf{Q}$ and $\mathbf{R}$. The maximization of

₆₀₉ this total likelihood has been applied for the estimation of deterministic and stochastic parameters

₆₁₀ (related to $\mathbf{Q}$) using a direct sequential optimization procedure (Delsole and Yang 2010). Ueno

₆₁₁ et al. (2010) used a grid-based procedure to estimate noise levels and spatial correlation lengths of

₆₁₂ $\mathbf{Q}$ and a noise level for $\mathbf{R}$. This grid-based method uses predefined sets of covariance parameters

₆₁₃ and evaluates the different combinations to find the one that maximizes the likelihood criterion.

₆₁₄ Brankart et al. (2010) also proposed a method using the same criterion but adding (at the initial

₆₁₅ time) information on scale and correlation length parameters of $\mathbf{Q}$ and $\mathbf{R}$. This information is only

₆₁₆ given the first time, and is progressively forgotten over time, using a decreasing exponential factor.

₆₁₇ The marginalization of the hidden state in Eq. (17) considers all the previous observations, and it

₆₁₈ requires the use of a filter. The maximization of the total likelihood $p\left(\mathbf{y}(1:K)|\boldsymbol{\theta}\right)$ to estimate

₆₁₉ model error covariance $\mathbf{Q}$ was conducted in Pulido et al. (2018), where they used a gradient-based

₆₂₀ optimization technique and the EnKF.

33

<sub>621</sub> The likelihood function given in Eq. (17) only depends on the observations **y**. This likelihood

<sub>622</sub> can be written in a different way, taking into account both the observations and the hidden state **x**.

<sub>623</sub> Indeed, the marginalization of the hidden state to obtain the total likelihood can be produced using

<sub>624</sub> the whole trajectory of the state from $k = 0$ to the last time step $K$ all at once. It is given by

$$p(\mathbf{y}(1:K)|\boldsymbol{\theta}) = \int p(\mathbf{x}(0:K), \mathbf{y}(1:K)|\boldsymbol{\theta}) \mathrm{d}\mathbf{x}(0:K). \tag{18}$$

<sub>625</sub> The maximization of the total likelihood as a function of statistical parameters $\boldsymbol{\theta}$ is not possible,

<sub>626</sub> since the total likelihood cannot be evaluated directly, nor its gradient with regard to the parameters

<sub>627</sub> (Pulido et al. 2018). Shumway and Stoffer (1982) proposed using an iterative procedure based on

<sub>628</sub> the expectation–maximization algorithm (hereinafter denoted as EM). They applied it to estimate

<sub>629</sub> the parameters of a linear state-space model, with linear dynamics, and a linear observational

<sub>630</sub> operator and Gaussian errors. The EM algorithm was introduced by Dempster et al. (1977).

<sub>631</sub> Each iteration of the EM algorithm consists of two steps. In the expectation step (E-step), the

<sub>632</sub> posterior density $p(\mathbf{x}(0:K)|\mathbf{y}(1:K), \boldsymbol{\theta}_{(n)})$ is determined conditioned on the batch of observations

<sub>633</sub> $\mathbf{y}(1:K)$ and given the parameters $\boldsymbol{\theta}_{(n)} = (\mathbf{Q}_{(n)}, \mathbf{R}_{(n)})$ from the previous iteration or initial guess.

<sub>634</sub> This is obtained through the application of a smoother like the EnKS. Then, the M-step relies on

<sub>635</sub> the maximization of an intermediate function, depending on the posterior density obtained in the

<sub>636</sub> E-step. The intermediate function is defined by the conditional expectation

$$\mathrm{E}\left[\log\left(p(\mathbf{x}(0:K), \mathbf{y}(1:K)|\boldsymbol{\theta})\right)|\mathbf{y}(1:K), \theta_{(n)}\right]. \tag{19}$$

<sub>637</sub> If as in Eqs. (1) and (2) the observational and model errors are assumed to be additive, unbiased

<sub>638</sub> and Gaussian, the expression for the logarithm of the joint density in Eq. (19) is given by

$$-\frac{1}{2}\{\sum_{k=1}^{K} \|\mathbf{x}(k) - \mathscr{M}(\mathbf{x}(k-1))\|_{\mathbf{Q}}^2 + \log|\mathbf{Q}|$$

$$+ \|\mathbf{y}(k) - \mathscr{H}(\mathbf{x}(k))\|_{\mathbf{R}}^2 + \log|\mathbf{R}|\} + c \tag{20}$$

34

639 where $\|\mathbf{v}\|_{\mathbf{A}}^2$ is defined to be equal to $\mathbf{v}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{v}$ and $c$ is a constant independent of $\mathbf{Q}$ and $\mathbf{R}$. In this

640 case, an analytic expression for the optimal error covariances at each iteration of the EM algorithm

641 can be obtained. The estimators of the parameters that maximize Eq. (19) using Eq. (20) are

$$\mathbf{Q}_{(n+1)} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{E}[(\mathbf{x}(k) - \mathscr{M}(\mathbf{x}(k-1)))$$
$$(\mathbf{x}(k) - \mathscr{M}(\mathbf{x}(k-1)))^{\mathrm{T}}|\mathbf{y}(1:K), \theta_{(n)}] \tag{21a}$$

642 and

$$\mathbf{R}_{(n+1)} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{E}[(\mathbf{y}(k) - \mathscr{H}(\mathbf{x}(k)))$$
$$(\mathbf{y}(k) - \mathscr{H}(\mathbf{x}(k)))^{\mathrm{T}}|\mathbf{y}(1:K), \theta_{(n)}]. \tag{21b}$$

643 The application of the EM algorithm for the estimation of $\mathbf{Q}$ and $\mathbf{R}$ is rather straightforward.

644 Starting from $\mathbf{Q}_{(1)}$ and $\mathbf{R}_{(1)}$, an ensemble Kalman smoother is applied with this first guess and

645 the batch of observations $\mathbf{y}(1:K)$ to obtain the posterior density $p(\mathbf{x}(0:K)|\mathbf{y}(1:K), \theta_{(1)})$. Then

646 Eqs. (21a) and (21b) are used to update and obtain $\mathbf{Q}_{(2)}$ and $\mathbf{R}_{(2)}$. Next, a new application of

647 the smoother is conducted using the parameters $\mathbf{Q}_{(2)}$ and $\mathbf{R}_{(2)}$ and the observations $\mathbf{y}(1:K)$, the

648 new resulting states are used in Eqs. (21a) and (21b) to estimate $\mathbf{Q}_{(3)}$ and $\mathbf{R}_{(3)}$, and so on. As

649 a diagnostic of convergence or as a stop criterion, the product of innovation likelihood functions

650 given in Eq. (17) is evaluated using a filter. The EM algorithm guarantees that the total likelihood

651 increases in each iteration and that the sequence $\theta_{(n)}$ converges to a local maximum (Wu 1983).

652 A summary of the EM method (using EnKF and EnKS) from Dreano et al. (2017) is given in the

653 appendix, Algorithm 4.

654 EM is a well-known algorithm used in the statistical community. This procedure is parameter-

655 free and robust, due to the large number of observations used to approximate the likelihood when

656 using a long batch period (Shumway and Stoffer 1982). Although the use of the EM algorithm is

35

still limited in DA, it is becoming more and more popular. Some studies have implemented the EM algorithm for estimating only the observation error matrix $\mathbf{R}$. For instance, Ueno and Nakamura (2014) used the model proposed in Zebiak and Cane (1987) and satellite altimetry observations, whereas Liu et al. (2017) used an air quality model for accidental pollutant source retrieval. But the estimation of only the observation error covariance is limited, and other studies have tried to jointly estimate model error $\mathbf{Q}$ and $\mathbf{R}$ matrices, for instance as in Tandeo et al. (2015) for an orographic subgrid-scale nonlinear observation operator. Then, Dreano et al. (2017) and Pulido et al. (2018) used the EM procedure to produce joint estimation of $\mathbf{Q}$ and $\mathbf{R}$ matrices in the Lorenz-63 and stochastic parameters of the Lorenz-96 systems, respectively. Recently, Yang and Mémin (2019) extended the EM procedure for the estimation of physical parameters in a one-dimensional shallow water model, more specifically for the identification of stochastic subgrid terms. Lastly, an online adaptation of the EM algorithm for the estimation of $\mathbf{Q}$ and $\mathbf{R}$ at each time step, after the filtering procedure, has been proposed in Cocucci et al. (2020). In this adaptive case, the likelihood is averaged locally over time, see Cappé (2011) for more details.

To our knowledge, EM has not been tested yet on operational systems with large observation- and state-space. In that case, the use of parametric forms for the matrices $\mathbf{Q}$ and $\mathbf{R}$ is essential to reduce the number of statistical parameters $\boldsymbol{\theta}$ to estimate. For instance, Dreano et al. (2017) and Liu et al. (2017) showed that in the particular cases where covariances are diagonal or of the form $\alpha\mathbf{A}$ with $\mathbf{A}$ a positive definite matrix, expressions in Eq. (21a) and Eq. (21b) are simplified, and a suboptimal $\boldsymbol{\theta}$ in the space of the parametric covariance form can be obtained.

## 5. Other methods

In this section, we describe other methods that have been used to estimate $\mathbf{Q}$ and $\mathbf{R}$, and that cannot be included in the categories presented in the previous sections. In particular, we report

36

680 here about methods that are applied either a posteriori, after DA cycles, or without applying any

681 DA algorithms.

*a. Analysis (or reanalysis) increment approach*

683 This first method is based on previous DA outputs. The key idea here is to use the analysis

684 (or reanalysis) increments to provide a realistic sample of model errors from which statistical

685 moments, such as the covariance matrix $\mathbf{Q}$, can be empirically estimated. This assumes that the

686 sequence of reanalysis $\mathbf{x}^s$ (or analysis $\mathbf{x}^a$) is the best available representation of the true process $\mathbf{x}$.

687 In that case, the following approximation in Eq. (1) is made:

$$
\begin{aligned}
\boldsymbol{\eta}(k) = & \mathscr{M}\left(\mathbf{x}(k-1)\right) - \mathbf{x}(k) \\
\approx & \mathscr{M}\left(\mathbf{x}^s(k-1)\right) - \mathbf{x}^s(k).
\end{aligned}
\tag{22}
$$

688 In this approximation, it is implicitly assumed that the estimated state is the truth, so that the initial

689 condition at time $k-1$ is neglected. A similar approximation of the true process by $\mathbf{x}^a$ or $\mathbf{x}^s$ in

690 Eq. (2) can be used to estimate the observation error covariance matrix $\mathbf{R}$.

691 Operationally, the analysis (or reanalysis) increment method is applied after a DA filter (or

692 smoother) to estimate the $\mathbf{Q}$ matrix. This method was originally introduced by Leith (1978), and

693 later used to account for model error in the context of ensemble Kalman filters, using analysis and

694 reanalysis increments by Mitchell and Carrassi (2015), and in the context of weak-constraint vari-

695 ational assimilation by Bowler (2017). Along this line, Rodwell and Palmer (2007) also proposed

696 evaluating the average of instantaneous analysis increments to represent the systematic forecast

697 tendencies of a model.

37

*b. Covariance matching*

The covariance matching method was introduced by Fu et al. (1993). It involves matching sample covariance matrices to their theoretical expectations. Thus, it is a method of moments, similar to the work in Desroziers et al. (2005), except that covariance matching is performed on a set of historical observations and numerical simulations (noted $\mathbf{x}^{sim}$), without applying any DA algorithms. It has been extended by Menemenlis and Chechelnitsky (2000) to time-lagged innovations, as first considered in Bélanger (1974).

In the case of a constant and linear observation operator $\mathbf{H}$, the basic idea in Fu et al. (1993) is to assume the following system

$$\begin{cases} \mathbf{x}^{sim}(k) = \mathbf{x}(k) + \boldsymbol{\eta}^{sim}(k), & \text{(23a)} \\[2mm] \boldsymbol{\eta}^{sim}(k) = \mathbf{A}\boldsymbol{\eta}^{sim}(k-1) + \boldsymbol{\eta}(k), & \text{(23b)} \\[2mm] \mathbf{H}\mathbf{x}^{sim}(k) - \mathbf{y}(k) = \mathbf{H}\boldsymbol{\eta}^{sim}(k) + \boldsymbol{\epsilon}(k), & \text{(23c)} \end{cases}$$

with $\mathbf{A}$ a transition matrix close to the identity matrix, assuming slow variations of the numerical simulation errors (noted $\boldsymbol{\eta}^{sim}$). In Eq. (23b) and Eq. (23c), the definitions of $\boldsymbol{\eta}$ and $\boldsymbol{\epsilon}$ errors remain similar, as in the general Eqs. (1) and (2).

Assuming that $\mathbf{Q}$ and $\mathbf{R}$ are constant over time, $\boldsymbol{\epsilon}$ is uncorrelated from $\mathbf{x}$ and from $\boldsymbol{\eta}^{sim}$, then Eq. (23c) and Eq. (23a) yield to the following estimates of $\mathbf{R}$ and $\mathbf{P}^{sim}$ (the latter represents the error covariance of the numerical simulations):

$$\begin{aligned} \widehat{\mathbf{R}} = \frac{1}{2}\{ &\mathrm{E}[(\mathbf{y}-\mathbf{H}\mathbf{x}^{sim})(\mathbf{y}-\mathbf{H}\mathbf{x}^{sim})^{\mathrm{T}}] \\ &- \mathrm{E}[(\mathbf{H}\mathbf{x}^{sim})(\mathbf{H}\mathbf{x}^{sim})^{\mathrm{T}}] + \mathrm{E}[\mathbf{y}\mathbf{y}^{\mathrm{T}}]\}, \end{aligned} \tag{24a}$$

$$\begin{aligned} \mathbf{H}\widehat{\mathbf{P}}^{sim}\mathbf{H}^{\mathrm{T}} = \frac{1}{2}\{ &\mathrm{E}[(\mathbf{y}-\mathbf{H}\mathbf{x}^{sim})(\mathbf{y}-\mathbf{H}\mathbf{x}^{sim})^{\mathrm{T}}] \\ &+ \mathrm{E}[(\mathbf{H}\mathbf{x}^{sim})(\mathbf{H}\mathbf{x}^{sim})^{\mathrm{T}}] - \mathrm{E}[\mathbf{y}\mathbf{y}^{\mathrm{T}}]\}. \end{aligned} \tag{24b}$$

38

$_{711}$ where E is the expectation operator over time. Then, an estimate of $\mathbf{Q}$ is obtained using Eq. (23b),

$_{712}$ Eq. (24b) and assuming that $\mathbf{P}^{sim}$ has a unique time-invariant limit.

$_{713}$ *c. Forecast sensitivity*

$_{714}$ In operational meteorology, it is critical to learn the sensitivity of the forecast accuracy to various

$_{715}$ parameters of a DA system, in particular the error statistics of both the model and the observations.

$_{716}$ This is why a significant portion of literature considers the tuning problem of $\mathbf{R}$ and $\mathbf{Q}$ through the

$_{717}$ lens of the sensitivity of the forecast to these parameters. The computation of those sensitivities can

$_{718}$ be seen as a first-order correction or diagnostic for such an estimation. The forecast sensitivities are

$_{719}$ computed either using the adjoint model (Daescu and Todling 2010; Daescu and Langland 2013)

$_{720}$ in the context of variational methods, or a forecast ensemble (Hotta et al. 2017) in the context of

$_{721}$ the EnKF.

$_{722}$ The basic idea is to compute at each assimilation cycle an innovation between forecast and anal-

$_{723}$ ysis, noted $\mathbf{d}^{f-a}(k) = \mathbf{x}^f(k) - \mathbf{x}^a(k)$. Then, the forecast sensitivity is given by $\mathbf{d}^{f-a}(k)^{\mathrm{T}}\mathbf{S}\mathbf{d}^{f-a}(k)$

$_{724}$ with $\mathbf{S}$ a diagonal scaling matrix, to normalize the components of $\mathbf{d}^{f-a}$. $\mathbf{Q}$ and $\mathbf{R}$ estimates are the

$_{725}$ matrices that minimize $\mathbf{d}^{f-a}(k)$. The adjoint or the ensemble are thus used to compute the partial

$_{726}$ derivatives of this forecast sensitivity. w.r.t. $\mathbf{Q}$ and $\mathbf{R}$.

$_{727}$ **6. Conclusions and perspectives**

$_{728}$ As often considered in data assimilation, this review paper also deals with model and observation

$_{729}$ errors that are assumed additive and Gaussian with covariance matrices $\mathbf{Q}$ and $\mathbf{R}$. The model error

$_{730}$ corresponds to the dynamic model deficiencies to represent the underlying physics, whereas the

$_{731}$ observation error corresponds to the instrumental noise and the representativity error. Model and

39

732 observation errors are assumed to be uncorrelated and white in time. The model and observations

733 are also assumed unbiased, a strong assumption for real data assimilation applications.

734 The discussion starts with the aid of an illustration of the individual and joint impacts of im-

735 properly calibrated covariances using a linear toy model. The experiments clearly showed that

736 to achieve reasonable filter accuracy (i.e., in terms of root mean squared error), it is crucial to

737 carefully define both $\mathbf{Q}$ and $\mathbf{R}$. The effect on the coverage probability of a mis-specification of

738 $\mathbf{Q}$ or $\mathbf{R}$ is also highlighted. This coverage probability is related to the estimated covariance of

739 the reconstructed state, and thus to the uncertainty quantification in data assimilation. After the

740 one-dimensional illustration, the core of the paper gives an overview of various methods to jointly

741 estimate the $\mathbf{Q}$ and $\mathbf{R}$ error covariance matrices: they are summarized and compared below.

## a. Comparison of existing methods for estimating $\mathbf{Q}$ and $\mathbf{R}$

743 We mainly focused in this review on four methodologies for the joint estimation of the error co-

744 variances $\mathbf{Q}$ and $\mathbf{R}$. The methods are summarized in Table 1. They correspond to classic estimation

745 methods, based on statistical moments or likelihoods. The main difference between the four meth-

746 ods comes from the innovations taken into account: the total innovation, as in the EM algorithm

747 proposed by Shumway and Stoffer (1982); lag innovations, following the idea given in Mehra

748 (1970); or different type of innovations in the observation space, as in Desroziers et al. (2005).

749 Additionally, to constrain the estimation, hierarchical Bayesian approaches use prior distributions

750 for the shape parameters of $\mathbf{Q}$ and $\mathbf{R}$.

751 Most of the methods estimate the model error $\mathbf{Q}$. The exception is the one using the Desroziers

752 diagnostic, dealing with different type of innovations in the observation space, which instead esti-

753 mates an inflation factor for $\mathbf{P}^f$. Moreover, the methods are mainly defined online, meaning that

754 they aim to estimate $\mathbf{Q}$ and $\mathbf{R}$ adaptively, together with the current state of the system. Conse-

40

quently, these methods require additional tunable parameters to smooth the estimated covariances over time. However, most of the methods presented in this review also have an offline variant. In that case, a batch of observations is used to estimate $\mathbf{Q}$ and $\mathbf{R}$. In some methods, such as the EM algorithm, the parameters are determined iteratively. These offline approaches avoid the use of additional smoothing parameters.

Throughout this review paper, as usually stated in DA, it is assumed that model error $\boldsymbol{\eta}$ and observation error $\boldsymbol{\epsilon}$, defined in Eqs. (1) and (2), are Gaussian. Consequently, the distribution of the innovations are also Gaussian. The four presented methods use this property to build estimates of $\mathbf{Q}$ and $\mathbf{R}$ adequately. But, if $\boldsymbol{\eta}$ and $\boldsymbol{\epsilon}$ are non-Gaussian, Desroziers diagnostic and lag-innovation methods are not suitable anymore. However, the EM procedures and Bayesian methods are still relevant, although they must be used with an appropriate filter (e.g., particle filters), not Kalman-based algorithms (i.e., assuming a Gaussian distribution of the state). Recently, the treatment of non-Gaussian error distributions in DA has been explored in Katzfuss et al. (2019), using hierarchical state-space models. This Bayesian framework allows to handle unknown variables that cannot be easily included in the state vector (e.g., parameters of $\mathbf{Q}$ and $\mathbf{R}$) and to model non-Gaussian observations.

The four methods have been applied at different levels of complexity. For instance, Bayesian inference methods (due to their algorithm complexity) and the EM algorithm (due to its computational cost) have so far only been applied to small dynamic models. However, the online version of the EM algorithm is less consuming and opens new perspectives of applications on larger models. On the other hand, methods using innovation statistics in the observation space have already been applied to NWP models.

777 The four methods summarized in Table 1 show differences in maturity in terms of applications

778 and methodological aspects. This review also shows that there are still remaining challenges and

779 possible improvements for the four methods.

780 *b. Remaining challenges for each method*

781 The first challenge concerns the improvements of adaptive techniques regarding additional pa-

782 rameters that control the variations of **Q** and **R** estimates over time. Instead of using fixed values

783 for these parameters, for instance fixed $\rho$ in the lag innovations or $\sigma_\lambda^2$ in the inflation methods,

784 we suggest using time-dependent adaptations. This adaptive solution could avoid the problems

785 of instabilities close to the solution. Another option could be to adapt these procedures, working

786 with stable parameter values (small $\rho$, low $\sigma_\lambda^2$) and iterating the procedures on a batch of obser-

787 vations, as in the EM algorithm. This offline variant was suggested and tested in Desroziers et al.

788 (2005) with encouraging results. To the best of our knowledge, it has not yet been tested with

789 lag-innovation methods.

790 The second challenge concerns considering time-varying error covariance matrices. The adap-

791 tive procedures, based on online estimations with temporal smoothing of **Q** and **R**, are supposed

792 to capture slowly evolving covariances. On the contrary, offline methods like the EM algorithm

793 are working on a batch of observations, assuming that **Q** and **R** are constant over the batch period.

794 Online solutions for the EM algorithm, with the likelihood averaged locally over time (Cocucci

795 et al. 2020), could also capture slow evolution of the covariances. Another simple solution could

796 be to work on small sets of observations, named as mini-batches, and to apply the EM algorithm

797 in each set using the previous estimates as an initial guess. These intermediate schemes are of

798 common use in machine learning.

799     A third challenge has to do with the assumption, used by all of the methods described herein, that

800   observation and model errors are mutually independent. Nevertheless, as pointed out in Berry and

801   Sauer (2018), observation and model error are often correlated in real data assimilation problems

802   (e.g., for satellite retrieval of Earth observations that uses model outputs in the inversion process).

803   Methods based on Bayesian inference can, in principle, exploit existing model-to-observation cor-

804   relations by using a prior joint distribution (i.e., not two individual ones). The explicit taking into

805   account of this correlation can then constrain the optimization procedure. This is not possible in

806   the other approaches described in this review, at least not in their standard known formulations,

807   and the presence of model-observation correlation can deteriorate their accuracy.

808     A fourth challenge is common to all the methods presented in this review. Iterative versions

809   of the presented algorithms need initial values or distributions for $\mathbf{R}$ and $\mathbf{Q}$ (or $\mathbf{B} = \mathbf{P}^f$ in the

810   case of Desroziers). But, as mentioned in Waller et al. (2016) for the Desrorziers diagnostics,

811   there is no guarantee that the algorithms will converge to the optimal solution. Indeed, in such

812   an optimization problem, there are possibly several local and non-optimal solutions. Suboptimal

813   specifications of $\mathbf{R}$, $\mathbf{Q}$, or $\mathbf{B}$ in the initial DA cycle will affect the final estimation results. There

814   are several solutions to avoid this convergence problem: initialize the covariance matrices using

815   physical expertise, execute the iterative algorithms several times with different initial covariance

816   matrices, or use stochastic perturbations in the optimization algorithms to avoid to be trapped in

817   local solutions. These aspects of convergence and sensitivity to initial conditions have so far been

818   poorly addressed. It is therefore necessary to check which method is robust operationally.

819     The last remaining challenge concerns the estimation of other statistical parameters of the state-

820   space model given in Eqs. (1) and (2) and associated filters. Indeed, the initial conditions $\mathbf{x}(0)$ and

821   $\mathbf{P}(0)$ are crucial for certain satellite retrieval problems and have to be estimated. This is the case,

822   for instance, when the time sequence of observations is short (i.e., shorter than the spinup time

43

of the filter with an uninformative prior) or when filtering and smoothing are repeated on various iterations, as in the EM algorithm. Estimation methods should also consider the estimation of systematic or time-varying biases, the deterministic part of $\eta$ and $\epsilon$. This was initially proposed by Dee et al. (1999a) and tested in Dee et al. (1999b) in the case of maximizing the innovation likelihood, in Dee (2005) in a state augmentation formulation, and was adapted to a Bayesian update formulation in Liu et al. (2017) and in Berry and Harlim (2017). Recently, the joint estimation of bias and covariance error terms, for the treatment of brightness temperatures from the European geostationary satellite, has been successfully applied in Merchant et al. (2020).

## c. *Perspectives for geophysical DA*

Beyond the aforementioned potential improvements in the existing techniques, specific research directions need to be taken by the data assimilation community. The main one concerns the realization of a comprehensive numerical evaluation of the different methods for the estimation of $\mathbf{Q}$ and $\mathbf{R}$, built on an agreed experimental framework and a consensus model. Such an effort would help to evaluate (i) the pros and cons of the different methods (including their capability to deal with high dimensionality, localization in ensemble methods, and their practical feasibility), (ii) their effects on different error statistics (RMSE, coverage probabilities, and other diagnostics), (iii) the potential combination of the various methods (especially those considering constant or adaptive covariances), and (iv) the capability to take into account other sources of error (due for instance to improper parameterizations, multiplicative errors, or forcing terms).

The use of a realistic DA problem, with a high-dimensional state-space and a limited and heterogeneous observational coverage should be addressed in the future. In that realistic case, the observational information per degree of freedom will be significantly lower, and the estimates of $\mathbf{Q}$ and $\mathbf{R}$ will deteriorate. Parametric versions of these error covariance matrices will therefore be

44

necessary. Among the parameters, some of them will control the variances, and will be different depending on the variable. Other parameters will control the spatial correlation lengths, that could be isotropic or anisotropic, depending on the region of interest and the considered variable. Cross-correlations between variables will also have to be considered. Consequently, $\mathbf{Q}$ and $\mathbf{R}$ will be block-matrices with as few parameters as possible.

A further challenge for future work is the evaluation of the feasibility of estimating non-additive, non-Gaussian, and time-correlated noises under the current estimation frameworks. In this way, the need for observational constraints for the stochastic perturbation methods in the NWP community could be considered within the estimation framework discussed in this review.

45

APPENDIX

**Four main algorithms to jointly estimate Q and R in data assimilation**

- initialize inflation factor (for instance $\lambda(1) = 1$);

**for** $k$ in 1:$K$ **do**

    **for** $i$ in 1:$N_e$ **do**

        - compute forecast $\mathbf{x}_i^f(k)$ using Eq. (4a);

        - compute innovation $\mathbf{d}_i(k)$ using Eq. (4b);

    **end**

    - compute empirical covariance $\tilde{\mathbf{P}}^f(k)$ of the $\mathbf{x}_i^f(k)$;

    - compute $\mathbf{K}^f(k)$ using Eq. (4c) where $\tilde{\mathbf{P}}^f(k)\mathscr{H}_k^{\mathrm{T}}$ and $\mathscr{H}_k\tilde{\mathbf{P}}^f(k)\mathscr{H}_k^{\mathrm{T}}$ are inflated by

     $\lambda(k)$;

    **for** $i$ in 1:$N_e$ **do**

        - compute analysis $\mathbf{x}_i^a(k)$ using Eq. (4d);

    **end**

    - compute mean innovations $\mathbf{d}^{o-f}(k)$ and $\mathbf{d}^{o-a}(k)$ with $\mathbf{d}_i^{o-f}(k) = \mathbf{y}(k) - \mathscr{H}_k(\mathbf{x}_i^f(k))$

     and $\mathbf{d}_i^{o-a}(k) = \mathbf{y}(k) - \mathscr{H}_k(\mathbf{x}_i^a(k))$;

    - update $\mathbf{R}(k)$ from Eq. (6b) using the cross-covariance between $\mathbf{d}_i^{o-f}(k)$ and $\mathbf{d}_i^{o-a}(k)$;

    - estimate $\tilde{\lambda}(k)$ using Eq. (8) where $\mathscr{H}_k\tilde{\mathbf{P}}^f(k)\mathscr{H}_k^{\mathrm{T}}$ is inflated by $\lambda(k)$;

    - update $\lambda(k+1)$ using temporal smoother;

**end**

**Algorithm 1:** Adaptive algorithm for the EnKF (Miyoshi et al. 2013)

- initialize $\mathbf{Q}(1)$ and $\mathbf{R}(1)$;

**for** <u>$k$ in 1:$K$</u> **do**

    **for** <u>$i$ in 1:$N_e$</u> **do**

        - compute forecast $\mathbf{x}_i^f(k)$ using Eq. (4a);

        - compute innovation $\mathbf{d}_i(k)$ using Eq. (4b);

    **end**

    - compute $\mathbf{K}^f(k)$ using Eq. (4c);

    **for** <u>$i$ in 1:$N_e$</u> **do**

        - compute analysis $\mathbf{x}_i^a(k)$ using Eq. (4d);

    **end**

    - apply Eq. (12a) to get $\tilde{\mathbf{P}}(k)$ using linearizations of $\mathbf{M}_k$ and $\mathbf{H}_k$ given in Eqs. (5a) and

     (5b);

    - estimate $\tilde{\mathbf{Q}}(k)$ using Eq. (12b);

    - estimate $\tilde{\mathbf{R}}(k)$ using Eq. (12c);

    - update $\mathbf{Q}(k+1)$ and $\mathbf{R}(k+1)$ using temporal smoothers;

**end**

**Algorithm 2:** Adaptive algorithm for the EnKF (Berry and Sauer 2013)

47

- define a priori distributions for $\boldsymbol{\theta}$ (shape parameters of $\mathbf{Q}$ and $\mathbf{R}$);

**for** $\underline{k \text{ in } 1{:}K}$ **do**

    **for** $\underline{i \text{ in } 1{:}N_e}$ **do**

        - draw samples $\boldsymbol{\theta}_i(k)$ from $p\left(\boldsymbol{\theta}|\mathbf{y}(1:k-1)\right)$;

        - compute forecast $\mathbf{x}_i^f(k)$ using Eq. (4a) with $\boldsymbol{\theta}_i(k)$;

        - compute innovation $\mathbf{d}_i(k)$ using Eq. (4b) with $\boldsymbol{\theta}_i(k)$;

    **end**

    - compute $\mathbf{K}^f(k)$ using Eq. (4c);

    **for** $\underline{i \text{ in } 1{:}N_e}$ **do**

        - compute analysis $\mathbf{x}_i^a(k)$ using Eq. (4d);

    **end**

    - approximate Gaussian likelihood of innovations $p\left(\mathbf{y}(k)|\mathbf{y}(1:k-1),\boldsymbol{\theta}(k)\right)$ using

    empirical mean $\bar{\mathbf{d}}(k) = \frac{1}{N_e}\sum_{i=1}^{N_e}\mathbf{d}_i(k)$ and empirical covariance

    $\Sigma(k) = \frac{1}{N_e-1}\sum_{i=1}^{N_e}\left(\mathbf{d}_i(k)-\bar{\mathbf{d}}(k)\right)\left(\mathbf{d}_i(k)-\bar{\mathbf{d}}(k)\right)^{\mathrm{T}}$ with $\mathbf{d}_i(k) = \mathbf{y}(k) - \mathscr{H}_k(\mathbf{x}_i^f(k))$;

    - update $p\left(\boldsymbol{\theta}|\mathbf{y}(1:k)\right)$ using Eq. (16);

**end**

**Algorithm 3:** Adaptive algorithm for the EnKF (Stroud et al. 2018)

48

**while** $\underline{p\left(\mathbf{y}(1:K)|\boldsymbol{\theta}_{(n)}\right) - p\left(\mathbf{y}(1:K)|\boldsymbol{\theta}_{(n-1)}\right) > \varepsilon}$ **do**

    **for** $\underline{k}$ in 1:$K$ **do**

        **for** $\underline{i}$ in 1:$N_e$ **do**

            - compute forecast $\mathbf{x}_i^f(k)$ using Eq. (4a);

            - compute innovation $\mathbf{d}_i(k)$ using Eq. (4b);

        **end**

        - compute $\mathbf{K}^f(k)$ using Eq. (4c);

        **for** $\underline{i}$ in 1:$N_e$ **do**

            - compute analysis $\mathbf{x}_i^a(k)$ using Eq. (4d);

        **end**

    **end**

    **for** $\underline{k}$ in $K$:1 **do**

        - compute $\mathbf{K}^s(k)$ using Eq. (4e);

        **for** $\underline{i}$ in 1:$N_e$ **do**

            - compute reanalysis $\mathbf{x}_i^s(k)$ using Eq. (4f);

        **end**

    **end**

    - increment $n \leftarrow n+1$;

    - estimate $\mathbf{Q}_{(n)}$ using Eq. (21a);

    - estimate $\mathbf{R}_{(n)}$ using Eq. (21b);

**end**

**Algorithm 4:** EM algorithm for the EnKF/EnKS (Dreano et al. 2017)

## References

Anderson, J. L., 2007: An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus A: Dynamic Meteorology and Oceanography*, **59 (2)**, 210–224.

Anderson, J. L., 2009: Spatially and temporally varying adaptive covariance inflation for ensemble filters. *Tellus, Series A: Dynamic Meteorology and Oceanography*, **61 (1)**, 72–83.

Anderson, J. L., and S. L. Anderson, 1999: A Monte Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts. *Monthly Weather Review*, **12 (127)**, 2741–2758.

Bélanger, P. R., 1974: Estimation of noise covariance matrices for a linear time-varying stochastic process. *Automatica*, **10 (3)**, 267–275.

Berry, T., and J. Harlim, 2017: Correcting Biased Observation Model Error in Data Assimilation. *Monthly Weather Review*, **145 (7)**, 2833–2853.

Berry, T., and T. Sauer, 2013: Adaptive ensemble Kalman filtering of non-linear systems. *Tellus, Series A: Dynamic Meteorology and Oceanography*, **65 (20331)**, 1–16.

Berry, T., and T. Sauer, 2018: Correlation between system and observation errors in data assimilation. *Monthly Weather Review*, **146 (9)**, 2913–2931.

Bishop, C. H., and E. A. Satterfield, 2013: Hidden error variance theory. Part I: Exposition and analytic model. *Monthly Weather Review*, **141 (5)**, 1454–1468.

Bishop, C. H., E. A. Satterfield, and K. T. Shanley, 2013: Hidden error variance theory. Part II: An instrument that reveals hidden error variance distributions from ensemble forecasts and observations. *Monthly Weather Review*, **141 (5)**, 1469–1483.

Blanchet, I., C. Frankignoul, and M. A. Cane, 1997: A comparison of adaptive kalman filters for a tropical pacific ocean model. *Monthly Weather Review*, **125 (1)**, 40–58.

Bocquet, M., 2011: Ensemble Kalman filtering without the intrinsic need for inflation. *Nonlinear Processes in Geophysics*, **18 (5)**, 735–750.

Bocquet, M., P. N. Raanes, and A. Hannart, 2015: Expanding the validity of the ensemble Kalman filter without the intrinsic need for inflation. *Nonlinear Processes in Geophysics*, **22**, 645–662.

Bocquet, M., and P. Sakov, 2012: Combining inflation-free and iterative ensemble Kalman filters for strongly nonlinear systems. *Nonlinear Processes in Geophysics*, **19**, 383–399.

Bormann, N., A. Collard, and P. Bauer, 2010: Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. II: Application to AIRS and IASI data. *Quarterly Journal of the Royal Meteorological Society*, **136 (649)**, 1051–1063.

Bowler, N. E., 2017: On the diagnosis of model error statistics using weak-constraint data assimilation. *Quarterly Journal of the Royal Meteorological Society*, **143 (705)**, 1916–1928.

Brankart, J.-M., E. Cosme, C.-E. Testut, P. Brasseur, and J. Verron, 2010: Efficient adaptive error parameterizations for square root or ensemble Kalman filters: Application to the control of ocean mesoscale signals. *Monthly Weather Review*, **138 (3)**, 932–950.

Buehner, M., 2010: Error statistics in data assimilation: Estimation and modelling. *Data Assimilation*, Springer, 93–112.

Campbell, W. F., E. A. Satterfield, B. Ruston, and N. L. Baker, 2017: Accounting for correlated observation error in a dual-formulation 4D variational data assimilation system. *Monthly Weather Review*, **145 (3)**, 1019–1032.

51

Cappé, O., 2011: Online Expectation-Maximisation. *Mixtures: Estimation and Applications*, Wiley Series in Probability and Statistics, 1–53.

Carrassi, A., M. Bocquet, L. Bertino, and G. Evensen, 2018: Data Assimilation in the Geosciences: An overview on methods, issues and perspectives. *WIREs Clim Change*, **9 (5)**, e535.

Chapnik, B., G. Desroziers, F. Rabier, and O. Talagrand, 2004: Properties and first application of an error-statistics tuning method in variational assimilation. *Quarterly Journal of the Royal Meteorological Society*, **130 (601)**, 2253–2275.

Cocucci, T. J., M. Pulido, M. Lucini, and P. Tandeo, 2020: Model error covariance estimation in particle and ensemble Kalman filters using an online expectation-maximization algorithm. *arXiv preprint arXiv:2003.02109*.

Corazza, M., E. Kalnay, D. J. Patil, R. Morss, M. Cai, I. Szunyogh, B. R. Hunt, and J. A. Yorke, 2003: Use of the breeding technique to estimate the structure of the analysis "errors of the day". *Nonlinear Processes in Geophysics*, **10 (3)**, 233–243.

Daescu, D. N., and R. H. Langland, 2013: Error covariance sensitivity and impact estimation with adjoint 4D-Var: theoretical aspects and first applications to NAVDAS-AR. *Quarterly Journal of the Royal Meteorological Society*, **139**, 226–241.

Daescu, D. N., and R. Todling, 2010: Adjoint sensitivity of the model forecast to data assimilation system error covariance parameters. *Quarterly Journal of the Royal Meteorological Society*, **136**, 2000–2012.

Daley, R., 1991: Atmospheric data analysis. Cambridge University Press, 457 pp.

Daley, R., 1992: Estimating Model-Error Covariances for Application to Atmospheric Data Assimilation. *Monthly Weather Review*, **120 (8)**, 1735–1746.

52

Dee, D. P., 1995: On-line estimation of error covariance parameters for atmospheric data assimilation. *Monthly Weather Review*, **123 (4)**, 1128–1145.

Dee, D. P., 2005: Bias and data assimilation. *Quarterly Journal of the Royal Meteorological Society*, **131 (613)**, 3323–3343.

Dee, D. P., S. E. Cohn, A. Dalcher, and M. Ghil, 1985: An efficient algorithm for estimating noise covariances in distributed systems. *IEEE Transactions on Automatic Control*, **30 (11)**, 1057–1065.

Dee, D. P., G. Gaspari, C. Redder, L. Rukhovets, and A. M. da Silva, 1999a: Maximum-likelihood estimation of forecast and observation error covariance parameters. Part I: Methodology. *Monthly Weather Review*, **127 (1992)**, 1822–1834.

Dee, D. P., G. Gaspari, C. Redder, L. Rukhovets, and A. M. da Silva, 1999b: Maximum-likelihood estimation of forecast and observation error covariance parameters. Part II: Applications. *Monthly Weather Review*, **127 (1992)**, 1835–1849.

Delsole, T., and X. Yang, 2010: State and parameter estimation in stochastic dynamical models. *Physica D: Nonlinear Phenomena*, **239 (18)**, 1781–1788.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977: Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39 (1)**, 1–38.

Desroziers, G., L. Berre, B. Chapnik, and P. Poli, 2005: Diagnosis of observation, background and analysis-error statistics in observation space. *Quarterly Journal of the Royal Meteorological Society*, **131 (613)**, 3385–3396.

Desroziers, G., and S. Ivanov, 2001: Diagnosis and adaptive tuning of observation-error parameters in a variational assimilation. *Quarterly Journal of the Royal Meteorological Society*, **127 (574)**, 1433–1452.

Dreano, D., P. Tandeo, M. Pulido, T. Chonavel, B. AIt-El-Fquih, and I. Hoteit, 2017: Estimating model error covariances in nonlinear state-space models using Kalman smoothing and the expectation-maximisation algorithm. *Quarterly Journal of the Royal Meteorological Society*, **143 (705)**, 1877–1885.

Duník, J., O. Straka, O. Kost, and J. Havlík, 2017: Noise covariance matrices in state-space models: A survey and comparison of estimation methods-Part I. *International Journal of Adaptive Control and Signal Processing*, **31 (11)**, 1505–1543.

El Gharamti, M., 2018: Enhanced Adaptive Inflation Algorithm for Ensemble Filters. *Monthly Weather Review*, **146**, 623–640.

Evensen, G., 2009: *Data assimilation: the ensemble Kalman filter*. Springer Science & Business Media.

Frei, M., and H. R. Künsch, 2012: Sequential State and Observation Noise Covariance Estimation Using Combined Ensemble Kalman and Particle Filters. *Monthly Weather Review*, **140 (5)**, 1476–1495.

Fu, L.-L., I. Fukumori, and R. N. Miller, 1993: Fitting dynamic models to the Geosat sea level observations in the tropical Pacific ocean. Part II: A linear, wind-driven model. *Journal of Physical Oceanography*, **23 (10)**, 2162–2181.

Ghil, M., and P. Malanotte-Rizzoli, 1991: Data assimilation in meteorology and oceanography.pdf. *Advances in Geophysics*, **33**, 141–266.

54

Guillet, O., A. T. Weaver, X. Vasseur, Y. Michel, S. Gratton, and S. Gürol, 2019: Modelling spatially correlated observation errors in variational data assimilation using a diffusion operator on an unstructured mesh. *Quarterly Journal of the Royal Meteorological Society*, doi:10.1002/qj.3537.

Harlim, J., 2018: Ensemble Kalman Filters. *Data-Driven Computational Methods*, Cambridge university press, 31–59.

Harlim, J., A. Mahdi, and A. J. Majda, 2014: An ensemble Kalman filter for statistical estimation of physics constrained nonlinear regression models. *Journal of Computational Physics*, **257**, 782–812.

Hollingsworth, A., and P. Lönnberg, 1986: The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: The wind field. *Tellus A*, **38 (2)**, 111–136.

Hotta, D., E. Kalnay, Y. Ota, and T. Miyoshi, 2017: EFSR: Ensemble forecast sensitivity to observation error covariance. *Monthly Weather Review*, **145**, 5015–5031.

Houtekamer, P. L., H. L. Mitchell, and X. Deng, 2009: Model Error Representation in an Operational Ensemble Kalman Filter. *Monthly Weather Review*, **137 (7)**, 2126–2143.

Houtekamer, P. L., and F. Zhang, 2016: Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation. *Monthly Weather Review*, **144 (12)**, 4489–4532.

Ide, K., P. Courtier, M. Ghil, and A. C. Lorenc, 1997: Unified Notation for Data Assimilation: Operational, Sequential and Variational. *Journal of the Meteorological Society of Japan*, **75 (1)**, 181–189.

Janjić, T., and Coauthors, 2018: On the representation error in data assimilation. *Quarterly Journal of the Royal Meteorological Society*, **144 (713)**, 1257–1278.

55

Jazwinski, A. H., 1970: *Stochastic processes and filtering theory*. Academic Press.

Kantas, N., A. Doucet, S. S. Singh, J. Maciejowski, and N. Chopin, 2015: On particle methods for parameter estimation in state-space models. *Statistical Science*, **30 (3)**, 328–351.

Katzfuss, M., J. R. Stroud, and C. K. Wikle, 2019: Ensemble Kalman methods for high-dimensional hierarchical dynamic space-time models. *Journal of the American Statistical Association*, 1–43.

Kotsuki, S., T. Miyoshi, K. Terasaki, G.-Y. Lien, and E. Kalnay, 2017a: Assimilating the global satellite mapping of precipitation data with the Nonhydrostatic Icosahedral Atmospheric Model (NICAM). *Journal of Geophysical Research: Atmospheres*, **122 (2)**, 631–650.

Kotsuki, S., Y. Ota, and T. Miyoshi, 2017b: Adaptive covariance relaxation methods for ensemble data assimilation: Experiments in the real atmosphere. *Quarterly Journal of the Royal Meteorological Society*, **143 (705)**, 2001–2015.

Leith, C. E., 1978: Objective methods for weather prediction. *Annual Review of Fluid Mechanics*, **10 (1)**, 107–128.

Li, H., E. Kalnay, and T. Miyoshi, 2009: Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter. *Quarterly Journal of the Royal Meteorological Society*, **135 (2)**, 523–533.

Liang, X., X. Zheng, S. Zhang, G. Wu, Y. Dai, and Y. Li, 2012: Maximum likelihood estimation of inflation factors on error covariance matrices for ensemble Kalman filter assimilation. *Quarterly Journal of the Royal Meteorological Society*, **138 (662)**, 263–273.

Liu, J., and M. West, 2001: Combined parameter and state estimation in simulation-based filtering. *Sequential Monte Carlo methods in practice*, Springer, 197–223.

56

Liu, Y., J.-M. Haussaire, M. Bocquet, Y. Roustan, O. Saunier, and A. Mathieu, 2017: Uncertainty quantification of pollutant source retrieval: comparison of Bayesian methods with application to the Chernobyl and Fukushima Daiichi accidental releases of radionuclides. *Quarterly Journal of the Royal Meteorological Society*, **143 (708)**, 2886–2901.

Mehra, R. K., 1970: On the identification of variances and adaptive Kalman filtering. *IEEE Transactions on Automatic Control*, **AC-15 (2)**, 175–184.

Mehra, R. K., 1972: Approaches to adaptive filtering. *IEEE Transactions on Automatic Control*, **17 (5)**, 693–698.

Ménard, R., 2016: Error covariance estimation methods based on analysis residuals: theoretical foundation and convergence properties derived from simplified observation networks. *Quarterly Journal of the Royal Meteorological Society*, **142 (694)**, 257–273.

Menemenlis, D., and M. Chechelnitsky, 2000: Error estimates for an ocean general circulation model from altimeter and acoustic tomography data. *Monthly Weather Review*, **128 (3)**, 763–778.

Ménétrier, B., and T. Auligné, 2015: Optimized localization and hybridization to filter ensemble-based covariances. *Monthly Weather Review*, **143**, 3931–3947.

Merchant, C. J., S. Saux-Picart, and J. Waller, 2020: Bias correction and covariance parameters for optimal estimation by exploiting matched in-situ references. *Remote Sensing of Environment*, **237**, 111 590.

Mitchell, H. L., and P. L. Houtekamer, 2000: An adaptive ensemble Kalman filter. *Monthly Weather Review*, **128 (2)**, 416–433.

Mitchell, L., and A. Carrassi, 2015: Accounting for model error due to unresolved scales within ensemble Kalman filtering. *Quarterly Journal of the Royal Meteorological Society*, **141 (689)**, 1417–1428.

Miyoshi, T., 2011: The Gaussian Approach to Adaptive Covariance Inflation and Its Implementation with the Local Ensemble Transform Kalman Filter. *Monthly Weather Review*, **139 (5)**, 1519–1535.

Miyoshi, T., E. Kalnay, and H. Li, 2013: Estimating and including observation-error correlations in data assimilation. *Inverse Problems in Science and Engineering*, **21 (3)**, 387–398.

Miyoshi, T., Y. Sato, and T. Kadowaki, 2010: Ensemble Kalman filter and 4D-Var intercomparison with the Japanese operational global analysis and prediction system. *Monthly Weather Review*, **138 (7)**, 2846–2866.

Pham, D. T., J. Verron, and M. C. Roubaud, 1998: A singular evolutive extended Kalman filter for data assimilation in oceanography. *Journal of Marine systems*, **16 (3-4)**, 323–340.

Pulido, M., P. Tandeo, M. Bocquet, A. Carrassi, and M. Lucini, 2018: Stochastic parameterization identification using ensemble Kalman filtering combined with maximum likelihood methods. *Tellus A: Dynamic Meteorology and Oceanography*, **70 (1)**, 1442 099.

Purser, R. J., and D. F. Parrish, 2003: A Bayesian technique for estimating continuously varying statistical parameters of a variational assimilation. *Meteorology and Atmospheric Physics*, **82 (1-4)**, 209–226.

Raanes, P. N., M. Bocquet, and A. Carrassi, 2019: Adaptive covariance inflation in the ensemble Kalman filter by Gaussian scale mixtures. *Quarterly Journal of the Royal Meteorological Society*, **145 (718)**, 53–75.

58

Rodwell, M. J., and T. N. Palmer, 2007: Using numerical weather prediction to assess climate models. *Quarterly Journal of the Royal Meteorological Society*, **133 (622)**, 129–146.

Ruiz, J. J., M. Pulido, and T. Miyoshi, 2013a: Estimating model parameters with ensemble-based data assimilation: A review. *Journal of the Meteorological Society of Japan*, **91**, 79–99.

Ruiz, J. J., M. Pulido, and T. Miyoshi, 2013b: Estimating model parameters with ensemble-based data assimilation: Parameter Covariance Treatment. *Journal of the Meteorological Society of Japan*, **91**, 453–469.

Rutherford, I. D., 1972: Data assimilation by statistical interpolation of forecast error fields. *Journal of the Atmospheric Sciences*, **29 (5)**, 809–815.

Satterfield, E. A., D. Hodyss, D. D. Kuhl, and C. H. Bishop, 2018: Observation-informed generalized hybrid error covariance models. *Monthly Weather Review*, **146**, 3605–3622.

Scheffler, G., J. Ruiz, and M. Pulido, 2019: Inference of stochastic parametrizations for model error treatment using nested ensemble Kalman filters. *Quarterly Journal of the Royal Meteorological Society*, doi:https://doi.org/10.1002/qj.3542.

Schmidt, S. F., 1966: Applications of state space methods to navigation problems. *Advances in Control Systems*, **3**, 293–340.

Shumway, R. H., and D. S. Stoffer, 1982: An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, **3 (4)**, 253–264.

Solonen, A., J. Hakkarainen, A. Ilin, M. Abbas, and A. Bibov, 2014: Estimating model error covariance matrix parameters in extended Kalman filtering. *Nonlinear Processes in Geophysics*, **21 (5)**, 919–927.

59

Stroud, J. R., and T. Bengtsson, 2007: Sequential state and variance estimation within the ensemble Kalman filter. *Monthly Weather Review*, **135 (9)**, 3194–3208.

Stroud, J. R., M. Katzfuss, and C. K. Wikle, 2018: A Bayesian adaptive ensemble Kalman filter for sequential state and parameter estimation. *Monthly Weather Review*, **146 (1)**, 373–386.

Tandeo, P., M. Pulido, and F. Lott, 2015: Offline parameter estimation using EnKF and maximum likelihood error covariance estimates: Application to a subgrid-scale orography parametrization. *Quarterly Journal of the Royal Meteorological Society*, **141 (687)**, 383–395.

Todling, R., 2015: A lag-1 smoother approach to system-error estimation: sequential method. *Quarterly Journal of the Royal Meteorological Society*, **141 (690)**, 1502–1513.

Ueno, G., T. Higuchi, T. Kagimoto, and N. Hirose, 2010: Maximum likelihood estimation of error covariances in ensemble-based filters and its application to a coupled atmosphere-ocean model. *Quarterly Journal of the Royal Meteorological Society*, **136 (650)**, 1316–1343.

Ueno, G., and N. Nakamura, 2014: Iterative algorithm for maximum-likelihood estimation of the observation-error covariance matrix for ensemble-based filters. *Quarterly Journal of the Royal Meteorological Society*, **140 (678)**, 295–315.

Ueno, G., and N. Nakamura, 2016: Bayesian estimation of the observation-error covariance matrix in ensemble-based filters. *Quarterly Journal of the Royal Meteorological Society*, **142 (698)**, 2055–2080.

Wahba, G., and J. Wendelberger, 1980: Some new mathematical methods for variational objective analysis using splines and cross validation. *Monthly weather review*, **108 (8)**, 1122–1143.

60

Waller, J. A., S. L. Dance, and N. K. Nichols, 2016: Theoretical insight into diagnosing observation error correlations using observation-minus-background and observation-minus-analysis statistics. *Quarterly Journal of the Royal Meteorological Society*, **142 (694)**, 418–431.

Waller, J. A., S. L. Dance, and N. K. Nichols, 2017: On diagnosing observation-error statistics with local ensemble data assimilation. *Quarterly Journal of the Royal Meteorological Society*, **143 (708)**, 2677–2686.

Wang, X., and C. H. Bishop, 2003: A Comparison of Breeding and Ensemble Transform Kalman Filter Ensemble Forecast Schemes. *Journal of the Atmospheric Sciences*, **60 (9)**, 1140–1158.

Weston, P. P., W. Bell, and J. R. Eyre, 2014: Accounting for correlated error in the assimilation of high-resolution sounder data. *Quarterly Journal of the Royal Meteorological Society*, **140 (685)**, 2420–2429.

Whitaker, J. S., and T. M. Hamill, 2012: Evaluating methods to account for system errors in ensemble data assimilation. *Monthly Weather Review*, **140 (9)**, 3078–3089.

Whitaker, J. S., T. M. Hamill, X. Wei, Y. Song, and Z. Toth, 2008: Ensemble data assimilation with the NCEP global forecast system. *Monthly Weather Review*, **136 (2)**, 463–482.

Winiarek, V., M. Bocquet, N. Duhanyan, Y. Roustan, O. Saunier, and A. Mathieu, 2014: Estimation of the caesium-137 source term from the Fukushima Daiichi nuclear power plant using a consistent joint assimilation of air concentration and deposition observations. *Atmospheric Environment*, **82**, 268–279.

Winiarek, V., M. Bocquet, O. Saunier, and A. Mathieu, 2012: Estimation of errors in the inverse modeling of accidental release of atmospheric pollutant: Application to the reconstruction of

the cesium-137 and iodine-131 source terms from the Fukushima Daiichi power plant. *Journal of Geophysical Research: Atmospheres*, **117 (D5)**.

Wu, C. F. J., 1983: On the convergence properties of the EM algorithm. *Annals of Statistics*, **11 (1)**, 95–103.

Yang, Y., and E. Mémin, 2019: Estimation of physical parameters under location uncertainty using an ensemble-expectation-maximization algorithm. *Quarterly Journal of the Royal Meteorological Society*, **145 (719)**, 418–433.

Ying, Y., and F. Zhang, 2015: An adaptive covariance relaxation method for ensemble data assimilation. *Quarterly Journal of the Royal Meteorological Society*, **141 (692)**, 2898–2906.

Zebiak, S. E., and M. A. Cane, 1987: A model El Nino–southern oscillation. *Monthly Weather Review*, **115 (10)**, 2262–2278.

Zhang, F., C. Snyder, and J. Sun, 2004: Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter. *Monthly Weather Review*, **132 (5)**, 1238–1253.

Zhen, Y., and J. Harlim, 2015: Adaptive error covariances estimation methods for ensemble Kalman filters. *Journal of Computational Physics*, **294**, 619–638.

63

TABLE 1. Comparison of several methods to estimate error covariance matrices **Q** and **R** in data assimilation.

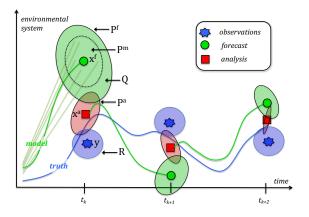| Estimation method | Criteria | Estimation of covariance **Q** | Suitable for non-Gaussian errors | Application to the highest complexity model |
|---|---|---|---|---|
| Method of moments | Innovation statistics in the observation space | No (inflation of $\mathbf{P}^f$ instead) | No | NWP |
| Method of moments | Lag innovation between consecutive times | Yes | No | Lorenz-96 |
| Likelihood methods | Bayesian update of the posterior distribution | No (or joint parameter with **R**) | Yes (using particle filters, not EnKF) | Shallow water |
| Likelihood methods | Maximization of the total likelihood | Yes | Yes (using particle filters, not EnKF) | Two-scale Lorenz-96 |

65

1162   FIG. 1. Sketch of sequential and ensemble data assimilation algorithms in the observation space (i.e., in the

1163   space of the observations **y**), where the observation operator $\mathscr{H}$ is omitted for simplicity. The ellipses represent

1164   the forecast $\mathbf{P}^f$ and analysis $\mathbf{P}^a$ error covariances, while the model $\mathbf{Q}$ and observation $\mathbf{R}$ error covariances are the

1165   unknown entries of the state-space model in Eqs. (1) and (2). The forecast error covariance matrix is written $\mathbf{P}^f$

1166   and is the sum of $\mathbf{P}^m$, the forecasted state $\mathbf{x}^f$ spread, and the model error $\mathbf{Q}$. This scheme is a modified version

1167   based on Fig. 1 from Carrassi et al. (2018).

66
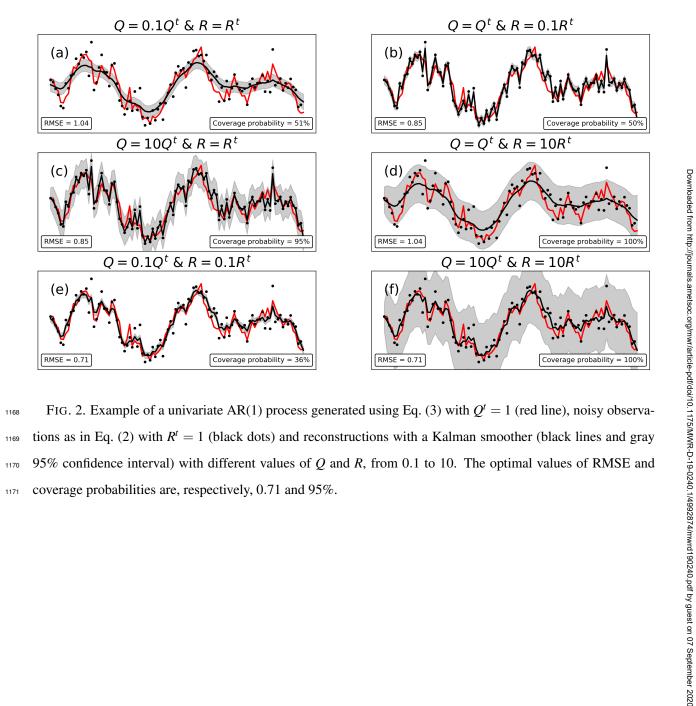
FIG. 2. Example of a univariate AR(1) process generated using Eq. (3) with $Q^t = 1$ (red line), noisy observations as in Eq. (2) with $R^t = 1$ (black dots) and reconstructions with a Kalman smoother (black lines and gray 95% confidence interval) with different values of $Q$ and $R$, from 0.1 to 10. The optimal values of RMSE and coverage probabilities are, respectively, 0.71 and 95%.
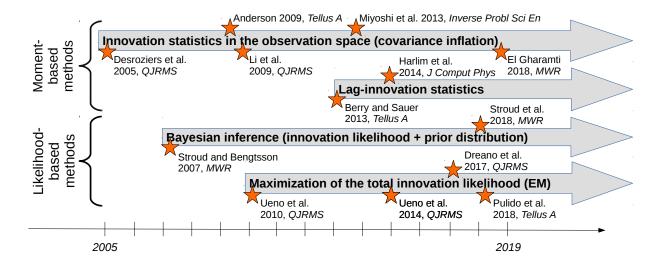
67

1172   FIG. 3. Timeline of the main methods used in geophysical data assimilation for the joint estimation of **Q** and

1173   **R** over the last 15 years. Dee (1995) and Desroziers and Ivanov (2001) are not represented here but are certainly

1174   the seminal work of this research field in data assimilation.

68