Chapter 6
The ensemble Kalman filter

In Chapter 3 our goal was to track the truth of the partially observed dynamical system,

$$\begin{cases} \mathbf{x}_{k+1} = \mathcal{M}_{k+1}(\mathbf{x}_k) + \boldsymbol{\epsilon}_k^{\mathrm{q}}, \\ \mathbf{y}_k = \mathcal{H}_k(\mathbf{x}_k) + \boldsymbol{\epsilon}_k^{\mathrm{o}}, \end{cases} \tag{6.1}$$

where $\mathbf{x}_k \in \mathbb{R}^n$ is the true state of the system at time t_k. The first equation is the forecast step. It simulates the evolution of the model from t_k to t_{k+1} using the model propagator $\mathcal{M}_{k+1}$. The term $\boldsymbol{\epsilon}_k^{\mathrm{q}}$ is the model error of $\mathcal{M}_{k+1}$ when compared to the true physical processes. It is assumed that this noise is unbiased, uncorrelated in time (white noise), and of model error covariance matrix $\mathbf{Q}_k \in \mathbb{R}^{n\times n}$, i.e.,

$$\mathrm{E}[\boldsymbol{\epsilon}_k^{\mathrm{q}}] = 0 \qquad \text{and} \qquad \mathrm{E}[\boldsymbol{\epsilon}_k^{\mathrm{q}}(\boldsymbol{\epsilon}_l^{\mathrm{q}})^{\mathrm{T}}] = \mathbf{Q}_k \delta_{kl}.$$

The second equation in (6.1) is the observation equation, where $\mathcal{H}_k$ is the observation operator at time t_k. The term $\boldsymbol{\epsilon}_k^{\mathrm{o}} \in \mathbb{R}^p$ is the observation error, which is assumed to be unbiased, uncorrelated in time (white noise), and of error covariance matrix $\mathbf{R}_k \in \mathbb{R}^{p\times p}$, i.e.,

$$\mathrm{E}[\boldsymbol{\epsilon}_k^{\mathrm{o}}] = 0 \qquad \text{and} \qquad \mathrm{E}\left[\boldsymbol{\epsilon}_k^{\mathrm{o}}(\boldsymbol{\epsilon}_l^{\mathrm{o}})^{\mathrm{T}}\right] = \mathbf{R}_k \delta_{kl}.$$

As an additional assumption, though quite a realistic one, we suppose that there is no correlation between model error and observation error. As a consequence,

$$\mathrm{E}\left[\boldsymbol{\epsilon}_k^{\mathrm{o}}(\boldsymbol{\epsilon}_l^{\mathrm{q}})^{\mathrm{T}}\right] = 0.$$

To track the truth, we used the extended Kalman filter (EKF) that was introduced in Section 3.6. For the clarity of this chapter, we begin by presenting the algorithm of the EKF in Algorithm 6.1.

There are two major drawbacks to the EKF. First of all, the EKF is numerically scarcely affordable in high-dimensional systems. One has to manipulate the error covariance matrix, which requires $n(n+1)/2$ scalars to be stored. Clearly this is impossible in high-dimensional systems for which the storage of a few state vectors is already challenging. Moreover, during the forecast step from t_k to t_{k+1}, one has to compute a forecast error covariance matrix (3.19),

$$\mathbf{P}_{k+1}^{\mathrm{f}} = \mathbf{M}_{k+1}\mathbf{P}_k^{\mathrm{a}}\mathbf{M}_{k+1}^{\mathrm{T}} + \mathbf{Q}_k,$$

Algorithm 6.1 Algorithm of the EKF

Require: For $k = 0, \ldots, K$: the observation error covariance matrices $\mathbf{R}_k$, the model error covariance matrices $\mathbf{Q}_k$, the observation models $\mathcal{H}_k$ and their tangent linear $\mathbf{H}_k$, the forward models $\mathcal{M}_k$ and their tangent linear $\mathbf{M}_k$.

1: Initialize system state $\mathbf{x}_0^{\mathrm{f}}$ and error covariance matrix $\mathbf{P}_0^{\mathrm{f}}$.
2: **for** $k = 0, \ldots, K$ **do**
3: Compute the gain: $\mathbf{K}_k = \mathbf{P}_k^{\mathrm{f}} \mathbf{H}_k^{\mathrm{T}} \left(\mathbf{H}_k \mathbf{P}_k^{\mathrm{f}} \mathbf{H}_k^{\mathrm{T}} + \mathbf{R}_k \right)^{-1}$.
4: Compute the state analysis: $\mathbf{x}_k^{\mathrm{a}} = \mathbf{x}_k^{\mathrm{f}} + \mathbf{K}_k \left(\mathbf{y}_k - \mathcal{H}_k[\mathbf{x}_k^{\mathrm{f}}] \right)$.
5: Compute the analysis error covariance matrix: $\mathbf{P}_k^{\mathrm{a}} = (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k) \mathbf{P}_k^{\mathrm{f}}$.
6: Compute the forecast state: $\mathbf{x}_{k+1}^{\mathrm{f}} = \mathcal{M}_{k+1}[\mathbf{x}_k^{\mathrm{a}}]$.
7: Compute the forecast error covariance matrix: $\mathbf{P}_{k+1}^{\mathrm{f}} = \mathbf{M}_{k+1} \mathbf{P}_k^{\mathrm{a}} \mathbf{M}_{k+1}^{\mathrm{T}} + \mathbf{Q}_k$.
8: **end for**

where $\mathbf{M}_{k+1}$ is the tangent linear model (TLM) of $\mathcal{M}_{k+1}$. In addition to the derivation of the TLM, such a computation would require the use of the TLM $2n$ times (a left matrix multiplication by $\mathbf{M}_{k+1}$, followed by a right matrix multiplication by $\mathbf{M}_{k+1}^{\mathrm{T}}$). For a high-dimensional model, this is likely to be much too costly, even when resorting to parallel computing.

Second, the EKF is approximate for nonlinear systems, which is another major drawback. It relies on the tangent linear of the model to describe error propagation. When the tangent linear approximation is breached, for instance when the time step between two consecutive updates is long enough so that the nonlinearity of the model can develop, the forecast error covariance matrix may become underestimated, which is likely to lead to the divergence of the filter.

6.1 ▪ The reduced-rank square root filter

Following ideas put forward in Chapter 5, reduced-rank KFs [Verlaan and Heemink, 1997; Pham et al., 1998; Segers et al., 2000; Pham, 2001; Verlaan and Heemink, 2001] offer a solution to the major problem of the dimensionality. We shall focus on a specific variant, the reduced-rank square root filter, denoted RRSQRT. In this filter, the issue of the computation and propagation of the error covariance matrices is astutely circumvented. The error covariance matrices are represented by their principal axes (those with the largest eigenvalues), that is to say by a limited set of modes. The update and the forecast step will therefore apply to a limited number of modes, a collection of $m \ll n$ vectors for a state space of dimension n, rather than on huge matrices of size $n \times n$.

Suppose the initial system state is $\mathbf{x}_0^{\mathrm{f}}$, with an error covariance matrix $\mathbf{P}_0^{\mathrm{f}}$. We assume a decomposition in terms of the principal modes of $\mathbf{P}_0^{\mathrm{f}} \in \mathbb{R}^{n \times n}$: $\mathbf{P}_0^{\mathrm{f}} \simeq \mathbf{S}_0^{\mathrm{f}} (\mathbf{S}_0^{\mathrm{f}})^{\mathrm{T}}$, where $\mathbf{S}_0^{\mathrm{f}}$ is a matrix of size $n \times m$ composed of m n-vector columns that coincide with the first m dominant modes of $\mathbf{P}_0^{\mathrm{f}}$. The representation of the background has radically changed: rather than thinking in terms of $\mathbf{P}_0^{\mathrm{f}}$, we now think about its dominant modes, $\mathbf{S}_0^{\mathrm{f}}$.

Let us focus on an analysis at a given time t_k. That is why the time index, k, is dropped in the following. We also consider the transformed matrix $\mathbf{Y}_{\mathrm{f}} = \mathbf{H}\mathbf{S}_{\mathrm{f}}$ of size $p \times m$ ranging in the observation space. $\mathbf{H}$ is either the observation operator when it is linear, or its tangent linear otherwise. These matrices appear when considering the

Kalman gain needed in the analysis step (3.16),

$$\begin{aligned}\mathbf{K} &= \mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}}+\mathbf{R}\right)^{-1}\\ &= \mathbf{S}_{\mathrm{f}}\mathbf{S}_{\mathrm{f}}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{S}_{\mathrm{f}}\mathbf{S}_{\mathrm{f}}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}+\mathbf{R}\right)^{-1}\\ &= \mathbf{S}_{\mathrm{f}}(\mathbf{H}\mathbf{S}_{\mathrm{f}})^{\mathrm{T}}\left[(\mathbf{H}\mathbf{S}_{\mathrm{f}})(\mathbf{H}\mathbf{S}_{\mathrm{f}})^{\mathrm{T}}+\mathbf{R}\right]^{-1}.\end{aligned}$$

Hence, the Kalman gain, computed at the analysis step, is simply expressed with the help of the $\mathbf{S}_{\mathrm{f}}$ (or $\mathbf{Y}_{\mathrm{f}}$) matrix of size $p \times m$:

$$\mathbf{K} = \mathbf{S}_{\mathrm{f}}\mathbf{Y}_{\mathrm{f}}^{\mathrm{T}}\left(\mathbf{Y}_{\mathrm{f}}\mathbf{Y}_{\mathrm{f}}^{\mathrm{T}}+\mathbf{R}\right)^{-1}.$$

The analysis estimate, $\mathbf{x}^{\mathrm{a}}$, can be obtained from this gain and the standard update formula (3.21).

Then, what happens to the formula for the analysis error covariance matrix $\mathbf{P}^{\mathrm{a}}$? We have (3.22)

$$\begin{aligned}\mathbf{P}^{\mathrm{a}} &= (\mathbf{I}_n - \mathbf{K}\mathbf{H})\mathbf{P}^{\mathrm{f}}\\ &= \left[\mathbf{I}_n - \mathbf{S}_{\mathrm{f}}\mathbf{Y}_{\mathrm{f}}^{\mathrm{T}}\left(\mathbf{Y}_{\mathrm{f}}\mathbf{Y}_{\mathrm{f}}^{\mathrm{T}}+\mathbf{R}\right)^{-1}\mathbf{H}\right]\mathbf{S}_{\mathrm{f}}\mathbf{S}_{\mathrm{f}}^{\mathrm{T}}\\ &= \mathbf{S}_{\mathrm{f}}\left[\mathbf{I}_m - \mathbf{Y}_{\mathrm{f}}^{\mathrm{T}}\left(\mathbf{Y}_{\mathrm{f}}\mathbf{Y}_{\mathrm{f}}^{\mathrm{T}}+\mathbf{R}\right)^{-1}\mathbf{Y}_{\mathrm{f}}\right]\mathbf{S}_{\mathrm{f}}^{\mathrm{T}},\end{aligned}$$

where $\mathbf{I}_n$ ($\mathbf{I}_m$) is the $n \times n$ ($m \times m$) identity matrix. We look for a *square root* matrix such that $\mathbf{S}_{\mathrm{a}}\mathbf{S}_{\mathrm{a}}^{\mathrm{T}} = \mathbf{P}^{\mathrm{a}}$. One such matrix of size $n \times m$ is

$$\mathbf{S}_{\mathrm{a}} = \mathbf{S}_{\mathrm{f}}\left[\mathbf{I}_m - \mathbf{Y}_{\mathrm{f}}^{\mathrm{T}}(\mathbf{Y}_{\mathrm{f}}\mathbf{Y}_{\mathrm{f}}^{\mathrm{T}}+\mathbf{R})^{-1}\mathbf{Y}_{\mathrm{f}}\right]^{\frac{1}{2}} \tag{6.2}$$

and represents a collection of m state vectors, that is to say a posterior ensemble. The square root matrix that we use in (6.2) is defined as follows. Given $\mathbf{A}$ a diagonalizable matrix, such that $\mathbf{A} = \mathbf{\Omega}\mathbf{\Lambda}\mathbf{\Omega}^{-1}$, where $\mathbf{\Lambda}$ is diagonal with nonnegative entries and $\mathbf{\Omega}$ is invertible, then the square root of $\mathbf{A}$ is defined by $\mathbf{A}^{\frac{1}{2}} = \mathbf{\Omega}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{\Omega}^{-1}$.

This factorization avoids a brutal calculation of the error covariance matrices. The computation of the square root matrix might look numerically costly. However, this is not the case since $\mathbf{Y}_{\mathrm{f}}$ is of reduced size and $\mathbf{I}_m - \mathbf{Y}_{\mathrm{f}}^{\mathrm{T}}(\mathbf{Y}_{\mathrm{f}}\mathbf{Y}_{\mathrm{f}}^{\mathrm{T}}+\mathbf{R})^{-1}\mathbf{Y}_{\mathrm{f}}$ is of dimension $m \times m$. In addition, the conditioning of the square root matrix is better than that of the initial error covariance matrix, as it becomes the square root of the original conditioning. This ensures finer numerical precision. This square root update in matrix form was first proposed by Andrews [1968].

After the analysis step, we seek to reduce the dimension of the system. We wish to shrink the number of modes from m to $m-q$. We first diagonalize $\mathbf{S}_{\mathrm{a}}^{\mathrm{T}}\mathbf{S}_{\mathrm{a}} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}$, where $\mathbf{V}$ is an orthogonal matrix and $\mathbf{\Lambda}$ is a diagonal matrix of nonnegative entries. We consider the first $m-q$ eigenmodes with the largest eigenvalues. We keep the first $m-q$ eigenvectors that correspond to the first $m-q$ columns of $\mathbf{V}$ if the diagonal entries of $\mathbf{\Lambda}$ are stored in decreasing order. These $m-q$ vectors are stored in the $m \times (m-q)$ matrix $\widetilde{\mathbf{V}}$. Then $\mathbf{S}_{\mathrm{a}}$ is reduced to $\widetilde{\mathbf{S}}_{\mathrm{a}} \equiv \mathbf{S}_{\mathrm{a}}\widetilde{\mathbf{V}}$, which is an $n \times (m-q)$ matrix.

At the forecast step, the analysis is forecast by $\mathbf{x}_{k+1}^{\mathrm{f}} = \mathcal{M}_{k+1}(\mathbf{x}_k^{\mathrm{a}})$. The square root matrix $\mathbf{S}_k^{\mathrm{a}}$ is propagated with the TLM. More precisely, the matrix is enlarged by

Algorithm 6.2 Algorithm for RRSQRT

Require: For $k = 0, \dots, K$: the observation error covariances $\mathbf{R}_k$, the model error covariances $\mathbf{Q}_k$, the observation models $\mathscr{H}_k$ and their tangent linear $\mathbf{H}_k$, the forward models $\mathscr{M}_k$ and their tangent linear $\mathbf{M}_k$.

Initialize system state $\mathbf{x}_0^{\mathrm{f}}$ and error covariance matrix $\mathbf{P}_0^{\mathrm{f}}$.

Decomposition: $\mathbf{Q}_k = \mathbf{T}_k \mathbf{T}_k^{\mathrm{T}}$ and $\mathbf{P}_0^{\mathrm{f}} = \mathbf{S}_0^{\mathrm{f}} \left(\mathbf{S}_0^{\mathrm{f}}\right)^{\mathrm{T}}$.

for $k = 0, \dots, K$ **do**

  Compute the gain: $\mathbf{Y}_k = \mathbf{H}_k \mathbf{S}_k^{\mathrm{f}}, \qquad \mathbf{K}_k = \mathbf{S}_k^{\mathrm{f}} \mathbf{Y}_k^{\mathrm{T}} \left(\mathbf{Y}_k \mathbf{Y}_k^{\mathrm{T}} + \mathbf{R}_k\right)^{-1}$.

  Compute the state analysis: $\mathbf{x}_k^{\mathrm{a}} = \mathbf{x}_k^{\mathrm{f}} + \mathbf{K}_k \left[\mathbf{y}_k - \mathscr{H}_k(\mathbf{x}_k^{\mathrm{f}})\right]$.

  Compute the (square root) matrix of the modes:

  $\mathbf{S}_k^{\mathrm{a}} = \mathbf{S}_k^{\mathrm{f}} \left(\mathbf{I}_m - \mathbf{Y}_k^{\mathrm{T}} (\mathbf{Y}_k \mathbf{Y}_k^{\mathrm{T}} + \mathbf{R}_k)^{-1} \mathbf{Y}_k\right)^{\frac{1}{2}}$, where $\mathbf{S}_k^{\mathrm{a}}$ has m modes.

  Diagonalization: $\mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^{\mathrm{T}} = \left(\mathbf{S}_k^{\mathrm{a}}\right)^{\mathrm{T}} \mathbf{S}_k^{\mathrm{a}}$.

  Selection of the first $m - q$ modes of $\mathbf{V} \to \widetilde{\mathbf{V}}$.

  Reduction of the ensemble: $\widetilde{\mathbf{S}}_k^{\mathrm{a}} = \mathbf{S}_k^{\mathrm{a}} \widetilde{\mathbf{V}}$. Then $\widetilde{\mathbf{S}}_k^{\mathrm{a}}$ has $m - q$ modes.

  Compute the forecast state: $\mathbf{x}_{k+1}^{\mathrm{f}} = \mathscr{M}_{k+1}[\mathbf{x}_k^{\mathrm{a}}]$

  Compute the forecast modes: $\mathbf{S}_{k+1}^{\mathrm{f}} = [\mathbf{M}_{k+1} \widetilde{\mathbf{S}}_k^{\mathrm{a}}, \mathbf{T}_k]$, where $\mathbf{S}_{k+1}^{\mathrm{f}}$ has m modes.

end for

adding q modes that are meant to introduce model error variability. The matrix of these augmented modes is of the form

$$\mathbf{S}_{k+1}^{\mathrm{f}} = [\mathbf{M}_{k+1} \widetilde{\mathbf{S}}_k^{\mathrm{a}}, \mathbf{T}_k],$$

where $\mathbf{T}_k$ introduces m n-vector perturbations. The matrix $\mathbf{S}_{k+1}^{\mathrm{f}}$ has m modes, so that the assimilation procedure can be cycled. The full RRSQRT scheme is summarized in Algorithm 6.2.

This type of filter has been proposed in hydrology and oceanography by Verlaan and Heemink [1997, 2001]. As described in Chapter 5, successful variants are known under the names of SEEK and SEIK [Pham et al., 1998; Pham, 2001]. It has been advocated for air quality forecasts, where the number of chemical species increases the state space dimension considerably [Segers, 2002; Hanea et al., 2004; Wu et al., 2008].

These filters clearly overcome the main drawback of the KF, provided the dynamics of the system can be represented with a limited ($m \ll n$) number of modes. However, by still making use of the TLM, the propagation of uncertainty remains approximate. In that respect, one can propose an even finer reduced KF, the ensemble Kalman filter (EnKF).

6.2 ▪ The EnKF: Principle and classification

The EnKF was proposed by G. Evensen in 1994 and later amended in 1998 [Evensen, 1994; Burgers et al., 1998; Houtekamer and Mitchell, 1998; Evensen, 2009]. Its semi-empirical justification could be disconcerting and initially not as obvious as that of the RRSQRT, in spite of a certain elegance. Nevertheless, the EnKF has proven very efficient on a large number of both academic and operational DA problems. It has become very popular over the past 20 years in its many forms.

The EnKF could be seen as a reduced-order KF, like the RRSQRT filter we described. Just as the EKF and the RRSQRT, it only handles error statistics up to second

order (i.e., mean and covariances), which is loosely summarized by designating it as a Gaussian filter. Because of this truncation of statistics, it has been shown that in the limit of a large number of particles, the EnKF does not solve the Bayesian filtering problem, as opposed to the particle filter (see Chapter 3), except when the models are linear and when the initial error distribution is Gaussian and is spanned by the ensemble deviations from the mean [Le Gland et al., 2011; Mandel et al., 2011]. Nonetheless, this does not prevent the EnKF from often being a good approximate algorithm for the filtering problem.

Just as with the particle filter and the RRSQRT filter, the EnKF is based on the concept of particles, a collection of state vectors, which are called the members of the ensemble. Rather than propagating huge covariance matrices, the errors are emulated by scattered particles, a collection of state vectors whose variability is meant to represent the uncertainty of the system's state, which comes from the forecaster's ignorance. Just like the particle filter, but unlike the RRSQRT, the members are to be propagated by the model, without any linearization. Not only does this avoid the derivation of the TLM, but it also circumvents the approximate linearization. Finally, as opposed to the particle filter, the EnKF does not irremediably suffer from the curse of dimensionality (see Chapter 3).

Essentially two main flavors of EnKF have been proposed. The EnKFs of the first class are termed *stochastic* because some random sampling noise is generated in the analysis. The EnKFs of the second class are termed *deterministic* since they do not use random perturbations, which is achieved by the square root formalism that was first described with the RRSQRT.

6.3 ▪ The stochastic EnKF

The focus is first on the stochastic EnKF and its analysis step.

6.3.1 ▪ The analysis step

The EnKF seeks to mimic the analysis step of the KF but with an ensemble of limited size in place of the cumbersome covariance matrices. The goal is to perform for each member of the ensemble an analysis of the form

$$\mathbf{x}_i^{\mathrm{a}} = \mathbf{x}_i^{\mathrm{f}} + \mathbf{K}\left[\mathbf{y} - \mathscr{H}(\mathbf{x}_i^{\mathrm{f}})\right], \tag{6.3}$$

where $i = 1, \ldots, m$ is the member index in the ensemble and $\mathbf{x}_i^{\mathrm{f}}$ is the forecast state vector i, which represents a background state or prior at the analysis time. To mimic the KF, $\mathbf{K}$ must be identified with the Kalman gain,

$$\mathbf{K} = \mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right)^{-1}, \tag{6.4}$$

that we wish to estimate from the ensemble statistics. First of all, we can estimate the forecast error covariance matrix as

$$\mathbf{P}^{\mathrm{f}} = \frac{1}{m-1}\sum_{i=1}^{m}\left(\mathbf{x}_i^{\mathrm{f}} - \overline{\mathbf{x}}^{\mathrm{f}}\right)\left(\mathbf{x}_i^{\mathrm{f}} - \overline{\mathbf{x}}^{\mathrm{f}}\right)^{\mathrm{T}}, \quad \text{with} \quad \overline{\mathbf{x}}^{\mathrm{f}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i^{\mathrm{f}}.$$

This forecast error covariance matrix can be factorized into

$$\mathbf{P}^{\mathrm{f}} = \mathbf{X}_{\mathrm{f}}\mathbf{X}_{\mathrm{f}}^{\mathrm{T}},$$

where $\mathbf{X}_\mathrm{f}$ is an $n \times m$ matrix whose columns are the *normalized anomalies* or *normalized perturbations*, i.e., for $i = 1, \ldots, m$,

$$[\mathbf{X}_\mathrm{f}]_i = \frac{\mathbf{x}_i^\mathrm{f} - \overline{\mathbf{x}}^\mathrm{f}}{\sqrt{m-1}}.$$

Here $\mathbf{X}_\mathrm{f}$ plays the role of the matrix of the reduced modes $\mathbf{S}_\mathrm{f}$ of the RRSQRT filter (see Section 6.1).

Thanks to (6.3), we can obtain a posterior ensemble, $\{\mathbf{x}_i^\mathrm{a}\}_{i=1,\ldots,m}$, from which we can compute the posterior statistics. Hence, the posterior state and an ensemble of posterior perturbations can be computed from

$$\overline{\mathbf{x}}^\mathrm{a} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i^\mathrm{a}, \quad [\mathbf{X}_\mathrm{a}]_i = \frac{\mathbf{x}_i^\mathrm{a} - \overline{\mathbf{x}}^\mathrm{a}}{\sqrt{m-1}}.$$

The normalized anomalies, $\mathbf{X}_i^\mathrm{a} \equiv [\mathbf{X}_\mathrm{a}]_i$, i.e., the normalized deviations of the ensemble members from the mean, are obtained from (6.3) minus the mean update,

$$\mathbf{X}_i^\mathrm{a} = \mathbf{X}_i^\mathrm{f} + \mathbf{K}\left(0 - \mathbf{H}\mathbf{X}_i^\mathrm{f}\right) = (\mathbf{I}_n - \mathbf{K}\mathbf{H})\,\mathbf{X}_i^\mathrm{f}, \tag{6.5}$$

where $\mathbf{X}_i^\mathrm{f} \equiv [\mathbf{X}_\mathrm{f}]_i$, which yields the analysis error covariance matrix

$$\mathbf{P}^\mathrm{a} = \mathbf{X}_\mathrm{a}\mathbf{X}_\mathrm{a}^\mathrm{T} = (\mathbf{I}_n - \mathbf{K}\mathbf{H})\mathbf{X}_\mathrm{f}\mathbf{X}_\mathrm{f}^\mathrm{T}(\mathbf{I}_n - \mathbf{K}\mathbf{H})^\mathrm{T} = (\mathbf{I}_n - \mathbf{K}\mathbf{H})\mathbf{P}^\mathrm{f}(\mathbf{I}_n - \mathbf{K}\mathbf{H})^\mathrm{T}. \tag{6.6}$$

However, theoretically, in order to mimic the BLUE analysis of the KF, we should have obtained instead (see Chapter 3)

$$\mathbf{P}^\mathrm{a} = (\mathbf{I}_n - \mathbf{K}\mathbf{H})\mathbf{P}^\mathrm{f}(\mathbf{I}_n - \mathbf{K}\mathbf{H})^\mathrm{T} + \mathbf{K}\mathbf{R}\mathbf{K}^\mathrm{T} = (\mathbf{I}_n - \mathbf{K}\mathbf{H})\mathbf{P}^\mathrm{f}. \tag{6.7}$$

Therefore, the error covariances of (6.6) are underestimated since the second positive term in (6.7), related to the observation errors, is ignored, which is likely to lead to the divergence of the EnKF when the scheme is cycled.

An elegant solution to this problem is to perturb the observation vector for each member: $\mathbf{y} \to \mathbf{y}_i \equiv \mathbf{y} + \mathbf{u}_i$, where $\mathbf{u}_i$ is drawn from the Gaussian distribution $\mathbf{u}_i \sim \mathcal{N}(0, \mathbf{R})$. Let us define $\overline{\mathbf{u}}$, the mean of the sampled $\mathbf{u}_i$, and the innovation perturbations

$$[\mathbf{Y}_\mathrm{f}]_i = \frac{\mathbf{H}\mathbf{x}_i^\mathrm{f} - \mathbf{u}_i - \mathbf{H}\overline{\mathbf{x}}^\mathrm{f} + \overline{\mathbf{u}}}{\sqrt{m-1}}, \tag{6.8}$$

which shows that the forecast observations $\mathbf{H}\mathbf{x}_i^\mathrm{f}$ are the quantities that are genuinely perturbed. The posterior anomalies are modified accordingly:

$$\mathbf{X}_i^\mathrm{a} = \mathbf{X}_i^\mathrm{f} - \mathbf{K}\mathbf{Y}_i^\mathrm{f} = (\mathbf{I}_n - \mathbf{K}\mathbf{H})\mathbf{X}_i^\mathrm{f} + \frac{\mathbf{K}(\mathbf{u}_i - \overline{\mathbf{u}})}{\sqrt{m-1}}. \tag{6.9}$$

Denoting $\mathbf{E}_i = (\mathbf{u}_i - \overline{\mathbf{u}})/\sqrt{m-1}$ and $\mathbf{E} = [\mathbf{E}_1, \ldots, \mathbf{E}_m]$, we have $\mathbf{X}_\mathrm{a} = \mathbf{X}_\mathrm{f} - \mathbf{K}\mathbf{Y}_\mathrm{f} = (\mathbf{I}_n - \mathbf{K}\mathbf{H})\mathbf{X}_\mathrm{f} + \mathbf{K}\mathbf{E}$, which yields the analysis error covariance matrix

$$\mathbf{P}^\mathrm{a} = (\mathbf{I}_n - \mathbf{K}\mathbf{H})\mathbf{P}^\mathrm{f}(\mathbf{I}_n - \mathbf{K}\mathbf{H})^\mathrm{T} + \mathbf{K}\mathbf{E}\mathbf{E}^\mathrm{T}\mathbf{K}^\mathrm{T} + (\mathbf{I}_n - \mathbf{K}\mathbf{H})\mathbf{X}_\mathrm{f}\mathbf{E}^\mathrm{T}\mathbf{K}^\mathrm{T} + \mathbf{K}\mathbf{E}\mathbf{X}_\mathrm{f}^\mathrm{T}(\mathbf{I}_n - \mathbf{K}\mathbf{H})^\mathrm{T},$$

whose expectation over the random noise gives the proper expected posterior covariances:

$$\begin{aligned}\mathrm{E}[\mathbf{P}^\mathrm{a}] &= (\mathbf{I}_n - \mathbf{K}\mathbf{H})\mathbf{P}^\mathrm{f}(\mathbf{I}_n - \mathbf{K}\mathbf{H})^\mathrm{T} + \mathbf{K}\mathrm{E}\left[\mathbf{E}\mathbf{E}^\mathrm{T}\right]\mathbf{K}^\mathrm{T} \\ &= (\mathbf{I}_n - \mathbf{K}\mathbf{H})\mathbf{P}^\mathrm{f}(\mathbf{I}_n - \mathbf{K}\mathbf{H})^\mathrm{T} + \mathbf{K}\mathbf{R}\mathbf{K} = (\mathbf{I}_n - \mathbf{K}\mathbf{H})\mathbf{P}^\mathrm{f}.\end{aligned}$$

Note that the gain can be formulated in terms of the anomaly matrices only,

$$\mathbf{K} = \mathbf{X}_\mathrm{f}\mathbf{Y}_\mathrm{f}^\mathrm{T}\left(\mathbf{Y}_\mathrm{f}\mathbf{Y}_\mathrm{f}^\mathrm{T}\right)^{-1}, \tag{6.10}$$

since $\mathbf{X}_\mathrm{f}\mathbf{Y}_\mathrm{f}^\mathrm{T}$ is a sample estimate for $\mathbf{P}^\mathrm{f}\mathbf{H}^\mathrm{T}$ and $\mathbf{Y}_\mathrm{f}\mathbf{Y}_\mathrm{f}^\mathrm{T}$ is a sample estimate for $\mathbf{H}\mathbf{P}^\mathrm{f}\mathbf{H}^\mathrm{T}+\mathbf{R}$. In this form, it is striking that the updated perturbations are linear combinations of the forecast perturbations. The new perturbations are sought within the *ensemble subspace* of the initial perturbations. Similarly, the state analysis is sought within the affine space $\overline{\mathbf{x}}^\mathrm{f} + \mathrm{Vec}\left(\mathbf{X}_1^\mathrm{f}, \mathbf{X}_2^\mathrm{f}, \ldots, \mathbf{X}_m^\mathrm{f}\right)$.

6.3.2 ▪ The forecast step

In the forecast step, the updated ensemble obtained at the analysis step is propagated by the model over a time step:

$$\text{for} \quad i = 1, \ldots, m, \quad \mathbf{x}_{i,k+1}^\mathrm{f} = \mathcal{M}_{k+1}(\mathbf{x}_{i,k}^\mathrm{a}).$$

A forecast can be computed from the mean of the forecast ensemble, while the forecast error covariances can be estimated from the forecast perturbations. Yet, these are only optional diagnostics in the scheme and they are not required in the cycling of the EnKF. It is important to observe that the use of the TLM operator is avoided. This makes a significant difference with the RRSQRT filter, especially in a significantly nonlinear regime. However, in this case, the EnKF is itself outdone by schemes known as the iterative EnKF (IEnKF) and the iterative ensemble Kalman smoother (IEnKS). These will be discussed in Chapter 7.

6.3.3 ▪ Avoiding the observation tangent linear operator

Like the particle filter or 3D-Var, the EnKF does not require the tangent linear of the forecast model. More interestingly, it can also avoid the computation of the adjoint of the observation operator. Let us consider the formula of the Kalman gain (6.10), which makes use of the adjoint of the observation operator via (6.8). The matrix products $\mathbf{P}^\mathrm{f}\mathbf{H}^\mathrm{T}$ and $\mathbf{H}\mathbf{P}^\mathrm{f}\mathbf{H}^\mathrm{T}$ both require the tangent linear of the observation operator. Instead, these can be estimated by an analogue of finite differences using the full observation model [Evensen, 1994; Houtekamer and Mitchell, 1998],

$$\begin{aligned}
\overline{\mathbf{y}}^\mathrm{f} &= \frac{1}{m}\sum_{i=1}^{m} \mathscr{H}\left(\mathbf{x}_i^\mathrm{f}\right), \\
\mathbf{P}^\mathrm{f}\mathbf{H}^\mathrm{T} &= \frac{1}{m-1}\sum_{i=1}^{m}\left(\mathbf{x}_i^\mathrm{f} - \overline{\mathbf{x}}^\mathrm{f}\right)\left[\mathbf{H}\left(\mathbf{x}_i^\mathrm{f} - \overline{\mathbf{x}}^\mathrm{f}\right)\right]^\mathrm{T} \\
&\simeq \frac{1}{m-1}\sum_{i=1}^{m}\left(\mathbf{x}_i^\mathrm{f} - \overline{\mathbf{x}}^\mathrm{f}\right)\left[\mathscr{H}\left(\mathbf{x}_i^\mathrm{f}\right) - \overline{\mathbf{y}}^\mathrm{f}\right]^\mathrm{T}, \\
\mathbf{H}\mathbf{P}^\mathrm{f}\mathbf{H}^\mathrm{T} &= \frac{1}{m-1}\left[\mathbf{H}\left(\mathbf{x}_i^\mathrm{f} - \overline{\mathbf{x}}^\mathrm{f}\right)\right]\left[\mathbf{H}\left(\mathbf{x}_i^\mathrm{f} - \overline{\mathbf{x}}^\mathrm{f}\right)\right]^\mathrm{T} \\
&\simeq \frac{1}{m-1}\sum_{i=1}^{m}\left[\mathscr{H}\left(\mathbf{x}_i^\mathrm{f}\right) - \overline{\mathbf{y}}^\mathrm{f}\right]\left[\mathscr{H}\left(\mathbf{x}_i^\mathrm{f}\right) - \overline{\mathbf{y}}^\mathrm{f}\right]^\mathrm{T},
\end{aligned} \tag{6.11}$$

where the key assumption in these derivations is the approximation

$$\mathbf{H}\left(\mathbf{x}_i^\mathrm{f} - \overline{\mathbf{x}}^\mathrm{f}\right) \simeq \mathscr{H}\left(\mathbf{x}_i^\mathrm{f}\right) - \overline{\mathbf{y}}^\mathrm{f},$$

Algorithm 6.3 Algorithm for the (stochastic) EnKF

Require: For $k = 0, \dots, K$: the observation error covariance matrices $\mathbf{R}_k$, the observation models $\mathscr{H}_k$, the forward models $\mathscr{M}_k$.

1: Initialize the ensemble $\left\{\mathbf{x}^{\mathrm{f}}_{i,0}\right\}_{=1,\dots,m}$.
2: **for** $k = 0, \dots, K$ **do**
3: Draw a statistically consistent observation set:
for $i = 1, .., m$: $\mathbf{y}_{i,k} = \mathbf{y}_k + \mathbf{u}_i$, with $\mathbf{u}_i \sim \mathcal{N}(0, \mathbf{R}_k)$
4: Compute the ensemble means
$\overline{\mathbf{x}}^{\mathrm{f}}_k = \frac{1}{m}\sum_{i=1}^m \mathbf{x}^{\mathrm{f}}_{i,k}$, $\overline{\mathbf{u}} = \frac{1}{m}\sum_{i=1}^m \mathbf{u}_i$, $\overline{\mathbf{y}}^{\mathrm{f}}_k = \frac{1}{m}\sum_{i=1}^m \mathscr{H}_k\left(\mathbf{x}^{\mathrm{f}}_{i,k}\right)$
and the normalized anomalies
$[\mathbf{X}_{\mathrm{f}}]_{i,k} = \frac{\mathbf{x}^{\mathrm{f}}_{i,k} - \overline{\mathbf{x}}^{\mathrm{f}}_k}{\sqrt{m-1}}$, $[\mathbf{Y}_{\mathrm{f}}]_{i,k} = \frac{\mathscr{H}_k\left(\mathbf{x}^{\mathrm{f}}_{i,k}\right) - \mathbf{u}_i - \overline{\mathbf{y}}^{\mathrm{f}}_k + \overline{\mathbf{u}}}{\sqrt{m-1}}$.
5: Compute the gain: $\mathbf{K}_k = \mathbf{X}^{\mathrm{f}}_k \left(\mathbf{Y}^{\mathrm{f}}_k\right)^{\mathrm{T}} \left\{\mathbf{Y}^{\mathrm{f}}_k \left(\mathbf{Y}^{\mathrm{f}}_k\right)^{\mathrm{T}}\right\}^{-1}$
6: Update of the ensemble:
for $i = 1, .., m$: $\mathbf{x}^{\mathrm{a}}_{i,k} = \mathbf{x}^{\mathrm{f}}_{i,k} + \mathbf{K}_k\left(\mathbf{y}_{i,k} - \mathscr{H}_k(\mathbf{x}^{\mathrm{f}}_{i,k})\right)$,
7: Compute the ensemble forecast:
for $i = 1, .., m$ $\mathbf{x}^{\mathrm{f}}_{i,k+1} = \mathscr{M}_{k+1}(\mathbf{x}^{\mathrm{a}}_{i,k})$.
8: **end for**

which can be seen as finite differences if the spread of the ensemble is small enough. This avoids the use of not only the adjoint model but also the tangent linear of the observation operator. This suggests modifying the definition of the perturbations as

$$[\mathbf{Y}_{\mathrm{f}}]_i = \frac{\mathscr{H}\left(\mathbf{x}^{\mathrm{f}}_i\right) - \mathbf{u}_i - \overline{\mathbf{y}}^{\mathrm{f}} + \overline{\mathbf{u}}}{\sqrt{m-1}}.$$

With the important exception of the Kalman gain's computation, all the operations on the ensemble members are independent. This implies that their parallelization can be carried out straightforwardly. This is one of the main reasons for the success and popularity of the EnKF. The algorithm of the stochastic EnKF is given in Algorithm 6.3.

The stochastic EnKF has one disadvantage over other deterministic variants of the EnKF (which we will investigate in Section 6.4): it introduces stochastic noise. This may impact the performance of the filter to an extent that depends on the precise DA setup; on the dynamics nonlinearity; and, above all, on the ensemble size.

6.3.4 ▪ Numerical illustration

The discrete model

$$x_0 = 0, \quad x_1 = 1, \quad \text{and for } 1 \le k \le N, \quad x_{k+1} - 2x_k + x_{k-1} = \omega^2 x_k - \lambda^2 x_k^3, \quad (6.12)$$

is a numerical implementation of the anharmonic oscillator of a material point,

$$\frac{\mathrm{d}^2 x}{\mathrm{d}\, t^2} - \Omega^2 x + \Lambda^2 x^3 = 0,$$

where Ω^2 is proportional to ω^2 and Λ^2 is proportional to λ^2. The related potential energy is $V(x) = -\frac{1}{2}\Omega^2 x^2 + \frac{1}{4}\Lambda^2 x^4$. The second term of $V(x)$ stabilizes the oscillator

and plays the role of a spring force, whereas the first term destabilizes the point of origin, $x = 0$, leading to two potential wells. It is a second-order discrete equation, with a state vector that can be advantageously written as

$$\mathbf{u}_k = \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix}. \tag{6.13}$$

From (6.12) and (6.13), the state-dependent transition matrix is

$$\mathcal{M}_{k+1} = \begin{bmatrix} 2 + \omega^2 - \lambda^2 x_k^2 & -1 \\ 1 & 0 \end{bmatrix},$$

so that $\mathbf{u}_{k+1} = \mathcal{M}_{k+1}(\mathbf{u}_k)$.

The EKF and the stochastic EnKF are tested with this nonlinear system in the absence of model error. We choose an ensemble of size $m = 10$, much greater than the state system dimension ($n = 2$). That is why the focus of this experiment is not on the impact of the reduction of dimensionality obtained by the limited size ensemble, but on the impact of the nonlinear dynamics on the filter's performance.

Figure 6.1 gives an example of synthetic (or twin) DA experiments. The true trajectory, *the truth*, is run as a reference, with $\omega = 3.5 \times 10^{-2}$ and $\lambda = 3 \times 10^{-4}$. Its trajectory is shown with a thick dashed black line. Obviously, the trajectory seems periodic. The truth is not meant to be directly accessible. However, it is observed through a set of observations that measure x_k, so that the observation operator is $\mathbf{H}_k = [1, 0]$. The observation equation is

$$y_k = \mathbf{H}_k \mathbf{u}_k + \epsilon_k.$$

Each noiseless observation, $\mathbf{H}_k \mathbf{u}_k$, is independently perturbed with an unbiased Gaussian noise, ϵ_k, of standard deviation $\sigma = 10$. The perturbed observations are indicated by circles. In the top panels, the system is observed every 25 time steps, whereas it is observed every 50 time steps in the bottom panels. Clearly, with 50 time steps between observations, the model nonlinearity has a stronger impact than with an interval of 25 time steps.

The assimilation is meant to track the truth using the perfect model, in particular with the same ω and λ, and the perturbed observations. However, the initial conditions are not known. The best estimate (i.e., through the analysis/forecast cycles) trajectory is plotted as a full line curve for the EKF and the EnKF. One can observe several abrupt corrections of the trajectory of the best estimate when an analysis takes place, especially when the innovation of the assimilation is stronger.

In the frequent-observation case (upper panels), both the EKF and the EnKF keep track of the truth. However, in the less frequent case (lower panels), the EKF loses track of the truth, while the EnKF still follows it. The EKF is said to have diverged (from the truth) even though the trajectory remains bounded. The divergence occurs because of the approximate propagation of the errors with the TLM, which could lead to an underestimation of the uncertainty. The EKF becomes increasingly but wrongly confident, which leads to the divergence. This behavior can be reproduced with other initial conditions, although the divergence of the EKF may happen at a different time. In the context of this experiment, the only significant difference between the two methods is the handling of nonlinearity, which is finer in the EnKF than in the EKF.

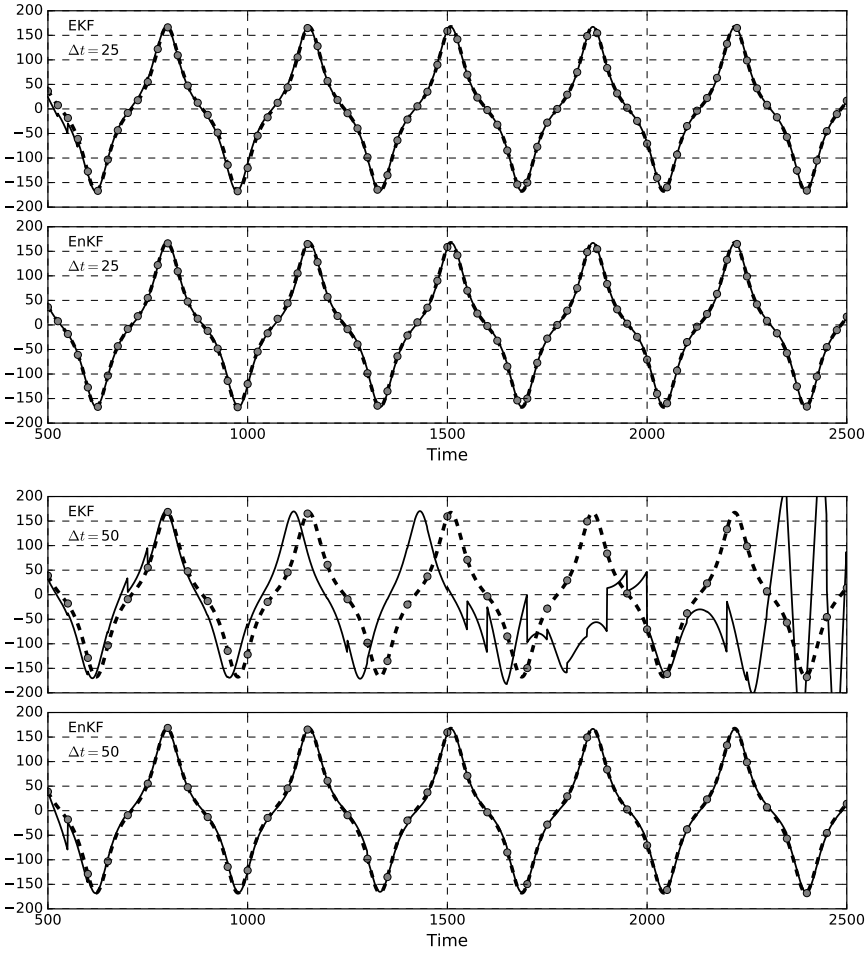

Figure 6.1. *Synthetic DA experiments with the anharmonic oscillator. The observations are represented by disks, the truth by a thick dashed line, and the EKF and EnKF best estimate trajectories by full line curves. The upper panels correspond to frequent observations (every* 25 *time steps), whereas the bottom panels correspond to half as frequent observations (every* 50 *time steps).*

6.4 ▪ The deterministic EnKF

The stochastic EnKF enables the individual tracking of each member of the ensemble, with the analysis step for the members to interact via the Kalman gain matrix. This nice property, central to stochastic filtering, comes with a price: we need to independently perturb the observation vector of each member. Although this seems elegant, it also introduces numerical noise when drawing the perturbations, $\mathbf{u}_i$. This can affect the performance of the stochastic EnKF.

An alternative idea for performing a statistically consistent EnKF analysis is to follow the square root approach implemented in the RRSQRT. This yields the *ensemble square root Kalman filter (EnSRKF)*. Since with this scheme the observations are not perturbed, it is sometimes called the *deterministic EnKF*. There are several variants of the EnSRKF [Tippett et al., 2003]. Be aware that they are, to a large extent, algebraically equivalent. Yet, their practical implementations may lead to discrepancies, especially

when localization is applied (see Section 6.5). In this chapter, we shall present the En-SRKF where the analysis is performed in state space [Whitaker and Hamill, 2002], as well as the so-called ensemble-transform variant of the deterministic EnKF, usually abbreviated *ETKF* [Bishop et al., 2001; Hunt et al., 2007]. In the latter case, rather than performing the linear algebra in state or observation space, the algebra is mostly performed in the ensemble subspace of small dimension.

6.4.1 ▪ Spanning the ensemble subspace

Let us define the ensemble subspace. Following the stochastic EnKF, we can write the forecast error covariance matrix as

$$\mathbf{P}^{\mathrm{f}} = \frac{1}{m-1}\sum_{i=1}^{m}\left(\mathbf{x}_i^{\mathrm{f}} - \overline{\mathbf{x}}^{\mathrm{f}}\right)\left(\mathbf{x}_i^{\mathrm{f}} - \overline{\mathbf{x}}^{\mathrm{f}}\right)^{\mathrm{T}} = \mathbf{X}_{\mathrm{f}}\mathbf{X}_{\mathrm{f}}^{\mathrm{T}},$$

where $\mathbf{X}_{\mathrm{f}}$ is an $n \times m$ matrix whose columns are the *normalized* anomalies

$$[\mathbf{X}_{\mathrm{f}}]_i = \frac{\mathbf{x}_i^{\mathrm{f}} - \overline{\mathbf{x}}^{\mathrm{f}}}{\sqrt{m-1}}.$$

We shall assume that the analysis belongs to the affine subspace whose vectors $\mathbf{x}$ are of the form

$$\mathbf{x} = \overline{\mathbf{x}}^{\mathrm{f}} + \mathbf{X}_{\mathrm{f}}\mathbf{w},$$

where $\mathbf{w}$ is a vector of coefficients in the ensemble subspace $\mathbb{R}^m$. The restriction to the ensemble subspace is an assumption shared with the stochastic EnKF. However, here, the state vector is parameterized in the ensemble subspace by $\mathbf{w}$. Note that the decomposition of $\mathbf{x}$ is not unique because the vector of coefficients $\mathbf{w} + \lambda\mathbf{1}$ with $\mathbf{1} = (1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^m$ yields the same state vector $\mathbf{x}$, since $\mathbf{X}_{\mathrm{f}}\mathbf{1} = \mathbf{0}$. Following the terminology used in physics, this degree of freedom is called a *gauge* degree of freedom.

Again we shall use the notation $\mathbf{Y}_{\mathrm{f}}$ to represent $\mathbf{H}\mathbf{X}_{\mathrm{f}}$ if the observation operator is linear (or linearized). If it is nonlinear, we consider $\mathbf{Y}_{\mathrm{f}}$ to be the matrix of the *observation anomalies* (see Section 6.3.3):

$$[\mathbf{Y}_{\mathrm{f}}]_i = \frac{\mathcal{H}(\mathbf{x}_i^{\mathrm{f}}) - \overline{\mathbf{y}}^{\mathrm{f}}}{\sqrt{m-1}} \quad \text{with} \quad \overline{\mathbf{y}}^{\mathrm{f}} = \frac{1}{m}\sum_{i=1}^{m}\mathcal{H}(\mathbf{x}_i^{\mathrm{f}}).$$

Note that no random perturbation was added to these anomalies, as opposed to the definition of $\mathbf{Y}_{\mathrm{f}}$ in the stochastic EnKF.

6.4.2 ▪ State analysis in state space and ensemble subspace

As opposed to the stochastic EnKF, we wish to perform a single analysis followed by the generation of a new set of perturbations centered on it, rather than performing an analysis for each member of the ensemble. For the state update, we can adapt the mean analysis of the stochastic EnKF

$$\mathbf{x}^{\mathrm{a}} = \overline{\mathbf{x}}^{\mathrm{f}} + \mathbf{K}\left[\mathbf{y} - \mathcal{H}(\overline{\mathbf{x}}^{\mathrm{f}})\right], \tag{6.14}$$

where we used $\mathbf{K} = \mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right)^{-1}$. Here, the Kalman gain is computed using $\mathbf{P}^{\mathrm{f}} = \mathbf{X}_{\mathrm{f}}\mathbf{X}_{\mathrm{f}}^{\mathrm{T}}$ and a known predetermined $\mathbf{R}$.

This analysis can be reformulated in the ensemble subspace. This means finding the corresponding optimal coefficient vector $\mathbf{w}^{\mathrm{a}}$ that parameterizes $\mathbf{x}^{\mathrm{a}}$,

$$\mathbf{x}^{\mathrm{a}} = \overline{\mathbf{x}}^{\mathrm{f}} + \mathbf{X}_{\mathrm{f}}\mathbf{w}^{\mathrm{a}}.$$

Inserting this decomposition into (6.14), and denoting the innovation vector as $\boldsymbol{\delta} = \mathbf{y} - \mathscr{H}(\overline{\mathbf{x}}^{\mathrm{f}})$, we obtain

$$\overline{\mathbf{x}}^{\mathrm{f}} + \mathbf{X}_{\mathrm{f}}\mathbf{w}^{\mathrm{a}} = \overline{\mathbf{x}}^{\mathrm{f}} + \mathbf{X}_{\mathrm{f}}\mathbf{X}_{\mathrm{f}}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{X}_{\mathrm{f}}\mathbf{X}_{\mathrm{f}}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right)^{-1}\boldsymbol{\delta},$$

which suggests

$$\mathbf{w}^{\mathrm{a}} = \mathbf{X}_{\mathrm{f}}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{X}_{\mathrm{f}}\mathbf{X}_{\mathrm{f}}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right)^{-1}\boldsymbol{\delta} = \mathbf{Y}_{\mathrm{f}}^{\mathrm{T}}\left(\mathbf{Y}_{\mathrm{f}}\mathbf{Y}_{\mathrm{f}}^{\mathrm{T}} + \mathbf{R}\right)^{-1}\boldsymbol{\delta}.$$

The gain, $\mathbf{K}$, is usually computed in the observation space. Using the Sherman–Morrison–Woodbury formula, it is possible to compute the gain in ensemble subspace. Remember that we learned in Section 2.4.3 that the gain, $\mathbf{K}$, can be written in state or observation space as

$$\mathbf{K} = \mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right)^{-1} = \left(\mathbf{P}^{\mathrm{f}-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}.$$

Using this identity and with the identifications $\mathbf{P}^{\mathrm{f}} \equiv \mathbf{I}_m$ and $\mathbf{H} \equiv \mathbf{Y}_{\mathrm{f}}$, we finally obtain

$$\mathbf{w}^{\mathrm{a}} = \left(\mathbf{I}_m + \mathbf{Y}_{\mathrm{f}}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y}_{\mathrm{f}}\right)^{-1}\mathbf{Y}_{\mathrm{f}}^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{\delta}.$$

The gain is now computed in the ensemble subspace rather than in the observation or state space. This is numerically efficient provided the ensemble is smaller than the number of observations. We refer to Hunt et al. [2007] for a detailed numerical complexity analysis.

6.4.3 ▪ Generating the posterior ensemble (ensemble subspace)

Writing the state analysis in ensemble subspace is merely a reformulation of the stochastic EnKF. The genuine difference between the deterministic EnKF and the stochastic EnKF appears in the generation of the posterior ensemble.

To generate a posterior ensemble of perturbations that would be representative of the posterior uncertainty, we would like to factorize $\mathbf{P}^{\mathrm{a}} = \mathbf{X}_{\mathrm{a}}\mathbf{X}_{\mathrm{a}}^{\mathrm{T}}$. We can repeat the derivation of the RRSQRT, (6.2):

$$\begin{aligned}
\mathbf{P}^{\mathrm{a}} &= (\mathbf{I}_n - \mathbf{K}\mathbf{H})\mathbf{P}^{\mathrm{f}} \\
&= \left(\mathbf{I}_n - \mathbf{X}_{\mathrm{f}}\mathbf{Y}_{\mathrm{f}}^{\mathrm{T}}\left(\mathbf{Y}_{\mathrm{f}}\mathbf{Y}_{\mathrm{f}}^{\mathrm{T}} + \mathbf{R}\right)^{-1}\mathbf{H}\right)\mathbf{X}_{\mathrm{f}}\mathbf{X}_{\mathrm{f}}^{\mathrm{T}} \\
&= \mathbf{X}_{\mathrm{f}}\left(\mathbf{I}_m - \mathbf{Y}_{\mathrm{f}}^{\mathrm{T}}\left(\mathbf{Y}_{\mathrm{f}}\mathbf{Y}_{\mathrm{f}}^{\mathrm{T}} + \mathbf{R}\right)^{-1}\mathbf{Y}_{\mathrm{f}}\right)\mathbf{X}_{\mathrm{f}}^{\mathrm{T}}.
\end{aligned}$$

That is why we choose

$$\mathbf{X}_{\mathrm{a}} = \mathbf{X}_{\mathrm{f}}\left(\mathbf{I}_m - \mathbf{Y}_{\mathrm{f}}^{\mathrm{T}}(\mathbf{Y}_{\mathrm{f}}\mathbf{Y}_{\mathrm{f}}^{\mathrm{T}} + \mathbf{R})^{-1}\mathbf{Y}_{\mathrm{f}}\right)^{\frac{1}{2}}\mathbf{U}, \tag{6.15}$$

with $\mathbf{U}$ an arbitrary orthogonal matrix. This expression can be simplified into

$$\begin{aligned}\mathbf{X}_\mathrm{a} &= \mathbf{X}_\mathrm{f}\left(\mathbf{I}_m - \mathbf{Y}_\mathrm{f}^\mathrm{T}(\mathbf{Y}_\mathrm{f}\mathbf{Y}_\mathrm{f}^\mathrm{T} + \mathbf{R})^{-1}\mathbf{Y}_\mathrm{f}\right)^{\frac{1}{2}}\mathbf{U} \\ &= \mathbf{X}_\mathrm{f}\left(\mathbf{I}_m - \left(\mathbf{I}_m + \mathbf{Y}_\mathrm{f}^\mathrm{T}\mathbf{R}^{-1}\mathbf{Y}_\mathrm{f}\right)^{-1}\mathbf{Y}_\mathrm{f}^\mathrm{T}\mathbf{R}^{-1}\mathbf{Y}_\mathrm{f}\right)^{\frac{1}{2}}\mathbf{U} \\ &= \mathbf{X}_\mathrm{f}\left[\left(\mathbf{I}_m + \mathbf{Y}_\mathrm{f}^\mathrm{T}\mathbf{R}^{-1}\mathbf{Y}_\mathrm{f}\right)^{-1}\left(\mathbf{I}_m + \mathbf{Y}_\mathrm{f}^\mathrm{T}\mathbf{R}^{-1}\mathbf{Y}_\mathrm{f} - \mathbf{Y}_\mathrm{f}^\mathrm{T}\mathbf{R}^{-1}\mathbf{Y}_\mathrm{f}\right)\right]^{\frac{1}{2}}\mathbf{U} \\ &= \mathbf{X}_\mathrm{f}\left(\mathbf{I}_m + \mathbf{Y}_\mathrm{f}^\mathrm{T}\mathbf{R}^{-1}\mathbf{Y}_\mathrm{f}\right)^{-\frac{1}{2}}\mathbf{U}, \end{aligned} \tag{6.16}$$

where we have used the Sherman–Morrison–Woodbury formula again between the first and the second lines. It is often called the *right transform* of the initial perturbations.

One critical property of the initial anomalies is that they are centered, which reads $\mathbf{X}_\mathrm{f}\mathbf{1} = \mathbf{0}$, where we recall that $\mathbf{1} = (1, \ldots, 1)^\mathrm{T}$. One would also want the updated perturbations to be centered on $\mathbf{x}^\mathrm{a}$. To ensure that $\mathbf{X}_\mathrm{a}\mathbf{1} = \mathbf{0}$, it is sufficient to satisfy $\mathbf{U}\mathbf{1} = \mathbf{1}$, which can easily be checked by right-multiplying the members of (6.15) by $\mathbf{1}$. Not only is this intuitively more appealing but it was also proven to lead to a more precise algorithm [Wang et al., 2004; Sakov and Oke, 2008a; Livings et al., 2008].

Moreover, even though this linear analysis is independent of the choice of $\mathbf{U}$, this choice may have an impact on the nonlinear ensemble forecast. The physical balance of the members of the ensemble can be affected differently by distinct $\mathbf{U}$'s. The $\mathbf{U}$ that minimizes the displacement between the prior perturbation and the updated perturbations is $\mathbf{U} = \mathbf{I}_m$ [Ott et al., 2004]. This is likely to mitigate the imbalance generated by the approximate linear analysis when dealing with nonlinear models.

This posterior ensemble of anomalies is all that we need to cycle the deterministic EnKF. Defining $\mathbf{T} = \left(\mathbf{I}_m + \mathbf{Y}_\mathrm{f}^\mathrm{T}\mathbf{R}^{-1}\mathbf{Y}_\mathrm{f}\right)^{-1}$, we can build the posterior ensemble by setting, for all $i = 1, \ldots, m$,

$$\mathbf{x}_i^\mathrm{a} = \mathbf{x}^\mathrm{a} + \sqrt{m-1}\mathbf{X}_\mathrm{f}\left[\mathbf{T}^{\frac{1}{2}}\mathbf{U}\right]_i = \overline{\mathbf{x}}^\mathrm{f} + \mathbf{X}_\mathrm{f}\left(\mathbf{w}^\mathrm{a} + \sqrt{m-1}\left[\mathbf{T}^{\frac{1}{2}}\mathbf{U}\right]_i\right). \tag{6.17}$$

The $\mathbf{T}$ matrix is acting on the right of the anomaly matrix $\mathbf{X}_\mathrm{f}$. Again, this shows that posterior perturbations are linear combinations of the initial perturbations.

The analysis update in ensemble subspace is followed by an ensemble forecast similar to the stochastic EnKF. The resulting algorithm of this ETKF is proposed in Algorithm 6.4 in the form of a pseudocode.

6.4.4 ▪ Generating the posterior ensemble (state space)

It can also be convenient, as we shall see later, to have an equivalent update formula but in state space, rather than algebraically obtained in the ensemble subspace. Hence, one seeks a linear transform acting on the left of the prior perturbations $\mathbf{X}_\mathrm{f}$. To find one, we will need a simple formula for matrices.

Lemma 6.1. *Assume that the scalar map $x \mapsto f(x)$ can be expanded as a formal power series: $f(x) = \sum_{k=0}^{\infty} \alpha_k x^k$. Then, we have, at a formal level, $\mathbf{A}f(\mathbf{BA}) = f(\mathbf{AB})\mathbf{A}$ for any two matrices $\mathbf{A}$ and $\mathbf{B}$ of compatible dimensions since*

$$\mathbf{A}f(\mathbf{BA}) = \mathbf{A}\sum_{k=0}^{\infty}\alpha_k(\mathbf{BA})^k = \sum_{k=0}^{\infty}\alpha_k\mathbf{A}(\mathbf{BA})^k = \sum_{k=0}^{\infty}\alpha_k(\mathbf{AB})^k\mathbf{A} = f(\mathbf{AB})\mathbf{A}.$$

Algorithm 6.4 Pseudocode for a complete cycle of the ETKF

Require: Observation operator $\mathscr{H}$ at current time; $\mathbf{E}$, the forecast ensemble at current time; $\mathbf{y}$, the observation at current time; $\mathbf{U}$, an orthogonal matrix in $\mathbb{R}^{m\times m}$ satisfying $\mathbf{U1}=\mathbf{1}$; $\mathscr{M}$, the model resolvent from current time to the next analysis time; $\mathbf{R}$, is the error covariance matrix.

$\overline{\mathbf{x}} = \mathbf{E1}/m$
$\mathbf{X} = \left(\mathbf{E} - \overline{\mathbf{x}}\mathbf{1}^{\mathrm{T}}\right)/\sqrt{m-1}$
$\mathbf{Z} = \mathscr{H}(\mathbf{E})$
$\overline{\mathbf{y}} = \mathbf{Z1}/m$
$\mathbf{S} = \mathbf{R}^{-\frac{1}{2}}\left(\mathbf{Z} - \overline{\mathbf{y}}\mathbf{1}^{\mathrm{T}}\right)/\sqrt{m-1}$
$\boldsymbol{\delta} = \mathbf{R}^{-\frac{1}{2}}(\mathbf{y} - \overline{\mathbf{y}})$
$\mathbf{T} = \left(\mathbf{I}_m + \mathbf{S}^{\mathrm{T}}\mathbf{S}\right)^{-1}$
$\mathbf{w} = \mathbf{T}\mathbf{S}^{\mathrm{T}}\boldsymbol{\delta}$
$\mathbf{E} := \overline{\mathbf{x}}\mathbf{1}^{\mathrm{T}} + \mathbf{X}\left(\mathbf{w}\mathbf{1}^{\mathrm{T}} + \sqrt{m-1}\mathbf{T}^{\frac{1}{2}}\mathbf{U}\right)$
$\mathbf{E} := \mathscr{M}(\mathbf{E})$

We apply Lemma 6.1 (the so-called matrix shift lemma), $\mathbf{A}f(\mathbf{BA}) = f(\mathbf{AB})\mathbf{A}$, to (6.16) with $\mathbf{A} = \mathbf{X}_{\mathrm{f}}$, $\mathbf{B} = \mathbf{Y}_{\mathrm{f}}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}$ and $f(x) = (1+x)^{-\frac{1}{2}}$:

$$\begin{aligned}\mathbf{X}_{\mathrm{a}} &= \mathbf{X}_{\mathrm{f}}\left(\mathbf{I}_m + \mathbf{Y}_{\mathrm{f}}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{X}_{\mathrm{f}}\right)^{-\frac{1}{2}}\mathbf{U}\\ &= \left(\mathbf{I}_n + \mathbf{X}_{\mathrm{f}}\mathbf{X}_{\mathrm{f}}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right)^{-\frac{1}{2}}\mathbf{X}_{\mathrm{f}}\mathbf{U}\\ &= \left(\mathbf{I}_n + \mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right)^{-\frac{1}{2}}\mathbf{X}_{\mathrm{f}}\mathbf{U}. \end{aligned} \tag{6.18}$$

Hence, the transform matrix $\mathbf{T} = \left(\mathbf{I}_n + \mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right)^{-\frac{1}{2}}$ acts on the left of the anomaly matrix, i.e., in state space. It is often called the *left transform*. This condensed form of the transform matrix was first suggested by Sakov and Bertino [2011].

This left transform can also be rewritten so that it looks like the update of the stochastic EnKF but with a modified Kalman gain. To do so, we need another matrix lemma.

Lemma 6.2. *Assume that the scalar map $x \mapsto f(x)$ can be expanded as a formal power series: $f(x) = \sum_{k=0}^{\infty}\alpha_k x^k$ with, in addition, $f(0) = 1$. Define $g(x) = (f(x)-1)/x$. The formal power series of g is $g(x) = \sum_{k=0}^{\infty}\alpha_{k+1}x^k$. If $\mathbf{A}$ is a matrix of size $n\times p$ and if $\mathbf{B}$ is a matrix of size $p\times n$, then*

$$f(\mathbf{AB}) = \sum_{k=0}^{\infty}\alpha_k(\mathbf{AB})^k = \mathbf{I}_n + \mathbf{A}\left(\sum_{k=0}^{\infty}\alpha_{k+1}(\mathbf{BA})^k\right)\mathbf{B} = \mathbf{I}_n + \mathbf{A}g(\mathbf{BA})\mathbf{B}.$$

In particular, if $f(x) = \frac{1}{\sqrt{1+x}}$, then $g(x) = -\frac{1}{1+x+\sqrt{1+x}}$. When applied to $\mathbf{AB}$, one obtains [Bocquet. 2016]

$$(\mathbf{I}_n + \mathbf{AB})^{-\frac{1}{2}} = \mathbf{I}_n - \mathbf{A}\left(\mathbf{I}_p + \mathbf{BA} + \sqrt{\mathbf{I}_p + \mathbf{BA}}\right)^{-1}\mathbf{B}. \tag{6.19}$$

Choosing $\mathbf{A} = \mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}}$ and $\mathbf{B} = \mathbf{R}^{-1}\mathbf{H}$ in (6.19), we find that

$$\mathbf{X}_{\mathrm{a}} = \left(\mathbf{I}_n - \widetilde{\mathbf{K}}\mathbf{H}\right)\mathbf{X}_{\mathrm{f}}\mathbf{U}, \tag{6.20}$$

with

$$\widetilde{\mathbf{K}} = \mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}}\left(\mathbf{R} + \mathbf{H}\mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}} + \mathbf{R}\sqrt{\mathbf{I}_p + \mathbf{R}^{-1}\mathbf{H}\mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}}}\right)^{-1},$$

which was put forward in Whitaker and Hamill [2002] following Andrews [1968]. Note that this can also be written using the original Kalman gain,

$$\widetilde{\mathbf{K}} = \mathbf{K}\left(\mathbf{I}_p + \left[\mathbf{I}_p + \mathbf{H}\mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\right]^{-\frac{1}{2}}\right)^{-1}. \tag{6.21}$$

6.5 ▪ Localization and inflation

We have traded the EKF for a seemingly considerably cheaper filter meant to achieve similar performances. But this comes with some significant drawbacks. Fundamentally, one cannot hope to represent the full error covariance matrix of a complex high-dimensional system with only a few modes $m \ll n$, usually from a few dozen to a few hundred. This implies large *sampling errors*, meaning that the error covariance matrix is only sampled by a limited number of modes. This rank deficiency is accompanied by spurious correlations at long distances that strongly affect the filter performance. Even though the unstable degrees of freedom of dynamical systems that we wish to control with a filter are usually far fewer than the dimension of the system, they often still represent a nonnegligible fraction of the total degrees of freedom. Forecasting an ensemble of this size is usually not affordable.

The consequence of this issue is always the divergence of the filter. Hence, the EnKF is useful on condition that efficient fixes are applied. In other words, to make it a viable algorithm, one first needs to cope with the rank deficiency of the filter and with its manifestations, i.e., sampling errors.

Fortunately, there are clever tricks to overcome this major issue, known as *localization* and *inflation*, which explains, ultimately, the broad success of the EnKF in geosciences and engineering.

6.5.1 ▪ Localization

Localization relies on the idea that, for most geophysical systems, distant observables are weakly correlated. In other words, two distant parts of the system are almost independent, at least for short time scales. It is possible to exploit this relative independence and spatially localize the analysis [Houtekamer and Mitchell, 2001; Hamill et al., 2001; Evensen, 2003; Ott et al., 2004]. This has naturally been termed *localization*.

6.5.1.1 ▪ Domain localization

Broadly speaking, there are two main ways to implement this idea. The first one is called *domain localization*, or *local analysis*. Instead of performing a global analysis valid at any location in the domain, we perform a local analysis to update the local state variables using local observations. Typically, one would update a state variable at a location by assimilating only the observations within a fixed range of this point. If this range, the *localization length*, is too long, then the analysis becomes global, as before, and may fail because of the rank deficiency. On the other hand, if the localization length is too small, some short- to medium-range correlation might be neglected, resulting in a viable but less precise DA system.

One would then update all the state variables performing these local analyses. This may seem a formidable computational task, but all these analyses can be carried out

in parallel (again we refer to Hunt et al. [2007] for an analysis of the numerical complexity). Besides, since the number of local observations is limited, the computation of the local gain is much faster than that of the global gain. All in all, such an approach is viable even on very large systems. Figure 6.2 is a schematic representation of a local update in the domain localization approach where only the surrounding observations of a grid cell to update are assimilated.

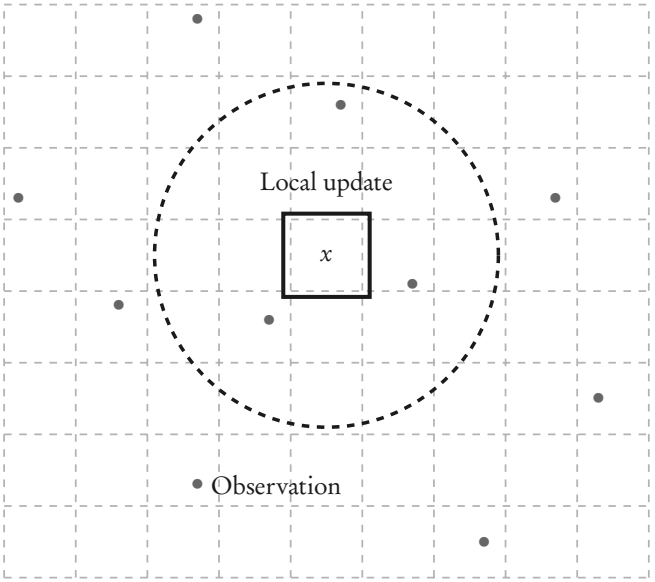

Figure 6.2. *Schematic representation of the local update of the variable defined over the framed grid cell with three surrounding observations (dots within the circle).*

The simplest strategy consists of selecting the observations within a disk around the center, $\mathbf{x}$, of the local analysis: the boxcar scheme. This is equivalent to making the variances and covariances of observations outside the domain go to infinity, i.e., replace, for each local analysis, $\mathbf{R}$ by a local $\mathbf{R}_{\mathbf{x}}$ restricted to the local observations.

Another analysis centered on a nearby point might share many observations but incorporate and exclude others, resulting in a discrete and not necessarily smooth transition between the two nearby analyses. To mitigate these transitions and reconstruction and obtain more consistent state and perturbation updates, it is possible to taper the inverse of the local observation error covariance matrix by a cutoff function that decreases with distance from $\mathbf{x}$. This way, the transition from one local domain to a closer one is smoother.

One possible cutoff function can be defined using the Gaspari–Cohn function [Gaspari and Cohn, 1999], which is a fifth-order piecewise rational function $G(r)$:

$$G(r) = \begin{cases} \text{if} \quad 0 \le r < 1: & 1 - \frac{5}{3}r^2 + \frac{5}{8}r^3 + \frac{1}{2}r^4 - \frac{1}{4}r^5, \\ \text{if} \quad 1 \le r < 2: & 4 - 5r + \frac{5}{3}r^2 + \frac{5}{8}r^3 - \frac{1}{2}r^4 + \frac{1}{12}r^5 - \frac{2}{3r}, \\ \text{if} \quad r \ge 2: & 0. \end{cases}$$

The cutoff function would be defined by $r \in \mathbb{R}^+ \mapsto G(r/c)$, where c is a length scale called the localization radius. It mimics a Gaussian distribution but vanishes beyond $r \ge 2c$, which is numerically efficient (G is compactly supported). This

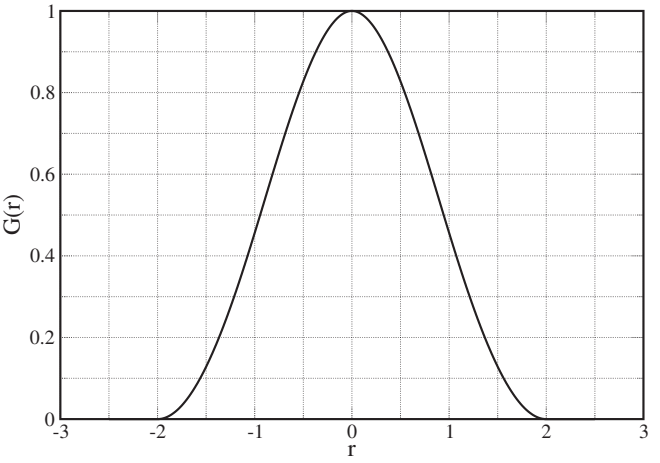

Figure 6.3. *Plot of the Gaspari–Cohn fifth-order piecewise rational function, which resembles a Gaussian distribution but has a compact support.*

approach is equivalent to tapering the ensemble members *and* the innovations by $r \mapsto \sqrt{G(r/c)}$, instead of modifying the local observation error covariances [Sakov and Bertino, 2011]. The Gaspari–Cohn cutoff function is represented in Figure 6.3.

6.5.1.2 ▪ Covariance localization

The second approach, called *covariance localization* or *Schur localization*, focuses on the forecast error covariance matrix. It is based on the remark that the forecast error covariance matrix $\mathbf{P}^{\mathrm{f}}$ is of low rank, at most $m-1$, and that this rank deficiency could be cured by filtering these empirical covariances. Physically, even though the empirical $\mathbf{P}^{\mathrm{f}}$ is probably a good approximation of the true error covariance matrix at short distances, the insufficient rank will induce long-range correlations. These correlations are *spurious*: they are likely to be nonphysical and are not present in the true forecast error covariance matrix. Hence, the idea is to regularize $\mathbf{P}^{\mathrm{f}}$ by smoothing out these long-range spurious correlations and, simultaneously, increasing the rank of $\mathbf{P}^{\mathrm{f}}$. To do so, one can regularize $\mathbf{P}^{\mathrm{f}}$ by multiplying it pointwise with a short-range predefined correlation matrix $\rho \in \mathbb{R}^{n\times n}$. The pointwise multiplication is called a Schur product and denoted by a $\circ$:

$$\left[\rho \circ \mathbf{P}^{\mathrm{f}}\right]_{i,j} = \left[\mathbf{P}^{\mathrm{f}}\right]_{i,j} [\rho]_{i,j}\,. \tag{6.22}$$

The Schur product theorem [Horn and Johnson, 2012] ensures that if ρ and $\mathbf{P}^{\mathrm{f}}$ are positive semidefinite (resp. positive definite), then the Schur product, $\rho \circ \mathbf{P}^{\mathrm{f}}$, is positive semidefinite (resp. positive definite). For instance, positive definite covariance matrices can be built with the Gaspari–Cohn function. This will ensure that the Schur product is positive semidefinite. As can be read from (6.22), the spurious correlations are leveled off by ρ if ρ is short ranged. This regularized $\rho \circ \mathbf{P}^{\mathrm{f}}$ will be used in the EnKF analysis as well as in the generation of the posterior ensemble of perturbations, as a replacement for $\mathbf{P}^{\mathrm{f}}$.

Figure 6.4 illustrates covariance localization. We consider a one-dimensional state space of $n = 200$ variables and a true error covariance matrix defined by $[\mathbf{P}]_{i,j} = e^{-|i-j|/L}$, where $i, j = 1, \ldots, n$ and $L = 10$ (correlation length). An ensemble of $m = 20$

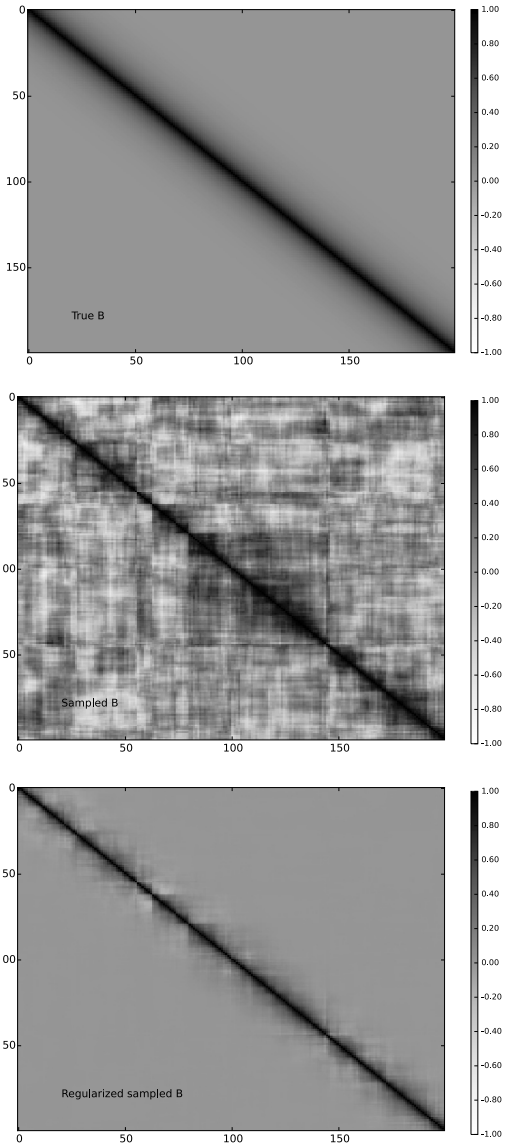

Figure 6.4. *Covariance localization. Upper panel: the true covariance matrix. Middle panel: the empirical covariance matrix from a* 20*-member ensemble. Bottom panel: the regularized empirical covariance matrix using Schur localization.*

members is generated from the exact Gaussian distribution of error covariance matrix **P**. The empirical error covariance matrix is computed from this ensemble, and a Schur localization is applied to the empirical covariance matrix to form a regularized covariance matrix. The Gaspari–Cohn correlation function is used to generate the regularizing correlation function. For the sake of illustration, we take c of the Gaspari–Cohn function to be the correlation length of the true covariance matrix, which does not necessarily guarantee an optimal localization.

6.5.1.3 ▪ Implementation of localization

For the stochastic EnKF, the main task is to compute the Kalman gain. The regularization of the sample forecast error covariance matrix should happen at this stage. Because localization will transform the sample $\mathbf{P}^{\mathrm{f}}$ into a full-rank $\rho \circ \mathbf{P}^{\mathrm{f}}$, the computation of $\left(\rho \circ \mathbf{P}^{\mathrm{f}}\right)\mathbf{H}^{\mathrm{T}}$, of the innovation statistics $\mathbf{R}+\mathbf{H}\left(\rho \circ \mathbf{P}^{\mathrm{f}}\right)\mathbf{H}^{\mathrm{T}}$, and of its inverse in the Kalman gain may become computationally unfavorable for high-dimensional systems. One can take benefit from an often significantly smaller number of observations compared to the number of control variables to reduce the complexity of the gain computation. Furthermore, assume that the observations are in situ. At the simplest, assume that observation i is located within the grid cell with index $l(i)$. Then

$$\left[\left(\rho \circ \mathbf{P}^{\mathrm{f}}\right)\mathbf{H}^{\mathrm{T}}\right]_{k,i} = [\rho]_{k,l(i)}\left[\mathbf{P}^{\mathrm{f}}\right]_{k,l(i)}[\mathbf{H}]_{i,l(i)} = \left[\rho \circ \left(\mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}}\right)\right]_{k,i},$$

which, in this case, explicitly extends the use of localization to observation space in direct correspondence with state space. Then, to estimate the Kalman gain, it often becomes much more efficient to compute instead

$$\left(\rho \circ \mathbf{P}^{\mathrm{f}}\right)\mathbf{H}^{\mathrm{T}} \longrightarrow \rho \circ (\mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}}), \qquad \mathbf{H}\left(\rho \circ \mathbf{P}^{\mathrm{f}}\right)\mathbf{H}^{\mathrm{T}} \longrightarrow \rho \circ (\mathbf{H}\mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}}).$$

Hence the Kalman gain could be estimated as [Houtekamer and Mitchell, 2001]

$$\mathbf{K} = \rho \circ (\mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}})\left[\mathbf{R}+\rho \circ (\mathbf{H}\mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}})\right]^{-1}.$$

The main drawback of localization is that it might also remove some physical and true long-distance correlations, thus inducing some physical imbalance in the analysis [Kepert, 2009; Greybush et al., 2011].

Covariance localization and localization by local analyses seem quite different. Indeed they are mathematically distinct, though they are both based on two sides of the same paradigm. However, it can be shown that they should yield very similar results if the innovation at each time step is not too strong [Sakov and Bertino, 2011], i.e., when the analysis remains close to the prior.

Because DA systems based on high-dimensional geophysical models very often yield inhomogeneous error fields, it would be useful to design localization schemes that are adaptive. They could depend on the local density of observations or on flow-dependent errors, etc. Preliminary statistical studies could be carried out to determine the optimal localization function for a given DA system in a given regime [Anderson and Lei, 2013]. The optimal localization function parameters could be determined on the fly from the ensemble only, using optimal linear filtering applied to the estimation of the regularized covariances [Ménétrier et al., 2015a,b].

The local ensemble transform Kalman filter (LETKF) has become an emblem of the success of the EnKF in conjunction with localization on significantly high-dimensional academic, but also operational, models. It was illustrated not only in several fields of the geosciences, such as atmosphere (including extraterrestrial atmospheres [Hoffman et al., 2010]), ocean, and solid earth sciences, but also in engineering, such as adaptive optics for future extra-large telescopes [Gray et al., 2014].

6.5.2 ▪ Inflation

Even when the analysis is made local, the error covariance matrices are still evaluated with an ensemble of limited size. This often leads to sampling errors and spurious

correlations. With a proper localization scheme, they might be significantly reduced. No matter how small the residual errors are, they will accumulate and they will carry over to the next cycles of the sequential EnKF scheme. As a consequence, there is always a risk that the filter may ultimately diverge. One way around this is to inflate the error covariance matrix by a factor λ^2 slightly greater than 1, before or after the analysis [Pham et al., 1998; Anderson and Anderson, 1999]. For instance, after the analysis,

$$\mathbf{P}^{\mathrm{a}} \longrightarrow \lambda^2 \mathbf{P}^{\mathrm{a}}.$$

Another way to achieve this is to inflate the ensemble,

$$\mathbf{x}_i^{\mathrm{a}} \longrightarrow \overline{\mathbf{x}}^{\mathrm{a}} + \lambda\left(\mathbf{x}_i^{\mathrm{a}} - \overline{\mathbf{x}}^{\mathrm{a}}\right),$$

which can alternatively be enforced on the prior (forecast) ensemble. This type of inflation is called *multiplicative inflation*.

In a perfect but nonlinear model context, the multiplicative inflation is meant to compensate for sampling errors that are the indirect consequence of the nonlinearity in the ensemble forecast [Bocquet, 2011; Whitaker and Hamill, 2012; Bocquet et al., 2015]. In this case, it helps cure an intrinsic source of error of the EnKF scheme.

Yet, inflation can also compensate for the underestimation of the errors due to the presence of unaccounted-for model error, a source of error external to the EnKF scheme. In this context, an additive type of inflation can also be used, either by

$$\mathbf{P}^{\mathrm{f}} \longrightarrow \mathbf{P}^{\mathrm{f}} + \mathbf{Q}$$

or by adding noise to the ensemble members,

$$\mathbf{x}_i^{\mathrm{f}} \longrightarrow \mathbf{x}_i^{\mathrm{f}} + \boldsymbol{\epsilon}_i \quad \text{with} \quad \mathrm{E}\left[\boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^{\mathrm{T}}\right] = \mathbf{Q}.$$

The need and magnitude of a proper inflation is very dependent on the system, on its dynamics, on the observation network, or on the particular EnKF variant. It can also be made variable in space. For instance, in the absence of observation and of an update in a specific area, there is no need to locally update the ensemble, so that no inflation is required, in contrast to densely observed regions.

Perhaps even more than localization, inflation is a trick, even though conveniently simple. There has been development to estimate the proper inflation required by the EnKF, often depending on a ratio of information content between the observation and the prior. Because it can be diagnosed from its environment, inflation can be made adaptive (in time and/or in space). Some information on the adaptive inflation can even be carried over from one cycle to the next. Efficient adaptive inflation methods have been proposed by, e.g., Wang and Bishop [2003], Anderson [2007], Li et al. [2009], Zheng [2009], Brankart et al. [2010], Bocquet [2011], Miyoshi [2011], Bocquet and Sakov [2012], Liang et al. [2012], and Ying and Zhang [2015] to account for sampling errors and possibly extrinsic model error.

6.6 ▪ Numerical illustrations with the Lorenz-95 model

The Lorenz-95 low-order model is a one-dimensional toy model introduced by the famous meteorologist Edward Lorenz in 1995 (later published in 1996 in the ECMWF's seminar proceedings and in 1998 as Lorenz and Emmanuel [1998]). It represents a

midlatitude zonal circle of the global atmosphere. It has $n = 40$ variables, $\{x_i\}_{i=1,\ldots,n}$. Its dynamics are given by the following set of ODEs:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = (x_{i+1} - x_{i-2})x_{i-1} - x_i + F$$

for $i = 1, \ldots, n$. The domain on which the 40 variables are defined is circle-like. Hence, it is assumed periodic, and the following definitions are enforced: $x_0 = x_{40}$, $x_{-1} = x_{39}$, $x_{41} = x_1$. The term $(x_{i+1} - x_{i-2})x_{i-1}$ is a nonlinear convective term. It preserves the energy $\sum_{i=1}^{n} x_i^2$. The term $-x_i$ is a dampening term, while F is an energy injection/extraction term depending on the sign of the variables. Here F is chosen to be 8. This makes the dynamics of this model chaotic. With this choice, the model has 13 positive Lyapunov exponents. This implies that, out of the 40 degrees of freedom of this model, 13 directions lead to growing modes. Moreover, one Lyapunov exponent is zero, which corresponds to a neutral mode. In practice and for short-term forecasts, roughly 13 directions (which change with time) out of the 40 in model state space make a small perturbation grow under the action of the model. A time step of 0.05 is meant to represent a time interval of 6 hours in the real atmosphere. Here, the model is integrated using a fourth-order Runge–Kutta scheme with a time step of $\delta t = 0.05$ because the scheme is precise and one can afford the numerical cost in this low-order context.

Figure 6.5 displays a trajectory of the model state. The model is characterized by about eight nonlinear waves that interact (this number depends on F). As can be seen, it seems difficult to predict the behavior of these waves except that they have the tendency to drift eastward in spite of a phase velocity shifting the peaks and lows westward. These waves are meant to represent large-scale Rossby waves.

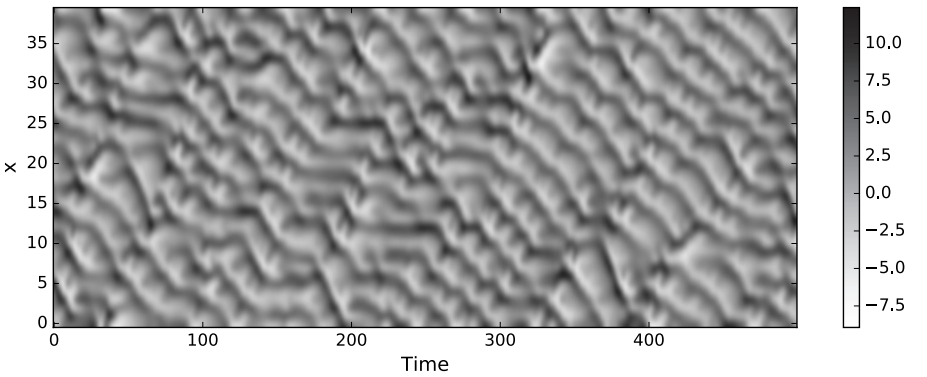

Figure 6.5. *Trajectory of a state of the Lorenz-95 model. The x coordinate represents time in units of* $\Delta t = 0.05$. *The y coordinate represents the* 40 *state variables.*

A twin experiment is conducted (see Section 6.3.4 for a definition). The truth is represented by a free model run, meant to be tracked by the DA system. The system is assumed to be fully observed ($p = 40$) every Δt, so that $\mathbf{H}_k \equiv \mathbf{I}_{40}$, with the observation error covariance matrix $\mathbf{R}_k \equiv \mathbf{I}_{40}$. The time interval between observational updates is set to $\Delta t = 0.05$, meant to be representative of a DA cycle of global meteorological models (6 hours), as we previously discussed. With such a value for Δt, the DA system is considered to be weakly nonlinear, yielding statistics of the errors weakly diverging from Gaussianity. The related synthetic observations are generated from the truth and perturbed by a Gaussian noise with the same distribution as the

observation error prior. In this experiment, the performance of a scheme is measured by the temporal mean of a root mean square difference between a state estimate ($\mathbf{x}^{\mathrm{a}}$) and the truth ($\mathbf{x}^{\mathrm{t}}$). Typically, one averages over time the following analysis root mean square error (RMSE):

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i^{\mathrm{a}}-x_i^{\mathrm{t}}\right)^2}. \tag{6.23}$$

All DA runs will extend over 10^5 cycles, after a burn-in period of 5×10^3 cycles, which guarantees satisfying convergence of the error statistics, due to ergodicity.

We vary the ensemble size from $m = 5$ to $m = 50$ and compare the performance in terms of RMSE of

- the EnKF without inflation and without localization,
- the EnKF with optimally tuned inflation and without localization,
- the EnKF without inflation and with optimally tuned localization, and
- the EnKF with optimally tuned inflation and with optimally tuned localization.

Optimally tuned means that the selected parameterization corresponds to the best RMSE. The EnKF variant is chosen to be the ETKF. The average RMSEs of the analyses over the long run are displayed in Figure 6.6.

If one agrees that the application of the EnKF to this low-order model captures several of the difficulties of realistic DA, it becomes clear from these numerical results that localization and inflation are indeed mandatory ingredients of a satisfying implementation of the EnKF. In particular, localization cannot be avoided with an ensemble size smaller than about 15, which is about the size of the unstable and neutral model subspace (equal to the number of nonnegative Lyapunov exponents).

6.7 ▪ Other important flavors of the EnKF

6.7.1 ▪ Variants of the EnKF

6.7.1.1 ▪ Ensemble adjustment Kalman filter

The ensemble adjustment Kalman filter (EAKF) [Anderson, 2001] is an EnSRKF where the ensemble update is computed by left-multiplication of the prior perturbations. Because the EAKF was developed quite early, the update formula's expression is not as compact as in (6.18) or (6.20), but the crucial ingredients are in it.

6.7.1.2 ▪ Serial EnKF

The serial EnKF is an EnKF where the observations are assimilated serially, i.e., one by one. Consider the scalar observation y, of error variance r, and the related observation operator $\mathbf{h} : \mathbb{R}^n \mapsto \mathbb{R}$. Then the analysis update is $\mathbf{x}^{\mathrm{a}} = \mathbf{x}^{\mathrm{f}} + \mathbf{K}\left[y - h(\mathbf{x}^{\mathrm{f}})\right]$, where the simplified Kalman gain of size $n \times 1$ is a vector that reads

$$\mathbf{K} = \mathbf{P}^{\mathrm{f}}\mathbf{h}^{\mathrm{T}}/(r + \mathbf{h}\mathbf{P}^{\mathrm{f}}\mathbf{h}^{\mathrm{T}}),$$

where the tangent linear, $\mathbf{h}$, of the observation operator is a row vector in $\mathbb{R}^n$. The modified gain, $\widetilde{\mathbf{K}}$, of the ensemble square root approach simplifies to

$$\widetilde{\mathbf{K}} = \frac{1}{1+1/\sqrt{1+r^{-1}\mathbf{h}\mathbf{P}^{\mathrm{f}}\mathbf{h}^{\mathrm{T}}}}\mathbf{K}.$$

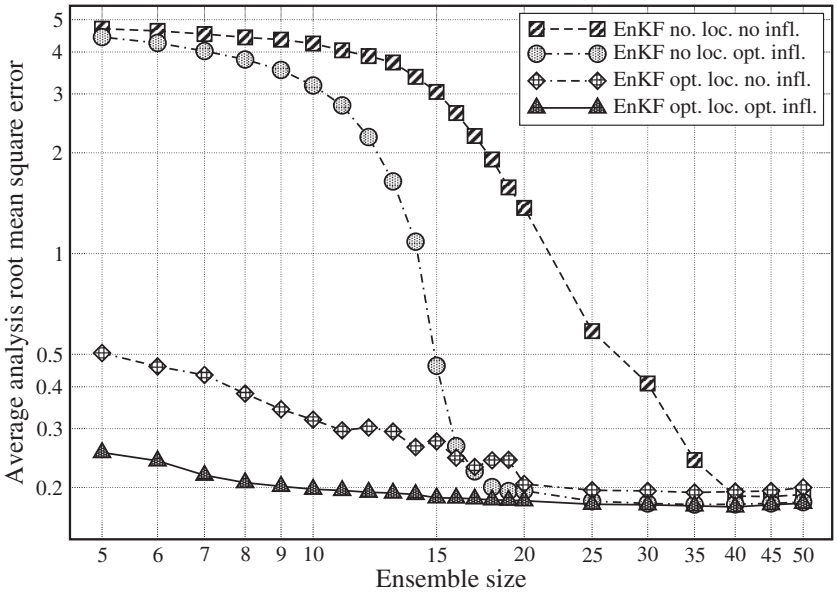

Figure 6.6. *Average analysis RMSE for a deterministic EnKF (ETKF) without localization and without inflation (boxes); without localization and with optimally tuned inflation (disks); with optimally tuned localization and no inflation (diamonds); with optimally tuned localization and with optimally tuned inflation (triangles).*

The correction due to the ensemble square root (Potter) scheme is conveniently reduced to a scalar.

If the ensemble is updated after each observation assimilation, then the serial EnKF should be mathematically equivalent to the *batch* EnKF, i.e., based on matrix update and the assimilation of observation vectors. Yet, this is only true for diagonal error covariance matrices.

Localization in the serial EnKF is elegantly simple. Focusing on scalar variables as the serial EnKF does, the covariance between the state variable x_i in grid cell i and the observation y, assumed to be local and located in grid cell j, is a scalar (a regression coefficient), which must be multiplied by the tapering correlation function $\rho_{i,j}$ for regularization. The Kalman gain of the stochastic EnKF is changed using

$$\mathbf{P}^{\mathrm{f}}\mathbf{h}^{\mathrm{T}} \longrightarrow \rho \circ \mathbf{P}^{\mathrm{f}}\mathbf{h}^{\mathrm{T}} \simeq \rho \circ \left(\mathbf{P}^{\mathrm{f}}\mathbf{h}^{\mathrm{T}}\right)$$

and using the approximation discussed in Section 6.5.1. Remarkably, $\mathbf{h}\mathbf{P}^{\mathrm{f}}\mathbf{h}^{\mathrm{T}}$ is unchanged, since $\rho_{i,i} = 1$. Moreover, the correction of the modified gain, $\widetilde{\mathbf{K}}$, of the ensemble square root approach is also unchanged for the same reason.

Finally, note that the equivalence between serial and batch assimilation in the EnKF is broken when localization is used.

6.7.1.3 ▪ DEnKF

We use the term *deterministic EnKF* to generally name the ensemble square root update scheme that avoids the need of stochastic sampling. Sakov and Oke [2008b] proposed a simplification of the EnSRKF, which they called the *deterministic EnKf or DEnKF*, because it astutely mimics the stochastic EnKF while being deterministic.

In the DEnKF, the state update is the same as in all previously discussed EnKF schemes. The perturbation update scheme, however, is an approximate but elegantly simple square root update scheme. Let us consider (6.21). Using the natural order on the positive definite matrices, we have

$$\mathbf{I}_p + \mathbf{H}\mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1} \geq \mathbf{I}_p,$$

so that

$$\left(\mathbf{I}_p + \left[\mathbf{I}_p + \mathbf{H}\mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\right]^{-\frac{1}{2}}\right)^{-1} \geq \frac{1}{2}\mathbf{I}_p. \tag{6.24}$$

The gain defined by $\widehat{\mathbf{K}} \equiv \frac{1}{2}\mathbf{K}$ could be used for the perturbation update as an approximation of the theoretical optimal gain $\mathbf{K}$. It is clear from (6.24) that the weaker this assimilation, i.e., $\mathbf{R} \gg \mathbf{H}\mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}}$, the more precise the approximation. Hence, the update formula of the DEnKF is the square root–free

$$\mathbf{X}_{\mathrm{a}} \simeq \left(\mathbf{I}_n - \frac{1}{2}\mathbf{K}\mathbf{H}\right)\mathbf{X}_{\mathrm{f}}. \tag{6.25}$$

This scheme avoids the need to compute any square root, making it very similar to the original stochastic EnKF but with a very simple modification. It is computationally very competitive. Using the approximate gain, $\widehat{\mathbf{K}}$, instead of the deterministic gain, $\widetilde{\mathbf{K}}$, defined by (6.21) impacts the posterior error covariances, making the scheme a suboptimal but cautious approximation; it can be shown [Sakov and Oke, 2008b] that the expected posterior error covariances of the DEnKF are

$$\widehat{\mathbf{P}}_{\mathrm{a}} = (\mathbf{I}_n - \mathbf{K}\mathbf{H})\,\mathbf{P}^{\mathrm{f}} + \frac{1}{4}\mathbf{K}\mathbf{H}\mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}}\mathbf{K}^{\mathrm{T}}.$$

The first term, $(\mathbf{I}_n - \mathbf{K}\mathbf{H})\,\mathbf{P}^{\mathrm{f}}$, is the EnKF optimal $\mathbf{P}^{\mathrm{a}}$. The extra term, which adds to the posterior covariances, is positive semidefinite. Hence $\widehat{\mathbf{P}}_{\mathrm{a}} \geq \mathbf{P}^{\mathrm{a}}$. We conclude that choosing $\widehat{\mathbf{K}}$ instead of $\widetilde{\mathbf{K}}$ impacts the posterior error covariances, making the scheme suboptimal but a safe approximation (the main risk being to underestimate the errors).

The state and perturbations scheme can even be reformulated as

$$\mathbf{x}_i^{\mathrm{a}} = \mathbf{x}_i^{\mathrm{f}} + \mathbf{K}\left[\mathbf{y} - \mathscr{H}\left(\frac{\mathbf{x}_i^{\mathrm{f}} + \overline{\mathbf{x}}^{\mathrm{f}}}{2}\right)\right],$$

to be compared to the stochastic EnKF update.

6.7.1.4 ▪ Mollified EnKF

One of the conceptual limitations of DA is the standpoint adopted earlier of the discretized representation of fields. This is obviously justified by the need for numerical implementation of the models and control variables. However, time continuous representations are often more satisfying from the conceptual standpoint, by ignoring implementation details. One interesting standpoint on the EnKF update step is offered by the mollified EnKF [Bergemann and Reich, 2010]. Instead of the algebraic update of the ensemble members as in the stochastic EnKF, or of the mean and the perturbations in the DEnKF, it is tempting to represent the update as the integration of ODEs (or possibly PDEs) from the prior to the posterior over a fictitious interval

of time [Simon, 2006; Bergemann and Reich, 2010]. One would update the ensemble members via

$$\frac{\mathrm{d}\mathbf{x}_i(s)}{\mathrm{d}s} = \frac{1}{2}\mathbf{P}(s)\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\left[2\mathbf{y} - \mathbf{H}\{\mathbf{x}_i(s) + \overline{\mathbf{x}}(s)\}\right], \tag{6.26}$$

where $\mathbf{x}_i$ is integrated from $s = 0$ to $s = 1$, with the initial condition $\mathbf{x}_i(0) = \mathbf{x}_i^{\mathrm{f}}$. Then, it can be shown that $\mathbf{x}_i(1) = \mathbf{x}_i^{\mathrm{a}}$. We note the clear resemblance with the DEnKF. This shows that the DEnKF update represents a linearized integration of (6.26), i.e., maintaining the value of $\mathbf{P}(s)$ and $\overline{\mathbf{x}}(s)$ at $\mathbf{P}^{\mathrm{f}}$ and $\overline{\mathbf{x}}^{\mathrm{f}}$ over the interval.

6.7.1.5 ▪ Assimilation in the unstable subspace

Let us consider a minimalist DA problem (already encountered, in various forms, in the first three chapters) with a one-variable, perfect, linear model $x_{k+1} = \alpha x_k$, with k the time index. If $\alpha^2 > 1$, the model is unstable, and if $\alpha^2 < 1$ it is stable. Let us denote by b_k the forecast/prior error variance, r the static observation error variance, and a_k the error analysis variance. Sequential DA implies the following recursions for the variances:

$$a_k^{-1} = b_k^{-1} + r^{-1} \quad \text{and} \quad b_{k+1} = \alpha^2 a_k,$$

whose asymptotic solution ($a_k \to a_\infty$) is

$$a_\infty = 0 \text{ if } \alpha^2 < 1 \quad \text{and} \quad a_\infty = (1 - 1/\alpha^2)r \text{ if } \alpha^2 \geq 1.$$

Very roughly, it tells us that only the growing modes need to be controlled, i.e., that DA should be targeted at preventing errors to increase indefinitely in the space generated by the growing modes. This paradigm is called *assimilation in the unstable space* or *AUS* [see Palatella et al., 2013, and references therein]. It is tempting to identify the unstable subspace with the time-dependent space generated by the Lyapunov vectors with nonnegative exponents, which, strictly speaking, is the unstable and neutral subspace. Applied to the KF and possibly the EnKF, it is intuitively known that the error covariance matrix tends to collapse to this unstable and neutral subspace [Trevisan and Palatella, 2011]. This can be made rigorous in the linear model Gaussian statistics case. The generalization of the paradigm to nonlinear dynamical systems is more speculative, but Ng et al. [2011] and Palatella and Trevisan [2015] put forward some enlightening arguments about it. A connection between the AUS paradigm and a justification of multiplicative inflation was established in Bocquet et al. [2015].

6.7.2 ▪ Variational analysis

We recall one of the key elementary results of DA seen in Chapters 1 and 3, namely that when the observation model is linear, the BLUE analysis is equivalent to a 3D variational problem (3D-Var). That is to say, evaluating the matrix formula

$$\mathbf{x}^{\mathrm{a}} = \mathbf{x}^{\mathrm{f}} + \mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}}\left(\mathbf{R} + \mathbf{H}\mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}}\right)^{-1}\left(\mathbf{y} - \mathbf{H}\mathbf{x}^{\mathrm{f}}\right)$$

is equivalent to solving the minimization problem

$$\mathbf{x}^{\mathrm{a}} = \underset{\mathbf{x}}{\operatorname{argmin}}\, \mathscr{L}(\mathbf{x}) \quad \text{with} \quad \mathscr{L}(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{H}\mathbf{x}\|_{\mathbf{R}}^2 + \frac{1}{2}\left\|\mathbf{x} - \overline{\mathbf{x}}^{\mathrm{f}}\right\|_{\mathbf{P}^{\mathrm{f}}}^2,$$

where $||\mathbf{x}||_{\mathbf{A}}^2 = \mathbf{x}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{x}$ for any symmetric positive definite matrix $\mathbf{A}$. We assume momentarily that $\mathbf{P}^{\mathrm{f}}$ is full rank so that it is invertible and positive-definite. This equivalence can be fruitful with high-dimensional systems, where tools of numerical optimization can be used in place of linear algebra. This equivalence is also of theoretical interest because it enables an elegant generalization of the BLUE update to the case where the observation operator is nonlinear. Simply put, the cost function is now replaced with

$$\mathscr{L}_{\mathrm{NL}}(\mathbf{x}) = \frac{1}{2}||\mathbf{y} - \mathscr{H}(\mathbf{x})||_{\mathbf{R}}^2 + \frac{1}{2}\left\|\mathbf{x} - \overline{\mathbf{x}}^{\mathrm{f}}\right\|_{\mathbf{P}^{\mathrm{f}}}^2,$$

where $\mathscr{H}$ is nonlinear.

6.7.2.1 ▪ The maximum likelihood ensemble filter

This equivalence was put to use in the maximum likelihood ensemble filter (MLEF) introduced by Zupanski [2005]. To describe the MLEF in a framework we have already detailed, let us formulate it in terms of the ETKF (following Bocquet and Sakov [2013]). Hence, we write the analysis of the MLEF in ensemble subspace closely following Section 6.4. However, we must first write the corresponding cost function. Recall that the state vector is parameterized in terms of the vector of coefficients $\mathbf{w}$ in $\mathbb{R}^m$, $\mathbf{x} = \overline{\mathbf{x}}^{\mathrm{f}} + \mathbf{X}\mathbf{w}$. The reduced cost function is denoted by

$$\mathscr{J}_{\mathrm{NL}}(\mathbf{w}) = \mathscr{L}_{\mathrm{NL}}(\overline{\mathbf{x}}^{\mathrm{f}} + \mathbf{X}\mathbf{w}).$$

Its first term can easily be written in the reduced ensemble subspace:

$$\mathscr{J}_{\mathrm{NL}}^{\mathrm{o}}(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{y} - \mathscr{H}(\overline{\mathbf{x}}^{\mathrm{f}} + \mathbf{X}\mathbf{w})\right\|_{\mathbf{R}}^2.$$

To proceed with the background term, $\mathscr{J}^{\mathrm{b}}(\mathbf{w})$, of the cost function, we first have to explain what the inverse of $\mathbf{P}^{\mathrm{f}} = \mathbf{X}_{\mathrm{f}}\mathbf{X}_{\mathrm{f}}^{\mathrm{T}}$ of incomplete rank is whenever $m \leq n$. Because for most EnKFs, the analysis is entirely set in ensemble subspace as we repeatedly pointed out, the inverse, $\mathbf{P}^{\mathrm{f}} = \mathbf{X}_{\mathrm{f}}\mathbf{X}_{\mathrm{f}}^{\mathrm{T}}$, must be the Moore–Penrose inverse [Golub and van Loan, 2013] of $\mathbf{P}^{\mathrm{f}}$, denoted by $\mathbf{P}_{\mathrm{f}}^{\dagger}$. Indeed, it is defined in the range of $\mathbf{P}^{\mathrm{f}}$. It is even more direct to introduce the SVD of the perturbation matrix $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}}$, where $\Sigma > 0$ is the diagonal matrix of the $m' \leq m$ positive singular values, $\mathbf{V}$ is of size $m \times m'$ such that $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}_{m'}$, and $\mathbf{U}$ is of size $n \times m'$ such that $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_{m'}$. Note that $\mathbf{P}_{\mathrm{f}}^{\dagger} = \mathbf{U}\Sigma^{-2}\mathbf{U}^{\mathrm{T}}$. Then we have

$$\begin{aligned}\mathscr{J}^{\mathrm{b}}(\mathbf{w}) &= \frac{1}{2}||\mathbf{X}_{\mathrm{f}}\mathbf{w}||_{\mathbf{P}^{\mathrm{f}}}^2 = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{X}_{\mathrm{f}}^{\mathrm{T}}\mathbf{P}_{\mathrm{f}}^{\dagger}\mathbf{X}_{\mathrm{f}}\mathbf{w} = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{V}\Sigma\mathbf{U}^{\mathrm{T}}\left(\mathbf{U}\Sigma^2\mathbf{U}^{\mathrm{T}}\right)^{\dagger}\mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}}\mathbf{w} \\ &= \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{V}\Sigma\mathbf{U}^{\mathrm{T}}\mathbf{U}\Sigma^{-2}\mathbf{U}^{\mathrm{T}}\mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}}\mathbf{w} = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{V}^{\mathrm{T}}\mathbf{w}.\end{aligned}$$

As was mentioned earlier, there is a freedom in $\mathbf{w}$ that makes the solution of the minimization problem degenerate. Clearly $\mathscr{J}_{\mathrm{NL}}^{\mathrm{o}}(\mathbf{w})$ is unchanged if $\mathbf{w}$ is shifted by $\lambda\mathbf{1}$. So is $\mathscr{J}^{\mathrm{b}}(\mathbf{w})$ because $\mathbf{1}$ is in the null space of $\mathbf{V}$ since $\mathbf{X}\mathbf{1} = 0$. One way to lift the degeneracy of the variational problem is to add a *gauge-fixing* term that will constrain the solution in the null space of $\mathbf{X}$, or $\mathbf{V}$. In practice, one can add

$$\mathscr{J}^{\mathrm{g}}(\mathbf{w}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}(\mathbf{I}_m - \mathbf{V}\mathbf{V}^{\mathrm{T}})\mathbf{w}$$

to the cost function $\mathscr{J}_{\mathrm{NL}}$ to obtain the regularized cost function

$$\mathscr{J}(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{y} - \mathscr{H}(\overline{\mathbf{x}}^{\mathrm{f}} + \mathbf{X}\mathbf{w})\right\|_{\mathbf{R}}^2 + \frac{1}{2}||\mathbf{w}||^2,$$

where $||\mathbf{w}||^2 = \mathbf{w}^{\mathrm{T}}\mathbf{w}$. The cost function still has the same minimum but it is achieved at a non-degenerate $\mathbf{w}^\star$ such that $\left(\mathbf{I}_m - \mathbf{V}\mathbf{V}^{\mathrm{T}}\right)\mathbf{w}^\star = \mathbf{0}$. This is the cost function of the ETKF [Hunt et al., 2007] but with a nonlinear observation operator.

The update step of this EnKF can now be seen as a nonlinear variational problem, which can be solved using a variety of iterative methods, such as a Gauss–Newton method, a quasi-Newton method, of a Levenberg–Marquardt method [Nocedal and Wright, 2006]. For instance, with the Gauss–Newton method, we would define the iterate, the gradient, and an approximation of the Hessian as

$$\begin{aligned}
\mathbf{x}^{(j)} &= \overline{\mathbf{x}}^{\mathrm{f}} + \mathbf{X}_{\mathrm{f}}\mathbf{w}^{(j)},\\
\nabla\mathscr{J}_{(j)} &= -\mathbf{Y}_{(j)}^{\mathrm{T}}\mathbf{R}^{-1}\left(\mathbf{y} - \mathscr{H}(\mathbf{x}^{(j)})\right) + \mathbf{w}^{(j)},\\
\mathrm{H}_{(j)} &= \mathbf{I}_m + \mathbf{Y}_{(j)}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y}_{(j)},
\end{aligned}$$

respectively. The Gauss–Newton iterations are indexed by j and given by

$$\mathbf{w}^{(j+1)} = \mathbf{w}^{(j)} - \mathrm{H}_{(j)}^{-1}\nabla\mathscr{J}_{(j)}.$$

The scheme is iterated until a satisfying convergence is reached, for instance when the norm of $\left\|\mathbf{w}^{(j+1)} - \mathbf{w}^{(j)}\right\|$ crosses below a given threshold. The vector $\mathbf{Y}_{(j)}$ is defined as the image of the initial ensemble perturbations $\mathbf{X}_{\mathrm{f}}$ through the tangent linear model of the observation operator computed at $\mathbf{x}^{(j)}$ by $\mathbf{Y}_{(j)} = \mathbf{H}_{|\mathbf{x}^{(j)}}\mathbf{X}_{\mathrm{f}}$. Following Sakov et al. [2012], there are at least two ways to compute these sensitivities. One explicitly mimics the tangent linear by a downscaling of the perturbations by ε such that $0 < \varepsilon \ll 1$ before application of the full nonlinear operator $\mathscr{H}$ followed by an upscaling by ε^{-1}. The operation reads

$$\mathbf{Y}_{(j)} \approx \frac{1}{\varepsilon}\mathscr{H}\left(\mathbf{x}^{(j)}\mathbf{1}^{\mathrm{T}} + \varepsilon\mathbf{X}_{\mathrm{f}}\right)\left(\mathbf{I}_m - \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{m}\right).$$

Note that ε accounts for a normalization factor of $\sqrt{m-1}$. The second way consists of avoiding resizing the perturbations because this implies applying the observation operator to the ensemble an extra time. Instead of downscaling the perturbations, we can (i) generate transformed perturbations by applying the right-multiplication operator

$$\mathbf{T} = \left[\mathbf{I}_m + \mathbf{Y}_{(j)}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y}_{(j)}\right]^{-\frac{1}{2}},$$

(ii) build a new ensemble from these transformed perturbations around $\mathbf{x}^{(j)}$, (iii) apply $\mathscr{H}$ to this ensemble, and finally (iv) rotate back the new perturbations around $\mathbf{x}^{(j+1)}$ by applying $\mathbf{T}^{-1}$. Through the $\mathbf{T}$-transformation the second scheme also ensures a resizing of the perturbations where $\mathscr{H}$ is in a close-to-linear regime. However, as opposed to the first scheme, the last propagation of the perturbation can be used to directly estimate the final approximation of the Hessian and the final updated set of perturbations, which can be numerically efficient.

Algorithm 6.5 Pseudocode for a complete cycle of the MLEF, as a variant in ensemble subspace following Zupanski [2005], Carrassi et al. [2009], Sakov et al. [2012], and Bocquet and Sakov [2013]

Require: Observation operator $\mathcal{H}$ at current time; algorithm parameters: e, $j_{\max}$, ε; $\mathbf{E}$, the prior ensemble; $\mathbf{y}$, the observation at current time; $\mathbf{U}$, an orthogonal matrix in $\mathbb{R}^{m\times m}$ satisfying $\mathbf{U1} = \mathbf{1}$; $\mathcal{M}$, the model resolvent from current time to the next analysis time.

$\overline{\mathbf{x}} = \mathbf{E1}/m$
$\mathbf{X} = \left(\mathbf{E} - \overline{\mathbf{x}}\mathbf{1}^{\mathrm{T}}\right)/\sqrt{m-1}$
$\mathbf{T} = \mathbf{I}_m$
$j = 0$, $\mathbf{w} = 0$
repeat
  $\mathbf{x} = \overline{\mathbf{x}} + \mathbf{Xw}$
  Bundle: $\mathbf{E} = \mathbf{x}\mathbf{1}^{\mathrm{T}} + \varepsilon\mathbf{X}$
  Transform: $\mathbf{E} = \mathbf{x}\mathbf{1}^{\mathrm{T}} + \sqrt{m-1}\mathbf{XT}$
  $\mathbf{Z} = \mathcal{H}(\mathbf{E})$
  $\overline{\mathbf{y}} = \mathbf{Z1}/m$
  Bundle: $\mathbf{Y} = \left(\mathbf{Z} - \overline{\mathbf{y}}\mathbf{1}^{\mathrm{T}}\right)/\varepsilon$
  Transform: $\mathbf{Y} = \left(\mathbf{Z} - \overline{\mathbf{y}}\mathbf{1}^{\mathrm{T}}\right)\mathbf{T}^{-1}/\sqrt{m-1}$
  $\nabla\mathcal{J} = \mathbf{w} - \mathbf{Y}^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y} - \overline{\mathbf{y}})$
  $\mathrm{H} = \mathbf{I}_m + \mathbf{Y}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y}$
  Solve $\mathrm{H}\Delta\mathbf{w} = \nabla\mathcal{J}$
  $\mathbf{w} := \mathbf{w} - \Delta\mathbf{w}$
  Transform: $\mathbf{T} = \mathrm{H}^{-\frac{1}{2}}$
  $j := j + 1$
until $||\Delta\mathbf{w}|| \le e$ **or** $j \ge j_{\max}$
Bundle: $\mathbf{T} = \mathrm{H}^{-\frac{1}{2}}$
$\mathbf{E} = \mathbf{x}\mathbf{1}^{\mathrm{T}} + \sqrt{m-1}\mathbf{XTU}$
$\mathbf{E} = \mathcal{M}(\mathbf{E})$

Note that each iteration amounts to solving an inner loop problem with the quadratic cost function

$$\mathcal{J}^{(j)}(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{y} - \mathcal{H}(\mathbf{x}^{(j)}) - \mathbf{Y}_{(j)}(\mathbf{w} - \mathbf{w}^{(j)})\right\|_{\mathbf{R}}^2 + \frac{1}{2}\left\|\mathbf{w} - \mathbf{w}^{(j)}\right\|^2.$$

The update of the perturbations follows that of the ETKF, i.e., equation (6.17). A full cycle of the algorithm is given in Algorithm 6.5 as a pseudocode. Either the bundle (finite differences with the ε-rescaling) scheme or the transform (using $\mathbf{T}$) scheme is needed to compute the sensitivities. Both are indicated in the algorithm. Inflation, possibly localization, should be added to the scheme to make it functional. In summary, the MLEF is an EnKF scheme that can use nonlinear observation operators in a consistent way using a variational analysis.

6.7.2.2 ▪ Numerical illustration

The (bundle) MLEF as implemented by Algorithm 6.5 is tested against the EnKF (ETKF implementation) using a setup similar to that of the Lorenz-95 model. To exhibit a difference of performance, the observation operator has been chosen to be

nonlinear. Each of the 40 variables is observed with the nonlinear observation operator

$$\mathcal{H}(\mathbf{x}) = \frac{\mathbf{x}}{2}\left\{1 + \left(\frac{|\mathbf{x}|}{10}\right)^{\gamma-1}\right\}, \tag{6.27}$$

where $|\mathbf{x}|$ is the componentwise absolute value of $\mathbf{x}$. The second nonlinear term in the brackets is meant to be of the order of magnitude of the first linear term to avoid numerical overflow. Obviously, γ tunes the nonlinearity of the observation operator, with $\gamma = 1$ corresponding to the linear case $\mathcal{H}(\mathbf{x}) = \mathbf{x}$. The prior observation error is chosen to be $\mathbf{R} \equiv \mathbf{I}_p$. The ensemble size is $m = 20$, which, in this context, makes localization unnecessary. For both the ETKF and the MLEF, the need for inflation is addressed either by using the Bayesian hierarchical scheme for the EnKF, known as the finite-size EnKF or EnKF-N, which we shall describe later, or by optimally tuning a uniform inflation (which comes with a significant numerical cost).

The MLEF is expected to offer strong performance in the first cycles of the DA scheme when the spread of the ensemble is large enough, over a span where the tangent linear observation model is not a good approximation. To measure this performance, the length of the DA run is set to 10^2 cycles and these runs are repeated 10^3 times, over which a mean analysis RMSE is computed. The spread of the initial ensemble is chosen to be 3, i.e., roughly the climatological variability of a single Lorenz-95 variable.

The overall performances of the schemes are computed as a function of γ, i.e., the nonlinearity strength of the observation operator, and reported in Figure 6.7.

Since the model is fully observed, an efficient DA scheme should have an RMSE smaller than the prior observation error, i.e., 1, which is indeed the case for all RMSEs computed in these experiments. As γ departs from $\gamma = 1$, the performance of

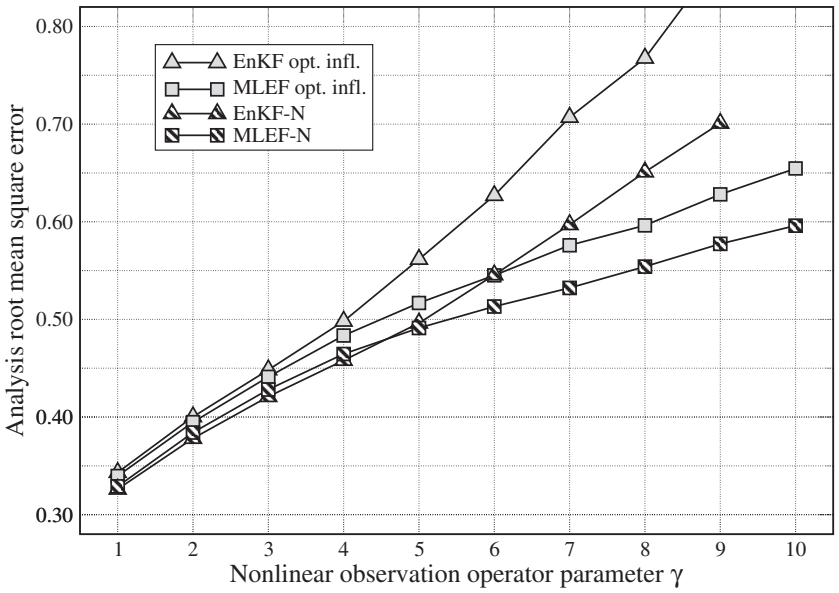

Figure 6.7. *Average analysis RMSE of a deterministic EnKF (ETKF) and of the MLEF with the Lorenz-95 model and the nonlinear observation operator equation* (6.27)*, as a function of* γ *the nonlinearity strength in the observation operator. For each RMSE,* 10^3 *DA experiments are run over* 10^2 *cycles. The final RMSE is the mean over those* 10^3 *experiments. The EnKF-N and MLEF-N are hierarchical filter counterparts to the EnKF and MLEF that will be discussed in Section* 6.7.3.

both filters consistently degrades. Beyond $\gamma > 9$, the EnKF diverges from the truth. However, the MLEF better handles the model nonlinearity, especially beyond $\gamma > 4$ and the gap between the EnKF and the MLEF increases with γ. We also note that the EnKF-N and MLEF-N offer better performance than an optimal but uniformly tuned EnKF. This is explained by the fact that the finite-size scheme is adaptive and adjusts the inflation as the ensemble spread decreases.

For much longer runs, the RMSE gap between the EnKF and the MLEF decreases significantly. Indeed, in the permanent regime of these experiments, the spread of the ensemble is significantly smaller than 1 and the system is closer to linearity, a regime where both the EnKF and the MLEF perform equally well. The gap between the finite-size and the optimally tuned ensemble filters also decreases since the adaptation feature is not necessarily important in a permanent regime.

The MLEF will be generalized later by including not only the observation operator but also the forecast model in the analysis over a 4D DA window, leading to the IEnKF and IEnKS.

6.7.2.3 ▪ The α control variable trick

The analysis using covariance localization with a localization matrix ρ can be equivalently formulated in terms of ensemble coefficients similar to the $\mathbf{w}$ of the ETKF, but that have been made space dependent. Instead of a vector $\mathbf{w}$ of size m that parameterizes the state vector $\mathbf{x}$,

$$\mathbf{x} = \overline{\mathbf{x}} + \sum_{i=1}^{m} w_i \frac{\mathbf{x}_i - \overline{\mathbf{x}}}{\sqrt{m-1}},$$

we choose a much larger control vector $\boldsymbol{\alpha}$ of size mn that parameterizes the state vector $\mathbf{x}$ as

$$\mathbf{x} = \overline{\mathbf{x}} + \sum_{i=1}^{m} \boldsymbol{\alpha}_i \circ \frac{\mathbf{x}_i - \overline{\mathbf{x}}}{\sqrt{m-1}},$$

where $\boldsymbol{\alpha}_i$ is a subvector of size n. This change of control variable is known as the $\boldsymbol{\alpha}$ control variable [Lorenc, 2003; Buehner, 2005; Wang et al., 2007]. It is better formulated in a variational setting. Consider the cost function

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}) = \frac{1}{2} \left\| \mathbf{y} - \mathcal{H} \left\{ \overline{\mathbf{x}} + \sum_{i=1}^{m} \boldsymbol{\alpha}_i \circ \frac{\mathbf{x}_i - \overline{\mathbf{x}}}{\sqrt{m-1}} \right\} \right\|_{\mathbf{R}}^2 + \frac{1}{2} \sum_{i=1}^{m} \|\boldsymbol{\alpha}_i\|_{\rho}^2, \tag{6.28}$$

which is formally equivalent to the Lagrangian,

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2} \|\mathbf{y} - \mathcal{H}(\mathbf{x})\|_{\mathbf{R}}^2 + \frac{1}{2} \sum_{i=1}^{m} \|\boldsymbol{\alpha}_i\|_{\rho}^2 + \boldsymbol{\beta}^{\mathrm{T}} \left\{ \mathbf{x} - \overline{\mathbf{x}} - \sum_{i=1}^{m} \boldsymbol{\alpha}_i \circ \frac{\mathbf{x}_i - \overline{\mathbf{x}}}{\sqrt{m-1}} \right\},$$

where $\boldsymbol{\beta}$ is a vector in $\mathbb{R}^n$ of Lagrange multipliers. Yet, as opposed to $\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha})$, the Lagrangian $\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is quadratic in $\boldsymbol{\alpha}$ and can be analytically minimized on $\boldsymbol{\alpha}$. The saddle point condition on $\boldsymbol{\alpha}_i$ is, denoting as before $\mathbf{X}_i = \frac{\mathbf{x}_i - \overline{\mathbf{x}}}{\sqrt{m-1}}$,

$$\boldsymbol{\alpha}_i = \rho \boldsymbol{\beta} \circ \mathbf{X}_i.$$

Its substitution in $\mathscr{L}(\mathbf{x}, \alpha, \beta)$ implies computing

$$\sum_{i=1}^{m} (\beta \circ \mathbf{X}_i)^{\mathrm{T}} \rho (\beta \circ \mathbf{X}_i) = \sum_{i=1}^{m} \sum_{k,l=1}^{n} \beta_k [\mathbf{X}_i]_k \rho_{k,l} \beta_l [\mathbf{X}_j]_l$$
$$= \sum_{k,l=1}^{n} \beta_k \left[\rho \circ \mathbf{X}\mathbf{X}^{\mathrm{T}} \right]_{k,l} \beta_l = \beta^{\mathrm{T}} \rho \circ \mathbf{P} \beta,$$

where $\mathbf{P}$ is the sample covariance matrix $\mathbf{P} = \frac{1}{m-1} \sum_{i=1}^{m} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}}$. This yields the new Lagrangian

$$\mathscr{L}(\mathbf{x}, \beta) = \frac{1}{2} ||\mathbf{y} - \mathscr{H}(\mathbf{x})||_{\mathbf{R}}^2 + \frac{1}{2} \beta^{\mathrm{T}} (\rho \circ \mathbf{P}) \beta + \beta^{\mathrm{T}} (\mathbf{x} - \overline{\mathbf{x}}).$$

This can be further optimized on β to yield the cost function

$$\mathscr{L}(\mathbf{x}) = \frac{1}{2} ||\mathbf{y} - \mathscr{H}(\mathbf{x})||_{\mathbf{R}}^2 + \frac{1}{2} ||\mathbf{x} - \overline{\mathbf{x}}||_{\rho \circ \mathbf{P}}^2 . \tag{6.29}$$

Hence we have obtained a formal equivalence between (6.28) and the cost function (6.29) that governs the analysis of the EnKF with covariance localization. Both approaches are used in the literature.

6.7.3 ▪ Hierarchical EnKFs

Key assumptions of the EnKF are that the mean and the error covariance matrix are exactly given by the sampled moments of the ensemble. As we have seen, this is bound to fail without fixes. Indeed, the uncertainty is underestimated and may often lead to the divergence of the filter. As seen in Section 6.5, inflation and localization are fixes that address this issue in a satisfying manner.

The use of a *Bayesian statistical hierarchy* has been more recently explored as a distinct route toward solving this issue. If $\mathbf{x}$ is a state vector that depends on parameters θ and is observed by $\mathbf{y}$, Bayes' rule (see section 3.2) tells us that the probability of the variables and parameters conditioned on the observations is

$$p(\mathbf{x}, \theta | \mathbf{y}) \propto p(\mathbf{y} | \theta, \mathbf{x}) p(\mathbf{x}, \theta),$$

where the proportionality factor only depends on $\mathbf{y}$. But if θ is further assumed to be an uncertain parameter vector obeying a prior distribution, $p(\theta)$, and if the likelihood of $\mathbf{y}$ only depends on θ through $\mathbf{x}$, one can further decompose this posterior distribution into

$$p(\mathbf{x}, \theta | \mathbf{y}) \propto p(\mathbf{y} | \mathbf{x}) p(\mathbf{x} | \theta) p(\theta).$$

We have created a hierarchy of random variables: $\mathbf{x}$ at the first level and θ at a second level. In this context, $p(\mathbf{x}|\theta)$ is still called a prior, while $p(\theta)$ is termed a *hyperprior* to emphasize that it operates at a second level of the Bayesian hierarchy [Gelman et al., 2014]. Applied to the EnKF, a Bayesian hierarchy could be enforced by seeing the moments of the true error distribution as multivariate random variables, rather than coinciding with the sampled moments of the ensemble.

One of the simplest ways to enforce this idea is to marginalize $p(\mathbf{x}, \theta | \mathbf{y})$, and hence $p(\mathbf{x}|\theta)$, over all potential θ. In the following, θ will be the ensemble sampled moments.

Specifically, Bocquet [2011] recognized that the ensemble mean $\overline{\mathbf{x}}$ and ensemble error covariance matrix $\mathbf{P}$ used in the EnKF may be different from the unknown first- and second-order moments of the true error distribution, $\mathbf{x}_b$ and $\mathbf{B}$, where $\mathbf{B}$ is a positive definite matrix. The mismatch is due to the finite size of the ensemble, which leads to sampling errors. It is claimed in Bocquet et al. [2015] that these errors are mainly induced by the nonlinear ensemble propagation in the forecast step.

Let us account for the uncertainty in $\mathbf{x}_b$ and in $\mathbf{B}$. As in Section 6.4.1, we denote by $\mathbf{E} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m]$ the ensemble of size m formatted as an $n \times m$ matrix; $\overline{\mathbf{x}} = \mathbf{E}\mathbf{1}/m$ the ensemble mean, where $\mathbf{1} = (1, \ldots, 1)^{\mathrm{T}}$; and $\mathbf{X} = (\mathbf{E} - \overline{\mathbf{x}}\mathbf{1}^{\mathrm{T}})/\sqrt{m-1}$ the normalized perturbation matrix. Hence, $\mathbf{P} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$ is the empirical covariance matrix of the ensemble. Marginalizing over all potential $\mathbf{x}_b$ and $\mathbf{B}$, the prior of $\mathbf{x}$ reads

$$p(\mathbf{x}|\mathbf{E}) = \int \mathrm{d}\mathbf{x}_b \mathrm{d}\mathbf{B}\, p(\mathbf{x}|\mathbf{E}, \mathbf{x}_b, \mathbf{B}) p(\mathbf{x}_b, \mathbf{B}|\mathbf{E}).$$

The symbol $\mathrm{d}\mathbf{B}$ corresponds to the Lebesgue measure on all independent entries $\prod_{i \leq j}^{n} \mathrm{d}[\mathbf{B}]_{ij}$, but the integration is restricted to the cone of positive definite matrices. Since $p(\mathbf{x}|\mathbf{E}, \mathbf{x}_b, \mathbf{B})$ is conditioned on the knowledge of the true prior statistics, it does not depend on $\mathbf{E}$, so that

$$p(\mathbf{x}|\mathbf{E}) = \int \mathrm{d}\mathbf{x}_b \mathrm{d}\mathbf{B}\, p(\mathbf{x}|\mathbf{x}_b, \mathbf{B}) p(\mathbf{x}_b, \mathbf{B}|\mathbf{E}).$$

Bayes' rule can be applied to $p(\mathbf{x}_b, \mathbf{B}|\mathbf{E})$, yielding

$$p(\mathbf{x}|\mathbf{E}) = \frac{1}{p(\mathbf{E})} \int \mathrm{d}\mathbf{x}_b \mathrm{d}\mathbf{B}\, p(\mathbf{x}|\mathbf{x}_b, \mathbf{B}) p(\mathbf{E}|\mathbf{x}_b, \mathbf{B}) p(\mathbf{x}_b, \mathbf{B}). \tag{6.30}$$

Assuming independence of the samples, the likelihood of the ensemble $\mathbf{E}$ can be written

$$p(\mathbf{E}|\mathbf{x}_b, \mathbf{B}) = \prod_{i=1}^{m} p(\mathbf{x}_i|\mathbf{x}_b, \mathbf{B}).$$

The last factor in (6.30), $p(\mathbf{x}_b, \mathbf{B})$, is the hyperprior. The distribution represents our beliefs about the forecast filter statistics, $\mathbf{x}_b$ and $\mathbf{B}$, prior to actually running any filter. We recall that this distribution is termed *hyperprior* because it represents a prior for the background information in the first stage of a Bayesian hierarchy.

Assuming one subscribes to this view of the EnKF, it shows that more information is actually required in the EnKF, in addition to the observations and the prior ensemble, which are potentially insufficient for an inference.

A simple choice was made in Bocquet [2011] for the hyperprior: the Jeffreys' prior is an analytically tractable and uninformative hyperprior of the form

$$p_{\mathrm{J}}(\mathbf{x}_b, \mathbf{B}) \propto |\mathbf{B}|^{-\frac{n+1}{2}}, \tag{6.31}$$

where $|\mathbf{B}|$ is the determinant of the background error covariance matrix $\mathbf{B}$ of dimension $n \times n$. A more sophisticated hyperprior meant to hold static information, the normal-inverse-Wishart distribution, was proposed in Bocquet et al. [2015].

With a given hyperprior, the marginalization over $\mathbf{x}_b$ and $\mathbf{B}$, (6.30), can in principle be carried out to obtain $p(\mathbf{x}|\mathbf{E})$. We choose to call it a *predictive prior* to comply with the traditional view that sees it as prior before assimilating the observations. Note,

however, that statisticians would rather call it a *predictive posterior* distribution as the outcome of a first-stage inference of a Bayesian hierarchy, where $\mathbf{E}$ is the data.

Using Jeffreys' hyperprior, Bocquet [2011] showed that the integral can be obtained analytically and that the predictive prior is a multivariate t-distribution,

$$p(\mathbf{x}|\mathbf{E}) \propto \left| \frac{(\mathbf{x}-\overline{\mathbf{x}})(\mathbf{x}-\overline{\mathbf{x}})^{\mathrm{T}}}{m-1} + \varepsilon_m \mathbf{P} \right|^{-\frac{m}{2}}, \tag{6.32}$$

where $|\cdot|$ denotes the determinant and $\varepsilon_m = 1 + 1/m$. The determinant is computed in the ensemble subspace $\xi = \overline{\mathbf{x}} + \mathrm{Vec}(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m)$, i.e., the affine space spanned by the perturbations of the ensemble so that it is not singular. Moreover, we impose $p(\mathbf{x}|\mathbf{E}) = 0$ if $\mathbf{x}$ is not in ξ. This distribution has fat tails, thus accounting for the uncertainty in $\mathbf{B}$. The factor ε_m is a result of the uncertainty in $\mathbf{x}_b$; if $\mathbf{x}_b$ were known to coincide with the ensemble mean $\overline{\mathbf{x}}$, then ε_m would be 1 instead. For a Gaussian process, $\varepsilon_m \mathbf{P}$ is an unbiased estimator of the squared error of the ensemble mean $\overline{\mathbf{x}}$ [Sacher and Bartello, 2008], where ε_m stems from the uncertain $\mathbf{x}_b$, which does not coincide with $\overline{\mathbf{x}}$. In the derivation of Bocquet [2011], the $\varepsilon_m \mathbf{P}$ correction comes from integrating out on $\mathbf{x}_b$. Therefore, ε_m can be seen as an inflation factor on the prior covariance matrix that should actually apply to any type of EnKF.

This non-Gaussian prior distribution can be seen as a mixture of Gaussian distributions weighted according to the hyperprior. It can be shown that (6.32) can be rearranged as

$$p(\mathbf{x}|\mathbf{E}) \propto \left\{ 1 + \frac{(\mathbf{x}-\overline{\mathbf{x}})^{\mathrm{T}} (\varepsilon_m \mathbf{P})^{\dagger} (\mathbf{x}-\overline{\mathbf{x}})}{m-1} \right\}^{-\frac{m}{2}} \tag{6.33}$$

for $\mathbf{x} \in \xi$ and $p(\mathbf{x}|\mathbf{E}) = 0$ if $\mathbf{x} \notin \xi$; $\mathbf{P}^{\dagger}$ is the Moore–Penrose inverse of $\mathbf{P}$.

In comparison, the traditional EnKF implicitly assumes that the hyperprior is $\delta(\mathbf{B}-\mathbf{P})\delta(\mathbf{x}_b - \overline{\mathbf{x}})$, where δ is a Dirac multidimensional distribution. In other words, the background statistics generated from the ensemble coincide with the true background statistics. As a result, one obtains in this case the Gaussian prior

$$p(\mathbf{x}|\mathbf{E}) \propto \exp\left\{ -\frac{1}{2} (\mathbf{x}-\overline{\mathbf{x}})^{\mathrm{T}} \mathbf{P}^{\dagger} (\mathbf{x}-\overline{\mathbf{x}}) \right\} \tag{6.34}$$

for $\mathbf{x} \in \xi$ and $p(\mathbf{x}|\mathbf{E}) = 0$ if $\mathbf{x} \notin \xi$.

From these predictive priors given in state space, it is possible to derive a formally simple prior in ensemble subspace, i.e., in terms of the coefficients $\mathbf{w}$ that we used for the ETKF and MLEF ($\mathbf{x} = \overline{\mathbf{x}} + \mathbf{X}\mathbf{w}$). In turn, this prior leads to an effective cost function in the ensemble subspace for the analysis [Bocquet et al., 2015]

$$\mathcal{J}(\mathbf{w}) = \frac{1}{2} ||\mathbf{y} - \mathcal{H}(\overline{\mathbf{x}} + \mathbf{X}\mathbf{w})||_{\mathbf{R}}^2 + \frac{m+1}{2} \ln\left(\varepsilon_m + \frac{||\mathbf{w}||^2}{m-1} \right), \tag{6.35}$$

which should be used in place of the related cost function of the ETKF, which reads [Hunt et al., 2007]

$$\mathcal{J}(\mathbf{w}) = \frac{1}{2} ||\mathbf{y} - \mathcal{H}(\overline{\mathbf{x}} + \mathbf{X}\mathbf{w})||_{\mathbf{R}}^2 + \frac{1}{2} ||\mathbf{w}||^2. \tag{6.36}$$

The EnKF that results from the effective cost function (6.35) has been called the finite-size EnKF because it sees the ensemble in the asymptotic limit, but as a finite set. It

is denoted *EnKF-N*, where the N indicates an explicit dependence on the size of the ensemble.

It was further shown in Bocquet and Sakov [2012] that it is enlightening to separate the angular degrees of freedom of $\mathbf{w}$, i.e., $\mathbf{w}/|\mathbf{w}|$, from its radial one $|\mathbf{w}|$ in the cost function. This amounts to defining a Lagrangian of the form

$$\mathscr{L}(\mathbf{w},\boldsymbol{\rho},\zeta)=\frac{1}{2}||\mathbf{y}-\mathscr{H}(\overline{\mathbf{x}}+\mathbf{X}\mathbf{w})||^2_{\mathbf{R}}+\frac{\zeta}{2}\left(\frac{||\mathbf{w}||^2}{m-1}-\rho\right)+\frac{m+1}{2}\ln(\varepsilon_m+\rho),$$

where ζ is a Lagrange parameter used to enforce the decoupling. When the observation operator is linear or linearized, this Lagrangian turns out to be equivalent to a dual cost function of the ζ parameter, which is

$$\mathscr{D}(\zeta)=\frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\left(\mathbf{R}+\frac{m-1}{\zeta}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right)^{-1}\boldsymbol{\delta}+\frac{\varepsilon_m\zeta}{2}+\frac{m+1}{2}\ln\frac{m+1}{\zeta}-\frac{m+1}{2}, \tag{6.37}$$

where $\boldsymbol{\delta}=\mathbf{y}-\mathscr{H}(\overline{\mathbf{x}})$ is the innovation vector. The dual cost function is defined over the interval $]0,(m+1)/\varepsilon_m]$. Although it is not necessarily convex, its global minimum can easily be found numerically because it is a one-dimensional optimization problem. To perform an EnKF-N analysis using this dual cost function, one would first minimize (6.37) to obtain the optimal ζ_{a}. The analysis is then

$$\mathbf{w}_{\mathrm{a}}=\left(\mathbf{Y}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y}+\frac{\zeta_{\mathrm{a}}}{m-1}\mathbf{I}_m\right)^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{\delta}=\mathbf{Y}^{\mathrm{T}}\left(\frac{\zeta_{\mathrm{a}}}{m-1}\mathbf{R}+\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right)^{-1}\boldsymbol{\delta}. \tag{6.38}$$

Based on the effective cost function (6.35), an updated set of perturbations can be obtained:

$$\mathbf{X}_{\mathrm{a}}=\mathbf{X}[\mathrm{H}_{\mathrm{a}}]^{-\frac{1}{2}}\mathbf{U}\quad\text{with}\quad\mathrm{H}_{\mathrm{a}}=\mathbf{Y}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y}+\frac{\zeta_{\mathrm{a}}}{m-1}\mathbf{I}_m-\frac{2}{m+1}\left(\frac{\zeta_{\mathrm{a}}}{m-1}\right)^2\mathbf{w}_{\mathrm{a}}\mathbf{w}_{\mathrm{a}}^{\mathrm{T}}. \tag{6.39}$$

The last term of the Hessian, $-\frac{2}{m+1}\left(\frac{\zeta_{\mathrm{a}}}{m-1}\right)^2\mathbf{w}_{\mathrm{a}}\mathbf{w}_{\mathrm{a}}^{\mathrm{T}}$, which is related to the covariances of the angular and radial degrees of freedom of $\mathbf{w}$, can very often be neglected. If so, the update equations are equivalent to those of the ETKF but with an inflation of the prior covariance matrix by a factor $(m-1)/\zeta_{\mathrm{a}}$. Hence the EnKF-N implicitly determines an adaptive optimal inflation.

If an SVD of $\mathbf{Y}$ is available, the minimization of $\mathscr{D}(\zeta)$ is immediate, using for instance a dichotomous search. Such a decomposition is often already available because it was meant to be used to compute (6.38) and (6.39).

Practically, it was found in several low-order models and in perfect model conditions that the EnKF-N does not require any inflation and that its performance is close to that of an equivalent ETKF, where a uniform inflation would have been optimally tuned to obtain the best performance of the filter. In a preliminary study by Bocquet et al. [2015], the use of a more informative hyperprior, such as the normal-inverse-Wishart distribution, was proposed to avoid the need for localization while still avoiding the need for inflation.

6.7.3.1 ▪ Numerical illustrations

The three-variable Lorenz model [Lorenz, 1963] (Lorenz-63 hereafter) is the emblem of chaotic systems. It is defined by the ODEs

$$\frac{dx}{dt} = \sigma(y-x),$$
$$\frac{dy}{dt} = \rho x - y - xz,$$
$$\frac{dz}{dt} = xy - \beta z,$$

where $\sigma = 10$, $\rho = 28$, and $\beta = 8/3$. This model is chaotic, with $(0.91, 0, -14.57)$ as its Lyapunov exponents. The model doubling time is 0.78 time units. Its attractor has the famous butterfly shape with two distinct wings, or lobes, which were illustrated in Section 2.5.1. It was used by Edward Lorenz to explain the finite horizon of predictability in meteorology.

To demonstrate the relevance of the EnKF-N with this model, a numerical twin experiment similar to that used with the Lorenz-95 model is designed. The system is assumed fully observed so that $\mathbf{H}_k \equiv \mathbf{I}_3$, with the observation error covariance matrix $\mathbf{R}_k \equiv 4\mathbf{I}_3$. The time interval between observational updates is varied from $\Delta t = 0.10$ to $\Delta t = 0.50$. The larger Δt is, the stronger the impact of model nonlinearity on the state estimation, and the stronger the need for an inflation correction. The performance of the DA schemes is measured by the analysis RMSE (6.23) averaged over a very long run. We test a standard ETKF where the uniform inflation is optimally tuned to minimize the RMSE (about 20 values are tested). We compare it to the EnKF-N, which does not require inflation. In both cases, the ensemble size is set to $m = 3$.

The skills of both filters are shown in Figure 6.8. It is remarkable that the EnKF-N achieves an even better performance than the optimally tuned ETKF, without any tuning. As Δt increases, the ETKF requires a significantly stronger inflation. This is mostly needed at the transition between the two lobes of the Lorenz-63 attractor. Within the lobes, the DA system is effectively much more linear and requires little inflation. By contrast, the EnKF-N, which is adaptive, applies a strong inflation only when needed, i.e., at the transition between lobes.

An analogous experiment can be carried out, but with the Lorenz-95 model. The setup of Section 6.6 is used. The performance of both filters is shown in the left panel of Figure 6.9 when the ensemble size is varied and when $\Delta t = 0.05$. The EnKF-N achieves the same performance but without any tuning, and hence with numerical efficiency. A similar experiment is performed but varying Δt, and hence system nonlinearity, and setting $m = 20$. The results are reported in the right panel of Figure 6.9. Again, it can be seen that the same performance can be reached without any tuning with the EnKF-N.

The adaptation of the EnKF-N is also illustrated by Figure 6.7, where the MLEF and ETKF were compared in the first stages of a DA experiment with the Lorenz-95 model and a nonlinear observation operator. In this context, the finite-size MLEF and ETKF were shown to outperform the standard MLEF and ETKF with optimally tuned uniform inflation.

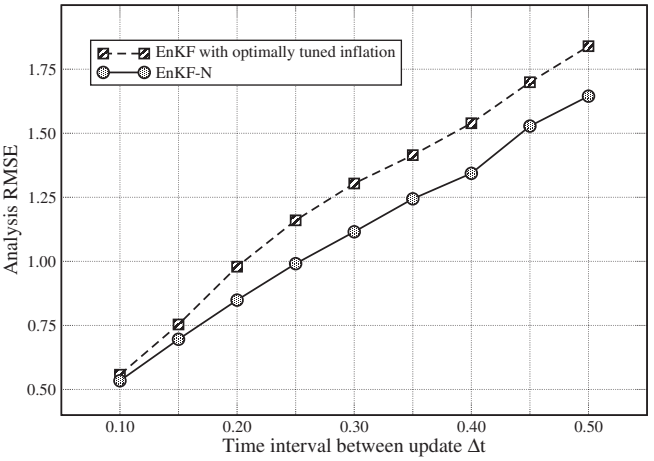

Figure 6.8. *Average analysis RMSE for a deterministic EnKF (ETKF) with optimally tuned inflation and for the EnKF-N. The model is Lorenz-*63*.*

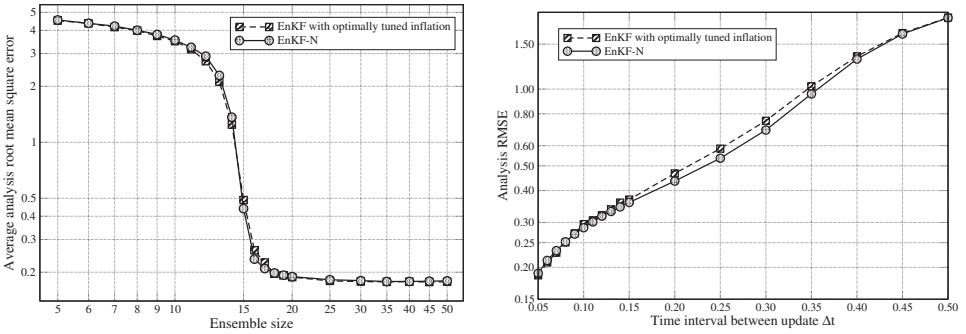

Figure 6.9. *Average analysis RMSE for a deterministic EnKF (ETKF) with optimally tuned inflation and for the EnKF-N. Left panel: the ensemble size is varied from* $m = 5$ *to* $m = 50$ *and* $\Delta t = 0.05$*. Right panel:* Δt *is varied from* 0.05 *to* 0.50 *and the ensemble size is set to* $m = 20$*. The model is Lorenz-*95*.*

6.7.3.2 ▪ Passing on hierarchical statistical information

The EnKF-N was built with the idea to remain algorithmically as close as possible to the EnKF. To avoid relying on any additional input, the hierarchy of information that was established on $\mathbf{x}_b$ and $\mathbf{B}$ was closed by marginalizing over all potential priors leading to effective cost functions.

However, it is actually possible to propagate the information that the filter carries about $\mathbf{x}_b$ and $\mathbf{B}$ from one cycle to the next. This idea was formalized in Myrseth and Omre [2010]. It relies on the natural conjugacy of the normal-inverse-Wishart distribution with the multivariate normal distribution. The ensemble can be seen as an observation set for the estimation of the moments of the true distribution, $\mathbf{x}_b$ and $\mathbf{B}$, which obeys a multivariate normal distribution. If $\mathbf{x}_b$ and $\mathbf{B}$ are supposed to follow a normal-inverse-Wishart distribution, then the posterior distribution will also follow a normal-inverse-Wishart distribution with parameters that are easily updated using the data, i.e., the ensemble members, thanks to the natural conjugacy. This defines a level-2 update scheme for the mean and the error covariances. The mean and error

covariances used on the level-1 update scheme, i.e., the EnKF, are subsequently drawn from this updated distribution. This scheme is quite different from a traditional EnKF and truly accounts for the uncertainty in the moments. Another and rather similar attempt was documented in Tsyrulnikov and Rakitkoa [2016].

6.8 ▪ The ensemble Kalman smoother

Filtering consists of assimilating observations as they become available, making the best estimate at the present time. If $\mathbf{y}_{K:1} = \mathbf{y}_K, \mathbf{y}_{K-1}, \ldots, \mathbf{y}_1$ is the collection of observations from t_1 to t_K, filtering aims at estimating the PDF $p(\mathbf{x}_K|\mathbf{y}_{K:1})$ at t_K. Only past and present observations are accounted for, which would necessarily be the case for real-time nowcasting and forecasting.

Smoothing, on the other hand, aims at estimating the state of the system (or a trajectory of it), using past, present, and possibly future observations. Indeed, assuming again t_K is the present time, one could also be interested in the distribution $p(\mathbf{x}_k|\mathbf{y}_{K:1})$, where $1 \le k \le K$. More generally, one would be interested in estimating $p(\mathbf{x}_{K:1}|\mathbf{y}_{K:1})$, where $\mathbf{x}_{K:1} = \mathbf{x}_K, \mathbf{x}_{K-1}, \ldots, \mathbf{x}_1$ is the collection of state vectors from t_1 to t_K, i.e., a trajectory. Note that $p(\mathbf{x}_k|\mathbf{y}_{K:1})$ is a marginal distribution of $p(\mathbf{x}_{K:1}|\mathbf{y}_{K:1})$ obtained by integrating out $\mathbf{x}_l$, with $l = 1, \ldots, k-1, k+1, \ldots, K$. This is especially useful for hindcasting and reanalysis problems that aim at retrospectively obtaining the best estimate for the model state using all available information.

The KF can be extended to address the smoothing problem, leading to a variety of Kalman smoother algorithms [Anderson and Moore, 1979]. It has been introduced in the geosciences and studied in this context by Cohn et al. [1994].

The generalization of the Kalman smoother to the family of *ensemble* Kalman filters followed the development of the EnKF. Even for linear systems, where we theoretically expect equivalent schemes, there are several implementations of the smoother, even for a given flavor of EnKF [Cosme et al., 2012]. Its implementation may also depend on whether one wishes to estimate a state vector or a trajectory, or on how long the backward analysis can go. In the following, we shall focus on the fixed-lag ensemble Kalman smoother (EnKS) that was introduced and used in Evensen and van Leeuwen [2000], Zhu et al. [2003], Khare et al. [2008], Cosme et al. [2010], and Nerger et al. [2014]. If Δt is the time interval between updates, fixed-lag means that the analysis goes backward by $L\Delta t$ in time. The variable L measures the length of the time window in units of Δt. We will implement it following the ensemble transform framework as used in Bocquet and Sakov [2013, 2014].

The EnKS is built on the EnKF, which will be its backbone. There are two steps. The first step consists of running the EnKF from t_{k-1} to t_k. The second step consists of updating the state vectors at t_{k-1}, t_{k-2} back to t_{k-L}.

Let us make L the lag of the EnKS and define t_L as the present time. Hence, we wish to retrospectively estimate $\mathbf{x}_{L:0} = \mathbf{x}_L, \mathbf{x}_{L-1}, \ldots, \mathbf{x}_0$. An EnKF has been run from the starting time to t_L. The collection of all posterior ensembles $\mathbf{E}_L, \mathbf{E}_{L-1}, \ldots, \mathbf{E}_0$ is assumed to be stored, which is a significant technical constraint. For each $\mathbf{E}_k$ there is a corresponding mean state $\overline{\mathbf{x}}_k$ and a normalized perturbation matrix $\mathbf{X}_k$. This need for memory (at least $L \times m \times n$ scalars) is the main requirement of the EnKS.

The EnKF provides an approximation for the PDF $p(\mathbf{x}_L|\mathbf{y}_{L:})$, where $\mathbf{y}_{L:}$ represents the collection of observation vectors from the beginning of the DA experiment (earlier than t_0) to t_L. For the EnKF, this PDF can be approximated as a Gaussian distribution. Now, let us describe the backward pass, starting from the latest dates. Let us first derive a retrospective update for $\mathbf{x}$ one time step backward. Hence, we wish to compute a

Gaussian approximation for $p(\mathbf{x}_{L-1}|\mathbf{y}_{L:})$. From Bayes' rule, we obtain

$$p(\mathbf{x}_{L-1}|\mathbf{y}_{L:}) \propto p(\mathbf{y}_L|\mathbf{x}_{L-1}, \mathbf{y}_{L-1:}) p(\mathbf{x}_{L-1}|\mathbf{y}_{L-1:}),$$

which relates the smoothing PDF to the current observation likelihood and to the filtering distribution at t_{L-1}. From the EnKF's standpoint, $p(\mathbf{x}_{L-1}|\mathbf{y}_{L-1:})$ is approximately Gaussian in the affine subspace centered at $\mathbf{x}^{\mathrm{a}}_{L-1}$ and spanned by the columns of $\mathbf{X}^{\mathrm{a}}_{L-1}$. The affine subspace can be parameterized by $\mathbf{x} = \mathbf{x}^{\mathrm{a}}_{L-1} + \mathbf{X}^{\mathrm{a}}_{L-1}\mathbf{w}$. Using the variational expression of the ETKF or the MLEF as defined in ensemble subspace, $p(\mathbf{x}_{L-1}|\mathbf{y}_{L-1:})$ is proportional to $\exp(-\frac{1}{2}||\mathbf{w}||^2)$ when written in terms of the coordinates, $\mathbf{w}$, of the ensemble subspace (see Section 6.4.2). Then $p(\mathbf{y}_L|\mathbf{x}_{L-1}, \mathbf{y}_{L-1:})$ is the likelihood and reads, in ensemble subspace,

$$p(\mathbf{y}_L|\mathbf{x}_{L-1}, \mathbf{y}_{L-1:}) \propto \exp\left\{-\frac{1}{2}\left\|\mathbf{y}_L - \mathscr{H}_L \circ \mathscr{M}_{L:L-1}(\mathbf{x}^{\mathrm{a}}_{L-1} + \mathbf{X}^{\mathrm{a}}_{L-1}\mathbf{w})\right\|^2_{\mathbf{R}_L}\right\}.$$

Hence the complete cost function for the retrospective analysis on t_{L-1} is

$$\mathscr{L}(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{y}_L - \mathscr{H}_L \circ \mathscr{M}_{L:L-1}(\mathbf{x}^{\mathrm{a}}_{L-1} + \mathbf{X}^{\mathrm{a}}_{L-1}\mathbf{w})\right\|^2_{\mathbf{R}_L} + \frac{1}{2}||\mathbf{w}||^2.$$

This type of potentially nonquadratic cost function will be at the heart of the IEnKF/IEnKS (see Chapter 7). Here, the update is expanded around $\mathbf{x}^{\mathrm{a}}_{L-1}$ using the TLM to make the cost function quadratic:

$$\begin{aligned}
\mathscr{L}(\mathbf{w}) &= \frac{1}{2}\left\|\mathbf{y}_L - \mathscr{H}_L \circ \mathscr{M}_{L:L-1}(\mathbf{x}^{\mathrm{a}}_{L-1} + \mathbf{X}^{\mathrm{a}}_{L-1}\mathbf{w})\right\|^2_{\mathbf{R}_L} + \frac{1}{2}||\mathbf{w}||^2 \\
&\simeq \frac{1}{2}\left\|\mathbf{y}_L - \mathscr{H}_L \circ \mathscr{M}_{L:L-1}(\mathbf{x}^{\mathrm{a}}_{L-1}) - \mathbf{H}_L\mathbf{M}_{L:L-1}\mathbf{X}^{\mathrm{a}}_{L-1}\mathbf{w}\right\|^2_{\mathbf{R}_L} + \frac{1}{2}||\mathbf{w}||^2 \\
&= \frac{1}{2}\left\|\mathbf{y}_L - \mathscr{H}_L(\mathbf{x}^{\mathrm{f}}_L) - \mathbf{H}_L\mathbf{X}^{\mathrm{f}}_L\mathbf{w}\right\|^2_{\mathbf{R}_L} + \frac{1}{2}||\mathbf{w}||^2 \\
&= \frac{1}{2}\left\|\boldsymbol{\delta}_L - \mathbf{Y}^{\mathrm{f}}_L\mathbf{w}\right\|^2_{\mathbf{R}_L} + \frac{1}{2}||\mathbf{w}||^2,
\end{aligned}$$

where $\boldsymbol{\delta}_L = \mathbf{y}_L - \mathscr{H}_L(\mathbf{x}^{\mathrm{f}}_L)$ and, as before, $\mathbf{Y}^{\mathrm{f}}_k = \mathbf{H}_k\mathbf{X}^{\mathrm{f}}_k$. From this cost function, the derivation of an ETKF-like analysis is immediate. We obtain

$$\mathbf{X}^{\mathrm{a},1}_{L-1} = \mathbf{X}^{\mathrm{a}}_{L-1}\sqrt{\boldsymbol{\Omega}_\star} \qquad \text{and} \qquad \mathbf{x}^{\mathrm{a},1}_{L-1} = \mathbf{x}^{\mathrm{a}}_{L-1} + \mathbf{X}^{\mathrm{a}}_{L-1}\mathbf{w}_\star, \tag{6.40}$$

with

$$\boldsymbol{\Omega}_\star = \left[\mathbf{I}_m + (\mathbf{Y}^{\mathrm{f}}_L)^{\mathrm{T}}\mathbf{R}^{-1}_L\mathbf{Y}^{\mathrm{f}}_L\right]^{-1} \qquad \text{and} \qquad \mathbf{w}_\star = \boldsymbol{\Omega}_\star\mathbf{Y}^{\mathrm{f}}_L\mathbf{R}^{-1}_L\boldsymbol{\delta}_L. \tag{6.41}$$

The superscript 1 indicates that the estimate of $\mathbf{x}_{L-1}$ now accounts for observation one time step ahead.

We can proceed backward and consider an updated estimation for $\mathbf{x}_{L-2}$. The EnKF had yielded $\mathbf{x}^{\mathrm{a}}_{L-2}$, two time steps earlier. The backward pass of the EnKS run, one time step earlier, must have updated $\mathbf{x}_{L-2}$ using the same formula that we derived for $\mathbf{x}_{L-1}$ a few lines above, yielding the estimate $\mathbf{x}^{\mathrm{a},1}_{L-2}$ and the updated ensemble $\mathbf{X}^{\mathrm{a},1}_{L-2}$. With $\mathbf{x}^{\mathrm{a},1}_{L-2}$, we have accounted for $\mathbf{y}_{L-1}$, but not $\mathbf{y}_L$ yet. Hence, using Bayesian estimation as a guide, we need to estimate

$$p(\mathbf{x}_{L-2}|\mathbf{y}_{L:}) \propto p(\mathbf{y}_L|\mathbf{x}_{L-2}, \mathbf{x}_{L-1:}) p(\mathbf{x}_{L-2}|\mathbf{y}_{L-1:}).$$

As above, we can derive from these distributions an approximate quadratic cost function for the retrospective analysis on $\mathbf{x}_{L-2}$. We formally obtain the cost function

$$\mathscr{L}(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{y}_L - \mathscr{H}_L \circ \mathscr{M}_{L:L-2}(\mathbf{x}^{\mathrm{a},1}_{L-2} + \mathbf{X}^{\mathrm{a},1}_{L-2}\mathbf{w})\right\|^2_{\mathbf{R}_L} + \frac{1}{2}\|\mathbf{w}\|^2,$$

where $\mathbf{x}_{L-2}$ is parameterized as $\mathbf{x}_{L-2} = \mathbf{x}^{\mathrm{a},1}_{L-2} + \mathbf{X}^{\mathrm{a},1}_{L-2}\mathbf{w}$. The outcome is formally the same,

$$\mathscr{L}(\mathbf{w}) = \frac{1}{2}\left\|\boldsymbol{\delta}_L - \mathbf{Y}^{\mathrm{f}}_L\mathbf{w}\right\|^2_{\mathbf{R}_L} + \frac{1}{2}\|\mathbf{w}\|^2,$$

with $\boldsymbol{\delta}_L = \mathbf{y}_L - \mathscr{H}_L(\mathbf{x}^{\mathrm{f}}_L)$ and $\mathbf{Y}^{\mathrm{f}}_L = \mathbf{H}_L\mathbf{X}^{\mathrm{f}}_L$, but where $\mathbf{w}$ is a vector of coefficients that applies to a different ensemble of perturbations ($\mathbf{X}^{\mathrm{a},1}_{L-2}$ instead of $\mathbf{X}^{a}_{L-1}$). From this cost function, an ETKF-like analysis is immediate. We obtain

$$\mathbf{x}^{\mathrm{a},2}_{L-2} = \mathbf{x}^{\mathrm{a},1}_{L-2} + \mathbf{X}^{\mathrm{a},1}_{L-2}\mathbf{w}_\star \qquad \text{and} \qquad \mathbf{X}^{\mathrm{a},2}_{L-2} = \mathbf{X}^{\mathrm{a},1}_{L-2}\sqrt{\boldsymbol{\Omega}_\star},$$

with

$$\boldsymbol{\Omega}_\star = \left[\mathbf{I}_m + (\mathbf{Y}^{\mathrm{f}}_L)^{\mathrm{T}}\mathbf{R}^{-1}_L\mathbf{Y}^{\mathrm{f}}_L\right]^{-1} \qquad \text{and} \qquad \mathbf{w}_\star = \boldsymbol{\Omega}_\star\mathbf{Y}^{\mathrm{f}}_L\mathbf{R}^{-1}_L\boldsymbol{\delta}_L.$$

The state vectors have been updated retrospectively down to two time steps in the past (fixed-lag EnKS of $L = 2$). The scheme could be extended backward for longer lags following the same rationale. The schematic Figure 6.10 of the EnKS depicts the EnKF step followed by a backward sweep for smoothing for a time window length of $L = 4$. Hence, provided one can afford to store the ensembles, the backward pass is rather cheap as it only performs the linear algebra of the analysis. The forecast model is only used in the EnKF pass. The algorithm of the EnKS in ensemble subspace is given in Algorithm 6.6 as a pseudocode.

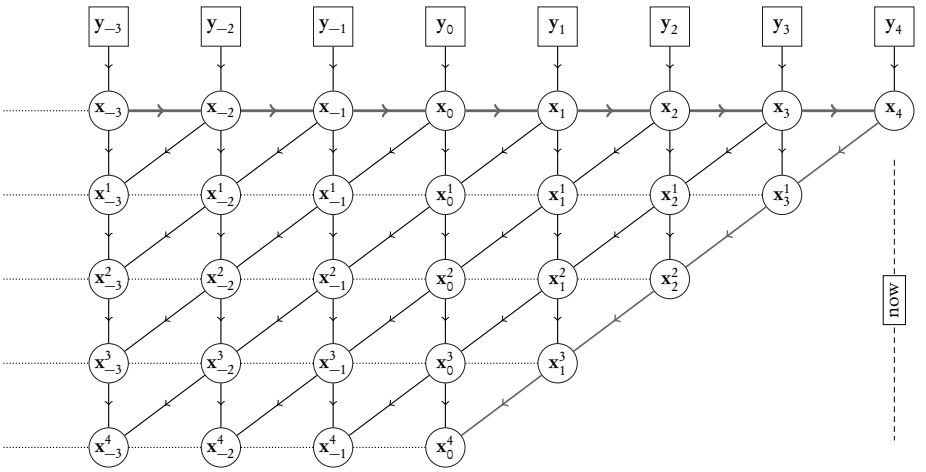

Figure 6.10. *Schematic of the EnKS. The far right of the figure (t_4) corresponds to the present time. The upper level corresponds to the EnKF pass, from left to right. Lower levels indicate updates that account for more recent observations. The backward smoothing pass follows a diagonal moving down and backward in time. The arrows indicate the fluxes of information.*

Algorithm 6.6 Pseudocode for a complete cycle of the EnKS in ensemble subspace.

Require: Observation operator $\mathscr{H}$ at current time; for $l = 0, L-1$: $\mathbf{E}^l$, the previous L analysis ensembles; $\mathbf{E}^L$, the forecast ensemble at current time; $\mathbf{y}$, the observation at current time; $\mathbf{U}$, an orthogonal matrix in $\mathbb{R}^{m\times m}$ satisfying $\mathbf{U1} = \mathbf{1}$; $\mathscr{M}$, the model resolvent from current time to the next analysis time.

$\mathbf{E} = \mathbf{E}^L$
$\overline{\mathbf{x}} = \mathbf{E1}/m$
$\mathbf{X} = \left(\mathbf{E} - \overline{\mathbf{x}}\mathbf{1}^{\mathrm{T}}\right)/\sqrt{m-1}$
$\mathbf{Z} = \mathscr{H}(\mathbf{E})$
$\overline{\mathbf{y}} = \mathbf{Z1}/m$
$\mathbf{S} = \mathbf{R}^{-\frac{1}{2}}\left(\mathbf{Z} - \overline{\mathbf{y}}\mathbf{1}^{\mathrm{T}}\right)/\sqrt{m-1}$
$\boldsymbol{\delta} = \mathbf{R}^{-\frac{1}{2}}(\mathbf{y} - \overline{\mathbf{y}})$
$\mathbf{T} = \left(\mathbf{I}_m + \mathbf{S}^{\mathrm{T}}\mathbf{S}\right)^{-1}$
$\mathbf{w} = \mathbf{T}\mathbf{S}^{\mathrm{T}}\boldsymbol{\delta}$
$\mathbf{E}^L = \overline{\mathbf{x}}\mathbf{1}^{\mathrm{T}} + \mathbf{X}\left(\mathbf{w}\mathbf{1}^{\mathrm{T}} + \sqrt{m-1}\mathbf{T}^{\frac{1}{2}}\mathbf{U}\right)$
for $l = 0, \ldots, L-1$ **do**
  $\overline{\mathbf{x}} = \mathbf{E}^l\mathbf{1}/m$
  $\mathbf{X} = \left(\mathbf{E}^l - \overline{\mathbf{x}}\mathbf{1}^{\mathrm{T}}\right)/\sqrt{m-1}$
  $\mathbf{E}^l = \overline{\mathbf{x}}\mathbf{1}^{\mathrm{T}} + \mathbf{X}\left(\mathbf{w}\mathbf{1}^{\mathrm{T}} + \sqrt{m-1}\mathbf{T}^{\frac{1}{2}}\mathbf{U}\right)$
end for
for $l = 0, \ldots, L-1$ **do**
  $\mathbf{E}^l := \mathbf{E}^{l+1}$
end for
$\mathbf{E}^L = \mathscr{M}(\mathbf{E}^{L-1})$

Using the same setup as in Section 6.6 and $m = 20$, we test the EnKS with the Lorenz-95 model. The lag is fixed to $L = 50$. The results are reported in Figure 6.11. The EnKS provides not only a filtering estimation but also a retrospective estimation of the state $l\Delta t$ in the past with $0 \leq l \leq L$. The analysis time-average RMSEs are computed as a function of $0 \leq l \leq L$. As expected, for $l = 0$, one recovers the filtering performance, while for $l = 50$, one recovers the maximal smoothing performance. In accordance with a few experiments in the literature on several models, a performance saturation is observed when the lag, L, is increased. This is due to the truncation of second-order statistics (the Gaussian assumption) so that information is necessarily lost with time.

One common mistake when implementing the EnKS is to apply inflation to account for sampling errors in both the EnKF and the backward updating pass. It must only be used in the EnKF, not the backward updating, because sampling errors are already accounted for in the EnKF and the updated ensembles once and for all. Hence, applying inflation in the backward pass would result in a suboptimal estimation, and a degradation of smoothing as L increases, or within the time window for fixed L.

Finally, let us mention that the ideas of the EnKS are quite useful in assimilating *asynchronous* observations within an EnKF DA system [Sakov et al., 2010]. Some DA systems, for instance the operational DA systems, continuously receive the observations they are meant to assimilate. Most of them have been obtained within the time window of the DA cycle, typically 3 to 6 hours for numerical weather forecasting systems. They are asynchronous with the update. Those observations, collected within

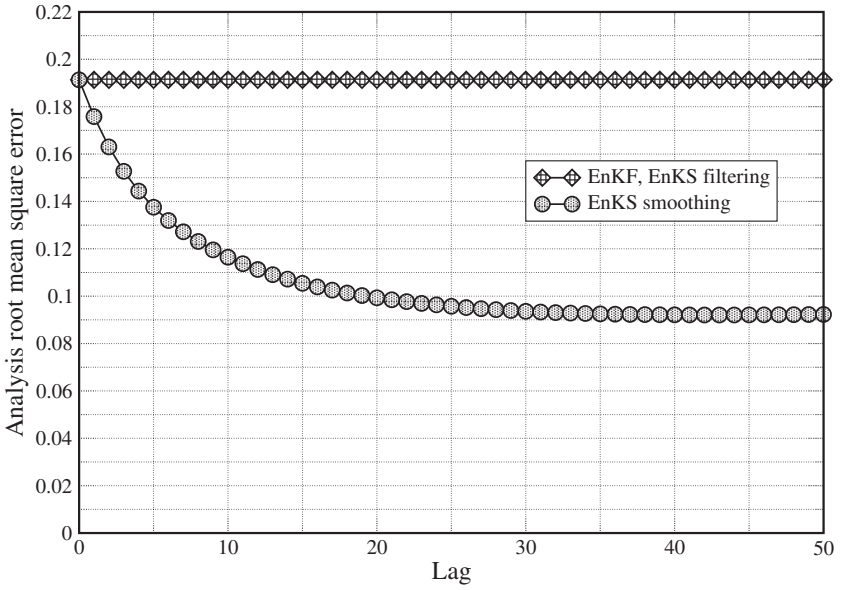

Figure 6.11. *Analysis RMSE of the EnKS with $L = 50$ using the same setup as in Section 6.6 and $m = 20$, applied to the Lorenz-95 model. The diamonds indicate the RMSE for the EnKF, which is also the filtering RMSE of the EnKS. The circles indicate the smoothing RMSE of the EnKS, i.e., the mean RMSE of the reanalysis of the state vector, $L\Delta t$, in the past.*

the time window $[t_k, t_{k+1}]$, can be used to update the ensemble at time t_k of the most recent scheduled update of an EnKF DA system. This can be achieved provided that an ensemble forecast throughout $[t_k, t_{k+1}]$ has been computed. Then, applying the smoothing equations of the EnKS, for instance (6.40) and (6.41), leads to the formulation of a consistent update at t_k. This, however, relies on linearity assumptions, which are indeed taken for granted in the EnKS.

6.9 ▪ A widespread and popular DA method

The EnKF, its variants, and precursors of the method have been successfully developed and implemented in meteorology and oceanography, including in operations. Because the method is simple to implement, it has been used in a huge number of studies in these fields. But it has spread out to other geoscience disciplines and beyond. For instance, to quote just a very few references, it has been applied in greenhouse gas inverse modeling [Kang et al., 2012], air quality forecasting [Wu et al., 2008], extraterrestrial atmosphere forecasting [Hoffman et al., 2010], detection and attribution in climate sciences [Hannart et al., 2016], geomagnetism reanalysis [Fournier et al., 2013], and ice-sheet parameter estimation and forecasting [Bonan et al., 2014]. It has also been used in petroleum reservoir estimation and in adaptive optics for extra-large telescopes [Gray et al., 2014].