3

A Tutorial on Bayesian Data Assimilation

Colin Grudzien and Marc Bocquet

Abstract: This chapter provides a broad introduction to Bayesian data assimilation that will be useful to practitioners in interpreting algorithms and results, and for theoretical studies developing novel schemes with an understanding of the rich history of geophysical data assimilation and its current directions. The simple case of data assimilation in a 'perfect' model is primarily discussed for pedagogical purposes. Some mathematical results are derived at a high level in order to illustrate key ideas about different estimators. However, the focus of this chapter is on the intuition behind these methods, where more formal and detailed treatments of the data assimilation problem can be found in the various references. In surveying a variety of widely used data assimilation schemes, the key message of this chapter is how the Bayesian analysis provides a consistent framework for the estimation problem and how this allows one to formulate its solution in a variety of ways to exploit the operational challenges in the geosciences.

3.1 Introduction

In applications such as short- to medium-range weather prediction, *data assimilation* (DA) provides a means to sequentially and recursively update forecasts of a time-varying physical process with newly incoming information (Daley, 1991; Kalnay, 2003; Asch et al., 2016), typically Earth observations. The Bayesian approach to DA is widely adopted (Lorenc, 1986) because it provides a unified treatment of tools from statistical estimation, non-linear optimisation, and machine learning for handling such a problem. This chapter illustrates how this approach can be utilised to develop and interpret a variety of widely used DA algorithms, both classical and those at the current state-of-the-art.

Suppose that the time-dependent physical states to be modelled can be written as a vector, $x_k \in \mathbb{R}^{N_x}$, where k labels some time t_k. Formally, the time-evolution of these states is represented with the non-linear map $\mathcal{M}$,

$$x_k = \mathcal{M}_k(x_{k-1}, \lambda) + \eta_k, \qquad (3.1)$$

where: (i) x_{k-1} is the state variable vector at an earlier time t_{k-1}; (ii) λ is a vector of uncertain static physical parameters but on which the time evolution depends; and (iii) η_k is an additive (for simplicity but other choices are possible), stochastic noise term, representing errors in our model for the physical process. Define $\Delta_t := t_k - t_{k-1}$ to be a fixed-length forecast horizon in this chapter, though none of the following results require this to be fixed in practice.

The basic goal of sequential DA is to estimate the random state vector x_k, given a prior distribution on (x_{k-1}, λ), and given knowledge of $\mathcal{M}_k$ and knowledge of how η_k is statistically distributed. At time t_{k-1}, a forecast is made for the distribution of x_k utilising the prior knowledge, which includes the physics-based model. For simplicity, most of this chapter is restricted to the case where λ is a known constant, and the forecast model is *perfect*, that is,

$$x_k = \mathcal{M}_k(x_{k-1}). \qquad (3.2)$$

However, a strength of Bayesian analysis is how it easily extends to include a general treatment of model errors and the estimation of model parameters (see, e.g., Asch et al. (2016) for a more general introduction).

While a forecast is made, one acquires a collection of observations of the real-world process. This is written as the observation vector $y_k \in \mathbb{R}^{N_y}$, which is related to the state vector by

$$y_k = \mathcal{H}_k(x_k) + \epsilon_k. \qquad (3.3)$$

The (possibly non-linear) map $\mathcal{H}_k : \mathbb{R}^{N_x} \to \mathbb{R}^{N_y}$ relates the physical states being modelled, x_k, to the values that are actually observed, y_k. Typically, in geophysical applications, observations are not 1 : 1 with the state variables; while the data dimension can be extremely large, $N_y \ll N_x$, so this information is sparse relative to the model state dimension. The term ϵ_k in Eq. (3.3) is an additive, stochastic noise term representing errors in the measurements, or

a mismatch between the state variable representation and the observation (the representation error, Janjić et al., 2018).

Therefore, sometime after the real-life physical system has reached time t_k, one has a forecast distribution for the states x_k, generated by the prior on $x_k - 1$ and the physics-based model $\mathcal{M}$, and the observations y_k with some associated uncertainty. The goal of Bayesian DA is to estimate the posterior distribution for x_k conditioned on y_k, or some statistics of this distribution.

3.2 Hidden Markov Models and Bayesian Analysis

3.2.1 The Observation-Analysis-Forecast Cycle

Recursive estimation of the distribution for x_k conditional on y_k (i.e. assuming y_k is known) can be described as an observation-analysis-forecast cycle (Trevisan and Uboldi, 2004). Given the forecast-prior for the model state, and the likelihood function for the observation, Bayes' law updates the prior for the modelled state to the posterior conditioned on the observation. Bayes' law is a simple re-arrangement of *conditional probability*, defined by Kolmogorov (2018) as

$$\mathcal{P}(A|B) := \frac{\mathcal{P}(A, B)}{\mathcal{P}(B)}, \tag{3.4}$$

for two events A and B where the probability of B is non-zero. Intuitively, this says that the probability of an event A, given knowledge that an event B occurs, is equal to the probability of A occurring relative to a sample space restricted to the event B. Using the symmetry in the joint event $\mathcal{P}(A, B) = \mathcal{P}(B, A)$, Bayes' law is written

$$\mathcal{P}(A|B) = \frac{\mathcal{P}(B|A)\mathcal{P}(A)}{\mathcal{P}(B)}. \tag{3.5}$$

In the observation-analysis-forecast cycle, A is identified with the state vector (seen as a random vector) taking its value in a neighbourhood of x_k, and B is identified with the observation vector (seen as a random vector) taking its value in a neighbourhood of y_k. The power of this statement is in how it describes an 'inverse' probability – while the *posterior*, $\mathcal{P}(A|B)$, on the left-hand-side may not be directly accessible, often the *likelihood* $\mathcal{P}(B|A)$ and the *prior* $\mathcal{P}(A)$ are easy to compute, and this is sufficient to develop a variety of probabilistic DA techniques.

A conceptual diagram of this process is pictured in Fig. 3.1. Given the initial first prior (represented as the first 'posterior' at time t_0), a forecast prior for the model state at t_1, $\mathcal{P}(A)$, is produced with the numerical model. At the update time, there is an (possibly indirect and noisy) observation of the physical state with an associated likelihood $\mathcal{P}(B|A)$. The posterior for the model state conditioned on this observation $\mathcal{P}(A|B)$, commonly denoted the *analysis* in geophysical DA, is used to initialise the subsequent
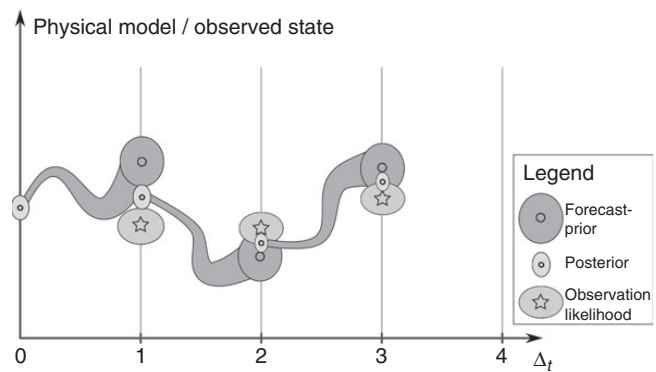

Figure 3.1 Conceptual diagram of the observation-analysis-forecast cycle. The *y*-axis represents a variable of the state and observation vectors; the *x*-axis represents time. Ellipses represent the spread of the observation/posterior/forecast errors. Original figure, adapted from Carrassi et al. (2018).

numerical forecast. Recursive estimates of the current modelled state can be performed in this fashion, but a related question regards the past states. The newly received observation gives information about the model states at past times and this allows one to produce a retrospective posterior estimate for the past states. Recursive estimation of the present state using these incoming observations is commonly known as *filtering*. Conditional estimation of a past state given a time series including future observations is commonly known as *smoothing*.

It is important to recognise that the filtering *probability density function* (pdf) for the current time $p(x_k|y_k, y_{k-1}, y_{k-2}, \ldots)$ is actually just a marginal of the joint posterior pdf over all states in the current *data assimilation window* (DAW) (i.e. the window of lagged past and current states being estimated). In Fig. 3.1, the DAW is the time window $\{t_1, t_2, t_3\}$. The conditional pdf for the model state at time t_3, given observations in the DAW, is written in terms of the joint posterior as

$$p(x_3|y_3, y_2, y_1) = \iiint p(x_3, x_2, x_1, x_0|y_3, y_2, y_1) \, dx_2 dx_1 dx_0, \tag{3.6}$$

by averaging out the past history of the model state from the joint posterior in the integrand. A smoothing estimate may be produced in a variety of ways, exploiting different formulations of the Bayesian problem. One may estimate only a *marginal pdf* as on the left-hand-side of Eq. (3.6), or the entire *joint posterior pdf* as in the integrand in Eq. 3.6 (Anderson and Moore, 1979; Cohn et al., 1994; Cosme et al., 2012). The approach chosen may strongly depend on whether the DAW is static or is advanced in time. Particularly, if one produces a smoothing estimate for the state at times t_1 through t_3, one may subsequently shift the DAW so that, in the new cycle, the posterior for times t_2

through t_4 is estimated: this type of analysis is known as *fixed-lag smoothing*. This chapter considers how one can utilise a Bayesian *maximum a posteriori* (MAP) formalism to efficiently solve the filtering and smoothing problems, using the various tools of statistical estimation, non-linear optimisation, and machine learning.

3.2.2 A Generic Hidden Markov Model

Recall the perfect physical process model and the noisy observation model,

$$\boldsymbol{x}_k = \mathcal{M}_k(\boldsymbol{x}_{k-1}), \tag{3.7a}$$

$$\boldsymbol{y}_k = \mathcal{H}_k(\boldsymbol{x}_k) + \boldsymbol{\epsilon}_k. \tag{3.7b}$$

Denote the sequence of the process model states and observation model states between time t_k and time t_l, for $k < l$, as

$$\boldsymbol{x}_{l:k} := \{\boldsymbol{x}_l, \boldsymbol{x}_{l-1}, \ldots, \boldsymbol{x}_k\}, \ \ \boldsymbol{y}_{l:k} := \{\boldsymbol{y}_l, \boldsymbol{y}_{l-1}, \ldots, \boldsymbol{y}_k\}. \tag{3.8}$$

For arbitrary $l \in \mathbb{N}$, assume that the sequence of observation error

$$\{\boldsymbol{\epsilon}_l, \boldsymbol{\epsilon}_{l-1}, \ldots, \boldsymbol{\epsilon}_{-1}\} \tag{3.9}$$

is independent-in-time (i.e. a white process).

This formulation is a type of *hidden Markov model*, where the dynamic state variables $\boldsymbol{x}_k$ are known as the hidden variables because they are not directly observed. A Markov model is a type of 'memoryless' process, described this way because of how the conditional probability for the state is represented between different times (Ross, 2014, chapter 4). Particularly, if $\boldsymbol{x}_{k:1}$ is a Markov process, the Markov property is defined as

$$p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1:0}) = p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}). \tag{3.10}$$

Equation (3.10) says that, given knowledge of the state $\boldsymbol{x}_{k-1}$, the conditional probability for $\boldsymbol{x}_k$ is independent of the past history of the state before time t_{k-1}, representing the probabilistic analogue of an initial value problem.

Applying the Markov property recursively with the definition of the conditional pdf yields

$$p(\boldsymbol{x}_{L:0}) = p(\boldsymbol{x}_0)\prod_{k=1}^{L} p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}). \tag{3.11}$$

Therefore, the joint pdf for a forecast of the model state can be written as the product of the first prior at time t_0, representing the uncertainty of the data used to initialise the model forecast, and the product of the *Markov transition pdfs*, describing the evolution of the state between discrete times.

With the perfect state model, as in Eq. (3.7a), the transition probability for some subset $\mathrm{d}\boldsymbol{x} \subset \mathbb{R}^{N_x}$ is written

$$\mathcal{P}(\boldsymbol{x}_k \in \mathrm{d}\boldsymbol{x}|\boldsymbol{x}_{k-1}) = \delta_{\mathcal{M}_k(\boldsymbol{x}_{k-1})}(\mathrm{d}\boldsymbol{x}), \tag{3.12}$$

with $\delta_{\boldsymbol{v}}$ referring to the *Dirac measure* at $\boldsymbol{v} \in \mathbb{R}^{N_x}$. The Dirac measure satisfies

$$\int f(\boldsymbol{x})\delta_{\boldsymbol{v}}(\mathrm{d}\boldsymbol{x}) = f(\boldsymbol{v}), \tag{3.13}$$

where this is a singular measure, to be understood by the integral equation. Accordingly, the transition pdf is often written proportional as

$$p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}) \propto \delta\{\boldsymbol{x}_k - \mathcal{M}_k(\boldsymbol{x}_{k-1})\}, \tag{3.14}$$

where δ represents the *Dirac distribution*. Heuristically, this is known as the 'function' defined by the property

$$\delta_\epsilon(\boldsymbol{x}) = \begin{cases} \dfrac{1}{\epsilon} & \boldsymbol{x} \in [-\epsilon, +\epsilon] \\ 0 & \text{else} \end{cases}, \quad \delta(\boldsymbol{x}) = \lim_{\epsilon \to 0^+} \delta_\epsilon(\boldsymbol{x}). \tag{3.15}$$

However, this is just a convenient abuse of notations, as the Dirac measure does not have a pdf with respect to the standard Lebesgue measure. Rather, the Dirac distribution is understood through the generalised function theory of distributions (Taylor, 1996, section 3.4) as a type of integral kernel satisfying

$$\int f(\boldsymbol{x}_k)\delta\{\boldsymbol{x}_k - \mathcal{M}_k(\boldsymbol{x}_{k-1})\}\mathrm{d}\boldsymbol{x}_k = f(\mathcal{M}_k(\boldsymbol{x}_{k-1})). \tag{3.16}$$

Equation (3.16) represents the existence and uniqueness of the solution to an initial value problem in deterministic systems of ordinary and partial differential equations, where the knowledge of the state $\boldsymbol{x}_{k-1}$ completely determines the state $\boldsymbol{x}_k$ via the perfect forecast model $\mathcal{M}_k$. Particularly, there is probability 1 of the state following the unique solution to the time evolution, and probability 0 of all other outcomes. Each Markov transition pdf represents the evolution of an initial condition with respect to the dynamical model, conditional ultimately on an uncertain outcome of the initial data from the first prior. While the perfect model assumption is used for simplicity, the decomposition of the forecast pdf can be derived for erroneous models under the additional assumption that the model errors are independent-in-time. Like the decomposition of the forecast pdf, for the observation likelihood we can write

$$p(\boldsymbol{y}_k|\boldsymbol{x}_k, \boldsymbol{y}_{k-1:1}) = p(\boldsymbol{y}_k|\boldsymbol{x}_k), \tag{3.17}$$

due to the independence assumption on the observation errors, and the relationship between $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$. This says that the knowledge of the physical state $\boldsymbol{x}_k$ completely determines the likelihood of the observation $\boldsymbol{y}_k$. Indeed,

$$\boldsymbol{\epsilon}_k = \boldsymbol{y}_k - \mathcal{H}_k(\boldsymbol{x}_k), \tag{3.18}$$

which follows a known distribution and is independent of the other observation error outcomes by assumption.

Consider thus how to estimate the filtering pdf $p(\boldsymbol{x}_k|\boldsymbol{y}_{k:1})$. Using the definition of the conditional pdf, one has

$$p(\boldsymbol{x}_k|\boldsymbol{y}_{k:1}) = \frac{p(\boldsymbol{y}_{k:1}, \boldsymbol{x}_k)}{p(\boldsymbol{y}_{k:1})}. \tag{3.19}$$

Rewriting these pdfs as conditional pdfs, and by using the independence assumption,

$$p(\boldsymbol{x}_k|\boldsymbol{y}_{k:1}) = \frac{p(\boldsymbol{y}_k, (\boldsymbol{x}_k, \boldsymbol{y}_{k-1:1}))}{p(\boldsymbol{y}_{k:1})} \tag{3.20a}$$

$$= \frac{p(\boldsymbol{y}_k|\boldsymbol{x}_k, \boldsymbol{y}_{k-1:1})p(\boldsymbol{x}_k, \boldsymbol{y}_{k-1:1})}{p(\boldsymbol{y}_{k:1})}$$

$$= \frac{p(\boldsymbol{y}_k|\boldsymbol{x}_k)p(\boldsymbol{x}_k, \boldsymbol{y}_{k-1:1})}{p(\boldsymbol{y}_{k:1})}. \tag{3.20b}$$

Writing the joint pdfs again in terms of conditional pdfs

$$p(\boldsymbol{x}_k|\boldsymbol{y}_{k:1}) = \frac{p(\boldsymbol{y}_k|\boldsymbol{x}_k)p(\boldsymbol{x}_k|\boldsymbol{y}_{k-1:1})p(\boldsymbol{y}_{k-1:1})}{p(\boldsymbol{y}_k|\boldsymbol{y}_{k-1:1})p(\boldsymbol{y}_{k-1:1})} \tag{3.21a}$$

$$= \frac{p(\boldsymbol{y}_k|\boldsymbol{x}_k)p(\boldsymbol{x}_k|\boldsymbol{y}_{k-1:1})}{p(\boldsymbol{y}_k|\boldsymbol{y}_{k-1:1})}. \tag{3.21b}$$

Now, suppose that the posterior pdf $p(\boldsymbol{x}_{k-1}|\boldsymbol{y}_{k-1:1})$ at the last observation time t_{k-1} is already computed – then the model forecast of this pdf is given by averaging over the state at time t_{k-1} with respect to the Markov transition pdf,

$$p(\boldsymbol{x}_k|\boldsymbol{y}_{k-1:1}) = \int p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})p(\boldsymbol{x}_{k-1}|\boldsymbol{y}_{k-1:1})\,\mathrm{d}\boldsymbol{x}_{k-1}, \tag{3.22}$$

yielding the forecast-prior. The filtering pdf, on the left-hand-side of Eq. (3.21) is written in terms of: (i) the likelihood of the observed data given the model forecast, $p(\boldsymbol{y}_k|\boldsymbol{x}_k)$; (ii) the forecast-prior given the last best estimate of the state, $p(\boldsymbol{x}_k|\boldsymbol{y}_{k-1:1})$; and (iii) the marginal of the joint pdf $p(\boldsymbol{y}_k, \boldsymbol{x}_k|\boldsymbol{y}_{k-1:1})$, integrating out the hidden variables,

$$p(\boldsymbol{y}_k|\boldsymbol{y}_{k-1:1}) = \int p(\boldsymbol{y}_k|\boldsymbol{x}_k)p(\boldsymbol{x}_k|\boldsymbol{y}_{k-1:1})\,\mathrm{d}\boldsymbol{x}_k. \tag{3.23}$$

This type of pdf, only depending on the observations, is called an *evidence* (e.g. Carrassi et al., 2017).

Typically, the pdf in the denominator of Eq. (3.21) is mathematically intractable. However, the denominator is independent of the hidden variable $\boldsymbol{x}_k$ by construction – the free argument in the pdf on the left-hand-side is the model state $\boldsymbol{x}_k$ and the purpose of the denominator on the right-hand-side is only to normalise the integral of the posterior pdf to 1. Instead, as a proportionality statement,

$$p(\boldsymbol{x}_k|\boldsymbol{y}_{k:1}) \propto p(\boldsymbol{y}_k|\boldsymbol{x}_k)p(\boldsymbol{x}_k|\boldsymbol{y}_{k-1:1}), \tag{3.24}$$

one can devise the Bayesian MAP estimate as the choice of $\overline{\boldsymbol{x}}_k$ that maximises the posterior pdf, but written in terms of the two right-hand-side components in Eq. (3.24). For the purpose of maximising the posterior

pdf, the denominator leads to insignificant constants that can be discarded. Thus, in order to compute the MAP sequentially and recursively in time, one can develop a recursion in proportionality. However, note that the evidence can be estimated and used for other significant purposes still within a Bayesian framework (Carrassi et al., 2017).

3.2.3 Linear-Gaussian Models

Generally, the filtering pdf $p(\boldsymbol{x}_k|\boldsymbol{y}_{k:1})$ has no analytical solution (i.e. no explicit expression). However, when the models are linear, that is, both the state and observation models are written as matrix actions

$$\boldsymbol{x}_k = \mathbf{M}_k\boldsymbol{x}_{k-1}, \tag{3.25a}$$

$$\boldsymbol{y}_k = \mathbf{H}_k\boldsymbol{x}_k + \boldsymbol{\epsilon}_k, \tag{3.25b}$$

and the error pdfs are *Gaussian*:

$$p(\boldsymbol{x}_0) = n(\boldsymbol{x}_0|\overline{\boldsymbol{x}}_0, \mathbf{B}_0),\ p(\boldsymbol{y}_k|\boldsymbol{x}_k) = n(\boldsymbol{y}_k|\mathbf{H}_k\boldsymbol{x}_k, \mathbf{R}_k), \tag{3.26}$$

$$n(\boldsymbol{z}|\overline{\boldsymbol{z}}, \mathbf{B}) := \frac{1}{\sqrt{(2\pi)^{N_z}\det(\mathbf{B})}}\exp\left\{-\frac{1}{2}(\overline{\boldsymbol{z}} - \boldsymbol{z})^{\mathrm{T}}\mathbf{B}^{-1}(\overline{\boldsymbol{z}} - \boldsymbol{z})\right\},$$
$$\tag{3.27}$$

then the forecast and posterior pdfs are Gaussian at all times, and are parametrised in terms of their mean and covariance. In particular, Gaussian distributions are closed under affine transformations, that is, maps of the form

$$\boldsymbol{f}(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x} + \boldsymbol{b}, \tag{3.28}$$

corresponding to a linear transformation when $\boldsymbol{b}$ is a vector of zeros, such that $\boldsymbol{b} = \mathbf{0}$ (Tong, 2012, see theorem 3.3.3). If $\boldsymbol{x}$ is distributed with pdf $p(\boldsymbol{x}) = n(\boldsymbol{x}|\overline{\boldsymbol{x}}, \mathbf{B})$, then the random vector $\boldsymbol{y} := \mathbf{A}\boldsymbol{x} + \boldsymbol{b}$ is distributed with pdf

$$p(\boldsymbol{y}) = n(\boldsymbol{y}|\mathbf{A}\overline{\boldsymbol{x}} + \boldsymbol{b}, \mathbf{A}\mathbf{B}\mathbf{A}^{\mathrm{T}}). \tag{3.29}$$

Suppose that the last analysis pdf is given as

$$p(\boldsymbol{x}_{k-1}|\boldsymbol{y}_{k-1:1}) = n(\boldsymbol{x}_{k-1}|\overline{\boldsymbol{x}}_{k-1}^{\mathrm{a}}, \mathbf{B}_{k-1}^{\mathrm{a}}), \tag{3.30}$$

parametrised in terms of the analysis mean $\overline{\boldsymbol{x}}_{k-1}^{\mathrm{a}}$ and analysis error covariance $\mathbf{B}_{k-1}^{\mathrm{a}}$. Then, the forecast-prior pdf is written as

$$p(\boldsymbol{x}_k|\boldsymbol{y}_{k-1:1}) = n(\boldsymbol{x}_k|\overline{\boldsymbol{x}}_{k-1}^{\mathrm{f}}, \mathbf{B}_k^{\mathrm{f}}), \tag{3.31}$$

where the forecast mean and forecast error covariance are defined by

$$\overline{\boldsymbol{x}}_k^{\mathrm{f}} := \mathbf{M}_k\overline{\boldsymbol{x}}_{k-1}^{\mathrm{a}}, \tag{3.32a}$$

$$\mathbf{B}_k^{\mathrm{f}} := \mathbf{M}_k\mathbf{B}_{k-1}^{\mathrm{a}}\mathbf{M}_k^{\mathrm{T}}, \tag{3.32b}$$

respectively.

Similarly, the conditional and marginal distributions of a Gaussian random vector are also Gaussian and their pdf has an analytical form. Suppose that $z \in \mathbb{R}^{N_z}$ is an arbitrary Gaussian random vector, partitioned as

$$z := \begin{pmatrix} x \\ y \end{pmatrix} \quad p(z) := n\left(z \middle| \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}\right), \quad (3.33)$$

with the dimensions given as $N_z = N_x + N_y$ and

$$x, \bar{x} \in \mathbb{R}^{N_x}, \quad y, \bar{y} \in \mathbb{R}^{N_y}, \quad \Sigma_{xx} \in \mathbb{R}^{N_x \times N_x}, \quad (3.34a)$$

$$\Sigma_{xy} = \Sigma_{yx}^{\mathrm{T}} \in \mathbb{R}^{N_x \times N_y}, \quad \Sigma_{yy} \in \mathbb{R}^{N_y \times N_y}. \quad (3.34b)$$

Then, the general form of the pdf for x conditioned on the outcome of y is given by the Gaussian

$$p(x|y) = n\left(x|\bar{x} + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \bar{y}), \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\right), \quad (3.35)$$

where the covariance matrix $\Sigma_{xx} - \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx}$ is called the *Schur complement* (see, e.g., theorem 3.3.4 of Tong, 2012). Noting the form of the linear observation model in Eq. (3.25), and the independence of the observation errors, it is easy to see that the vector composed of $\begin{pmatrix} x_k^\top & y_k^\top \end{pmatrix}^\top$ is jointly Gaussian. Relying on the identifications

$$\Sigma_{xy} = \mathbf{B}_k^{\mathrm{f}}\mathbf{H}_k^\top, \quad \Sigma_{yy} = \mathbf{H}_k\mathbf{B}_k^{\mathrm{f}}\mathbf{H}_k^\top + \mathbf{R}_k, \quad (3.36)$$

the posterior pdf at time t_k is derived as the Gaussian with analysis mean and analysis error covariance given by

$$\bar{x}_k^{\mathrm{a}} := \bar{x}_k^{\mathrm{f}} + \mathbf{B}_k^{\mathrm{f}}\mathbf{H}_k^\top \left(\mathbf{H}_k\mathbf{B}_k^{\mathrm{f}}\mathbf{H}_k^\top + \mathbf{R}_k\right)^{-1}(y_k - \mathbf{H}_k\bar{x}_k^{\mathrm{f}}), \quad (3.37a)$$

$$\mathbf{B}_k^{\mathrm{a}} := \mathbf{B}_k^{\mathrm{f}} - \mathbf{B}_k^{\mathrm{f}}\mathbf{H}_k^\top \left(\mathbf{H}_k\mathbf{B}_k^{\mathrm{f}}\mathbf{H}_k^\top + \mathbf{R}_k\right)^{-1}\mathbf{H}_k\mathbf{B}_k^{\mathrm{f}}. \quad (3.37b)$$

Defining $\mathbf{K}_k := \mathbf{B}_k^{\mathrm{f}}\mathbf{H}_k^\top \left(\mathbf{H}_k\mathbf{B}_k^{\mathrm{f}}\mathbf{H}_k^\top + \mathbf{R}_k\right)^{-1}$ as the Kalman gain, Eqs. (3.37a,b) yield the classical *Kalman filter* (KF) update, and this derivation inductively defines the forecast and posterior distribution for x_k at all times. The KF is recognised thus as the parametric representation of a linear-Gaussian hidden Markov model, providing a recursion on the first two moments for the forecast and posterior.

This analysis extends, as with the classical KF, easily to incorporate additive model errors as in Eq. (3.1); see, for example, Anderson and Moore (1979). However, there are many ways to formulate the KF and there are some drawbacks of this approach. Even when the model forecast equations themselves are linear, if they functionally depend on an uncertain parameter vector, defined $\mathbf{M}_k(\lambda)$, the joint estimation problem of (x_k, λ) can become highly non-linear and an iterative approach to the joint estimation may be favourable. Therefore, this chapter develops the subsequent extensions to the KF with the MAP approach, which coincides with the development in least-squares and non-linear optimisation.

3.3 Least-Squares and Non-linear Optimisation

3.3.1 The 3-D Cost Function from Gaussian Statistics

As seen in the previous section, the linear-Gaussian analysis admits an analytical solution for the forecast and posterior pdf at all times. However, an alternative approach for deriving the KF is formed using the MAP estimation in proportionality. Consider that the *natural logarithm* (log) is monotonic, so that an increase in the input of the argument corresponds identically to an increase in the output of the function. Therefore, maximising the posterior pdf as in Eq. (3.24) is equivalent to maximising

$$\log(p(y_k|x_k)p(x_k|y_{k-1:1})) = \log(p(y_k|x_k)) \\ + \log(p(x_k|y_{k-1:1})). \quad (3.38)$$

Given the form of the multivariate Gaussian pdf in Eq. (3.27), maximising the log-posterior Eq. (3.38) is equivalent to minimising the following least-squares cost function, derived in proportionality to the minus-log-posterior:

$$\mathcal{J}_{\mathrm{KF}}(x_k) = \frac{1}{2}\|\bar{x}_k^{\mathrm{f}} - x_k\|_{\mathbf{B}_k^{\mathrm{f}}}^2 + \frac{1}{2}\|y_k - \mathbf{H}_k x_k\|_{\mathbf{R}_k}^2. \quad (3.39)$$

For an arbitrary positive definite matrix $\mathbf{A}$, the *Mahalanobis distance* (Mahalanobis, 1936) with respect to $\mathbf{A}$ is defined as

$$\|v\|_{\mathbf{A}} := \sqrt{v^{\mathrm{T}}\mathbf{A}^{-1}v}. \quad (3.40)$$

These distances are weighted Euclidean norms with: (i) $\|\circ\|_{\mathbf{B}_k^{\mathrm{f}}}$ weighting relative to the forecast spread; and (ii) $\|\circ\|_{\mathbf{R}_k}$ weighting relative to the observation imprecision. The MAP state thus interpolates the forecast mean and the observation relative to the uncertainty in each piece of data. Due to the unimodality of the Gaussian, and its symmetry about its mean, it is clear that the conditional mean $\bar{x}_k^{\mathrm{a}}$ is also the MAP state.

While this cost function analysis provides the solution to finding the first moment of the Gaussian posterior, this does not yet address how to find the posterior error covariance. In this linear-Gaussian setting, it is easily shown that the analysis error covariance is actually given by

$$\mathbf{B}_k^{\mathrm{a}} := \left[(\mathbf{B}_k^{\mathrm{f}})^{-1} + \mathbf{H}_k^\top\mathbf{R}_k^{-1}\mathbf{H}_k\right]^{-1} = \Xi_{\mathcal{J}_{\mathrm{KF}}}^{-1}, \quad (3.41)$$

where $\Xi_{\mathcal{J}_{\mathrm{KF}}}$ refers to the *Hessian* of the least-squares cost function in Eq. (3.39), that is, the matrix of its mixed second partial derivatives in the state vector variables. This is a fundamental result that links the recursive analysis in time to the MAP estimation performed with linear least-squares; see, for example, section 6.2 of Reich and Cotter (2015) for further details.

A more general result from maximum likelihood estimation extends this analysis as an approximation to non-linear state and observation models, and to non-Gaussian error distributions. Define $\hat{\theta}$ to be the *maximum likelihood estimator* (MLE) for some unknown parameter vector $\theta \in \mathbb{R}^{N_\theta}$,

where $\hat{\boldsymbol{\theta}}$ depends on the realisation of some arbitrarily distributed random sample of observed data $\{z_i\}_{i=1}^{N}$. Then, under fairly general regularity conditions, the MLE satisfies

$$\mathcal{I}(\boldsymbol{\theta})^{-\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \rightarrow_d \mathcal{N}(\mathbf{0}, \mathbf{I}_{N_\theta}), \tag{3.42}$$

where: (i) Eq. (3.42) refers to convergence in distribution of the random vector $\mathcal{I}(\boldsymbol{\theta})^{-\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ as the sample size $N \rightarrow \infty$; (ii) $\mathbf{I}_{N_\theta}$ is the identity matrix in the dimension of $\boldsymbol{\theta}$; (iii) $\mathcal{N}(\mathbf{0}, \mathbf{I}_{N_\theta})$ is the multivariate Gaussian distribution with mean zero and covariance equal to the identity matrix, that is, with pdf $n(\boldsymbol{x}|\mathbf{0}, \mathbf{I}_{N_\theta})$; and (iv) $\mathcal{I}(\boldsymbol{\theta})$ is the *Fisher information matrix*. The Fisher information matrix is defined as the expected value of the Hessian of the minus-log-likelihood, taken over the realisations of the observed data $\{z_i\}_{i=1}^{N}$, and with respect to the true parameter vector $\boldsymbol{\theta}$. It is common to approximate the distribution of the MLE as

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}, I(\hat{\boldsymbol{\theta}})), \tag{3.43}$$

where $I(\hat{\boldsymbol{\theta}})$ is the observed Fisher information (i.e. the realisation of the Hessian of the minus-log-likelihood given the observed sample data). This approximation improves in large sample sizes like the central limit theorem; see, for example, chapter 9 of Pawitan (2001) for the details of this discussion. In non-linear optimisation, the approximate distribution of the MLE therefore relates the geometry in the neighbourhood of a local minimiser to the variation in the optimal estimate. Particularly, the curvature of the cost function level contours around the local minimum is described by the Hessian while the spread of the distribution of the MLE is described by the analysis error covariance.

3.3.2 3D-VAR and the Extended Kalman Filter

One of the benefits of the aforementioned cost function approach is that this immediately extends to handle a non-linear observation operator $\mathcal{H}_k$ as a problem of non-linear least-squares. When the observations are related non-linearly to the state vector, the Bayesian posterior is no longer generally Gaussian, but the interpretation of the analysis state $\bar{\boldsymbol{x}}_k^{\mathrm{a}}$ interpolating between a background proposal $\bar{\boldsymbol{x}}_k^{\mathrm{f}}$ and the observed data $\boldsymbol{y}_k$ relative to their respective uncertainties remains valid. The use of the non-linear least-squares cost function can be considered as making a Gaussian approximation, similar to the large sample theory in maximum likelihood estimation. However, the forecast and analysis background error covariances, $\mathbf{B}_k^{\mathrm{f/a}}$, take on different interpretations depending on the DA scheme.

In full-scale geophysical models, the computation and storage of the background error covariance $\mathbf{B}_k^{\mathrm{f/a}}$ is rarely feasible due to its large size; in practice, this is usually treated abstractly in its action as an operator by preconditioning the optimisation (Tabeart et al., 2018). One traditional approach to handle this reduced representation is to use a static-in-time background for the cost function in Eq.

(3.39), rendering the *three-dimensional, variational* (3D-VAR) cost function

$$\mathcal{J}_{\mathrm{3D-VAR}}(\boldsymbol{x}) := \frac{1}{2}\|\bar{\boldsymbol{x}}_k^{\mathrm{f}} - \boldsymbol{x}\|_{\mathbf{B}_{\mathrm{3D-VAR}}}^2 + \frac{1}{2}\|\boldsymbol{y}_k - \mathcal{H}_k(\boldsymbol{x})\|_{\mathbf{R}_k}^2. \tag{3.44}$$

The 3D-VAR background error covariance is typically defined as a posterior, 'climatological' covariance, taken with respect to a long-time average over the modelled state. One way that this is roughly estimated in meteorology is by averaging over a *reanalysis data set*, which is equivalent to averaging with respect to a joint-smoothing posterior over a long history for the system (Kalnay, 2003, and references therein). If one assumes that the DA cycle represents a stationary, ergodic process in its posterior statistics, and that this admits an invariant, ergodic measure, π^*, one would define the 3D-VAR background error covariance matrix with the expected value with respect to this measure as

$$\bar{\boldsymbol{x}}^* := \mathbb{E}_{\pi^*}[\boldsymbol{x}], \tag{3.45a}$$

$$\mathbf{B}_{\mathrm{3D-VAR}} := \mathbb{E}_{\pi^*}\left[(\boldsymbol{x} - \bar{\boldsymbol{x}}^*)(\boldsymbol{x} - \bar{\boldsymbol{x}}^*)^\top\right]. \tag{3.45b}$$

The 3D-VAR cycle is defined where $\bar{\boldsymbol{x}}_k^{\mathrm{a}}$ is the *minimising argument* (argmin) of Eq. (3.44) and subsequently $\bar{\boldsymbol{x}}_{k+1}^{\mathrm{f}} := \mathcal{M}_{k+1}(\bar{\boldsymbol{x}}_k^{\mathrm{a}})$, meaning, the forecast mean is estimated with the background control trajectory propagated through the fully non-linear model. Therefore, at every iteration of the algorithm, 3D-VAR can be understood to treat the proposal $\bar{\boldsymbol{x}}_k^{\mathrm{f}}$ as a random draw from the invariant, climatological-posterior measure of the system; the optimised solution $\bar{\boldsymbol{x}}_k^{\mathrm{a}}$ is thus the state that interpolates the forecast mean and the currently observed data as if the forecast mean was drawn randomly from the climatological-posterior.

For illustration, consider a cost-effective, explicit formulation of 3D-VAR where the background error covariance matrix is represented as a Cholesky factorisation. Here, one performs the optimisation by writing the modelled state as a perturbation of the forecast mean,

$$\boldsymbol{x}_k := \bar{\boldsymbol{x}}_k^{\mathrm{f}} + \Sigma\boldsymbol{w}, \tag{3.46}$$

where $\mathbf{B}_{\mathrm{3D-VAR}} := \Sigma\Sigma^{\mathrm{T}}$. The weight vector $\boldsymbol{w}$ gives the linear combination of the columns of the matrix factor that describe $\boldsymbol{x}_k$ as a perturbation. One can iteratively optimise the 3D-VAR cost function as such with a locally quadratic approximation. Let $\bar{\boldsymbol{x}}_k^i$ be the i-th iterate for the proposal value at time t_k, defining $\bar{\boldsymbol{x}}_k^0 := \bar{\boldsymbol{x}}_k^{\mathrm{f}}$. The $i + 1$-st iteration is defined (up to a subtlety pointed out by Ménétrier and Auligné (2015)) as

$$\bar{\boldsymbol{x}}_k^{i+1} = \bar{\boldsymbol{x}}_k^i + \Sigma\bar{\boldsymbol{w}}^{i+1}, \tag{3.47}$$

where $\bar{\boldsymbol{w}}^{i+1}$ is the argmin of the following *incremental cost function*

$$\mathcal{J}_{\text{3D-VARI}}(\boldsymbol{w}) := \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{1}{2}\|\boldsymbol{y}_k - \mathcal{H}_k(\overline{\boldsymbol{x}}_k^{\text{i}}) - \mathbf{H}_k\boldsymbol{\Sigma}\boldsymbol{w}\|_{\mathbf{R}_k}^2. \tag{3.48}$$

This approximate cost function follows from truncating the Taylor expansion

$$\mathcal{H}_k(\boldsymbol{x}) = \mathcal{H}_k(\overline{\boldsymbol{x}}_k^{\text{i}}) + \mathbf{H}_k\boldsymbol{\Sigma}\boldsymbol{w} + \mathcal{O}(\|\boldsymbol{w}\|^2) \tag{3.49}$$

and the definition of the Mahalanobis distance. Notice that $\overline{\boldsymbol{w}}^{i+1}$ is defined uniquely as Eq. (3.48) is quadratic with respect to the weights, allowing one to use fast and adequate methods such as the conjugate gradient (Nocedal and Wright, 2006, chapter 5). The iterations are set to terminate when $\|\overline{\boldsymbol{w}}^{\text{i}}\|$ is sufficiently small, representing a negligible change from the last proposal. This approach to non-linear optimisation is known as the incremental approach, and corresponds mathematically to the Gauss-Newton optimisation method, an efficient simplification of Newton's descent method for non-linear least-squares (Nocedal and Wright, 2006, chapter 10). As described, it consists of an iterative linearisation of the non-linear optimisation problem, which is a concept utilised throughout this chapter.

The 3D-VAR approach described here played an important role in early DA methodology, and aspects of this technique are still widely used in ensemble-based DA when using the regularisation technique known as covariance hybridisation (Hamill and Snyder, 2000; Lorenc, 2003; Penny, 2017). However, optimising the state with the climatological-posterior background neglects important information in the time-dependent features of the forecast spread, described sometimes as the 'errors of the day' (Corazza et al., 2003). The *extended Kalman filter* (EKF) can be interpreted to generalise the 3D-VAR cost function by including a time-dependent background, like the KF, by approximating its time-evolution at first order with the tangent linear model as follows.

Suppose that the equations of motion are generated by a non-linear function, independent of time for simplicity,

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x} := \boldsymbol{f}(\boldsymbol{x}), \quad \boldsymbol{x}_k = \mathcal{M}_k(\boldsymbol{x}_{k-1}) := \int_{t_{k-1}}^{t_k}\boldsymbol{f}(\boldsymbol{x})\,\mathrm{d}t + \boldsymbol{x}_{k-1}. \tag{3.50}$$

One can extend the linear-Gaussian approximation for the forecast pdf by modelling the state as a perturbation of the analysis mean,

$$\boldsymbol{x}_{k-1} := \overline{\boldsymbol{x}}_{k-1}^{\text{a}} + \boldsymbol{\delta}_{k-1} \sim \mathcal{N}(\overline{\boldsymbol{x}}_{k-1}^{\text{a}}, \mathbf{B}_{k-1}^{\text{a}}), \tag{3.51a}$$

$$\Leftrightarrow \boldsymbol{\delta}_{k-1} \sim \mathcal{N}(\mathbf{0}, \mathbf{B}_{k-1}^{\text{a}}). \tag{3.51b}$$

The evolution of the perturbation $\boldsymbol{\delta}_{k-1}$ is written via Taylor's theorem as

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\delta}_{k-1} := \frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{x}_{k-1} - \overline{\boldsymbol{x}}_{k-1}^{\text{a}}) \tag{3.52a}$$

$$= \boldsymbol{f}(\boldsymbol{x}_{k-1}) - \boldsymbol{f}(\overline{\boldsymbol{x}}_{k-1}^{\text{a}}) \tag{3.52b}$$

$$= \nabla_{\boldsymbol{x}}\boldsymbol{f}(\overline{\boldsymbol{x}}_{k-1}^{\text{a}})\boldsymbol{\delta}_{k-1} + \mathcal{O}(\|\boldsymbol{\delta}_{k-1}\|^2), \tag{3.52c}$$

where $\nabla_{\boldsymbol{x}}\boldsymbol{f}(\overline{\boldsymbol{x}}_{k-1}^{\text{a}})$ is the Jacobian equation with dependence on the underlying analysis mean state. The linear evolution defined by the truncated Taylor expansion about the underlying reference solution $\overline{\boldsymbol{x}}$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\delta} := \nabla_{\boldsymbol{x}}\boldsymbol{f}(\overline{\boldsymbol{x}}) \cdot \boldsymbol{\delta}, \tag{3.53}$$

is known as the *tangent linear model*.

Making the approximation of the tangent linear model for the first order evolution of the modelled state, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x} \approx \boldsymbol{f}(\overline{\boldsymbol{x}}) + \nabla_{\boldsymbol{x}}\boldsymbol{f}(\overline{\boldsymbol{x}}) \cdot \boldsymbol{\delta} \tag{3.54a}$$

$$\Rightarrow \int_{t_{k-1}}^{t_k}\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}\,\mathrm{d}t \approx \int_{t_{k-1}}^{t_k}\boldsymbol{f}(\overline{\boldsymbol{x}})\,\mathrm{d}t + \int_{t_{k-1}}^{t_k}\nabla_{\boldsymbol{x}}\boldsymbol{f}(\overline{\boldsymbol{x}}) \cdot \boldsymbol{\delta}\,\mathrm{d}t \tag{3.54b}$$

$$\Rightarrow \boldsymbol{x}_k \approx \mathcal{M}_k(\overline{\boldsymbol{x}}_{k-1}) + \mathbf{M}_k\boldsymbol{\delta}_{k-1}, \tag{3.54c}$$

where $\mathbf{M}_k$ is the resolvent of the tangent linear model. Given that Gaussians are closed under affine transformations, the EKF approximation for the (perfect) evolution of the state vector is defined as

$$p(\boldsymbol{x}_k|\boldsymbol{y}_{k-1:1}) \approx n\left(\boldsymbol{x}_k|\mathcal{M}_k(\overline{\boldsymbol{x}}_{k-1}^{\text{a}}), \mathbf{M}_k\mathbf{B}_{k-1}^{\text{a}}\mathbf{M}_k^\top\right). \tag{3.55}$$

Respectively, define the EKF analysis cost function as

$$\mathcal{J}_{\text{EKF}}(\boldsymbol{x}) := \frac{1}{2}\|\overline{\boldsymbol{x}}_k^{\text{f}} - \boldsymbol{x}\|_{\mathbf{B}_k^{\text{f}}}^2 + \frac{1}{2}\|\boldsymbol{y}_k - \mathcal{H}_k(\boldsymbol{x})\|_{\mathbf{R}_k}^2, \tag{3.56}$$

where

$$\overline{\boldsymbol{x}}_k^{\text{f}} := \mathcal{M}_k(\overline{\boldsymbol{x}}_{k-1}^{\text{a}}), \tag{3.57a}$$

$$\mathbf{B}_k^{\text{f}} := \mathbf{M}_k\mathbf{B}_{k-1}^{\text{a}}\mathbf{M}_k^\top. \tag{3.57b}$$

A similar, locally quadratic, weight–space optimisation, can be performed versus the non-linear observation operator using a matrix decomposition for the time-varying background forecast error covariance, defining the state as a perturbation using this matrix factor similar to Eq. (3.46). This gives the square root EKF, which was used historically to improve the stability over the direct approach (Tippett et al., 2003).

The accuracy and stability of this EKF depends strongly on the length of the forecast horizon Δ_t (Miller et al., 1994). For short-range forecasting, the perturbation dynamics of the tangent linear model can be an adequate approximation, and is an underlying approximation for most operational

DA (Carrassi et al., 2018). However, the explicit linear approximation can degrade quickly, and especially when the mean state is not accurately known. The perturbation dynamics in the linearisation about the approximate mean can differ substantially from the non-linear dynamics of the true system as the approximate mean state diverges from the modelled system (Grewal and Andrews, 2014, see chapters 7 and 8). These stability issues, and the computational cost of explicitly representing the evolution of the background covariance in the tangent linear model, limit the use of the EKF for geophysical DA. Nonetheless, this technique provides important intuition that is developed later in this chapter.

3.3.3 The Gaussian 4-D Cost Function

Consider again the perfect, linear-Gaussian model represented by Eq. (3.25). Using the Markov assumption, Eq. (3.10), and the independence of observation errors, Eq. (3.17), recursively for the hidden Markov model, the joint posterior decomposes as

$$p(\boldsymbol{x}_{L:0}|\boldsymbol{y}_{L:1}) \propto \underbrace{p(\boldsymbol{x}_0)}_{(i)} \underbrace{\left[\prod_{k=1}^{L} p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}) \right]}_{(ii)} \underbrace{\left[\prod_{k=1}^{L} p(\boldsymbol{y}_k|\boldsymbol{x}_k) \right]}_{(iii)},$$

(3.58)

where (i) is the prior for the initial state $\boldsymbol{x}_0$; (ii) is the free-forecast with the perfect model $\mathbf{M}_k$, depending on some outcome drawn from the prior; and (iii) is the joint likelihood of the observations in the DAW, given the background forecast.

Define the composition of the linear model forecast from time t_{k-1} to t_l as

$$\mathbf{M}_{l:k} := \mathbf{M}_l \cdots \mathbf{M}_k, \qquad \mathbf{M}_{k:k} := \mathbf{I}_{N_x}. \quad (3.59)$$

Using the perfect, linear model hypothesis, note that

$$\boldsymbol{x}_k := \mathbf{M}_{k:1}\boldsymbol{x}_0 \quad (3.60)$$

for every k. Therefore, the transition pdfs in Eq. (3.58) are reduced to a trivial condition by re-writing

$$p(\boldsymbol{x}_{L:0}|\boldsymbol{y}_{L:1}) \propto p(\boldsymbol{x}_0) \left[\prod_{k=1}^{L} p(\boldsymbol{x}_k|\mathbf{M}_{k-1:1}\boldsymbol{x}_0) \right] \left[\prod_{k=1}^{L} p(\boldsymbol{y}_k|\boldsymbol{x}_k) \right]$$

(3.61a)

$$\propto p(\boldsymbol{x}_0) \left[\prod_{k=1}^{L} p(\boldsymbol{y}_k|\mathbf{M}_{k:1}\boldsymbol{x}_0) \right], \quad (3.61b)$$

as this pdf evaluates to zero whenever $\boldsymbol{x}_k \neq \mathbf{M}_{k:1}\boldsymbol{x}_0$. Given a Gaussian prior, and Gaussian observation error distributions, the minus-log-posterior *four-dimensional* (4-D) cost function is derived as

$$\mathcal{J}_{4\mathrm{D}}(\boldsymbol{x}_0) := \frac{1}{2}\|\bar{\boldsymbol{x}}_0 - \boldsymbol{x}_0\|_{\mathbf{B}_0}^2 + \frac{1}{2}\sum_{k=1}^{L}\|\boldsymbol{y}_k - \mathbf{H}_k\mathbf{M}_{k:1}\boldsymbol{x}_0\|_{\mathbf{R}_k}^2.$$

(3.62)

Notice that the 4-D cost function in Eq. (3.62) is actually quadratic with respect to the initial condition $\boldsymbol{x}_0$. Therefore, the smoothing problem in the perfect, linear-Gaussian model has a unique, optimal initial condition that minimises the sum-of-square deviations from the prior mean, with distance weighted with respect to the prior covariance, and the observations, with distance weighted with respect to the observation error covariances. The optimal, smoothed initial condition using observation information up to time t_L is denoted $\bar{\boldsymbol{x}}_{0|L}^{\mathrm{s}}$; this gives the smoothed model states at subsequent times by the perfect model evolution,

$$\bar{\boldsymbol{x}}_{k|L}^{\mathrm{s}} := \mathbf{M}_{k:1}\bar{\boldsymbol{x}}_{0|L}^{\mathrm{s}}. \quad (3.63)$$

This derivation is formulated for a smoothing problem with a static-in-time DAW, in which an entire time series of observations $\boldsymbol{y}_{L:1}$ is available, and for which the estimation may be performed 'offline'. However, a simple extension of this analysis accommodates DAWs that are sequentially shifted in time, allowing an 'online' smoothing estimate to be formed analogously to sequential filtering.

Fixed-lag smoothing sets the length of the DAW, L, to be fixed for all time, while the underlying states are cycled through this window as it shifts forward in time. Given the *lag* of length L, suppose that a *shift* S is defined for which $1 \leq S \leq L$. It is convenient to consider an algorithmically stationary DAW, referring to the time indices $\{t_1, \cdots, t_L\}$. In a given cycle, the joint posterior $p(\boldsymbol{x}_{L:1}|\boldsymbol{y}_{L:1})$ is estimated. After the estimate is produced, the DAW is subsequently shifted in time by $S \times \Delta_t$ and all states are re-indexed by $t_k \leftarrow t_{k+S}$ to begin the next cycle. For a lag of L and a shift of S, the observation vectors at times $\{t_{L-S+1}, \cdots, t_L\}$ correspond to observations newly entering the DAW for the analysis performed up to time t_L. When $S = L$, the DAWs are disconnected and adjacent in time, whereas for $S < L$ there is an overlap between the estimated states in sequential DAWs. Figure 3.2 provides a schematic of how the DAW is shifted for a lag of $L = 5$ and shift $S = 2$. Following the common convention in DA that there is no observation at time zero, in addition to the DAW, $\{t_1, \cdots, t_L\}$, states at time t_0 may be estimated or utilised in order to connect estimates between adjacent/overlapping DAWs.

Consider the algorithmically stationary DAW, and suppose that the current analysis time is t_L, where the joint posterior pdf $p(\boldsymbol{x}_{L-S:1-S}|\boldsymbol{y}_{L-S:1-S})$ is available from the last fixed-lag smoothing cycle at analysis time t_{L-S}. Using the independence of observation errors and the Markov assumption recursively,
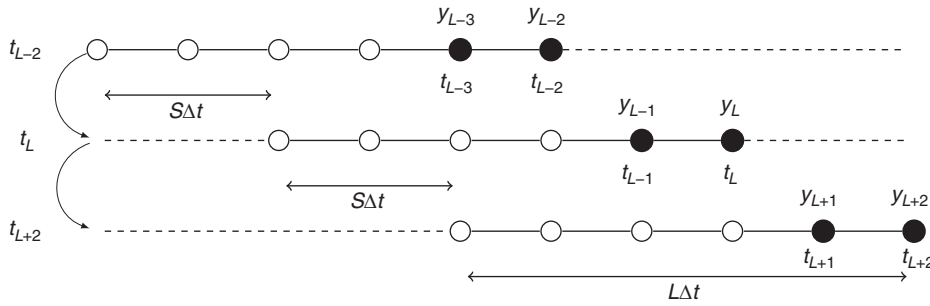
Figure 3.2 Three cycles of a shift $S = 2$, lag $L = 5$ fixed-lag smoother, cycle number is increasing top to bottom. Time indices on the left-hand margin indicate the current time for the associated cycle of the algorithm. New observations entering the current DAW are shaded black. Source: Grudzien and Bocquet (2021), adapted from Asch et al. (2016).

$$p(\boldsymbol{x}_{L:1}|\boldsymbol{y}_{L:1-S}) \propto$$

$$\int \underbrace{\mathrm{d}\boldsymbol{x}_0 \, p(\boldsymbol{x}_0|\boldsymbol{y}_{L-S:1-S})}_{(i)} \underbrace{\left[\prod_{k=1}^{L} p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})\right]}_{(ii)} \underbrace{\left[\prod_{k=L-S+1}^{L} p(\boldsymbol{y}_k|\boldsymbol{x}_k)\right]}_{(iii)},$$

$$(3.64)$$

where: (i) now represents averaging out the initial condition at time t_0 with respect to the marginal smoothing pdf for $\boldsymbol{x}_{0|L-S}^{\mathrm{s}}$ over the last DAW; (ii) represents the free forecast of the smoothed estimate for $\boldsymbol{x}_{0|L-S}^{\mathrm{s}}$; and term (iii) represents the joint likelihood of the newly incoming observations to the DAW given the forecasted model state. Noting that $p(\boldsymbol{x}_{L:1}|\boldsymbol{y}_{L:1}) \propto p(\boldsymbol{x}_{L:1}|\boldsymbol{y}_{L:1-S})$, this provides a recursive form of the 4-D cost function, shifted sequentially in time,

$$\mathcal{J}_{\mathrm{4D-seq}}(\boldsymbol{x}_0) := \frac{1}{2}\|\overline{\boldsymbol{x}}_{0|L-S}^{\mathrm{s}} - \boldsymbol{x}_0\|_{\mathbf{B}_{0|L-S}^{\mathrm{s}}}^2$$

$$+ \frac{1}{2}\sum_{k=L-S+1}^{L} \|\boldsymbol{y}_k - \mathbf{H}_k \mathbf{M}_{k:1}\boldsymbol{x}_0\|_{\mathbf{R}_k}^2, \quad (3.65)$$

where $\mathbf{B}_{0|L-S}^{\mathrm{s}}$ refers to the smoothed error covariance from the last DA cycle. As with the KF cost function, the posterior error covariance, conditioning on observations up to time t_L, is identified with

$$\mathbf{B}_{0|L}^{\mathrm{s}} := \boldsymbol{\Xi}_{\mathcal{J}_{\mathrm{4D-seq}}}^{-1}. \quad (3.66)$$

Given the perfect, linear-Gaussian model assumption, the mean and covariance are propagated to time t_S via

$$\overline{\boldsymbol{x}}_{S|L}^{\mathrm{s}} := \mathbf{M}_{S:1}\overline{\boldsymbol{x}}_{0|L}^{\mathrm{s}}, \quad (3.67a)$$

$$\mathbf{B}_{S|L}^{\mathrm{s}} := \mathbf{M}_{S:1}\mathbf{B}_{0|L}^{\mathrm{s}}\mathbf{M}_{S:1}^{\mathrm{T}}, \quad (3.67b)$$

and states are re-indexed as $t_{k+S}\leftarrow t_k$ to initialise the next cycle. This provides a 4-D derivation of the *Kalman smoother* (KS), assuming a perfect model, though not all developments take this approach, using a global analysis over all observations in the current DAW at once. Other formulations use an alternating (i) forward sequential filtering pass to update the current

forecast; and (ii) backward-in-time sequential filtering pass over the DAW to condition lagged states on the new information. Examples of smoothers that follow this alternating analysis include the *ensemble Kalman smoother* (EnKS) (Evensen and Leeuwen, 2000) and the ensemble *Rauch–Tung–Striebel* (RTS) smoother (Raanes, 2016), with the EnKS to be discussed in Section 3.4.3.

3.3.4 Incremental 4D-VAR

The method of incremental *four-dimensional, variational* (4D-VAR) data assimilation (Le Dimet and Talagrand, 1986; Talagrand and Courtier, 1987; Courtier et al., 1994) is a classical and widely used DA technique that extends the linear-Gaussian 4-D analysis to non-linear settings, both for fixed and sequential DAWs. Modern formulations of the 4D-VAR analysis furthermore handle model errors, as in weak-constraint 4D-VAR (Trémolet, 2006; Desroziers et al., 2014; Laloyaux et al., 2020), and may include time-varying background error covariances by combining an ensemble of 4D-VAR (Bonavita et al., 2012), which is known as *ensemble of data assimilation* (EDA).

Assuming now that the state and observation models are non-linear, as in Eq. (3.7), denote the composition of the non-linear model forecast from time t_{k-1} to time t_l as

$$\mathcal{M}_{l:k} := \mathcal{M}_l \circ \cdots \circ \mathcal{M}_k, \quad \mathcal{M}_{k:k} := \mathbf{I}_{N_x}. \quad (3.68)$$

Then, the linear-Gaussian 4-D cost function is formally extended as a Gaussian approximation in 4D-VAR with

$$\mathcal{J}_{\mathrm{4D-VAR}}(\boldsymbol{x}_0) := \frac{1}{2}\|\overline{\boldsymbol{x}}_0 - \boldsymbol{x}_0\|_{\mathbf{B}_{\mathrm{4D-VAR}}}^2$$

$$+ \frac{1}{2}\sum_{k=1}^{L} \|\boldsymbol{y}_k - \mathcal{H}_k \circ \mathcal{M}_{k:1}(\boldsymbol{x}_0)\|_{\mathbf{R}_k}^2, \quad (3.69)$$

where $\mathbf{B}_{\mathrm{4D-VAR}}\leftarrow\mathbf{B}_{\mathrm{3D-VAR}}$. A similar indexing to equation to Eq. (3.65) gives the sequential form over newly incoming observations.

The incremental linearisation that was performed in 3D-VAR in Eq. (3.48) forms the basis for the classical technique

of incremental 4D-VAR. Suppose again that an explicit Cholesky factor for the background covariance is given $\mathbf{B}_{4D-VAR} := \Sigma\Sigma^\top$, where the state is written as a perturbation of the i-th iterate as

$$\boldsymbol{x}_0 := \overline{\boldsymbol{x}}_0^i + \Sigma\boldsymbol{w}. \tag{3.70}$$

Taking a Taylor expansion of the composition of the non-linear observation model and the non-linear state model at the i-th iterate,

$$\mathcal{H}_k \circ \mathcal{M}_{k:1}(\boldsymbol{x}_0) = \mathcal{H} \circ \mathcal{M}_{k:1}(\overline{\boldsymbol{x}}_0^i) + \mathbf{H}_k\mathbf{M}_{k:1}\Sigma\boldsymbol{w} + \mathcal{O}(\|\boldsymbol{w}\|^2), \tag{3.71}$$

the incremental cost function is rendered

$$\mathcal{J}_{4D-VARI}(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{1}{2}\sum_{k=1}^{L}\|\boldsymbol{y}_k - \mathcal{H}_k \circ \mathcal{M}_{k:1}(\overline{\boldsymbol{x}}_0^i) - \mathbf{H}_k\mathbf{M}_{k:1}\Sigma\boldsymbol{w}\|^2_{\mathbf{R}_k}. \tag{3.72}$$

The argmin of Eq. (3.72) is defined as $\overline{\boldsymbol{w}}^i$, where $\overline{\boldsymbol{x}}^{i+1}$ is defined as in Eq. (3.47).

It is important to remember that in Eq. (3.72), the term $\mathbf{M}_{k:1}$ involves the calculation of the tangent linear model with reference to the underlying non-linear solution $\mathcal{M}_{L:1}(\overline{\boldsymbol{x}}_0^i)$, simulated freely over the entire DAW. A key aspect to the efficiency of the incremental 4D-VAR approach is in the adjoint model computation of the gradient to this cost function, commonly known as *backpropagation* in statistical and machine learning (Rojas, 2013, see chapter 7). The *adjoint model* with respect to the proposal is defined as

$$\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\boldsymbol{\delta}} = \left(\nabla_{\boldsymbol{x}}\boldsymbol{f}(\overline{\boldsymbol{x}}^i)\right)^\top\widetilde{\boldsymbol{\delta}}. \tag{3.73}$$

The solution to the linear adjoint model then defines the adjoint resolvent matrix $\mathbf{M}_k^\top$ (i.e. the transpose of the tangent linear resolvent). Notice that for the composition of the tangent linear model forecasts with $k \leq l$, the adjoint is given as

$$\mathbf{M}_{l:k}^\top := (\mathbf{M}_l\mathbf{M}_{l-1}\cdots\mathbf{M}_{k+1}\mathbf{M}_k)^\top = \mathbf{M}_k^\top\mathbf{M}_{k+1}^\top\cdots\mathbf{M}_{l-1}^\top\mathbf{M}_l^\top. \tag{3.74}$$

This means that the adjoint model variables $\widetilde{\boldsymbol{\delta}}l$ are propagated by the linear resolvents of the adjoint model, but applied in reverse chronological order from the tangent linear model from last time t_l to the initial time t_{k-1},

$$\widetilde{\boldsymbol{\delta}}_{k-1} := \mathbf{M}_{l:k}^\top\widetilde{\boldsymbol{\delta}}_l \tag{3.75}$$

transmitting the sensitivity from a future time back to a perturbation of (Kalnay, 2003, see section 6.3).

Define the *innovation* vector of the i-th iterate and the k-th observation vector as

$$\overline{\boldsymbol{\delta}}_k^i := \boldsymbol{y}_k - \mathcal{H}_k \circ \mathcal{M}_{k:1}(\overline{\boldsymbol{x}}_0^i). \tag{3.76}$$

Then the gradient of Eq. (3.72) with respect to the weight vector is written

$$\nabla_{\boldsymbol{w}}\mathcal{J}_{4D-VARI} = \boldsymbol{w} - \sum_{k=1}^{L}\Sigma^\top\mathbf{M}_{k:1}^\top\mathbf{H}_k^\top\mathbf{R}_k^{-1}\left[\overline{\boldsymbol{\delta}}_k^i - \mathbf{H}_k\mathbf{M}_{k:1}\Sigma\boldsymbol{w}\right]. \tag{3.77}$$

Making the substitution

$$\boldsymbol{\Delta}_k := \mathbf{R}_k^{-1}\left[\overline{\boldsymbol{\delta}}_k^i - \mathbf{H}_k\mathbf{M}_{k:1}\Sigma\boldsymbol{w}\right], \tag{3.78}$$

notice that the gradient is written in Horner factorisation as

$$\nabla_{\boldsymbol{w}}\mathcal{J}_{4D-VARI} = \boldsymbol{w} - \Sigma^\top\mathbf{M}_1^\top\left[\mathbf{H}_1^\top\boldsymbol{\Delta}_1 + \mathbf{M}_2^\top[\mathbf{H}_2^\top\boldsymbol{\Delta}_2 + \cdots + [\mathbf{M}_L^\top\mathbf{H}_L^\top\boldsymbol{\Delta}_L]]\right]. \tag{3.79}$$

Defining the adjoint variable as $\widetilde{\boldsymbol{\delta}}_L := \mathbf{H}_L^\top\boldsymbol{\Delta}_L$, with recursion in t_k,

$$\widetilde{\boldsymbol{\delta}}_k := \mathbf{H}_k^\top\boldsymbol{\Delta}_k + \mathbf{M}_{k+1}^\top\widetilde{\boldsymbol{\delta}}_{k+1}, \tag{3.80}$$

the gradient of the cost function is written as

$$\nabla_{\boldsymbol{w}}\mathcal{J}_{4D-VARI} := \boldsymbol{w} - \Sigma\widetilde{\boldsymbol{\delta}}_0. \tag{3.81}$$

Using these definitions, the gradient of the incremental cost function is computed from the following steps: (i) a forward pass of the free evolution of $\boldsymbol{x}_0^i$ under the non-linear forecast model, computing the innovations as in Eq. (3.76); (ii) the propagation of the perturbation $\Sigma\boldsymbol{w}$ in the tangent linear model and the linearised observation operator with respect to the proposal, $\overline{\boldsymbol{x}}_k^i$, in order to compute the terms in Eq. (3.78); and (iii) the back propagation of the sensitivities in the adjoint model, Eq. (3.73), to obtain the adjoint variables recursively as in Eq. (3.80) and the gradient from Eq. (3.81).

The benefit of this approach is that it gives an extremely efficient calculation of the gradient, provided that the tangent linear and adjoint models are available, which, in turn, is key to very efficient numerical optimisations of cost functions. However, the trade-off is that the tangent linear and adjoint models of a full-scale geophysical state model, and of the observation model, require considerable development time and expertise. Increasingly, these models can be computed abstractly by differentiating a computer program alone, in what is known as *automatic differentiation* of code (Griewank, 1989; Griewank and Walther, 2003; Hascöet, 2014; Baydin et al., 2018). When tangent linear and adjoint models are not available, one alternative is to use ensemble sampling techniques in the fully non-linear model $\mathcal{M}$ alone. The ensemble-based analysis provides a complementary approach to the explicit tangent linear and adjoint model analysis – this approach can be developed independently, or hybridised with the use of the tangent

linear and adjoint models as in various flavours of hybrid *ensemble-variational* (EnVAR) techniques (Asch et al., 2016; Bannister, 2017). This chapter focuses on the independent development of EnVAR estimators in the following section.

3.4 Bayesian Ensemble-Variational Estimators

3.4.1 The Ensemble Transform Kalman Filter

Consider once again the perfect, linear-Gaussian model in Eq. (3.25). Rather than explicitly computing the evolution of the background mean and error covariance in the linear model as in Eq. (3.32) and the KF, one can alternatively estimate the state mean and the error covariances using a statistical sampling approach. Let $\left\{ x_{k,i}^{\mathrm{f/a}} \right\}_{i=1}^{N_e}$ be replicates of the model state, *independently and identically distributed* (iid) according to the distribution

$$x_{k,i}^{\mathrm{f/a}} \sim \mathcal{N}(\overline{x}_k^{\mathrm{f/a}}, \mathbf{B}_k^{\mathrm{f/a}}). \tag{3.82}$$

Given the iid assumption, the ensemble-based mean and the ensemble-based covariance

$$\hat{x}_k^{\mathrm{f/a}} := \frac{1}{N_e} \sum_{i=1}^{N_e} x_{k,i}^{\mathrm{f/a}}, \tag{3.83a}$$

$$\mathbf{P}_k^{\mathrm{f/a}} := \frac{1}{N_e - 1} \sum_{i=1}^{N_e} \left(x_{k,i}^{\mathrm{f/a}} - \hat{x}_k^{\mathrm{f/a}} \right)\left(x_{k,i}^{\mathrm{f/a}} - \hat{x}_k^{\mathrm{f/a}} \right)^{\mathrm{T}}, \tag{3.83b}$$

are (asymptotically) consistent estimators of the background, that is,

$$\mathbb{E}\left[\hat{x}_k^{\mathrm{f/a}} \right] = \overline{x}_k^{\mathrm{f/a}}, \qquad \lim_{N_e \to \infty} \mathbb{E}\left[\mathbf{P}_k^{\mathrm{f/a}} \right] = \mathbf{B}_k^{\mathrm{f/a}}, \tag{3.84}$$

where the expectation is over the possible realisations of the random sample. Particularly, the multivariate central limit theorem gives that

$$\left(\mathbf{P}_k^{\mathrm{f/a}} \right)^{-\frac{1}{2}} \left(x_k^{\mathrm{f/a}} - \hat{x}_k^{\mathrm{f/a}} \right) \to_d \mathcal{N}(\mathbf{0}, \mathbf{I}_{N_x}), \tag{3.85}$$

referring to convergence in distribution as $N_e \to \infty$ (Härdle et al., 2017, see section 6.2).

Using the relationships in Eq. (3.83), these estimators are efficiently encoded as linear operations on the ensemble matrix. Define the *ensemble matrix* and the *perturbation matrix* as

$$\mathbf{E}_k^{\mathrm{f/a}} := \begin{pmatrix} x_{k,1}^{\mathrm{f/a}} & \cdots & x_{k,N_e}^{\mathrm{f/a}} \end{pmatrix} \in \mathbb{R}^{N_x \times N_e}, \tag{3.86}$$

$$\mathbf{X}_k^{\mathrm{f/a}} := \begin{pmatrix} x_{k,1}^{\mathrm{f/a}} - \hat{x}_k^{\mathrm{f/a}} & \cdots & x_{k,N_e}^{\mathrm{f/a}} - \hat{x}_k^{\mathrm{f/a}} \end{pmatrix} \in \mathbb{R}^{N_x \times N_e}, \tag{3.87}$$

that is, as the arrays with the columns given as the ordered replicates of the model state and their deviations from the ensemble mean respectively. For $\mathbf{1}$, defined as the vector composed entirely of ones, define the following linear operations conformally in their dimensions

$$\hat{x}_k^{\mathrm{f/a}} = \mathbf{E}_k^{\mathrm{f/a}} \mathbf{1}/N_e, \tag{3.88a}$$

$$\mathbf{X}_k^{\mathrm{f/a}} = \mathbf{E}_k^{\mathrm{f/a}}(\mathbf{I}_{N_e} - \mathbf{1}\mathbf{1}^{\mathrm{T}}/N_e), \tag{3.88b}$$

$$\mathbf{P}_k^{\mathrm{f/a}} = \left(\mathbf{X}_k^{\mathrm{f/a}} \right)\left(\mathbf{X}_k^{\mathrm{f/a}} \right)^{\mathrm{T}}/(N_e - 1). \tag{3.88c}$$

The operator $\mathbf{I}_{N_e} - \mathbf{1}\mathbf{1}^{\mathrm{T}}/N_e$ is the orthogonal complementary projection operator to the span of the vector of ones, known as the *centring operator* in statistics (Härdle et al., 2017, see section 6.1); this has the effect of transforming the ensemble to mean zero.

Notice, when $N_e \leq N_x$, the ensemble-based error covariance has rank of at most $N_e - 1$, irrespective of the rank of the background error covariance, corresponding to the one degree of freedom lost in computing the ensemble mean. Therefore, to utilise the ensemble error covariance in the least-squares optimisation as with the KF in Eq. (3.39), a new construction is necessary. For a generic matrix $\mathbf{A} \in \mathbb{R}^{N \times M}$ with full column rank M, define the (left Moore–Penrose) *pseudo-inverse* (Meyer, 2000, see page 423)

$$\mathbf{A}^{\dagger} := (\mathbf{A}^{\top}\mathbf{A})^{-1}\mathbf{A}^{\top}. \tag{3.89}$$

In particular, $\mathbf{A}^{\dagger}\mathbf{A} = \mathbf{I}_M$ (and the orthogonal projector into the column span of $\mathbf{A}$) is defined by $\mathbf{A}\mathbf{A}^{\dagger}$. When $\mathbf{A}$ has full column rank as here, define the Mahalanobis 'distance' with respect to $\mathbf{G} = \mathbf{A}\mathbf{A}^{\top}$ as

$$\|v\|_{\mathbf{G}} := \sqrt{(\mathbf{A}^{\dagger}v)^{\top}(\mathbf{A}^{\dagger}v)}. \tag{3.90}$$

Note that in the case that $\mathbf{G}$ does not have full column rank (i.e. $N > M$), this is not a true norm on $\mathbb{R}^N$ as it is degenerate in the null space of $\mathbf{A}^{\dagger}$. This instead represents a lift of a non-degenerate norm in the column span of $\mathbf{A}$ to $\mathbf{R}^N$. In the case that v is in the column span of $\mathbf{A}$, one can equivalently write

$$v = \mathbf{A}w, \tag{3.91a}$$

$$\|v\|_{\mathbf{G}} = \|w\|, \tag{3.91b}$$

for a vector of weights $w \in \mathbf{R}^M$.

The *ensemble Kalman filter* (EnKF) cost function for the linear-Gaussian model is defined

$$\mathcal{J}_{\mathrm{EnKF}}(x_k) := \frac{1}{2}\|\hat{x}_k^{\mathrm{f}} - x_k\|_{\mathbf{P}_k^{\mathrm{f}}}^2 + \frac{1}{2}\|y_k - \mathbf{H}_k x_k\|_{\mathbf{R}_k}^2 \tag{3.92a}$$

$$\Leftrightarrow \mathcal{J}_{\mathrm{EnKF}}(w) := \frac{1}{2}(N_e - 1)\|w\|^2 + \frac{1}{2}\|y_k - \mathbf{H}_k\hat{x}_k^{\mathrm{f}} - \mathbf{H}_k\mathbf{X}_k^{\mathrm{f}}w\|_{\mathbf{R}_k}^2, \tag{3.92b}$$

where the model state is written as a perturbation of the ensemble mean

$$x_k = \hat{x}_k^{\mathrm{f}} + \mathbf{X}_k^{\mathrm{f}} w. \tag{3.93}$$

Notice, $w \in \mathbb{R}^{N_e}$, giving the linear combination of the ensemble perturbations used to represent the model state.

Define $\hat{w}$ to be the argmin of the cost function in Eq. (3.92b). Hunt et al. (2007) and Bocquet (2011) demonstrate that, up to a gauge transformation, $\hat{w}$ yields the argmin of the state-space cost function, Eq. (3.92a), when the estimate is restricted to the ensemble span. Equation (3.92b) is quadratic in w and can be solved to render

$$\hat{w} := \mathbf{0} - \mathbf{\Xi}_{\mathcal{J}_{\mathrm{EnKF}}}^{-1} \nabla_w \mathcal{J}_{\mathrm{EnKF}}|_{w=\mathbf{0}}, \tag{3.94a}$$

$$\mathrm{T} := \mathbf{\Xi}_{\mathcal{J}_{\mathrm{EnKF}}}^{-\frac{1}{2}}, \tag{3.94b}$$

$$\mathbf{P}_k^{\mathrm{a}} = (\mathbf{X}_k^{\mathrm{f}}\mathrm{T})(\mathbf{X}_k^{\mathrm{f}}\mathrm{T})^{\top}/(N_e - 1), \tag{3.94c}$$

corresponding to a single iteration of Newton's descent algorithm in Eq. (3.94a), initialised with the ensemble mean, to find the optimal weights.

The linear *ensemble transform Kalman filter* (ETKF) equations (Bishop et al., 2001; Hunt et al., 2007) are then given by

$$\mathbf{E}_k^{\mathrm{f}} = \mathbf{M}_k \mathbf{E}_{k-1}^{\mathrm{a}}, \tag{3.95a}$$

$$\mathbf{E}_k^{\mathrm{a}} = \hat{x}_k^{\mathrm{f}} \mathbf{1}^{\top} + \mathbf{X}_k^{\mathrm{f}}\left(\hat{w}\mathbf{1}^{\top} + \sqrt{N_e - 1}\mathrm{T}\mathbf{U}\right), \tag{3.95b}$$

where $\mathbf{U} \in \mathbb{R}^{N_e \times N_e}$ can be any mean-preserving, orthogonal transformation (i.e. $\mathbf{U}\mathbf{1} = \mathbf{1}$). The simple choice of $\mathbf{U} := \mathbf{I}_{N_e}$ is sufficient, but it has been demonstrated that choosing a random, mean-preserving orthogonal transformation at each analysis can improve the accuracy and robustness of the ETKF, smoothing out higher-order artefacts in the empirical covariance estimate (Sakov and Oke, 2008).

Notice that Eq. (3.95b) is written equivalently as a single right ensemble transformation:

$$\mathbf{E}_k^{\mathrm{a}} = \mathbf{E}_k^{\mathrm{f}}\mathbf{\Psi}_k, \tag{3.96a}$$

$$\mathbf{\Psi}_k := \mathbf{1}\mathbf{1}^{\top}/N_e + (\mathbf{I}_{N_e} - \mathbf{1}\mathbf{1}^{\top}/N_e)\left(\hat{w}\mathbf{1}^{\top} + \sqrt{N_e - 1}\mathrm{T}\mathbf{U}\right), \tag{3.96b}$$

where the columns are approximately distributed as in Eq. (3.82), with the (asymptotic) consistency as with the central limit theorem, Eq. (3.85). However, in the small feasible sample sizes for realistic geophysical models, this approximation can lead to a systematic underestimation of the uncertainty of the analysis state, where the ensemble-based error covariance can become overly confident in its own estimate with artificially small variances. Covariance inflation is a technique that is widely used to regularise the ensemble-based error covariance by increasing the empirical variances of the estimate. This can be used to handle the inaccuracy of the estimator due to the finite sample size

approximation of the background mean and error covariance as in Eq. (3.85), as well as inaccuracies due to a variety of other sources of error (Carrassi et al., 2018; Raanes et al., 2019a; Tandeo et al., 2020).

A Bayesian hierarchical approach can model the inaccuracy in the approximation error due to the finite sample size by including a prior additionally on the background mean and error covariance $p(\overline{x}_k^{\mathrm{f}}, \mathbf{B}_k^{\mathrm{f}})$, as in the finite-size ensemble Kalman filter formalism of Bocquet (2011), Bocquet and Sakov (2012), and Bocquet et al. (2015). Mathematical results demonstrate that covariance inflation can ameliorate the systematic underestimation of the variances due to model error in the presence of a low-rank ensemble (Grudzien et al., 2018). In the presence of significant model error, the finite-size analysis is extended by the variant developed by Raanes et al. (2019a).

The linear transform of the ensemble matrix in Eq. (3.96) is key to the efficiency of the (deterministic) EnKF presented in this section (Sect. 3.4.1). The ensemble-based cost function Hessian $\mathbf{\Xi}_{\mathcal{J}_{\mathrm{EnKF}}} \in \mathbb{R}^{N_e \times N_e}$, where $N_e \ll N_x$ for typical geophysical models. Particularly, the cost of computing the optimal weights, as in Eq. (3.94a), and computing the transform matrix $\mathbf{T}$ in Eq. (3.94b) are both subordinate to the cost of the eigenvalue decomposition of the Hessian at $\mathcal{O}(N_e^3)$ *floating point operations* (flops), or to a randomised singular value decomposition (Farchi and Bocquet, 2019). However, the extremely low ensemble size means that the correction to the forecast is restricted to the low-dimensional ensemble span, which may fail to correct directions of rapidly growing errors due to the rank deficiency. While chaotic, dissipative dynamics implies that the background covariance $\mathbf{B}_k^{\mathrm{f/a}}$ has spectrum concentrated on a reduced rank subspace (Carrassi et al., 2022, and referenced therein), covariance hybridisation (Penny, 2017) or localisation (Sakov and Bertino, 2011) are used in practice to regularise the estimator's extreme rank deficiency, and the spurious empirical correlations that occur as a result of the degenerate sample size.

When extended to non-linear dynamics, as in Eq. (3.7a), the EnKF can be seen to make an ensemble-based approximation to the EKF cost function, where the forecast ensemble is defined by

$$x_{k,i}^{\mathrm{f}} := \mathcal{M}_{k-1}(x_{k-1,i}^{\mathrm{a}}), \quad \mathbf{E}_k^{\mathrm{f}} := \left(x_{k,1}^{\mathrm{f}} \quad \cdots \quad x_{k,N_e}^{\mathrm{f}} \right). \tag{3.97}$$

The accuracy of the linear-Gaussian approximation to the dynamics of the non-linear evolution of the ensemble, like the approximation of the EKF, depends strongly on the length of the forecast horizon Δ_t. When the ensemble mean is a sufficiently accurate approximation of the mean state, and if the ensemble spread is of the same order as the error in the mean estimate (Whitaker and Loughe, 1998), a similar approximation can be made for the ensemble evolution at first order, as

with Eq. (3.53), but linearised about the ensemble mean as discussed later in Eq. (3.112). Despite the similarity to the EKF, in the moderately non-linear dynamics present in medium- to longer-range forecast horizons, the EnKF does not suffer from the same inaccuracy as the EKF in truncating the time-evolution at first order (Evensen, 2003). The forecast ensemble members themselves evolve fully non-linearly, tracking the higher-order dynamics, but the analysis update based on the Gaussian approximation becomes increasingly biased and fails to discriminate features like multimodality of the posterior, even though this is occasionally an asset with higher-order statistical artefacts (see the discussion in Lawson and Hansen, 2004).

3.4.2 The Maximum Likelihood Ensemble Filter

The EnKF filter analysis automatically accommodates weak non-linearity in the state model without using the tangent linear model, as the filtering cost function has no underlying dependence on the state model. However, the EnKF analysis must be adjusted to account for non-linearity in the observation model. When a non-linear observation operator is introduced, as in Eq. (3.7b), the EnKF cost function can be re-written in the incremental analysis as

$$\mathcal{J}_{\text{EnKFI}}(\boldsymbol{w}) := \frac{1}{2}(N_e - 1)\|\boldsymbol{w}\|^2 + \frac{1}{2}\|\boldsymbol{y}_k - \mathcal{H}_k\left(\hat{\boldsymbol{x}}_k^{i,\text{f}}\right)$$
$$-\mathbf{H}_k\mathbf{X}_k^{\text{f}}\boldsymbol{w}\|_{\mathbf{R}_k}^2, \tag{3.98}$$

where $\hat{\boldsymbol{x}}_k^{i,\text{f}}$ refers to the *i*-th iteration for the forecast mean. Note that this does not refer to an estimate derived by an iterated ensemble forecast through the non-linear state model – rather, this is an iteration only with respect to the estimate of the optimal weights for the forecast perturbations. If $\hat{\boldsymbol{w}}^i$ is defined as the argmin of the cost function in Eq. (3.98), then the iterations of the ensemble mean are given as

$$\hat{\boldsymbol{x}}_k^{i+1,\text{f}} := \hat{\boldsymbol{x}}_k^{i,\text{f}} + \mathbf{X}_k^{\text{f}}\hat{\boldsymbol{w}}^i. \tag{3.99}$$

When $\|\hat{\boldsymbol{w}}^i\|$ is sufficiently small, the optimisation terminates and the transform and the ensemble update can be performed as with the ETKF as in Eqs. (3.94b) and (3.96). However, the direct, incremental approach here used the computation of the Jacobian of the observation operator $\mathbf{H}_k$.

The *maximum likelihood ensemble filter* (MLEF) of Zupanski (2005) and Zupanski et al. (2008) is an estimator designed to perform incremental analysis in the ETKF formalism, but without taking an explicit Taylor expansion for the observation operator. The method is designed to approximate the directional derivative of the non-linear observation operator with respect to the ensemble perturbations,

$$\mathbf{H}_k\mathbf{X}_k^{\text{f}} := \nabla|_{\hat{\boldsymbol{x}}_k^{i,\text{f}}}[\mathcal{H}_k]\mathbf{X}_k^{\text{f}}, \tag{3.100}$$

equivalent to computing the ensemble sensitivities of the map linearised about the ensemble-based mean. This is often performed with an explicit finite differences approximation between the ensemble members and the ensemble mean, mapped to the observation space. Particularly, one may write

$$\mathcal{H}_k\left(\boldsymbol{x}_k^{i,\text{f}}\right) \approx \hat{\boldsymbol{y}}_k^i := \mathcal{H}_k\left(\hat{\boldsymbol{x}}_k^{i,\text{f}}\mathbf{1}^\top + \epsilon\mathbf{X}_k^{\text{f}}\right)\mathbf{1}/N_e, \tag{3.101a}$$

$$\mathbf{H}_k\mathbf{X}_k^{\text{f}} \approx \widetilde{\mathbf{Y}}_k := \frac{1}{\epsilon}\mathcal{H}_k\left(\hat{\boldsymbol{x}}_k^{i,\text{f}}\mathbf{1}^\top + \epsilon\mathbf{X}_k^{\text{f}}\right)(\mathbf{I}_{N_e} - \mathbf{1}\mathbf{1}^\top/N_e),$$
$$\tag{3.101b}$$

where ϵ is a small constant that re-scales the ensemble perturbations to approximate infinitesimals about the mean, and rescales the finite differences about the ensemble mean in the observation space. This technique is used, for example, in a modern form of the MLEF algorithm, based on the analysis of the iterative ensemble Kalman filter and smoother (Asch et al., 2016, see section 6.7.2.1). The approximation in Asch et al. (2016) easily generalises to a 4-D analysis, taking the directional derivative with respect to the state and observation models simultaneously, as in Eq. (3.111). However, in the smoothing problem, one can also extend the analysis in terms of an alternating forward-filtering pass and backward-filtering pass to estimate the joint posterior. This chapter returns to the 4-D analysis with the iterative ensemble Kalman filter and smoother in Section 3.4.4, after an interlude on the retrospective analysis of the EnKS in the following section.

3.4.3 The Ensemble Transform Kalman Smoother

The EnKS extends the filter analysis in the ETKF over the smoothing DAW by sequentially reanalysing past states with future observations with an additional filtering pass over the DAW backward-in-time. This analysis is performed retrospectively in the sense that the filter cycle of the ETKF is left unchanged, while an additional inner-loop of the DA cycle performs an update on the estimated lagged state ensembles within the DAW, stored in memory. This can be formulated both for a fixed DAW and for fixed-lag smoothing, where only minor modifications are needed. Consider here the algorithmically stationary DAW $\{t_1, \ldots, t_L\}$ of fixed-lag smoothing, with a shift S and lag L, and where it is assumed that $S = 1 \leq L$. The fixed-lag smoothing cycle of the EnKS begins by estimating the joint posterior pdf $p(\boldsymbol{x}_{L:1}|\boldsymbol{y}_{L:1})$ recursively, given the joint posterior estimate over the last DAW $p(\boldsymbol{x}_{L-1:0}|\boldsymbol{y}_{L-1:0})$.

Given $p(\boldsymbol{x}_{L-1:0}, \boldsymbol{y}_{L-1:0})$, one can write the filtering pdf up to proportionality:

$$p(\boldsymbol{x}_L|\boldsymbol{y}_{L:0}) \propto p(\boldsymbol{y}_L|\boldsymbol{x}_L, \boldsymbol{y}_{L-1:0})\, p(\boldsymbol{x}_L, \boldsymbol{y}_{L-1:0}) \tag{3.102a}$$

$$\propto \underbrace{p(\boldsymbol{y}_L|\boldsymbol{x}_L)}_{(i)} \underbrace{\int p(\boldsymbol{x}_L|\boldsymbol{x}_{L-1})p(\boldsymbol{x}_{L-1:0}|\boldsymbol{y}_{L-1:0})\mathrm{d}\boldsymbol{x}_{L-1:0}}_{(ii)}, \quad (3.102\text{b})$$

as the product of (i) the likelihood of the observation $\boldsymbol{y}_L$ given $\boldsymbol{x}_L$; and (ii) the forecast for $\boldsymbol{x}_L$ using the transition pdf on the last joint posterior estimate, marginalising out the past history of the model state $\boldsymbol{x}_{L-1:0}$. Recalling that $p(\boldsymbol{x}_L|\boldsymbol{y}_{L:1}) \propto p(\boldsymbol{x}_L|\boldsymbol{y}_{L:0})$, this provides a means to estimate the filter marginal of the joint posterior. An alternating filtering pass, backward-in-time, completes the smoothing cycle by estimating the joint posterior pdf $p(\boldsymbol{x}_{L:1}, \boldsymbol{y}_{L:1})$.

Consider that the marginal smoother pdf is proportional to

$$p(\boldsymbol{x}_{L-1}|\boldsymbol{y}_{L:0}) \propto p(\boldsymbol{y}_L|\boldsymbol{x}_{L-1}, \boldsymbol{y}_{L-1:0})p(\boldsymbol{x}_{L-1}, \boldsymbol{y}_{L-1:0}) \quad (3.103\text{a})$$

$$\propto \underbrace{p(\boldsymbol{y}_L|\boldsymbol{x}_{L-1})}_{(i)} \underbrace{p(\boldsymbol{x}_{L-1}|\boldsymbol{y}_{L-1:0})}_{(ii)}, \quad (3.103\text{b})$$

where (i) is the likelihood of the observation $\boldsymbol{y}_L$ given the past state $\boldsymbol{x}_{L-1}$, and (ii) is the marginal pdf for $\boldsymbol{x}_{L-1}$ from the last joint posterior. The corresponding linear-Gaussian Bayesian MAP cost function is given for the retrospective analysis of the KS as

$$\mathcal{J}_{\text{KS}}(\boldsymbol{x}_{L-1}) = \frac{1}{2}\|\boldsymbol{x}_{L-1} - \overline{\boldsymbol{x}}^{\text{s}}_{L-1|L-1}\|^2_{\mathbf{B}^{\text{s}}_{L-1|L-1}} + \frac{1}{2}\|\boldsymbol{y}_L$$
$$- \mathbf{H}_L\mathbf{M}_L\boldsymbol{x}_{L-1}\|^2_{\mathbf{R}_L}, \quad (3.104)$$

where $\overline{\boldsymbol{x}}^{\text{s}}_{L-1|L-1}$ and $\mathbf{B}^{\text{s}}_{L-1|L-1}$ are the mean and covariance of the marginal smoother pdf $p(\boldsymbol{x}_{L-1}|\boldsymbol{y}_{L-1:0})$. Define the matrix decomposition with the factorisation (e.g. a Cholesky decomposition):

$$\mathbf{B}^{\text{s}}_{L-1|L-1} = \boldsymbol{\Sigma}^{\text{s}}_{L-1|L-1}(\boldsymbol{\Sigma}^{\text{s}}_{L-1|L-1})^\top, \quad (3.105)$$

and write $\boldsymbol{x}_{L-1} = \overline{\boldsymbol{x}}^{\text{s}}_{L-1|L-1} + \boldsymbol{\Sigma}^{\text{s}}_{L-1|L-1}\boldsymbol{w}$, rendering the cost function as

$$\mathcal{J}_{\text{KS}}(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{1}{2}\|\boldsymbol{y}_L - \mathbf{H}_L\mathbf{M}_L(\overline{\boldsymbol{x}}^{\text{s}}_{L-1|L-1} + \boldsymbol{\Sigma}^{\text{s}}_{L-1|L-1}\boldsymbol{w})\|^2_{\mathbf{R}_L}$$
$$(3.106\text{a})$$

$$= \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{1}{2}\|\boldsymbol{y}_L - \mathbf{H}_L\overline{\boldsymbol{x}}^{\text{f}}_L - \mathbf{H}_L\boldsymbol{\Sigma}^{\text{f}}_L\boldsymbol{w}\|^2_{\mathbf{R}_L}. \quad (3.106\text{b})$$

Let $\overline{\boldsymbol{w}}$ now denote the argmin of Eq. (3.106). It is important to recognise that

$$\boldsymbol{x}_L := \mathbf{M}_L\left(\overline{\boldsymbol{x}}^{\text{s}}_{L-1|L-1} + \boldsymbol{\Sigma}^{\text{s}}_{L-1|L-1}\boldsymbol{w}\right) \quad (3.107\text{a})$$

$$= \overline{\boldsymbol{x}}^{\text{f}}_L + \boldsymbol{\Sigma}^{\text{f}}_L\boldsymbol{w}, \quad (3.107\text{b})$$

such that the argmin for the smoothing problem $\overline{\boldsymbol{w}}$ is also the argmin for the filtering MAP analysis.

The ensemble-based approximation,

$$\boldsymbol{x}_{L-1} = \hat{\boldsymbol{x}}^{\text{s}}_{L-1|L-1} + \mathbf{X}^{\text{s}}_{L-1|L-1}\boldsymbol{w}, \quad (3.108\text{a})$$

$$\mathcal{J}_{\text{EnKS}}(\boldsymbol{w}) = \frac{1}{2}(N_e - 1)\|\boldsymbol{w}\|^2 + \frac{1}{2}\|\boldsymbol{y}_L - \mathbf{H}_L\hat{\boldsymbol{x}}^{\text{f}}_L - \mathbf{H}_L\mathbf{X}^{\text{f}}_L\boldsymbol{w}\|^2, \quad (3.108\text{b})$$

to the exact smoother cost function in Eq. (3.106), yields the retrospective analysis of the EnKS as

$$\hat{\boldsymbol{w}} := \boldsymbol{0} - \boldsymbol{\Xi}^{-1}_{\mathcal{J}_{\text{EnKS}}}\nabla\mathcal{J}_{\text{EnKS}}|_{\boldsymbol{w}=\boldsymbol{0}}, \quad (3.109\text{a})$$

$$\mathbf{T} := \boldsymbol{\Xi}^{-\frac{1}{2}}_{\mathcal{J}_{\text{EnKS}}}, \quad (3.109\text{b})$$

$$\mathbf{E}^{\text{s}}_{L-1|L} = \hat{\boldsymbol{x}}^{\text{s}}_{L-1|L-1}\mathbf{1}^\top + \mathbf{X}^{\text{s}}_{L-1|L-1}\left(\hat{\boldsymbol{w}}\mathbf{1}^\top + \sqrt{N_e-1}\mathbf{T}\mathbf{U}\right),$$
$$= \mathbf{E}^{\text{s}}_{L-1|L-1}\boldsymbol{\Psi}_L, \quad (3.109\text{c})$$

where $\boldsymbol{\Psi}_L$ is the ensemble transform as defined for the filtering update as in Eq. (3.96).

These equations generalise for arbitrary indices $k|L$ over the DAW, providing the complete description of the inner-loop between each filter cycle of the EnKS. After each new observation is assimilated with the ETKF analysis step, a smoother inner-loop makes a backward pass over the DAW applying the transform and the weights of the ETKF filter update to each past ensemble state stored in memory. This analysis easily generalises to the case where there is a shift of the DAW with $S > 1$, though the EnKS alternating forward- and backward-filtering passes must be performed in sequence over the observations, ordered-in-time, rather than making a global analysis over $\boldsymbol{y}_{L:L-S+1}$. Finally, this easily extends to accommodate a non-linear observation model by using the MLEF filtering step to obtain the optimal ensemble transform, and by applying this recursively backward-in-time to the lagged ensemble states.

A schematic of the EnKS cycle for a lag of $L = 4$ and a shift of $S = 1$ is pictured in Fig. 3.3. Time moves forward from left to right in the horizontal axis with a step size of Δ_t. At each analysis time, the ensemble forecast from the last filter pdf is combined with the observation to produce the ensemble transform update. This transform is then utilised to produce the posterior estimate for all lagged ensemble states, conditioned on the new observation. The information in the posterior estimate thus flows in reverse time to the lagged states stored in memory, but the information flow is unidirectional in this scheme. This type of retrospective analysis maintains the computational cost of the fixed-lag EnKS described here at a comparable level to the EnKF, with the only significant additional cost being the storage of the ensemble at lagged times to be reanalysed.
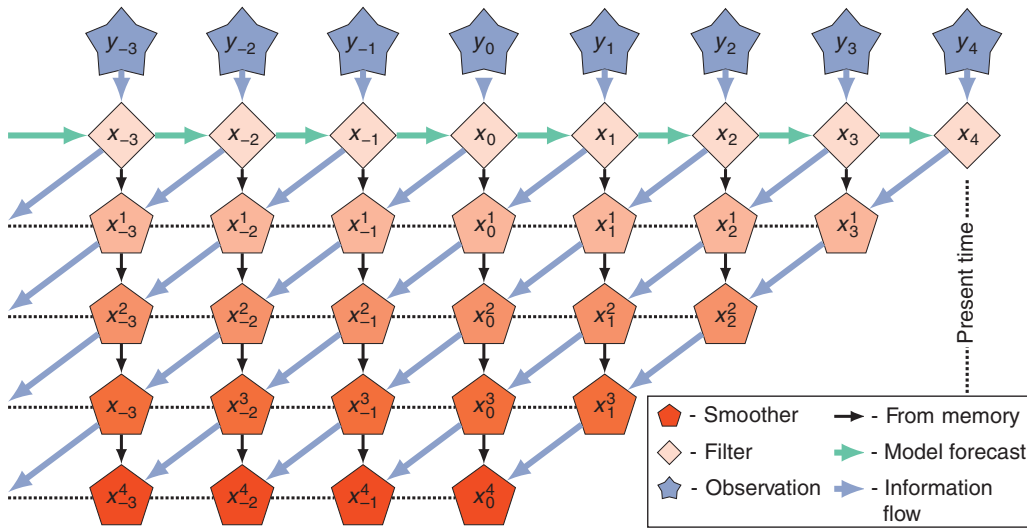
Figure 3.3 $L = 4$ (lag), $S = 1$ (shift) EnKS. Observations are assimilated sequentially via the filter cost function and a retrospective reanalysis is applied to all ensemble states within the lag window stored in memory. Source: Grudzien and Bocquet (2021), adapted from Asch et al. (2016).

3.4.4 The Iterative Ensemble Kalman Filter and Smoother

While the EnKS is computationally efficient, its retrospective analysis has some drawbacks compared to the 4-D analysis in terms of forecast accuracy. The 4-D fixed-lag smoothing analysis re-initialises each cycle with a reanalysed estimate for the initial data, transmitting the observations' information forward-in-time through the non-linear dynamics. The challenge with the 4-D analysis in the absence of the tangent linear and adjoint models is in devising how to efficiently and accurately compute the gradient of the 4-D cost function. Building on Zupanski (2005) and Liu et al. (2008), the analysis of the *iterative ensemble Kalman filter* (IEnKF) and the *iterative ensemble Kalman smoother* (IEnKS) extends the ensemble transform method of the ETKF to iteratively optimise the 4-D cost function.

Recall the quadratic cost function at the basis of incremental 4D-VAR, Eq. (3.72) – making an ensemble-based approximation for the background mean and error covariance, and writing the model state as an ensemble perturbation of the *i*-th proposal for the ensemble mean, one has

$$
\mathcal{J}_{\text{IEnKS}}(\boldsymbol{w}) := \frac{1}{2} \|\hat{\boldsymbol{x}}_0^{\text{i}} - \hat{\boldsymbol{x}}_0^{\text{i}} - \mathbf{X}_0 \boldsymbol{w}\|^2_{\mathbf{P}_0}
$$
$$
+ \frac{1}{2} \sum_{k=1}^{L} \|\boldsymbol{y}_k - \mathcal{H}_k \circ \mathcal{M}_{k:1}(\hat{\boldsymbol{x}}_0^i) - \mathbf{H}_k \mathbf{M}_{k:1} \mathbf{X}_0 \boldsymbol{w}\|^2_{\mathbf{R}_k}
$$
(3.110a)

$$
= \frac{1}{2}(N_e - 1)\|\boldsymbol{w}\|^2
$$
$$
+ \frac{1}{2} \sum_{k=1}^{L} \|\boldsymbol{y}_k - \mathcal{H}_k \circ \mathcal{M}_{k:1}(\hat{\boldsymbol{x}}_0^i) - \mathbf{H}_k \mathbf{M}_{k:1} \mathbf{X}_0 \boldsymbol{w}\|^2_{\mathbf{R}_k}.
$$
(3.110b)

Applying the finite differences approximation as given in Eq. (3.101a,b), but with respect to the composition of the non-linear observation operator with the non-linear state model

$$
\mathcal{H}_k \circ \mathcal{M}_{k:1}(\hat{\boldsymbol{x}}_0^i) \approx \hat{\boldsymbol{y}}_k := \mathcal{H}_k \circ \mathcal{M}_{k:1}(\hat{\boldsymbol{x}}_0^i + \mathbf{X}_0 \epsilon) \mathbf{1}/N_e,
$$
(3.111a)

$$
\mathbf{H}_k \mathbf{M}_{k:1} \mathbf{X}_0 \approx \widetilde{\mathbf{Y}}_k := \frac{1}{\epsilon} \mathcal{H}_k \circ \mathcal{M}_{k:1}(\hat{\boldsymbol{x}}_0^i + \mathbf{X}_0 \epsilon)(\mathbf{I}_{N_e} - \mathbf{11}^{\text{T}}/N_e),
$$
(3.111b)

this sketches the 'bundle' formulation of the IEnKS (Bocquet and Sakov, 2013; Bocquet and Sakov, 2014). This indexing refers to the case where the smoothing problem is performed offline, with a fixed DAW, though a similar indexing to Eq. (3.65) gives the sequential form over newly incoming observations.

The sequential form of the IEnKS can be treated as a non-linear sequential filter, which is the purpose that the method was originally devised for. Indeed, setting the number of lagged states $L = 1$, this provides a direct extension of the EnKF cost function, Eq. (3.92b), but where there is an additional dependence on the initial conditions for the ensemble forecast. This lag-1 iterative filtering scheme is called the IEnKF (Bocquet and Sakov, 2012; Sakov et al., 2012), which formed the original basis for the IEnKS. Modern forms of the IEnKF/S analysis furthermore include the treatment of model errors, like weak-constraint 4D-VAR (Sakov and Bocquet, 2018; Sakov et al., 2018; Fillion et al., 2020). Alternative formulations of this analysis, based on the original stochastic, perturbed observation EnKF (Evensen, 1994; Burgers et al., 1998), also have a parallel development in the *ensemble randomised maximum likelihood method* (EnRML) of Gu and Oliver (2007), Chen and Oliver (2012), and Raanes et al. (2019b). A similar ensemble-variational estimator based on the EnKF analysis is the *ensemble Kalman inversion* (EKI) of Iglesias et al. (2013), Schillings and Stuart (2018), and Kovachki and Stuart (2019).

3.4.5 The Single-Iteration Ensemble Kalman Smoother

The sequential smoothing analysis so far has presented two classical approaches: (i) a 4-D approach using the global analysis of all new observations available within a DAW at once, optimising an initial condition for the lagged model state; and (ii) the 3-D approach based upon alternating forward- and backward-filtering passes over the DAW. Each of these approaches has strengths and weaknesses in a computational cost/forecast-accuracy trade-off. Particularly, the 4-D approach, as in the IEnKS, benefits from the improved estimate of the initial condition when producing the subsequent forecast statistics in a shifted DAW; however, each step of the iterative optimisation comes at the cost of simulating the ensemble forecast over the entire DAW in the fully non-linear state model, which is typically the greatest numerical expense in geophysical DA. On the other hand, the 3-D approach of the classical EnKS benefits from a low computational cost, requiring only a single ensemble simulation of the non-linear forecast model over the DAW, while the retrospective analysis of lagged states is performed with the filtering transform without requiring additional model simulations (though at a potentially large memory storage cost). It should be noted that in the perfect, linear-Gaussian model Bayesian analysis, both approaches produce equivalent estimates of the joint posterior. However, when non-linearity is present in the DA cycle, the approaches produce distinct estimates: for this reason, the source of the non-linearity in the DA cycle is of an important practical concern.

Consider the situation in which the forecast horizon Δ_t is short, so that the ensemble time-evolution is weakly non-linear, and where the ensemble mean is a good approximation for the mean of the filtered distribution. In this case, the model forecast dynamics are well-approximated by the tangent linear evolution of a perturbation about the ensemble mean, that is,

$$\boldsymbol{x}_k = \mathcal{M}_k(\boldsymbol{x}_{k-1}) \approx \mathcal{M}_k(\hat{\boldsymbol{x}}_{k-1}) + \mathbf{M}_k \boldsymbol{\delta}_{k-1}. \tag{3.112}$$

For such short-range forecasts, non-linearity in the observation-analysis-forecast cycle may instead be dominated by the non-linearity in the observation operator $\mathcal{H}_k$, or in the optimisation of hyper-parameters of the filtering cost function, and not by the model forecast itself. In this situation, an iterative simulation of the forecast dynamics as in the 4-D approach may not produce a cost-effective reduction in the analysis error as compared to, for example, the MLEF filter analysis optimising the filtering cost function alone. However, one may still obtain the benefits of re-initialisation of the forecast with a reanalysed prior by using a simple hybridisation of the 3-D and 4-D smoothing analyses.

Specifically, in a given fixed-lag smoothing cycle, one may iteratively optimise the sequential filtering cost functions for a given DAW corresponding to the new observations at times $\{t_{L-S+1}, \ldots, t_L\}$ as with the MLEF. These filtering ensemble transforms not only condition the ensemble at the corresponding observation time instance, but also produce a retrospective analysis of a lagged state as in Eq. (3.109c) with the EnKS. However, when the DAW itself is shifted, one does not need to produce a forecast from the latest filtered ensemble – instead one can initialise the next ensemble forecast with the lagged, retrospectively smoothed ensemble at the beginning of the last DAW. This hybrid approach utilising the retrospective analysis, as in the classical EnKS, and the ensemble simulation over the lagged states while shifting the DAW, as in the 4-D analysis of the IEnKS, was recently developed by Grudzien and Bocquet (2022), and is called the *single-iteration ensemble Kalman smoother* (SIEnKS). The SIEnKS is named as such because it produces its forecast, filter, and reanalysed smoother statistics with a single iteration of the ensemble simulation over the DAW in a fully consistent Bayesian analysis. By doing so, it seeks to minimise the leading order cost of EnVAR smoothing (i.e. the ensemble simulation in the non-linear forecast model). However, the estimator is free to iteratively optimise the filter cost function for any single observation vector without additional iterations of the ensemble simulation. In observation-analysis-forecast cycles in which the forecast error dynamics are weakly non-linear, yet other aspects of the cycle are moderately to strongly non-linear, this scheme is shown to produce a forecast accuracy comparable to, and at times better than, the 4-D approach but with an overall lower leading-order computational burden (Grudzien and Bocquet, 2022).

A schematic of the SIEnKS cycle for a lag of $L = 4$ and a shift of $S = 2$ is pictured in Fig. 3.4. This demonstrates how the sequential analysis of the filter cost function and sequential, retrospective reanalysis for each incoming observation differs from the global analysis of the 4-D approach of the IEnKS. Other well-known DA schemes combining a retrospective reanalysis and re-initialisation of the ensemble forecast include the *running in place* (RIP) smoother of Kalnay and Yang (2010) and Yang et al. (2013) and the *one step ahead* (OSA) smoother of Desbouvries et al. (2011) and Ait-El-Fquih and Hoteit (2022). It can be shown that, with an ETKF-style filter analysis, a single iteration of the ensemble over the DAW, a perfect model assumption, and a lag of $L = S = 1$, the SIEnKS, RIP, and OSA smoothers all coincide (Grudzien and Bocquet, 2022).

3.5 Machine Learning and Data Assimilation in a Bayesian Perspective

This final section demonstrates how the Bayesian DA framework can be extended to estimate more than the state vector, including key parameters of the model and the observation-analysis-forecast cycle. In particular, the Bayesian framework can be used to formulate techniques
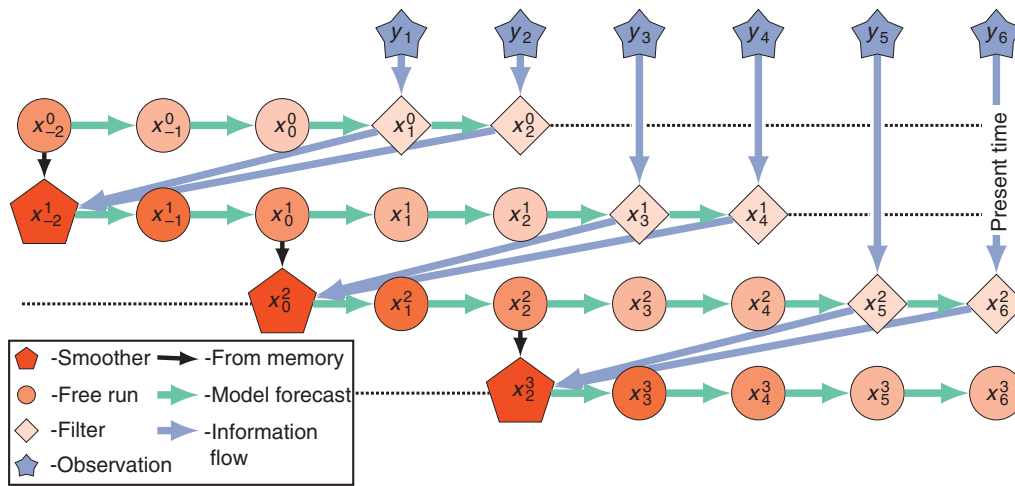
Figure 3.4 $L = 4$ (lag), $S = 2$ (shift) SIEnKS diagram. An initial condition from the last smoothing cycle initialises a forecast simulation over the current DAW of $L = 4$ states. New observations entering the DAW are assimilated sequentially via the filter cost function. After each filter analysis, a retrospective reanalysis is applied to the initial ensemble, and this reanalysed initial condition is evolved via the model S analysis times forward to begin the next cycle. Source: Grudzien and Bocquet (2021).

for learning both the state vector and part of, if not the full, dynamical model. If this objective was always in the scope of classical DA, it was actually made possible, significantly beyond linear regression, by the introduction of *machine learning* (ML) techniques to supplement traditional DA schemes.

Incorporating ML into DA algorithms was suggested quite early by Hsieh and Tang (1998), who clearly advocated the use of *neural networks* (NNs) in a variational DA framework. More recently this was put forward and illustrated by Abarbanel et al. (2018) and Bocquet et al. (2019) with a derivation of the generalised cost function from Bayes' law, showing that the classical ML cost function for learning a surrogate model of the dynamical model is a limiting case of the DA framework, as in Eq. (3.117). This formalism was further generalised by Bocquet et al. (2020), using classical DA notation, for learning the state vector, the dynamical model, and the error statistics attached to this full retrieval. This section follows this latest paper, showing how to derive the cost function of the generalised estimation problem. Note that going even beyond this approach, attempting to learn the optimisation scheme of the cost function, or even the full DA procedure, an approach called *end-to-end* in ML, is a subject of active investigations (Fablet et al., 2021; Peyron et al., 2021).

3.5.1 Prior Error Statistics
For the sake of simplicity, again assume Gaussian statistics for the observation errors $p(\boldsymbol{y}_k|\boldsymbol{x}_k) = n(\boldsymbol{y}_k|\boldsymbol{x}_k, \mathbf{R}_k)$, where the observation error covariance matrices $\mathbf{R}_{L:1} := \{\mathbf{R}_L, \mathbf{R}_{L-1}, \dots, \mathbf{R}_1\}$ are supposed to be known. The dynamical model is meant to be learned or approximated and thus stands as a surrogate model for the unknown true physical dynamics. Assuming that the model does not explicitly depend on time, its resolvent is defined by

$$\boldsymbol{x}_k := \mathbf{F}_{\mathbf{A}}^k(\boldsymbol{x}_{k-1}) + \boldsymbol{\eta}_k, \tag{3.113}$$

depending on a (possibly very large) set of parameters $\mathbf{A}$. Prototypically, $\mathbf{A}$ represents the weights and biases of an NN, which are learned from the observations alongside the state vectors within the DAW. The distribution for model error, such as $\boldsymbol{\eta}_k$ in Eq. (3.113), is also assumed Gaussian such that

$$p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}, \mathbf{A}, \mathbf{Q}_k) := n\Big(\boldsymbol{x}_k|\mathbf{F}_{\mathbf{A}}^k(\boldsymbol{x}_{k-1}), \mathbf{Q}_k\Big), \tag{3.114}$$

where $\mathbf{Q}_{L:1} := \{\mathbf{Q}_L, \mathbf{Q}_{L-1}, \dots, \mathbf{Q}_1\}$ are not necessarily known. Further assume that these Gaussian errors are white-in-time and that the observation and model errors are mutually independent.

Note the intriguing status of $\mathbf{Q}_{L:1}$ since it depends, *a posteriori*, on how well the surrogate model is estimated. This calls for an adaptive estimation of the model error statistics $\mathbf{Q}_{L:1}$, as the surrogate model, parametrised by $\mathbf{A}$, is better approximated.

3.5.2 Joint Estimation of the Model, Its Error Statistics, and the State Trajectory
Following the Bayesian formalism, one may form a MAP estimate for the joint pdf in $\mathbf{A}$ and $\boldsymbol{x}_{L:0}$ conditioned on the observations, together with the model error statistics. This generalised conditional pdf is expressed in the hierarchy

$$p(\mathbf{A}, \mathbf{Q}_{L:1}, \boldsymbol{x}_{L:0}|\boldsymbol{y}_{k:1}, \mathbf{R}_{K:0}) = $$
$$\frac{p(\boldsymbol{y}_{L:1}|\boldsymbol{x}_{L:0}, \mathbf{R}_{L:0})p(\boldsymbol{x}_{L:0}|\mathbf{A}, \mathbf{Q}_{L:1})p(\mathbf{A}, \mathbf{Q}_{L:1})}{p(\boldsymbol{y}_{L:1}, \mathbf{R}_{L:0})}, \tag{3.115}$$

where the mutual independence of the observation and model error is used. Once again, this remarkably stresses how powerful and general the Bayesian framework can be.

The first term in the numerator of the right-hand side is the usual likelihood of the observations. The second term in the numerator is the prior on the trajectory, given a known model and known model error statistics. The
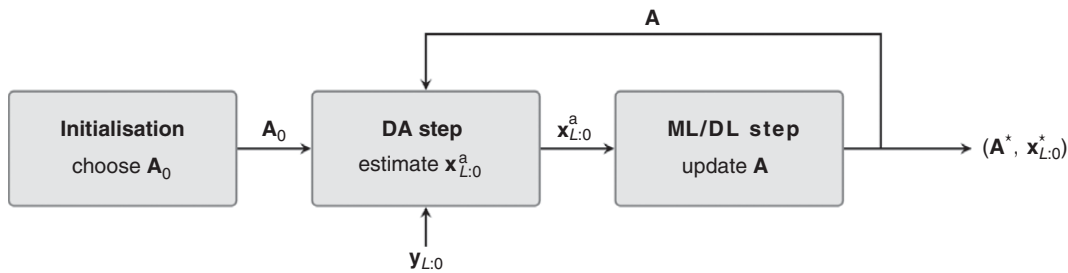
Figure 3.5 Estimation of both model and state trajectory using coordinate descent by alternately optimising on the state trajectory using DA and on the NN model parameters using ML/DL. The iterative loop stops when an accuracy criterion is met. Source: Farchi et al. (2021).

final term of the numerator is the joint prior of the model and the model error statistics, described as a *hyperprior*. The associated cost function is derived proportional to

$$
\begin{aligned}
\mathcal{J}_{\mathrm{DA-ML}}(\mathbf{A}, \boldsymbol{x}_{L:0}, \mathbf{Q}_{L:1}) &= -\log p(\mathbf{A}, \mathbf{Q}_{L:1}, \boldsymbol{x}_{L:0}|\boldsymbol{y}_{L:1}, \mathbf{R}_{L:1}) \\
&= \frac{1}{2}\sum_{k=1}^{L}\left\{ \|\boldsymbol{y}_k - \mathcal{H}_k(\boldsymbol{x}_k)\|_{\mathbf{R}_k}^2 \right. \\
&\qquad + \log(|\mathbf{R}_k|) \Big\} \\
&\qquad + \frac{1}{2}\sum_{k=1}^{L}\left\{ \|\boldsymbol{x}_k - \mathbf{F}_{\mathbf{A}}^k(\boldsymbol{x}_{k-1})\|_{\mathbf{Q}_k}^2 \right. \\
&\qquad + \log(|\mathbf{Q}_k|) \Big\} - \log p(\boldsymbol{x}_0, \mathbf{A}, \mathbf{Q}_{L:1}).
\end{aligned}
$$

(3.116)

Note the resemblance of (3.116) with the weak-constraint 4D-VAR cost function of classical DA (Trémolet, 2006). Very importantly, this Bayesian formulation allows for a rigorous treatment of partial and noisy observations. The classical ML cost function that uses noiseless, complete observations of the physical system is derived from Eq. (3.116), assuming that $\mathbf{Q}_k$ is known, $\mathbf{H}_k \equiv \mathbf{I}_{N_x}$, and setting $\mathbf{R}_k$ to go to $\mathbf{0}$. Associating the initial data as $\boldsymbol{y}_0 \leftarrow \boldsymbol{x}_0$, $\mathcal{J}_{\mathrm{DA-ML}}(\mathbf{A}, \boldsymbol{x}_{L:0}, \mathbf{Q}_{L:1})$ becomes, in its limit,

$$
\mathcal{J}_{\mathrm{ML}}(\mathbf{A}) = \frac{1}{2}\sum_{k=1}^{L}\|\boldsymbol{y}_k - \mathbf{F}_{\mathbf{A}}^k(\boldsymbol{y}_{k-1})\|_{\mathbf{Q}_k}^2 - \log p(\boldsymbol{y}_0, \mathbf{A}, \mathbf{Q}_{L:1}).
$$

(3.117)

Such connections between ML and DA were first highlighted by Hsieh and Tang (1998), Abarbanel et al. (2018), and Bocquet et al. (2019).

Solving the combined DA/ML problem, that is, minimising Eq. (3.116), leads to several key remarks. First, as mentioned earlier, this formalism allows one to learn a surrogate model of the true dynamics using partial, possibly sparse and noisy observations, as opposed to off-the-shelf ML techniques. This is obviously critical for geophysical systems. Second, in this framework, one fundamentally looks for a stochastically additive surrogate model of the form in Eq. (3.113) rather than a deterministic surrogate model, since $\boldsymbol{\eta}_k$ is drawn from the normal distribution of covariance matrices $\mathbf{Q}_k$. Third, there

are many ways to carry out this minimisation, as discussed by Bocquet et al. (2019) and Bocquet et al. (2020). Because the state vectors and the model parameters $\mathbf{A}$ are of fundamentally different nature, and yet are statistically interdependent, one idea is to minimise Eq. (3.116) through a coordinate descent (i.e. alternating minimisations on $\mathbf{A}$ and $\boldsymbol{x}_{L:0}$), as illustrated by Fig. 3.5. This was first suggested and successfully implemented by Brajard et al. (2020), by using an EnKF for the assimilation step (i.e. minimising on $\boldsymbol{x}_{L:0}$), and a *deep learning* (DL) optimiser for the ML step (i.e. minimising on $\mathbf{A}$). This work was extended to 4-D variational analysis using a 4D-VAR assimilation step with a DL optimiser for the ML step by Farchi et al. (2021).

3.5.3 Joint Estimation of the Model and the Error Statistics

A slightly different objective is to obtain a MAP estimate for the surrogate model, irrespective of any model state realisation, if, for example, one is interested in the MAP of the marginal conditional pdf

$$
p(\mathbf{A}, \mathbf{Q}_{L:1}|\boldsymbol{y}_{L:1}, \mathbf{R}_{L:1}) = \int p(\mathbf{A}, \mathbf{Q}_{L:1}, \boldsymbol{x}_{L:0}|\boldsymbol{y}_{L:1}, \mathbf{R}_{L:1})\,\mathrm{d}\boldsymbol{x}_{L:0},
$$

(3.118)

which is theoretically obtained by minimising

$$
\mathcal{J}(\mathbf{A}, \mathbf{Q}_{L:1}) = -\log p(\mathbf{A}, \mathbf{Q}_{L:1}|\boldsymbol{y}_{L:1}, \mathbf{R}_{L:1}). \qquad (3.119)
$$

As pointed out by Bocquet et al. (2019), the marginal pdf Eg. (3.118) can be approximately related to the joint pdf through a Laplace approximation of the integral. Here, however, one is interested in the full solution to this problem. This can be solved numerically by using the *expectation-maximisation* (EM) statistical algorithm (Dempster et al., 1977) jointly with the variational numerical solution of Eq. (3.116). This was suggested in Ghahramani and Roweis (1999) and Nguyen et al. (2019) and implemented and validated by Bocquet et al. (2020).

3.6 Conclusions

Unifying techniques from statistical estimation, non-linear optimisation, and even machine learning, the Bayesian approach to DA provides a consistent treatment of

a variety of topics in the classical DA problem, discussed throughout this chapter. Furthermore, the Bayesian approach can be reasonably extended to treat additional challenges not fully considered in this chapter, such as estimating process model parameters, handling significant modelling errors, and including the optimisation of various hyper-parameters of the observation-analysis-forecast cycle. Estimation may even include learning the dynamical process model itself, as in the emerging topic of DA/ML hybrid algorithms. In this context especially, a Bayesian analysis provides a coherent treatment of the problem where either DA or ML techniques themselves may be insufficient. This is particularly relevant where surrogate ML models are used to augment traditional physics-based dynamical models for, e.g. simulating unresolved dynamics at scales too fine to be dynamically represented. The hybrid approach in the Bayesian analysis provides a means to combine the dynamical and surrogate simulations with real-world observations, and to produce an analysis of the state and parameters for which subsequent simulations depend. This chapter thus presents one framework for interpreting the DA problem, both in its classical formulation and in the directions of the current state-of-the-art. In surveying a variety of widely used DA schemes, the key message of this chapter is how the Bayesian analysis provides a consistent framework for the estimation problem and how this allows one to formulate its solution in a variety of ways to exploit the operational challenges in the geosciences.

Acknowledgements

References

Abarbanel, H. D. I., Rozdeba, P. J., and Shirman, S. (2018). Machine learning: Deepest learning as statistical data assimilation problems. *Neural Computation*, 30(8), 2025–55. https://doi.org/10.1162/neco_a_01094.

Ait-El-Fquih, B., and Hoteit, I. (2022). Filtering with one-step-ahead smoothing for efficient data assimilation. In S. K. Park and L. Xu eds., *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications*, vol. IV. Cham: Springer, pp. 69–96. https://doi.org/10.1007/978–3–030–77722-7_1.

Anderson, B. D. O., and Moore, J. B. (1979). *Optimal Filtering*. Englewood Cliffs, NJ: Prentice-Hall.

Asch, M., Bocquet, M., and Nodet, M. (2016). *Data Assimilation: Methods, Algorithms, and Applications*. Philadelphia, PA: Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611974546.

Bannister, R. N. (2017). A review of operational methods of variational and ensemble-variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 143, 607–33. https://doi.org/10.1002/qj.2982.

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. (2018). Automatic differentiation in machine learning: A survey. *Journal of Machine Learning Research*, 18, 1–43. https://doi.org/0.5555/3122009.3242010.

Bishop, C. H., Etherton, B. J., and Majumdar, S. J. (2001). Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Monthly Weather Review*, 129, 420–36. https://doi.org/10.1175/1520–0493

Bocquet, M. (2011). Ensemble Kalman filtering without the intrinsic need for inflation. *Nonlinear Processes in Geophysics*, 18, 735–50. https://doi.org/10.5194/npg-18–735–2011.

Bocquet, M., Brajard, J., Carrassi, A., and Bertino, L. (2019). Data assimilation as a learning tool to infer ordinary differential equation representations of dynamical models. *Nonlinear Processes in Geophysics*, 26, 143–62. https://doi.org/10.5194/npg-26-143-2019.

Bocquet, M., Brajard, J., Carrassi, A., and Bertino, L. (2020). Bayesian inference of chaotic dynamics by merging data assimilation, machine learning and expectation-maximization. *Foundations of Data Science*, 2, 55–80. https://doi.org/10.3934/fods.2020004.

Bocquet, M., Raanes, P. N., and Hannart, A. (2015). Expanding the validity of the ensemble Kalman filter without the intrinsic need for inflation. *Nonlinear Processes in Geophysics*, 22, 645–62. https://doi.org/10.5194/npg22–645-2015.

Bocquet, M., and Sakov, P. (2012). Combining inflation-free and iterative ensemble Kalman filters for strongly nonlinear systems. *Nonlinear Processes in Geophysics*, 19, 383–99. https://doi.org/10.5194/npg-19-383-2012.

Bocquet, M., and Sakov, P. (2013). Joint state and parameter estimation with an iterative ensemble Kalman smoother. *Nonlinear Processes in Geophysics*, 20, 803–18. https://doi.org/10.5194/npg-20-803-2013.

Bocquet, M., and Sakov, P. (2014). An iterative ensemble Kalman smoother. *Quarterly Journal of the Royal Meteorological Society*, 140, 1521–35. https://doi.org/10.1002/qj.2236.

Bonavita, M., Isaksen, L., and Hólm, E. (2012). On the use of EDA background error variances in the ECMWF 4D-VAR.

Quarterly Journal of the Royal Meteorological Society, 138, 1540–59. https://doi.org/10.1002/qj.1899.

Brajard, J., Carrassi, A., Bocquet, M., and Bertino, L. (2020). Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the Lorenz 96 model. *Journal of Computational Science*, 44, 101171. https://doi.org/10.1016/j.jocs.2020.101171.

Burgers, G., van Leeuwen, P. J., and Evensen, G. (1998). Analysis scheme in the ensemble Kalman filter. *Monthly Weather Review*, 126, 1719–24. https://doi.org/10.1175/1520-0493(1998)126<1719:ASITEK>2.0.CO;2.

Carrassi, A., Bocquet, M., Bertino, L., and Evensen, G. (2018). Data assimilation in the geosciences: An overview on methods, issues, and perspectives. *WIREs Climate Change*, 9, e535. https://onlinelibrary.wiley.com/doi/abs/10.1002/wcc.535.

Carrassi, A., Bocquet, M., Demaeyer, J., Gruzien, C., Raanes, P. N. and Vannitsem, S. (2022). Data assimilation for chaotic dynamics. In S. K. Park and L. Xu eds., *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications*, vol. IV. Cham: Springer, pp. 1–42. https://doi.org/10.1007/978-3-030–77722–7_1.

Carrassi, A., Bocquet, M., Hannart, A., and Ghil, M. (2017). Estimating model evidence using data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 143, 866–80. https://doi.org/10.1002/qj.2972.

Chen, Y., and Oliver, D. S. (2012). Ensemble randomized maximum likelihood method as an iterative ensemble smoother. *Mathematical Geosciences*, 44, 1–26. https://doi.org/10.1007/s11004-011-9376-z.

Cohn, S. E., Sivakumaran, N. S., and Todling, R. (1994). A fixed-lag Kalman smoother for retrospective data assimilation. *Monthly Weather Review*, 122, 2838–67. https://doi.org/10.1175/1520–0493(1994)122<2838:AFLKSF>2.0.CO;2.

Corazza, M., Kalnay, E., Patil, D. J. et al. (2003). Use of the breeding technique to estimate the structure of the analysis errors of the day. *Nonlinear Processes in Geophysics*, 10, 233–43. https://doi.org/10.5194/npg-10-233-2003.

Cosme, E., Verron, J., Brasseur, P., Blum, J., and Auroux, D. (2012). Smoothing problems in a Bayesian framework and their linear Gaussian solutions. *Monthly Weather Review*, 140, 683–95. https://doi.org/10.1175/MWR-D-10-05025.1.

Courtier, P., Thépaut, J.-N., and Hollingsworth, A. (1994). A strategy for operational implementation of 4D-VAR, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, 120, 1367–87. https://doi.org/10.1002/qj.49712051912.

Daley, R. (1991). *Atmospheric Data Analysis*. New York: Cambridge University Press.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 1–38. https://doi.org/10.1111/j.2517–6161.1977.tb01600.x.

Desbouvries, F., Petetin, Y., and Ait-El-Fquih, B. (2011). Direct, prediction- and smoothing-based Kalman and particle filter algorithms. *Signal Processing*, 91, 2064–77. https://doi.org/10.1016/j.sigpro.2011.03.013.

Desroziers, G., Camino, J.-T., and Berre, L. (2014). 4DEnVar: link with 4D state formulation of variational assimilation and different possible implementations. *Quarterly Journal of the Royal Meteorological Society*, 140, 2097–110. https://doi.org/10.1002/qj.2325.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, 99, 10143–62. https://doi.org/10.1029/94JC00572.

Evensen, G. (2003). The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, 53, 343–67. https://doi.org/10.1007/s10236-003–0036–9.

Evensen, G., and van Leeuwen, P. J. (2000). An ensemble Kalman smoother for nonlinear dynamics. *Monthly Weather Review*, 128, 1852–67. https://doi.org/10.1175/1520–0493(2000)128<1852:AEKSFN>2.0.CO;2.

Fablet, R., Chapron, B., Drumetz, L. et al. (2021). Learning variational data assimilation models and solvers. *Journal of Advances in Modeling Earth Systems*, 13, e2021MS002572. https://doi.org/10.1029/2021MS002572.

Farchi, A., and Bocquet, M. (2019). On the efficiency of covariance localisation of the ensemble Kalman filter using augmented ensembles. *Frontiers in Applied Mathematics and Statistics*, 5, 3. https://doi.org/10.3389/fams.2019.00003.

Farchi, A., Laloyaux, P., Bonavita, M., and Bocquet, M. (2021). Using machine learning to correct model error in data assimilation and forecast applications. *Quarterly Journal of the Royal Meteorological Society*, 147(739), 3067–84. https://doi.org/10.1002/qj.4116.

Fillion, A., Bocquet, M., Gratton, S., Gürol, S. and Sakov, P. (2020). An iterative ensemble Kalman smoother in presence of additive model error. *SIAM/ASA Journal of Uncertainty Quantification*, 8, 198–228. https://doi.org/10.1137/19M1244147.

Ghahramani, Z., and Roweis, S. T. (1999). Learning nonlinear dynamical systems using an EM algorithm. In M. Kearns, S. Solla, and D. Cohn, eds., *Advances in Neural Information Processing Systems*, vol. 11. Cambridge, MA: MIT Press, pp. 431–7. https://doi.org/0.5555/340534.340696.

Grewal, M. S., and Andrews, A. P. (2014). *Kalman Filtering: Theory and Practice with MATLAB*. Hoboken, NJ; John Wiley & Sons. https://doi.org/10.1002/9781118984987.

Griewank, A. (1989). On automatic differentiation. *Mathematical Programming: Recent Developments and Applications*, 6, 83–107.

Griewank, A., and Walther, A. (2003). Introduction to automatic differentiation. *Proceedings in Applied Mathematics and Mechanics*, 2(1), 45–9. https://doi.org/10.1002/pamm.200310012.

Grudzien, C., and Bocquet, M. (2022). A fast, single-iteration ensemble Kalman smoother for sequential data assimilation. *Geoscientific Model Development*, 15, 7641–7681. https://doi.org/10.5194/gmd–15-7641-2022.

Grudzien, C., Carrassi, A., and Bocquet, M. (2018). Chaotic dynamics and the role of covariance inflation for reduced rank Kalman filters with model error. *Nonlinear Processes in Geophysics*, 25, 633–48. https://doi.org/10.5194/npg-25-633-2018.

Gu, Y., and Oliver, D. S. (2007). An iterative ensemble Kalman filter for multiphase fluid flow data assimilation. *SPE Journal*, 12, 438–46. https://doi.org/10.2118/108438-PA.

Hamill, T. M., and Snyder, C. (2000). A hybrid ensemble Kalman filter-3D variational analysis scheme. *Monthly Weather Review*, 128, 2905–19. https://doi.org/10.1175/1520-0493 (2000)128<2905:AHEKFV>2.0.CO;2.

Härdle, W. K., Okhrin, O., and Okhrin, Y. (2017). *Basic Elements of Computational Statistics*. Cham: Springer. https://link .springer.com/book/10.1007/978-3-319-55336-8.

Hascoët, L. (2014). An introduction to inverse modelling and parameter estimation for atmosphere and ocean sciences. In É. Blayo, M. Bocquet, E. Cosme, and L. F. Cugliandolo, eds., *Advanced Data Assimilation for Geosciences*. Oxford: Oxford University Press, pp. 349–69.

Hsieh, W. W., and Tang, B. (1998). Applying neural network models to prediction and data analysis in meteorology and oceanography. *Bulletin of the American Meteorological Society*, 79, 1855–70. https://doi.org/10.1175/1520-0477 (1998)079<1855: ANNMTP>2.0.CO;2.

Hunt, B. R., Kostelich, E. J., and Szunyogh, I. (2007). Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D*, 230, 112–26. https://doi .org/10.1016/j.physd.2006.11.008.

Iglesias, M. A., Law, K. J. H., and Stuart, A. M. (2013). Ensemble Kalman methods for inverse problems. *Inverse Problems*, 29, 045001. https://doi.org/10.1088/0266-5611/ 29/4/045001.

Janjić, T., Bormann, N., Bocquet, M. et al. (2018). On the representation error in data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 144, 1257–78. https://doi.org/ 10.1002/qj.3130.

Kalnay, E. (2003). *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511802270.

Kalnay, E., and Yang, S.-C. (2010). Accelerating the spin-up of ensemble Kalman filtering. *Quarterly Journal of the Royal Meteorological Society*, 136, 1644–51. https://doi.org/10 .1002/qj.652.

Kolmogorov, A. N. (2018). *Foundations of the Theory of Probability*, 2nd English ed. Mineola, NY: Dover Publications.

Kovachki, N. B., and Stuart, A. M. (2019). Ensemble Kalman inversion: a derivative-free technique for machine learning tasks. *Inverse Problems*, 35, 095005. https://doi.org/10.1088/ 1361-6420/ab1c3a.

Laloyaux, P., Bonavita, M., Dahoui, M., et al. (2020). Towards an unbiased stratospheric analysis. *Quarterly Journal of the Royal Meteorological Society*, 146, 2392–409. https://doi.org/ 10.1002/qj.3798.

Lawson, W. G., and Hansen, J. A. (2004). Implications of stochastic and deterministic filters as ensemble-based data assimilation methods in varying regimes of error growth. *Monthly Weather Review*, 132, 1966–81. https://doi.org/10.1175/1520-0493(2004)132<1966:IOSADF>2.0.CO;2.

Le Dimet, F.-X., and Talagrand, O. (1986). Variational algorithms for analysis and assimilation of meteorological observations: Theoretical aspects. *Tellus A*, 38, 97–110. https://doi.org/10 .3402/tellusa.v38i2.11706.√

Liu, C., Xiao, Q. and Wang, B. (2008). An ensemble-based four-dimensional variational data assimilation scheme. Part I: Technical formulation and preliminary test. *Monthly Weather Review*, 136, 3363–73. https://doi.org/10.1175/ 2008MWR2312.1.√

Lorenc, A. C. (1986). Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 112, 1177–94. https://doi.org/10.1002/qj.49711247414.√

Lorenc, A. C. (2003). The potential of the ensemble Kalman filter for NWP: A comparison with 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, 129, 3183–203. https://doi .org/10.1256/qj.02.132.√

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Science of India*, 2, 49–55.

Ménétrier, B., and Auligné, T. (2015). An overlooked issue of variational data assimilation. *Monthly Weather Review*, 143, 3925–30. https://doi.org/10.1175/MWRD-14–00404.1.

Meyer, C. D. (2000). *Matrix Analysis and Applied Linear Algebra*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Miller, R. N., Ghil, M., and Gauthiez, F. (1994). Advanced data assimilation in strongly nonlinear dynamical systems. *Journal of Atmospheric Sciences*, 51, 1037–56. https://doi.org/10.1175/1520-0469(1994)051<1037:ADAISN>2.0.CO;2.√

Nguyen, V. D., Ouala, S., Drumetz, L., and Fablet, R. (2019). EM-like learning chaotic dynamics from noisy and partial observations. *arXiv:1903.10335*. https://arxiv.org/abs/1903.10335.

Nocedal, J., and Wright, S. (2006). *Numerical Optimization*. New York: Springer. https://doi.org/10.1007/978-0-387-40065-5.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford: Oxford University Press.

Penny, S. G. (2017). Mathematical foundations of hybrid data assimilation from a synchronization perspective. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27, 126801. https://doi.org/10.1063/1.5001819.√

Peyron, M., Fillion, A., Gürol, S., et al. (2021). Latent space data assimilation by using deep learning. *Quarterly Journal of the Royal Meteorological Society*, 147, 3759–77. https://doi.org/ 10.1002/qj.4153.√

Raanes, P. N. (2016). On the ensemble Rauch–Tung–Striebel smoother and its equivalence to the ensemble Kalman smoother. *Quarterly Journal of the Royal Meteorological Society*, 142, 1259–64. https://doi.org/10.1002/qj.2728.√

Raanes, P. N., Bocquet, M., and Carrassi, A. (2019a). Adaptive covariance inflation in the ensemble Kalman filter by Gaussian scale mixtures. *Quarterly Journal of the Royal Meteorological Society*, 145, 53–75. https://doi.org/10.1002/qj .3386.

Raanes, P. N., Stordal, A. S., and Evensen, G. (2019b). Revising the stochastic iterative ensemble smoother. *Nonlinear Processes in Geophysics*, 26, 325–38. https://doi.org/10.5194/ npg-26-325-2019.

Reich, S., and Cotter, C. (2015). *Probabilistic Forecasting and Bayesian Data Assimilation*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781107706804.

Rojas, R. (2013). *Neural Networks: A Systematic Introduction*. New York: Springer. https://doi.org/10.1007/978-3-642-61068-4.

Ross, S. M. (2014). *Introduction to Probability Models*, *11th edition*. San Diego, CA: Academic Press. https://doi.org/10.1016/C2012-0-03564-8.

Sakov, P., and Bertino, L. (2011). Relation between two common localisation methods for the EnKF. *Computational Geosciences*, 15, 225–37. https://doi.org/10.1007/s10596-010–9202–6.

Sakov, P., and Bocquet, M. (2018). Asynchronous data assimilation with the EnKF in presence of additive model error. *Tellus A*, 70, 1414545. https://doi.org/10.1080/16000870.2017.1414545.

Sakov, P., Haussaire, J.-M., and Bocquet, M. (2018). An iterative ensemble Kalman filter in presence of additive model error. *Quarterly Journal of the Royal Meteorological Society*, 144, 1297–309. https://doi.org/10.1002/qj.3213.

Sakov, P., and Oke, P. R. (2008). Implications of the form of the ensemble transformation in the ensemble square root filters. *Monthly Weather Review*, 136, 1042–53. https://doi.org/10.1175/2007MWR2021.1.

Sakov, P., Oliver, D. S., and Bertino, L. (2012). An iterative EnKF for strongly nonlinear systems. *Monthly Weather Review*, 140, 1988–2004. https://doi.org/10.1175/MWR-D-11–00176.1.

Schillings, C., and Stuart, A. M. (2018). Convergence analysis of ensemble Kalman inversion: The linear, noisy case. *Applicable Analysis*, 97, 107–23. https://doi.org/10.1080/00036811.2017.1386784.

Tabeart, J. M., Dance, S. L., Haben, S. A., et al. (2018). The conditioning of least-squares problems in variational data assimilation. *Numerical Linear Algebra with Applications*, 25, e2165. https://doi.org/10.1002/nla.2165.

Talagrand, O., and Courtier, P. (1987). Variational assimilation of meteorological observation with the adjoint vorticity equation. I: Theory. *Quarterly Journal of the Royal Meteorological Society*, 113, 1311–28. https://doi.org/10.1002/qj.49711347812.

Tandeo, P., Ailliot, P., Bocquet, M., et al. (2020). A review of innovation-based approaches to jointly estimate model and observation error covariance matrices in ensemble data assimilation. *Monthly Weather Review*, 148, 3973–94. https://doi.org/10.1175/MWRD-19–0240.1.

Taylor, M. E. (1996). *Partial Differential Equations. Vol. 1: Basic Theory*. New York: Springer. https://doi.org/10.1007/978-1-4419-7055-8.

Tippett, M. K., Anderson, J. L., Bishop, C. H., Hamill, T. M., and Whitaker, J. S. (2003). Ensemble square root filters. *Monthly Weather Review*, 131, 1485–90. https://doi.org/10.1175/1520-0493(2003)131<1485:ESRF>2.0.CO;2.

Tong, Y. L. (2012). *The Multivariate Normal Distribution*. New York: Springer. https://doi.org/10.1007/978-1-4613-9655-0.

Trémolet, Y. (2006). Accounting for an imperfect model in 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, 132, 2483–504. https://doi.org/10.1256/qj.05.224.

Trevisan, A., and Uboldi, F. (2004). Assimilation of standard and targeted observations within the unstable subspace of the observation-analysis-forecast cycle. *Journal of the Atmospheric Sciences*, 61, 103–13. https://doi.org/10.1175/1520-0469(2004)061<0103:AOSATO>2.0.CO;2.

Whitaker, J. S., and Loughe, A. F. (1998). The relationship between ensemble spread and ensemble mean skill. *Monthly Weather Review*, 126, 3292–302. https://doi.org/10.1175/1520-0493(1998)126<3292:TRBESA>2.0.CO;2.

Yang, S.-C., Lin, K.-J., Miyoshi, T., and Kalnay, E. (2013). Improving the spinup of regional EnKF for typhoon assimilation and forecasting with Typhoon Sinlaku (2008). *Tellus A*, 65, 20804. https://doi.org/10.3402/tellusa.v65i0.20804.

Zupanski, M. (2005). Maximum likelihood ensemble filter: Theoretical aspects. *Monthly Weather Review*, 133, 1710–26. https://doi.org/10.1175/MWR2946.1.

Zupanski, M., Navon, I. M., and Zupanski, D. (2008). The maximum likelihood ensemble filter as a non-differentiable minimization algorithm. *Quarterly Journal of the Royal Meteorological Society*, 134, 1039–50. https://doi.org/10.1002/qj.251.