Improvements to Ensemble Methods for Data Assimilation in the Geosciences

Patrick Nima Raanes

St Hugh's College

University of Oxford

A thesis submitted for

Doctor of Philosophy

Michaelmas Term 2015

Acknowledgements

I am highly grateful to three different institutions, and their people, for their contributions to this thesis.

At the University of Oxford, Chris Farmer and Irene Moroz have provided me with invaluable support in their capacity as my academic supervisors. The Mathematical Institute houses some wonderful mathematicians, and the discussions I have had there with colleagues and friends have been very fruitful. My college, St Hugh's, has provided a beautiful backdrop where I have found many friends.

Half of my research has been conducted at the Nansen Environmental and Remote Sensing Center, where Laurent Bertino has been excellent as my main supervisor. From the same institution, I must also thank Alberto Carrassi, with whom I worked closely for the chapter on model error incorporation, and Francois Counillon and Mohamad Gharamti for their friendly mentoring and overall helpfulness. I also wish to thank Marc Bocquet, with whom I have collaborated closely.

I am most grateful for the grant initiated by Geir Evensen and provided by Statoil Petroleum AS, which houses another set of influential, academic mentors, including Remus Hanea, Matthew Owoyemi, and Jan Arild Skjervheim.

Finally, I must thank my parents, on whom I have always relied.

Improvements to Ensemble Methods for Data Assimilation in the Geosciences

Patrick Nima Raanes

St Hugh's College
University of Oxford

A thesis submitted for
Doctor of Philosophy

Michaelmas Term 2015

Data assimilation considers the problem of using a variety of data to calibrate model-based estimates of dynamic variables and static parameters. Geoscientific examples include (i) satellite observations and atmospheric models for weather forecasting, and (ii) well-log data and reservoir flow simulators for oil production optimization. Approximate solutions are provided by the set of techniques deriving from the ensemble Kalman filter (EnKF), which combines a Monte Carlo approach with assumptions of linearity and Gaussianity. This thesis proposes some improvements to the accuracy and understanding of such ensemble methods.

Firstly, a new scheme is developed to account for model noise in the forecast step of the EnKF. The main aim is to eliminate the sampling errors of additive, simulated noise. The scheme is based on the previously developed "square root" schemes for the analysis step, but requires further consideration due to the limited subspace spanned by the ensemble. The properties of the square root scheme in general are surveyed.

Secondly, the "finite size" ensemble Kalman filter (EnKF-N) is reviewed. The EnKF-N explicitly considers the uncertainty in the forecast moments (mean and covariance), thereby not requiring the multiplicative inflation commonly used to compensate for an intrinsic bias of the analysis step of the standard EnKF. Thus, in the perfect model setting, it avoids the process of tuning the inflation factor. This presentation consolidates the earlier literature on the EnKF-N, substantiates the scalar inflation perspective, and rectifies a deficiency.

Thirdly, two ensemble "smoothers" expressed by different recursions, used in different applications, and hitherto thought to yield different results, are shown to be equivalent. The theory is revisited under practical considerations, where equivalence is broken due to inflation and localization, but the methods remain equally capable.

In each case, the theory is tested and the accuracy performance is benchmarked against standard methods using numerical twin experiments.

Contents

List of Figures

Notation

Abbreviations

Abbrev.	Meaning	Reference
DA	data assimilation	section 1.1
EnKF	ensemble Kalman filter	chapter 2
EnKS	ensemble Kalman smoother	chapter 7
EnRTS	ensemble RTS (smoother)	chapter 7
HMM	hidden Markov model	section 1.2.1
iid	independent and identically distributed	
KF	Kalman filter	section 2.1
pdf	probability density function	
RMSE	root mean square error	section 3.1
RRSQRT	reduced-rank square root filter	page 25
RTS	Rauch-Tung-Striebel	chapter 7
SEIK	singular evolutive interpolated Kalman filter	page 50
SPD	symmetric, positive-definite	page 37
SVD	singular value decomposition	section E.1
Th.	theorem	

Symbols in frequent use

Symbol	Definition	Reference
N	Ensemble size.	chapter 2
m	State vector size.	section 1.2
p	Observation vector size.	section 1.2
$k_1{:}k_2$	The sequence of integers from k_1 to k_2.	
$x_{k_1:k_2}$	The sequence $\{x_k \ ; \ k = k_1, \ldots, k_2\}$.	
$p()$	Probability density function of the random variable indicated by its argument.	
$\mathbf{I}_k$	Identity matrix of size $k \times k$.	
$\mathbb{1}$	Vector of ones. Length: N.	
$\boldsymbol{e}_i$	Unit coordinate vector in the i-th dimension.	
$\mathbf{M}^\mathsf{T}$	Transpose of the arbitrary matrix $\mathbf{M}$.	
$\mathbf{M}^+$	Pseudoinverse of $\mathbf{M}$.	section E.3
$\mathbf{M}^{1/2}, \mathbf{M}^{\mathsf{T}/2}$	A matrix square root of $\mathbf{M}$, and its transpose. $\mathbf{M}^{1/2}\mathbf{M}^{\mathsf{T}/2} = \mathbf{M}$.	section 4.2
$\mathbf{M}^{1/2}_{\text{sym}}$	The symmetric square root of the SPD matrix $\mathbf{M}$.	section 4.2
$[\mathbf{M}]_{i,j}$	Element (i,j) of $\mathbf{M}$.	
$\mathbb{E}(\boldsymbol{x})$	$= \int \boldsymbol{x}\, p(\boldsymbol{x})\, d\boldsymbol{x}$. Expectation of the random variable $\boldsymbol{x}$.	
$\mathbb{V}\text{ar}(\boldsymbol{x})$	Variance of the random variable $\boldsymbol{x}$. If $\boldsymbol{x}$ is a vector, then the output is a covariance *matrix*. $\mathbb{V}\text{ar}(\boldsymbol{x}) = \int (\boldsymbol{x} - \mathbb{E}(\boldsymbol{x}))(\boldsymbol{x} - \mathbb{E}(\boldsymbol{x}))^\mathsf{T}\, p(\boldsymbol{x})\, d\boldsymbol{x}$.	
$\mathbb{C}\text{ov}(\boldsymbol{x}, \boldsymbol{y})$	Cross-covariance of $\boldsymbol{x}$ and $\boldsymbol{y}$. $\mathbb{C}\text{ov}(\boldsymbol{x}, \boldsymbol{y}) = \int (\boldsymbol{x} - \mathbb{E}(\boldsymbol{x}))(\boldsymbol{y} - \mathbb{E}(\boldsymbol{y}))^\mathsf{T}\, p(\boldsymbol{x}, \boldsymbol{y})\, d\boldsymbol{x}\, d\boldsymbol{y}$.	
tr	Trace of a matrix.	
diag	Diagonal of a matrix.	
$\boldsymbol{x}_t$	The state variable at time index t.	section 1.2
$\boldsymbol{y}_t$	The observation at time index t.	section 1.2
$f(\boldsymbol{x}_t)$	The dynamical forecast model.	section 1.2
$h(\boldsymbol{x}_t)$	The observation model.	section 1.2
$\boldsymbol{x}^a, \boldsymbol{x}^f$	The analysis/forecast *exact* mean. $\boldsymbol{x}^a = \mathbb{E}(\boldsymbol{x}_t \vert \boldsymbol{y}_{1:t})$, $\boldsymbol{x}^f = \mathbb{E}(\boldsymbol{x}_t \vert \boldsymbol{y}_{1:t-1})$.	section 2.1
$\mathbf{P}^a, \mathbf{P}^f$	The analysis/forecast *exact* covariance matrix. $\mathbf{P}^a = \mathbb{V}\text{ar}(\boldsymbol{x}_t \vert \boldsymbol{y}_{1:t})$, $\mathbf{P}^f = \mathbb{V}\text{ar}(\boldsymbol{x}_t \vert \boldsymbol{y}_{1:t-1})$.	section 2.1
$\boldsymbol{x}_n$	An ensemble member/realisation. Typically associated with a given state (and thus time index), $\boldsymbol{x}_t$, as well as a given conditioning, e.g. $\vert \boldsymbol{y}_{1:t}$. If this is of relevance, it is indicated by superscripts a and f, as described below.	section 2.2

Symbol	Definition	Reference
$\mathbf{E}, \mathbf{E}^a, \mathbf{E}^f$	The ensemble matrix, whose columns are the individual members. $\mathbf{E} = \begin{bmatrix} \boldsymbol{x}_1, & \dots & \boldsymbol{x}_n, & \dots & \boldsymbol{x}_N \end{bmatrix} \in \mathbb{R}^{m \times N}$. The superscript a (resp. f) indicates that it is an "analysis" (resp. "forecast") entity: it estimates $\boldsymbol{x}_t$ based on $\boldsymbol{y}_{1:t}$ (resp. $\boldsymbol{y}_{1:t-1}$).	section 2.2
$\mathbf{A}, \mathbf{A}^a, \mathbf{A}^f$	The ensemble anomalies matrix. $\mathbf{E} = \boldsymbol{x}\mathbb{1}^\mathsf{T} + \mathbf{A}$.	section 2.2
$\bar{\boldsymbol{x}}, \bar{\boldsymbol{x}}^a, \bar{\boldsymbol{x}}^f$	The *ensemble* mean. $\bar{\boldsymbol{x}} = \frac{1}{N}\mathbf{E}\mathbb{1}$.	section 2.2
$\bar{\mathbf{P}}, \bar{\mathbf{P}}^a, \bar{\mathbf{P}}^f$	The *ensemble* covariance. $\bar{\mathbf{P}} = \frac{1}{N-1}\mathbf{A}\mathbf{A}^\mathsf{T}$.	section 2.2
$\bar{\mathbf{K}}, \mathbf{K}$	The ensemble and exact Kalman gain matrices.	section 2.3
$\bar{\boldsymbol{\delta}}$	Mean innovation.	section 4.1
$\mathbf{Y}$	Anomalies of the observed ensemble.	section 4.1
$\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}, \tilde{\boldsymbol{\xi}}$	Random variables drawn from $\mathcal{N}(0, \mathbf{I}_m)$.	chapter 5
$\Xi, \hat{\Xi}, \tilde{\Xi}$	Matrices consisting of N realisations of $\boldsymbol{\xi}$ (or $\hat{\boldsymbol{\xi}}$ or $\tilde{\boldsymbol{\xi}}$) concatenated together.	chapter 5
$\boldsymbol{\Pi}$	Orthogonal projection matrix.	page 13
$\|\boldsymbol{v}\|^2_{\mathbf{M}}$	$= \boldsymbol{v}^\mathsf{T}\mathbf{M}^{-1}\boldsymbol{v}$. 2-norm of the generic vector $\boldsymbol{v}$ weighted by the inverse invertible matrix $\mathbf{M}$.	
$\|\boldsymbol{v}\|^2_{\mathbf{M}}$	$= \boldsymbol{v}^\mathsf{T}\mathbf{M}^{-1}\boldsymbol{v}$. 2-norm of the generic vector $\boldsymbol{v}$ weighted by the inverse invertible matrix $\mathbf{M}$.	
$\boldsymbol{w}$	Coefficients for linear combinations of the ensemble anomalies such that $\boldsymbol{x}(\boldsymbol{w}) = \bar{\boldsymbol{x}} + \mathbf{A}\boldsymbol{w}$.	section 6.2
$\boldsymbol{b}, \mathbf{B}$	Exact mean and covariance of the forecast distribution.	chapter 6
$\bar{\boldsymbol{x}}, \bar{\mathbf{B}}$	Mean and covariance of the forecast ensemble.	chapter 6
ε_N	$= 1 + 1/N$.	section 6.3.1
c_g	$= N + g$, where g is the dimensionality of the null space of $\mathbf{A}$.	section 6.4

Standard probability distributions

Symbol. Name.	Probability density function. Moments.	
$\mathcal{N}(\boldsymbol{x}\|\boldsymbol{\mu},\boldsymbol{\Sigma})$	$= \|2\pi\boldsymbol{\Sigma}\|^{-1/2}\exp\Big(-\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}}\Big)$.	
Gaussian/Normal.	Mode, Mean $= \boldsymbol{\mu}$. Var: $\boldsymbol{\Sigma}$.	
$\boldsymbol{t}_m(\boldsymbol{x}\|\nu;\boldsymbol{\mu},\boldsymbol{\Sigma})$	$= c\|\pi\boldsymbol{\Sigma}\|^{-1/2}\Big(1+\frac{1}{\nu}\|\boldsymbol{x}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}}\Big)^{-(\nu+m)/2}$,	$c = \frac{\Gamma(\frac{\nu+m}{2})}{\nu^{m/2}\Gamma(\nu/2)}$.
t distribution.	Mode, Mean $= \boldsymbol{\mu}$. Var: $\frac{\nu}{\nu-2}\boldsymbol{\Sigma}$.	
$\mathcal{W}_m^{+1}(\mathbf{S}\|\nu,\boldsymbol{\Sigma})$	$= c\|\boldsymbol{\Sigma}\|^{-\frac{\nu}{2}}\|\mathbf{S}\|^{(\nu-m-1)/2}\exp\Big(-\frac{1}{2}\mathsf{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1})\Big)$,	$c = \frac{1}{2^{\frac{\nu m}{2}}\Gamma_m(\nu/2)}$.
Wishart.	Mode $= (\nu-m-1)\boldsymbol{\Sigma}$. Mean $= \nu\boldsymbol{\Sigma}$.	
$\mathcal{W}_m^{-1}(\boldsymbol{\Sigma}\|\nu,\mathbf{S})$	$= c\|\mathbf{S}\|^{\frac{\nu}{2}}\|\boldsymbol{\Sigma}\|^{-(\nu+m+1)/2}\exp\Big(-\frac{1}{2}\mathsf{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1})\Big)$,	c: as above.
Inv. Wishart.	Mode $= \frac{1}{\nu+m+1}\mathbf{S}$. Mean $= \frac{1}{\nu-m-1}\mathbf{S}$.	
$\chi^{+2}(s\|\nu)$	$= cs^{\nu/2-1}e^{-\nu s/2}$,	$c = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}$.
Chi-square.	Mode $= \frac{\nu-2}{\nu}$. Mean $= 1$. Var $= 2/\nu$.	
$\chi^{-2}(s\|\nu)$	$= cs^{-\nu/2-1}e^{-\frac{1}{2}\nu/s}$,	c: as above.
Inv. chi-square.	Mode $= \frac{\nu}{\nu+2}$. Mean $= \frac{\nu}{\nu-2}$. Var $= \frac{2\nu^2}{(\nu-2)^2(\nu-4)}$.	

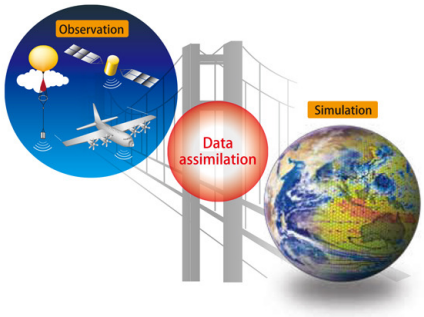
The $m \times m$ matrices $\boldsymbol{\Sigma}$ and $\mathbf{S}$ are restricted to being symmetric, positive-definite, and the scalar s must be positive. Further details are provided by Gelman et al. [72, appx. A].

Chapter 1

Introduction

1.1 Data assimilation

Data assimilation (DA) is the process of combining model forecasts with observation data; it is a set of techniques specialized for sequential, statistical inference on dynamic model variables, in addition to static parameters [208]. The typical application of DA is forecast initialization, but it is also used for reanalysis, i.e. the estimation of historical states.



DA "bridges" data and models.[1]

Modern DA builds on "state estimation" techniques developed in control systems engineering. For example, the Kalman filter (section 2.1) was originally developed to steer the Apollo mission spacecraft [78, 131]. In this case, the state variables to be estimated are the location and the (linear and angular) velocity and acceleration; the dynamics are the kinematic equations; and the observations come from accelerometers, gyroscopes, and sextant measurements. In DA, the aim is to condition the estimate on all of the available data; indeed, if the velocity (resp. acceleration) of the spacecraft example is assumed to be a fixed but unknown parameter, then the Kalman filter is the recursive, on-line solution to the first (resp. second) degree polynomial linear regression problem for the trajectory of the spacecraft [178].

In contrast to the above example, in geoscience, the state variables are often the result of the discretization of a physical field, and may number in the billions [149], and the dynamics are typically nonlinear. Examples of geoscientific DA include

[1] Source and permission: Data Assimilation Research Team, `www.aics.riken.jp`.

- "History matching" in reservoir engineering for oil production planning and optimization [1, 43, 63, 157, 167]. Here, the state variables include pressure and saturation fields; fixed, but unknown, parameters include permeability and porosity fields; the dynamics are modelled using a combination of Darcy's law, material balance, and thermodynamic equations of state.
- Meteorology and weather forecast initialization [16, 118, 124]. Here, the state variables include velocity, pressure, and temperature fields; the dynamics consist of the Navier-Stokes equations, the first law of thermodynamics and the ideal gas law; observations come from diverse satellite instruments, weather stations, and buoys.
- Epidemiology analysis and forecast initialization [12, 90, 148, 181]. Here, the state variable is the number count of people susceptible, infected, and recovered; the dynamics are modelled by a set of coupled differential equations, which may be spatially distributed using agent-based modelling or diffusion; observation include compiled reports from government health agencies and "Google Flu Trends".

Other application areas of DA include oceanography [19, 21, 174, 210], climatology on Earth [13] and Mars [89], hydrology [166, 197], atmospheric chemistry and air quality analysis [83], and forest fire prediction [132].

1.2 Principles of sequential inference

Several approaches to the DA problem can be classified as Bayesian, sequential inference. Its main principles are introduced here.

1.2.1 The hidden Markov model

Following Jazwinski [102], suppose the state and observation, $\boldsymbol{x}_t \in \mathbb{R}^m$ and $\boldsymbol{y}_t \in \mathbb{R}^p$ respectively, are generated for sequentially increasing t, by a dynamical model, $f : \mathbb{R}^m \to \mathbb{R}^m$, and a measurement model, $h : \mathbb{R}^m \to \mathbb{R}^p$, as follows:

$$\boldsymbol{x}_{t+1} = f(\boldsymbol{x}_t) + \boldsymbol{q}_t\,, \qquad t = 0, 1, \ldots\,, \tag{1.1}$$

$$\boldsymbol{y}_t = h(\boldsymbol{x}_t) + \boldsymbol{r}_t\,, \qquad t = 1, 2, \ldots\,, \tag{1.2}$$

where the Gaussian white noise processes $\{\boldsymbol{q}_t \ ; \ t = 0, 1, \ldots\}$ and $\{\boldsymbol{r}_t \ ; \ t = 1, 2, \ldots\}$, and the initial condition, $\boldsymbol{x}_0$, are specified by:

$$\boldsymbol{x}_0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \mathbf{P}_0)\,, \qquad \boldsymbol{q}_t \sim \mathcal{N}(0, \mathbf{Q})\,, \qquad \boldsymbol{r}_t \sim \mathcal{N}(0, \mathbf{R})\,, \tag{1.3}$$

where $\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ is the multivariate Gaussian probability law with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$. Note that we do not use upper and lowercase to distinguish between random variables and their realizations. Generalizations to time-dependent $\mathbf{Q}, \mathbf{R}, f$, and h are straightforward for all of the theory developed in this thesis. The models and parameters, f, h, $\boldsymbol{\mu}_0$, $\mathbf{P}_0, \mathbf{Q}$ and $\mathbf{R}$, are all assumed known.

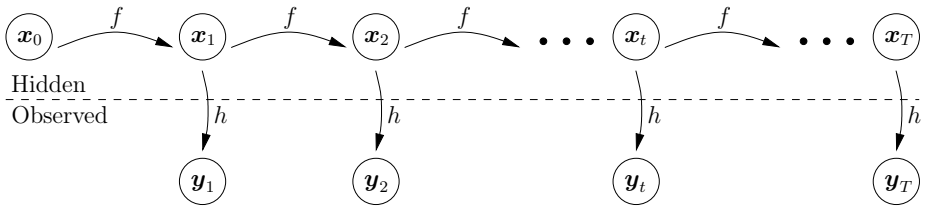


Figure 1.1: Diagram of the HMM of eqns. (1.1) and (1.2). The arrows indicate causality.

The process $\{(\boldsymbol{x}_t, \boldsymbol{y}_t) \ ; \ t = 1, 2, \ldots\}$, illustrated in Fig. 1.1, constitutes a hidden Markov model (HMM): a sequence of (possibly) hidden states linked together by a forecast law, such as eqn. (1.1), which are only observed through an observation law, such as eqn. (1.2). For generic random variables (possibly vectors) $\boldsymbol{u}$ and $\boldsymbol{v}$, denote $p(\boldsymbol{u}|\boldsymbol{v})$ the probability density function (pdf) of $\boldsymbol{u}$ conditioned on $\boldsymbol{v}$. The fundamental properties of a HMM, which explain its name, and can be derived from eqns. (1.1) to (1.3), are the independence relations

$$p(\boldsymbol{x}_{t+1} \mid \boldsymbol{x}_{0:t}) = p(\boldsymbol{x}_{t+1} \mid \boldsymbol{x}_t)\,, \tag{1.4}$$

$$p(\boldsymbol{y}_t \mid \boldsymbol{x}_{0:T}) = p(\boldsymbol{y}_t \mid \boldsymbol{x}_t)\,, \tag{1.5}$$

for any two time indices $t \leq T$, where the colon is used to indicate a sequence, e.g. $\boldsymbol{x}_{0:T} = \{\boldsymbol{x}_t \ ; \ t = 0, 1, \ldots, T\}$. Being very general, but possessive of the above properties (and more [32]), the HMM is a useful abstraction for building inference techniques for dynamical systems [142, §10.2.2].

An advantage of ensemble DA methods is that they are non-invasive; indeed, it is not uncommon that the dynamical model, f, is only available as a binary executable. Nevertheless, though f is then treated as a generic "black-box", the dynamics of the forecast law, eqn. (1.1), are typically a discretization of an underlying time-continuous, physical system. For example, if $f(\boldsymbol{x}) = \boldsymbol{x} + \Delta t\, u(\boldsymbol{x})$, for some function u, and if $\mathbf{Q}$

scales with $\sqrt{\Delta t}$, then the stochastic difference equation, eqn. (1.1), may be regarded as the Euler-Maruyama discretization of the time-continuous, m-vector, Itô stochastic differential equation known as the (simplified) Langevin equation,

$$d\boldsymbol{x} = u(\boldsymbol{x}, t)\, dt + d\boldsymbol{q}, \quad t > 0\,. \tag{1.6}$$

Conditions for the existence of solutions to eqn. (1.6) are discussed by Jazwinski [102, §4.4], while Gardiner [71, §10] provides a discussion of the convergence of eqn. (1.1) to eqn. (1.6) as $\Delta t \to 0$. Although $\boldsymbol{x}(t)$ is a random variable (process), its pdf, $p(\boldsymbol{x}; t)$, evolves *deterministically* in time, according to the Fokker-Planck equation [102, §4.4],

$$\frac{\partial p}{\partial t} = -\nabla.(up) + \frac{1}{2}\nabla.(\mathbf{Q}\nabla p)\,, \tag{1.7}$$

with far-field boundary conditions $p(\boldsymbol{x}; t) \to 0$ as $\|\boldsymbol{x}\|_2 \to \infty$ for all $t > 0$. An example solution to eqn. (1.7) with a scalar $\boldsymbol{x}$ is illustrated in Fig. 1.2.

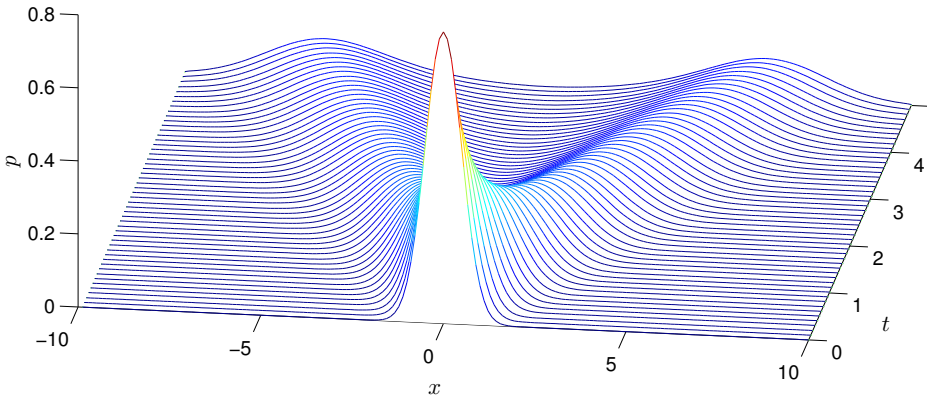


Figure 1.2: Illustration of the integration of the Fokker-Planck equation (1.6) with $u(x) = \mathsf{arctan}(x)$, $\mathbf{Q} = 0.5$, and $p(x) = \mathcal{N}(x|0, 1/4)$ at $t = 0$.

1.2.2 Bayesian data assimilation

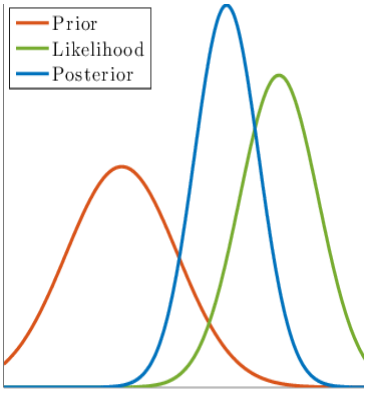
DA is now formulated as statistical inference on a HMM. The principal objective in Bayesian inference is to compute $p(\boldsymbol{x}|\boldsymbol{y})$, resulting from Bayes' rule:

$$p(\boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x})\, p(\boldsymbol{x})\,, \tag{1.8}$$

where $\boldsymbol{y}$ is any data available, and $\boldsymbol{x}$ is the unknown that we wish to estimate. The constant of proportionality is $1/\,p(\boldsymbol{y})$. In other words, the proportionality is "with respect to $\boldsymbol{x}$". The pdf $p(\boldsymbol{x}|\boldsymbol{y})$ is known as the *posterior*, resulting from the pointwise

(a) P.S. Laplace, 1749 – 1827. Pioneer of inference.

(b) C.F. Gauss, 1777 – 1855. Pioneer of inference.



(c) Illustration of Bayes' rule, eqn. (1.8), with Gaussian distributions.

multiplication of the *likelihood*, $p(\boldsymbol{y}|\boldsymbol{x})$, our knowledge from the observations, and the *prior*, $p(\boldsymbol{x})$, a pdf in $\boldsymbol{x}$ quantifying our prior information about $\boldsymbol{x}$. Note the usage of "overloading" notation: p is the pdf of the random variable identified by the symbol of its input argument.

In DA, the unknowns are the state vectors of the dynamics, $\boldsymbol{x}_{0:T}$, and the data are the observation vectors, $\boldsymbol{y}_{1:T}$. The objective is then to compute the pdf, $p(\boldsymbol{x}_{0:t}|\boldsymbol{y}_{1:T})$, though some of these time indices may be dropped, depending on the application of the estimation. Suppose we are only interested in the marginal pdf, $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:T})$; then the estimation problem is called

- smoothing if $t < T$;
- filtering if $t = T$;
- forecasting (or prediction) if $t > T$.

The nomenclature is also used when the aim is to compute a joint pdf (i.e. where there are multiple time indices on $\boldsymbol{x}$). As shown in section 1.2.3 and chapter 7, filtering is typically part of the procedures for forecasting and smoothing. Hence some ideas are expressed through modifications to the filtering algorithm only, even though they also apply to the filtering component of prediction or smoothing algorithms.

Before deriving the filter recursions, suppose we want to obtain the joint posterior, $p(\boldsymbol{x}_{0:T}|\boldsymbol{y}_{1:T})$, for sequentially increasing time indices, T. Using Bayes' rule (1.8),

$$p(\boldsymbol{x}_{0:T}|\boldsymbol{y}_{1:T}) \propto p(\boldsymbol{x}_{0:T})\, p(\boldsymbol{y}_{1:T}|\boldsymbol{x}_{0:T})\,. \tag{1.9}$$

Next, eqns. (1.4) and (1.5) imply that

$$p(\boldsymbol{x}_{0:T}|\boldsymbol{y}_{1:T}) \propto p(\boldsymbol{x}_0) \prod_{t=1}^{T} p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})\, p(\boldsymbol{y}_t|\boldsymbol{x}_t)\,, \tag{1.10}$$

inducing the recursion

$$p(\boldsymbol{x}_{0:t}|\boldsymbol{y}_{1:t}) \propto p(\boldsymbol{y}_t|\boldsymbol{x}_t)\, p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})\, p(\boldsymbol{x}_{0:t-1}|\boldsymbol{y}_{1:t-1})\,. \tag{1.11}$$

Thus joint, sequential inference can be performed by recursively appending the factor $p(\boldsymbol{y}_t|\boldsymbol{x}_t)\, p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ to the previous posterior.

1.2.3 Filtering

Filtering is the procedure whereby one computes the pdf $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$, typically in order to initialize a probabilistic forecast. The equations may be derived by marginalization of the joint inference equations (1.11), but is abbreviated here by deliberately seeking a recursion in the marginal pdf. To that end, assume that $p(\boldsymbol{x}_{t-1}|\boldsymbol{y}_{1:t-1})$ is available, or $p(\boldsymbol{x}_0)$ if $t = 1$. Then, by the Markov property, eqn. (1.4), the "forecast" pdf is given by the Chapman-Kolmogorov equation:

$$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1}) = \int p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})\, p(\boldsymbol{x}_{t-1}|\boldsymbol{y}_{1:t-1})\, d\boldsymbol{x}_{t-1}\,, \tag{1.12}$$

where the integral is over the state space, $\mathbb{R}^m$. Next, by virtue of the other conditional independence property of the HMM, eqn. (1.5), and Bayes' rule, the filtered, posterior, or "analysis", pdf is given by:

$$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t}) \propto p(\boldsymbol{y}_t|\boldsymbol{x}_t)\, p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1})\,. \tag{1.13}$$

These two equations constitute a cycle that can be repeated sequentially in time to obtain the filtered pdfs.

As mentioned in section 1.2.2, filtering is the archetypical sequential inference procedure. Indeed, in the linear-Gaussian case, the recursions of eqns. (1.12) and (1.13) reduce to a famous set of matrix formulae known as the Kalman filter. In geoscientific DA, however, $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ may be very high-dimensional, and f and h may be highly nonlinear (section 1.1). Therefore, only approximate solution methods are possible.

The family of techniques deriving from the ensemble Kalman filter (EnKF) provides a set of such methods. Since their inception [57], they have acquired a prominent role in geoscientific DA, having proved both capable and practical in some applications. However, while the ensemble approach is promising, there is still room for improvement in formalism, understanding, and performance. This thesis proposes some improvements to ensemble DA methods, notably in the areas of model noise incorporation, and the correction of an intrinsic bias. It is also shown that two different ensemble smoothing algorithms are equivalent.

1.3 Organization of the thesis

Chapter 2 presents the Kalman filter. A derivation and interpretation is given in appendix A. The ensemble Kalman filter is then introduced. Subsequently, its main properties are discussed, and a survey of related efforts for solving the filtering problem is provided. Chapter 3 describes the setup of the numerical experiments, including the metric used to assess the DA methods, and the dynamical models used. Benchmarks generated from these experiments are provided throughout the rest of the thesis.

Chapter 4 presents the "square root" scheme for the analysis step of the EnKF, In particular, a survey of technical details, important for chapter 5, is provided. Motivated by the scheme of chapter 4, the main objective of chapter 5 is to develop a new method to account for model noise in the forecast step. It begins with a discussion on the relevance of model noise as a DA modelling device, as well as a survey of methods to account for model noise in the EnKF. After the core method has been derived, other methods are surveyed. Subsequently, rank-related problems are illuminated, and two possible solutions are proposed. In addition to studying the properties of the square root method in general, several theoretical results concerning the new techniques are derived, some of which are given in appendix B.

Chapter 6 reviews the "finite size" ensemble Kalman filter (EnKF-N). The introduction consists of a re-derivation of the standard EnKF analysis step by the simple but explicit assumption that the ensemble estimates of the forecast moments are exact. The EnKF-N is then developed by rejecting this assumption; the prior and posterior distributions are derived with careful attention to the parameterization. The EnKF-N is then reverse engineered to reveal its intimate connection with inflation. Alternative approaches to defining the inflation factor are discussed in appendix C. The theoretical choice of the inflation value is then discussed, leading to a correction of a deficiency in the performance of the EnKF-N.

Chapter 7 studies the ensemble formulation of the Rauch-Tung-Striebel (RTS) smoother and proves its equivalence to the ensemble Kalman smoother. First, the general, linear-Gaussian, and ensemble formulations of both smoothers are described. The equivalence is then proved, and subsequently re-examined under practical considerations. The accompanying appendix D provides a derivation of the linear-Gaussian formulation of the RTS smoother.

1.4 Original contributions

Section 2.1 is a review of standard results, but the accompanying derivation of the Kalman filter equations of appendix A is a synthesis with the aim of being self-contained and didactic, despite accommodating a large part of the myriad of possible formulations. The rest of chapter 2 then discusses the EnKF, drawing on multiple sources, hoping to extend their width and precision. For example, considerations of the type of Lemma 2.1 are rarely mentioned, while the topics of section 2.4, such as sampling error, bias, and consistency are sometimes confused. It is also hoped that the holistic effort of section 2.5 to explain why the EnKF works, in spite of a limited ensemble size, is useful as a review.

The description of the experiments, chapter 3, is only a synthesis. However, the number of baseline methods that are described and included in the benchmarks is a point of merit, and both the depth and width of section 3.2 is rarely encountered in the EnKF literature, though it is well known in general.

Chapter 4 on the analysis step square root method is mainly a synthesis, though some modest elements of section 4.4 are original. Chapter 5 on the forecast step square root method is mainly original. A somewhat abridged version has been published as a journal article in the Monthly Weather Review [160]. Unless otherwise stated, the original idea and theoretical developments (including appendix B) are entirely my own work. All of the numerical experiments and writing is my own, but I have benefited from much guidance and feedback from Alberto Carrassi and Laurent Bertino during the process.

Chapter 6 on the EnKF-N is a collaboration with the inventor of the method, Professor Marc Bocquet, where I have taken the lead on several items, as described below. Although section 6.2 is only a re-derivation of the square root analysis update for the EnKF, it is original in its rigour and technical details. Sections 6.3 and 6.4 re-derive the EnKF-N. Original contributions include composing the integral developments (based on the Jeffreys prior) of the earlier literature; clarifying the discussion

on the optimization of the posterior; a more careful treatment of the restricted dimensionality of the ensemble subspace and of its parameterization; and including more references to (and understanding from) related results in the statistics literature. Section 6.5 and the accompanying appendix C is original and not previously published. Although all of its developments are my own, independent work, the original idea and perspective is attributable to Prof. Bocquet. Half of the experimental tests in section 6.6 are reproductions. The other half are original. Section 6.7 is mainly my independent work. It was published, along with some of the thoughts of sections 6.2 to 6.4 in an article in the Nonlinear Processes in Geophysics [30], which was selected as a highlighted article by the 2016 edition of the EGU conference. The rest of the chapter is considered for future publication.

Unless otherwise stated, originality is claimed for the entirety of chapter 7. A slightly abridged version has been published in the Quarterly Journal of the Royal Meteorological Society [161].

Appendix E is original as a synthesis emphasizing the connection between the singular value decomposition, linear inverse problems, and the pseudoinverse.

Chapter 2

The ensemble Kalman filter

The recursions established in section 1.2.3 avoid repeating the calculations in full after each time step. The recursions represent the fundamental exploitation of the time-sequential structure of the estimation problem on HMMs, thus reducing the problem's dimensionality and complexity. Nevertheless, the size and nonlinearity of the systems remain challenging, necessitating the use of approximate solution methods. One approximate solution method is the ensemble Kalman filter (EnKF).

This thesis is mainly concerned with improving aspects of the EnKF. The EnKF builds on the Kalman filter (KF), which is developed in section 2.1. Sections 2.2 and 2.3 develop the basic theory of the EnKF. Sections 2.4 and 2.5 discuss its properties, including bias and consistency. An initial survey of related efforts for solving the filtering problem is provided in section 2.6.

2.1 The Kalman filter

The KF is the closed-form solution to the Bayesian filtering equations (1.12) and (1.13) obtained in the linear, Gaussian case. This section provides a summary of the KF; a complete derivation can be found in section A.2. Suppose the HMM, eqns. (1.1) and (1.2), is linear, i.e. that $f(\boldsymbol{x}) = \mathbf{F}\boldsymbol{x}$ and $h(\boldsymbol{x}) = \mathbf{H}\boldsymbol{x}$, where $\mathbf{F} \in \mathbb{R}^{m^2}$ and $\mathbf{H} \in \mathbb{R}^{p \times m}$. It can then be shown that the pdfs involved remain Gaussian throughout the filtering process. Thus, the pdfs are fully characterized by their first two moments, labelled as follows:

$$\boldsymbol{x}^f = \mathbb{E}(\boldsymbol{x}_t | \boldsymbol{y}_{1:t-1})\,, \qquad \mathbf{P}^f = \mathbb{V}\mathrm{ar}(\boldsymbol{x}_t | \boldsymbol{y}_{1:t-1})\,, \tag{2.1}$$

$$\boldsymbol{x}^a = \mathbb{E}(\boldsymbol{x}_t | \boldsymbol{y}_{1:t})\,, \qquad \mathbf{P}^a = \mathbb{V}\mathrm{ar}(\boldsymbol{x}_t | \boldsymbol{y}_{1:t})\,. \tag{2.2}$$

where $\mathbb{E}(.)$ and $\mathbb{V}\text{ar}(.)$ are the multivariate expectation and variance operators, respectively (i.e. $\mathbb{V}\text{ar}(.)$ outputs a square *matrix*). As is convention [100], the superscripts f or a are used to distinguish between forecast and analysis objects, respectively. The explicit time indices are dropped because the focus will typically be restricted to a single index anyway.

Assume initially that $p(\boldsymbol{x}_{t-1} \mid \boldsymbol{y}_{1:t-1}) = \mathcal{N}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}^a, \mathbf{P}^a)$, where $\boldsymbol{x}^a$ and $\mathbf{P}^a$ are associated with the time index, $t-1$. As shown in section A.2, the forecast equation (1.12) yields $p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t-1}) = \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{x}^f, \mathbf{P}^f)$ with

$$\boldsymbol{x}^f = \mathbf{F}\boldsymbol{x}^a \,, \tag{2.3}$$

$$\mathbf{P}^f = \mathbf{F}\mathbf{P}^a\mathbf{F}^\mathsf{T} + \mathbf{Q} \,, \tag{2.4}$$

where the time interval between the *indices* t and $t-1$ should be included as a scaling of $\mathbf{Q}$. Next, the analysis equation (1.13) implies that $p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t}) = \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{x}^a, \mathbf{P}^a)$, where $\boldsymbol{x}^a$ and $\mathbf{P}^a$ are now associated with t, and are given by

$$\boldsymbol{x}^a = \boldsymbol{x}^f + \mathbf{K}[\boldsymbol{y}_t - \mathbf{H}\boldsymbol{x}^f] \,, \tag{2.5}$$

$$\mathbf{P}^a = [\mathbf{I}_m - \mathbf{K}\mathbf{H}]\mathbf{P}^f \,, \tag{2.6}$$

where the "Kalman gain matrix", $\mathbf{K} \in \mathbb{R}^{m \times p}$, is

$$\mathbf{K} = \mathbf{P}^f\mathbf{H}^\mathsf{T}\left(\mathbf{H}\mathbf{P}^f\mathbf{H}^\mathsf{T} + \mathbf{R}\right)^{-1} \,. \tag{2.7}$$

The KF thus consists in repeating the matrix computations of eqns. (2.3) to (2.6) for sequentially increasing time indices t.

The Kalman gain can also be derived from the perspective of optimal point estimation. Assuming Gaussianity as before, any convex loss function yields the posterior mean (or median or mode) as its optimum [182], and thereby the KF equations are recovered. Alternatively, without assuming Gaussianity, it can be derived as the minimum square error *linear estimator* [5, Ths. 2.1,2.3]. Derivations under the heading of orthogonal projections or recursive least squares [102, §7.3] can be formulated through one of the above [5, Th. 2.5].

The applicability of the KF equations can be extended to the case of nonlinear models by linearization. That is, while the formulae (2.3) and (2.5) are computed using the full nonlinear models, formulae (2.4), (2.6) and (2.7) are computed with $\mathbf{F}$ and $\mathbf{H}$ being the linearization of f and h respectively. This first order approximation to the exact solution is known as the extended Kalman filter.

2.2 Ensemble preliminaries

The EnKF is an algorithm that approximately generates an ensemble, i.e. an iid sample, $\boldsymbol{x}_{1:N} = \{\boldsymbol{x}_n \ ; \ n \in 1{:}N\}$, from the Bayesian filtering distributions (1.12) and (1.13), recursively in time, for sequentially increasing time indices, t. More vaguely, it is also sometimes helpful to think of the ensemble as a "cloud" in the phase space of $\boldsymbol{x}$ representing its pdf.

The positive integer N is used to denote ensemble size, while m and p are used to denote state and observation vector sizes, and a colon is used to indicate an integer sequence. For convenience, we concatenate all of the state realizations, or "ensemble members", into the "ensemble matrix":

$$\mathbf{E} = \Big[\boldsymbol{x}_1, \ \ \ldots \ \ \boldsymbol{x}_n, \ \ \ldots \ \ \boldsymbol{x}_N\Big] \in \mathbb{R}^{m \times N} \,. \tag{2.8}$$

The overhead bar is used to designate the ensemble-estimate counterparts to the exact mean and covariance matrix of eqns. (2.1) and (2.2);

$$\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n \,, \qquad\qquad \bar{\mathbf{P}} = \frac{1}{N-1}\sum_{n=1}^{N} (\boldsymbol{x}_n - \bar{\boldsymbol{x}})\,(\boldsymbol{x}_n - \bar{\boldsymbol{x}})^{\mathsf{T}} \,. \tag{2.9}$$

These are the canonical, unbiased [e.g. 141, §3.1] sample estimators. However, in EnKF contexts they should be regarded mainly as a conventional notational tool; indeed, a significant body of research (e.g. chapter 6) deals with adjusting these estimates. For a given ensemble, its "anomalies" are defined as

$$\mathbf{A} = \Big[\boldsymbol{x}_1 - \bar{\boldsymbol{x}}, \ \ \ldots \ \ \boldsymbol{x}_n - \bar{\boldsymbol{x}}, \ \ \ldots \ \ \boldsymbol{x}_N - \bar{\boldsymbol{x}}\Big] \,. \tag{2.10}$$

Note that the ensemble mean, anomalies, and covariance matrix can be conveniently expressed and computed as

$$\bar{\boldsymbol{x}} = \frac{1}{N}\mathbf{E}\mathbb{1} \,, \qquad \mathbf{A} = \mathbf{E}\Big(\mathbf{I}_N - \mathbb{1}\mathbb{1}^{\mathsf{T}}/N\Big) \,, \qquad (N-1)\bar{\mathbf{P}} = \mathbf{A}\mathbf{A}^{\mathsf{T}} \,, \tag{2.11}$$

where $\mathbb{1}$ is the vector of ones of length N, and $\Big(\mathbf{I}_N - \mathbb{1}\mathbb{1}^{\mathsf{T}}/N\Big) \in \mathbb{R}^{N^2}$ should be recognized as $\mathbf{\Pi}_{\mathbb{1}}^{\perp}$, the orthogonal projector[1] onto $\mathsf{range}(\mathbb{1})^{\perp}$, the orthogonal complement space to $\mathsf{range}(\mathbb{1})$.

[1] A matrix $\mathbf{\Pi}$ is an orthogonal projection matrix if it satisfies $\mathbf{\Pi}^2 = \mathbf{\Pi}$ and $\mathbf{\Pi} = \mathbf{\Pi}^{\mathsf{T}}$.

As will be seen, these relations are not just helpful because they (i) abbreviate eqns. (2.9) and (2.10) and avoid the member index, n; they also (ii) highlight linearity aspects of the operations involved; (iii) give insight into subspace rank issues; and (iv) emphasize that sometimes the ensemble may be seen as a deterministic parameterization of the pdf.

2.3 The EnKF algorithm

As with the KF, the EnKF consists of the recursive application of a forecast step and an analysis step. This section follows the traditional [57] template, presenting the EnKF as the "the Monte Carlo version of the KF where the state covariance is estimated by the ensemble covariance". It is not obvious that this postulated method should work; indeed, it is only justified upon inspection of its properties, deferred to section 2.4. An improved *derivation* is given in section 6.2.

As in eqn. (2.1), the time indices of the state and conditioning are implied by the superscript f or a for the ensemble. This indicates that $\boldsymbol{x}^f_{1:N}$ (resp. $\boldsymbol{x}^a_{1:N}$) is a forecast (resp. analysis) ensemble, and is also used for the derivative objects, $\mathbf{E}, \mathbf{A}, \bar{\boldsymbol{x}}, \bar{\mathbf{P}}$.

2.3.1 Forecast step

For a given, implicit, time index, t, assume $\boldsymbol{x}^a_{1:N}$ is an iid sample from $p(\boldsymbol{x}_{t-1}|\boldsymbol{y}_{1:t-1})$, which is not necessarily Gaussian. The forecast step of the EnKF consists of a Monte Carlo simulation of eqn. (1.1): for each $n \in 1{:}N$, $\boldsymbol{x}^a_n$ is propagated in time through the forecast, dynamical model, f, and a simulated noise realization, $\boldsymbol{q}_n$ is added to it

$$\forall n, \quad \boldsymbol{x}^f_n = f(\boldsymbol{x}^a_n) + \boldsymbol{q}_n \,, \tag{2.12}$$

$$\text{or}, \quad \mathbf{E}^f = f(\mathbf{E}^a) + \mathbf{D}_{\text{mod}} \,, \tag{2.13}$$

where f is applied column-wise to $\mathbf{E}^a$, and the columns of $\mathbf{D}_{\text{mod}}$ are sampled iid from $\mathcal{N}(0, \mathbf{Q})$. The ensemble, $\boldsymbol{x}^f_{1:N}$, is then an iid sample from the forecast pdf, $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1})$.

Lemma 2.1.
Assuming the ensemble (forecast or analysis) is iid drawn from a non-degenerate pdf, then, almost surely,

$$\mathsf{rank}(\mathbf{E}) = \mathsf{min}(m, N) \,, \qquad \mathsf{rank}(\mathbf{A}) = \mathsf{min}(m, N-1) \,. \tag{2.14}$$

Proof. The size of $\mathbf{E}$ and $\mathbf{A}$ imply that their ranks are bounded by m and N. For $N \leq m$, any strict subset of the ensemble members spans a strict subspace of $\mathbb{R}^m$, which is of (probability) measure zero. This can be used inductively to prove almost sure linear independence between the ensemble members. However, the fact that $\mathbf{A}\mathbb{1} = 0$, as seen from eqn. (2.11), reduces the rank of $\mathbf{A}$ by one. More details are given by Gupta and Nagar [79, Th. 3.2.1] and Muirhead [141, Th. 3.1.4]. □

Because of possible nonlinear dynamics, and finite precision, the assumption of Lemma 2.1 is frequently not satisfied in practice. But it provides useful upper bounds on the ranks that are convenient to keep in mind when manipulating the ensemble matrices.

2.3.2 Analysis step

The analysis update of the ensemble is given by:

$$\forall n, \quad \boldsymbol{x}_n^a = \boldsymbol{x}_n^f + \bar{\mathbf{K}}\left\{\boldsymbol{y} - \boldsymbol{r}_n - h(\boldsymbol{x}_n^f)\right\}, \tag{2.15}$$

$$\text{or,} \quad \mathbf{E}^a = \mathbf{E}^f + \bar{\mathbf{K}}\left\{\boldsymbol{y}\mathbb{1}^\mathsf{T} - \mathbf{D}_{\text{obs}} - h(\mathbf{E}^f)\right\}, \tag{2.16}$$

where the "observation perturbations", $\boldsymbol{r}_n$, are sampled iid from $\mathcal{N}(0, \mathbf{R})$ and form the columns of $\mathbf{D}_{\text{obs}}$, and h is applied column-wise to $\mathbf{E}^f$. If the forecast distribution

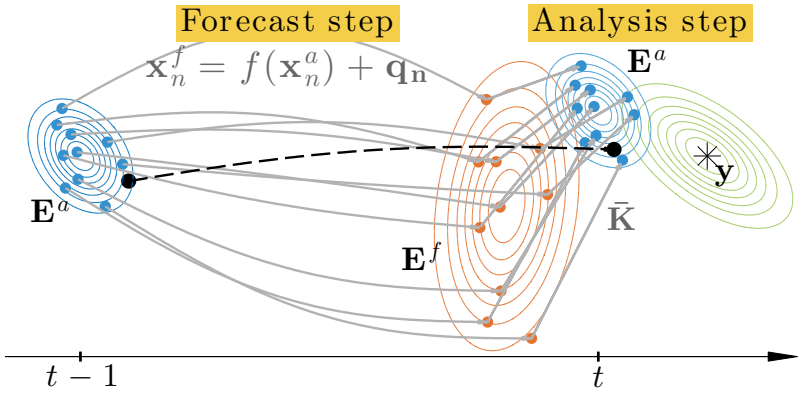


Figure 2.1: Illustration of one assimilation cycle of the EnKF. The ensemble members are shown propagated by the dynamics, eqn. (2.13), and then updated by eqn. (2.16). The colouring corresponds to figure (c) on page 5. The truth (black) is assumed drawn from the same distribution as the ensemble.

were Gaussian, h linear, and if $\bar{\mathbf{K}}$ (detailed below) were the actual Kalman gain, $\mathbf{K}$ of eqn. (2.7), then the columns of $\mathbf{E}^a$ would provide an updated ensemble, $\boldsymbol{x}_{1:N}^a$, sampled iid from the analysis distribution, $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$. This is demonstrated along the lines of

section 2.4.1 by showing that the analysis ensemble members maintain Gaussianity and that $\mathbb{E}(\boldsymbol{x}_n^a) = \boldsymbol{x}^a$ and $\mathbb{V}\mathrm{ar}(\boldsymbol{x}_n^a) = \mathbf{P}^a$ [150]. In any case, $\bar{\mathbf{K}}$ is not exact, but rather the ensemble estimate obtained by substituting $\bar{\mathbf{P}}^f$ for $\mathbf{P}^f$ in eqn. (2.7), and therefore the analysis ensemble is only approximately drawn from $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$. If $\mathbf{H}$ is defined, then the ensemble-estimated Kalman gain is obtained by substituting $\bar{\mathbf{P}}^f$ for $\mathbf{P}^f$ in eqn. (2.7)

$$\bar{\mathbf{K}} = \bar{\mathbf{P}}^f\mathbf{H}^\mathsf{T}(\mathbf{H}\bar{\mathbf{P}}^f\mathbf{H}^\mathsf{T} + \mathbf{R})^{-1}\,. \tag{2.17}$$

However, $\bar{\mathbf{K}}$ can also be estimated even if h is nonlinear, and $\mathbf{H}$ undefined; instead of substituting for $\mathbf{P}^f$, the ensemble is used to individually estimate (i) $\mathbf{P}^f\mathbf{H}^\mathsf{T}$: the cross-covariance between $\boldsymbol{x}^f$ and $h(\boldsymbol{x}^f)$, and (ii) $\mathbf{H}\mathbf{P}^f\mathbf{H}^\mathsf{T}$: the covariance matrix of $h(\boldsymbol{x}^f)$. Using the canonical estimators, as with $\mathbf{P}^f$ of eqn. (2.9), yields

$$\bar{\mathbf{K}} = \mathbf{A}\mathbf{Y}^\mathsf{T}(\mathbf{Y}\mathbf{Y}^\mathsf{T} + (N-1)\mathbf{R})^{-1}\,, \tag{2.18}$$

where $\mathbf{Y} = h(\mathbf{E}^f)\left(\mathbf{I}_N - \mathbb{1}\mathbb{1}^\mathsf{T}/N\right) \in \mathbb{R}^{p\times N}$ are the anomalies of the *observed* ensemble. When eqn. (2.17) is defined, eqn. (2.18) agrees with it.

Note that $\bar{\mathbf{P}}^f$ is not explicitly computed by eqn. (2.18). Thus, even if $\mathbf{H}$ is defined, eqn. (2.18) saves significant computation compared to eqn. (2.17) if $p < m$. As described in section 4.1, there is also an ensemble-space formulation of $\bar{\mathbf{K}}$, requiring the inversion of an $N \times N$ matrix. If $\mathbf{R}$ is readily invertible [31] this can be significantly more computationally cheap and stable, compared to formula (2.18). Unless p is significantly larger than N, the time savings might not be crucial, however, because the forecast model integration typically takes up much more time.

The connection between eqn. (2.17) and eqn. (2.18) indicates that the ensemble somehow performs statistical *linearization* of h. The connection to linear regression was recognized by Anderson [7], and its understanding is important for iterative methods [44]. However, it is worth noting that the nonlinearities of h can also be concealed by an augmented forecast model, yielding a state vector that includes the observations. The resulting observation operator then becomes merely the matrix that selects the observation components of the augmented state vector. The analysis formulae of this linear, augmented system, can then be dismantled into blocks, whereupon eqn. (2.18) is recovered. This thesis is not particularly concerned with nonlinear observation operators, h. However, all of the developments employ statistical linearization, and are thus fully applicable in the nonlinear case.

2.4 Properties in the linear-Gaussian case

Given the assumptions of the EnKF, it is desirable that $\bar{\mathbf{P}}^{f/a} = \mathbf{P}^{f/a}$ and $\bar{\boldsymbol{x}}^{f/a} = \boldsymbol{x}^{f/a}$ throughout the DA process, at least in the linear-Gaussian case, where the KF solves the Bayesian filtering equations exactly. This section provides the justification for the EnKF analysis update, eqn. (2.16), by showing that it satisfies these conditions in the expected sense, where the expectation is with respect to $\mathbf{D}_{\text{obs}}$. Related concepts are also discussed. The subscript on $\mathbf{D}$ is dropped in this section.

2.4.1 Conformality and unbiasedness

By "conformality" we mean that the ensemble estimates satisfy the relations obtained by *individually* replacing, in the KF equations, the exact moments of the forecast distribution by their ensemble estimates. It is not a particularly distinguished property in itself, but is used as a lemma for showing other results such as unbiasedness.

Proposition 2.1 – Unbiasedness of the EnKF in the mean**.**
If the column-mean of $\mathbf{D}$ *is zero, i.e.*

$$\mathbf{D}\mathbb{1} = 0\,, \tag{2.19}$$

then the mean EnKF update conforms to the KF mean update. Furthermore, with expectation over $\mathbf{D}$*, the EnKF analysis update is unbiased in the mean.*

Proof. If eqn. (2.19) holds, then eqn. (2.16) yields

$$\bar{\boldsymbol{x}}^a = \frac{1}{N}\mathbf{E}^a\mathbb{1} = \frac{1}{N}\mathbf{E}^f\mathbb{1} + \frac{1}{N}\bar{\mathbf{K}}\left(\boldsymbol{y}\mathbb{1}^\mathsf{T} - \mathbf{D} - \mathbf{H}\mathbf{E}^f\right)\mathbb{1} = \bar{\boldsymbol{x}}^f + \bar{\mathbf{K}}\left[\boldsymbol{y} - \mathbf{H}\bar{\boldsymbol{x}}^f\right]\,, \tag{2.20}$$

and the mean EnKF update conforms to the KF mean update (2.5). By virtue of the linearity of eqn. (2.20) with respect to eqn. (2.19), which holds in the expected sense, the mean EnKF update is unbiased with respect to $\mathbf{D}$. □

Proposition 2.2 – Unbiasedness of the EnKF in the covariance**.**
If, in addition to the assumption of Proposition 2.1, the sample variance of $\mathbf{D}$ *is* exactly $\mathbf{R}$*, and its sample cross-covariance with the anomalies zero, i.e.*

$$\frac{1}{N-1}\mathbf{D}\mathbf{D}^\mathsf{T} = \mathbf{R}\,, \tag{2.21a}$$

$$\mathbf{A}^f\mathbf{D}^\mathsf{T} = 0\,, \tag{2.21b}$$

then the resulting covariance update of the EnKF conforms with the KF covariance update. Furthermore, with expectation over $\mathbf{D}$*, the EnKF analysis update is unbiased in the covariance.*

Proof. First, compute the updated anomalies, $\mathbf{A}^a$, by inserting eqn. (2.16) for $\mathbf{E}^a$:

$$\mathbf{A}^a = \mathbf{E}^a\Big(\mathbf{I}_N - \mathbb{1}\mathbb{1}^\mathsf{T}/N\Big) \tag{2.22}$$

$$= \mathbf{A}^f - \bar{\mathbf{K}}\left[\mathbf{D} + \mathbf{H}\mathbf{A}^f\right], \tag{2.23}$$

where eqn. (2.19) has been used. Hence, the updated ensemble covariance matrix is

$$\bar{\mathbf{P}}^a = \frac{1}{N-1}\mathbf{A}^a\mathbf{A}^{a\mathsf{T}} \tag{2.24}$$

$$= \frac{1}{N-1}\left(\mathbf{A}^f - \bar{\mathbf{K}}\left[\mathbf{D} + \mathbf{H}\mathbf{A}^f\right]\right)\left(\mathbf{A}^f - \bar{\mathbf{K}}\left[\mathbf{D} + \mathbf{H}\mathbf{A}^f\right]\right)^\mathsf{T} \tag{2.25}$$

$$= \bar{\mathbf{P}}^f + \bar{\mathbf{K}}\mathbf{H}\bar{\mathbf{P}}^f\mathbf{H}^\mathsf{T}\bar{\mathbf{K}}^\mathsf{T} - \bar{\mathbf{P}}^f\mathbf{H}^\mathsf{T}\bar{\mathbf{K}}^\mathsf{T} - \bar{\mathbf{K}}\mathbf{H}\bar{\mathbf{P}}^f \tag{2.26}$$

$$- \frac{1}{N-1}\left[\mathbf{A}^f\mathbf{D}^\mathsf{T}\bar{\mathbf{K}}^\mathsf{T} + \bar{\mathbf{K}}\mathbf{D}\mathbf{A}^{f\mathsf{T}} - \bar{\mathbf{K}}\mathbf{D}\mathbf{A}^{f\mathsf{T}}\mathbf{H}^\mathsf{T}\bar{\mathbf{K}}^\mathsf{T} - \bar{\mathbf{K}}\mathbf{H}\mathbf{A}^f\mathbf{D}^\mathsf{T}\bar{\mathbf{K}}^\mathsf{T} - \bar{\mathbf{K}}\mathbf{D}\mathbf{D}^\mathsf{T}\bar{\mathbf{K}}^\mathsf{T}\right]$$

$$= (\mathbf{I}_m - \bar{\mathbf{K}}\mathbf{H})\bar{\mathbf{P}}^f(\mathbf{I}_m - \mathbf{H}^\mathsf{T}\bar{\mathbf{K}}^\mathsf{T}) \tag{2.27}$$

$$- \frac{1}{N-1}\left[(\mathbf{I}_m - \bar{\mathbf{K}}\mathbf{H})\mathbf{A}^f\mathbf{D}^\mathsf{T}\bar{\mathbf{K}}^\mathsf{T} + \bar{\mathbf{K}}\mathbf{D}\mathbf{A}^{f\mathsf{T}}(\mathbf{I}_m - \mathbf{H}^\mathsf{T}\bar{\mathbf{K}}^\mathsf{T}) - \bar{\mathbf{K}}\mathbf{D}\mathbf{D}^\mathsf{T}\bar{\mathbf{K}}^\mathsf{T}\right],$$

where the identity $(N-1)^{-1}\mathbf{A}^f\mathbf{A}^{f\mathsf{T}} = \bar{\mathbf{P}}^f$ has been used. Substituting eqns. (2.21) into eqn. (2.26) yields

$$\bar{\mathbf{P}}^a = \bar{\mathbf{P}}^f + \bar{\mathbf{K}}\mathbf{H}\bar{\mathbf{P}}^f\mathbf{H}^\mathsf{T}\bar{\mathbf{K}}^\mathsf{T} - \bar{\mathbf{P}}^f\mathbf{H}^\mathsf{T}\bar{\mathbf{K}}^\mathsf{T} - \bar{\mathbf{K}}\mathbf{H}\bar{\mathbf{P}}^f + \bar{\mathbf{K}}\mathbf{R}\bar{\mathbf{K}}^\mathsf{T} \tag{2.28}$$

$$= (\mathbf{I}_m - \bar{\mathbf{K}}\mathbf{H})\bar{\mathbf{P}}^f + \bar{\mathbf{K}}(\mathbf{H}\bar{\mathbf{P}}^f\mathbf{H}^\mathsf{T} + \mathbf{R})\bar{\mathbf{K}}^\mathsf{T} - \bar{\mathbf{P}}^f\mathbf{H}^\mathsf{T}\bar{\mathbf{K}}^\mathsf{T} \tag{2.29}$$

$$= (\mathbf{I}_m - \bar{\mathbf{K}}\mathbf{H})\bar{\mathbf{P}}^f + \bar{\mathbf{P}}^f\mathbf{H}^\mathsf{T}\bar{\mathbf{K}}^\mathsf{T} - \bar{\mathbf{P}}^f\mathbf{H}^\mathsf{T}\bar{\mathbf{K}}^\mathsf{T} \tag{2.30}$$

$$= (\mathbf{I}_m - \bar{\mathbf{K}}\mathbf{H})\bar{\mathbf{P}}^f, \tag{2.31}$$

and thus the covariance of the EnKF update conforms to the KF covariance update (2.6). By virtue of the linearity of eqn. (2.26) with respect to eqns. (2.21), which hold in the expected sense, the covariance of the EnKF update is unbiased with respect to $\mathbf{D}$. □

For completeness it is worth mentioning that before [35, 95] the original EnKF [57] used the analysis update scheme

$$\mathbf{E}^a = \mathbf{E}^f + \bar{\mathbf{K}}\left(\boldsymbol{y}\mathbb{1}^\mathsf{T} - \mathbf{H}\mathbf{E}^f\right), \tag{2.32}$$

which lacks the perturbations, $\mathbf{D}$ of eqn. (2.16), and has the effect on eqn. (2.27) of removing its entire second line. Thus $(\mathbf{I}_m - \bar{\mathbf{K}}\mathbf{H})$ is effectively applied twice on the covariance, meaning that the reduction in the spread of the ensemble is too significant.

The Deterministic EnKF, proposed by Sakov and Oke [171], also uses no perturbations, i.e. $\mathbf{D} = 0$. However, instead of operating on the ensemble, as eqns. (2.16) and (2.32), it updates the ensemble mean, $\bar{\boldsymbol{x}}$ and anomalies, $\mathbf{A}$, separately. This allows it to shift the anomalies only halfway compared to eqn. (2.23), implying

$$\begin{aligned}
\bar{\mathbf{P}}^a &= \left(\mathbf{I}_m - \frac{1}{2}\bar{\mathbf{K}}\mathbf{H}\right)\bar{\mathbf{P}}^f\left(\mathbf{I}_m - \frac{1}{2}\mathbf{H}^\mathsf{T}\bar{\mathbf{K}}^\mathsf{T}\right) && (2.33)\\
&= \bar{\mathbf{P}}^f - \frac{1}{2}\bar{\mathbf{K}}\mathbf{H}\bar{\mathbf{P}}^f - \frac{1}{2}\bar{\mathbf{P}}^f\mathbf{H}^\mathsf{T}\bar{\mathbf{K}}^\mathsf{T} + \bar{\mathbf{K}}\mathbf{H}\bar{\mathbf{P}}^f\mathbf{H}^\mathsf{T}\bar{\mathbf{K}}^\mathsf{T} && (2.34)\\
&= (\mathbf{I}_m - \bar{\mathbf{K}}\mathbf{H})\bar{\mathbf{P}}^f + \bar{\mathbf{K}}\mathbf{H}\bar{\mathbf{P}}^f\mathbf{H}^\mathsf{T}\bar{\mathbf{K}}^\mathsf{T}, && (2.35)
\end{aligned}$$

where the symmetry of $\bar{\mathbf{K}}\mathbf{H}\bar{\mathbf{P}}^f$ can be shown by inserting eqn. (2.18). Thus, if $\bar{\mathbf{K}}\mathbf{H}$ is small (loosely speaking, if the observation uncertainty is larger than the state uncertainty) then the term that is quadratic in $\bar{\mathbf{K}}\mathbf{H}$ is dominated by the linear term, and hence $\bar{\mathbf{P}}^a$ is only slightly larger than its desired value, $(\mathbf{I}_m - \bar{\mathbf{K}}\mathbf{H})\mathbf{P}^f$.

2.4.2 Convergence

Propositions 2.1 and 2.2 and Slutsky's theorem [133] suffice to show that

$$\lim_{N\to\infty} \bar{\boldsymbol{x}}^a = \boldsymbol{x}^a, \tag{2.36a}$$

$$\lim_{N\to\infty} \bar{\mathbf{P}}^a = \mathbf{P}^a, \tag{2.36b}$$

with convergence "in probability", provided the same holds for the forecast moments. In other words, the EnKF analysis update is consistent.

At first sight, in the linear-Gaussian case, it would appear that the convergence of the EnKF moments to those of the KF in eqns. (2.36) can be inductively cascaded through time so as to conclude that the EnKF converges to the exact posterior for any time index. Note, however, that $\bar{\mathbf{K}}$ is computed from the ensemble, which is then updated by the same $\bar{\mathbf{K}}$. This destroys the linearity with respect to $\mathbf{E}^f$ of the

analysis update (2.16) (that would exist with the true $\mathbf{K}$ and a linear h), and thus the apparent Gaussianity and independence of the resulting ensemble members, $\mathbf{E}^a$. This should give pause before drawing conclusions across multiple time steps. Nevertheless, Le Gland et al. [113] showed that (i) under suitable continuity conditions on f, moment conditions on $p(\boldsymbol{x}_0)$, and with a linear observation operator, the moments of the empirical distribution of the ensemble do converge, and (ii) in the linear-Gaussian case, this limit is that of the actual posterior. The latter was also shown by Mandel et al. [133], whose proof depends on the notion of exchangeability of the ensemble members: as $N \to \infty$, the irrelevance of the ordering of the ensemble is sufficient for a particular law of large numbers to hold, which, in turn, yields convergence.

2.4.3 Sampling error and bias

The conditions of eqns. (2.19), (2.21a) and (2.21b) are never realized for finite N. The same applies for the ensemble estimates of the forecast moments, $\bar{\boldsymbol{x}}^f$ and $\bar{\mathbf{P}}^f$ (as compared to $\boldsymbol{x}^f$ and $\mathbf{P}^f$). In the EnKF literature, this discrepancy is known as sampling error. Chapters 4 to 6 are all concerned with sampling error in some way.

The sampling error in the forecast moments communicates to the moments of the ensemble updated by the analysis, eqn. (2.16). Moreover, in contrast to consistency, expectation (and hence unbiasedness) does not commute with nonlinear operations. In other words, individually inserting expected values into a formula, such as eqn. (2.17), does not yield the expected value of the formula. The nonlinearity in $\bar{\mathbf{P}}^f$ of eqn. (2.26) therefore induces Proposition 2.3. Note that it does not contradict Propositions 2.1 and 2.2: the expectation referred to by Propositions 2.1 and 2.2 is only with respect to $\mathbf{D}$, while Proposition 2.3 also averages over the forecast ensemble.

Proposition 2.3 – Bias of the EnKF update**.**
Even if the forecast moments are unbiased, i.e. $\mathbb{E}(\bar{\boldsymbol{x}}^f) = \boldsymbol{x}^f$ *and* $\mathbb{E}(\bar{\mathbf{P}}^f) = \mathbf{P}^f$*, the same does not hold for the analysis estimates,* $\bar{\boldsymbol{x}}^a$ *and* $\bar{\mathbf{P}}^a$*. More specifically,*

$$\mathbb{E}(\mathsf{tr}(\bar{\mathbf{P}}^a)) < \mathsf{tr}(\mathbf{P}^a)\,. \tag{2.37}$$

Proof. The proof is adapted from Furrer and Bengtsson [70]. It only considers a simplified, scalar case. More details on and additional results are provided therein and by [169] and [195].

Suppose $\mathbf{H} = 1$, $\mathbf{R} = 1$, and assume that $\mathbf{P}^f$ is known. Let $g(u) = u/(1+u)$, so that the true and estimated Kalman gains, eqns. (2.7) and (2.17), are $\mathbf{K} = g\Big(\mathbf{P}^f\Big)$ and $\bar{\mathbf{K}} = g\Big(\bar{\mathbf{P}}^f\Big)$. Then, by Jensen's inequality [114, Th. 7.5], $\mathbb{E}(\bar{\mathbf{K}}) < \mathbf{K}$, since g is a

concave function and since $\mathbb{E}(\bar{\mathbf{P}}^f) = \mathbf{P}^f$. Furthermore, $\mathbf{P}^a = \mathbf{K}$ (for these particular values of $\mathbf{R}$ and $\mathbf{H}$) and, as can be shown from eqn. (2.31), $\mathbb{E}(\bar{\mathbf{P}}^a) = \mathbb{E}(\bar{\mathbf{K}})$, where the expectation is *also* over $\mathbf{D}$. Hence $\mathbb{E}(\bar{\mathbf{P}}^a) = \mathbb{E}(\bar{\mathbf{K}}) < \mathbf{K} = \mathbf{P}^a$. □

Proposition 2.3 means that the sampling error in the forecast moments yields a systematic underestimation of $\bar{\mathbf{P}}^a$. This may be compensated for by multiplicative "inflation", either manually tuned, as in section 5.7, or estimated on-line, as is the topic of chapter 6. Another possibility is to split the ensemble into smaller sub-ensembles [95].

2.5 The effects of a small ensemble size

Rank deficiency refers to the fact that N might be a lot smaller than m, and that therefore the low-order "parameterization" of the pdf by the ensemble cannot represent the *directions* of uncertainty of the distribution (as measured by the second order moment, the covariance matrix), nor its growth through the models. Conceptually, at least, this should be distinguished from sampling error, which affects the *accuracy* of the pdf representations and the model linearizations. Whether due to rank deficiency or sampling error, the smaller the ensemble size, N, the worse the ensemble estimates. However, even for large geoscientific models, experiments indicate that an ensemble size in the range of 10 to 100 is often sufficient [4, 11, 61, 96, 137, 207]. How can this be? The following subsections provide a semi-heuristic explanation.

2.5.1 Rank deficiency

With respect to the forecast step of the EnKF, it may be explained by observing that the dynamics typically only have a relatively small number of fast-growing "modes" [151, 196]. The growth in uncertainty can therefore be emulated by the growth in spread of the ensemble, if this is big enough [154, 158]. The number of growing modes of the dynamics can be estimated by the number of positive Lyapunov exponents [106], which is therefore sometimes used to estimate the required size of the ensemble [80, 98, 156].

Note that the EnKF analysis step, eqn. (2.16), does not explicitly change the subspace (i.e. directions) spanned by the ensemble. Instead, it is the propagation through the dynamical forecast model that is trusted with tracking the growing modes.

As such, it is hoped that the ensemble constitutes an adaptive set of vectors that span the subspace of uncertainty (outside of which the uncertainty is close to zero), and that its left singular vectors form a "sparse" basis thereof [37].

This ability of the ensemble has enabled recommendations [42, 99] not to overly reduce the parameterization of the models.[2] This not only reduces the required preparation for DA, but may also reduce the propensity of the ensemble analysis update to produce dynamical inconsistencies [151].

With respect to the analysis step, it is helpful to think of the EnKF as trading bias for variance [188]. More specifically, the EnKF analysis step builds on the KF, which is a linear estimator. Although the inflexibility due to the linearity of the KF estimator yields a bias in the non-Gaussian setting, it also protects against large, random variations [23, §3.2]. The connection to rank deficiency is that the parameterization of the k-th order moment requires on the order of m^k degrees of freedom. So although the ensemble may not be rank deficient compared to the covariance, it still is with respect to the third or forth order moments. In this view, the EnKF's "refusal to acknowledge" its rank deficiency (by neglecting higher order moments) is what guards it from the "curse of dimensionality" that so afflicts the particle filter [196].

2.5.2 Spurious correlations and localization

The limited ensemble size also causes problems through an effect known as spurious correlations. Recall that the ensemble is subject to stochastic variation due to the initial sampling, and the sampling of $\mathbf{D}_{\text{obs}}$ and $\mathbf{D}_{\text{mod}}$. Model nonlinearity may also be included in this reckoning [30]. This means that the forecast moments are subject to stochastic error, or sampling error, as defined in section 2.4.3. The errors *off* the diagonal of $\bar{\mathbf{P}}^f$ are known as spurious correlations.

To illustrate their effect, suppose that, based on physical arguments, one knows that two state variables that are far apart[3] have zero correlation. However, the variance of the estimate of this correlation is of the order of $1/N$, i.e. not zero. This, in turn, reduces the size of $\bar{\mathbf{P}}^a$, as summarized by Proposition 2.3 (which also concerns the effects of errors *on* the diagonal of $\bar{\mathbf{P}}^f$). It was already mentioned that inflation, studied in chapter 6, may be used to compensate for the systematic underestimation of $\bar{\mathbf{P}}^a$. By contrast, "localization" attempts to prevent spurious correlations.

[2] Reduced parameterizations restricts the DA process to a select subset of the state variables and parameters. A central concern is which basis is more adept for this selection. Sometimes simplified dynamics are also used [191]. Reduced parameterizations is also sometimes used for the EnKF [65].

[3] Distance is the simplest example, but more advanced considerations may also be used.

There are two types of localization: local analysis and covariance tapering. Both techniques are outlined below, for completeness, but neither one is employed in any of the theory or experiments of this thesis.

2.5.2.1 Local analysis

The local analysis approach is to perform the EnKF analysis update separately for different regions of a distance-based partition of the state vector, omitting observations that are beyond the area of influence of the region under consideration. The omission can be effectuated smoothly by multiplying the observation error variance (i.e. $\mathbf{R}$) of distant locations by factors larger than 1. Spurious correlations are diminished by omitting observation subsets that are known to carry little information on the region under consideration.

Furthermore, unlike the global update (as discussed in section 5.3.1), a local analysis allows the analysis ensemble to consist of different linear combinations of the ensemble members in different regions. Hence the composite global analysis is not confined to the N-dimensional ensemble subspace, but is free to explore a much higher-dimensional space [98]. Thus, local analysis also constitutes a remedy to the issue of rank deficiency.

2.5.2.2 Covariance tapering

The covariance tapering localization technique consists of the Schur (element-wise) product $\bar{\mathbf{P}}^f \mapsto \rho \circ \bar{\mathbf{P}}^f$ where ρ is a distance-based $m \times m$ correlation matrix. The effect of applying ρ is to taper the off-diagonal elements of $\bar{\mathbf{P}}^f$, thus making it more banded, and increasing its rank.

Sakov and Bertino [170] showed that the effects of the two localization approaches are highly similar. However, covariance tapering has the theoretical benefit that it is similar to Tikhonov regularization [85], which in this case would consist in *adding* a banded matrix to $\bar{\mathbf{P}}^f$. Since Tikhonov regularization can be seen as a way of including prior information [66], the link to Tikhonov regularization provides some formalism for localization [30], whose justification, as given above, is otherwise rather ad hoc.

2.6 Summary and discussion

The principal reason for employing the EnKF as an approximate filtering algorithm is its capacity to deal with nonlinearity despite its low-order ensemble cloud representation of the pdfs [27]. For example, in contrast with the extended KF, the EnKF does not require pre-derived linearizations of f and h. Furthermore, if the state vector length, m, is on the order of 10^9, as it may be in DA (section 1.1), then working with the state covariance matrix requires at least 8 terabytes of memory. This is currently infeasible, ruling out naive application of the extended KF. On the other hand, as seen from eqns. (2.13), (2.16) and (2.18), the EnKF does not explicitly compute $\bar{\mathbf{P}}^{f/a}$, but works with matrices of size $m \times N$, $p \times N$, and $p \times p$. The latter can be exchanged for a matrix of size $N \times N$, as shown in chapter 4.

Another advantage of the EnKF is that it is relatively easy to understand and implement, depending only on matrix algebra libraries. Moreover, in contrast to the extended KF and variational methods such as 4D-Var,[4] the EnKF is non-invasive: the forecast and observation models, f and h, are operated as black boxes, because the ensemble provides approximations to their linearized sensitivities (section 2.3.2). Furthermore, the EnKF is of a Bayesian nature: the background covariance estimate is provided by the forecast and is thus state-dependent, and the ensemble provides multiple possible realizations rather than a single estimate. Nevertheless, in some applications, 4D-Var is still preferred to the EnKF, mainly because its iterative smoothing formulation makes it more accurate for forecast initialization. However, this formulation can also be adapted by the EnKF [26, 175], and significant efforts are currently focused on merging the variational and the ensemble approaches, giving rise to methods such as 4D-EnVar, En-4D-Var [124, 200], and the ensemble "randomized maximum likelihood" method [42].

Another advantage of the EnKF is that it is extensively, and trivially, parallelizable. The model integration for the forecast step can be carried out on individual computers for each ensemble member, with no intercommunication required before the analysis step. If local analysis localization is used, then the analysis can also be distributed to individual computers, each one performing the update for its subset/region of the state vector.

[4] Very briefly, 4D-Var [122, 123] can be described as the method of iteratively optimizing (for a single point) $p(\boldsymbol{x}_{t-L} \mid \boldsymbol{y}_{t-L:t})$, for some lag length L, using the full nonlinear models, f and h, and their gradients. The prior, $p(\boldsymbol{x}_{t-L})$, is typically a climatological average.

In summary, while it is equivalent to the KF [59] and thus also 4D-Var under idealistic assumptions [27, 67], the main advantages of the EnKF in practice are that it (i) has a natural Bayesian interpretation; (ii) uses state-dependent error estimates in its observation analysis; (iii) does not require an adjoint tangent linear model of the dynamics; (iv) requires moderate working memory storage; and (v) is highly parallelizable.

Two algorithms closely related to the EnKF are the Unscented KF [104] and the reduced-rank square root filter (RRSQRT) [86]. Both of these can be viewed as versions of the EnKF where the ensemble members are re-initialized after each analysis so as to parameterize the filtered pdf according to particular criteria. This is in contrast with the analysis update of the EnKF, where each updated ensemble member is a linear combination of the forecasted ensemble members, which is a relatively benign updating mechanism (section 5.3) compared to the re-initializations of the Unscented KF and the RRSQRT.

Attempts have been made to construct EnKF-derived filters that converge to the exact distributions also in the non-Gaussian case. One such method is the rank histogram filter [10]. It is only rigorous in one-dimensional DA problems, but attempts have been made to generalize it to higher dimensions [136]. Another possible avenue is to merge the EnKF with the particle filter in some way. For example, the moment-matching EnKF replaces the analysis ensemble mean with that from the particle filter [115, 211]; higher-order corrections are also possible; performance improvements have been shown for large N [165]. Other attempts include particle filters which use the EnKF as a proposal distribution [145, 196], as well as Gaussian mixture filters [188].

Chapter 3

Numerical twin experiments

DA methods may be benchmarked using "twin experiments": an artificial "truth" is simulated, and its trajectory is subsequently estimated by the DA methods. Figure 3.1 gives an illustration. Although tracking the truth is not the formal objective of DA, which is to compute the Bayesian posterior distribution of the truth, the idea behind such twin experiments is intuitive. Furthermore, it is possible that the two objectives can be theoretically related by the connection between the "log score" and the average Kullback-Leibler divergence [203]. More details are given in section 3.2.

In this thesis, a linear advection model, the Lorenz-63 model, and the Lorenz-96 model, described in sections 3.3 to 3.5, are used to test the performance of the different methods.

In addition to the ensemble methods, the following DA methods are sometimes provided as baselines:

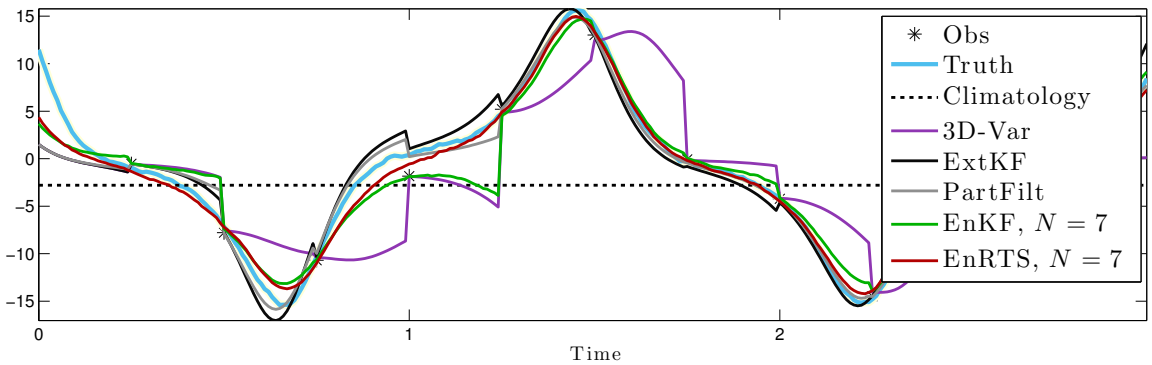


Figure 3.1: Illustration of a twin experiment, obtained with the Lorenz-63 system. Each line represents the method's mean estimate of the "x" dimension. Also included is the simulated, synthetic truth, and the noisy observations of it.

- The extended Kalman Filter (ExtKF): described in section 2.1; for linear systems, it provides the exact posterior.
- The particle filter (PartFilt): see [187, 196]; only applied for the Lorenz-63 model, where it is used with a sufficiently large number of particles ($N = 10^4$) to practically converge to the exact posterior.
- Climatology: the stationary Gaussian distribution whose mean and covariance are the averages of the dynamics over a long period of time.
- 3D-Var: like the extended KF, except that the forecast prior (i.e. $\boldsymbol{x}^f$ and $\mathbf{P}^f$) is taken to be the climatology.

However, these baselines will not draw too much of our attention; their purpose is to help identify the experiments that are relevant. For example, it is not very interesting to evaluate improvements to the EnKF under conditions (i.e. small N) where the EnKF is largely outperformed by 3D-Var. The baselines also help in appraising the significance of the differences between different versions of the EnKF.

3.1 RMSE averages

The performance of the ensemble methods is measured by their accuracy, as quantified by the root-mean-square error (RMSE) metric, between the ensemble mean and truth. For a given experiment with truth trajectory $\{\boldsymbol{x}_t \,;\, t = 0, 1, \ldots\}$, at a specific time, t, the RMSE of the ensemble mean, $\bar{\boldsymbol{x}}_t$, is

$$\text{RMSE} = \sqrt{\frac{1}{m}\|\bar{\boldsymbol{x}}_t - \boldsymbol{x}_t\|_2^2}\,. \tag{3.1}$$

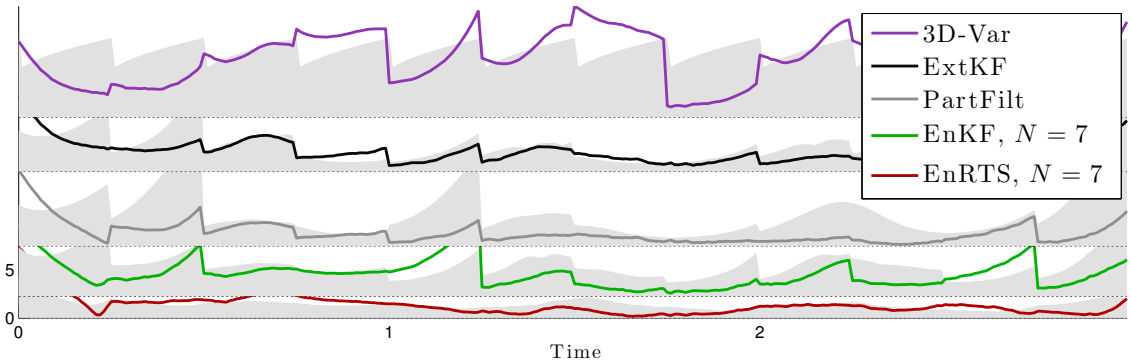


Figure 3.2: Illustration of the RMSE in a twin experiment. The vertical scale is indicated for the bottom plot only, due to the axes overlap. The shaded areas show the "root-mean variance" of the corresponding method, such as $\mathsf{tr}(\bar{\mathbf{P}}_t)/m$ for the ensemble methods.

Figure 3.2 shows the RMSE values and the root mean variance values corresponding to Fig. 3.1. Plots such as Figs. 3.1 and 3.2 can be invaluable as diagnostic tools, for debugging, and to gain understanding. However, time series plots of individual or even domain-averaged variables, statistics, or ensemble trajectories should be interpreted with caution, as they are always subject to chance occurrences. The same can be said for instantaneous snapshots of the state amplitudes/fields or scatter plots of ensembles.

The RMSE is therefore also averaged across time, with the exclusion of an initial transitory period whose duration is established beforehand by studying the time series. By convention, the RMSE is measured (only) immediately following each analysis update. This choice (as opposed to measuring *before* analysis, for example) was generally found to have little impact on the relative scores of the different EnKF methods tested. Each experiment is repeated 16 times with different initial random seeds. The empirical variance of the experiment RMSEs is computed and checked to ensure satisfying convergence.

By comparing with two of the baselines mentioned on page 27, the RMSE averages may be normalized to "skill scores" between 0 and 1 [209, §7.1.4]. We have chosen to not do so here so that the absolute RMSE values can serve as some measure of the difficulty of the DA problem of a given figure.

3.2 Considerations on the metric

The RMSE only directly evaluates the performance of the ensemble mean, $\bar{\boldsymbol{x}}$. However, the ensemble approach produces probabilistic forecast distributions, and not just point estimates. Therefore, it would be more appropriate [103] to use probabilistic scoring rules, provided they are proper.[1] One example is the log score, which evaluates the logarithm of the estimated pdf at the point of the truth. However, the computation of probabilistic scores for ensembles typically requires some form of density estima-

[1] A "proper" scoring rule cannot be "gamed": a forecaster always best served by listing his or her actual beliefs [75], rather than erring on the side of caution, optimism, or in some other way. Additionally, all proper scores can be understood as a sum of measures of uncertainty, reliability, and resolution [34]. Famous examples include the Brier score [33] and the log score [22]. Beyond propriety, Bickel [22] showed that the log score has some useful advantages compared to other proper scores. Shuford Jr. et al. [183] showed that it is the only "local" proper score. Benedetti [18] recognizes this as a fundamental scientific "likelihood" principle, and shows that the log score can be derived from this result. Benedetti [18], Weijs et al. [203] show that the Brier score is the second-order approximation to the log score.

tion, which can be complicated in high dimensions. Possible approximations include marginal totals [188] and low-order moment truncations [46].

Another set of metrics that are more suitable for probabilistic forecasts are the distance measures for distributions, including the Kolmogorov-Smirnov distance [143], and the Kullback-Leibler divergence [45]. However, these metrics require the actual posterior for reference, which is typically too costly to compute. Some medium-scale examples exist [56, 143], but are burdened by the question of their serendipity. Moreover, this metric requires density estimation across the entire state domain. Therefore, attempts have been made to derive cloud-to-cloud formulations, but this introduces additional variance [198].

By comparison to scoring rules and distribution distances, the RMSE score is very simple. Yet it can be argued that the average RMSE score should be quite reliable in DA for assessing the skill of an ensemble system in a twin experiment. This is because, by the sequentiality of DA, it indirectly assesses other aspects than just the ensemble mean. For example, while the spread of the ensemble does not directly impact the concurrent RMSE score, it will impact how the observation is weighted vis-a-vis the prior at the *next* analysis step, and hence the next RMSE score. Additionally, an ensemble that poorly represents the actual distribution will undergo different, nonlinear propagation than an accurately representative ensemble, hence also impacting the mean and the RMSE. Lastly, the RMSE is a highly standard metric in the EnKF literature, and its use therefore facilitates comparison and reproducibility.

The following sections describe the models used in the experiments.

3.3 Linear advection

Upon discretization with the first-order upwind scheme, the nondimensionalized, one-dimensional advection equation yields

$$[\boldsymbol{x}_{t+1}]_i = \Delta t\,[\boldsymbol{x}_t]_{i-1} + (1 - \Delta t\,)[\boldsymbol{x}_t]_i\,. \tag{3.2}$$

The customary [e.g. 62] model time step is used, namely $\Delta t = 1$, coinciding with the CFL limit [e.g. 192, §4]. Additionally, to counteract the growth due to the additive model noise, the state is multiplied by a dissipation factor of 0.98 after each time step. Equation (3.2) therefore becomes

$$[\boldsymbol{x}_{t+1}]_i = 0.98[\boldsymbol{x}_t]_{i-1}\,, \tag{3.3}$$

which is run for $t = 0, \ldots, 2000$, $i \in 1{:}m$, with $m = 1000$, and periodic boundary conditions. It is illustrated in Fig. 3.3. Direct observations of the truth are taken at $p = 40$ equidistant locations, with $\mathbf{R} = 0.01\mathbf{I}_p$, every fifth time step.

Similarly to [172], the initial ensemble members, $\{\boldsymbol{x}_{0,n}\ ;\ n \in 1{:}N\}$, as well as the truth, $\boldsymbol{x}_0$, are generated as a sum of 25 sinusoids of random amplitude and phase,

$$[\boldsymbol{x}_{0,n}]_i = \frac{1}{c_n} \sum_{k=1}^{25} a_n^k \sin\left(2\pi k \left(i/m + \varphi_n^k\right)\right)\,, \tag{3.4}$$

where a_n^k and φ_n^k are drawn independently and uniformly from the interval $(0, 1)$ for each n and k, and the normalization constant, c_n, is such that the standard deviation of each $\boldsymbol{x}_{0,n}$ is 1. Note that the (spatial) mean of each realization of eqn. (3.4) is zero.

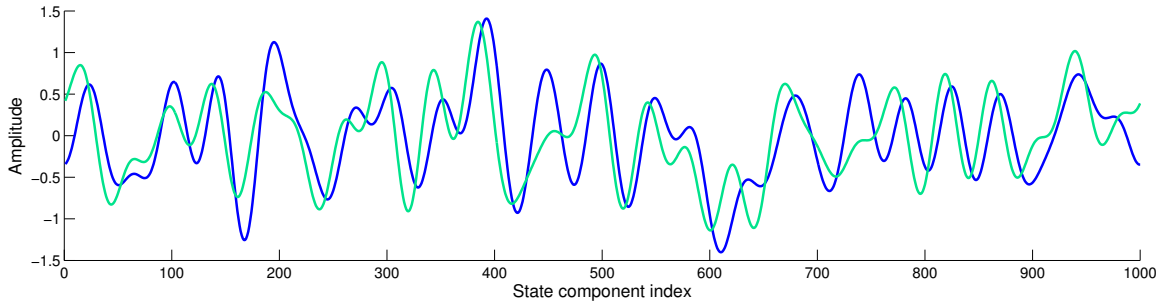


Figure 3.3: Snapshots of amplitudes during a free run of the linear advection system of eqns. (3.3) and (3.4). The first snapshot (turquoise) is taken at $t = 140$, The last (blue) snapshot is taken 10 time steps later, at $t = 150$. Because of model noise, the advection is not a pure translation.

3.4 Lorenz-63

The Lorenz-63 model [127, 176], is a three-dimensional model that has become ubiquitous in the field of nonlinear dynamics. Its equations,

$$\dot{x} = \sigma(y - x)\,, \tag{3.5}$$

$$\dot{y} = rx - y - xz\,, \tag{3.6}$$

$$\dot{z} = xy - bz\,, \tag{3.7}$$

are a leading-order approximation to Rayleigh-Bénard flow [87, §C], where an incompressible fluid of Rayleigh number r and Prandtl number σ circulates in a box of aspect ratio b due to a temperature gradient. The variables represent flow strength, x; temperature perturbation strength, y; and deviation from a nonlinear temperature profile, z. We use the common parameter settings of $r = 28$, $\sigma = 10$, and $b = 8/3$, in which case the system is nonlinear with a maximum Lyapunov exponent of 0.9 [e.g. 38]. In this case there are no stable fixed points or limit cycles; instead the system has a "strange attractor" and the solutions never repeat, but bear resemblance to a butterfly when plotted over time, as illustrated in Fig. 3.4.

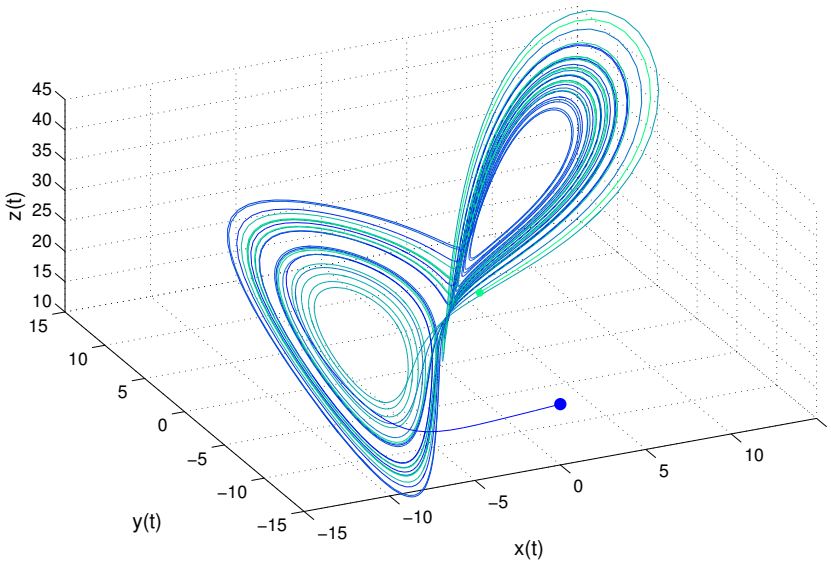


Figure 3.4: Trajectory of a free run of the Lorenz-63 system (3.5) in phase space. The trajectory traces a run from $t = 0$ to $t = 40$, with $\boldsymbol{x}_0 = [0, -15, 20]^\mathsf{T}$.

Equation (3.5) is integrated using the fourth-order Runge-Kutta numerical time stepping scheme with a time step of $\Delta t = 0.01$ for $T = 50\,000$ time steps. Unless

otherwise stated, direct observations of the entire state vector are taken $\Delta t_{\text{obs}} = 0.25$ apart, with error covariance $\mathbf{R} = 2\mathbf{I}_3$.

3.5 Lorenz-96

The Lorenz-96 model [128], given by the coupled set of ordinary differential equations,

$$\frac{d[\boldsymbol{x}]_i}{dt} = ([\boldsymbol{x}]_{i+1} - [\boldsymbol{x}]_{i-2})\,[\boldsymbol{x}]_{i-1} - [\boldsymbol{x}]_i + F\,, \tag{3.8}$$

applied for $t > 0$, and $i \in 1{:}m$, with $m = 40$ and periodic boundary conditions. It is a nonlinear, chaotic model that mimics the atmosphere at a certain latitude circle. Unlike its spiritual predecessor, the Lorenz-63 model, it was not derived by truncating a more comprehensive set of meteorological equations. It was *designed* [129] as a simplistic, symmetric system where the variables can be thought of as a some equidistant discretization of a scalar, meteorological quantity such as temperature or vorticity, and (i) the nonlinear terms, intended to simulate advection, are quadratic and together conserve $\|\boldsymbol{x}\|_2^2$, the total energy; (ii) the linear terms, representing mechanical or thermal dissipation, decrease the total energy; and (iii) the constant terms, representing external forcing, prevent the total energy from decaying to zero.

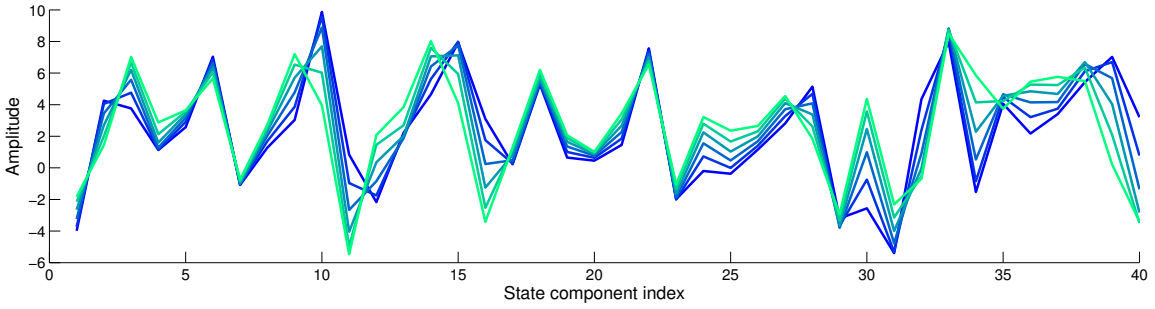


Figure 3.5: Snapshots of amplitudes during a free run of the Lorenz-96 system (3.8). Although the initial sample is sampled from $\mathcal{N}(0, \mathbf{I}_{40})$, the first snapshot (dark blue) is taken at $t = 10$, so that the transient effects have had time to expire. The last (turquoise) snapshot is at $t = 10.20$, corresponding to the analogous atmospheric time span of 1 day. Although wavenumber 8 appears prominently in the spectrum, the individual "highs" and "lows" alter their shapes and intensities rather irregularly, as they progress slowly and not invariably westward (towards lower dimension indices).

With $m = 40$, the steady-state solution, $[\boldsymbol{x}]_1 = \ldots = [\boldsymbol{x}]_m = F$, is stable for $F < 0.895$ [130]. For $0.895 < F < 4$ the system instead gradually settles into a wavenumber 8 periodic solution travelling in the direction of decreasing i. For $F > 4$

the system is chaotic, although wavenumber 8 still dominates its spectrum (Fig. 3.5). In this case the time-average mean is 2.3 and the standard deviation 3.6 (for all of the state variables). Furthermore, the system has 13 positive Lyapunov exponents, which Bocquet and Sakov [26] connected to the minimum useful ensemble size for the EnKF without localization, and the Kaplan-Yorke dimension of its fractal attractor is 27.05. The leading Lyapunov exponent corresponds to a doubling time of 0.42 time units; equating this to the doubling time of global atmospheric circulation models, estimated at 2.1 days by Lorenz [126], a unit of non-dimensional time can be identified with 5 days. Using this identification means that the leading exponent is 0.336 $(\text{days})^{-1}$ [39].

Equation (3.8) is integrated using the fourth-order Runge-Kutta scheme with a time step of $\Delta t = 0.05$ for $T = 10\,000$ time steps. Unless otherwise stated, direct observations of the entire state vector are taken $\Delta t_{\text{obs}} = 0.05$ apart (representative of 6 hours) with error covariance $\mathbf{R} = \mathbf{I}_{40}$.

Chapter 4

The square root method in the analysis step

Square root KFs [159] primarily improve on the numerical stability of the KF [105]. They inherently preserve the symmetry and non-negativity of the state covariance matrix, and the condition number of the matrices involved is approximately the square root of that of the standard KF [e.g. 5, 118].

Introduced to the analysis step of the EnKF by references [6, 24, 204], the square root analysis update was soon recognized and connected to the original square root KFs [190]. However, for the EnKF, the main purpose is not numerical stability, but rather to do away with the stochasticity and the accompanying sampling errors (non-conformality) of the perturbed-observation analysis update of the original EnKF.

Recall, in particular, that $\mathbf{D}_{\text{obs}} \in \mathbb{R}^{p \times N}$ is a matrix whose columns are drawn independently from $\mathcal{N}(0, \mathbf{R})$. Unfortunately, as was shown in section 2.4.1, the perturbed-observation analysis update of section 2.3.2,

$$\mathbf{E}^a = \mathbf{E}^f + \bar{\mathbf{K}} \left\{ \boldsymbol{y} \mathbb{1}^\mathsf{T} - \mathbf{D}_{\text{obs}} - h(\mathbf{E}^f) \right\} , \tag{4.1}$$

only yields the intended, conformal covariance, eqn. (2.31), on average:[1][2]

$$\mathbb{E}(\bar{\mathbf{P}}^a) = [\mathbf{I}_m - \bar{\mathbf{K}}\mathbf{H}]\bar{\mathbf{P}}^f . \tag{4.2}$$

[1] In accordance with Propositions 2.1 and 2.2, the expectation, $\mathbb{E}$, is taken with respect to $\mathbf{D}_{\text{obs}}$, but not with respect to the forecast ensemble, which is considered fixed in this chapter.

[2] Nonlinearity is a separate issue, and not the topic of this chapter. If h is nonlinear, this discussion and eqn. (4.2) should be interpreted either with $\mathbf{H}$ as in the state-observation augmentation trick, or by replacing $\mathbf{H}\bar{\mathbf{P}}^f$ by $\frac{1}{N-1}\mathbf{Y}\mathbf{A}^{f\mathsf{T}}$, as outlined in section 2.3.2.

4.1 Method

By contrast, the square root analysis update, described immediately below, satisfies eqn. (2.31) exactly. This is possible because $\mathbf{A}^f$ factorizes out from the right hand side of eqn. (2.31). Define the notation $\overline{h(\mathbf{E}^f)} = \frac{1}{N} h(\mathbf{E}^f)\mathbb{1}$, and let

$$\bar{\boldsymbol{\delta}} = \boldsymbol{y} - \overline{h(\mathbf{E}^f)} \qquad \in \mathbb{R}^p \,, \tag{4.3}$$

$$\mathbf{Y} = h(\mathbf{E}^f)\Big(\mathbf{I}_N - \mathbb{1}\mathbb{1}^\mathsf{T}/N\Big) \in \mathbb{R}^{p\times N} \,, \tag{4.4}$$

be the mean "innovation" and the anomalies of the *observed* ensemble, respectively. Using eqn. (2.11), it may be shown that eqns. (2.20) and (2.31) are satisfied if

$$\bar{\boldsymbol{x}}^a = \bar{\boldsymbol{x}}^f + \mathbf{A}^f\mathbf{Y}^\mathsf{T}\Big(\mathbf{Y}\mathbf{Y}^\mathsf{T} + (N{-}1)\mathbf{R}\Big)^{-1}\bar{\boldsymbol{\delta}} \,, \tag{4.5}$$

$$\mathbf{A}^a\mathbf{A}^{a\mathsf{T}} = \mathbf{A}^f\mathbf{G}^a{\mathbf{A}^f}^\mathsf{T} \,, \tag{4.6}$$

$$\text{where} \quad \mathbf{G}^a = \mathbf{I}_N - \mathbf{Y}^\mathsf{T}\Big(\mathbf{Y}\mathbf{Y}^\mathsf{T} + (N{-}1)\mathbf{R}\Big)^{-1}\mathbf{Y} \,. \tag{4.7}$$

Corollaries A.1 and A.2 may be used to rewrite eqns. (4.5) and (4.7) as

$$\mathbf{G}^a = \Big(\tfrac{1}{N-1}\mathbf{Y}^\mathsf{T}\mathbf{R}^{-1}\mathbf{Y} + \mathbf{I}_N\Big)^{-1} \,, \tag{4.8}$$

$$\bar{\boldsymbol{x}}^a = \bar{\boldsymbol{x}}^f + \tfrac{1}{N-1}\mathbf{A}^f\mathbf{G}^a\mathbf{Y}^\mathsf{T}\mathbf{R}^{-1}\bar{\boldsymbol{\delta}} \,, \tag{4.9}$$

the calculation of which is computationally cheaper if $N < p$ and $\mathbf{R}$ is readily inverted.

Therefore, if $\mathbf{A}^a$ is computed by

$$\mathbf{A}^a = \mathbf{A}^f\mathbf{T}^a \,, \tag{4.10}$$

with $\mathbf{T}^a$ being a matrix square root of $\mathbf{G}^a$, then $\mathbf{A}^a$ satisfies eqn. (2.31) exactly. Note, however, as indicated by eqn. (4.2), that the square root update only addresses the issue of sampling error in $\mathbf{D}_{\text{obs}}$. That is, the EnKF with the square root analysis method is still afflicted by the issue of Proposition 2.3.

The ensemble is obtained by recombining the anomalies and the mean:

$$\mathbf{E}^a = \bar{\boldsymbol{x}}^a\mathbb{1}^\mathsf{T} + \mathbf{A}^a \,. \tag{4.11}$$

The term "square root update" is henceforth used to refer to any update of the anomalies through the right-multiplication of a transform matrix, as in eqn. (4.10).

4.2 The symmetric square root

Equation (4.8) implies that $\mathbf{G}^a$ is symmetric, positive-definite (SPD). The matrix $\mathbf{T}^a$ is a square root of $\mathbf{G}^a$ if it satisfies

$$\mathbf{G}^a = \mathbf{T}^a \mathbf{T}^{a\mathsf{T}} . \tag{4.12}$$

However, by substitution into eqn. (4.12) it is clear that $\mathbf{T}^a\mathbf{\Omega}$ is also a square root of $\mathbf{G}^a$, for any orthogonal matrix $\mathbf{\Omega}$. There are therefore infinitely many square roots. Nevertheless, some have properties that make them unique. For example, the Cholesky factor is unique as the only triangular square root with positive diagonal entries.

Here, however, the square root of most interest is the symmetric one, $\mathbf{T}^a_{\text{sym}} = \mathbf{V}\mathbf{\Sigma}^{1/2}\mathbf{V}^\mathsf{T}$, where $\mathbf{V}\mathbf{\Sigma}\mathbf{V}^\mathsf{T} = \mathbf{G}^a$ is the eigendecomposition of $\mathbf{G}^a$, and $\mathbf{\Sigma}^{1/2}$ is defined as the entry-wise *positive* square root of $\mathbf{\Sigma}$ [91, Th. 7.2.6]. Its existence follows from the spectral theorem, and its uniqueness from that of the eigendecomposition. Note its distinction by the "sym" subscript.

It was gradually discovered that the symmetric square root choice has several advantageous properties for its use in eqn. (4.10), one of which is that it does not affect the ensemble mean [e.g. 59, 199], which is updated by eqn. (4.5) apart from the anomalies. Further advantages are surveyed in section 5.3, providing strong justification for choosing the symmetric square root, and strong motivation to extend the square root approach to the forecast step.

4.3 Efficient computation

It is necessary to compute $\mathbf{G}^a$ for the mean update, eqn. (4.9), and $\mathbf{T}^a_{\text{sym}}$, for the analysis update, eqn. (4.10). Fortunately, $\mathbf{T}^a_{\text{sym}}$ comes at practically no additional cost: with $\mathbf{V} \in \mathbb{R}^{N^2}$ as the *right* singular vectors of $((N-1)\mathbf{R})^{-1/2}\mathbf{Y}$, and $\mathbf{\Sigma} \in \mathbb{R}^{p\times N}$ as the diagonal matrix of singular values,

$$\mathbf{G}^a = \mathbf{V}(\mathbf{I}_N + \mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma})^{-1}\mathbf{V}^\mathsf{T} , \tag{4.13}$$

$$\mathbf{T}^a_{\text{sym}} = \mathbf{V}(\mathbf{I}_N + \mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma})^{-1/2}\mathbf{V}^\mathsf{T} . \tag{4.14}$$

Evidently, $\mathbf{V}$ is also the matrix of eigenvectors of $\mathbf{G}^a$ and $\mathbf{T}^a_{\text{sym}}$, while $(\mathbf{I}_N + \boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma})$ contains the eigenvalues. However, as long as only the reduced SVD (Definition E.2) is computed, it does not matter much which of these paths is used to obtain $\mathbf{V}$.

4.4 Random rotations

Although Sakov and Oke [172] found the symmetric square root to be superior to other square root choices, it was also noted that for sufficiently large ensemble sizes, N, the RMSE performance may actually deteriorate with increasing N. This counter-intuitive phenomenon was attributed to a tendency of the symmetric square root choice to build up "outlier" ensemble members. In response, it was found that it is sometimes beneficial to "scramble" the ensemble from time to time by drawing a random $\boldsymbol{\Omega}$ and right-multiplying the anomalies by it after the update of eqn. (4.10). As in section 4.2, $\boldsymbol{\Omega}$ should be orthogonal, and (in order to preserve the mean) should have $\mathbb{1}$ as an eigenvector. As noted by Pham [158], $\boldsymbol{\Omega}$ can be generated by sampling an $(N-1)\times(N-1)$ noise sample, where each is element is drawn independently from $\mathcal{N}(0,1)$, and then computing how the sample is rotated compared to some reference.[3]

The deterioration and its remedy are still somewhat mysterious, nevertheless, because it remains unexplained why the build-up of outliers occurs. A possible explanation is proposed in the following. As detailed in section 5.3.3, the symmetric choice has the property that it minimizes the "transport" of the ensemble through a square root update. This is not only beneficial for the dynamical consistency of the ensemble, but also for its statistical properties, because it preserves higher order moments of the prior ensemble beyond its covariance [59]. However, there is such a thing as too much preservation: after all, the likelihood also contains higher order moment information (typically they are all assumed zero, by Gaussianity). The outliers may be a reflection of this over-preservation.

[3]For example, the angles may be defined by how the singular vectors of the sample compare to the coordinate vectors [62]. However, it is less costly to use the orthogonal factor from the QR decomposition. Moreover, if such scrambling is used, then it is of no consequence which square root choice is used for $\mathbf{T}^a$ [69, Lemma 4.5.5]. In terms of "flops" and standard matrix decomposition algorithms, it is then computationally cheaper [by a factor of 4, 194] to use the Cholesky factor rather than the symmetric square root.

Chapter 5

Extending the square root method to account for additive forecast noise in ensemble methods

This chapter considers a square root approach for the problem of accounting for model noise in the forecast step of the ensemble Kalman filter (EnKF) and related algorithms. The primary aim is to replace the method of simulated, pseudo-random, additive noise of eqn. (2.12) so as to eliminate the associated sampling errors. The core method is based on the analysis step of ensemble square root filters, and consists in the deterministic computation of a transform matrix. The theoretical advantages regarding dynamical consistency are surveyed, applying equally well to the square root method in the analysis step. A fundamental problem due to the limited size of the ensemble subspace is discussed, and novel solutions that complement the core method are suggested and studied. Benchmarks from twin experiments with simple, low-order dynamics indicate improved performance over standard approaches.

5.1 Introduction

Recall that the recursive filtering process is usually broken down into two steps: the forecast step, whose output is denoted by the superscript f, and the analysis step, whose output is denoted using the superscript a. Furthermore, $\mathbf{A}$ (without any superscript) is henceforth used to refer to the anomalies at an intermediate stage in the forecast step, before model noise incorporation. In summary, the superscript

usage of the EnKF cycle is illustrated by

$$\overbrace{\mathbf{A}^a \xrightarrow[\text{eqn. (5.6)}]{\text{Model integration,}} \mathbf{A} \xrightarrow[\text{incorporation}]{\text{Model noise}} \underbrace{\mathbf{A}^f \xrightarrow[\text{eqns. (4.5) and (4.10)}]{\text{Analysis}} \mathbf{A}^a}_{\text{Analysis step}}}^{\text{Forecast step}} \tag{5.1}$$

Although the first $\mathbf{A}^a$ of the diagram is associated with the time step before that of $\mathbf{A}$, $\mathbf{A}^f$, and the latter $\mathbf{A}^a$, this ambiguity becomes moot by focusing on the analysis step and the forecast step separately.

In this chapter we focus on the process in the middle of the diagram above. That is, we consider the question of how to best incorporate the uncertainty due to the model noise, $\boldsymbol{q}_t$, in the forecast step of the EnKF; assuming that we have a sample of N realizations of the random variable $f(\boldsymbol{x}_t)$, and that we know the statistics of $\boldsymbol{q}_t$, we ask how to best transform, or update, the realizations so that they represent a sample from $f(\boldsymbol{x}_t) + \boldsymbol{q}_t$. In particular, we derive a deterministic square root method for this purpose which is motivated by the square root method for the analysis step of the EnKF.

5.1.1 Relevance and scope

While uncertainty quantification is an important end product of any estimation procedure, it is paramount in DA due to the sequentiality and the need to correctly weight the observations at the next time step. The two main sources of uncertainty in a forecast are the initial conditions and model error [186]. Accounting for model error is therefore essential in DA.

Model error, the discrepancy between nature and computational model, can be due to incomplete understanding, linearization, truncation, sub-grid-scale processes, and numerical imprecision [120, 147]. For the purposes of DA, however, model error is frequently described as a stochastic, additive, stationary, zero-centred, spatially correlated, Gaussian white noise process. This is highly unrealistic, yet defensible in view of the multitude of unknown error sources, the central limit theorem, and tractability [102, §3.8]. Another issue is that the size and complexity of geoscientific models makes it infeasible to estimate the model error statistics to a high degree of detail and accuracy, necessitating further reduction of its parameterizations [50].

The model error in this study adheres to all of the above assumptions. This, however, renders it indistinguishable from a noise process, even from the point of view taken in twin experiments. Thus, this study effectively also pertains to natural noises

not generally classified as model error, such as inherent stochasticity (e.g. quantum mechanics) and stochastic, external forcings (e.g. cosmic microwave radiation). Therefore, while model error remains the primary motivation, model *noise* is henceforth the designation most used. It is left to future studies to recuperate more generality by scaling back on the assumptions.

Several studies in the literature are concerned with the estimation of model error, as well as its treatment in a DA scheme [49, 138, 214]. The scope of this study is more restricted, addressing the treatment only. To that end, it is functional to assume that the noise statistics, namely the mean and covariance, are perfectly known. This unrealistic assumption is therefore made, allowing us to focus solely on the problem of incorporating or *accounting for* model noise in the EnKF.

5.1.2 Model noise treatment in the EnKF

From its inception, the EnKF has explicitly considered model noise and accounted for it in a Monte Carlo way: adding simulated, pseudo-random noise to the state realizations [57]. A popular alternative technique is multiplicative inflation, where the spread of the ensemble is increased by some "inflation factor". Several comparisons of these techniques exist in the literature [e.g. 52, 82, 207].

Quite frequently, however, model noise is not explicitly accounted for, but treated simultaneously with other system errors, notably sampling error and errors in the specification of the noise statistics [9, 96, 97, 206]. This is because (i) inflation can be used to compensate for these system errors too, and (ii) tuning separate inflation factors seems wasteful or too difficult. Nevertheless, even in realistic settings, it can be rewarding to treat model error explicitly. For example, Whitaker and Hamill [205] show evidence that, in the presence of multiple sources of error, a tuned combination of a multiplicative technique and additive noise is superior to either technique used alone.

Section 5.4 discusses the EnKF model noise incorporation techniques most relevant to this study. However, the scope of this study is not to provide a full comparison of all of the alternatives under all relevant circumstances, but to focus on the square root approach. Techniques not considered any further here include using more complicated stochastic parameterizations [15, 20], physics-based forcings such as stochastic kinetic energy backscatter [184], relaxation [213], and boundary condition forcings.

5.1.3 Outline

The proposed methods to account for model noise builds on the square root method of the analysis step described in chapter 4. The core of the proposed methods is set forth in section 5.2. Properties of both methods are analysed in section 5.3. Other techniques, against which the proposed method is compared, are outlined in section 5.4. Based on these alternatives, section 5.6 introduces methods to account for the residual noise resulting from the core method. It therefore connects to, and completes, section 5.2. The setup and results from numerical experiments are given in section 5.7. A summary is provided, along with final discussions, in section 5.8. Appendix B provides additional details on the properties of the proposed square root methods.

5.2 The square root method in the forecast step

Chapter 4 reviewed the square root update method for the analysis step of the EnKF. In view of its improvements over the Monte Carlo method, it is expected that a similar scheme for incorporating the model noise into the forecast ensemble, $\mathbf{E}^f$, would be beneficial. Section 5.2.2 derives such a scheme: Sqrt-Core. First, however, section 5.2.1 illuminates the motivation: model noise sampling error.

5.2.1 Forecast sampling errors in the standard EnKF

Assume linear dynamics, $f : \boldsymbol{x} \mapsto f(\boldsymbol{x}) = \mathbf{F}\boldsymbol{x}$, for ease of exposition. The Monte Carlo simulation (2.13) of eqn. (1.1) can then be written

$$\mathbf{E}^f = \mathbf{F}\mathbf{E}^a + \mathbf{D}\,, \tag{5.2}$$

where the columns of $\mathbf{D}$ are drawn from $\mathcal{N}(0, \mathbf{Q})$ by

$$\mathbf{D} = \mathbf{Q}^{1/2}\mathbf{\Xi}\,, \tag{5.3}$$

where $\mathbf{\Xi} = \left[\boldsymbol{\xi}_1, \ \ldots \ \boldsymbol{\xi}_n, \ \ldots \ \boldsymbol{\xi}_N\right]$, and each $\boldsymbol{\xi}_n$ is independently drawn from $\mathcal{N}(0, \mathbf{I}_m)$. Note that different choices of the square root, say $\mathbf{Q}^{1/2}$ and $\mathbf{Q}^{1/2}\mathbf{\Omega}$, yield equally-distributed random variables, $\mathbf{Q}^{1/2}\boldsymbol{\xi}$ and $\mathbf{Q}^{1/2}\mathbf{\Omega}\boldsymbol{\xi}$. Therefore the choice does not matter, and is left unspecified. It is typical to eliminate sampling error of the first order by centring the model noise perturbations so that $\mathbf{D}\mathbb{1} = 0$. This introduces de-

pendence between the members and reduces their variance. The latter is compensated for by rescaling by a factor of $\sqrt{N/(N-1)}$. The result is that

$$\begin{aligned}\bar{\mathbf{P}}^f &= \mathbf{F}\bar{\mathbf{P}}^a\mathbf{F}^\mathsf{T} + \mathbf{Q} \\ &+(\bar{\mathbf{Q}} - \mathbf{Q}) - \frac{1}{N-1}\left(\mathbf{F}\mathbf{A}^a\mathbf{D}^\mathsf{T} + \mathbf{D}(\mathbf{F}\mathbf{A}^a)^\mathsf{T}\right),\end{aligned} \tag{5.4}$$

as per eqn. (2.11), where $\bar{\mathbf{Q}} = (N-1)^{-1}\mathbf{D}\mathbf{D}^\mathsf{T}$. But, for the same reasons as for the analysis step, ideally:

$$\bar{\mathbf{P}}^f = \mathbf{F}\bar{\mathbf{P}}^a\mathbf{F}^\mathsf{T} + \mathbf{Q}\,. \tag{5.5}$$

Thus, the second line of eqn. (5.4) constitutes a stochastic discrepancy from the desired relations (5.5).

5.2.2 The square root method for model noise – Sqrt-Core

As illustrated in "eqn." (5.1), define $\mathbf{A}$ as the anomalies of the propagated ensemble before noise incorporation:

$$\mathbf{A} = f(\mathbf{E}^a)\left(\mathbf{I}_N - \mathbb{1}\mathbb{1}^\mathsf{T}/N\right), \tag{5.6}$$

where f is applied column-wise to $\mathbf{E}^a$. Then the desired relation (5.5) is satisfied if $\mathbf{A}^f$ satisfies:

$$\mathbf{A}^f\mathbf{A}^{f\mathsf{T}} = \mathbf{A}\mathbf{A}^\mathsf{T} + (N-1)\mathbf{Q}\,. \tag{5.7}$$

However, $\mathbf{A}^f$ can only have N columns. Thus, the problem of finding an $\mathbf{A}^f$ that satisfies eqn. (5.7) is ill-posed, since the right hand side of eqn. (5.7) is of rank m for arbitrary, full-rank $\mathbf{Q}$, while the left hand side is of rank N or less. To render the problem well-posed, the conditions of eqn. (5.7) must be reduced in some way.

Motivated by the connection between factorization and inversion (e.g. $\mathbf{Q} = \mathbf{M}[\mathbf{M}^{-1}\mathbf{Q}]$ for some invertible $\mathbf{M}$), let $\mathbf{A}^+$ be the Moore-Penrose pseudoinverse (section E.3) of $\mathbf{A}$, denote $\mathbf{\Pi}_\mathbf{A} = \mathbf{A}\mathbf{A}^+$ the orthogonal projector onto the column space of $\mathbf{A}$, and define $\hat{\mathbf{Q}} = \mathbf{\Pi}_\mathbf{A}\mathbf{Q}\mathbf{\Pi}_\mathbf{A}$ the "two-sided" projection of $\mathbf{Q}$. Note that the orthogonality of the projector, $\mathbf{\Pi}_\mathbf{A}$, induces its symmetry. Instead of eqn. (5.7), the core square root model noise incorporation method proposed here, Sqrt-Core, only

aims to satisfy

$$\mathbf{A}^f \mathbf{A}^{f\mathsf{T}} = \mathbf{A}\mathbf{A}^\mathsf{T} + (N-1)\hat{\mathbf{Q}}\,. \tag{5.8}$$

By virtue of the projection, eqn. (5.8) can be written as

$$\mathbf{G}^f = \mathbf{I}_N + (N-1)\mathbf{A}^+\mathbf{Q}(\mathbf{A}^+)^\mathsf{T}\,, \tag{5.9}$$

$$\mathbf{A}^f \mathbf{A}^{f\mathsf{T}} = \mathbf{A}\mathbf{G}^f\mathbf{A}^\mathsf{T}\,. \tag{5.10}$$

Thus, with $\mathbf{T}^f$ being a square root of $\mathbf{G}^f$, the update

$$\mathbf{A}^f = \mathbf{A}\mathbf{T}^f \tag{5.11}$$

accounts for the component of the noise quantified by $\hat{\mathbf{Q}}$. The difference between the right hand sides of eqns. (5.7) and (5.8), $(N-1)[\mathbf{Q} - \hat{\mathbf{Q}}]$, is henceforth referred to as the "residual noise" covariance matrix. Accounting for it is not trivial. This discussion is resumed in section 5.6.

As for the analysis step, we choose to use the symmetric square root, $\mathbf{T}^f_{\text{sym}}$, of $\mathbf{G}^f$. Note that *two* SVDs are required to perform this step: one to calculate $\mathbf{A}^+$, and one to calculate the symmetric square root of $\mathbf{G}^f$. Fortunately, both are relatively computationally inexpensive, needing only to calculate $N-1$ singular values and vectors. For later use, define the square root "additive equivalent":

$$\hat{\mathbf{D}} = \mathbf{A}^f - \mathbf{A} = \mathbf{A}[\mathbf{T}^f_{\text{sym}} - \mathbf{I}_N]\,. \tag{5.12}$$

5.2.3 Preservation of the mean

The square root update is a deterministic scheme that satisfies the covariance update relations exactly (in the space of $\mathbf{A}$). But in updating the anomalies, the mean should remain the same. For Sqrt-Core, this can be shown to hold true in the same way as Livings et al. [121] did for the analysis step, with the addition of Lemma 5.1.

Lemma 5.1 – A pseudoinverse identity**.**
For any real matrix $\mathbf{M}$*,*

$$\mathbf{M}^+ = \mathbf{M}^\mathsf{T}(\mathbf{M}\mathbf{M}^\mathsf{T})^+\,, \tag{5.13}$$

which can be shown algebraically [17, §1.6], or using the SVD.

Proposition 5.1 – Mean preservation**.**
If $\mathbf{A}^f = \mathbf{A}\mathbf{T}^f_{\text{sym}}$*, then*

$$\mathbf{A}^f \mathbb{1} = 0\,. \tag{5.14}$$

I.e. the symmetric square root choice for the model noise transform matrix preserves the ensemble mean.

Proof. Applying Lemma 5.1, and the definition (5.6),

$$\mathbf{G}^f \mathbb{1} = \mathbb{1} + (N-1)\mathbf{A}^+ \mathbf{Q} (\mathbf{A}\mathbf{A}^\mathsf{T})^+ \mathbf{A}\mathbb{1} = \mathbb{1}\,. \tag{5.15}$$

But the eigenvectors of the square of a diagonalizable matrix are the same as for the original matrix, with squared eigenvalues. Thus eqn. (5.15) implies $\mathbf{A}^f \mathbb{1} = \mathbf{A}\mathbf{T}^f_{\text{sym}} \mathbb{1} = \mathbf{A}\mathbb{1} = 0$. □

5.3 Dynamical consistency of square root updates

Many dynamical systems embody "balances" or constraints on the state space [196]. For reasons of complexity and efficiency these concerns are often not encoded in the prior [202]. They are therefore not considered by the statistical updates, resulting in state realizations that are inadmissible because of a lack of dynamical consistency or physical feasibility. Typical consequence of breaking such constraints include unbounded growth ("blow up"), exemplified by the quasi-geostrophic model of [171], or failure of the model to converge, after resuming simulations, as exemplified by reservoir simulators [44, 185].

This section provides a formal review of the properties of the square root update as regards dynamical consistency, presenting theoretical support for the square root method. The discussion concerns any square root update, and is therefore relevant for the square root method in the analysis step as well as for Sqrt-Core.

5.3.1 Affine subspace confinement

The fact that the square root update $\mathbf{A} \mapsto \mathbf{A}\mathbf{T}$ is a right-multiplication means that each column of the updated anomalies is a linear combination of the original anomalies. On the other hand, $\mathbf{T}$ itself depends on $\mathbf{A}$. In recognition of these two aspects, Evensen

[58] called such an update a "weakly nonlinear combination". However, it is more accurate is to say that the update is confined to the ensemble subspace: the affine space $\{\boldsymbol{x} \in \mathbb{R}^m \ : \ [\boldsymbol{x} - \bar{\boldsymbol{x}}] \in \mathsf{range}(\mathbf{A})\}$.

5.3.2 Satisfying equality constraints

It seems reasonable to assume that the updated ensemble, being in the subspace of the original one, could be approximately dynamically consistent. However, if consistency can be described by equality constraints, then discussions thereof can be made more formal and specific, as is the purpose of this subsection. In so doing, it uncovers some interesting, hitherto unnoticed advantages of the symmetric square root choice.

Suppose the original ensemble, $\boldsymbol{x}_{1:N}$, or $\mathbf{E}$, satisfies $\mathbf{C}\boldsymbol{x}_n = \boldsymbol{d}$ for all $n \in 1{:}N$, i.e.

$$\mathbf{C}\mathbf{E} = \boldsymbol{d}\mathbb{1}^\mathsf{T} . \tag{5.16}$$

One example is conservation of mass, in which case the state, $\boldsymbol{x}$, would contain grid-block densities, while the constraint coefficients, $\mathbf{C}$, would be a row vector of the corresponding volumes, and $\boldsymbol{d}$ would be the total mass. Another example is geostrophic balance [e.g. 88], in which case $\boldsymbol{x}$ would hold horizontal velocity components and sea surface heights, while $\mathbf{C}$ would concatenate the identity and a discretized horizontal differentiation operator, and $\boldsymbol{d}$ would be zero.

The constraints (5.16) should hold also after the update. Visibly, if $\boldsymbol{d}$ is zero, any right-multiplication of $\mathbf{E}$, i.e. any combination of its columns, will also satisfy the constraints. This provides formal justification for the proposition of [58], that the "linearity" of the EnKF update implicitly ensures respecting linear constraints.

One can also write

$$\mathbf{C}\bar{\boldsymbol{x}} = \boldsymbol{d} , \tag{5.17}$$

$$\mathbf{C}\mathbf{A} = 0\mathbb{1}^\mathsf{T} , \tag{5.18}$$

implying (5.16) provided $\mathbf{E} = \bar{\boldsymbol{x}}\mathbb{1}^\mathsf{T} + \mathbf{A}$ holds. Equations (5.17) and (5.18) show that the ensemble mean and anomalies can be thought of as particular and homogeneous solutions to the constraints. They also indicate that in a square root update, even if $\boldsymbol{d}$ is not zero, one only needs to ensure that the mean constraints are satisfied, because the homogeneity of eqn. (5.18) means that any right-multiplying update to $\mathbf{A}$ will satisfy the anomaly constraints. However, as mentioned above, unless it preserves

the mean, it might perturb eqn. (5.17). A corollary of Proposition 5.1 is therefore that the symmetric choice for the square root update also satisfies inhomogeneous constraints.

Finally, in the case of nonlinear constraints, e.g. $\mathscr{C}(\boldsymbol{x}_n) = \boldsymbol{d}$, truncating the Taylor expansion of $\mathscr{C}$ yields

$$\mathbf{C}\mathbf{A} \approx [\boldsymbol{d} - \mathscr{C}(\bar{\boldsymbol{x}})]\mathbb{1}^\mathsf{T}\,, \tag{5.19}$$

where $\mathbf{C} = \frac{\partial \mathscr{C}}{\partial \boldsymbol{x}}(\bar{\boldsymbol{x}})$. Contrary to eqn. (5.18), the approximate constraints of eqn. (5.19), are not homogeneous, and therefore not satisfied by any right-multiplying update. Again, however, by Proposition 5.1, the symmetric square root appears an advantageous choice, because it has $\mathbb{1}$ as an eigenvector with eigenvalue 1, and therefore satisfies the (approximate) constraints.

5.3.3 Optimality of the symmetric choice

A number of related properties on the optimality of the symmetric square root exist scattered in the literature. However, to the best of our knowledge, these have yet to be brought together in a unified discussion. Similarly, considerations on their implications on DA have so far not been collected. These are the aims of this subsection.

Theorem 5.1 – Minimal ensemble displacement**.**
Consider the ensemble anomalies $\mathbf{A}$ *with ensemble covariance matrix* $\bar{\mathbf{P}}$*, and let* $\boldsymbol{q}_n$ *be column* n *of* $\mathbf{D} = \mathbf{A}\mathbf{T} - \mathbf{A}$*: the displacement of the* n*-th anomaly through a square root update. The symmetric square root,* $\mathbf{T}_{\text{sym}}$*, minimizes*

$$J(\mathbf{T}) = \frac{1}{N-1}\sum_n \|\boldsymbol{q}_n\|^2_{\bar{\mathbf{P}}} \tag{5.20}$$

$$= \mathsf{tr}\left(\left[\mathbf{A}\mathbf{T} - \mathbf{A}\right]^\mathsf{T}\left(\mathbf{A}\mathbf{A}^\mathsf{T}\right)^{+}\left[\mathbf{A}\mathbf{T} - \mathbf{A}\right]\right) \tag{5.21}$$

among all $\mathbf{T} \in \mathbb{R}^{N^2}$ *such that* $\mathbf{A}\mathbf{T}\mathbf{T}^\mathsf{T}\mathbf{A}^\mathsf{T} = \mathbf{A}\mathbf{G}\mathbf{A}^\mathsf{T}$*, for some SPD matrix* $\mathbf{G}$*. Equation* (5.21) *coincides with eqn.* (5.20) *if* $\bar{\mathbf{P}}^{-1}$ *exists, but is also valid if not.*

Theorem 5.1 was proven by Ott et al. [155], and later restated by Hunt et al. [98] as the constrained optimum of the Frobenius norm of $[\mathbf{T} - \mathbf{I}_N]$. Another interesting and desirable property of the symmetric square root is the fact that the updated ensemble members are all equally likely realizations of the estimated posterior [135, 201]. More recently, the choice of mapping between the original and the updated ensembles has been formulated through optimal transport theory [150, 165]. However, the cost

functions therein typically use a different weighting than $J(\mathbf{T})$ for the norm, in one case yielding an optimum that is the symmetric *left*-multiplying transform matrix – not to be confused with the right-multiplying one of Theorem 5.1.

Theorem 5.1 and the related properties should benefit the performance of filters employing the square root update, whether for the analysis step, the model noise incorporation, or both. In part, this is conjectured because minimizing the displacement of an update means that the ensemble cloud should retain some of its shape, and with it higher-order, non-Gaussian information, as illustrated in Fig. 5.1.

A different set of reasons to expect strong performance from the symmetric square root choice is that it should promote dynamical consistency, particularly regarding inequality constraints, such as the inherent positivity of concentration variables, as well as nonlinear equality constraints, initially discussed in section 5.3.2. In either case it stands to reason that smaller displacements are less prone to break the constraints, and therefore that their minimization should inhibit it. Additionally, it is important when using "local analysis" localization that the ensemble is updated similarly at nearby grid points. Statistically, this is ensured by employing smoothly decaying localization functions, so that $\mathbf{G}$ does not jump too much from one grid point to the next. But, as pointed out by Hunt et al. [98], in order to translate this smoothness to dynamical consistency, it is also crucial that the square root is continuous in $\mathbf{G}$. Furthermore, even if $\mathbf{G}$ does jump from one grid point to the next, it still seems plausible that the minimization of displacement might restrain the creation of dynamical inconsistencies.

5.4 Alternative methods

This section describes the model noise incorporation methods most relevant to this study. Table 5.1 summarizes the methods that will be used in numerical comparison experiments. Add-Q is the standard method detailed in section 5.2.1. Mult-1 and Mult-m are multiplicative inflation methods. The rightmost column relates the different methods to each other by succinctly expressing the degree to which they satisfy eqn. (5.7); it can also be used as a starting point for their derivation. Note that Mult-1 only satisfies one degree of freedom of eqn. (5.7), while Mult-m satisfies m degrees, and would therefore be expected to perform better in general. It is clear that Mult-1 and Mult-m will generally not provide an exact statistical update, no matter how big N is, while Add-Q reproduces *all* of the moments almost surely as $N \to \infty$. By comparison, Sqrt-Core guarantees obtaining the correct first two moments for any $N > m$, but does not guarantee the accuracy of higher order moments.

Table 5.1: Comparison of some model noise incorporation methods.

Label	$\mathbf{A}^f =$	where	thus satisfying
ADD-Q	$\mathbf{A} + \mathbf{D}$	$\mathbf{D}$ is a centred sample from $\mathcal{N}(0, \mathbf{Q})$	$\mathbb{E}_{\mathbf{D}}$(eqn. (5.7))
MULT-1	$\lambda\mathbf{A}$	$\lambda^2 = \mathsf{tr}(\bar{\mathbf{P}})^{-1}\,\mathsf{tr}(\bar{\mathbf{P}} + \mathbf{Q})$	tr(eqn. (5.7))
MULT-m	$\mathbf{\Lambda A}$	$\mathbf{\Lambda}^2 = \mathsf{diag}(\bar{\mathbf{P}})^{-1}\,\mathsf{diag}(\bar{\mathbf{P}} + \mathbf{Q})$	diag(eqn. (5.7))
SQRT-CORE	$\mathbf{AT}$	$\mathbf{T} = \left(\mathbf{I}_N + (N-1)\mathbf{A}^{+}\mathbf{Q}\mathbf{A}^{+\mathsf{T}}\right)^{1/2}_{\mathrm{sym}}$	$\mathbf{\Pi_A}$(eqn. (5.7))$\mathbf{\Pi_A}$

Figure 5.1 illustrates the different techniques using a large ensemble size. The cloud of ADD-Q is clearly more dispersed than any of the other methods. The cloud of MULT-1 is just a scaled version of the black cloud, while that of MULT-m is stretched. The cloud of SQRT-CORE is also sheared, thus satisfying the off-diagonal covariance relations.

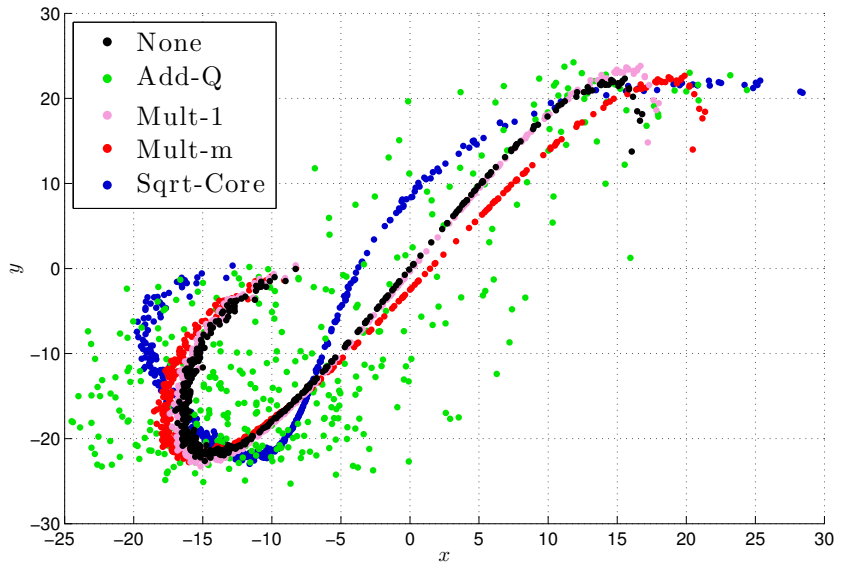


Figure 5.1: Scatter plot of ensemble forecasts with the Lorenz-63 system [127] using different schemes to account for the model noise, which is specified by $\Delta t\, Q = \mathsf{diag}([36, 3.6, 1.08])$ and makes up approximately 30% of the total spread of the updated ensembles. Each dot corresponds to the "(x, y)" coordinate of one realization among $N = 400$.

5.5 Omitted methods

Continuing the list from section 5.1.2, the following details other pertinent alternatives that are not further included in the investigation, except to the extent that they share some similarity with the square root methods proposed here.

One alternative is to resample the ensemble fully from $\mathcal{N}(0, \mathbf{A}\mathbf{A}^\mathsf{T}/(N-1) + \mathbf{Q})$. However, this incurs larger sampling errors than Add-Q, and is more liable to cause dynamical inconsistencies.

Second-order exact sampling [158] attempts to sample noise under the restriction that all of the terms on the second line of eqn. (5.4) be zero. It requires a very large ensemble size ($N > 2m$), and is therefore typically not applicable, though recent work indicate that this might be circumvented [92].

The singular evolutive interpolated Kalman (SEIK) filter [93] has a slightly less primitive and intuitive formalism than the EnKF, typically working with matrices of size $m \times (N-1)$. Moreover, it does not have a separate step to deal with model noise, treating it instead implicitly, as part of the analysis step. This lack of modularity has the drawback that the frequency of model noise incorporation is not controllable: in case of multiple model integration steps between observations, the noise should be incorporated at each step in order to evolve with the dynamics; under different circumstances, skipping the treatment of noise for a few steps can be cost efficient [64]. Nevertheless, a stand-alone model noise step can be distilled from the SEIK algorithm as a whole. Its forecast covariance matrix, $\bar{\mathbf{P}}^f$, would equal that of Sqrt-Core: $\mathbf{\Pi}_\mathbf{A}(\bar{\mathbf{P}} + \mathbf{Q})\mathbf{\Pi}_\mathbf{A}$. However, unlike Sqrt-Core, which uses the symmetric square root, the SEIK uses random rotation matrices to update the ensemble. Also, the SEIK filter uses a "forgetting factor". Among other system errors, this is intended to account for the residual noise covariance, $[\mathbf{Q} - \hat{\mathbf{Q}}]$. As outlined in section 5.1.2, however, this factor is not explicitly a function of $[\mathbf{Q} - \hat{\mathbf{Q}}]$; it is instead obtained from manual tuning. Moreover, it is only applied in the update of the ensemble mean.

Another method is to include only the $N-1$ largest eigenvalue components of $\bar{\mathbf{P}} + \mathbf{Q}$, as in reduced-rank square root filters [197], and some versions of the unscented Kalman filter [41]. This method can be referred to as T-SVD because the update can be implemented via a truncated SVD of $[\bar{\mathbf{P}}^{1/2}, \mathbf{Q}^{1/2}]$, where the choices of square roots do not matter. It captures more of the total variance than Sqrt-Core, but also changes the ensemble subspace. Moreover, it is not clear how to choose the updated ensemble. For example, one would suspect dynamical inconsistencies to arise from using the ordered sequence of the truncated SVD. Right-multiplying by random

rotation matrices, as in the SEIK, might be a good solution. Or, if computed in terms of a *left*-multiplying transform matrix, the symmetric choice is probably a good one. Building on T-SVD, the "partially orthogonal" EnKF and the COFFEE algorithm of [83, 86] also recognize the issue of the residual noise, referencing earlier reduced-rank efforts [191]. In contrasts with the additive treatments proposed in section 5.6, these methods introduce a complementary ensemble to account for it, something which seems rather inefficient.

5.6 Improving Sqrt-Core: Accounting for the residual noise

As explained in section 5.3.1, Sqrt-Core can only incorporate noise components that are in the span (range) of $\mathbf{A}$. This leaves a residual noise component unaccounted for, orthogonal to the span of $\mathbf{A}$, seemingly with $[\mathbf{Q} - \hat{\mathbf{Q}}]$ as its covariance matrix.

Section B.1.4 discusses the reason *why* this issue arises for model noise in the forecast but not for observation noise in the analysis. This section addresses the question of *how* to deal with it. It is assumed that Sqrt-Core, eqn. (5.11), has already been performed. The techniques proposed thus *complement* Sqrt-Core, but do not themselves possess the beneficial properties of Sqrt-Core discussed in section 5.3. Also, the notation of the previous section is reused. Thus, the aim of this section is to find an $\mathbf{A}^f \in \mathbb{R}^{m \times N}$ that satisfies, in some limited sense

$$\mathbf{A}^f \mathbf{A}^{f\mathsf{T}} = \mathbf{A}\mathbf{A}^\mathsf{T} + (N-1)[\mathbf{Q} - \hat{\mathbf{Q}}]\,. \tag{5.22}$$

5.6.1 Complementary, additive sampling – Sqrt-Add-Z

Let $\mathbf{Q}^{1/2}$ be any matrix square root of $\mathbf{Q}$, and define

$$\hat{\mathbf{Q}}^{1/2} = \mathbf{\Pi}_\mathbf{A} \mathbf{Q}^{1/2}\,, \tag{5.23}$$

$$\mathbf{Z} = (\mathbf{I}_m - \mathbf{\Pi}_\mathbf{A})\mathbf{Q}^{1/2}\,, \tag{5.24}$$

the orthogonal projection of $\mathbf{Q}^{1/2}$ onto the column space of $\mathbf{A}$, and the complement, respectively. A first attempt to account for the residual noise is to use one of the techniques of section 5.4, with $[\mathbf{Q} - \hat{\mathbf{Q}}]$ taking the place of the full $\mathbf{Q}$ in their formulae.

In particular, with ADD-Q in mind, the fact that

$$\mathbf{Q}^{1/2} = \hat{\mathbf{Q}}^{1/2} + \mathbf{Z} \tag{5.25}$$

motivates sampling the residual noise using $\mathbf{Z}$. That is, in addition to $\hat{\mathbf{D}}$ of SQRT-CORE, eqn. (5.12), which accounts for $\hat{\mathbf{Q}}$, one also adds $\tilde{\mathbf{D}} = \mathbf{Z}\tilde{\mathbf{\Xi}}$ to the ensemble, where the columns of $\tilde{\mathbf{\Xi}}$ are drawn independently from $\mathcal{N}(0, \mathbf{I}_m)$. We call this technique SQRT-ADD-Z.

Note that $\hat{\mathbf{Q}}^{1/2}$, defined by eqn. (5.23), is a square root of $\hat{\mathbf{Q}}$. By contrast, multiplying eqn. (5.25) with its own transpose yields

$$\mathbf{Z}\mathbf{Z}^\mathsf{T} = [\mathbf{Q} - \hat{\mathbf{Q}}] - \hat{\mathbf{Q}}^{1/2}\mathbf{Z}^\mathsf{T} - \mathbf{Z}\hat{\mathbf{Q}}^{\mathsf{T}/2}\,, \tag{5.26}$$

and reveals that $\mathbf{Z}$ is not a square root of $[\mathbf{Q} - \hat{\mathbf{Q}}]$. Therefore, with expectation over $\tilde{\mathbf{\Xi}}$, SQRT-ADD-Z does not respect $\mathbb{E}$(eqn. (5.22)), as one would hope.

Thus, SQRT-ADD-Z has a bias equal to the sum of the cross terms, $\hat{\mathbf{Q}}^{1/2}\mathbf{Z}^\mathsf{T} + \mathbf{Z}\hat{\mathbf{Q}}^{\mathsf{T}/2} = [\mathbf{Q} - \hat{\mathbf{Q}}] - \mathbf{Z}\mathbf{Z}^\mathsf{T}$. Notwithstanding this problem, Corollary B.1 shows that the sum of the cross terms has a spectrum symmetric around 0, and thus zero trace. To some extent, this exonerates SQRT-ADD-Z, since it means that the expected total variance is correct.

5.6.2 The underlying problem: replacing a single draw with two independent draws

Since any element of $\hat{\mathbf{Q}}$ is smaller than the corresponding element in $\mathbf{Q}$, either one of the multiplicative inflation techniques can be applied to account for $[\mathbf{Q} - \hat{\mathbf{Q}}]$ without second thoughts. Using MULT-1 would satisfy tr(eqn. (5.22)), while MULT-m would satisfy diag(eqn. (5.22)). However, the problem highlighted for SQRT-ADD-Z is not just a technicality. In fact, as shown in section B.1.2, $[\mathbf{Q} - \hat{\mathbf{Q}}]$ has negative eigenvalues because of the cross terms. It is therefore not a valid covariance matrix in the sense that it has no real square root: samples with covariance $[\mathbf{Q} - \hat{\mathbf{Q}}]$ will necessarily be complex numbers; this would generally be physically unrealizable and therefore inadmissible. This underlying problem seems to question the validity of the whole approach of splitting up $\mathbf{Q}$ and dealing with the parts $\hat{\mathbf{Q}}$ and $[\mathbf{Q} - \hat{\mathbf{Q}}]$ separately.

Let us emphasize the word "independently", because that is, to a first approximation, what we are attempting to do: replacing a single draw from $\mathcal{N}(0, \mathbf{Q})$ by one from $\mathcal{N}(0, \hat{\mathbf{Q}})$ plus another, independent draw from $\mathcal{N}(0, [\mathbf{Q} - \hat{\mathbf{Q}}])$. Rather than

considering N anomalies, let us now focus on a single one, and drop the n index. Define the two random variables,

$$\boldsymbol{q} = \hat{\mathbf{Q}}^{1/2}\boldsymbol{\xi} + \mathbf{Z}\boldsymbol{\xi}\,, \tag{5.27}$$

$$\boldsymbol{q}^{\perp\!\!\!\perp} = \hat{\mathbf{Q}}^{1/2}\hat{\boldsymbol{\xi}} + \mathbf{Z}\tilde{\boldsymbol{\xi}}\,, \tag{5.28}$$

where $\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}, \tilde{\boldsymbol{\xi}}$ are random variables independently drawn from $\mathcal{N}(0, \mathbf{I}_m)$. By eqn. (5.25), and design, $\boldsymbol{q}$ can be identified with any of the columns of $\mathbf{D}$ of eqn. (5.3) and, furthermore, $\mathbb{V}\mathrm{ar}(\boldsymbol{q}) = \mathbf{Q}$. On the other hand, while $\boldsymbol{q}$ originates in a single random draw, $\boldsymbol{q}^{\perp\!\!\!\perp}$ is the sum of two independent draws.

The dependence between the terms of $\boldsymbol{q}$, and the lack thereof for $\boldsymbol{q}^{\perp\!\!\!\perp}$, yields the following discrepancy between the variances:

$$\mathbb{V}\mathrm{ar}(\boldsymbol{q}) = \hat{\mathbf{Q}} + \mathbf{Z}\mathbf{Z}^{\mathsf{T}} + \hat{\mathbf{Q}}^{1/2}\mathbf{Z}^{\mathsf{T}} + \mathbf{Z}\hat{\mathbf{Q}}^{\mathsf{T}/2}\,, \tag{5.29}$$

$$\mathbb{V}\mathrm{ar}(\boldsymbol{q}^{\perp\!\!\!\perp}) = \hat{\mathbf{Q}} + \mathbf{Z}\mathbf{Z}^{\mathsf{T}}\,. \tag{5.30}$$

Formally, this is the same problem that was identified with eqn. (5.26), namely that of finding a real square root of $[\mathbf{Q} - \hat{\mathbf{Q}}]$, or eliminating the cross terms. But eqns. (5.29) and (5.30) show that the problem arises from the more primal problem of trying to emulate $\boldsymbol{q}$ by $\boldsymbol{q}^{\perp\!\!\!\perp}$. Vice versa, $\hat{\mathbf{Q}}^{1/2}\mathbf{Z}^{\mathsf{T}} = 0$ would imply that the ostensibly dependent terms, $\hat{\mathbf{Q}}^{1/2}\boldsymbol{\xi}$ and $\mathbf{Z}\boldsymbol{\xi}$, are independent, and thus $\boldsymbol{q}^{\perp\!\!\!\perp}$ is emulated by $\boldsymbol{q}$.

5.6.3 Reintroducing dependence – Sqrt-Dep

As already noted, though, making the cross terms zero is not possible for general $\mathbf{A}$ and $\mathbf{Q}$. However, the perspective of $\boldsymbol{q}$ and $\boldsymbol{q}^{\perp\!\!\!\perp}$ hints at another approach: reintroducing dependence between the draws. In this section we will reintroduce dependence by making the residual sampling depend on the square root equivalent, $\hat{\mathbf{D}}$ of eqn. (5.12).

The trouble with the cross terms is that $\mathbf{Q}$ "gets in the way" between $\boldsymbol{\Pi}_{\mathbf{A}}$ and $(\mathbf{I}_m - \boldsymbol{\Pi}_{\mathbf{A}})$, whose product would otherwise be zero. Although less ambitious than emulating $\boldsymbol{q}$ with $\boldsymbol{q}^{\perp\!\!\!\perp}$, it is possible to emulate a single draw from $\mathcal{N}(0, \mathbf{I}_m)$, e.g. $\boldsymbol{\xi}$, with two independent draws:

$$\boldsymbol{\xi}^{\perp\!\!\!\perp} = \boldsymbol{\Pi}\hat{\boldsymbol{\xi}} + (\mathbf{I}_m - \boldsymbol{\Pi})\tilde{\boldsymbol{\xi}}\,, \tag{5.31}$$

where, as before, $\hat{\boldsymbol{\xi}}$ and $\tilde{\boldsymbol{\xi}}$ are independent random variables with distribution $\mathcal{N}(0, \mathbf{I}_m)$, and $\boldsymbol{\Pi}$ is some orthogonal projection matrix. Then, as the cross terms cancel,

$$\boldsymbol{\Pi}\boldsymbol{\Pi}^{\mathsf{T}} + (\mathbf{I}_m - \boldsymbol{\Pi})(\mathbf{I}_m - \boldsymbol{\Pi})^{\mathsf{T}} = \mathbf{I}_m \,, \tag{5.32}$$

and thus $\mathbb{V}\mathrm{ar}(\boldsymbol{\xi}^{\perp\!\!\!\perp}) = \mathbb{V}\mathrm{ar}(\boldsymbol{\xi})$.

We can take advantage of this emulation possibility by choosing $\boldsymbol{\Pi}$ as the orthogonal projector onto the *rows* of $\hat{\mathbf{Q}}^{1/2}$, labelled here as $\boldsymbol{\Pi}_Q$. Instead of eqn. (5.27), redefine $\boldsymbol{q}$ as

$$\boldsymbol{q} = \mathbf{Q}^{1/2}\boldsymbol{\xi}^{\perp\!\!\!\perp} \,. \tag{5.33}$$

Then, since $\mathbb{V}\mathrm{ar}(\boldsymbol{\xi}^{\perp\!\!\!\perp}) = \mathbf{I}_m$,

$$\mathbb{V}\mathrm{ar}(\boldsymbol{q}) = \mathbf{Q}^{1/2}\mathbf{I}_m\mathbf{Q}^{\mathsf{T}/2} = \mathbf{Q} \,, \tag{5.34}$$

as desired. But also

$$\begin{aligned} \boldsymbol{q} &= (\hat{\mathbf{Q}}^{1/2} + \mathbf{Z})\left(\boldsymbol{\Pi}_Q\hat{\boldsymbol{\xi}} + (\mathbf{I}_m - \boldsymbol{\Pi}_Q)\tilde{\boldsymbol{\xi}}\right) && (5.35)\\ &= \hat{\mathbf{Q}}^{1/2}\hat{\boldsymbol{\xi}} + \mathbf{Z}\left(\boldsymbol{\Pi}_Q\hat{\boldsymbol{\xi}} + (\mathbf{I}_m - \boldsymbol{\Pi}_Q)\tilde{\boldsymbol{\xi}}\right) \,. && (5.36) \end{aligned}$$

The point is that, while maintaining $\mathbb{V}\mathrm{ar}(\boldsymbol{q}) = \mathbf{Q}$, and despite the reintroduction of dependence between the two terms in eqn. (5.36), the influence of $\tilde{\boldsymbol{\xi}}$ has been confined to $\mathsf{range}(\mathbf{Z}) = \mathsf{range}(\mathbf{A})^{\perp}$. The above reflections suggest the following algorithm, labelled Sqrt-Dep:

1. Perform the core square root update for $\hat{\mathbf{Q}}$, eqn. (5.11);
2. Find $\hat{\boldsymbol{\Xi}}$ such that $\hat{\mathbf{Q}}_{\mathrm{sym}}^{1/2}\hat{\boldsymbol{\Xi}} = \hat{\mathbf{D}}$ of eqn. (5.12). Components in the kernel of $\hat{\mathbf{Q}}_{\mathrm{sym}}^{1/2}$ are inconsequential;
3. Sample $\tilde{\boldsymbol{\Xi}}$ by drawing each column independently from $\mathcal{N}(0, \mathbf{I}_m)$;
4. Compute the residual noise, $\tilde{\mathbf{D}}$, and add it to the ensemble anomalies;

$$\tilde{\mathbf{D}} = \mathbf{Z}\left(\boldsymbol{\Pi}_Q\hat{\boldsymbol{\Xi}} + (\mathbf{I}_m - \boldsymbol{\Pi}_Q)\tilde{\boldsymbol{\Xi}}\right) \,. \tag{5.37}$$

Unfortunately, this algorithm requires the additional SVD of $\hat{\mathbf{Q}}^{1/2}$ in order to compute $\boldsymbol{\Pi}_Q$ and $\hat{\boldsymbol{\Xi}}$. Also, despite the reintroduction of dependence, Sqrt-Dep is not fully consistent, as discussed in section B.2.

5.7 Benchmark results

The methods are here benchmarked using twin experiments, as described in chapter 3.

The only difference between the ensemble DA methods is their model noise incorporation method. As indicated by eqns. (1.1) and (1.2), stochastic noise is added to the truth after each model integration and observation, respectively. Moreover, as we assumed from the beginning, the DA system has perfect knowledge of the specifications of eqns. (1.1) and (1.2), including $\mathbf{Q}$,$\mathbf{R}$,$\boldsymbol{\mu}_0$, and $\mathbf{P}_0$. The model time step, Δt, is the duration between successive time indices. Contrary to what eqn. (1.2) indicates, observations are not taken at every time index, but after a duration, Δt_{obs}, called the DA window, that is a multiple of Δt. The observation process, eqn. (1.2), is fully characterized by specifying $\mathbf{H}$, $\mathbf{R}$, and the duration between observation times, Δt_{obs}. The forecast process, eqn. (1.1), is fully characterized by specifying f and $\mathbf{Q}$. As defined in eqn. (1.1), $\mathbf{Q}$ implicitly includes a scaling of Δt.

The experiment is conducted for a range of values of a control setting, and the results are graphed. The collection of experiments for a typical figure takes around 24 hours to generate on a modern desktop computer through MATLAB, with the dynamics implemented in C. For all of the methods under comparison, the analysis step is performed using the symmetric square root update, described in chapter 4. The random rotations of section 4.4 are employed in the Lorenz-63 experiments when $N > 7$, and in the Lorenz-96 experiments when $N > 50$. Covariance localization is not used. Following each analysis update, the ensemble anomalies are rescaled by a scalar inflation factor intended to compensate for any intrinsic system errors [11, 204]. The necessity for this inflation was discussed in section 2.4.3. Prior to each experiment, this inflation factor was approximately optimally tuned, as measured by the RMSE performance. It should be noted that it is quite laborious to manually tune the inflation parameter [e.g. 9, 119]; several trials, with long time series, is required. The difficulty is increased by the fact that the optimal inflation value is found just above a threshold beyond which the performance deteriorates drastically [e.g. 172, figure 3]. For reproducibility, the values are listed in Table F.1. In this tuning process the ADD-Q method was used for the forecast noise incorporation, putting it at a slight advantage relative to the other methods.

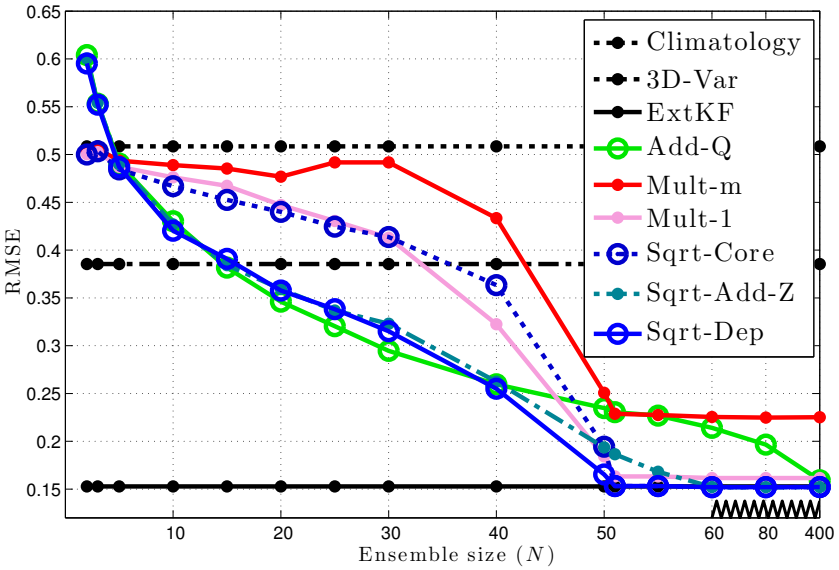


Figure 5.2: Performance benchmarks as a function of the ensemble size, N, obtained with the linear advection system. The scale has been irregularly compressed for $N > 60$.

5.7.1 Linear advection

The linear advection model was described in section 3.3. The model noise is given by

$$\mathbf{Q} = 0.01\,\mathbb{V}\mathrm{ar}(\boldsymbol{x}_0)\,, \tag{5.38}$$

where $\boldsymbol{x}_0$ is specified by eqn. (3.4), which has a maximum wavenumber of $k = 25$. Thus, by virtue of the linearity of the model, and the design of $\mathbf{P}_0$ and $\mathbf{Q}$, the dynamics take place in a subspace of rank 50, even though $m = 1000$. This is clearly reflected in the graphs of the square root methods, which all converge to the optimal performance of the Kalman filter (0.15) as N approaches 51, and $\mathbf{Z}$ goes to zero in Fig. 5.2. The curve of Sqrt-Add-Z takes a little longer to converge because of numerical error. The multiplicative inflation curves are also constant for $N \geq 51$, but they do not achieve the same level of performance. As one would expect, Add-Q converges to the Kalman filter for $N \to \infty$

Interestingly, despite Mult-m satisfying eqn. (5.7) to a higher degree than Mult-1, the latter performs distinctly better across the whole range of N. This is possibly due to the fact that Mult-m has the adverse effect of changing the subspace of the ensemble, though it is unclear why its worst performance occurs near $N = 25$.

Add-Q clearly outperforms Mult-1 in the intermediate range of N, indicating that the loss of nuance in the covariance matrices of Mult-1 is more harmful than the

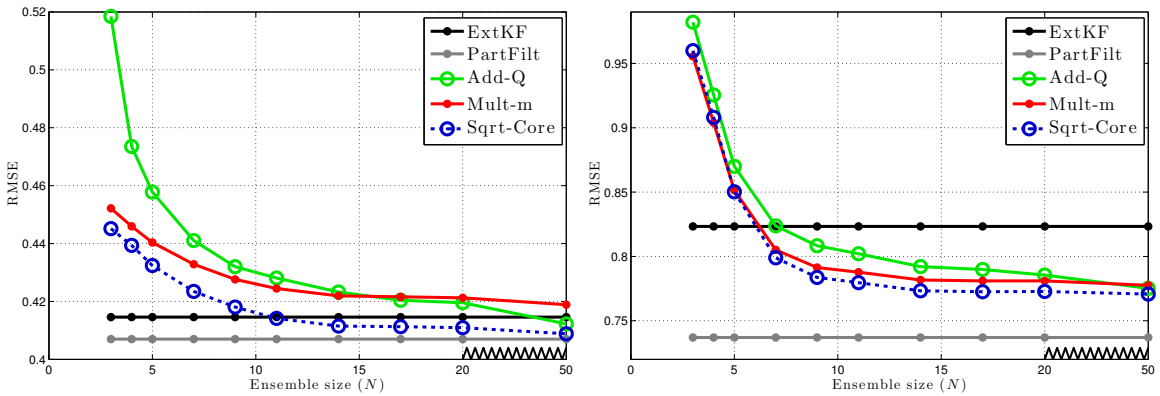
sampling error incurred by ADD-Q. But, for $45 < N < 400$, MULT-1 beats ADD-Q. It is not clear why this reversal happens.

SQRT-CORE performs quite similarly to MULT-1. In the intermediate range, it is clearly deficient compared to the square root methods that account for residual noise, illustrating the importance of doing so. The performance of SQRT-DEP is almost uniformly superior to all of the other methods. The only exception is around $N = 25$, where ADD-Q slightly outperforms it. The computationally cheaper SQRT-ADD-Z is beaten by ADD-Q for $N < 40$, but has a surprisingly robust performance nevertheless.

5.7.2 Lorenz-63

The Lorenz-63 model was described in section 3.4. The model noise is specified by

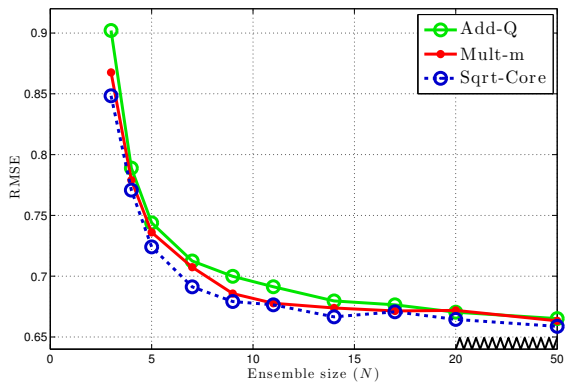
$$\mathbf{Q} = 0.1 \begin{bmatrix} 10 & -2 & 3 \\ -2 & 5 & 3 \\ 3 & 3 & 5 \end{bmatrix} . \tag{5.39}$$



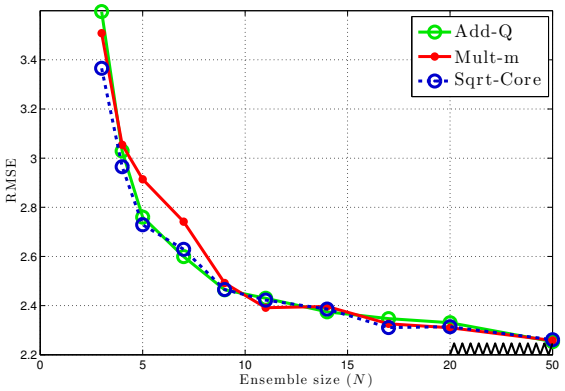
(a) All of $\boldsymbol{x} = (x, y, z)$ is observed. **(b)** Only x is observed.

Figure 5.3: Performance benchmarks as a function of the ensemble size, N, obtained with the Lorenz-63 system and $\Delta t_{\text{obs}} = 0.05$. 3D-Var averages an RMSE of 1.25 in the case of (a), and 5.0 in the case of (b). of the corresponding methods.

Figures 5.3 and 5.4 graphs RMSE versus ensemble size. Figure 5.4 is obtained with a larger Δt_{obs}, however, making the transformations of the uncertainties more nonlinear and less suitable for approximation by Gaussianity assumptions. This is reflected in the inferiority of the EnKF relative to the particle filter, which is larger for Fig. 5.4 than for Fig. 5.3. The same is observed to a larger extent for the extended KF.



(a) All of $\boldsymbol{x} = (x, y, z)$ is observed. Particle filter RMSE: 0.57. Extended Kalman filter: 1.4.



(b) Only x is observed. Particle filter RMSE: 1.23. Extended Kalman filter tends to diverge.

Figure 5.4: Same as Fig. 5.3, except that $\Delta t_{\text{obs}} = 0.25$.

Also note that restricting the observations to only the x-variable (Fig. (5.3b)) results in a stronger deterioration of the filters that rely more on Gaussianity. This is because the average state uncertainty larger, in turn possibly (i) making the uncertainty propagation more nonlinear, and (ii) decreasing the proportion of the error growth due to the additive noise, relative to the dynamics. Both effects typically decrease the Gaussianity of the distributions.

Among the different EnKFs, however, the qualitative inferences are the same in all figures: Sqrt-Core is uniformly superior to the other methods. Pointwise, it can be observed that, in Fig. (5.3b) for example, the performance of Sqrt-Core at $N = 7$ is only matched by Add-Q at $N = 17$. By contrast, Mult-m beats Add-Q for $N < 20$, but the situation is reversed for larger N as the deleterious effects of sampling error become small enough.

Each figure takes around 24 hours to generate with a 4.2 GHz CPU. Yet, as evidenced by the jitteriness in Fig. (5.4b), here the RMSE scores therein have not converged satisfactorily. In fact, a closer look at the RMSE time series, for $N = 11$ say, reveals that "filter divergence" is quite common in the case of Fig. (5.4b). Moreover, a histogram of the time series reveals that a scaling of the distribution (approximately chi-square) towards higher values is accompanied by a significant thickening of its tail. Therefore, the volatility of the RMSE time series scales more than linearly with respect to the mode value. The result it that, for some values of N in Fig. (5.4b), the 1-sigma confidence interval of the RMSE averages is of magnitude comparable to the differences between the EnKFs, making it difficult to draw any inferences from it.

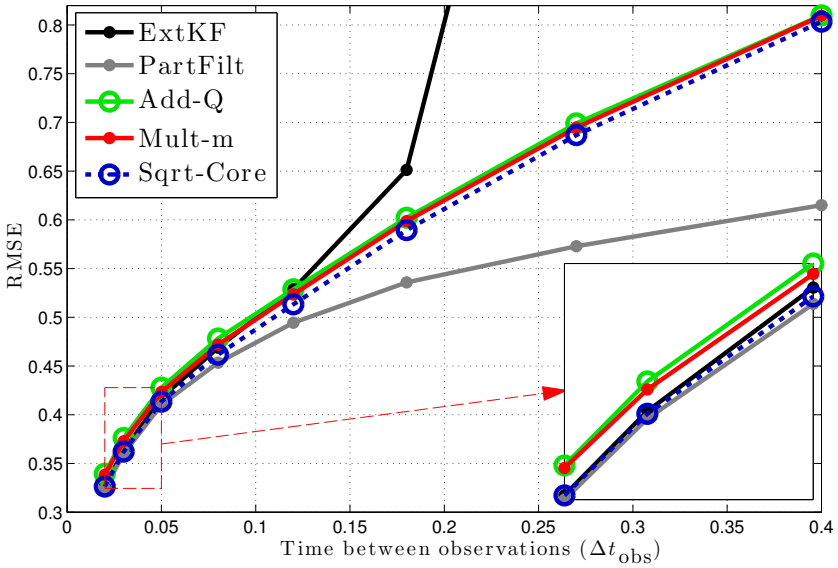


Figure 5.5: Performance benchmarks as a function of the data assimilation window, Δt_{obs}, obtained with the Lorenz-63 system and $N = 12$. The climatology and 3D-Var average the same as in Fig. 5.3.

Figure 5.5 plots the RMSE versus Δt_{obs}. Sqrt-Core averages uniformly lower RMSE values than Mult-m, which in turn averages uniformly lower values than Add-Q. Compared to the extended KF and the particle filter, however, there is little growth in the relative differences between the EnKFs.[1]

Figure 5.6 investigates the performance as a function of the strength of the noise. As one would expected, the EnKF graphs gradually converge as the noise decreases. This is not clearly legible from the figure, however, because the variance of the averages is quite large when $\mathbf{Q}$ is small. It is also notable that the EnKF is not able to capitalize on the decrease in noise as well as the particle filter.

[1]Similarly to the EnKF methods, the extended KF also requires inflation tuning. In the figures where the control variables is not N, the tuning for the extended KF should vary for the various experiments. However, unlike the EnKF methods, the tuning for the extended KF has only been undertaken with a single control variable setting.

In Fig. 5.6 this has the surprising consequence that the RMSE performance of the extended KF actually deteriorates for very small $\mathbf{Q}$. This is counter-intuitive because the entropy of the sum of two independent random variables is always larger than the entropy of either one alone [110], and in that sense, the DA problem should be easier as $\mathbf{Q}$ decreases. But the extended KF actually needs larger inflation values for very small $\mathbf{Q}$. A possible explanation is that the error due to linearization does not scale down as much as the total uncertainty for smaller $\mathbf{Q}$. Another possibility is that the actual distributions become less Gaussian and that the deleterious effects of this through the dynamics are stronger than the benefit of less noise.

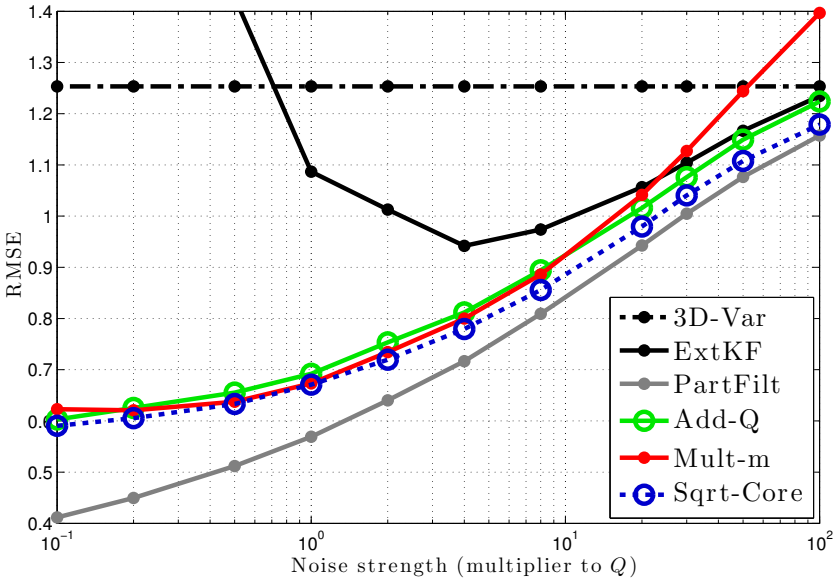


Figure 5.6: Performance benchmarks as a function of the noise strength obtained with the Lorenz-63 system and $N = 12$. The climatology and 3D-Var average about the same as in Fig. 5.3. On average, when $\mathbf{Q}$ is multiplied by 10^{-1} (resp. $10^0, 10^1$), the model noise makes up approximately 5 (resp. 30, 82) percent of the growth in the spread of the ensemble.

Mult-m mostly performs better than Add-Q, but the situation is reversed in the region where the extended KF performs almost as well as the EnKF. By contrast, Sqrt-Core performs uniformly best among the EnKFs.

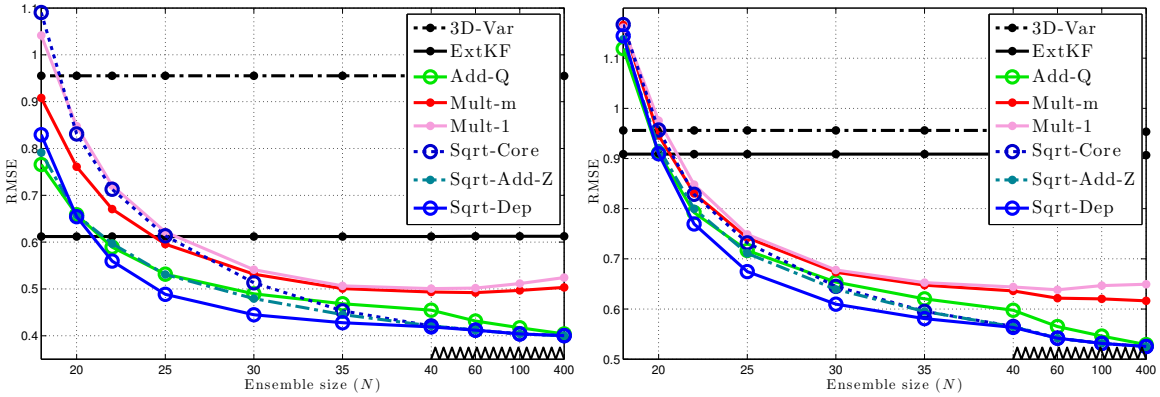
5.7.3 Lorenz-96

The Lorenz-96 model was described in section 3.5. The model noise used here is spatially homogeneous, generated using a Gaussian autocovariance function,

$$[\mathbf{Q}]_{i,j} = \mathsf{exp}\left(-1/30\|i-j\|_2^2\right) + 0.1\delta_{i,j}\,, \tag{5.40}$$

where the Kronecker delta, $\delta_{i,j}$, has been added for numerical stability. In some experiments $\mathbf{Q}$ is rescaled by a factor of 9 to increase its impact.

Figure 5.7 explores performance as a function of the ensemble size. Using a larger Δt_{obs}, the transformations of the uncertainties in the forecast and analyses involved are more nonlinear in subfigure (b) than in (a). This is reflected in the higher RMSE values of all of the graphs, and relatively worse performance of the extended KF. Even though the relative differences in performance for the EnKF noise incorporation methods are smaller in subfigure (b), the inferences drawn about their performances are the same for both figures.



(a) With $\Delta t_{\text{obs}} = 0.05$. **(b)** With $\Delta t_{\text{obs}} = 0.15$.

Figure 5.7: Performance benchmarks as a function of the ensemble size, N, obtained with the Lorenz-96 system. The climatology averages an RMSE of 3.69 for both figures.

As with the linear advection model, the curves of the square root schemes are coincident when $\mathbf{Z} = 0$, which here happens for $N > m = 40$. In contrast to the linear advection system, however, the square root methods still improve as N increases beyond m. This is because a larger ensemble is better able to characterize the non-Gaussianity of the distributions and the nonlinearity of the models. On the other hand, the performance of the multiplicative inflation methods stagnates around $N = m$, and even slightly deteriorates for larger N. This can possibly be attributed to the effects discussed in section 4.4.

Unlike the more ambiguous results of the Lorenz-63 and linear advection systems, here Add-Q uniformly beats the multiplicative inflation methods. Again, the importance of accounting for the residual noise is highlighted by the poor performance of Sqrt-Core for $N < 40$. However, even though Sqrt-Add-Z is biased, it outperforms Add-Q for $N > 25$, and approximately equals it for smaller N.

The performance of Sqrt-Dep is nearly uniformly the best, the exception being at $N = 18$, where it is marginally beaten by Add-Q and Sqrt-Add-Z. The existence of this occurrence can possibly be attributed to the slight suboptimality discussed in section B.2. as well as the advantage gained by Add-Q from using it to tune the analysis inflation. Note, though, that this region is hardly interesting, since results lie above the baseline of the extended KF.

Add-Q asymptotically attains the performance of the square root methods. In fact, though it would have been imperceptible if added to Fig. (5.7a), experiments show that Add-Q beats Sqrt-Dep by an average RMSE difference of 0.005 at $N = 800$, as predicted in section 5.4.

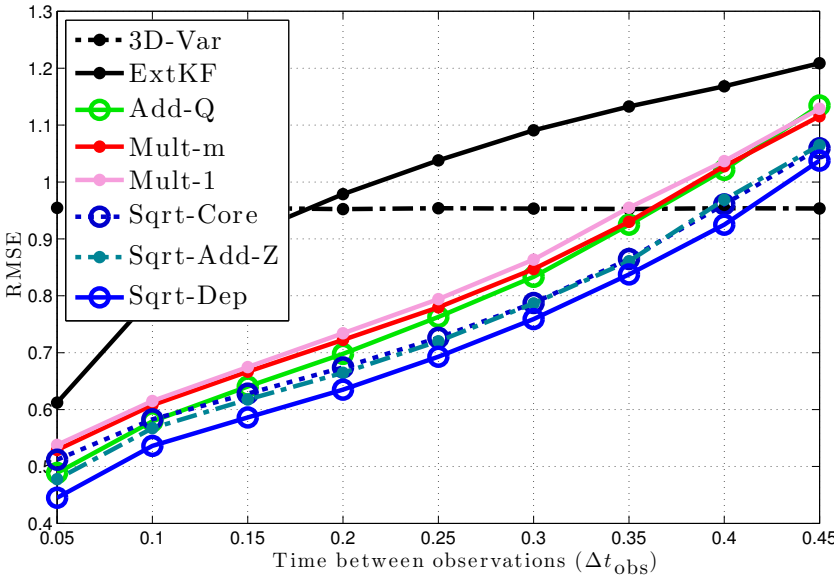


Figure 5.8: Performance benchmarks as a function of the data assimilation window, Δt_{obs}, obtained with the Lorenz-96 system and $N = 30$. The climatology averages an RMSE of 3.7.

Figure 5.8 plots the RMSE versus the DA window, Δt_{obs}. The performance of Add-Q clearly deteriorates more than that of all of the deterministic methods as Δt_{obs} increases. Indeed, the curves of Sqrt-Core and Add-Q cross at $\Delta t_{\text{obs}} \approx 0.1$, beyond which Sqrt-Core outperforms Add-Q. Sqrt-Core even gradually attains the performance of Sqrt-Add-Z, though this happens in an interval where all of the EnKF methods are beaten by 3D-Var. Again, however, Sqrt-Dep is uniformly superior, while Sqrt-Add-Z is uniformly the second best. Similar tendencies were observed in experiments (not shown) with $N = 25$.

Figure 5.9 plots the RMSE versus the amplitude of the noise. Towards the left, the curves converge to the same value as the noise approaches zero. At the higher end of the range, the curves of Mult-m and Sqrt-Core are approximately twice as steep as that of Sqrt-Dep. Again, Sqrt-Dep performs uniformly superior to the rest, with Sqrt-Add-Z performing second best. In contrasts, Add-Q performs worse than Mult-m for a noise strength multiplier smaller than 0.2, but better as the noise gets stronger.

Figure 5.10 plots the RMSE versus the forcing strength, F. The performances of the EnKFs are relatively unchanged as long as the dynamics are stable, $F < 0.895$ (section 3.5), after which they monotonically degrade. The dynamics are chaotic for $F = 4$, but the performance degradation continues also beyond this point. This can be expected, as the positive Lyapunov exponents still grow in magnitude and

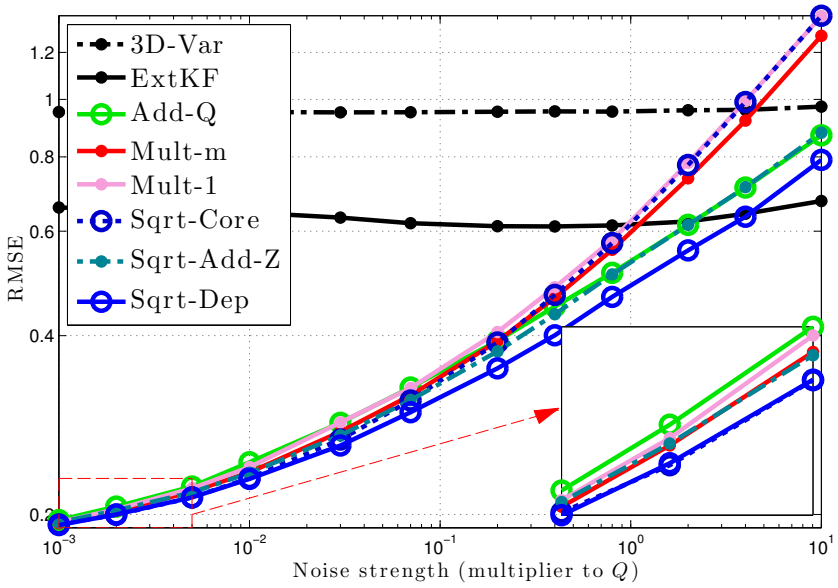


Figure 5.9: Performance benchmarks as a function of the noise strength, obtained with the Lorenz-96 system and $N = 25$. Both axes are logarithmic. On average, when $\mathbf{Q}$ is multiplied by 10^{-3} (resp. $10^{-2}, 10^{-1}, 10^{0}, 10^{1}$), the model noise makes up approximately 0.5 (resp. $4, 20, 70, 90$) percent of the growth in the spread of the ensemble. The climatology averages an RMSE score of approximately 4.

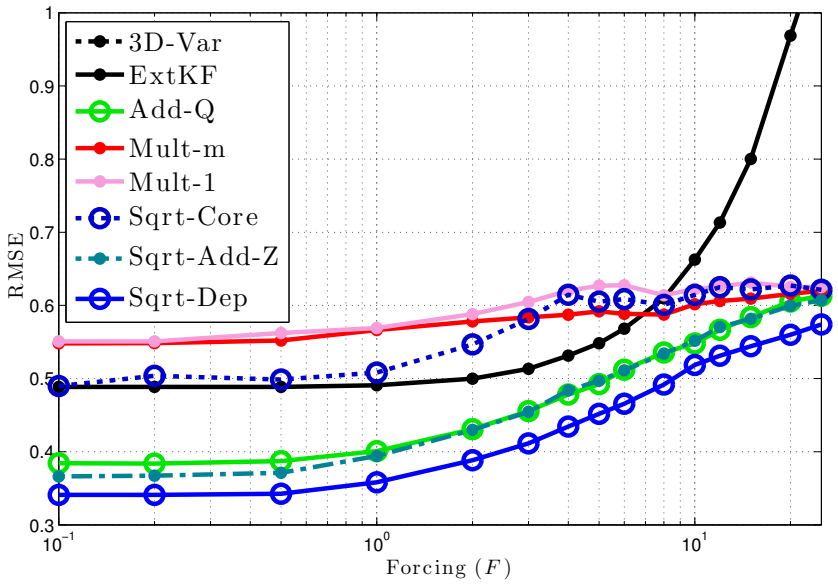


Figure 5.10: Performance benchmarks as a function of the forcing parameter, F, obtained with the Lorenz-96 system and $N = 25$.

quantity [40]. As F gets larger, the contribution to the growth in uncertainty from the dynamics increases in importance relative to that from model noise, explaining the coming together of the EnKF curves. The performance of the multiplicative methods and Sqrt-Core is very poor for small values of the forcing, but quite acceptable in the higher range. As in Fig. 5.8, Sqrt-Add-Z performs uniformly second best, after Sqrt-Dep.

5.8 Summary and discussion

The main effort of this study has been to extend the square root approach of the EnKF analysis step to the forecast step in order to account for model noise. Although the main motivation is to eliminate the need for simulated, stochastic perturbations, the core method, Sqrt-Core, was also found to possess several other desirable properties, which it shares with the analysis square root update. In particular, a review of these features revealed that the symmetric square root choice for the transform matrix can be beneficial in regards to dynamical consistency.

Yet, since it does not account for the residual noise, Sqrt-Core was found to be deficient in case the noise is strong and the dynamics relatively linear. In dealing with the residual noise, exploratory experiments (not shown) suggested that an additive approach works better than a multiplicative approach, similar to the forgetting factor of the SEIK. This is probably a reflection of the relative performances of Add-Q and Mult-m, as well as the findings of Whitaker and Hamill [205], which indicate that the additive approach is better suited to account for model error. Therefore, two additive techniques were proposed to complement Sqrt-Core, namely Sqrt-Add-Z and Sqrt-Dep. Adding simulated noise with no components in the ensemble subspace, Sqrt-Add-Z is computationally relatively cheap as well as intuitive. However, it was shown to yield biased covariance updates due to the presence of cross terms. By reintroducing dependence between the Sqrt-Core update and the sampled, residual noise, Sqrt-Dep remedies this deficiency at the cost of an additional SVD.

The utility of the noise integration methods proposed will depend on the properties of the system under consideration. However, Sqrt-Dep was found to perform robustly (nearly uniformly) better than all of the other methods. Moreover, the computationally less expensive method Sqrt-Add-Z was also found to have robust performance. These findings are further supported by omitted experiments using fewer observations, larger observation error, and different models.

The model noise square root approach has shown significant promise on low-order models, but has not yet been tested on realistic systems. It is also not clear how this approach performs with more realistic forms of model error.

As discussed in section B.2, a more shrewd choice of $\mathbf{Q}^{1/2}$ might improve Sqrt-Dep. This choice impacts $\hat{\mathbf{\Xi}}$, but not the core method, as shown in section B.1.3, and should not be confused with the choice of $\mathbf{T}^f$. While the Cholesky factor yielded worse performance than the symmetric choice, other options should be contemplated.

Nakano [144] proposed a method that is distinct, yet quite similar to Sqrt-Core, this should be explored further, in particular with regard to the residual noise.

Chapter 6

The EnKF-N and inflation

The recently developed "finite-size" ensemble Kalman filter (EnKF-N) explicitly considers the forecast moments and their (Bayesian) uncertainty. As such, it does not individually insert covariance estimates in the formula for the Kalman gain. It therefore avoids the inherent bias of the EnKF stated by Proposition 2.3.[1] Thus, in the perfect model setting, the EnKF-N does not require the multiplicative inflation commonly used to compensate for the bias, and avoids the burden of tuning it.

Here, the EnKF-N is derived in full, collecting the developments and corrections of the existing literature, as well as providing additional details and insights. Subsequently, the EnKF-N is reverse engineered into a scalar, scale mixture. This thoroughly explains why inflation may be used to compensate for sampling error, and helps in understanding the EnKF-N. Lastly, a deficiency apparent in the filter's performance in quasi-linear contexts is explained and rectified.

6.1 Framework

Suppose the forecast ensemble, $\mathbf{E} = \begin{bmatrix} \boldsymbol{x}_1, & \ldots & \boldsymbol{x}_n, & \ldots & \boldsymbol{x}_N \end{bmatrix}$, is the sole source of information on the true state, $\boldsymbol{x}$. As before, it is *assumed* that the ensemble is an iid sample from the same Gaussian distribution as the truth, i.e. $\boldsymbol{x}, \boldsymbol{x}_n \sim \mathcal{N}(\boldsymbol{b}, \mathbf{B})$. However, the true moments, $\boldsymbol{b}$ and $\mathbf{B}$ are *not known*.

[1] Proposition 2.3 is a statement on the average statistics of sampling error. Sampling error is an intrinsic problem of the EnKF, and should not be confused with the extrinsic problem of model noise. Also recall that the bias is present whether employing a perturbed-observation update or a square root update.

The EnKF-N is concerned with the analysis step only, enabling the simplified notation used above. Specifically, the notation of the true forecast moments, $\boldsymbol{x}^f$ and $\mathbf{P}^f$, is here simplified to $\boldsymbol{b}$ and $\mathbf{B}$, respectively; similarly, the forecast ensemble is denoted without the f superscript, i.e. $\mathbf{E}$, while its mean and covariance estimates are denoted $\bar{\boldsymbol{x}}$ and $\bar{\mathbf{B}}$. The probability distributions $\mathcal{N}$, $\boldsymbol{t}_m$, $\mathcal{W}_m^{+1}$, $\mathcal{W}_m^{-1}$, χ^{+2}, χ^{-2} are specified in the front matter, page x.

6.1.1 Outline

As an introduction to the concepts of this chapter, section 6.2 re-derives the standard EnKF from the simple but explicit assumption that the ensemble estimates of the forecast moments are exact. The EnKF-N is then developed by rejecting this assumption. The prior is derived in section 6.3, including careful attention to its parameterization in sections 6.3.2 and 6.3.3. The posterior is discussed in section 6.4. Section 6.4.1 considers the optimization for its mode, and section 6.4.2 deals with the update to the ensemble, and summarizes the EnKF-N algorithm. Section 6.5 then reverse engineers the EnKF-N to reveal its intimate connection with inflation. Alternative approaches to defining the inflation factor are discussed in appendix C. An illustration useful in understanding the EnKF-N is given in section 6.5.5. Section 6.6 presents some benchmarks of the EnKF-N performance. Section 6.7 discusses the optimality of various theoretical choices for the inflation value, and uses this perspective to rectify a deficiency in the performance of the EnKF-N. Concluding remarks are made in section 6.8.

6.2 Deriving the EnKF by assuming that the forecast moments are exact

The EnKF analysis step, as presented in chapter 2, substitutes $\bar{\mathbf{B}}$ in place of $\mathbf{B}$ to compute the Kalman gain, eqn. (2.7). The result is used, along with stochastic perturbations, to perform the analysis update to the ensemble. Section 2.4.1 then shows that this satisfies the KF equations in a particular statistical sense. This treatment has the drawbacks that it requires a leap of faith rather than being a self-driven derivation, and it "only" shows that the ensemble is a reasonable Monte Carlo sample of the Bayesian state distribution, but does not enable one to speak of the distribution itself.

Alternative derivations include that proceeding from Gaussian mixtures [188], and interpreting the analysis update as the first iteration in a Gauss-Newton minimization of the posterior of the randomized maximum likelihood method [42, 109]. Another approach is that of the Unscented Kalman filter [104], which is similar to the SEIK [158], and "cost function/variational" literature [154] in its explicit consideration of covariance matrices. The derivation below shares much in common with the latter. What is particular about it is that it proceeds from the *assumption* $\mathbf{B} = \bar{\mathbf{B}}$ only. It is not always fully recognized just how powerful this assumption is; this section studies its consequences to the fullest. The derivation also serves as an introduction to the EnKF-N.

What, explicitly, is $p(\boldsymbol{x}|\mathbf{E})$? Although this seems like a highly pertinent question, the answer is neither explicitly nor adequately provided by chapter 2. Nevertheless, an EnKF practitioner will typically propose that it is "the Gaussian with mean $\bar{\boldsymbol{x}}$ and covariance $\bar{\mathbf{B}}$", and will perhaps write $p(\boldsymbol{x}|\mathbf{E}) \propto \mathsf{exp}(-\frac{1}{2}[\boldsymbol{x} - \bar{\boldsymbol{x}}]^\mathsf{T}\bar{\mathbf{B}}^{-1}[\boldsymbol{x} - \bar{\boldsymbol{x}}])$. But, with $N \leq m$, this is invalid, as $\bar{\mathbf{B}}^{-1}$ does not exist since, recalling eqn. (2.11), $(N-1)\bar{\mathbf{B}} = \mathbf{A}\mathbf{A}^\mathsf{T}$, and $\mathbf{A}$ is rank deficient. Consider, instead, the generalized density,

$$p(\boldsymbol{x}|\mathbf{E}) \propto \mathsf{exp}\left(-\tfrac{1}{2}[\boldsymbol{x} - \bar{\boldsymbol{x}}]^\mathsf{T}\bar{\mathbf{B}}^{+}[\boldsymbol{x} - \bar{\boldsymbol{x}}]\right)\delta(\tilde{\mathbf{U}}^\mathsf{T}[\boldsymbol{x} - \bar{\boldsymbol{x}}])\,, \tag{6.1}$$

where δ is the Dirac delta function, $\bar{\mathbf{B}}^{+}$ is the pseudoinverse (section E.3) of $\bar{\mathbf{B}}$. and the columns of $\tilde{\mathbf{U}}$ form a basis of $\mathsf{range}(\mathbf{A})^\perp$. It can be checked that eqn. (6.1) yields $\mathbb{E}(\boldsymbol{x}|\mathbf{E}) = \bar{\boldsymbol{x}}$ and $\mathbb{V}\mathrm{ar}(\boldsymbol{x}|\mathbf{E}) = \bar{\mathbf{B}}$ [162, §8a.4].

Note that the assumptions $\boldsymbol{b} = \bar{\boldsymbol{x}}$ and $\mathbf{B} = \bar{\mathbf{B}}$ can, without doing any formal mathematics, be seen to *include* the assumption that there is zero uncertainty in the subspace $\mathsf{range}(\mathbf{A})^\perp$, meaning that $\boldsymbol{x}$ is restricted to the ensemble subspace.[2] But the pseudoinverse makes the exponential flat along directions of $[\boldsymbol{x} - \bar{\boldsymbol{x}}]$ in $\mathsf{range}(\mathbf{A})^\perp$, and thus has the opposite meaning. The use of the delta function(s) in eqn. (6.1) are therefore necessary in order to enforce the intended assumption, and to render the distribution proper.[3]

However, explicitly including the delta functions and worrying about the pseudoinverse is cumbersome. Consider, therefore, the parameterization $\boldsymbol{x}(\boldsymbol{w}) = \bar{\boldsymbol{x}} + \mathbf{A}\boldsymbol{w}$. This

[2] The ensemble subspace is the affine space $\{\boldsymbol{x} \in \mathbb{R}^m \,:\, [\boldsymbol{x} - \bar{\boldsymbol{x}}] \in \mathsf{range}(\mathbf{A})\}$.

[3] A "proper" distribution is one that is integrable; an improper distribution is one that is not, and therefore cannot be sampled from. The delta functions in eqn. (6.1) enforce the opposite extreme of impropriety, rendering the distribution "degenerate", meaning that some components of $\boldsymbol{x} - \bar{\boldsymbol{x}}$ have a certain (with probability 1) outcome.

It should be noted that caution is warranted when using the pseudoinverse elsewhere as well. For example, the substitution $\bar{\mathbf{B}}$ for $\mathbf{B}$ in eqn. (A.17a) yields the same result ($\mathbf{P}$) as in eqn. (A.18a), but not in eqn. (A.16a), using the pseudoinverse.

inherently restricts $\boldsymbol{x}$ to the ensemble subspace: $\boldsymbol{\Pi}_{\mathbf{A}}[\boldsymbol{x}(\boldsymbol{w}) - \bar{\boldsymbol{x}}] = [\boldsymbol{x}(\boldsymbol{w}) - \bar{\boldsymbol{x}}]$ for any $\boldsymbol{w} \in \mathbb{R}^N$. Note, however, that this does *not* introduce any additional approximations; as mentioned above, the restriction is already implied by the assumptions that $\boldsymbol{b} = \bar{\boldsymbol{x}}$ and $\mathbf{B} = \bar{\mathbf{B}}$. In terms of $\boldsymbol{w}$,

$$p(\boldsymbol{x}(\boldsymbol{w})|\mathbf{E}) \propto \mathsf{exp}\left(-\tfrac{1}{2}\boldsymbol{w}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}\bar{\mathbf{B}}^{+}\mathbf{A}\boldsymbol{w}\right) = \mathsf{exp}\left(-\tfrac{1}{2}(N{-}1)\boldsymbol{w}^{\mathsf{T}}\boldsymbol{\Pi}_{\mathbf{A}^{\intercal}}\boldsymbol{w}\right) . \tag{6.2}$$

The last equality follows from Lemma 5.1 and $\boldsymbol{\Pi}_{\mathbf{A}^{\intercal}} = \mathbf{A}^{+}\mathbf{A}$. The Jacobian[4] is constant, and so $p(\boldsymbol{w}|\mathbf{E}) \propto p(\boldsymbol{x}(\boldsymbol{w})|\mathbf{E})$. However, since any component of $\boldsymbol{w}$ in $\mathsf{ker}(\mathbf{A})$ has no bearing on $\boldsymbol{x}$, it can be ignored.[5] Therefore $(\mathbf{I}_N - \boldsymbol{\Pi}_{\mathbf{A}^{\intercal}})$ can be added to $\boldsymbol{\Pi}_{\mathbf{A}^{\intercal}}$ in eqn. (6.2), yielding simply $\mathbf{I}_N$, and

$$p(\boldsymbol{w}|\mathbf{E}) \propto \mathsf{exp}\left(-\tfrac{1}{2}(N{-}1)\|\boldsymbol{w}\|^2\right), \tag{6.3}$$

i.e. $p(\boldsymbol{w}|\mathbf{E}) = \mathcal{N}(\boldsymbol{w}|0, \frac{1}{N-1}\mathbf{I}_N)$. This simplification originated from the variational perspective [98], but this context, being fully Bayesian, is broader.

The above derivation of the prior required some effort with technicalities in order to maintain rigour, but the resulting prior, eqn. (6.3) is comparatively simple and intuitive. By Bayes' rule, the posterior is

$$p(\boldsymbol{w}|\boldsymbol{y},\mathbf{E}) \propto p(\boldsymbol{y}|\boldsymbol{w})\, p(\boldsymbol{w}|\mathbf{E}) \propto \mathsf{exp}\,{-\tfrac{1}{2}}\left(\|\boldsymbol{y} - h(\boldsymbol{x}(\boldsymbol{w}))\|^2_{\mathbf{R}} + (N{-}1)\|\boldsymbol{w}\|^2_{\mathbf{I}_N}\right) . \tag{6.4}$$

Recalling the definitions of the mean innovation and the anomalies of the observed ensemble, $\bar{\boldsymbol{\delta}}$ and $\mathbf{Y}$ of eqns. (4.3) and (4.4) respectively,

$$\boldsymbol{y} - h(\boldsymbol{x}(\boldsymbol{w})) = \boldsymbol{y} - h(\bar{\boldsymbol{x}} + \mathbf{A}\boldsymbol{w}) \approx \bar{\boldsymbol{\delta}} - \mathbf{Y}\boldsymbol{w} . \tag{6.5}$$

As discussed in section 2.3.2, this linearization of h by the ensemble is more accurately seen as a statical linearization rather than a finite difference approximation. It yields

$$p(\boldsymbol{w}|\boldsymbol{y},\mathbf{E}) \propto \mathsf{exp}\,{-\tfrac{1}{2}}\left(\|\mathbf{Y}\boldsymbol{w} - \bar{\boldsymbol{\delta}}\|^2_{\mathbf{R}} + (N{-}1)\|\boldsymbol{w}\|^2_{\mathbf{I}_N}\right) , \tag{6.6}$$

[4] Theorem 1 of Olkin and Roy [153] details the calculation of Jacobians for transformation to *sub*spaces. There is also a subtle complication due to the rank deficiency of $\mathbf{A}$, but this is shown in section 6.3.3 to be inconsequential in the Gaussian case.

[5] See section 6.3.3 for a thorough treatment.

Similarly as for the KF, Lemma A.2 then yields

$$p(\boldsymbol{w}|\boldsymbol{y},\mathbf{E}) = \mathcal{N}\Big(\boldsymbol{w} \mid \tfrac{1}{N-1}\mathbf{G}^a\mathbf{Y}^\mathsf{T}\mathbf{R}^{-1}\bar{\boldsymbol{\delta}},\ \tfrac{1}{N-1}\mathbf{G}^a\Big)\,, \tag{6.7}$$

$$\text{i.e.}\quad p(\boldsymbol{x}|\boldsymbol{y},\mathbf{E}) = \mathcal{N}\Big(\boldsymbol{x} \mid \bar{\boldsymbol{x}} + \tfrac{1}{N-1}\mathbf{A}\mathbf{G}^a\mathbf{Y}^\mathsf{T}\mathbf{R}^{-1}\bar{\boldsymbol{\delta}},\ \tfrac{1}{N-1}\mathbf{A}\mathbf{G}^a\mathbf{A}^\mathsf{T}\Big)\,, \tag{6.8}$$

where $\frac{1}{N-1}\mathbf{G}^a = \Big(\mathbf{Y}^\mathsf{T}\mathbf{R}^{-1}\mathbf{Y} + (N{-}1)\mathbf{I}_N\Big)^{-1}$ is the posterior covariance matrix of $\boldsymbol{w}$. A posterior ensemble can be generated for this distribution by computing a square root of $\mathbf{G}$, just as for the square root method, chapter 4. Indeed, the parameters of the Gaussian in eqn. (6.8) were also found in eqns. (4.6), (4.8) and (4.9). Note, however, that in the present derivation, by virtue of the parameterization in terms of $\boldsymbol{w}$, (i) it was not necessary to "discover" that $\mathbf{A}^f$ factorizes out on either side of eqn. (2.31) since this is anticipated by $\mathbb{V}\mathrm{ar}(\boldsymbol{x}) = \mathbb{V}\mathrm{ar}(\mathbf{A}\boldsymbol{w}) = \mathbf{A}\,\mathbb{V}\mathrm{ar}(\boldsymbol{w})\mathbf{A}^\mathsf{T}$, and (ii) the Woodbury identity is not required since eqn. (6.6) is already specified in an N-dimensional space. It also bears repeating that the derivation proceeds solely from the fundamental assumption that $\mathbf{B} = \bar{\mathbf{B}}$, and that it allows contemplating actual distributions, and not just Monte Carlo samples thereof.

The rest of this chapter is concerned with relaxing the assumption $\mathbf{B} = \bar{\mathbf{B}}$, and assumes nothing beyond conditional Gaussianity.

6.3 The EnKF-N prior

The standard EnKF assumes that the sample moments, $\bar{\boldsymbol{x}}$ and $\bar{\mathbf{B}}$, coincide with the true moments of the forecast distribution, $\boldsymbol{b}$ and $\mathbf{B}$, so that $p(\boldsymbol{x}|\mathbf{E}, \boldsymbol{b} = \bar{\boldsymbol{x}}, \mathbf{B} = \bar{\mathbf{B}}) = \mathcal{N}(\boldsymbol{x}|\bar{\boldsymbol{x}}, \bar{\mathbf{B}})$. Recognizing the fallacy of doing so, Bocquet [25] proposed the "finite-size" EnKF (the EnKF-N). It does not assume that $\boldsymbol{b} = \bar{\boldsymbol{x}}$ and $\mathbf{B} = \bar{\mathbf{B}}$, but instead asks what $p(\boldsymbol{x}|\mathbf{E})$ is, merely assuming the ensemble, $\mathbf{E} = \Big[\boldsymbol{x}_1,\ \ldots\ \boldsymbol{x}_n,\ \ldots\ \boldsymbol{x}_N\Big]$, is sampled from the same Bayesian distribution as the truth. The EnKF-N does not, however, relinquish the assumption of Gaussianity, conditional on knowing the moments, $\boldsymbol{b}$ and $\mathbf{B}$:

$$p(\boldsymbol{x}|\boldsymbol{b},\mathbf{B}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{b},\mathbf{B})\,. \tag{6.9}$$

Given the assumption of Gaussianity, finding $p(\boldsymbol{x}|\mathbf{E})$ is a parametric, hierarchical Bayesian inference problem, where $p(\boldsymbol{x}|\mathbf{E})$ is obtained by marginalizing out the conditional distribution of the moments. This was recognized by Bocquet et al. [30] as the well known problem [e.g. 72, p. 88] of obtaining the "posterior predictive" distri-

bution. However, the problem is significantly complicated by the fact that typically $N < p, m$. Furthermore, the conditioning on the observation, $\boldsymbol{y}$, and generation of a posterior ensemble, is not standard, but was first discussed in depth by Bocquet et al. [28]. Since $p(\boldsymbol{x}|\mathbf{E}, \boldsymbol{y})$ is the resulting posterior, $p(\boldsymbol{x}|\mathbf{E})$ is henceforth referred to as the "prior predictive" pdf. The following derivation of the EnKF-N benefits from the integral developments on the EnKF-N [25, 28, 30], but also contains details and insights beyond these publications.

6.3.1 Averaging over hyperparameters

By marginalization

$$p(\boldsymbol{x}|\mathbf{E}) = \int p(\boldsymbol{x}|\boldsymbol{b}, \mathbf{B}, \mathbf{E})\, p(\boldsymbol{b}, \mathbf{B}|\mathbf{E})\, d\boldsymbol{b}\, d\mathbf{B}\,, \tag{6.10}$$

where the integration region of $\mathbf{B} \in \mathbb{R}^{m\times m}$ is the subset of SPD matrices equipped with the standard Lebesgue measure on the $\frac{1}{2}m(m+1)$ distinct elements, which constitutes a topological cone [141, §2.1.2]. But given the true moments, the truth is independent of the ensemble, i.e. $p(\boldsymbol{x}|\boldsymbol{b}, \mathbf{B}, \mathbf{E}) = p(\boldsymbol{x}|\boldsymbol{b}, \mathbf{B})$. Now, the posterior for the moments can be expressed through Bayes' rule.

$$\begin{aligned} p(\boldsymbol{x}|\mathbf{E}) &\propto \int p(\boldsymbol{x}|\boldsymbol{b}, \mathbf{B})\, p(\mathbf{E}|\boldsymbol{b}, \mathbf{B})\, p(\boldsymbol{b}, \mathbf{B})\, d\boldsymbol{b}\, d\mathbf{B} && (6.11)\\ &= \int \mathcal{N}(\boldsymbol{x}|\boldsymbol{b}, \mathbf{B}) \left(\prod_{n=1}^{N} \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{b}, \mathbf{B}) \right) p(\boldsymbol{b}, \mathbf{B})\, d\boldsymbol{b}\, d\mathbf{B}\,. && (6.12) \end{aligned}$$

For generality and tractability, $\boldsymbol{b}$ and $\mathbf{B}$ are assumed a priori independent, and their hyperpriors are taken to be the common Jeffreys reference priors: $p(\boldsymbol{b}) \propto 1$ and $p(\mathbf{B}) \propto |\mathbf{B}|^{-(m+1)/2}$. Thus,

$$p(\boldsymbol{x}|\mathbf{E}) = \int |\mathbf{B}|^{-(N+m+2)/2}\, \mathsf{exp}\, -\tfrac{1}{2}\Big(\|\boldsymbol{x} - \boldsymbol{b}\|^2_{\mathbf{B}} + \sum_n \|\boldsymbol{x}_n - \boldsymbol{b}\|^2_{\mathbf{B}} \Big)\, d\boldsymbol{b}\, d\mathbf{B}\,, \tag{6.13}$$

where the parentheses around the exponent have been omitted to avoid clutter. The exponent is reformulated using $\bar{\boldsymbol{x}} - \bar{\boldsymbol{x}} = 0$ and Lemma A.2, thus gathering the terms dependent on $\boldsymbol{b}$

$$\begin{aligned} -2 \cdot \text{Exponent} &= \|\boldsymbol{x} - \boldsymbol{b}\|^2_{\mathbf{B}} + N\|\bar{\boldsymbol{x}} - \boldsymbol{b}\|^2_{\mathbf{B}} + \sum_n \|\boldsymbol{x}_n - \bar{\boldsymbol{x}}\|^2_{\mathbf{B}} && (6.14)\\ &= \left\| \boldsymbol{b} - \frac{N\bar{\boldsymbol{x}} + \boldsymbol{x}}{N+1} \right\|^2_{\mathbf{B}/(N+1)} + \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2_{\varepsilon_N \mathbf{B}} + \sum_n \|\boldsymbol{x}_n - \bar{\boldsymbol{x}}\|^2_{\mathbf{B}}\,, && (6.15) \end{aligned}$$

where $\varepsilon_N = \frac{N+1}{N}$. The first quadratic form in eqn. (6.15) depends on $\boldsymbol{x}$, but its result after the integration in eqn. (6.13) over $\boldsymbol{b}$ does not, its only relevant effect being a factor of $|\mathbf{B}|^{1/2}$. The latter two quadratic forms can be rewritten in terms of the sample covariance matrix, eqn. (2.9), and the identity $\|\boldsymbol{v}\|^2_{\mathbf{M}} = \mathsf{tr}(\boldsymbol{v}\boldsymbol{v}^\mathsf{T}\mathbf{M}^{-1})$, yielding

$$p(\boldsymbol{x}|\mathbf{E}) = \int |\mathbf{B}|^{-(N+m+1)/2} \, \mathsf{exp}\left(-\tfrac{1}{2}\,\mathsf{tr}(\mathbf{J}\mathbf{B}^{-1})\right) d\mathbf{B}\,, \tag{6.16}$$

where

$$\mathbf{J}(\boldsymbol{x}) = \varepsilon_N^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}})(\boldsymbol{x} - \bar{\boldsymbol{x}})^\mathsf{T} + (N-1)\bar{\mathbf{B}}\,. \tag{6.17}$$

The integrand can be recognized as the pdf of the inverse Wishart distribution $p(\boldsymbol{x}, \mathbf{B}|\mathbf{E}) \propto \mathcal{W}_m^{-1}(\mathbf{B}|N, \mathbf{J}(\boldsymbol{x}))$, the conjugate prior for the covariance matrix of a Gaussian sampling distribution [72]. But this pdf is improper (footnote 3, page 69) if $N \leq m-1$, indicating that it is necessary to supply further prior information. This is not a surprise; the Jeffreys prior is improper, and conditioning on the ensemble is not sufficient to make it proper in all dimensions.

6.3.2 Restricted parameterization

A more informative hyperprior, based on climatology, was explored by Bocquet et al. [30], effectively deriving localization from this rigorous framework. Here, however, the assumption is made that the state, $\boldsymbol{x}$, is in the ensemble subspace (footnote 2, page 69). As noted in section 6.2, this is a subset of the assumptions of the standard EnKF. The state dimensionality is thus reduced to $\hat{m} = N - g$, where g is the dimensionality of the kernel of $\mathbf{A}$ (typically $g = 1$, ref. Lemma 2.1). This assumption is normally warranted in applications of the EnKF, as discussed in section 2.5. If not, then neither the EnKF-N nor the standard EnKF will perform satisfactorily. Formalism that explicitly recognizes the assumption is employed. Specifically, the change of variables $\hat{\boldsymbol{w}} \mapsto \boldsymbol{x} = \bar{\boldsymbol{x}} + \hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\boldsymbol{w}}$ is used, where $\hat{\boldsymbol{w}} \in \mathbb{R}^{\hat{m}}$, $\mathbf{A} = \hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{V}}^\mathsf{T}$ is the reduced SVD of $\mathbf{A}$ (i.e. $\hat{\mathbf{\Sigma}} \in \mathbb{R}^{\hat{m}^2}$). This being an affine transformation, the Jacobian is a constant (footnote 4, page 70). Note that the ensemble mean is 0 in the space of $\hat{\boldsymbol{w}}$. Redoing the above derivation one finds $p(\hat{\boldsymbol{w}}, \mathbf{B}|\mathbf{E}) \propto \mathcal{W}_{\hat{m}}^{-1}(\mathbf{B}|N, \mathbf{J}(\hat{\boldsymbol{w}}))$, where $\mathbf{B}$ is of size $\hat{m} \times \hat{m}$ and, instead of eqn. (6.17),

$$\mathbf{J}(\hat{\boldsymbol{w}}) = \varepsilon_N^{-1}\hat{\boldsymbol{w}}\hat{\boldsymbol{w}}^\mathsf{T} + \mathbf{I}_{\hat{m}}\,, \tag{6.18}$$

The integration over $\mathbf{B}$ is then performed by the change of variables $\mathbf{C} \mapsto \mathbf{B} = \mathbf{J}^{1/2}\mathbf{C}\mathbf{J}^{\mathsf{T}/2}$, whose Jacobian is $|\mathbf{J}|^{(\hat{m}+1)/2}$ [79, §1.3], exactly cancelling out the reference prior and yielding

$$p(\hat{\boldsymbol{w}}|\mathbf{E}) = \int \left|\mathbf{J}^{1/2}\mathbf{C}\mathbf{J}^{\mathsf{T}/2}\right|^{-(N+m+1)/2} \exp\left(-\tfrac{1}{2}\operatorname{tr}(\mathbf{J}^{1/2}\mathbf{C}^{-1}\mathbf{J}^{-1/2})\right) |\mathbf{J}|^{(m+1)/2}\, d\mathbf{C} \quad (6.19)$$

$$= |\mathbf{J}|^{-N/2} \int |\mathbf{C}|^{-(N+m+1)/2} \exp\left(-\tfrac{1}{2}\operatorname{tr}(\mathbf{C}^{-1})\right) d\mathbf{C}\,. \quad (6.20)$$

The integral above is just a constant, so that $p(\hat{\boldsymbol{w}}|\mathbf{E}) \propto |\mathbf{J}|^{-N/2}$. The determinant, $|\mathbf{J}|$, is then rewritten in terms of the identity, valid for any vector $\boldsymbol{v}$ and any invertible matrix $\mathbf{M}$, $\left|\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}} + \mathbf{M}\right| = |\mathbf{M}|(1 + \|\boldsymbol{v}\|^2_{\mathbf{M}})$ [142, Th. 4.3.1], yielding

$$p(\hat{\boldsymbol{w}}|\mathbf{E}) \propto |\mathbf{J}(\hat{\boldsymbol{w}})|^{-N/2} \propto \left(\varepsilon_N + \|\hat{\boldsymbol{w}}\|^2_{\mathbf{I}_{\hat{m}}}\right)^{-N/2}\,. \quad (6.21)$$

This can be recognized as the $\hat{m}$-dimensional t distribution with $\nu = g$ degrees of freedom [111]. How does the t distribution look? If $\nu = 1$, it is also known as the Cauchy distribution, and, in the scalar case, it can be recognized as the Runge function [193, §5]:

$$\boldsymbol{t}_1(x|1;0,1) \propto \frac{1}{1+x^2}\,. \quad (6.22)$$

Illustrated in Fig. 6.1, the t distribution clearly has the aspect of a Gaussian distribution. However, the hallmark of the t distribution is its heavy tails, making it adept

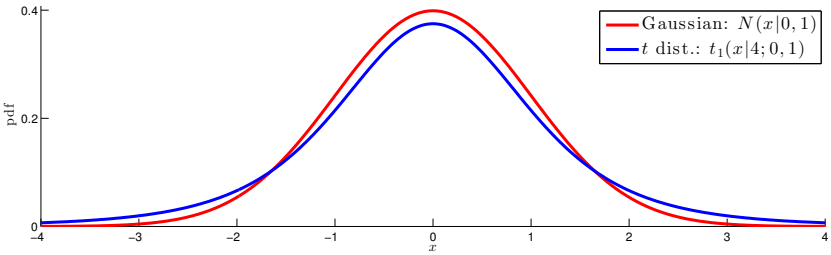


Figure 6.1: Illustration of the t distribution. Note the heavy tails of the t distribution.

for robust inference [73, §17.5]. Its heavy tails have the effect that the t distribution only has $\nu - 1$ moments. Thus, in our typical case, not even the mean exists.

6.3.3 Redundant parameterization

As is conventional in transform-based algorithms [154], this section derives the distribution of the "ensemble weights" $\boldsymbol{w} \in \mathbb{R}^N$ through the parameterization $\boldsymbol{x} = \bar{\boldsymbol{x}} + \mathbf{A}\boldsymbol{w}$, so that the n-th anomaly corresponds to the n-th unit vector: $(\boldsymbol{x}_n - \bar{\boldsymbol{x}}) = \mathbf{A}\boldsymbol{e}_n$. As for $\hat{\boldsymbol{w}}$, this parameterization also enforces the restriction of $\boldsymbol{x}$ to the ensemble subspace, but it also has the advantage that computing the SVD of $\mathbf{A}$ is not required. However, there is no theoretical necessity for this reparameterization; after all, the underlying probability space (i.e. $p(\boldsymbol{x}|\mathbf{E})$) does not change, and it can be shown that the associated symmetric square root transform update of the ensemble (chapter 4) is the same whether derived from $p(\boldsymbol{w}|\mathbf{E})$ or $p(\hat{\boldsymbol{w}}|\mathbf{E})$.

Note that $\boldsymbol{w} = \boldsymbol{w}(\hat{\boldsymbol{w}}, \tilde{\boldsymbol{w}}) = \hat{\mathbf{V}}\hat{\boldsymbol{w}} + \tilde{\mathbf{V}}\tilde{\boldsymbol{w}}$, for some $\tilde{\boldsymbol{w}} \in \mathbb{R}^g$, where the columns of $\tilde{\mathbf{V}} \in \mathbb{R}^{N\times g}$ are orthonormal vectors of the kernel of $\mathbf{A}$. Therefore the parameterization in $\boldsymbol{w}$ has g redundant degrees of freedom. This requires close scrutiny, but is not a novel concept (the simplest example of it is to express a single random variable as the sum of two) Indeed, since $\int p(\hat{\boldsymbol{w}}, \tilde{\boldsymbol{w}})\, d\tilde{\boldsymbol{w}} = p(\hat{\boldsymbol{w}})$ and $\left|\frac{\partial \boldsymbol{w}}{\partial(\hat{\boldsymbol{w}},\tilde{\boldsymbol{w}})}\right|$ is constant,

$$\int p(\boldsymbol{w}{=}\hat{\mathbf{V}}\hat{\boldsymbol{w}}{+}\tilde{\mathbf{V}}\tilde{\boldsymbol{w}})\, d\tilde{\boldsymbol{w}} = p(\hat{\boldsymbol{w}})\,. \tag{6.23}$$

Note that we temporarily dropped the "$|\mathbf{E}$" from all of the pdfs for notational simplicity, and used the abuse of notation $p(\boldsymbol{w}{=}\ldots)$ to clarify that the integrand is the pdf of $\boldsymbol{w}$. Using the change of variables $\boldsymbol{u} \mapsto \tilde{\boldsymbol{w}} = \sqrt{\varepsilon_N + \|\hat{\boldsymbol{w}}\|^2}\boldsymbol{u}$, it can be shown that the N-dimensional t distribution, again with g degrees of freedom,

$$p(\boldsymbol{w}|\mathbf{E}) \propto \left(\varepsilon_N + \|\boldsymbol{w}\|^2_{\mathbf{I}_N}\right)^{-(N+g)/2}, \tag{6.24}$$

satisfies eqn. (6.23). Note that the distribution chosen for the redundant degrees of freedom is not unique. For example, composition with affine transformations of the space of $\tilde{\boldsymbol{w}}$ also yield the correct marginal, $p(\hat{\boldsymbol{w}})$. Again, however, it can be shown that the associated update of the ensemble is not affected, and the distribution in eqn. (6.24) is chosen for its simplicity. Also note that the redundant parameterization is deceptively simple in the Gaussian case: $\mathcal{N}(\hat{\boldsymbol{w}}|0, \mathbf{I}_{\hat{m}}) = \int \mathcal{N}(\boldsymbol{w}|0, \mathbf{I}_N)\, d\tilde{\boldsymbol{w}}$, which is why the subtlety surrounding redundant parameterization is only a "conceptual difficulty" for Hunt et al. [98], and was not recognized to have an effect on the EnKF-N before Bocquet et al. [30]. However, as is shown in section 6.5.3, more care is needed when there are scalings on $\boldsymbol{w}$.

6.4 The EnKF-N posterior

Let $J(\boldsymbol{w}) = -2 \log p(\boldsymbol{w}|\mathbf{E}, \boldsymbol{y})$. By Bayes' rule,

$$J(\boldsymbol{w}) = \|\boldsymbol{y} - h(\boldsymbol{x}(\boldsymbol{w}))\|^2_{\mathbf{R}} + c_g \log\left(\varepsilon_N + \|\boldsymbol{w}\|^2_{\mathbf{I}_N}\right), \tag{6.25}$$

plus an inconsequential constant, and where $c_g = N + g$. One way of obtaining a posterior ensemble is to sample an ensemble of observation perturbations and optimize $J(\boldsymbol{w})$ for each individual realization [e.g. 152]. A less costly way is to approximate $p(\boldsymbol{w}|\mathbf{E}, \boldsymbol{y})$ by the Gaussian with the correct mode and the corresponding Hessian [25]. Bocquet et al. [28] used the L-BFGS-B minimizer of Byrd et al. [36] to find the optimum, and the ensemble-linearized h of eqn. (6.5) to approximate the gradient and the Hessian. The computational cost is then typically not much larger than for the standard EnKF.

6.4.1 Optimization

Despite the linearization of h, finding the critical point of $J(\boldsymbol{w})$ is not trivial. However, note that $J(\boldsymbol{w})$ is only non-quadratic through the norm of $\boldsymbol{w}$. Taking advantage of this through a change of variables that separates the radial and angular degrees of freedom of $\boldsymbol{w}$, Bocquet et al. [28] was able to reduce the optimization problem to a scalar one, as shown here.

Instead of employing spherical coordinates, the parameter space is augmented by ρ, and $\boldsymbol{w}$ is constrained to have a norm of ρ. A convenient, corresponding Lagrangian function is then

$$L(\boldsymbol{w}, \rho, \zeta) = \|\mathbf{Y}\boldsymbol{w} - \bar{\boldsymbol{\delta}}\|^2_{\mathbf{R}} + c_g \log(\varepsilon_N + \rho) + \zeta(\|\boldsymbol{w}\|^2_{\mathbf{I}_N} - \rho), \tag{6.26}$$

where $\rho, \zeta \in \mathbb{R}$.

By construction, $L(\boldsymbol{w}, \rho, \zeta)$ equals $J(\boldsymbol{w})$ along the constraint, $\rho = \|\boldsymbol{w}\|^2$. Instead of explicitly requiring $\frac{\partial L}{\partial \zeta} = 0$, however, note that $L(\boldsymbol{w}, \rho, \zeta)$ *is* quadratic in $\boldsymbol{w}$; indeed, the terms $\|\mathbf{Y}\boldsymbol{w} - \bar{\boldsymbol{\delta}}\|^2_{\mathbf{R}} + \zeta\|\boldsymbol{w}\|^2_{\mathbf{I}_N}$ can be recognized as the exponent of the standard EnKF, eqn. (6.6), adjusted with the inflation

$$\lambda^2 = \frac{N-1}{\zeta} \tag{6.27}$$

on the prior covariance estimate. Meanwhile, the condition $\frac{\partial L}{\partial \rho} = 0$ yields the unique minimum $\rho = \rho_\star(\zeta) = c_g/\zeta - \varepsilon_N$. Insert $\rho = \rho_\star(\zeta)$ and apply Lemma A.2 to obtain

$$L(\boldsymbol{w}, \rho_\star(\zeta), \zeta) = \|\boldsymbol{w} - \boldsymbol{w}_\star(\zeta)\|^2_{\mathbf{G}_\star(\zeta)} + \|\bar{\boldsymbol{\delta}}\|^2_{\mathbf{Y}\mathbf{Y}^\mathsf{T}/\zeta + \mathbf{R}} - c_g \log \zeta + \varepsilon_N \zeta \,, \tag{6.28}$$

plus an inconsequential constant, where

$$\mathbf{G}_\star(\zeta) = \left(\zeta \mathbf{I}_N + \mathbf{Y}^\mathsf{T} \mathbf{R}^{-1} \mathbf{Y}\right)^{-1} , \tag{6.29}$$

$$\boldsymbol{w}_\star(\zeta) = \mathbf{G}_\star(\zeta) \mathbf{Y}^\mathsf{T} \mathbf{R}^{-1} \bar{\boldsymbol{\delta}} = \mathbf{Y}^\mathsf{T} \left(\mathbf{Y}\mathbf{Y}^\mathsf{T} + \zeta \mathbf{R}\right)^{-1} \bar{\boldsymbol{\delta}} \,, \tag{6.30}$$

which can again be recognized in comparison with eqns. (4.5) and (4.8). Clearly, $\frac{\partial L}{\partial \boldsymbol{w}} = 0$ if and only if $\boldsymbol{w} = \boldsymbol{w}_\star(\zeta)$. Making this substitution as well,

$$D(\zeta) = L(\boldsymbol{w}_\star(\zeta), \rho_\star(\zeta), \zeta) = \|\bar{\boldsymbol{\delta}}\|^2_{\mathbf{Y}\mathbf{Y}^\mathsf{T}/\zeta + \mathbf{R}} + c_g \log \frac{1}{\zeta} + \varepsilon_N \zeta \,. \tag{6.31}$$

Since it is a function of the scalar ζ only, the minimization of D, recognizable as the "Lagrangian dual problem" [28], is computationally negligible provided the SVD of $\mathbf{R}^{-1/2}\mathbf{Y}$ has been precomputed, as detailed in section 4.3. Moreover, by the chain rule, $\frac{dD}{d\zeta}(\zeta) = \frac{\partial L}{\partial \zeta}(\boldsymbol{w}_\star(\zeta), \rho_\star(\zeta), \zeta)$ and therefore $\frac{\partial L}{\partial \zeta}(\boldsymbol{w}_\star(\zeta_\star), \rho_\star(\zeta_\star), \zeta_\star) = 0$, where $\zeta_\star$ is the minimizing argument of D. In summary, all of the partial derivatives of the Lagrangian are zero at $(\boldsymbol{w}_\star(\zeta_\star), \rho_\star(\zeta_\star), \zeta_\star)$ and $\boldsymbol{w}_\star(\zeta_\star)$ is a local minimum of J.

Note that the minimum of $\zeta \mapsto c_g \log \frac{1}{\zeta} + \varepsilon_N \zeta$ occurs at $\zeta = c_g/\varepsilon_N$, and that $\zeta \mapsto \|\bar{\boldsymbol{\delta}}\|^2_{\mathbf{Y}\mathbf{Y}^\mathsf{T}/\zeta + \mathbf{R}}$ is monotonically increasing. Hence the minimization of D can be safely limited to $[0, c_g/\varepsilon_N]$. However, D and hence J may have multiple local minima. While this is only a minor practical challenge, and seems to be a rare occurrence, it does pose some troubling theoretical questions. The issue is further discussed in section 6.4.2.

6.4.2 Analysis ensemble

The likelihood, $p(\boldsymbol{y}|\boldsymbol{x})$, is not conjugate with the t distribution predictive prior, and there is no simplified expression for the posterior beyond eqn. (6.25). Optimizing to find the mode of the posterior was described in section 6.4.1. However, its moments do not generally exist. Section 6.3.1 mentioned that the prior does not have any moments, for $g = 1$. The same applies to the posterior, unless the likelihood makes it proper. As seen from eqn. (6.25), however, this only happens in the row space of $\mathbf{Y}$, which might be smaller than the ensemble space if $p < N$. Moreover, even if they should exist, the meaning of the moments is not as significant as for Gaussians: the

mean does not generally coincide with the mode, and the variance is typically not the same as the inverse Hessian.

6.4.2.1 Using one of the candidate Gaussians

Recall, however, that the predictive prior is a continuous (multivariate) Gaussian mixture: a weighted collection of "candidate" Gaussians. Thus, by bringing the likelihood into the integral of eqn. (6.12), the posterior can also be identified as a Gaussian mixture:

$$p(\boldsymbol{x}|\boldsymbol{y},\mathbf{E}) \propto \int \mathcal{N}(\boldsymbol{x}|\boldsymbol{b}+\mathbf{K}[\boldsymbol{y}-\mathbf{H}\boldsymbol{b}],[\mathbf{I}-\mathbf{K}\mathbf{H}]\mathbf{B})\, p(\boldsymbol{b},\mathbf{B}|\mathbf{E})\, d\boldsymbol{b}\, d\mathbf{B}\,. \tag{6.32}$$

where $\mathbf{K}$ is also a function of $\mathbf{B}$. In view of this link, it seems reasonable to make the approximation that the posterior is a Gaussian, and more specifically one of the candidates from the mixture. The criterion used to chose among the candidate Gaussian posteriors is then that it has the correct mode, $\boldsymbol{w}_\star(\zeta_\star)$. This, as shown in section 6.4.1, corresponds[6] uniquely to the posterior obtained with a prior inflation $\lambda_\star$ such that $\lambda_\star^2 = (N-1)/\zeta_\star$. Let $q(\boldsymbol{w})$ be the pdf of this candidate Gaussian. The inverted Hessian of $-\log q(\boldsymbol{w})$ is its covariance, which can then be used to specify the posterior ensemble. It can be seen from eqns. (6.28) and (6.29) that the Hessian of $-\log q(\boldsymbol{w})$ is $(\mathbf{G}_\star(\zeta_\star))^{-1} = \zeta_\star \mathbf{I}_N + \mathbf{Y}^\mathsf{T}\mathbf{R}^{-1}\mathbf{Y}$. A summary of the EnKF-N analysis update algorithm developed above is:

1. Optimize $D(\zeta)$, eqn. (6.31) to obtain $\zeta_\star$, i.e. $\frac{N-1}{\lambda_\star^2}$.
2. Compute $\mathbf{G}_\star(\zeta_\star)$ and $\boldsymbol{w}_\star(\zeta_\star)$, eqns. (6.29) and (6.30).
3. Reconstitute the ensemble: $\mathbf{E}^a = [\bar{\boldsymbol{x}} + \mathbf{A}\boldsymbol{w}_\star(\zeta_\star)]\,\mathbb{1}^\mathsf{T} + \sqrt{N-1}\mathbf{A}(\mathbf{G}_\star(\zeta_\star))^{1/2}_{\text{sym}}$.

6.4.2.2 Using the Laplace approximation

Alternatively, instead of selecting a specific candidate among those making up $p(\boldsymbol{x}|\boldsymbol{y},\mathbf{E})$, another approach, known as the Laplace approximation [23, §4.4], is to fit the true posterior by a Gaussian with the same mode and Hessian (at the mode). Although the Laplace approximation will share the same mode as q, the Hessian of eqn. (6.25)

[6] The word "corresponds" is used, because the link to inflation of Gaussians that was provided in section 6.4.1 can only be spoken of using the conjunction "as if". Section 6.4.1 does not fully derive the EnKF-N posterior from an inflation starting point.

is

$$\frac{1}{2}\frac{\partial^2 \mathbf{J}}{\partial \boldsymbol{w}^2} = \mathbf{Y}^\mathsf{T}\mathbf{R}^{-1}\mathbf{Y} + c_g \left\{ \frac{\mathbf{I}_N}{\varepsilon_N + \|\boldsymbol{w}\|^2_{\mathbf{I}_N}} - \frac{2\boldsymbol{w}\boldsymbol{w}^\mathsf{T}}{\left(\varepsilon_N + \|\boldsymbol{w}\|^2_{\mathbf{I}_N}\right)^2} \right\}, \tag{6.33}$$

which, evaluated at $\boldsymbol{w}_\star(\zeta_\star)$, is

$$\frac{1}{2}\frac{\partial^2 \mathbf{J}}{\partial \boldsymbol{w}^2}(\boldsymbol{w}_\star(\zeta_\star)) = \mathbf{Y}^\mathsf{T}\mathbf{R}^{-1}\mathbf{Y} + \zeta_\star \mathbf{I}_N - \frac{2\zeta_\star^2}{c_g}\boldsymbol{w}_\star(\zeta_\star)\boldsymbol{w}_\star(\zeta_\star)^\mathsf{T}, \tag{6.34}$$

where $\zeta_\star = c_g/(\varepsilon_N + \|\boldsymbol{w}_\star(\zeta_\star)\|^2_{\mathbf{I}_N})$ has been used (derived from the Lagrangian optimality conditions). Note that the last term is not present in the Hessian of $-\log q(\boldsymbol{w})$. Bocquet et al. [30] showed that it originates in the interdependence of the radial and angular degrees of freedom of $p(\boldsymbol{w}|\boldsymbol{y}, \mathbf{E})$. It is not evident which approximation is better: q or that of Laplace. Anyway, numerical experiments (figure 3 of [30]) indicate that the performance difference is negligible. The Laplace approximation is not employed in this text; instead, only the algorithm using a candidate Gaussian, described previously, is employed.

6.5 The EnKF-N as a scale mixture

It is striking that (providing h is linear) the posterior distribution of the EnKF-N is only non-Gaussian in a single degree of freedom (section 6.4.1). Indeed, though the EnKF-N was derived to account for sampling error in the forecast ensemble, it turned out that it effectively acts like a method of adaptive inflation (section 6.4.1). This section reverse engineers the EnKF-N as a method of (scalar) inflation, putting this perspective on firm ground. The derivation is arguably more accessible, allowing for simpler interpretation (section 6.5.5) and closer attention to the specific choice of inflation value (section 6.7).

6.5.1 The scaled, Gaussian random variable as a covariance scale mixture

Suppose the random variable $\boldsymbol{x}$ is defined as the product $\boldsymbol{x} = \sqrt{s}\boldsymbol{v}$, where $\boldsymbol{v}$ is the m-dimensional Gaussian random variable: $\boldsymbol{v} \sim \mathcal{N}(0, \mathbf{B})$, and s is an independent random variable that follows the inverse chi-square distribution with ν degrees of freedom: $s \sim \chi^{-2}(\nu)$. Now $p(\boldsymbol{x}) = \int p(\boldsymbol{x}|s)\, p(s)\, ds$, where the integration domain is

the positive real line. But

$$p(\boldsymbol{x}|s) = \left|\frac{\partial \boldsymbol{v}}{\partial \boldsymbol{x}}\right| p(\boldsymbol{v}{=}\boldsymbol{x}/\sqrt{s}) = s^{-m/2}\mathcal{N}(\boldsymbol{x}/\sqrt{s}|0,\mathbf{B}) = \mathcal{N}(\boldsymbol{x}|0,s\mathbf{B})\,, \tag{6.35}$$

and therefore

$$p(\boldsymbol{x}) = \int \mathcal{N}(\boldsymbol{x}|0,s\mathbf{B})\,p(s)\,ds\,. \tag{6.36}$$

Thus the product, $\boldsymbol{x}$, can also be seen as a covariance scale mixture.

6.5.2 The t distribution as a scale mixture of Gaussians

This mixture is analytically integrable. This development closely follows the derivation of the original EnKF-N prior, but is significantly simpler.

$$p(\boldsymbol{x}) = \int \mathcal{N}(\boldsymbol{x}|0,s\mathbf{B})\,\chi^{-2}(s|\nu)\,ds \tag{6.37}$$

$$\propto \int s^{-m/2}e^{-\frac{1}{2}\|\boldsymbol{x}\|^2_{s\mathbf{B}}}\,s^{-(\nu+2)/2}e^{-\frac{1}{2}\nu/s}\,ds \tag{6.38}$$

$$= \int s^{-(m+\nu+2)/2}e^{-\frac{1}{2}J(x)/s}\,ds\,, \tag{6.39}$$

where $J(\boldsymbol{x}) = (\|\boldsymbol{x}\|^2_{\mathbf{B}} + \nu)$. The change of variables $u \mapsto s = J(\boldsymbol{x})u$ then yields

$$p(\boldsymbol{x}) = J(\boldsymbol{x})^{-(m+\nu)/2}\int \ldots du \tag{6.40}$$

$$\propto \left(1 + \tfrac{1}{\nu}\|\boldsymbol{x}\|^2_{\mathbf{B}}\right)^{-(m+\nu)/2} \tag{6.41}$$

$$\propto \boldsymbol{t}_m(\boldsymbol{x}|0;\mathbf{B},\nu)\,. \tag{6.42}$$

It was mentioned in the context of Fig. 6.1 that the t distribution has much heavier (polynomial) tails than the exponentially decaying tails of the Gaussian. In light of the above, this can be seen to be inherited from the heavy tail of the inverse chi-square distribution, illustrated in Fig. 6.2. A heuristic explanation based in eqn. (6.38) is that a differential increase in $\|\boldsymbol{x}\|^2_{\mathbf{B}}$ is hardly "felt" by the exponential, $\mathsf{exp}(-\frac{1}{2}\|\boldsymbol{x}\|^2_{s\mathbf{B}})$, because it is mainly accounted for by an increase in s.

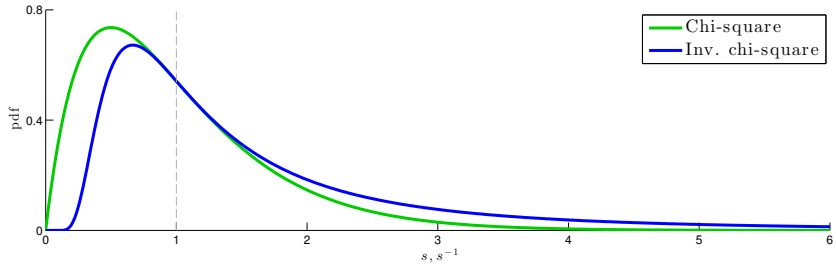


Figure 6.2: Illustration of the chi-square and inverse chi-square distributions, both with the parameter value $\nu = 4$; as such, they are related by the change of variables $s \mapsto 1/s$. Although the distributions have a similar shape, note that the inverse chi-square pdf has a heavier (thicker) tail. Indeed, it only decays polynomially, while the chi-square pdf decays exponentially. For both distributions, the (position of the) mode tends to 1 and the peak becomes sharper as ν increases.

6.5.3 Identification of $D(\zeta)$ as a probability density function

The EnKF-N prior was derived assuming the truth and ensemble is drawn from the same conditional Gaussian distribution, but with unknown mean and covariance, whose probabilities must therefore be averaged out. The resulting distribution, the predictive prior, is

$$p(\boldsymbol{w}|\mathbf{E}) \propto \left(\varepsilon_N + \|\boldsymbol{w}\|^2_{\mathbf{I}_N}\right)^{-c_g/2} \tag{6.43}$$

$$\text{i.e.} \quad p(\boldsymbol{w}|\mathbf{E}) = \boldsymbol{t}_N(\boldsymbol{w}|g; 0, \tfrac{\varepsilon_N}{g}\mathbf{I}_N)\,, \tag{6.44}$$

by identification. But, as shown in section 6.5.2, the multivariate t distribution can be interpreted as a scale mixture of Gaussians, i.e.

$$p(\boldsymbol{w}|\mathbf{E}) \propto \int \mathcal{N}(\boldsymbol{w}|0, s\tfrac{\varepsilon_N}{g}\mathbf{I}_N)\, \chi^{-2}(s|g)\, ds\,. \tag{6.45}$$

Now, condition on the observations and, as in the "empirical Bayes" approach [168, §10.4.2], include the likelihood in the integral:

$$p(\boldsymbol{w}|\mathbf{E}, \boldsymbol{y}) \propto \int \mathcal{N}(\bar{\boldsymbol{\delta}}|\mathbf{Y}\boldsymbol{w}, \mathbf{R})\, \mathcal{N}(\boldsymbol{w}|0, s\tfrac{\varepsilon_N}{g}\mathbf{I}_N)\, \chi^{-2}(s|g)\, ds \tag{6.46}$$

$$\propto \int \exp -\tfrac{1}{2}\left(\|\mathbf{Y}\boldsymbol{w} - \bar{\boldsymbol{\delta}}\|^2_{\mathbf{R}} + \|\boldsymbol{w}\|^2_{s\frac{\varepsilon_N}{g}\mathbf{I}_N} + \frac{g}{s} + (c_g + 2)\log s\right) ds\,. \tag{6.47}$$

The change of variables $\zeta \mapsto s = \frac{g}{\varepsilon_N}/\zeta$ then yields the Lagrangian of eqn. (6.28):

$$p(\boldsymbol{w}|\mathbf{E},\boldsymbol{y}) \propto \int \mathsf{exp}-\tfrac{1}{2}\Big(\|\mathbf{Y}\boldsymbol{w}-\bar{\boldsymbol{\delta}}\|^2_{\mathbf{R}} + \zeta\|\boldsymbol{w}\|^2_{\mathbf{I}_N} + \varepsilon_N\zeta + (c_g-2)\,\mathsf{log}\,\frac{1}{\zeta}\Big)\,d\zeta \tag{6.48}$$

$$\approx c\,\mathsf{exp}-\tfrac{1}{2}\Big(\|\mathbf{Y}\boldsymbol{w}-\bar{\boldsymbol{\delta}}\|^2_{\mathbf{R}} + \zeta_\star\|\boldsymbol{w}\|^2_{\mathbf{I}_N} + \varepsilon_N\zeta_\star + c_g\,\mathsf{log}\,\frac{1}{\zeta_\star}\Big) \tag{6.49}$$

$$= c\,\mathsf{exp}-\tfrac{1}{2}\Big(\|\boldsymbol{w}-\boldsymbol{w}_\star(\zeta_\star)\|^2_{\mathbf{G}_\star(\zeta_\star)} + \|\bar{\boldsymbol{\delta}}\|^2_{\mathbf{Y}\mathbf{Y}^\intercal/\zeta_\star+\mathbf{R}} + \varepsilon_N\zeta_\star - c_g\,\mathsf{log}\,\zeta_\star\Big) \tag{6.50}$$

$$= c\,\mathsf{exp}-\tfrac{1}{2}L(\boldsymbol{w},\rho_\star(\zeta_\star),\zeta_\star)\,. \tag{6.51}$$

Equation (6.49) is the Laplace approximation of eqn. (6.48), including $|\mathcal{H}/(2\pi)|^{-1/2} = c\zeta_\star$ for some constant c, where $\mathcal{H}$ is the second derivative, with respect to ζ, of the exponent of the integrand, and $\zeta_\star$ is a particular value for ζ. Equation (6.50) is obtained by Lemma A.2. The mode of $p(\boldsymbol{w}|\mathbf{E},\boldsymbol{y})$ is $\boldsymbol{w}_\star$ of eqn. (6.30), and $\mathbf{G}_\star(\zeta_\star)$ is specified by eqn. (6.29). Furthermore, by the law of total probability (i.e. marginalization), the integrand in eqn. (6.48) can be identified as $p(\boldsymbol{w},\zeta|\mathbf{E},\boldsymbol{y})$. Thus, the dual objective function, $D(\zeta)$, from the Lagrangian dual problem of section 6.4.1 is identified as[7]

$$D(\zeta) \approx -2\,\mathsf{log}\,\Big(\mathcal{H}^{-1/2}\,p(\boldsymbol{w}_\star,\zeta|\mathbf{E},\boldsymbol{y})\Big)\,, \tag{6.52}$$

plus an inconsequential constant.[8]

Alternatively, from eqn. (6.47) again, the change of variables $\lambda^2 \mapsto s = \frac{g}{(N-1)}\lambda^2$ yields

$$p(\boldsymbol{w}|\mathbf{E},\boldsymbol{y}) \propto \int \mathsf{exp}-\tfrac{1}{2}\left(\|\mathbf{Y}\boldsymbol{w}-\bar{\boldsymbol{\delta}}\|^2_{\mathbf{R}} + \frac{N-1}{\varepsilon_N\lambda^2}\|\boldsymbol{w}\|^2_{\mathbf{I}_N} + \frac{N-1}{\lambda^2} + (c_g+2)\,\mathsf{log}\,\lambda^2\right)d\lambda^2$$

$$\propto \int \mathcal{N}(\bar{\boldsymbol{\delta}}|\mathbf{Y}\boldsymbol{w},\mathbf{R})\,\mathcal{N}(\boldsymbol{w}/\lambda|0,\tfrac{\varepsilon_N}{N-1}\mathbf{I}_N)\,\chi^{-2}(\lambda^2|N{-}1)\,\lambda^{-g}\,d\lambda\,, \tag{6.53}$$

and, again, the integrand can be identified as the pdf $p(\boldsymbol{w},\lambda^2|\mathbf{E},\boldsymbol{y})$. The last factor, λ^{-g}, is there so that $\int p(\hat{\boldsymbol{w}},\tilde{\boldsymbol{w}},\lambda^2)\,d\tilde{\boldsymbol{w}} = p(\hat{\boldsymbol{w}},\lambda^2)$, as explained in section 6.3.3. Hence

$$p(\hat{\boldsymbol{w}}|\mathbf{E},\boldsymbol{y}) \propto \int \mathcal{N}(\bar{\boldsymbol{\delta}}|\mathbf{Y}\hat{\boldsymbol{w}},\mathbf{R})\,\mathcal{N}(\hat{\boldsymbol{w}}/\lambda|0,\tfrac{\varepsilon_N}{N-1}\mathbf{I}_{\hat{m}})\,\chi^{-2}(\lambda^2|N{-}1)\,d\lambda\,, \tag{6.54}$$

[7] This identification proceeds from the laws of probability, and should be distinguished from that of section 6.4.1, which only notes the striking resemblance of the Lagrangian objective function to the exponent of a posterior pdf obtained with an inflated Gaussian prior.

[8] Finally, it can be argued that $\zeta_\star$ should be the optimum of D (i.e. including the Hessian factor) by noting that $\boldsymbol{w}_\star$ depends on ζ and therefore $\boldsymbol{w}$ should not be seen as a constant in applying the Laplace approximation. This is possibly related to the discussion in [30] on the coupling on the radial and angular degrees of freedom.

i.e. the EnKF-N is a scale mixture with the scale prior $p(\lambda^2|\mathbf{E}) = \chi^{-2}(\lambda^2|N{-}1)$. The prior is explained in the following.

6.5.4 Defining λ and deriving its prior

We now revert back to the familiar notation of $\boldsymbol{x}$ and $\mathbf{B}$, although the following argumentation should be re-done, as in section 6.3.2, in terms of $\hat{\boldsymbol{w}}$ and its covariance matrix. Suppose

$$p(\boldsymbol{x}|\mathbf{B}) = \mathcal{N}(\boldsymbol{x}|\bar{\boldsymbol{x}}, \mathbf{B}) = |2\pi\mathbf{B}|^{-1/2} \exp\left(-\tfrac{1}{2}\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2_{\mathbf{B}}\right). \tag{6.55}$$

We define $\lambda > 0$ so that $\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2_{\mathbf{B}} = \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2_{\lambda^2\bar{\mathbf{B}}}$, i.e.

$$\lambda^2 = \frac{\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2_{\bar{\mathbf{B}}}}{\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2_{\mathbf{B}}}, \tag{6.56}$$

and hence

$$p(\boldsymbol{x}|\lambda^2, \mathbf{E}) = |2\pi\mathbf{B}|^{-1/2} \exp\left(-\tfrac{1}{2}\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2_{\lambda^2\bar{\mathbf{B}}}\right) \propto \mathcal{N}(\boldsymbol{x}/\lambda|\bar{\boldsymbol{x}}, \bar{\mathbf{B}}), \tag{6.57}$$

as in eqn. (6.54), where the proportionality holds with respect to λ^2 as well as $\boldsymbol{x}$. The following Theorem [79, 141, Ths. 3.2.8 and 3.3.11, respectively] will be used to derive the prior for λ^2.

Theorem 6.1 – A chi-square statistic**.**
Let $\boldsymbol{u} \neq 0$ be any m-dimensional vector, or an (almost never zero) random vector. If $\mathbf{M} \sim \mathcal{W}^{+1}_m(\nu, \boldsymbol{\Sigma})$ is independent of $\boldsymbol{u}$, then

$$\frac{1}{\nu}\frac{\boldsymbol{u}^\mathsf{T}\mathbf{M}\boldsymbol{u}}{\boldsymbol{u}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{u}} \sim \chi^{+2}(\nu), \tag{6.58}$$

which is also independent of $\boldsymbol{u}$.

But, as shown for eqn. (C.4), $\mathbf{B}^{-1}|\mathbf{E}$ has the Wishart distribution

$$p(\mathbf{B}^{-1}|\mathbf{E}) = \mathcal{W}^{+1}_m\left(\mathbf{B}^{-1}|N{-}1, ((N{-}1)\bar{\mathbf{B}})^{-1}\right) \tag{6.59}$$

$$\propto \left|\mathbf{B}^{-1}\right|^{+(N-m)/2-1} \exp\left(-\tfrac{1}{2}(N{-}1)\,\mathsf{tr}(\bar{\mathbf{B}}\mathbf{B}^{-1})\right). \tag{6.60}$$

Applying Theorem 6.1 to the definition of λ^2 with $\boldsymbol{u} = \boldsymbol{x} - \bar{\boldsymbol{x}}$ and $\mathbf{M} = \mathbf{B}^{-1}$ yields

$$p(\lambda^2|\mathbf{E}) = \chi^{-2}(\lambda^2|N{-}1)\,. \tag{6.61}$$

Thus, by marginalization,

$$p(\boldsymbol{x}|\mathbf{E}) = \int p(\boldsymbol{x}|\lambda, \mathbf{E})\, p(\lambda|\mathbf{E})\, d\lambda \tag{6.62}$$

$$= \int \mathcal{N}(\boldsymbol{x}/\lambda|\bar{\boldsymbol{x}}, \bar{\mathbf{B}})\, \chi^{-2}(\lambda^2|N{-}1)\, d\lambda\,, \tag{6.63}$$

again, as indicated by eqn. (6.54). Note that in the above application of Theorem 6.1, the vector $\boldsymbol{u} = \boldsymbol{x} - \bar{\boldsymbol{x}}$ is a constant, because its value is fixed by the left hand side of eqn. (6.62).

Alternative definitions based on the ratios of "total variance", "generalized variance", marginal variance, and conditional variance are explored in appendix C.

6.5.5 Interpretation

Figure 6.3 illustrates the EnKF-N in the scalar case. The (thin orange-red) *candidate* prior Gaussian curves all sum to 1. They are distinguished by their variance, s, which have been sampled from the inverse chi-square distribution of Fig. 6.2. The "sampling" is deterministic so that the cumulative probability between two subsequent scales is a constant, namely 1/17, since 17 is the number of candidates plotted. Therefore, the (thick orange-red) *effective* prior t distribution is the average of the candidate priors, as described in section 6.5.2.

The candidate posterior Gaussian curves (thin blue) do not sum to one, but are simply obtained by multiplying pointwise the candidate priors with the likelihood, as in eqn. (6.46). The effective posterior curve (thick blue) is again the average of the candidate posteriors, but with a subsequent normalization; it is also the pointwise product of the effective prior and the likelihood.

Additionally, note that although the effective posterior is again a continuous Gaussian mixture, the mixture is now across locations as well as scales. Therefore the effective posterior is not a t distribution. It also has the interesting consequence that the variance of the effective posterior will depend on the innovation $\bar{\boldsymbol{\delta}} = \boldsymbol{y} - \overline{h(\mathbf{E}^f)}$. Computationally, this is achieved in the EnKF-N by the inflation factor estimation. It is interesting because it is a sure sign of non-Gaussian effects in the inference, as discussed in section A.3.

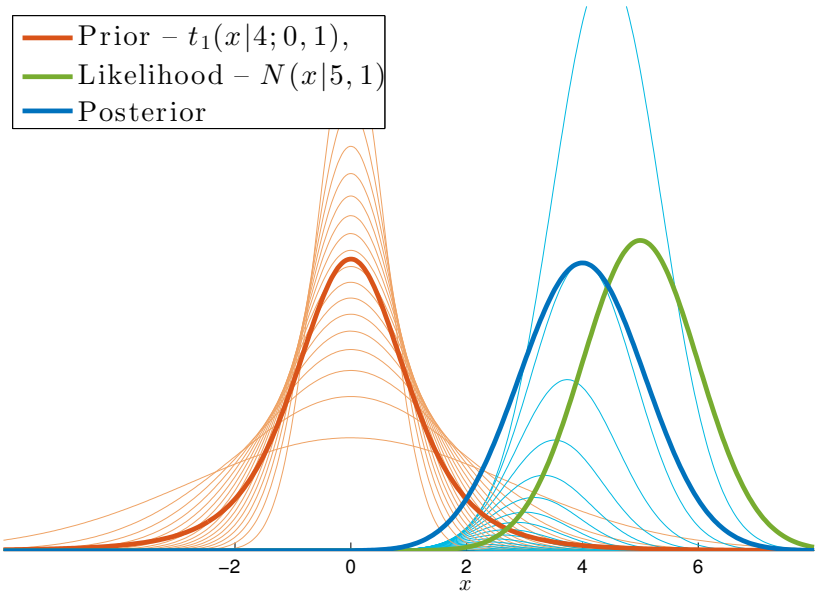


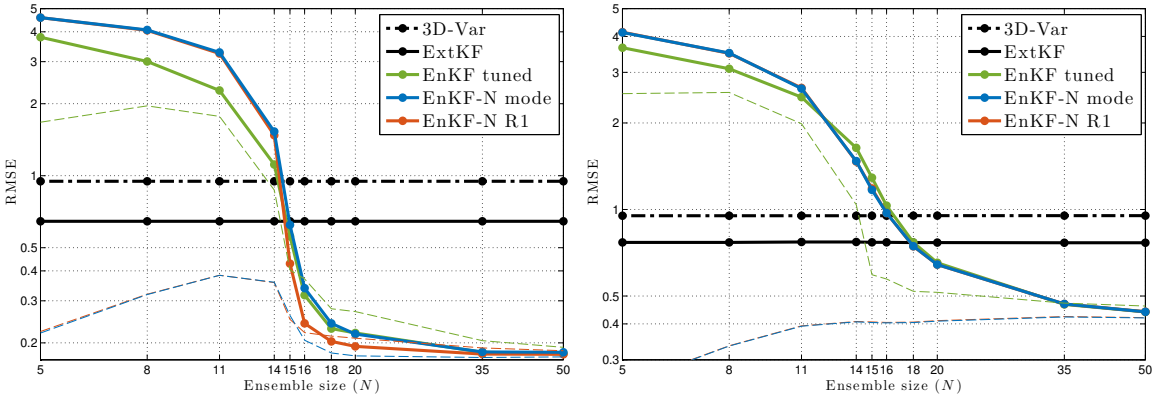
Figure 6.3: Illustration of the EnKF-N as a scale mixture of Gaussian distributions.

The set of candidate posterior Gaussians have been vertically scaled so that the one that shares the same mode as the effective posterior also has the same height. This makes it easy to see that no candidate Gaussian is coincident with the effective posterior, and are only approximations, as discussed in section 6.4.2.1. On the other hand, the Laplace approximation (not plotted) of section 6.4.2.2 is very nearly coincident with the effective posterior. However, this is not always the case: if the likelihood had been closer to the prior (but not on top of it) then the effective posterior would have been more asymmetric, and the Laplace approximation would have been less suited.

6.6 Benchmark results

The EnKF-N is benchmarked using twin experiments, as described in chapter 3. It is labelled "EnKF-N mode" in the experiments, in reference to how the inflation value is obtained as the optimum of $D(\zeta)$. The method labelled "EnKF-N R1" will be discussed in section 6.7. The reference comparison method, labelled "EnKF tuned", is the EnKF with the symmetric square root analysis update and optimally-tuned inflation, initially described in section 5.7. All of the ensemble methods use additive, simulated noise to account for model noise. Thus, the optimally tuned EnKF is the same as the method referred to as Add-Q in chapter 5. The inflation values used in the experiments are listed in Table F.2.

Consider Fig. 6.4, which plots performance as a function of the ensemble size, N. Subfigure (a) is essentially a reproduction of figure 3 of [25], with some improvements. Subfigure (b) uses the same settings as Fig. (5.7a).



(a) No model noise. **(b)** With model noise, specified by eqn. (5.40).

Figure 6.4: Performance benchmarks as a function of the ensemble size, N, obtained with the Lorenz-96 system.

In the linear-Gaussian case, the RMSE should equal ε_N times the average root-mean variance (the thin, dashed lines) of the ensemble. Outside of this context, it is not generally clear how the root-mean variance should relate to the RMSE; nevertheless, some degree of resemblance (e.g. monotonicity) is to be expected. An instantaneous (i.e. not time-averaged statistics) occurrence of a large discrepancy of these two statistics is called "filter divergence". It effectively sabotages the filter's use of the dynamics, consequently ruining its performance [68]. Indeed, as can be seen, for very small ensembles, the EnKF-N utterly fails to estimate the actual uncertainty, and the RMSE performance is worse than that of 3D-Var.

However, for this range of experiments, the optimally tuned EnKF also performs worse than 3D-Var and the extended Kalman filter. On the other hand, in the experiments with intermediate and high ensemble sizes, i.e. $N \geq 15$, the EnKF-N can be seen to attain approximately the same performance as the optimally tuned EnKF, and the estimated variances relate reasonably to the RMSE values.

The inflation factors estimated by the EnKF-N during the twin experiment with $N = 20$ of Fig. (6.4a) are tallied in Fig. 6.5. The histogram has the aspect of the distribution of an inverse chi-square variable illustrated in Fig. 6.2; however, it presumably deviates from this shape somewhat, as it is modulated by the likelihood, and should also be affected by non-Gaussianity of the underlying ensembles.

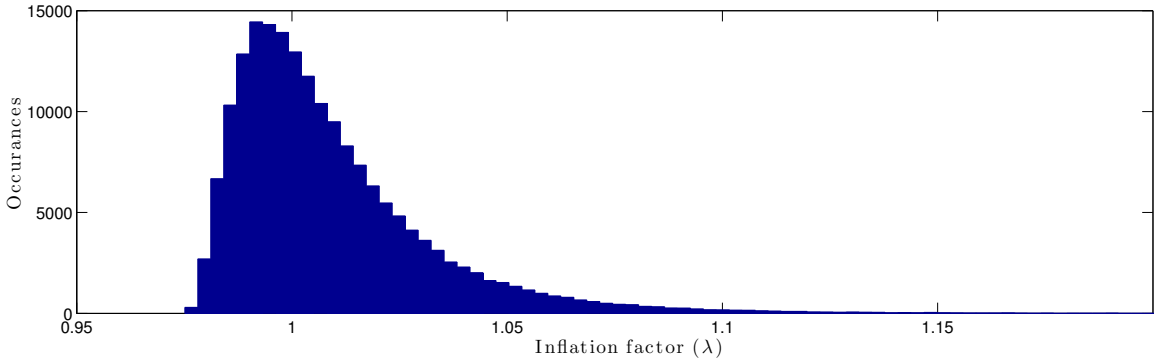
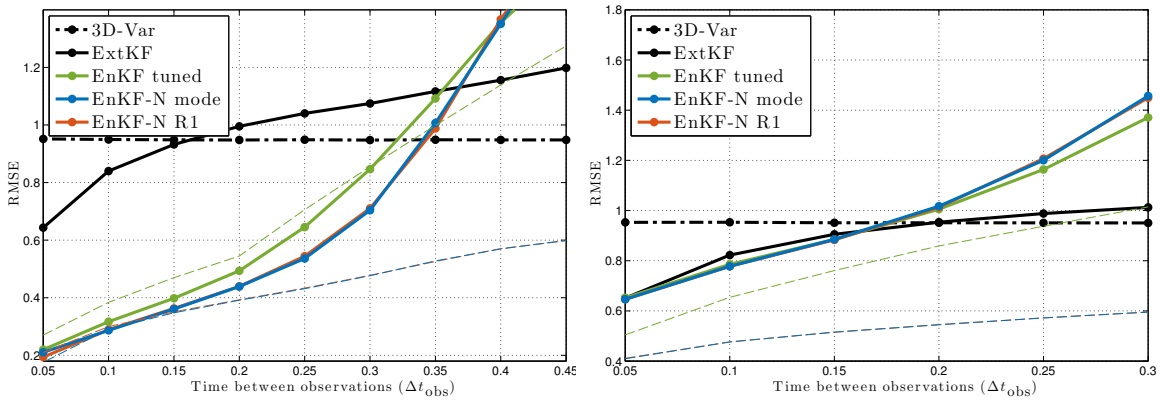


Figure 6.5: Inflation factors estimated by the EnKF-N "mode" in the experiment with $N = 20$ of Fig. (6.4a).

Figure 6.6 graphs the RMSE versus the DA window, Δt_{obs}. Subfigure (a) is essentially a reproduction of figure 4 of [25], with some improvements. Subfigure (b) uses the same settings as Fig. 5.8. Again, the EnKF-N can be seen to perform as well as the optimally tuned EnKF for experiments where the EnKF performance is better than that of 3D-Var or the extended Kalman filter. Similar tendencies were observed in experiments (not shown) with $N = 25$.



(a) No model noise. **(b)** With model noise, specified by eqn. (5.40).

Figure 6.6: Performance benchmarks as a function of the data assimilation window, Δt_{obs}, obtained with the Lorenz-96 system and $N = 20$.

It is remarkable that the estimated variances are significantly lower for the EnKF-N than for the optimally-tuned EnKF, both for Fig. 6.4 and Fig. 6.6, even though the RMSE performances are nearly coincident. This can likely be explained by the fact that the inflation of the EnKF-N is estimated on-line during the twin experiment, allowing for more confidence in certain instances.

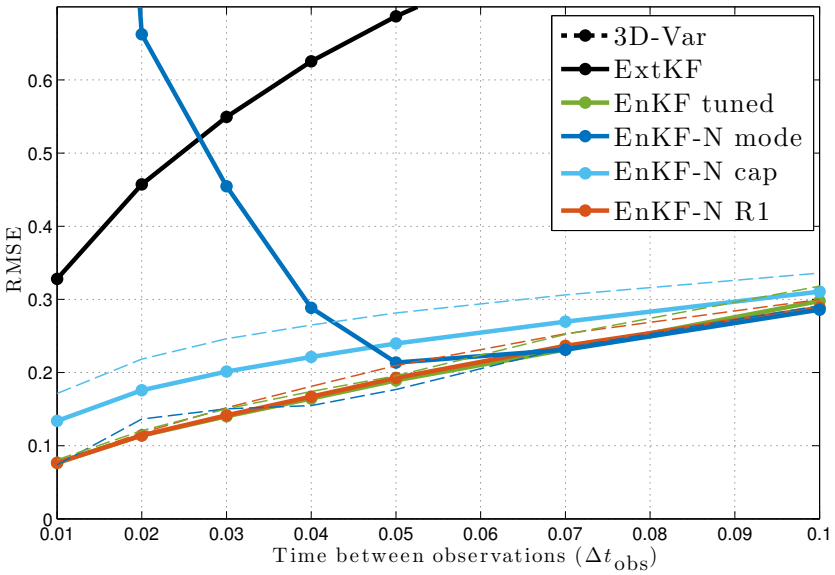


Figure 6.7: As Fig. (6.6a), but investigating a range of lower values of Δt_{obs}. The minor discrepancies compared to that figure can be attributed to the fact that it is necessary to use the shorter integration time step ($\Delta t = 0.01$) for this figure, effectively altering the dynamics.

Further experiments documenting the robust and strong performance of the EnKF-N is similar settings can be found in [28].

6.7 Choosing the inflation value

The experimental results presented in section 6.6, show that the EnKF-N performs quite satisfactorily in most cases. Strangely, however, the performance of the EnKF-N visibly deteriorates towards the left end of Fig. (6.6a), i.e. in setting where the uncertainty propagation is quasi-linear. Note that the point $\Delta t_{\text{obs}} = 0.05$ corresponds to the entirety of Fig. (6.4a). Figure 6.7 investigates this issue further, and shows that the performance becomes catastrophic for very low values of Δt_{obs}.

This was first diagnosed by Bocquet et al. [28] as occurring because the inflation was systematically being estimated at too low values. It was rectified by restricting the search space of $D(\zeta)$ to $\zeta < N - 1$, i.e. $\lambda > 1$. The result is the method plotted in Fig. 6.7 as "EnKF-N cap". It has robust performance across all of the experiments it was tested in. However, it remains unclear why the Gaussian framework of the EnKF-N should produce worse results in the more Gaussian setting.

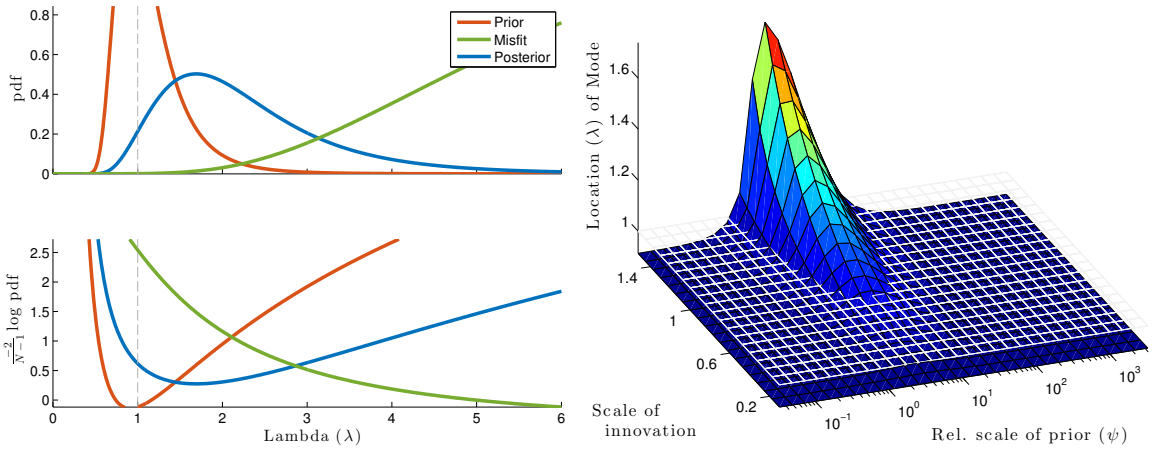
But section 6.5.3 made it clear that the dual objective function is an actual probability distribution, namely $D(\zeta) = -2 \log p(\zeta, \boldsymbol{w}_\star | \mathbf{E}, \boldsymbol{y})$. It is illustrated as the

posterior in Fig. (6.8a). This lucidity enables a deeper analysis of the strategy for choosing the value of λ from eqn. (6.49), which in turn yields a specific candidate Gaussian posterior approximation. This section presents the above quasi-linear problem as being caused by making an unsuitable choice of λ. The explanation is linked to the non-dimensional quantity ψ, measuring the relative confidence of the prior as compared to the likelihood;

$$\psi = \sqrt{\mathsf{tr}(\mathbf{H}\bar{\mathbf{B}}\mathbf{H}^{\mathsf{T}}\,\mathbf{R}^{-1})}\,, \tag{6.64}$$

which is, again, readily available with the SVD of $\mathbf{R}^{-1/2}\mathbf{Y}$ precomputed.

As seen in Fig. (6.8a), the posterior is not symmetric. It therefore does not have a distinguished point estimate [182]. Nevertheless, the "EnKF-N mode" method only uses a single point, namely the mode, as its inflation factor.



(a) Here, $\psi = 1$ and the innovation scale is 1.5 times larger than its expected value $\mathbf{R} + \frac{1}{N-1}\mathbf{Y}\mathbf{Y}^{\mathsf{T}}$. The line $\lambda = 1$ is added as a visual aid.

(b) The location of the posterior's mode mapped over a range of innovation scales and relative prior confidences. The white mesh is the plane $\lambda = 1$.

Figure 6.8: Subfigure (a) illustrates the prior $p(\lambda^2|\mathbf{E}) = \chi^{-2}(\lambda^2|N{-}1)$, the misfit, $p(\boldsymbol{y}, \boldsymbol{w}_\star|\mathbf{E}, \lambda^2) = \mathsf{exp}\left(-\frac{1}{2}\|\bar{\boldsymbol{\delta}}\|^2_{\mathbf{Y}\mathbf{Y}^{\mathsf{T}}/\zeta+\mathbf{R}}\right)$, and the posterior $p(\boldsymbol{w}_\star, \zeta|\mathbf{E}, \boldsymbol{y}) = \mathsf{exp}\left(-\frac{1}{2}\,D(\zeta)\right)$, where $\zeta = \frac{N-1}{\lambda^2}$. The probability distributions are normalized, while the misfit has been scaled (resp. offset) in the upper (resp. lower) axes for visibility. For both subfigures, $N = 6$, $\mathbf{H} = \mathbf{R} = \mathbf{I}_{10}$, and $\mathbf{Y}$ has been generated so that $\mathbf{H}\bar{\mathbf{B}}\mathbf{H}^{\mathsf{T}} \approx \psi^2\mathbf{R}$. The "innovation scale" is defined as $\|\bar{\boldsymbol{\delta}}\|^1_{\mathbf{Y}\mathbf{Y}^{\mathsf{T}}+\mathbf{R}}$, and is adjusted by scaling $\bar{\boldsymbol{\delta}}$.

The obvious alternatives, the mean and the median, are a bit more costly to compute (requiring quadrature) and perform significantly worse in twin experiment benchmarks. Mixed approaches such as using the "skewness" of the posterior to weight the mean and the mode have fared similarly. A different option is the Monte-Carlo

approach of sampling N realizations $\{\lambda_n \ ; \ n \in 1{:}N\}$ of the inflation factor from $D(\frac{N-1}{\lambda^2})$ so that each ensemble member uses a different prior variance, namely $\lambda_n^2 \bar{\mathbf{B}}$. However, this risks causing a significant amount of sampling error in itself. Therefore, an alternative that was considered is to condensing the N inflation realizations into a single one by computing their geometric mean: $\sqrt[N]{\prod_n \lambda_n}$. In the limit of large N, this becomes $\mathsf{exp}(\mathbb{E}\,\mathsf{log}(\lambda^2))$, which can be computed with quadrature. However, also this fails to attain the same performance as the choice of the mode in most benchmark tests.

A more rigorous approach to point estimation is provided by decision theory. The optimal point estimate is then the point, say $d > 0$, that minimizes the expected loss, $\mathbb{E}\, l(\lambda^2, d)$, where the expectation is taken with respect to the posterior, $p(\lambda^2, \boldsymbol{w}_\star | \mathbf{E}, \boldsymbol{y})$. Unfortunately, it is not clear which loss function to use.[9] Furthermore, the optimization is quite costly, because it requires quadrature to compute the expected loss for each choice, d, that is tried out. For these reasons, we revert to using special choices, such as the mode, as the point estimate.

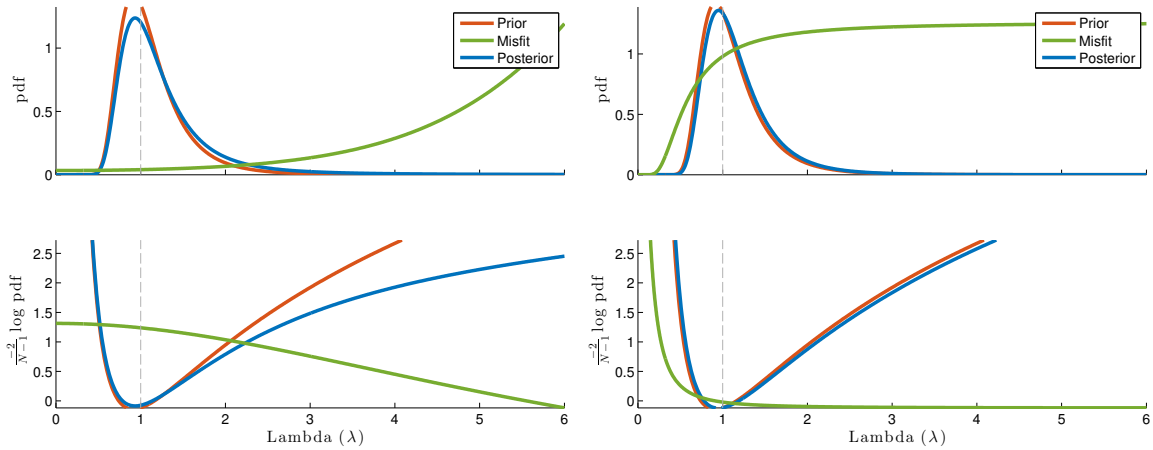
Fig. (6.8b) maps the location of the mode in a wide range of contexts. Note that, for any fixed-ψ cross section, the location of the mode increases in the direction of increasing innovation scale. By contrast the mode location is not monotonic with respect to the relative scale of the prior. Furthermore, it is only notably sensitive to the innovation in the band of intermediate values of ψ. Thinking in terms of feedback systems, the non-monotonicity and the restriction of the region of sensitivity are both somewhat disconcerting features of the map. However, as Fig. 6.9 illustrates, these features are not a consequence of using the mode as the point estimate, but rather a property of the full distributions. As the figures show, for very small or large ψ, the misfit is nearly flat in the interesting intervals, and therefore fails to impact the prior.

Thus, outside of the context of intermediate ψ, the prior and posterior are nearly equal. But, even with the factor ε_N (which can be juggled between the Gaussian and the scale priors as in section 6.5.3), the mode of the prior is

$$\lambda_b^2 = \frac{N-1}{N+1}\varepsilon_N = \frac{N-1}{N}\,, \tag{6.65}$$

which is always less than 1. One context in which this is inadmissible is in the limit of large $\mathbf{R}$, i.e. small ψ. What does an infinite $\mathbf{R}$ represent? That the observations

[9] We have tested the k-power loss function, $l(\lambda^2, d) = \left|\lambda^2 - d\right|^k$, with several values of k, and the "natural loss function for variances" [168, Ex. 4.2.7], $l(\lambda^2, d) = \lambda^{-4}(\lambda^2 - d)^2$. Neither have performed particularly well, nor provided a map of locations in the sense of Fig. (6.8b) that seems particularly useful.



(a) Here, $\psi = 0.1$. Inflection point: $\lambda = 3.6$. **(b)** Here, $\psi = 10$. Inflection point: $\lambda = 0.034$.

Figure 6.9: As Fig. (6.8a), but here with two opposite, extreme values of ψ. The misfit curve approximately consists of two pieces of opposite convexity, but its inflection point depends largely on ψ, as illustrated by the markedly different misfit curves of (a) and (b).

contain no information. Thus, if $\mathbf{R}$ is infinitely large, the ensemble should not "receive" any update whatsoever, and the inflation factor must be 1.[10] It is therefore undertaken to modify the *prior*, $p(\lambda^2|\mathbf{E}) = \chi^{-2}(\lambda^2|N{-}1)$, so that its mode tends to 1 as $\psi \to 0$. This is *not* a correction of the prior, but rather a computational trick that amends the *posterior* point estimate, while maintaining the relative computational simplicity of mode finding. Outside of the limit of $\psi \to 0$, it is not possible to pre-assert what the inflation factor should be. However, as was shown in section 6.6, the (unmodified) posterior mode there provides good performance, and it is therefore retained.

The modified prior, $q_\alpha(\lambda^2)$, is obtained by the rescaling: $q_\alpha(\lambda^2) = \chi^{-2}(\alpha\lambda^2|N{-}1)$, where the factor α is given by

$$\alpha = \left(\lambda_b^2\right)^{\frac{1}{1+\psi}} \ . \tag{6.66}$$

This formula is not derived, but engineered so that (i) the mode of q_α tends to 1 as $\psi \to 0$, but (ii) $\alpha \to 1$ and the mode is unaffected as $\psi \to \infty$. As such, the specific expression of eqn. (6.66) is not very important; indeed other formulae that fulfil the same requirements have been tested and yield similar performance [30]. As regards

[10] Note that this context occurs with the quasi-linear context above: as Δt_{obs} decreases and becomes very small, so will the relative confidence of the prior ψ. This can be seen by comparing the RMSE values (which are closely related to $\bar{\mathbf{B}}$, as discussed in section 6.6) which is $\mathbf{I}_{40}$ in all of the experiments.

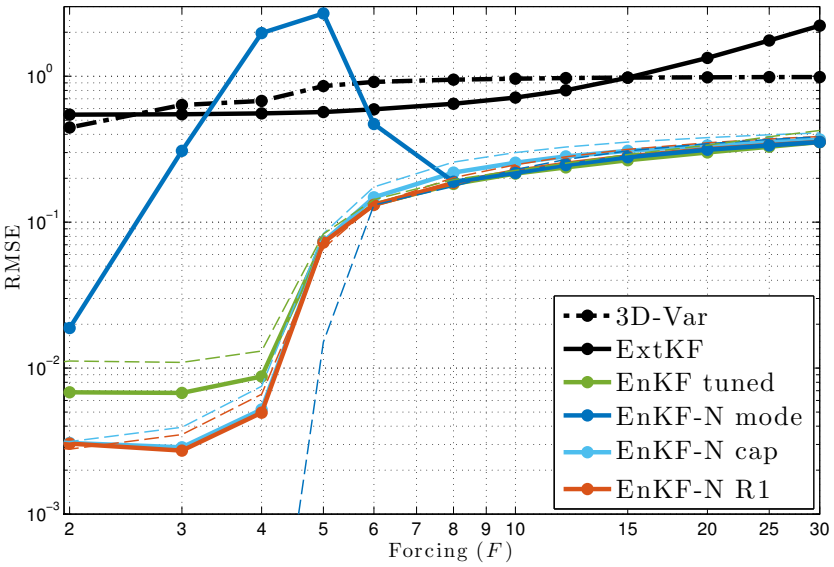


Figure 6.10: Performance benchmarks as a function of the forcing parameter, F, obtained with the Lorenz-96 system and $N = 20$.

the map of the locations of the mode, as in Fig. (6.8b), the effect of this modification is to smoothly elevate the left edge of the map to $\lambda = 1$.

As regards performance, the modification yields good results. Its benchmarks are plotted as "EnKF-N R1". It can be seen in the benchmarks of section 6.6 that it maintains the strong performance of the original EnKF-N in moderately nonlinear contexts. Figures 6.7 and 6.10 show that is also performs well in the quasi-linear context, sometimes even slightly better than "EnKF-N cap".

6.8 Summary and discussion

This chapter has discussed the estimation of the inflation parameter to compensate for sampling error. This has been achieved through the theory of the EnKF-N: a hierarchical, Gaussian framework to the EnKF analysis step. The theory, in particular the reverse engineered scalar perspective, explains why inflation may be used to compensate for sampling error, and helps in understanding the EnKF-N. Twin experiment benchmarks have indicated strong and robust performance, except in contexts where the observations are very uncertain compared to the prior. Lastly, a deficiency apparent in the filter's performance in quasi-linear contexts has been explained, and a rectification has been proposed and shown to work.

In the future, this proof of concept should be tested on real-world models; trials have already been conducted [2, 29], with moderate success, but should be comple-

mented by further investigation. One advantage of the scalar, scale mixture framework of section 6.5 is that it simplifies extending the EnKF-N framework so as to also compensate for model error (in addition to sampling error); though not included here, exploratory attempts at making such a hybrid method have been made, showing some promise, and should be further pursued. A drawback of the scalar framework is that it is less general than that of section 6.3, which marginalizes over full covariance matrices; indeed, Bocquet et al. [30] showed that it is possible to implement localization by employing more informative covariance priors; furthering this research constitutes another future work direction.

Chapter 7

The ensemble Rauch-Tung-Striebel smoother and its equivalence to the ensemble Kalman smoother

In contrast to filtering, the smoothing problem is also concerned with estimates of earlier states (section 1.2.2). The Rauch-Tung-Striebel (RTS) smoother is a linear-Gaussian solution to the smoothing problem that is well known in the engineering community. This chapter is a study of its ensemble formulation (EnRTS). An online expression is derived and discussed. In particular, it is used to show that the EnRTS is equivalent to the ensemble Kalman smoother (EnKS), even in the nonlinear, non-Gaussian case. The theory is re-examined under practical considerations and equability is illustrated by numerical experiments, even though equivalence is broken by inflation and localization.

7.1 Introduction

In the literature concerning ensemble methods, the Rauch-Tung-Striebel smoother [163] is much less discussed than the "augmented filter smoothers" [57, 60, 64], including the ensemble Kalman filter (EnKF) itself, the batch and sequential ensemble Kalman smoothers (EnKS), the fixed-lag smoother, and the fixed-point smoother. Nevertheless, the RTS smoother is the more recognized linear-Gaussian smoother[1]

[1]The relative obscurity of the linear-Gaussian augmented filters might be because they are feasible only for limited time augmentations, due to the growth in size of the covariance matrices. Indeed, as opposed to the ensemble convention, in the linear-Gaussian case, it is the RTS smoother that typically carries the moniker "Kalman smoother" [e.g. 23].

[134], and an ensemble formulation (EnRTS) was invented early on by Lermusiaux and Robinson [117]. It has since been reinvented and has found more popularity in the engineering community through its "unscented" formulation[2] [177, 179].

The RTS smoother is one of three classes of (non-iterative) smoothers surveyed by Cosme et al. [48], the other two being the augmented filters and the "two-filter" smoother. It is possible to formulate other smoothing recursions [e.g. 3, 53], but it is not clear if this can lead to competitive algorithms.

The RTS smoother only updates each state estimate twice: once during its forward, filtering "pass", and once during its similarly recursive, backward smoothing pass. Since assimilating new observations (seemingly) requires redoing the backward pass, it is labelled "off-line": not suitable for applications where the estimates should be repeatedly updated with new, incoming data.

By contrast, the smoothers classified as augmented filters may be efficiently applied on-line (as well as off-line). As indicated by their designation, they can all be derived and implemented as one, solely by augmenting the state vector of the filter with the states of previous times. Moreover, in their general formulation, linear-Gaussian formulation, and ensemble formulations, they are all equivalent in the sense that, when estimating the same entities, their estimates are equal.

Equivalence between the EnRTS and the EnKS is not obvious. Establishing the equivalence fosters understanding, and makes the choice between the two algorithms guided solely by practical considerations such as ease of implementation and operation.

7.1.1 Relevance

Smoothing refers to state estimation utilizing observations made both before and after the time of the state of interest. Because of their hindsight, smoothers typically improve on filter estimates. In geophysics, this makes smoothers ideally suited for reanalysis, either with the EnKS [54, 107, 212] or the EnRTS [116, 189]. Related applications include initial condition estimation and asynchronous filtering [173].

For approximate estimation techniques, like the ensemble approach, iterating is one way to deal with nonlinear dependencies [74, 102]. If the iterations span multiple time steps, smoothing is then inherent, even if the sole purpose of the estimation is forecast initialization, because the conditioning information has to be

[2]Very briefly, the unscented approach has the following contrasts to the ensemble approach: covariance matrices may be explicitly computed and the deterministic ensemble parameterizations of the uncertainty is repeatedly re-initialized rather than updated. Both of these contrasts make it unsuited for most geoscience applications, as mentioned in section 2.6.

communicated backward in time.[3] Note that nonlinearity is especially bountiful in parameter estimation problems, where the parameter-variable dependencies have been established over several time steps of dynamics.

7.1.2 Outline

The RTS smoother is described in its general Bayesian, linear-Gaussian, and ensemble formulations in section 7.2. The ensemble Kalman smoother is briefly reviewed in section 7.3. Equivalence is proved in section 7.4, and revisited in light of practical considerations in section 7.5.

7.1.3 Notation

The time index variable T is always used so that $t \leq T$. The case of $t = T$ defines empty sequences, for which the convention is used that the sequence product is the identity, and its sum is zero.

In the context of smoothing, there are more than just the "forecast" and "analysis" cases of conditioning. In order to distinguish these, the simplified notation using the superscripts f and a are abandoned in favour of *sub*scripts explicitly denoting the relevant time indices of the state and the conditioning. This is detailed in section 7.2.2.

As previously, $\mathbf{\Pi}_{\mathbf{M}}$ is the orthogonal projection matrix onto the column space of the (generic) matrix $\mathbf{M}$, and $\mathbf{\Pi}_{\mathbf{M}}^{\perp} = \mathbf{I} - \mathbf{\Pi}_{\mathbf{M}}$ the orthogonal complementary projector. We shall work with primitive variables as much as possible, so as to avoid the myriad of variables that come with matrix decompositions. For example, using the pseudoinverse, $\mathbf{M}^{+}$, the projection is given by $\mathbf{\Pi}_{\mathbf{M}} = \mathbf{M}\mathbf{M}^{+}$, though in practice the pseudoinverse will not be formed in the computations, which will only use the SVD.

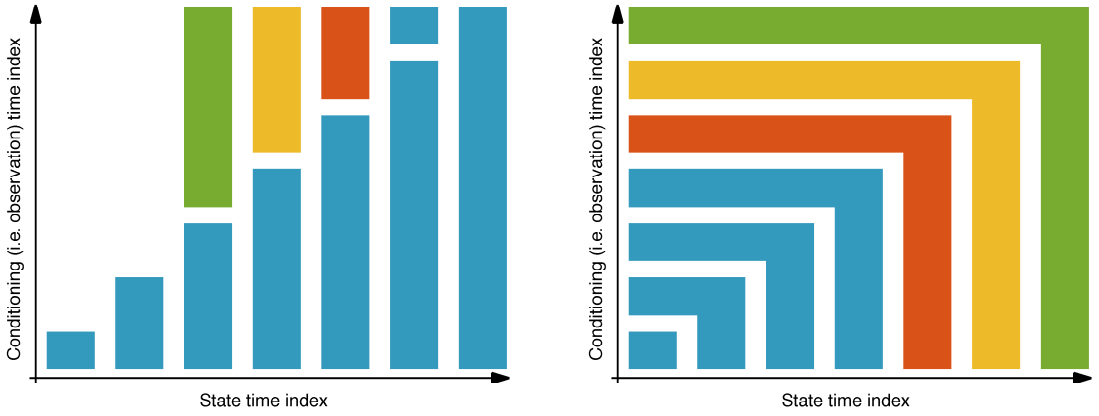
[3] Current iterative smoothers [e.g. 44, 55] are typically built on the ensemble (batch) smoother, which does not exploit the time-sequential dependency structure of the data assimilation problem. This is mainly because of issues related to dynamical consistency, as discussed in section 5.3. Another reason is that forecasts of summation variables such as the integral oil production are less sensitive to the dynamics than initial variables and parameters. Nevertheless, in principle there are improvements to be gained in recognizing the time-sequentiality.

7.2 The RTS smoother

The RTS smoother, which computes $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:T})$, is a marginal smoother, in contrast to the augmented filters, which are all joint. It is first expressed in the most general, Bayesian sense. It is then formulated with linear-Gaussian assumptions, and lastly using the ensemble. A more introductory presentation of this framework is presented by Wikle and Berliner [208].

The outline of the algorithm, in any formulation, is as follows: The forward pass of the RTS smoother is just the filter, detailed in section 1.2.3 and chapter 2. The subsequent backward pass then performs the conditioning on the future observations, $\boldsymbol{y}_{t+1:T}$. The backward pass is recursive, for decreasing time index t, so that T can be considered fixed (whereby the "off-line" attribute).

Anticipating the formal description, Fig. (7.1a) illustrates the progression of the conditioning of the RTS smoother. It should be contrasted to that of the augmented filters (section 7.3) in Fig. (7.1b).



(a) RTS smoother (i.e. EnRTS). **(b)** Augmented filter smoothers (i.e. EnKS).

Figure 7.1: Visualization of the processing steps of the smoothing algorithms. Each (non-broken) slab represents all of the conditioning that was performed at a given processing step.[4] The next (resp. current, previous) processing step is coloured green (resp. yellow, red). All steps before that are coloured blue.

[4]Formally, a slab's horizontal (resp. vertical) extent symbolizes the difference, in the time indices on $\boldsymbol{x}$ (resp. $\boldsymbol{y}$) from one processing step to the next.

7.2.1 The general formulation

The backward pass of the RTS smoother computes $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:T})$ for sequentially decreasing t. Similarly to how the Chapman-Kolmogorov equation (1.12) forecasts the filtered distribution, the following "hindcasts" the smoothed distribution:

$$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:T}) = \int p(\boldsymbol{x}_t|\boldsymbol{x}_{t+1}, \boldsymbol{y}_{1:T})\, p(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:T})\, d\boldsymbol{x}_{t+1}\,, \tag{7.1}$$

as follows from the law of total probability. But with $\boldsymbol{x}_{t+1}$ known, $\boldsymbol{x}_t$ does not depend on $\boldsymbol{y}_{t+1:T}$, i.e. $p(\boldsymbol{x}_t|\boldsymbol{x}_{t+1}, \boldsymbol{y}_{1:T}) = p(\boldsymbol{x}_t|\boldsymbol{x}_{t+1}, \boldsymbol{y}_{1:t})$. This is intuitive looking at Fig. 1.1, and can also be shown inductively using eqns. (1.4) and (1.5).

Still, the direction of the conditioning, $\boldsymbol{x}_t|\boldsymbol{x}_{t+1}$, is non-causal; it is inverted using Bayes' rule:

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t+1}, \boldsymbol{y}_{1:t}) = \frac{p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)\, p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})}{p(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t})}\,. \tag{7.2}$$

where the denominator is *not* a constant, and cannot be neglected. Note that all of the distributions required to compute $p(\boldsymbol{x}_t|\boldsymbol{x}_{t+1}, \boldsymbol{y}_{1:t})$ from eqns. (7.1) and (7.2) are known either from the recursion or from the pre-run filter. More details are provided by the references [108, 179].

7.2.2 The linear-Gaussian formulation

For any two time indices, t_1, t_2, define

$$\boldsymbol{\mu}_{t_1|t_2} = \mathbb{E}(\boldsymbol{x}_{t_1} \mid \boldsymbol{y}_{1:t_2})\,, \tag{7.3}$$

$$\mathbf{P}_{t_1|t_2} = \mathbb{V}\mathrm{ar}(\boldsymbol{x}_{t_1} \mid \boldsymbol{y}_{1:t_2})\,. \tag{7.4}$$

where $\mathbb{E}(.)$ and $\mathbb{V}\mathrm{ar}(.)$ are the (multivariate) expectation and variance operators. Note that the conditioning subscript abbreviates an index *range*.

Assume that f is linear, $f : \boldsymbol{x} \mapsto \mathbf{F}\boldsymbol{x}$. The backward pass of the linear-Gaussian RTS smoother is derived from eqns. (7.1) and (7.2) as shown in appendix D. It is given by the following recursion, for decreasing t:

$$\mathbf{J}_t = \mathbf{P}_{t|t}\mathbf{F}^\mathsf{T}\mathbf{P}_{t+1|t}^{-1}\,, \tag{7.5}$$

$$\boldsymbol{\mu}_{t|T} = \boldsymbol{\mu}_{t|t} + \mathbf{J}_t\Big[\boldsymbol{\mu}_{t+1|T} - \boldsymbol{\mu}_{t+1|t}\Big]\,, \tag{7.6}$$

$$\mathbf{P}_{t|T} = \mathbf{P}_{t|t} - \mathbf{J}_t\Big[\mathbf{P}_{t+1|t} - \mathbf{P}_{t+1|T}\Big]\mathbf{J}_t^\mathsf{T}\,. \tag{7.7}$$

By subtracting $\boldsymbol{\mu}_{t|t}$ from eqn. (7.6) it is seen that the role of the "backward gain" matrix, $\mathbf{J}_t$, is to map the mean increment (smoother − filter) at time $t+1$ to that of time t. Furthermore, consider the strict separation of the conditioning information: $\left[\boldsymbol{\mu}_{t+1|T} - \boldsymbol{\mu}_{t+1|t}\right]$ contains all of the information from $\boldsymbol{y}_{t+1:T}$, while $\boldsymbol{\mu}_{t|t}$ already contains the information from $\boldsymbol{y}_{1:t}$. This is also reflected in the covariance equation (7.7).

Using the simplified notation $\boldsymbol{u} = \boldsymbol{x}_t|\boldsymbol{y}_{1:t}$ and $\boldsymbol{v} = \mathbf{F}\boldsymbol{u}$, one can identify $\mathbf{J}_t$ of eqn. (7.5) as

$$\mathbf{J}_t = \mathbb{C}\mathrm{ov}(\boldsymbol{u}, \boldsymbol{v}) \left(\mathbb{V}\mathrm{ar}(\boldsymbol{v})\right)^{-1} . \tag{7.8}$$

This is a familiar formula, recognizable from the Kalman filter and Simple Kriging [14]; it is the coefficient matrix of the best linear unbiased estimator for $\boldsymbol{u}$ conditioned on $\boldsymbol{v}$. In other words, $\mathbf{J}_t$ is a statistical linearization of the relationship from $\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t}$ to $\boldsymbol{x}_t|\boldsymbol{y}_{1:t}$. The understanding gained from the principles of (i) no overlapping information and (ii) statistical linearization, together motivate the ensemble formulation.

7.2.3 The ensemble formulation

The conditioning subscript notation of section 7.2.2 is extended to apply to ensembles: $\mathbf{E}_{t_1|t_2}$ is the ensemble estimating $\boldsymbol{x}_{t_1}|\boldsymbol{y}_{1:t_2}$, and $\mathbf{A}_{t_1|t_2} = \mathbf{E}_{t_1|t_2}\boldsymbol{\Pi}_{\mathbb{1}}^{\perp}$ are the corresponding anomalies.

7.2.3.1 The forward filtering pass

As mentioned, the forward pass of the RTS smoother is just the filter, meaning the forecasts, and the analyses. Here, for convenience, the analysis formula (2.16) is abbreviated as

$$\mathbf{E}_{T|T} = \mathbf{E}_{T|T-1} + \mathbf{A}_{T|T-1}\boldsymbol{\Gamma}_T \,, \tag{7.9}$$

for increasing T, where $\mathbf{A}_{T|T-1} = \mathbf{E}_{T|T-1}\boldsymbol{\Pi}_{\mathbb{1}}^{\perp}$ and

$$\begin{aligned}\boldsymbol{\Gamma}_T = \mathbf{Y}_T^{\mathsf{T}} \Big(\mathbf{Y}_T\mathbf{Y}_T^{\mathsf{T}} + (N-1)\mathbf{R}\Big)^{-1} \\ \Big\{\boldsymbol{y}_T\mathbb{1}^{\mathsf{T}} - h(\mathbf{E}_{T|T-1}) - \text{perturbations}\Big\} \,.\end{aligned} \tag{7.10}$$

with $h(\mathbf{E}_{T|T-1})$ the observed forecast ensemble, and $\mathbf{Y}_T = h(\mathbf{E}_{T|T-1})\boldsymbol{\Pi}_{\mathbb{1}}^{\perp}$ its anomalies. Note that square root versions [24, 97, 158] of the EnKFs can also be written as eqn. (7.9), with a different $\boldsymbol{\Gamma}_T$ [47].

7.2.3.2 The backward smoothing pass

The backward pass of the EnRTS consists of the sequential application, for decreasing t, of the formula

$$\mathbf{E}_{t|T} = \mathbf{E}_{t|t} + \bar{\mathbf{J}}_t \left[\mathbf{E}_{t+1|T} - \mathbf{E}_{t+1|t} \right] , \tag{7.11}$$

with $\bar{\mathbf{J}}_t$ specified below. Equation (7.11) can be (i) motivated by the discussion in section 7.2.2, (ii) postulated by the leap of faith "replace the true moments by their ensemble estimates", and (iii) justified by showing its conformality to the linear-Gaussian equations. Alternatively, it can be derived in the manner of section 6.2.

As with the filter, eqn. (7.11) is an application of an ensemble-estimated gain to each individual ensemble member. Note that no "perturbations" need be furnished, as the conditioning object, $\boldsymbol{x}_{t+1|T}$, is already represented by an ensemble: $\mathbf{E}_{t+1|T}$. The ensemble-estimated gain matrix is derived from eqn. (7.8):

$$\bar{\mathbf{J}}_t = \mathbf{A}_{t|t}\mathbf{A}_{t+1|t}^{\mathsf{T}} \left(\mathbf{A}_{t+1|t}\mathbf{A}_{t+1|t}^{\mathsf{T}} \right)^{+} \tag{7.12}$$

$$= \mathbf{A}_{t|t}\mathbf{A}_{t+1|t}^{+} . \tag{7.13}$$

The second line follows from Lemma 5.1. The pseudoinverse is typically computed through the singular value decomposition (SVD). For efficiency, only the reduced SVD should be computed (Definition E.2).

7.2.3.3 Potential improvements

Conformality to the linear-Gaussian relations consists in

$$\bar{\mathbf{P}}_{t|T} = \bar{\mathbf{P}}_{t|t} - \bar{\mathbf{J}}_t \left[\bar{\mathbf{P}}_{t+1|t} - \bar{\mathbf{P}}_{t+1|T} \right] \bar{\mathbf{J}}_t^{\mathsf{T}} \tag{7.14}$$

for the covariance, and can be shown along the lines of section 2.4.1. If $N > m$, however, then eqn. (7.14) will not be exactly satisfied. This can be rectified by using a square root scheme, factoring out $\mathbf{A}_{t|t}$ on both edges of the right hand side of eqn. (7.14), and does not appear to have been mentioned yet in the ensemble literature, though it is inherent in the unscented formulation.[5] In any case, it is not clear if significant benefits could be derived from such a scheme.

[5] See footnote 2 of section 7.1.

The EnRTS is afflicted by the problem that the gain is computed using the same ensemble that it updates, just as with the EnKF [94], illustrated in Fig. (7.1b). Addressing this issue, Frei [69] found that resampling the smoothed ensemble yielded better results.

If model noise is being accounted for by adding simulated noise to the ensemble, one might consider computing $\bar{\mathbf{J}}$ before this addition, using the exact noise covariance matrix:

$$\tilde{\mathbf{J}}_t = \mathbf{A}_{t|t}\tilde{\mathbf{A}}^\mathsf{T}\left(\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\mathsf{T} + (N-1)\mathbf{Q}\right)^+ , \tag{7.15}$$

where $\tilde{\mathbf{A}}$ is the anomalies of the model-propagated ensemble before the addition of the simulated noise. This might be beneficial since it reduces the sources of sampling error in $\bar{\mathbf{J}}$. Another alternative is to use methods that reduce sampling error when accounting for model noise, which is the topic chapter 5. Similar considerations were discussed for the EnKS by Cosme et al. [47] and Nerger et al. [146]. We will not consider $\tilde{\mathbf{J}}_t$ any further.

7.3 The EnKS

The general formulation of this smoother was given in section 1.2.2. The linear-Gaussian formulation consists in applying the KF, section 2.1, to the augmented state, $\boldsymbol{x}_{0:T}$, and dismantling the resulting block matrices. The ensemble formulation, the EnKS, is just an extension of the EnKF analysis, eqn. (7.9), to all of the previous states as well: the observations at time T are assimilated into the ensemble of time t by multiplying with the appropriate anomalies:

$$\mathbf{E}^{\mathrm{KS}}_{t|T} = \mathbf{E}^{\mathrm{KS}}_{t|T-1} + \mathbf{A}^{\mathrm{KS}}_{t|T-1}\boldsymbol{\Gamma}_T \,. \tag{7.16}$$

Note that this means that the filter ensembles, $\mathbf{E}_{T|T}$ and $\mathbf{E}_{T+1|T}$, are the same as for the EnRTS for any T; only for smoothed ensembles, $\mathbf{E}_{t|T}$ with $T > t$, is it necessary to distinguish between the two methods using the "KS" superscript as in eqn. (7.16). An essential characteristic of the EnKS is that it is on-line: eqn. (7.16) is recursive for increasing T, trivially updating all of the previous states with the incoming observations. Equation (7.16) may be applied in an inner loop for any $t = 0, \ldots, T$ or any subset thereof.

The "$\mathbf{X}_5$" expression of Evensen [58, his eqn. 104] provides the means for writing eqn. (7.16) without the explicit forward recursion, namely

$$\mathbf{E}^{\mathrm{KS}}_{t|T} = \mathbf{E}^{\mathrm{KS}}_{t|T-1}\left(\mathbf{I}_N + \mathbf{\Pi}^{\perp}_{\mathbb{1}}\mathbf{\Gamma}_T\right) \tag{7.17}$$

$$= \mathbf{E}^{\mathrm{KS}}_{t|t}\prod_{\tau=t+1}^{T}\left(\mathbf{I}_N + \mathbf{\Gamma}_\tau\right)\,. \tag{7.18}$$

The projection "disappears" because it is already implicit in $\mathbf{\Gamma}_t$. The sequential product is to be understood as proceeding from left to right. Note that

$$\prod_{\tau=t}^{T}\mathbf{M}_\tau = \mathbf{M}_t\prod_{\tau=t+1}^{T}\mathbf{M}_\tau \tag{7.19}$$

for any sequence of square matrices, and thus eqn. (7.18) can be rewritten as a *backward* recursion. This led Ravela and McLaughlin [164] to formulate a more efficient implementation of the EnKS for off-line application, whose conditioning is like that of the EnRTS, i.e. Fig. (7.1a).

7.4 Equivalence of the EnKS and the EnRTS

The (linear-Gaussian) RTS and KS are equivalent; this follows from the conformality of marginal and joint Gaussian distributions. This section proves that the EnKS and the EnRTS are equivalent, i.e. that they generate equal ensembles. The main interest lies in the validity of the result even in the nonlinear case.[6] The result requires the very mild condition $N \leq m$, (typically for geoscientific applications, $N \ll m$) but even this can be dropped if the models are linear, tying in with what is to be expected based on the linear-Gaussian case.

[6]Unless otherwise stated, no assumptions of linearity of f or h are made. It is, however, assumed that they do not introduce (additional) linear dependence in an ensemble (when applied to each ensemble member). This assumption is needed in order to maintain the rank of the ensemble, similarly to Lemma 2.1 (which is based purely on sampling theory). Note that the assumption implies injectivity, but the converse is not true. Injectivity (being one-to-one) should hold in principle with continuous dynamics (or observing processes). However, if the dynamics are approximated using explicit time stepping schemes, then the dynamics are not injective. Nevertheless, in practice, for short time spans, it seems highly unlikely that the dynamics actually render different ensemble members linearly dependent, thus upholding the results dependent on this property.

Lemma 7.1 – The EnRTS as an on-line smoother**.**
Unconditionally,

$$\mathbf{E}_{t|T} = \mathbf{E}_{t|t} + \sum_{k=t+1}^{T} \left(\prod_{\tau=t}^{k-1} \bar{\mathbf{J}}_\tau \right) \left[\mathbf{E}_{k|k} - \mathbf{E}_{k|k-1} \right] . \tag{7.20}$$

Proof by induction, for decreasing t. Omitted.

What is remarkable about Lemma 7.1 is that it provides a formulation of the EnRTS estimates conveniently applied sequentially for increasing T, for any t, just as eqn. (7.16) is applied for the EnKS. Indeed, subtracting eqn. (7.20) by its previous T iterate yields the increment at time t due to the incoming observations:

$$\mathbf{E}_{t|T} - \mathbf{E}_{t|T-1} = \left(\prod_{\tau=t}^{T-1} \bar{\mathbf{J}}_\tau \right) \left[\mathbf{E}_{T|T} - \mathbf{E}_{T|T-1} \right] , \tag{7.21}$$

thus showing how the EnRTS ensembles can be computed in an on-line way. It should be appreciated that this is possible, without repeating all of the previous calculations, only because of the linearity of the recursion (7.11).

By comparing eqn. (7.21) with eqn. (7.11), notice that $\prod_\tau \bar{\mathbf{J}}_\tau$ appears to extend the role of $\bar{\mathbf{J}}_t$ to span multiple time indices. In recognition of this, define

$$\bar{\mathbf{J}}_{t \lhd T} = \prod_{\tau=t}^{T-1} \bar{\mathbf{J}}_\tau \, , \tag{7.22}$$

as a generalization of $\bar{\mathbf{J}}_t$. It is clear that $\bar{\mathbf{J}}_{t \lhd t+1} = \bar{\mathbf{J}}_t$. Building on eqn. (7.21), however, the following Lemma extends this identity, thus substantiating the generalization.

Lemma 7.2 – Understanding $\bar{\mathbf{J}}_{t \lhd T}$**.**
If $N \leq m$ or f is linear, then

$$\mathbf{A}_{t|T-1} = \bar{\mathbf{J}}_{t \lhd T} \mathbf{A}_{T|T-1} \, , \tag{7.23}$$

and thus

$$\bar{\mathbf{J}}_{t \lhd T} = \mathbf{A}_{t|T-1} \mathbf{A}^{+}_{T|T-1} \, . \tag{7.24}$$

Proof. Multiply eqn. (7.21) on the right by $\mathbf{\Pi}^{\perp}_{\mathbb{1}}$ and rearrange to obtain

$$\mathbf{A}_{t|T} - \bar{\mathbf{J}}_{t \lhd T} \mathbf{A}_{T|T} = \mathbf{A}_{t|T-1} - \bar{\mathbf{J}}_{t \lhd T} \mathbf{A}_{T|T-1} \, . \tag{7.25}$$

The right hand side is zero by the inductive assumption, eqn. (7.23), and therefore

$$\mathbf{A}_{t|T} - \bar{\mathbf{J}}_{t\lhd T}\mathbf{A}_{T|T} = 0\,. \tag{7.26}$$

But the null spaces of $\mathbf{A}_{T|T}$ and $\mathbf{A}_{T+1|T}$ coincide if $N \leq m$ or if f is linear.[7] Therefore $\mathbf{A}_{T|T} = \mathbf{A}_{T|T}\mathbf{\Pi}_{\mathbf{A}^\intercal_{T+1|T}}$ or, in terms of the pseudoinverse,

$$\mathbf{A}_{T|T} = \mathbf{A}_{T|T}\mathbf{A}^+_{T+1|T}\mathbf{A}_{T+1|T}\,. \tag{7.27}$$

Substituting this into eqn. (7.26) yields

$$\mathbf{A}_{t|T} = \bar{\mathbf{J}}_{t\lhd T}\mathbf{A}_{T|T}\mathbf{A}^+_{T+1|T}\mathbf{A}_{T+1|T}\,. \tag{7.28}$$

But from the definitions (7.13) and (7.22),

$$\bar{\mathbf{J}}_{t\lhd T+1} = \bar{\mathbf{J}}_{t\lhd T}\mathbf{A}_{T|T}\mathbf{A}^+_{T+1|T}\,, \tag{7.29}$$

and therefore eqn. (7.28) reduces to

$$\mathbf{A}_{t|T} = \bar{\mathbf{J}}_{t\lhd T+1}\mathbf{A}_{T+1|T}\,, \tag{7.30}$$

which is the inductive assumption, eqn. (7.23), at the next T iterate.

Equation (7.24) follows from eqn. (7.23) by the identity $\mathbf{M}^+\mathbf{M}\mathbf{M}^+ = \mathbf{M}^+$. □

Equation (7.23) is the central inductive result, but the offshoot, eqn. (7.24), is more intuitive. With eqn. (7.24), $\bar{\mathbf{J}}_{t\lhd T}$ can now be understood as a generalization of $\bar{\mathbf{J}}_t$, not just from its role in eqn. (7.21), but by direct comparison to the definition, eqn. (7.13). As discussed in relation to eqn. (7.8), eqn. (7.24) also lets us identify $\bar{\mathbf{J}}_{t\lhd T}$ as the best gain for the increments at time t based on increments at time T, conditioned on observations up to $T-1$.

Theorem 7.1 – Equivalence**.**
If $N \leq m$, or f is linear, then the EnRTS and the EnKS are equivalent:

$$\mathbf{E}_{t|T} = \mathbf{E}^{\mathrm{KS}}_{t|T}\,. \tag{7.31}$$

[7] If $N \leq m$, then the null space of both $\mathbf{A}_{T|T}$ and $\mathbf{A}_{T+1|T}$ is the span of $\mathbb{1}$ because f does not introduce additional linear dependence (footnote 6, page 103).

Alternatively, (regardless of N) if f is linear, say $f : \boldsymbol{z} \mapsto \mathbf{F}\boldsymbol{z}$, then $\mathbf{F}\boldsymbol{z} = 0$ only if $\boldsymbol{z} = 0$ due to injectivity. I.e. $\mathbf{F}\mathbf{A}\boldsymbol{x} = 0$ only if $\boldsymbol{x}$ is in the null space of $\mathbf{A}$. The opposite inclusion is trivial.

Proof. Starting from eqn. (7.21), the innovations can be inserted using eqn. (7.9):

$$\mathbf{E}_{t|T} = \mathbf{E}_{t|T-1} + \bar{\mathbf{J}}_{t\lhd T}\mathbf{A}_{T|T-1}\mathbf{\Gamma}_T \,. \tag{7.32}$$

Applying eqn. (7.23) yields

$$\mathbf{E}_{t|T} = \mathbf{E}_{t|T-1} + \mathbf{A}_{t|T-1}\mathbf{\Gamma}_T \,. \tag{7.33}$$

The inductive hypothesis then yields the "KS" superscripts. □

There is a subtlety about the equivalence worthwhile expounding on: the proof is an induction for increasing T, involving the associated $\mathbf{E}_{t|T}$. But these ensembles are not actually computed by the EnRTS, unless T is the endpoint (the "final") time index. However, by virtue of Lemma 7.1, they are implicitly but consistently defined, and thus valid intermediaries in proving the equality of the endpoint ensemble.

One scenario that has not yet been explicitly considered is when observations are absent for one or more time indices. It might then be tempting to think that the cascading, marginal EnRTS, might do better than the EnKS since it traverses the non-observed interval one step at a time. However, as seen directly from the definitions (7.13) and (7.22), removing the observations for some interval $\tau \in \{t+1, \ldots, T-1\}$ merely yields

$$\bar{\mathbf{J}}_{t\lhd T} = \mathbf{A}_{t|t}\left(\prod_{\tau=t+1}^{T-1} \mathbf{\Pi}_{\mathbf{A}^{\mathsf{T}}_{\tau|t}}\right)\mathbf{A}^{+}_{T|t} \,, \tag{7.34}$$

which reduces to $\bar{\mathbf{J}}_{t\lhd T} = \mathbf{A}_{t|t}\mathbf{A}^{+}_{T|t}$ under the same conditions as for Lemma 7.2. In other words, just as for the EnKS, the communication of the conditioning information backward across the interval is not influenced (and therefore not improved) by the intermediate time steps.

7.5 In practice

Note that most of the developments of section 7.4 are concerned only with the backward pass of the EnRTS; Lemmas 7.1 and 7.2 make no mention of the filtering analysis or the EnKS, and their conclusions are independent of how these updates are carried out. Ultimately, however, Theorem 7.1 also depends on the EnKF and EnKS updates, eqns. (7.9) and (7.16). Slight deviations from these forms is the reason why equality is typically not observed in practice. For example, a localized analysis update generally cannot be written in the form of eqn. (7.9), invalidating the equivalence.

Similarly, though it is possible to include post-analysis inflation in the form of eqn. (7.9) by including it in $\mathbf{\Gamma}_T$, this alteration[8] cannot be used for the EnKS smoothing updates (7.16), because applying the same inflation to the smoothed ensembles would result in deleterious, compounded inflation [107]. Therefore post-analysis inflation destroys the equivalence.

Vice-versa it becomes apparent that the post-analysis inflation of the forward pass of the EnRTS will compound during the backward pass, necessitating a compensating mechanism of some form. One possibility is to use *pre*-analysis inflation instead; this will not affect the equivalence, since it can be interpreted as a part of the forecast model, f, which has hitherto been considered as a black box. Another possibility is to multiply $\bar{\mathbf{J}}_t$ with a "deflation factor", say $(1-\delta)$, where $0 \leq \delta \ll 1$. In practice, this will be beneficial anyhow, since it acts to localize the smoothing in time. The δ parameter should then be tuned similarly to how the lag length is tuned in the EnKS [47, 146]. Based on a geometric series analysis (and neglecting inflation) one would expect the effective lag length to be on the order of $1/\delta$ time steps.

Figure 7.2 plots the performance of the smoothers for a range of different forcing parameters, F, for the Lorenz-96 system. The experimental setup is described in chapter 3, and the tuned inflation, deflation and lag lengths are listed in Table F.3. Here, no model noise is used (i.e. $\mathbf{Q}$ is zero), and $\Delta t_{\text{obs}} = 0.15$. Using the ensemble size $N = 25$, but no localization, it is necessary to use inflation in the filtering. As explained above, this destroys the equivalence; still, there is no particular reason to suspect that either method should do better than the other. This is supported by the experimental results: there is very little to differentiate the performances of the optimally tuned EnKS and EnRTS. Further experiments (not shown) with different systems support these findings. The slightly superior scores of EnRTS for small F

[8] However, this effective alteration is more complicated than merely a scalar multiplication of $\mathbf{\Gamma}_T$ by the inflation factor.

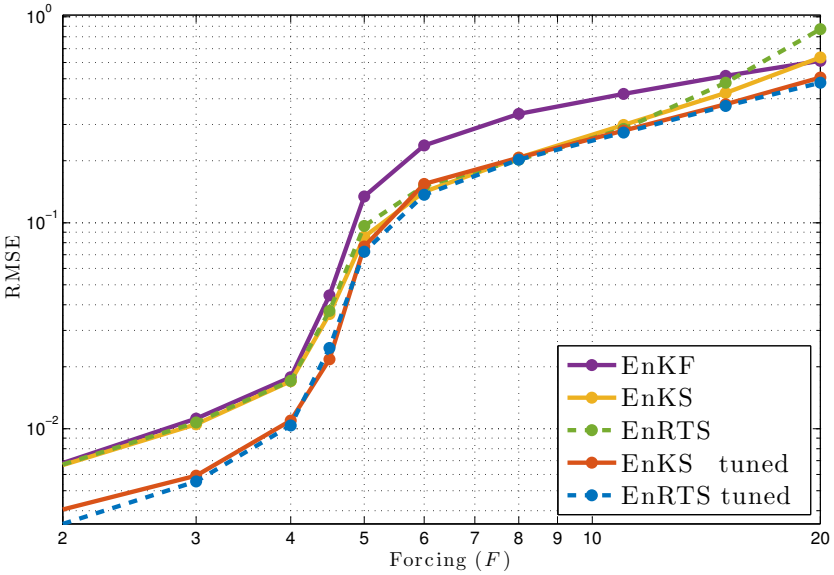


Figure 7.2: Performance benchmarks as a function of the forcing parameter, F, obtained with the Lorenz-96 system. The EnKS (resp. EnRTS) "tuned" uses a lag (resp. deflation) setting that has been tuned for each individual value of F. By contrast, the yellow and green curves use the same setting for all values of F, namely a lag of 2 and a deflation factor of 0.99 which – as can be seen from the intersection of the curves in the figure – are optimal for $F = 8$.

might be explained by the fact that the deflation approach yields smoothly decaying correlations, while the lag approach is a sharp cut-off.

Let us summarize the relevant considerations in choosing between the smoothers. (i) On the one hand, the EnRTS should be more recognizable for those familiar with linear-Gaussian state estimation. On the other, the derivation of the EnKS is remarkably simple. (ii) Previously it may have seemed that if the smoothing application is on-line, then one must necessarily use the EnKS. However, Lemma 7.1 showed that the EnRTS could be applied in the same fashion. Vice-versa, it might have been argued that if the optimal lag length is long, then the EnRTS will be less computationally costly, since it only updates the ensembles once. However, it has long been known that the EnKS can be formulated in the backward recursive way as well, as discussed in relation to eqn. (7.19). (iii) Tuning the deflation parameter is no more or less complicated than tuning the lag length for the EnKS. (iv) It thus seems that the main point of differentiation between the algorithms is that the EnRTS is slightly more complicated in operation than the EnKS, requiring the storage of the forecast ensemble during the forward run, and the computation of its reduced SVD for the pseudoinverse.

7.6 Summary

This chapter has shed some new light on the EnRTS. Consideration was given to how it works, its suboptimality for $N > m$, and a possible avenue for improvement (eqn. (7.15)). A sequential formulation of the smoother has been derived (Lemma 7.1) and used to understand how information communicates across multiple time indices (Lemma 7.2). These developments made it possible to show theoretical equivalence between the EnRTS and the EnKS (Theorem 7.1), a non-trivial result that surprisingly is valid even in the presence of nonlinearity. In practice equivalence will typically not hold due to inflation and localization, yet numerical experiments illustrated that the optimally tuned performances of both methods are very close. This leaves the slightly more complicated operation of the EnRTS as the main difference between the smoothers.

Chapter 8

Conclusion

This thesis has studied the EnKF and the related smoothers. After discussing the DA problem, the EnKF was introduced; taking the traditional approach, its presentation is intended to be accessible, broad, and precise. Next, the square root approach was investigated for the treatment of additive model noise; though only yielding modest improvements at the expense of incurring additional computation cost, the investigation fills a theoretical void, and it also prompted a review of the properties of the square root approach in general. Subsequently, the hierarchical framework of the EnKF-N was developed and then reverse engineered as a scale mixture; it is hoped that the additional detail and the original perspective are useful in understanding the method. Lastly, it was shown that two different ensemble smoothing algorithms are equivalent; this means that the choice between the two can be guided solely by practical considerations. More detailed summaries and discussions are given at the end of the respective chapters.

One enticing avenue for future research is the further development of the EnKF-N, in particular its hybridization with methods to estimate model error, and its potential to make localization techniques rigorous. Another intriguing avenue is the development of theory formalizing the link, in twin experiments, between the RMSE point metric and the distribution metrics. On a broader scale, the DA problem and the ensemble methods are academically interesting because they combine several technical aspects with a probabilistic, mathematical framework. Thus, future work in the field of ensemble methods will continue to advance the general practicality of real-world DA applications.

Appendix A

The Kalman filter in detail

This appendix complements section 2.1 by filling in the details of the derivation of the KF, as well as providing better understanding of its significance.

A.1 Matrix inversion identities

Matrix algebra is complicated by the fact that matrix multiplication is not commutative, and matrices not generally invertible. It is therefore necessary to catalogue some identities.

Lemma A.1 – Woodbury identity**.**
For any $m, p \in \mathbb{N}_+$*, any invertible* $\mathbf{B} \in \mathbb{R}^{m^2}$ *and* $\mathbf{R} \in \mathbb{R}^{p^2}$*, and any* $\mathbf{V}, \mathbf{U} \in \mathbb{R}^{p \times m}$ *such that* $\mathbf{V}^\mathsf{T}\mathbf{R}^{-1}\mathbf{U} + \mathbf{B}^{-1}$ *is invertible,*

$$\left(\mathbf{V}^\mathsf{T}\mathbf{R}^{-1}\mathbf{U} + \mathbf{B}^{-1}\right)^{-1} = \mathbf{B} - \mathbf{B}\mathbf{V}^\mathsf{T}\left(\mathbf{R} + \mathbf{U}\mathbf{B}\mathbf{V}^\mathsf{T}\right)^{-1}\mathbf{U}\mathbf{B}\,. \tag{A.1}$$

Proof. We will contend with a direct proof, multiply the inverse of the left hand side by the right hand side.

$$\begin{aligned}
&\left(\mathbf{B}^{-1} + \mathbf{V}^\mathsf{T}\mathbf{R}^{-1}\mathbf{U}\right)\left[\mathbf{B} - \mathbf{B}\mathbf{V}^\mathsf{T}\left(\mathbf{R} + \mathbf{U}\mathbf{B}\mathbf{V}^\mathsf{T}\right)^{-1}\mathbf{H}\mathbf{P}\right] \\
&\quad = \mathbf{I}_m + \mathbf{V}^\mathsf{T}\mathbf{R}^{-1}\mathbf{H}\mathbf{P} - (\mathbf{V}^\mathsf{T} + \mathbf{V}^\mathsf{T}\mathbf{R}^{-1}\mathbf{U}\mathbf{B}\mathbf{V}^\mathsf{T})(\mathbf{R} + \mathbf{U}\mathbf{B}\mathbf{V}^\mathsf{T})^{-1}\mathbf{H}\mathbf{P} \\
&\quad = \mathbf{I}_m + \mathbf{V}^\mathsf{T}\mathbf{R}^{-1}\mathbf{H}\mathbf{P} - \mathbf{V}^\mathsf{T}\mathbf{R}^{-1}(\mathbf{R} + \mathbf{U}\mathbf{B}\mathbf{V}^\mathsf{T})(\mathbf{R} + \mathbf{U}\mathbf{B}\mathbf{V}^\mathsf{T})^{-1}\mathbf{H}\mathbf{P} \\
&\quad = \mathbf{I}_m + \mathbf{V}^\mathsf{T}\mathbf{R}^{-1}\mathbf{H}\mathbf{P} - \mathbf{V}^\mathsf{T}\mathbf{R}^{-1}\mathbf{H}\mathbf{P} = \mathbf{I}_m
\end{aligned}$$

□

This identity, also known as the matrix inversion lemma, is significant because it shows that the inversion of a "rank-p update" ($\mathbf{V}^\mathsf{T}\mathbf{R}^{-1}\mathbf{U}$) to a rank-$m$ matrix

($\mathbf{B}^{-1}$), whose inverse ($\mathbf{B}$) is already known, only requires computing a $p \times p$ inverse. Although the identity had been derived earlier through the inversion of a block matrix, this significance was only fully recognized when the identity was obtained as a generalization on rank-1 updated inverses [81]. Here, Lemma A.1 is used through Corollaries A.1 and A.2, both of which embody the useful change of dimensionality of the inversion.

Corollary A.1.
For any symmetric, positive-definite $\mathbf{R}$ *and* $\mathbf{B}$*, and* $\mathbf{U} = \mathbf{V} = \mathbf{H}$*, the invertibility conditions in Lemma A.1 are already satisfied, and*

$$\left(\mathbf{H}^\mathsf{T}\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1}\right)^{-1} = \mathbf{B} - \mathbf{B}\mathbf{H}^\mathsf{T}\left(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^\mathsf{T}\right)^{-1}\mathbf{H}\mathbf{B} \tag{A.2}$$

Proof. For all $\boldsymbol{x} \in \mathbb{R}^m$, $\boldsymbol{x}^\mathsf{T}\mathbf{B}^{-1}\boldsymbol{x} > 0$ and $\boldsymbol{x}^\mathsf{T}\mathbf{H}^\mathsf{T}\mathbf{R}^{-1}\mathbf{H}\boldsymbol{x} \geq 0$ implies $\boldsymbol{x}^\mathsf{T}(\mathbf{H}^\mathsf{T}\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1})\boldsymbol{x} > 0$. Thus all eigenvalues of this and the other SPD matrices are positive, and they are all invertible. □

Corollary A.2.
For the same matrices as in Corollary A.1,

$$\left(\mathbf{H}^\mathsf{T}\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1}\right)^{-1}\mathbf{H}^\mathsf{T}\mathbf{R}^{-1} = \mathbf{B}\mathbf{H}^\mathsf{T}\left(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^\mathsf{T}\right)^{-1}. \tag{A.3}$$

Proof. A straightforward validation of eqn. (A.3) is obtained by "cross-multiplying" out the inverses, but a more satisfying exercise is to derive it from eqn. (A.2), starting by right-multiplying by $\mathbf{H}^\mathsf{T}$. □

A.2 Kalman filter derivation

Our derivation of the KF consists in applying the linearity assumptions of section 2.1 to the Bayesian filtering equations, (1.12) and (1.13). Focusing on one particular time index, t, assume that initially $\boldsymbol{x}_{t-1}|\boldsymbol{y}_{1:t-1} \sim \mathcal{N}(\boldsymbol{x}^a, \mathbf{P}^a)$, where the moments, $\boldsymbol{x}^a$ and $\mathbf{P}^a$ are known.

A.2.1 Forecast step

The linear forecast process, $\boldsymbol{x}_t = \mathbf{F}\boldsymbol{x}_{t-1} + \boldsymbol{q}_{t-1}$, amounts to a summation of Gaussian random variables. It therefore maintains the Gaussianity, and we can write

$$\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1} \sim \mathcal{N}(\boldsymbol{x}^f, \mathbf{P}^f)\,. \tag{A.4}$$

where the forecast moments, $\boldsymbol{x}^f$ and $\mathbf{P}^f$, remain to be derived. This may be done without any reference to Gaussianity; using the law of total expectation, the forecast step of the KF becomes:

$$\begin{aligned}\boldsymbol{x}^f &= \mathbb{E}\left(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1}\right) = \mathbb{E}\Big(\mathbb{E}(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1}, \boldsymbol{x}_{t-1})|\boldsymbol{y}_{1:t-1}\Big) &\text{(A.5)}\\ &= \mathbb{E}\Big(\mathbb{E}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})|\boldsymbol{y}_{1:t-1}\Big) = \mathbb{E}(\mathbf{F}\boldsymbol{x}_{t-1}|\boldsymbol{y}_{1:t-1}) = \mathbf{F}\,\mathbb{E}(\boldsymbol{x}_{t-1}|\boldsymbol{y}_{1:t-1}) = \mathbf{F}\boldsymbol{x}^a\,. &\text{(A.6)}\end{aligned}$$

Similarly, the law of total variance is used in the following:

$$\begin{aligned}\mathbf{P}^f &= \mathbb{V}\mathrm{ar}\left(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1}\right) = \mathbb{V}\mathrm{ar}\Big(\mathbb{E}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})|\boldsymbol{y}_{1:t-1}\Big) + \mathbb{E}\Big(\mathbb{V}\mathrm{ar}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})|\boldsymbol{y}_{1:t-1}\Big) &\text{(A.7)}\\ &= \mathbb{V}\mathrm{ar}\Big(\mathbf{F}\boldsymbol{x}_{t-1}|\boldsymbol{y}_{1:t-1}\Big) + \mathbb{E}\Big(\mathbf{Q}|\boldsymbol{y}_{1:t-1}\Big) = \mathbf{F}\mathbf{P}^a\mathbf{F}^\mathsf{T} + \mathbf{Q}\,. &\text{(A.8)}\end{aligned}$$

Alternatively eqn. (A.4) can be derived by inserting the Gaussian pdfs $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ and $p(\boldsymbol{x}_{t-1}|\boldsymbol{y}_{1:t-1})$ in the Chapman-Kolmogorov equation (1.12).

A.2.2 Analysis step

As was done with the forecast step, it can be shown that the linear observation process, $\boldsymbol{y}_t = \mathbf{H}\boldsymbol{x}_t + \boldsymbol{r}_t$, means that $\boldsymbol{y}_t|\boldsymbol{x}_t \sim \mathcal{N}(\mathbf{H}\boldsymbol{x}_t, \mathbf{R})$. Inserting this and eqn. (A.4) into Bayes' rule (1.13) yields

$$\begin{aligned}-2\log p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t}) &= \|\mathbf{H}\boldsymbol{x}_t - \boldsymbol{y}_t\|^2_{\mathbf{R}} + \left\|\boldsymbol{x}_t - \boldsymbol{x}^f\right\|^2_{\mathbf{P}^f} &\text{(A.9)}\\ &= \|\boldsymbol{x}_t - \boldsymbol{x}^a\|^2_{\mathbf{P}^a} + \text{const.} &\text{(A.10)}\end{aligned}$$

where the second line, including the specifications of $\boldsymbol{x}^a$, and $\mathbf{P}^a$, is given by Lemma A.2, and the constant is with respect to $\boldsymbol{x}_t$. Thus,

$$\boldsymbol{x}_t|\boldsymbol{y}_{1:t} \sim \mathcal{N}(\boldsymbol{x}^a, \mathbf{P}^a)\,, \tag{A.11}$$

and the forecast-analysis cycle can start over again. Lemma A.2 gives the moments, $\boldsymbol{x}^a$ and $\mathbf{P}^a$, in several forms, each with a particular utility. Note that both Corollaries A.1

and A.2 are employed. Since Lemma A.2 is pure matrix algebra, the notation is simplified to

$$\boldsymbol{x}_t = \boldsymbol{x} \qquad \boldsymbol{y}_t = \boldsymbol{y} \qquad \boldsymbol{x}^f = \boldsymbol{b} \qquad \mathbf{P}^f = \mathbf{B} \qquad \boldsymbol{x}^a = \boldsymbol{\mu} \qquad \mathbf{P}^a = \mathbf{P}\,. \tag{A.12}$$

Lemma A.2 – Completing the square**.**
For any $\mathbf{R}$*,* $\mathbf{B}$*, and* $\mathbf{H}$ *as in Corollary A.1,*[1] *and* $\boldsymbol{x}, \boldsymbol{b} \in \mathbb{R}^m$ *and* $\boldsymbol{y} \in \mathbb{R}^p$*,*

$$J = \|\mathbf{H}\boldsymbol{x} - \boldsymbol{y}\|^2_{\mathbf{R}} + \|\boldsymbol{x} - \boldsymbol{b}\|^2_{\mathbf{B}} \;=\; \|\boldsymbol{x} - \boldsymbol{\mu}\|^2_{\mathbf{P}} + \|\boldsymbol{y} - \mathbf{H}\boldsymbol{b}\|^2_{\mathbf{C}}\,. \tag{A.13}$$

where $\boldsymbol{\mu}$*,* $\mathbf{P}$*, and* $\mathbf{C}$ *are given in the derivation below.*

Proof. Expanding the squares of the left hand side of eqn. (A.13), gathering terms, and completing the square in $\boldsymbol{x}$ yields

$$J = \boldsymbol{x}^\mathsf{T}\left(\mathbf{H}^\mathsf{T}\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1}\right)\boldsymbol{x} - 2\boldsymbol{x}^\mathsf{T}\left[\mathbf{H}^\mathsf{T}\mathbf{R}^{-1}\boldsymbol{y} + \mathbf{B}^{-1}\boldsymbol{b}\right] + \|\boldsymbol{b}\|^2_{\mathbf{B}} + \|\boldsymbol{y}\|^2_{\mathbf{R}} \tag{A.14}$$

$$= \|\boldsymbol{x} - \boldsymbol{\mu}\|^2_{\mathbf{P}} - \|\boldsymbol{\mu}\|^2_{\mathbf{P}} + \|\boldsymbol{b}\|^2_{\mathbf{B}} + \|\boldsymbol{y}\|^2_{\mathbf{R}}\,, \tag{A.15}$$

where we have defined

$$\mathbf{P} = (\mathbf{H}^\mathsf{T}\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1})^{-1}\,, \tag{A.16a}$$

$$\boldsymbol{\mu} = \mathbf{P}\left[\mathbf{H}^\mathsf{T}\mathbf{R}^{-1}\boldsymbol{y} + \mathbf{B}^{-1}\boldsymbol{b}\right]\,, \tag{A.16b}$$

which we can identify as the posterior mean and variance by comparison to eqn. (A.10). Numerically, it is better to compute the moments by

$$\mathbf{P} = (\mathbf{B}\mathbf{H}^\mathsf{T}\mathbf{R}^{-1}\mathbf{H} + \mathbf{I}_m)^{-1}\mathbf{B}\,, \tag{A.17a}$$

$$\boldsymbol{\mu} = (\mathbf{B}\mathbf{H}^\mathsf{T}\mathbf{R}^{-1}\mathbf{H} + \mathbf{I}_m)^{-1}\left[\mathbf{B}\mathbf{H}^\mathsf{T}\mathbf{R}^{-1}\boldsymbol{y} + \boldsymbol{b}\right]\,, \tag{A.17b}$$

as this avoids the inversion of $\mathbf{B}^{-1}$. However, if $p < m$, then the Kalman gain form, which inverts a $p \times p$ symmetric matrix, is generally preferable; applying Corollary A.1

[1] With regards to the KF, $\mathbf{P}^f$ can be assumed invertible because the prior, $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1})$, is assumed proper and not degenerate (footnote 3, page 69). This will hold if $\mathbf{F}$ is invertible, and $p(\boldsymbol{x}_{t-1}|\boldsymbol{y}_{1:t-1})$ is proper and not degenerate, etc. On the other hand, $\mathbf{R}$ can be assumed invertible, because zero eigenvalues would imply perfect observations, which is unrealistic, while infinite eigenvalues would correspond to a component of the observation carrying no information at all which can in any case be circumvented by removing this (known) component from the observation (i.e. redefining the observation operator

to eqn. (A.16a) yields eqn. (A.18a)

$$\mathbf{P} = [\mathbf{I}_m - \mathbf{KH}]\mathbf{B}\,, \tag{A.18a}$$
$$\boldsymbol{\mu} = \boldsymbol{b} + \mathbf{K}\left[\boldsymbol{y} - \mathbf{H}\boldsymbol{b}\right]\,, \tag{A.18b}$$

(eqn. (A.18b) is derived further below) where

$$\mathbf{C} = \mathbf{HBH}^\mathsf{T} + \mathbf{R}\,, \tag{A.19a}$$
$$\mathbf{K} = \mathbf{BH}^\mathsf{T}\mathbf{C}^{-1}\,, \tag{A.19b}$$

are the total forecast observation variance and the Kalman gain, respectively. Comparing eqns. A.18 and A.19 to the scalar expressions, eqns. (A.41), it is seen that the Kalman gain is essentially a ratio of covariances. Two useful identities for this ratio,

$$\mathbf{K} = \mathbf{BH}^\mathsf{T}\mathbf{C}^{-1} = \mathbf{PH}^\mathsf{T}\mathbf{R}^{-1}\,, \tag{A.20a}$$
$$\mathbf{I}_m - \mathbf{KH} = \mathbf{PB}^{-1}\,, \tag{A.20b}$$

follow from Corollary A.2 and eqn. (A.18a) respectively. Eqns. (A.20) applied to eqn. (A.16b) can then be used to derive eqn. (A.18b). See Lewis et al. [118, p. 497] and Anderson and Moore [5, p. 147] for further forms of, and discussions on, the KF formulae.

For the purpose of deriving the KF formulae, it is not necessary to develop eqn. (A.15) any further, because $\|\boldsymbol{\mu}\|^2_{\mathbf{P}} + \|\boldsymbol{b}\|^2_{\mathbf{B}} + \|\boldsymbol{y}\|^2_{\mathbf{R}}$ is constant with respect to $\boldsymbol{x}$. However, for the derivation of the RTS smoother, or the finite-size EnKF, the remainder is still of interest. Substituting eqns. (A.16) into eqn. (A.15), expanding, and gathering terms yields

$$J = \|\boldsymbol{x} - \boldsymbol{\mu}\|^2_{\mathbf{P}} - \left\|\mathbf{H}^\mathsf{T}\mathbf{R}^{-1}\boldsymbol{y} + \mathbf{B}^{-1}\boldsymbol{b}\right\|^2_{\mathbf{P}^{-1}} + \|\boldsymbol{b}\|^2_{\mathbf{B}} + \|\boldsymbol{y}\|^2_{\mathbf{R}} \tag{A.21}$$
$$= \|\boldsymbol{x} - \boldsymbol{\mu}\|^2_{\mathbf{P}} + \boldsymbol{y}^\mathsf{T}\left\{\mathbf{R}^{-1} - (\mathbf{H}^\mathsf{T}\mathbf{R}^{-1})^\mathsf{T}\mathbf{P}(\mathbf{H}^\mathsf{T}\mathbf{R}^{-1})\right\}\boldsymbol{y} - 2\boldsymbol{y}^\mathsf{T}(\mathbf{H}^\mathsf{T}\mathbf{R}^{-1})^\mathsf{T}\mathbf{PB}^{-1}\boldsymbol{b} \tag{A.22}$$
$$- \|\boldsymbol{b}\|^2_{\mathbf{BP}^{-1}\mathbf{B}} + \|\boldsymbol{b}\|^2_{\mathbf{B}}.$$

Completing the square on $\boldsymbol{y}$ yields

$$J = \|\boldsymbol{x} - \boldsymbol{\mu}\|^2_{\mathbf{P}} + \|\boldsymbol{y} - \boldsymbol{d}\|^2_{\tilde{\mathbf{C}}} - \|\boldsymbol{d}\|^2_{\tilde{\mathbf{C}}} - \|\boldsymbol{b}\|^2_{\mathbf{BP}^{-1}\mathbf{B}} + \|\boldsymbol{b}\|^2_{\mathbf{B}}\,, \tag{A.23}$$

with the definitions

$$\tilde{\mathbf{C}} = \left\{\mathbf{R}^{-1} - (\mathbf{H}^\mathsf{T}\mathbf{R}^{-1})^\mathsf{T}\mathbf{P}(\mathbf{H}^\mathsf{T}\mathbf{R}^{-1})\right\}^{-1}, \tag{A.24a}$$

$$\boldsymbol{d} = \tilde{\mathbf{C}}(\mathbf{H}^\mathsf{T}\mathbf{R}^{-1})^\mathsf{T}\mathbf{P}\mathbf{B}^{-1}\boldsymbol{b}. \tag{A.24b}$$

Equation (A.20a) and eqn. (A.19a) are used to show what the notation hints at,

$$\begin{aligned}
\tilde{\mathbf{C}} &= \left\{\mathbf{R}^{-1} - (\mathbf{H}^\mathsf{T}\mathbf{R}^{-1})^\mathsf{T}\mathbf{B}\mathbf{H}^\mathsf{T}\mathbf{C}^{-1}\right\}^{-1} && \text{(A.25)}\\
&= \left\{\mathbf{R}^{-1} - \mathbf{R}^{-1}\left(\mathbf{H}\mathbf{B}\mathbf{H}^\mathsf{T}\right)\mathbf{C}^{-1}\right\}^{-1} && \text{(A.26)}\\
&= \mathbf{C}\left\{\mathbf{C} - \mathbf{H}\mathbf{B}\mathbf{H}^\mathsf{T}\right\}^{-1}\mathbf{R} && \text{(A.27)}\\
&= \mathbf{C}\mathbf{R}^{-1}\mathbf{R} && \text{(A.28)}\\
&= \mathbf{C}, && \text{(A.29)}
\end{aligned}$$

while $\boldsymbol{d}$ is simplified using eqns. (A.16a) and (A.19a),

$$\begin{aligned}
\boldsymbol{d} &= (\mathbf{H}\mathbf{B}\mathbf{H}^\mathsf{T} + \mathbf{R})\mathbf{R}^{-1}\mathbf{H}(\mathbf{H}^\mathsf{T}\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1})^{-1}\mathbf{B}^{-1}\boldsymbol{b} && \text{(A.30)}\\
&= (\mathbf{H}\mathbf{B}\mathbf{H}^\mathsf{T}\mathbf{R}^{-1}\mathbf{H} + \mathbf{H})(\mathbf{B}\mathbf{H}^\mathsf{T}\mathbf{R}^{-1}\mathbf{H} + \mathbf{I}_m)^{-1}\boldsymbol{b} && \text{(A.31)}\\
&= \mathbf{H}(\mathbf{B}\mathbf{H}^\mathsf{T}\mathbf{R}^{-1}\mathbf{H} + \mathbf{I}_m)(\mathbf{B}\mathbf{H}^\mathsf{T}\mathbf{R}^{-1}\mathbf{H} + \mathbf{I}_m)^{-1}\boldsymbol{b} && \text{(A.32)}\\
&= \mathbf{H}\boldsymbol{b}. && \text{(A.33)}
\end{aligned}$$

Thus, eqn. (A.23) becomes

$$\begin{aligned}
J &= \|\boldsymbol{x} - \boldsymbol{\mu}\|^2_{\mathbf{P}} + \|\boldsymbol{y} - \mathbf{H}\boldsymbol{b}\|^2_{\mathbf{C}} - \|\mathbf{H}\boldsymbol{b}\|^2_{\mathbf{C}} - \|\boldsymbol{b}\|^2_{\mathbf{B}\mathbf{P}^{-1}\mathbf{B}} + \|\boldsymbol{b}\|^2_{\mathbf{B}} && \text{(A.34)}\\
&= \|\boldsymbol{x} - \boldsymbol{\mu}\|^2_{\mathbf{P}} + \|\boldsymbol{y} - \mathbf{H}\boldsymbol{b}\|^2_{\mathbf{C}} + \boldsymbol{b}^\mathsf{T}\mathbf{D}^{-1}\boldsymbol{b}, && \text{(A.35)}
\end{aligned}$$

with

$$\begin{aligned}
\mathbf{D}^{-1} &= \mathbf{B}^{-1} - \mathbf{B}^{-1}\mathbf{P}\mathbf{B}^{-1} + \mathbf{H}^\mathsf{T}\mathbf{C}^{-1}\mathbf{H} && \text{(A.36)}\\
&= \mathbf{B}^{-1} - \mathbf{B}^{-1}(\mathbf{I}_m - \mathbf{K}\mathbf{H}) + \mathbf{B}^{-1}\mathbf{K}\mathbf{H} && \text{(A.37)}\\
&= 0, && \text{(A.38)}
\end{aligned}$$

where eqn. (A.20b) and eqn. (A.19b) were used on the second line. □

A.3 Interpreting the scalar KF

Comprehension of matrix identities can often be enhanced by investigating the scalar case. This section interprets the various KF equations by doing so, with $\boldsymbol{x}, \boldsymbol{b}, \boldsymbol{y} \in \mathbb{R}$, $\mathbf{H} = 1$, and $\mathbf{B}, \mathbf{R} > 0$. Equation (A.13) is then

$$\frac{(\boldsymbol{x} - \boldsymbol{b})^2}{\mathbf{B}} + \frac{(\boldsymbol{x} - \boldsymbol{y})^2}{\mathbf{R}} = \frac{(\boldsymbol{x} - \boldsymbol{\mu})^2}{\mathbf{P}} + \frac{(\boldsymbol{y} - \boldsymbol{b})^2}{\mathbf{R} + \mathbf{B}}\,, \tag{A.39}$$

where $\boldsymbol{\mu}$ and $\mathbf{P}$ are given by eqns. (A.16),

$$\mathbf{P} = \frac{1}{1/\mathbf{B} + 1/\mathbf{R}}\,, \tag{A.40a}$$

$$\boldsymbol{\mu} = \mathbf{P}(\boldsymbol{b}/\mathbf{B} + \boldsymbol{y}/\mathbf{R})\,. \tag{A.40b}$$

This formulation shows that the posterior moments are "weighted averages": the posterior mean is a weighted average of the observation and the prior mean, and the posterior precision is the sum of the observation precision and the background precision.

Next, consider the Kalman gain formulation of the posterior moments:

$$\mathbf{K} = \frac{\mathbf{B}}{\mathbf{B} + \mathbf{R}}\,, \tag{A.41a}$$

$$\mathbf{P} = (1 - \mathbf{K})\mathbf{B}\,, \tag{A.41b}$$

$$\boldsymbol{\mu} = \boldsymbol{b} + \mathbf{K}(\boldsymbol{y} - \boldsymbol{b})\,. \tag{A.41c}$$

Here, $\boldsymbol{\mu}$ is not obtained as a weighted average, but rather as a linear interpolation between the prior mean and the observations, controlled by the "gain" matrix $\mathbf{K}$. The definition (A.41a) of $\mathbf{K}$ is straightforward: it is the ratio of one variance to the total variance. As seen from eqn. (A.41b), the gain also linearly controls the reduction in variance.

In the multivariate case, the weighted average and the gain formulations are related through the Woodbury corollaries, A.1 and A.2. In the scalar case, however, both of these are just a step away from the trivial identity

$$\frac{1}{1/\mathbf{B} + 1/\mathbf{R}} = \frac{\mathbf{B}\mathbf{R}}{\mathbf{B} + \mathbf{R}}\,. \tag{A.42}$$

Similarly, the identity eqn. (A.20a) becomes just another ratio as well:

$$\mathbf{K} = \mathbf{P}/\mathbf{R} = \frac{1/\mathbf{R}}{1/\mathbf{R} + 1/\mathbf{B}} \,. \tag{A.43}$$

Consider again the right hand side of eqn. (A.39) or (A.13). The second term, $(\boldsymbol{y} - \boldsymbol{b})^2/(\mathbf{R} + \mathbf{B})$, is referred to as "model evidence" [84]. It is the mismatch between the prior mean and the data, normalized by their total variance. Although insignificant to the KF, the term plays a role for the RTS smoother (appendix D), it carries useful information for adjusting inflation parameters (chapter 6), and it may serve to interpret the stopping criteria for iterative methods [112, 125].

Now consider the first term, $(\boldsymbol{x} - \boldsymbol{\mu})^2/\mathbf{P}$. (i) It gathers all of the dependence in $\boldsymbol{x}$, facilitating the integration over $\boldsymbol{x}$. (ii) It is again quadratic, meaning that the posterior pdf of the KF is Gaussian. (iii) The posterior variance, $\mathbf{P}$, does not depend on the value of the model evidence. The last item is a remarkable [e.g. 5, example 3.6] property of Gaussian variables: the posterior variance is always reduced. This is puzzling; if the prior is very far from the data relative to $\mathbf{B}$, then maybe $\mathbf{B}$ was set too small? It is countered in chapter 6 by a hierarchical approach. In contrast to the Gaussian case, in general, only the expected variance is reduced, and similarly, only the expected Shannon information is increased.

Appendix B

Theoretical developments for chapter 5

B.1 The residual noise

B.1.1 The cross terms

Let $\mathbf{C}$ be the sum of the two cross terms:

$$\mathbf{C} = \hat{\mathbf{Q}}^{1/2}\mathbf{Z}^\mathsf{T} + \mathbf{Z}\hat{\mathbf{Q}}^{\mathsf{T}/2} \tag{B.1}$$

$$= \mathbf{\Pi_A}\mathbf{Q}(\mathbf{I}_m - \mathbf{\Pi_A}) + (\mathbf{I}_m - \mathbf{\Pi_A})\mathbf{Q}\mathbf{\Pi_A}\,. \tag{B.2}$$

Note that $\mathsf{range}(\hat{\mathbf{Q}}^{1/2}\mathbf{Z}^\mathsf{T}) \subseteq \mathsf{range}(\mathbf{A}) \subseteq \mathsf{ker}(\hat{\mathbf{Q}}^{1/2}\mathbf{Z}^\mathsf{T})$, and therefore $\hat{\mathbf{Q}}^{1/2}\mathbf{Z}^\mathsf{T}$ (and its transpose) only has the eigenvalue 0. Alternatively one can show that it is nilpotent of degree 2. By contrast, the nature of the eigenvalues of $\mathbf{C}$ is quite different.

Theorem B.1 – Properties of $\mathbf{C}$.
The symmetry of $\mathbf{C} \in \mathbb{R}^{m^2}$ implies, by the spectral theorem, that its spectrum is real. Suppose that λ is a non-zero eigenvalue of $\mathbf{C}$, with eigenvector $\boldsymbol{v} = \boldsymbol{v}_A + \boldsymbol{v}_B$, where $\boldsymbol{v}_A = \mathbf{\Pi_A}\boldsymbol{v}$ and $\boldsymbol{v}_B = (\mathbf{I}_m - \mathbf{\Pi_A})\boldsymbol{v}$. Then (i) $\boldsymbol{u} = \boldsymbol{v}_A - \boldsymbol{v}_B$ is also an eigenvector, (ii) its eigenvalue is $-\lambda$, and (iii) neither $\boldsymbol{v}_A$ nor $\boldsymbol{v}_B$ are zero.

Proof. Note that

$$\mathbf{C}\boldsymbol{v}_A = (\mathbf{I}_m - \mathbf{\Pi_A})\mathbf{Q}\boldsymbol{v}_A \in \mathsf{range}(\mathbf{A})^\perp\,, \tag{B.3}$$

$$\mathbf{C}\boldsymbol{v}_B = \mathbf{\Pi_A}\mathbf{Q}\boldsymbol{v}_B \qquad\quad \in \mathsf{range}(\mathbf{A})\,. \tag{B.4}$$

As $\mathbf{C}\boldsymbol{v} = \lambda[\boldsymbol{v}_A + \boldsymbol{v}_B]$, eqns. (B.3) and (B.4) imply that

$$\mathbf{C}\boldsymbol{v}_A = \lambda \boldsymbol{v}_B \,, \tag{B.5}$$

$$\mathbf{C}\boldsymbol{v}_B = \lambda \boldsymbol{v}_A \,. \tag{B.6}$$

Therefore,

$$\mathbf{C}\boldsymbol{u} = \mathbf{C}[\boldsymbol{v}_A - \boldsymbol{v}_B] = \lambda \boldsymbol{v}_B - \lambda \boldsymbol{v}_A = -\lambda[\boldsymbol{v}_A - \boldsymbol{v}_B] = -\lambda \boldsymbol{u} \qquad \square$$

Equations (B.5) and (B.6) can also be seen to imply (iii).

Corollary B.1 – $\mathsf{tr}(\mathbf{C}) = 0$**.**
This follows from the fact that the trace of a matrix equals the sum of its eigenvalues.

Corollary B.2 – $\|\boldsymbol{v}_A\|_2^2 = \|\boldsymbol{v}_B\|_2^2$**.**
This follows from the fact that $\boldsymbol{v}^\mathsf{T}\boldsymbol{u} = (\boldsymbol{v}_A + \boldsymbol{v}_B)^\mathsf{T}(\boldsymbol{v}_A - \boldsymbol{v}_B) = \boldsymbol{v}_A^\mathsf{T}\boldsymbol{v}_A - \boldsymbol{v}_B^\mathsf{T}\boldsymbol{v}_B$ *should be zero by the spectral theorem.*

Interestingly, imaginary, skew-symmetric matrices also have the property that their eigenvalues, all of which are real, come in positive/negative pairs. These matrices can all be written $\mathbf{M} - \mathbf{M}^\mathsf{T}$ for some $\mathbf{M} \in i\mathbb{R}^{m^2}$, which is very reminiscent of $\mathbf{C}$. However, it is not clear if these parallels can be used to prove Theorem B.1 because $\mathbf{M} - \mathbf{M}^\mathsf{T}$ only has zeros on the diagonal, while $\mathbf{C}$ generally does not (by symmetry, it can be seen that this would imply $\mathbf{C} = 0$). Also, Theorem B.1 depends on the fact that the cross terms are "flanked" by orthogonal projection matrices, whereas there are no requirements on $\mathbf{M}$.

B.1.2 The residual covariance matrix

The residual, $[\mathbf{Q} - \hat{\mathbf{Q}}]$, differs from the symmetric, positive matrix $\mathbf{Z}\mathbf{Z}^\mathsf{T}$ by the cross terms, $\mathbf{C}$. The following theorem establishes a problematic consequence.

Theorem B.2 – $[\mathbf{Q} - \hat{\mathbf{Q}}]$ is not a covariance matrix**.**
Provided $\mathbf{C} \neq 0$*, the residual "covariance" matrix,* $[\mathbf{Q} - \hat{\mathbf{Q}}]$*, has negative eigenvalues.*

Proof. Since $\mathbf{C}$ is symmetric, and thus orthogonally diagonalizable, the assumption that $\mathbf{C} \neq 0$ implies that $\mathbf{C}$ has non-zero eigenvalues. Let $\boldsymbol{v}$ be the eigenvector of a non-zero eigenvalue, and write $\boldsymbol{v} = \boldsymbol{v}_A + \boldsymbol{v}_B$, with $\boldsymbol{v}_A \in \mathsf{range}(\mathbf{A})$ and $\boldsymbol{v}_B \in \mathsf{range}(\mathbf{A})^\perp$.

Then $\boldsymbol{v}^\mathsf{T}\mathbf{C}\boldsymbol{v} = 2\boldsymbol{v}_A^\mathsf{T}\mathbf{Q}\boldsymbol{v}_B \neq 0$. Define $\boldsymbol{v}_\alpha = \boldsymbol{v}_B + \alpha\boldsymbol{v}_A$. Then:

$$\boldsymbol{v}_\alpha^\mathsf{T}[\mathbf{Q} - \hat{\mathbf{Q}}]\boldsymbol{v}_\alpha = \boldsymbol{v}_\alpha^\mathsf{T}[\mathbf{Z}\mathbf{Z}^\mathsf{T} + \mathbf{C}]\boldsymbol{v}_\alpha \tag{B.7}$$

$$= \boldsymbol{v}_B^\mathsf{T}\mathbf{Q}\boldsymbol{v}_B + 2\alpha\boldsymbol{v}_A^\mathsf{T}\mathbf{Q}\boldsymbol{v}_B\,. \tag{B.8}$$

The second term can always be made negative, but larger in magnitude than the first, merely by choosing the sign of α and making it sufficiently large. □

B.1.3 Eliminating the cross terms

Can the cross terms be entirely eliminated in some way? section 5.65.6.2 already answered this question in the negative: there is no particular choice of the square root of $\mathbf{Q}$, inducing a choice of $\hat{\mathbf{Q}}^{1/2}$ and $\mathbf{Z}$ through eqns. (5.23) and (5.24), that eliminates the cross terms, $\mathbf{C}$.

But suppose we allow changing the ensemble subspace. For example, suppose the partition $\mathbf{Q}^{1/2} = \hat{\mathbf{Q}}^{1/2} + \mathbf{Z}$ uses the projector onto the N largest-eigenvalue eigenvectors of $\mathbf{Q}$ instead of $\mathbf{\Pi_A}$. It can then be shown that the cross terms are eliminated: $\hat{\mathbf{Q}}^{1/2}\mathbf{Z}^\mathsf{T} = 0$, and hence $\mathbf{C} = 0$ and $\mathbb{V}\mathrm{ar}(\boldsymbol{q}^{\perp\!\!\!\perp}) = \mathbf{Q}$. A similar situation arises in the case of the COFFEE algorithm (section 5.5), explaining why it does not have the cross term problem. Another particular rank-N square root that yields $\mathbf{C} = 0$ is the lower-triangular Cholesky factor of $\mathbf{Q}$ with the last $m - N$ columns set to zero.

Unfortunately, for general $\mathbf{Q}$ and $\mathbf{A}$, the ensemble subspace will not be that of the rank-N truncated Cholesky or eigenvalue subspace. Therefore neither of these options can be carried out using a right-multiplying square root.

B.1.4 The essential difference to the analysis step

The model noise square root method is motivated by the square root methods of the analysis step. However, even if the observation covariance matrix, $\mathbf{R} \in \mathbb{R}^{p^2}$, is full-rank, with $p > N$, one does not encounter the problem of noise components outside of the span of the anomalies in the analysis step.

This begs the question: why? Note that the desired analysis step update, eqn. (2.31), can be written as $\bar{\mathbf{P}}^a = \bar{\mathbf{P}}^f - \bar{\mathbf{K}}(\mathbf{H}\bar{\mathbf{P}}^f\mathbf{H}^\mathsf{T} + \mathbf{R})\bar{\mathbf{K}}^\mathsf{T}$. As can be seen $\mathbf{R}$ is "confined" between two Kalman gains, $\bar{\mathbf{K}}$. But $\bar{\mathbf{K}}$ implicitly contains linear regression, as discussed in section 2.3.2, which implicitly performs a projection onto the subspace spanned by the anomalies. This explains, mechanically, *how* a high-rank $\mathbf{R}$ does not

pose any problems in the analysis step, but not *why*. In other words, is this just a lucky coincidence? And could we, in fact, improve the analysis step by treating $\mathbf{R}$ more carefully? Or, vice versa, could we justify neglecting the residual, i.e. the complementary projection, based on this comparison?

The answer is no, and the reason lies in the fundamental difference that the analysis step is removing uncertainty, while the forecast step is adding to it. This essential difference is manifested not just as a sign difference in the covariance equations, but also by the absence of any residual of $\mathbf{R}$. In fact, the EnKF implicitly assumes that there is zero uncertainty in the space not spanned by the anomalies. Since uncertainty, quantified through variance, is necessarily positive, nothing can be removed from these components, and any residual $\mathbf{R}$ must disappear. This is not so when adding uncertainty, quantified by $\mathbf{Q}$.

B.2 Consistency of Sqrt-Dep

Sqrt-Core ensures that eqn. (5.8) is satisfied, i.e. that

$$\frac{1}{N-1}[\mathbf{A}+\hat{\mathbf{D}}][\mathbf{A}+\hat{\mathbf{D}}]^{\mathsf{T}} = \bar{\mathbf{P}} + \hat{\mathbf{Q}}\,, \tag{B.9}$$

where $(N-1)\bar{\mathbf{P}} = \mathbf{A}\mathbf{A}^{\mathsf{T}}$. However, this does not imply that $\hat{\mathbf{D}}\hat{\mathbf{D}}^{\mathsf{T}} = (N-1)\hat{\mathbf{Q}}$. Therefore, in reference to Sqrt-Dep, $\hat{\mathbf{\Xi}}\hat{\mathbf{\Xi}}^{\mathsf{T}} \neq \mathbf{I}_m$. Instead, the magnitudes of $\hat{\mathbf{D}}$ and $\hat{\mathbf{\Xi}}$ are minimized as much as possible, as per Theorem 5.1.

However, Sqrt-Dep is designed assuming that $\hat{\mathbf{\Xi}}$ is stochastic, with its columns drawn independently from $\mathcal{N}(0, \mathbf{I}_m)$. If this were the case, then Sqrt-Dep would be consistent in the sense of

$$\frac{1}{N-1}\,\mathbb{E}\left([\mathbf{A}+\hat{\mathbf{D}}+\tilde{\mathbf{D}}][\mathbf{A}+\hat{\mathbf{D}}+\tilde{\mathbf{D}}]^{\mathsf{T}}\right) = \bar{\mathbf{P}} + \mathbf{Q}\,, \tag{B.10}$$

where the expectation is with respect to $\tilde{\mathbf{\Xi}}$ and $\hat{\mathbf{\Xi}}$. This follows from the consistency of $\boldsymbol{q}$ as defined in eqn. (5.33), which has $\mathbb{V}\mathrm{ar}(\boldsymbol{q}) = \mathbf{Q}$, because each column of $\mathbf{D} = \hat{\mathbf{D}} + \tilde{\mathbf{D}}$ is sampled in the same manner as $\boldsymbol{q}$.

The fact that $\hat{\mathbf{D}}$ is in fact not stochastic, as Sqrt-Dep assumes, but typically of a much smaller magnitude, suggests some possible avenues for future improvement. For example we speculate that inflating $\hat{\mathbf{\Xi}}$ by a factor larger than one, possibly estimated in a similar fashion to [50]. The value of $\hat{\mathbf{\Xi}}$ also depends on the choice of square root for $\hat{\mathbf{Q}}^{1/2}$. It may therefore be a good idea to choose $\hat{\mathbf{Q}}^{1/2}$ somewhat randomly, so as

to induce more randomness in the square root "noise", $\hat{\boldsymbol{\Xi}}$. One way of doing so is to apply a right-multiplying rotation matrix to $\hat{\mathbf{Q}}^{1/2}$. Cursory experiments indicate that there may be improvements using either of the above two suggestions.

B.3 Left-multiplying formulation of Sqrt-Core

Lemma B.1.
The row (and column) space of $\mathbf{T}^f_{\text{sym}} = (\mathbf{G}^f)^{1/2}_{\text{sym}}$ *is the row space of* $\mathbf{A}$.

Proof. Let $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\mathsf{T}$ be the SVD of $\mathbf{A}$. Then:

$$\begin{aligned}\mathbf{G}^f &= \mathbf{I}_N + (N-1)\mathbf{A}^+\mathbf{Q}(\mathbf{A}^+)^\mathsf{T} \qquad \text{(B.11)}\\ &= \mathbf{V}\left(\mathbf{I}_N + (N-1)\boldsymbol{\Sigma}^+\mathbf{U}^\mathsf{T}\mathbf{Q}\mathbf{U}(\boldsymbol{\Sigma}^+)^\mathsf{T}\right)\mathbf{V}^\mathsf{T} \qquad \square\end{aligned}$$

In view of Lemma B.1 it seems reasonable that there should be a left-multiplying update, $\mathbf{A}^f = \mathbf{L}\mathbf{A}$, such that it equals the right-multiplying update, $\mathbf{A}^f = \mathbf{A}\mathbf{T}^f_{\text{sym}}$. Although $N \ll m$ in most applications of the EnKF, the left-multiplying update would be a lot less costly to compute than the right-multiplying one in such cases if $N \gg m$. The following derivation of an explicit formula for $\mathbf{L}$ is very close to that of [172], except for the addition of eqn. (5.13). Lemma B.2 will also be of use.

Lemma B.2.
For any matrices, $\mathbf{A} \in \mathbb{R}^{m\times N}$, $\mathbf{M} \in \mathbb{R}^{m^2}$, *and any positive integer,* k,

$$\mathbf{A}(\mathbf{A}^\mathsf{T}\mathbf{M}\mathbf{A})^k = (\mathbf{A}\mathbf{A}^\mathsf{T}\mathbf{M})^k\mathbf{A}\,. \qquad \text{(B.12)}$$

Proposition B.1 – Left-multiplying transformation**.**
For any ensemble anomaly matrix, $\mathbf{A} \in \mathbb{R}^{m\times N}$, *and any SPD matrix* $\mathbf{Q} \in \mathbb{R}^{m^2}$,

$$\mathbf{A}\mathbf{T}^f_{\text{sym}} = \mathbf{L}\mathbf{A} \qquad \text{(B.13)}$$

where

$$\mathbf{T}^f_{\text{sym}} = \left(\mathbf{I}_N + (N-1)\mathbf{A}^+\mathbf{Q}(\mathbf{A}^+)^\mathsf{T}\right)^{1/2}_{\text{sym}}\,, \qquad \text{(B.14)}$$

$$\mathbf{L} = \left(\mathbf{I}_m + (N-1)\mathbf{A}\mathbf{A}^+\mathbf{Q}(\mathbf{A}\mathbf{A}^\mathsf{T})^+\right)^{1/2}\,. \qquad \text{(B.15)}$$

In case $N > m$, eqn. (B.15) *reduces to*

$$\mathbf{L} = \left(\mathbf{I}_m + (N-1)\mathbf{Q}(\mathbf{A}\mathbf{A}^\mathsf{T})^{-1}\right)^{1/2} . \tag{B.16}$$

Note that $(\mathbf{I}_m + \mathbf{A}\mathbf{A}^+\mathbf{Q}(\mathbf{A}\mathbf{A}^\mathsf{T})^+)$ is not a symmetric matrix. We can nevertheless define its square root as the square root obtained from its eigendecomposition, as was done for the symmetric square root in section 4.2.

Proof. Assuming $\mathbf{A}^+\mathbf{Q}(\mathbf{A}^+)^\mathsf{T}$ has eigenvalues less than 1, we can express the square root, $(\mathbf{A}^+\mathbf{Q}(\mathbf{A}^+)^\mathsf{T})^{1/2}$, through its Taylor expansion [76, Th. 9.1.2]. Applying eqn. (5.13), followed by Lemma B.2 with $\mathbf{M} = (\mathbf{A}\mathbf{A}^\mathsf{T})^+(N-1)\mathbf{Q}(\mathbf{A}\mathbf{A}^\mathsf{T})^+$, and eqn. (5.13) the other way again, one obtains eqn. (B.15).

If $N > m$, then $\mathsf{rank}(\mathbf{A}) = m$, unless the dynamics have made some of the anomalies collinear. Hence $\mathsf{rank}(\mathbf{A}\mathbf{A}^\mathsf{T}) = m$ and so $\mathbf{A}\mathbf{A}^\mathsf{T}$ is invertible, and $\mathbf{A}\mathbf{A}^+ = \mathbf{I}_m$. Thus, eqn. (B.15) reduces to eqn. (B.16). □

Note that the existence of a left-multiplying formulation of the right multiplying operation $\mathbf{A} \mapsto \mathbf{A}\mathbf{T}^f_{\text{sym}}$ could be used as a proof for Proposition 5.1, because $\mathbf{L}\mathbf{A}\mathbb{1} = 0$ by the definition (5.6) of $\mathbf{A}$. Finally, Proposition B.2 provides an indirect formula for $\mathbf{L}$.

Proposition B.2 – Indirect left-multiplying formula**.**
If we have already calculated the right-multiplying transform matrix $\mathbf{T}^f_{\text{sym}}$, then the we can obtain a corresponding left-multiplying matrix, $\mathbf{L}$, from:

$$\mathbf{L} = \mathbf{A}\mathbf{T}^f_{\text{sym}}\mathbf{A}^+ \tag{B.17}$$

Proof. We need to show that $\mathbf{L}\mathbf{A} = \mathbf{A}\mathbf{T}^f_{\text{sym}}$. Note that $\mathbf{A}^+\mathbf{A}$ is the orthogonal (and hence symmetric) projector onto the row space of $\mathbf{A}$, which Lemma B.1 showed is also the row and column space of $\mathbf{T}^f_{\text{sym}}$. Therefore $\mathbf{T}^f_{\text{sym}}(\mathbf{A}^+\mathbf{A}) = \mathbf{T}^f_{\text{sym}}$, and $\mathbf{L}\mathbf{A} = \mathbf{A}\mathbf{T}^f_{\text{sym}}(\mathbf{A}^+\mathbf{A}) = \mathbf{A}\mathbf{T}^f_{\text{sym}}$. □

Appendix C

Considerations on the definition of the inflation factor, λ

The definition of λ of section 6.5.4 is intrinsic to the problem statement, emerging as the relative mismatch between $\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2_{\mathbf{B}}$ and $\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2_{\bar{\mathbf{B}}}$. This appendix considers alternative, extrinsic definitions of λ. This is valuable because similar definitions are used in the literature, but without the analysis enabled by the EnKF-N framework. It is also useful because the different approaches are intertwined, necessitating detailed analysis to distinguish between them, and avoid "running in circles".

C.1 Introduction

One issue with several of the efforts in the literature towards on-line estimated adaptive inflation is that the definition of the inflation factor is somewhat vague [e.g. 8, 119, 139]. An estimator is typically provided, as is the recipe for its use (to multiply the anomalies by λ or the covariance by λ^2), but not the formal, mathematical definition of its estimand (what exactly does it estimate?). For example, if given the value of (i) the true state covariance matrix, $\mathbf{B}$, and (ii) the value of the estimated state covariance matrix, $\bar{\mathbf{B}}$, it should be possible to say unequivocally what the value of λ is. This is not possible, however, if all that is provided is an estimator. This lack of rigour impairs both understanding and applicability.

The absence of a clear definition is likely partly due to inflation being, historically [11], a manually-tuned parameter that accounts for multiple sources of error, as mentioned in section 5.1.2. A second obstacle is the fact that there is no unique generalization of the (scalar) ratio for matrices.

Moreover, consider for instance the estimator of Wang and Bishop [199], $[\bar{\boldsymbol{\delta}}^{\mathsf{T}}\bar{\boldsymbol{\delta}} - \mathsf{tr}(\mathbf{R})]/\,\mathsf{tr}(\mathbf{H}\bar{\mathbf{B}}\mathbf{H}^{\mathsf{T}})$. It is a point estimator derived solely from the *expected value* of the innovation, making no statement of the Bayesian distribution of the covariance matrix (before or after conditioning). By contrast our approach is to first obtain the distribution, $p(\lambda^2 \mid \mathbf{E})$, and then to "weight" it by Bayes' rule (ref. section C.6) against the likelihood of the observation, $\boldsymbol{y}$, *before* choosing a point estimate for λ^2. In other words, we are attempting to be more considerate, about this one degree of freedom, before fixing it.

As in chapter 6, the ensemble is assumed an iid sample from a Gaussian, with no additional information (such as climatology) available, so that the Jeffreys prior for $\mathbf{B}$ appropriate. As noted below eqn. (6.17), it is then necessary to restrict our attention to the ensemble subspace only. As further noted in section 6.3.2, the state then ought to be parameterized in terms of $\hat{\boldsymbol{w}}$, a vector of length $\hat{m} = N - g$. However, the state is not explicitly considered here, only its hyperparameter, the covariance matrix $\mathbf{B}$, of size $\hat{m} \times \hat{m}$. It should be noted that the sample covariance matrix, $\bar{\mathbf{B}}$, has the value $\frac{1}{\nu}\hat{\mathbf{V}}^{\mathsf{T}}\hat{\mathbf{V}} = \frac{1}{\nu}\mathbf{I}_{\hat{m}}$, as in eqn. (6.18), where $\nu = N - 1$.

C.2 The marginal precision ratio

It can be shown along the lines of section 6.3.1 that, with the Jeffreys prior $p(\mathbf{B}) \propto |\mathbf{B}|^{-(\hat{m}+1)/2}$,

$$p(\mathbf{B} \mid \mathbf{E}) \propto |\mathbf{B}|^{-(\hat{m}+1+\nu)/2} \exp\left(-\tfrac{1}{2}\,\mathsf{tr}(\nu\bar{\mathbf{B}}\mathbf{B}^{-1})\right), \tag{C.1}$$

$$\text{i.e.} \quad p(\mathbf{B} \mid \mathbf{E}) = \mathcal{W}_{\hat{m}}^{-1}\left(\mathbf{B} \mid \nu, \nu\bar{\mathbf{B}}\right). \tag{C.2}$$

The Jacobian of $\mathbf{B} \mapsto \mathbf{B}^{-1}$ being $|\mathbf{B}|^{-(\hat{m}+1)}$ [79, §1.3],

$$p(\mathbf{B}^{-1} \mid \mathbf{E}) \propto \left|\mathbf{B}^{-1}\right|^{+(\nu-\hat{m}-1)/2} \exp\left(-\tfrac{1}{2}\nu\,\mathsf{tr}(\bar{\mathbf{B}}\mathbf{B}^{-1})\right), \tag{C.3}$$

$$\text{i.e.} \quad p(\mathbf{B}^{-1} \mid \mathbf{E}) = \mathcal{W}_{\hat{m}}^{+1}\left(\mathbf{B}^{-1} \mid \nu, (\nu\bar{\mathbf{B}})^{-1}\right). \tag{C.4}$$

Consider an arbitrary state vector element index, $i \in 1{:}\hat{m}$. The corresponding element on the diagonal of the precision matrix, $[\mathbf{B}^{-1}]_{ii}$, has the marginal Wishart distribution [79, Th. 3.3.8] or [180, Th. 2.2]

$$p([\mathbf{B}^{-1}]_{ii} \mid \mathbf{E}) = \mathcal{W}_1^{+1}\left([\mathbf{B}^{-1}]_{ii} \mid \nu, [(\nu\bar{\mathbf{B}})^{-1}]_{ii}\right) \tag{C.5}$$

$$\propto [\mathbf{B}^{-1}]_{ii}^{\nu/2-1} \exp\left(-\tfrac{1}{2}\nu[\bar{\mathbf{B}}^{-1}]_{ii}^{-1}[\mathbf{B}^{-1}]_{ii}\right), \tag{C.6}$$

recognizable as a chi-square distribution. Define the inflation factor as the ratio,

$$\lambda^2 = [\bar{\mathbf{B}}^{-1}]_{ii}/[\mathbf{B}^{-1}]_{ii}\,. \tag{C.7}$$

Then

$$p(\lambda^2 \mid \mathbf{E}) \propto (\lambda^2)^{-\nu/2-1} e^{-\frac{1}{2}\nu/\lambda^2}\,, \tag{C.8}$$

$$\text{i.e.} \quad p(\lambda^2 \mid \mathbf{E}) = \chi^{-2}(\lambda^2 \mid \nu)\,, \tag{C.9}$$

as was obtained with the main definition of λ of eqn. (6.56).

Note that although this definition uses the ratio of a given couple of marginal elements, the index i is arbitrary. Therefore, no particular state dimension is (unjustly) favoured. Still, with this definition, knowing λ and $\bar{\mathbf{B}}$ only specifies $[\mathbf{B}]_{ii} = \lambda^2[\bar{\mathbf{B}}]_{ii}$, but not the other elements of $\mathbf{B}$. Therefore, the functional relationship between λ^2 and (the other elements of) $\mathbf{B}$ must be *assigned*; it is of course taken to be $\mathbf{B} = \lambda^2\bar{\mathbf{B}}$.

C.3 The marginal variance ratio

Beware that $[\bar{\mathbf{B}}^{-1}]_{ii}^{-1}$ is not simply $[\bar{\mathbf{B}}]_{ii}$, the i-th element on the diagonal of $\mathbf{B}$. Indeed, by contrast to the approach above, the distribution of the marginal elements of $\mathbf{B}$ is [79, Th. 3.4.2]

$$p([\mathbf{B}]_{ii} \mid \mathbf{E}) = \mathcal{W}_1^{-1}([\mathbf{B}]_{ii} \mid \nu-(\hat{m}-1), \nu[\bar{\mathbf{B}}]_{ii})\,. \tag{C.10}$$

In other words, with the alternative definition $\lambda^2 = [\mathbf{B}]_{ii}/[\bar{\mathbf{B}}]_{ii}$,

$$p(\lambda^2/\nu \mid \mathbf{E}) \propto \chi^{-2}(\lambda^2/\nu \mid \nu-(\hat{m}-1))\,, \tag{C.11}$$

yielding a mode and a mean for λ^2 on the order of ν rather than, as is the case for eqn. (C.9), 1. That is, the individual elements of the covariance matrix are much larger than the inverse of the elements of the precision matrix. Which is a better definition: that yielding eqn. (C.9), or that of eqn. (C.11) ? It is not quite clear, since these definitions are not inherent to the problem statement, as discussed on page 127. However, the fact that the covariance appears in the Gaussian pdf through its inverse (disregarding the normalization constant) does give some credence to the inverse-marginal-precision definition.

C.4 Conditional approach

Suppose, instead of considering the ratio of marginal elements, one investigates $p(\mathbf{B} \mid \mathbf{E})$ under the restriction $\mathbf{B} \propto \bar{\mathbf{B}}$. This can be performed by inspecting $p(\mathbf{B} \mid \mathbf{E})$ along $\mathbf{B} = \mathbf{B}(\lambda) = \lambda^2 \bar{\mathbf{B}}$. This approach is well established [e.g. 50, 51]. However, it should be recognized that inspecting the joint distribution along a constraint amounts to conditioning; in other words, it should be recognized that approximations[1] amount to the introduction of assumptions. The question arises: is it wise to use a conditional distribution?

On the one hand, under this assumption,

$$\frac{\mathsf{tr}(\mathbf{B})}{\mathsf{tr}(\bar{\mathbf{B}})} = \frac{\hat{m}}{\mathsf{tr}(\bar{\mathbf{B}}\mathbf{B}^{-1})} = \left(|\mathbf{B}|/|\bar{\mathbf{B}}|\right)^{1/\hat{m}} = [\mathbf{B}]_{ii}/[\bar{\mathbf{B}}]_{ii} = [\bar{\mathbf{B}}^{-1}]_{ii}/[\mathbf{B}^{-1}]_{ii}\,. \tag{C.12}$$

That is, all of these possible definitions of λ^2 are equivalent, which, for working purposes, is highly alluring. Additionally, it is advantageous that the functional relationship between λ^2 and $\mathbf{B}$ (ref. page 129) is given rather than assigned.

On the other hand, the prior, $p(\lambda^2 \mid \mathbf{E})$, is to be weighted by Bayes' rule (ref. section C.6) against the likelihood of the observation, $\boldsymbol{y}$, and so conditioning on information that is not actually present is dangerous, as it induces overconfidence.[2]

Let us investigate further the formal results of this approach. First it should be specified what exactly is meant by $\mathbf{B} \propto \bar{\mathbf{B}}$. To alleviate notation, define $\mathbf{C} = \mathbf{B}/[\mathbf{B}]_{ii}$. The assumption $\mathbf{B} \propto \bar{\mathbf{B}}$ is then taken to mean $\mathbf{C} = \bar{\mathbf{B}}/[\bar{\mathbf{B}}]_{ii}$. This being fixed,

$$p([\mathbf{B}]_{ii} \mid \mathbf{E}, \mathbf{C}) \propto p([\mathbf{B}]_{ii}, \mathbf{C} \mid \mathbf{E})\,. \tag{C.13}$$

But $\mathbf{B} = [\mathbf{B}]_{ii}\mathbf{C}$, so $[\mathbf{B}]_{ii}$ and $\mathbf{C}$ together fully specify $\mathbf{B}$. Nevertheless, the parameterization in terms of $[\mathbf{B}]_{ii}$ and $\mathbf{C}$ should be distinguished from the familiar "element-wise" parameterization of $\mathbf{B}$.[3] Overlooking the resulting Jacobian would be an instance of the same type of carelessness that gives rise to the Borel-Kolmogorov paradox [140,

[1] The approximation being to restrict our attention and inspection of the pdf to $\mathbf{B} = \mathbf{B}(\lambda)$.

[2] It should be noted that while this is not recognized by [50, 51], (i) they only consider the information of a single likelihood, and this is not subsequently weighted against a different object, and (ii) their "restriction" is based in physics, and is therefore more justified than in our sampling theoretic case.

[3] The reparameterization is analogous to that of section 6.5.1. A similar parameterization is that of the scaled inverse Wishart distribution [73, §3.6].

§2.5]. Since $\mathbf{C}$ is fixed, $\frac{d\mathbf{B}}{d([\mathbf{B}]_{ii},\mathbf{C})} \propto [\mathbf{B}]_{ii}^{\hat{m}(\hat{m}+1)/2-1}$ and so, from eqn. (C.13),

$$p([\mathbf{B}]_{ii} \mid \mathbf{E}, \mathbf{C}) \propto [\mathbf{B}]_{ii}^{\hat{m}(\hat{m}+1)/2-1}\, p(\mathbf{B}{=}[\mathbf{B}]_{ii}\mathbf{C} \mid \mathbf{E})\,. \tag{C.14}$$

Inserting $p(\mathbf{B} \mid \mathbf{E})$ from eqn. (C.1) yields

$$p([\mathbf{B}]_{ii} \mid \mathbf{E}, \mathbf{C}) \propto [\mathbf{B}]_{ii}^{\hat{m}(\hat{m}+1)/2-1} |[\mathbf{B}]_{ii}\mathbf{C}|^{-(\hat{m}+1+\nu)/2} \,\mathsf{exp}\left(-\tfrac{1}{2}\,\mathsf{tr}(\nu\bar{\mathbf{B}}([\mathbf{B}]_{ii}\mathbf{C})^{-1})\right) \tag{C.15}$$

$$\propto [\mathbf{B}]_{ii}^{\hat{m}\nu/2-1} \,\mathsf{exp}\left(-\tfrac{1}{2}\hat{m}\nu[\bar{\mathbf{B}}]_{ii}/[\mathbf{B}]_{ii}\right)\,. \tag{C.16}$$

Thus, with the definition $\lambda^2 = [\mathbf{B}]_{ii}/[\bar{\mathbf{B}}]_{ii}$,

$$p(\lambda^2 \mid \mathbf{E}, \mathbf{C}) = \chi^{-2}(\lambda^2 \mid \hat{m}\nu)\,. \tag{C.17}$$

Contrast this with the prior from the definition by marginal precisions, eqn. (C.9), or the prior of the primary definition, eqn. (6.61), and see Fig. 6.2 for understanding of the distributions. The precision (inverse variance) of the inverse chi-square distribution, $\chi^{-2}(.|k)$, is asymptotically $1/k$. Thus eqn. (C.17) has a "confidence" $\hat{m}$ times too strong, relative to the primary definition

It is interesting to note that the distribution $\chi^{-2}(.|\nu)$ is the (pointwise) $\hat{m}$-th root of $\chi^{-2}(.|\hat{m}\nu)$. This might be a sign that eqn. (C.17) may be seen as the (pointwise) product of $\hat{m}$ equal distributions, resulting possibly from an assumption of independence of $\hat{m}$ different instances of λ^2.

C.5 Acknowledging that one should "spend" information when conditioning

Section C.4 showed that conditioning on the information that $\mathbf{C} = \bar{\mathbf{B}}/[\bar{\mathbf{B}}]_{ii}$ yielded an overconfident prior, $p(\lambda^2 \mid \mathbf{E}, \mathbf{C}) = \chi^{-2}(\lambda^2 \mid \hat{m}\nu)$. Nonetheless, the approach had some practical advantages, notably those discussed in relation to eqn. (C.12), and the simplicity with which the inverse Wishart distribution reduced to the inverse chi-square distribution. This section attempts to salvages the approach by rectifying its central flaw, namely that it conditions on $\mathbf{C} = \bar{\mathbf{B}}/[\bar{\mathbf{B}}]_{ii}$ twice.

Jaynes [101] calls probability theory "the logic of science". Its rules, based on Cox's axioms, are self-consistent, and ensure that one does not use the same information twice over. Avoiding such missteps is fairly straightforward in the case of the KF: the fact that the cost function of eqn. (A.9) is a sum of *two* (i.e. not three or four)

quadratic forms can be motivated without the use of Bayes' rule to derive $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$. But in more complicated information decompositions, such as in section 7.2.1, the probabilistic formalism is quite necessary.

It should be recognized that by imposing $\mathbf{B} \propto \bar{\mathbf{B}}$ one is already "spending" a lot of the information contained in the ensemble. One should therefore not pretend that $\bar{\mathbf{B}}/[\bar{\mathbf{B}}]_{ii}$ is just some arbitrary constant, because it is extracted from the ensemble. This information should therefore not be used a second time, unlike what is done in eqns. (C.13) to (C.17).

The line of reasoning in this section has been heuristic. Somewhat more formally, decompose the ensemble, $\mathbf{E}$, into the pieces of information $\{\bar{\boldsymbol{x}}, \bar{\mathbf{B}}/[\bar{\mathbf{B}}]_{ii}, [\bar{\mathbf{B}}]_{ii}\}$, as well as a rotation matrix and a specification of the ordering of the ensemble (both of which are inconsequential due the Gaussian and independence assumptions). Then $p(\lambda^2 \mid \mathbf{E}, \mathbf{C}) = p(\lambda^2 \mid \bar{\boldsymbol{x}}, \bar{\mathbf{B}}/[\bar{\mathbf{B}}]_{ii}, [\bar{\mathbf{B}}]_{ii}, \mathbf{C}) \approx p(\lambda^2 \mid \bar{\boldsymbol{x}}, [\bar{\mathbf{B}}]_{ii}, \mathbf{C})$, since $\bar{\mathbf{B}}/[\bar{\mathbf{B}}]_{ii}$ contains information primarily concerning $\mathbf{C}$, which is already specified. In other words, given $\mathbf{C}$, the outcome $\bar{\mathbf{B}}/[\bar{\mathbf{B}}]_{ii}$ carries no information, because it is deterministic; it is what it had to be. Thus, besides $\bar{\boldsymbol{x}}$, the only remaining information is the scalar, chi-square statistic $[\bar{\mathbf{B}}]_{ii}$, yielding a distribution akin to $p(\lambda^2 \mid \mathbf{E}) = \chi^{-2}(\lambda^2 \mid \nu)$.

C.6 Total and generalized variance ratios

Consider the definition based on total variance, $\lambda^2 = \hat{m}/\operatorname{tr}(\bar{\mathbf{B}}\mathbf{B})$. Averaging reduces the variance because errors tend to cancel out [101, §7.4]; since $\hat{m}/\operatorname{tr}(\bar{\mathbf{B}}\mathbf{B})$ turns out to be the sample average of $\hat{m}$ independent random variables, the variance is reduced by a factor of $\hat{m}$. Indeed, theorem 3.2.20 of Muirhead [141] yields $p(\lambda^2 \mid \mathbf{E}) = \chi^{-2}(\lambda^2 \mid \hat{m}\nu)$. As described in section C.4, this is $\hat{m}$ times too confident. Could this have been anticipated? In other words, why is this average definition this inappropriate?

The explanation is as follows. With this definition of λ^2, the functional relationship between $\mathbf{B}$ and λ^2 must be assigned (ref. page 129). But as shown in section 6.5.3, the prior is weighted against the likelihood by Bayes' rule. Indeed, consider

$$-2\log p(\hat{\boldsymbol{w}}, \boldsymbol{y}|\lambda^2, \mathbf{E}) = \|\mathbf{Y}\hat{\mathbf{V}}\hat{\boldsymbol{w}} - \bar{\boldsymbol{\delta}}\|^2_{\mathbf{R}} + (N-1)\|\hat{\boldsymbol{w}}\|^2_{\lambda^2\mathbf{I}_{\hat{m}}}\,, \tag{C.18}$$

which can be extracted from eqn. (6.54). It can be seen that the importance (magnitude) of the likelihood and prior terms scale with their respective dimensionality. Thus, the penalization with respect to λ^2 already scales with $\hat{m}$. In other words, the

averaging already takes place as the effect of summing over the dimensions of $\hat{\boldsymbol{w}}$. It is therefore not appropriate that the (hyper)prior, $p(\lambda^2|\mathbf{E})$, should be based on an averaging operation on top of this.

A definition based on the general variance, $\lambda^2 = \left(|\mathbf{B}|/|\bar{\mathbf{B}}|\right)^{1/\hat{m}}$, will suffer from the same problem, this being an average as well, namely the geometric average of the eigenvalues. However, with this definition, $p(\lambda^2|\mathbf{E})$ is not quite tractable. Nevertheless, it is known that $\lambda^{-2}|\mathbf{E}$ has the distribution of the product of $\hat{m}$ independent χ^{+2} variables, and that its precision scales asymptotically with $\hat{m}$ [141, Th. 3.2.15 and 3.2.16], as expected.

Appendix D

Derivation of the linear-Gaussian RTS smoother

From eqn. (7.2),

$$-2\log p(\boldsymbol{x}_t|\boldsymbol{x}_{t+1},\boldsymbol{y}_{1:t}) = \left\|\boldsymbol{x}_{t+1}-\mathbf{F}\boldsymbol{x}_t\right\|^2_{\mathbf{Q}} + \left\|\boldsymbol{x}_t-\boldsymbol{\mu}_{t|t}\right\|^2_{\mathbf{P}_{t|t}} - \left\|\boldsymbol{x}_{t+1}-\boldsymbol{\mu}_{t+1|t}\right\|^2_{\mathbf{P}_{t+1|t}} + C\,,$$

where $\mathbf{F}$ is the linear forecast model and C is a constant with respect to $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t+1}$. Gather the terms in $\boldsymbol{x}_t$ using Lemma A.2. By contrast to the Kalman filter, the "model evidence" term should not be discarded, but used to cancel the third quadratic form above, by identifying $\mathbf{P}_{t+1|t} = \mathbf{F}\mathbf{P}_{t|t}\mathbf{F}^\mathsf{T} + \mathbf{Q}$ and $\boldsymbol{\mu}_{t+1|t} = \mathbf{F}\boldsymbol{\mu}_{t|t}$. Hence,

$$-2\log p(\boldsymbol{x}_t|\boldsymbol{x}_{t+1},\boldsymbol{y}_{1:t}) = \left\|\boldsymbol{x}_t - \left(\boldsymbol{\mu}_{t|t} + \mathbf{J}[\boldsymbol{x}_{t+1}-\boldsymbol{\mu}_{t+1|t}]\right)\right\|^2_{(\mathbf{I}-\mathbf{JF})\mathbf{P}_{t|t}} + C\,, \tag{D.1}$$

where $\mathbf{J}$ has been defined as

$$\mathbf{J} = \mathbf{P}_{t|t}\mathbf{F}^\mathsf{T}\mathbf{P}^{-1}_{t+1|t}\,. \tag{D.2}$$

Substituting eqn. (D.1) into the hindcast equation (7.1),

$$\begin{aligned} &-2\log p(\boldsymbol{x}_t,\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:T}) \qquad\qquad (D.3)\\ &\quad = \left\|\left(\boldsymbol{x}_t - \boldsymbol{\mu}_{t|t} + \mathbf{J}\boldsymbol{\mu}_{t+1|t}\right) - \mathbf{J}\boldsymbol{x}_{t+1}\right\|^2_{(\mathbf{I}-\mathbf{JF})\mathbf{P}_{t|t}} + \left\|\boldsymbol{x}_{t+1}-\boldsymbol{\mu}_{t+1|T}\right\|^2_{\mathbf{P}_{t+1|T}} + D\,, \end{aligned}$$

where D is another constant. Again, by expanding and gathering terms in $\boldsymbol{x}_{t+1}$, it can be shown that

$$
\begin{aligned}
&-2 \log p(\boldsymbol{x}_t, \boldsymbol{x}_{t+1} | \boldsymbol{y}_{1:T}) && \text{(D.4)} \\
&\quad = \left\| \boldsymbol{x}_{t+1} - \left(\boldsymbol{\mu}_{t+1|T} + \mathbf{P}_{t|T} \mathbf{J}^\mathsf{T} \mathbf{P}_{t|T}^{-1} \left[\boldsymbol{x}_t - \boldsymbol{\mu}_{t|t} + \mathbf{J} \boldsymbol{\mu}_{t+1|t} \right] \right) \right\|^2_{\left(\mathbf{I} - \mathbf{P}_{t|T} \mathbf{J}^\mathsf{T} \mathbf{P}_{t|T}^{-1} \mathbf{J} \right) \mathbf{P}_{t+1|T}} \\
&\qquad + \left\| \left(\boldsymbol{x}_t - \boldsymbol{\mu}_{t|t} + \mathbf{J} \boldsymbol{\mu}_{t+1|t} \right) - \mathbf{J} \boldsymbol{\mu}_{t+1|T} \right\|^2_{\mathbf{J} \mathbf{P}_{t+1|T} \mathbf{J}^\mathsf{T} + (\mathbf{I} - \mathbf{J}\mathbf{F}) \mathbf{P}_{t|t}} + D \,. && \text{(D.5)}
\end{aligned}
$$

The first line of eqn. (D.5) does not matter, as the marginalization of $\boldsymbol{x}_{t+1}$ in eqn. (7.1) does not leave any dependency in $\boldsymbol{x}_t$. Thus, by identification,

$$
\begin{aligned}
\boldsymbol{\mu}_{t|T} &= \boldsymbol{\mu}_{t|t} - \mathbf{J} \boldsymbol{\mu}_{t+1|t} + \mathbf{J} \boldsymbol{\mu}_{t+1|T} && \text{(D.6)} \\
&= \boldsymbol{\mu}_{t|t} + \mathbf{J} \left(\boldsymbol{\mu}_{t+1|T} - \boldsymbol{\mu}_{t+1|t} \right) && \text{(D.7)} \\
\mathbf{P}_{t|T} &= \mathbf{J} \mathbf{P}_{t+1|T} \mathbf{J}^\mathsf{T} + (\mathbf{I} - \mathbf{J}\mathbf{F}) \mathbf{P}_{t|t} && \text{(D.8)} \\
&= \mathbf{J} \mathbf{P}_{t+1|T} \mathbf{J}^\mathsf{T} + \mathbf{P}_{t|t} - \mathbf{J} \mathbf{P}_{t+1|t} \mathbf{J}^\mathsf{T} && \text{(D.9)} \\
&= \mathbf{P}_{t|t} - \mathbf{J} (\mathbf{P}_{t+1|t} - \mathbf{P}_{t+1|T}) \mathbf{J}^\mathsf{T} && \text{(D.10)}
\end{aligned}
$$

where eqn. (D.2) was used for eqn. (D.9).

A similar[1] derivation is given by Murphy [142, §18.3.2]. Alternative derivations are given by Jazwinski [102, §7.3] and Särkkä [179, §. 8.2].

[1] Except that instead of using Bayes' rule to invert the conditioning $\boldsymbol{x}_t | \boldsymbol{x}_{t+1}$, he uses the expression for conditioning the jointly Gaussian $(\boldsymbol{x}_t, \boldsymbol{x}_{t+1})$. This is only available in the Gaussian case; the starting point is therefore less general than ours.

Appendix E

The SVD, linear inverse problems, and the pseudoinverse

This appendix presents the basic theory of the singular value decomposition (SVD), linear least squares, and the pseudoinverse.

E.1 The SVD

The SVD is important both as a computational and analytic tool. The proofs for this section are given by Trefethen and Bau [194].

Definition E.1 – The (full) SVD**.**
For any positive integers p and m, the SVD of any $\mathbf{H} \in \mathbb{R}^{p\times m}$ is the factorization

$$\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{T}}, \tag{E.1}$$

*where $\mathbf{U} \in \mathbb{R}^{p^2}$ and $\mathbf{V} \in \mathbb{R}^{m^2}$ are orthogonal (i.e. their columns $\{\boldsymbol{u}_j \in \mathbb{R}^p \ ; \ j = 1, \ldots, p\}$ and $\{\boldsymbol{v}_j \in \mathbb{R}^m \ ; \ j = 1, \ldots, m\}$, are both ortho*normal*) and $\mathbf{\Sigma} \in \mathbb{R}^{p\times m}$ is diagonal and* non-negative*. The columns of $\mathbf{U}$ are known as the* left *singular vectors of $\mathbf{H}$. The columns of $\mathbf{V}$ are known as the* right *singular vectors of $\mathbf{H}$. The diagonal entries of $\mathbf{\Sigma}$, $\{\sigma_j \geq 0 \ ; \ j = 1, \ldots, \mathsf{min}(\{p, m\})\}$, are known as the singular values of $\mathbf{H}$, and are ordered in non-increasing order by convention.*

Proposition E.1 – Existence and uniqueness**.**
Every matrix $\mathbf{H} \in \mathbb{R}^{p\times m}$ *has an SVD. By virtue of the sorting convention, and the non-negativity, the matrix* $\mathbf{\Sigma}$ *is uniquely determined. By virtue of the restriction that* $\mathbf{U}$ *and* $\mathbf{V}$ *be real,* $\boldsymbol{u}_j$ *and* $\boldsymbol{v}_j$ *are uniquely determined up to a multiplication by* -1 *of both, provided that the* j*-th singular value is unique, i.e.* $\sigma_{j+1} < \sigma_j < \sigma_{j-1}$*).*

Note that, as opposed to the eigenvalue decomposition, the SVD exists for rectangular and/or rank-deficient matrices. However, for SPD matrices, the SVD and the eigenvalue decomposition coincide. Proposition E.2 catalogues some properties.

Proposition E.2 – SVD properties**.**

- $\|\mathbf{H}\|_2 = \max_{\boldsymbol{x}\in\mathbb{R}^m} \|\mathbf{H}\boldsymbol{x}\|_2 = \sigma_1$.
- *If* $k = \mathsf{rank}(\mathbf{H})$*, then* $k = \mathsf{max}\{j\ ;\ \sigma_j > 0\}$*, and*
- *the range of* $\mathbf{H}$ *is the span of the* k *first left singular vectors,* $(\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k\})$*.*
- *If* $\mathbf{H}$ *is square,* m*-by-*m*, then* $|\mathbf{H}| = \prod_{j=1}^{m} \sigma_j$.

In addition, by virtue of their orthogonality, the singular vectors (and the singular values) can have important physical or statistical interpretations, as illustrated by "empirical orthogonal eigenfunctions", or "principal component analysis" [209].

Defined below, the reduced SVD only computes the singular vectors with non-zero singular values. There may be significant computational cost savings in only calculating the reduced SVD instead of the full SVD. Golub and Van Loan [77, §8.6.4] and [194, §1.4] provide computational details.

Definition E.2 – The reduced SVD**.**
Let $k = \mathsf{rank}(\mathbf{H})$*. The reduced SVD of* $\mathbf{H}$ *is the factorization*

$$\mathbf{H} = \hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{V}}^{\mathsf{T}}\,, \tag{E.2}$$

where $\hat{\mathbf{U}} \in \mathbb{R}^{p\times k}$*,* $\hat{\mathbf{V}} \in \mathbb{R}^{m\times k}$ *and* $\hat{\mathbf{\Sigma}} \in \mathbb{R}^{k^2}$ *can be derived from* $\mathbf{U}$*,* $\mathbf{V}$ *and* $\mathbf{\Sigma}$ *by removing the appropriate number of rows and columns corresponding to any zero singular values. By construction, existence and uniqueness follow from the full SVD.*

Note that $\hat{\mathbf{U}}\hat{\mathbf{U}}^{\mathsf{T}} \neq \mathbf{I}_p$, even though $\hat{\mathbf{U}}^{\mathsf{T}}\hat{\mathbf{U}} = \mathbf{I}_k$. I.e., even though its columns are orthonormal, $\hat{\mathbf{U}}$ is not orthogonal. The same applies for $\hat{\mathbf{V}}$.

E.2 Linear inverse problems

This section compiles ideas from Wunsch [210] and Ben-Israel and Greville [17]. Consider the linear system of equations

$$\mathbf{H}\boldsymbol{x} = \boldsymbol{y}\,, \tag{E.3}$$

where $\mathbf{H} \in \mathbb{R}^{p\times m}$ again. As is typically the case for inverse problems, $\mathbf{H}$ is rectangular, i.e. $p \neq m$, and therefore not invertible. However, it is assumed, for the moment, that $\mathbf{H}$ is full-rank: $k = \mathsf{rank}(\mathbf{H}) = \mathsf{min}(p, m)$.

E.2.1 The overdetermined case

The system (E.3) is "overdetermined" when $p > m$. In other words, there are more observations in $\boldsymbol{y}$ (and hence equations) than degrees of freedom in $\boldsymbol{x}$, and therefore eqn. (E.3) has no solution. One approach that renders the problem well-posed is to formulate it as a minimization problem on the residual:

$$\underset{\boldsymbol{x}\in\mathbb{R}^m}{\mathsf{minimize}} \quad J(\boldsymbol{x}) = \|\mathbf{H}\boldsymbol{x} - \boldsymbol{y}\|_2^2\,. \tag{E.4}$$

This can be reformulated as the constrained optimization problem

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^m}{\mathsf{minimize}} \quad & J(\boldsymbol{x}, \boldsymbol{r}) = \|\boldsymbol{r}\|_2^2 \\ \mathsf{sub.\,to} \quad & \mathbf{H}\boldsymbol{x} + \boldsymbol{r} = \boldsymbol{y}\,. \end{aligned} \tag{E.5}$$

The problem (E.5) reduces to (E.4) by substitution. But the explicit residual, $\boldsymbol{r}$, serves to make the connection with a Gaussian noise probability (ref. section A.2.2), and thus motivates the constrained optimization approach. Moreover, writing the constraint as $\begin{bmatrix}\mathbf{H}\,, \mathbf{I}_p\end{bmatrix}\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{r}\end{bmatrix} = \boldsymbol{y}$, it is seen that the problem is, in a way, *under*determined. The solution may be found by differentiation of eqn. (E.4), or by the geometric argument that the shortest line between a point and a (hyper)plane (the range of $\mathbf{H}$) is the perpendicular one: $0 = \mathbf{H}^\mathsf{T}\boldsymbol{r} = \mathbf{H}^\mathsf{T}(\boldsymbol{y} - \mathbf{H}\boldsymbol{x}_\star)$, yielding

$$\mathbf{H}^\mathsf{T}\mathbf{H}\boldsymbol{x}_\star = \mathbf{H}^\mathsf{T}\boldsymbol{y}\,, \tag{E.6}$$

where $\boldsymbol{x}_\star$ denotes the minimizing argument, obtained by $\boldsymbol{x}_\star = \mathbf{H}^L \boldsymbol{y}$, where

$$\mathbf{H}^L = \left(\mathbf{H}^\mathsf{T}\mathbf{H}\right)^{-1}\mathbf{H}^\mathsf{T} \tag{E.7}$$

is the "left-inverse" of $\mathbf{H}$.

E.2.2 The underdetermined case

The system (E.3) is "underdetermined" when $p < m$, in which case there are infinitely many solutions. A related problem that is well-posed is to find the solution with the minimal norm, i.e.

$$\begin{aligned} &\underset{\boldsymbol{x}\in\mathbb{R}^m}{\mathsf{minimize}} \quad J(\boldsymbol{x},\boldsymbol{q}) = \|\boldsymbol{q}\|_2^2 \\ &\mathsf{sub.\,to} \qquad \begin{cases} \boldsymbol{x} + \boldsymbol{q} = 0 \\ \mathbf{H}\boldsymbol{x} = \boldsymbol{y}\,. \end{cases} \end{aligned} \tag{E.8}$$

In this case, the constraint $\mathbf{H}\boldsymbol{x} = \boldsymbol{y}$ cannot be eliminated by substitution. But, using Lagrangian multipliers, i.e. $\lambda \in \mathbb{R}^p$, the stationarity conditions yield

$$\boldsymbol{x}_\star - \mathbf{H}^\mathsf{T}\lambda = 0\,, \tag{E.9}$$

$$\mathbf{H}\boldsymbol{x}_\star - \boldsymbol{y} = 0\,, \tag{E.10}$$

resulting in $\boldsymbol{x}_\star = \mathbf{H}^R\boldsymbol{y}$, where $\mathbf{H}^R$ is the "right-inverse",

$$\mathbf{H}^R = \mathbf{H}^\mathsf{T}\left(\mathbf{H}\mathbf{H}^\mathsf{T}\right)^{-1}\,, \tag{E.11}$$

to be compared with eqn. (E.7).

E.3 The pseudoinverse

This section shows that, remarkably, both of the above "generalized inverses", $\mathbf{H}^R$ and $\mathbf{H}^L$, and hence their optimality properties, are unified by the pseudoinverse.

E.3.1 The SVD pseudoinverse

The formulae (E.7) and (E.11), for the left and right inverses of $\mathbf{H}$ reduce to the same when written in terms of the reduced SVD:

$$\mathbf{H}^+ = \hat{\mathbf{V}}\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{U}}^\mathsf{T}\,. \tag{E.12}$$

In this sense, the SVD unifies the left and right inverses. But the formula (E.12) is also defined when $k = \mathsf{rank}(\mathbf{H}) < \mathsf{min}(m, p)$, in which case both $\mathbf{H}\mathbf{H}^\mathsf{T}$ and $\mathbf{H}^\mathsf{T}\mathbf{H}$ are rank deficient. In effect, the matrix rank problems of $\mathbf{H}\mathbf{H}^\mathsf{T}$ and $\mathbf{H}^\mathsf{T}\mathbf{H}$ are automatically handled by the definition of the reduced SVD. This SVD inverse, $\mathbf{H}^+$, is called the pseudoinverse. The following theorem, from [76, Th. 5.5.1], summarizes these findings.

Theorem E.1 – Pseudoinverse optimality**.**
$\boldsymbol{x}_\star = \mathbf{H}^+\boldsymbol{y}$ *minimizes (i) the 2-norm of the residual,* $\|\mathbf{H}\boldsymbol{x} - \boldsymbol{y}\|_2^2$ *and (ii) the 2-norm of* $\boldsymbol{x}$ *itself, and it does so in that order of priority.*

It can be shown that for a diagonal matrix, such as $\mathbf{\Sigma}$, the pseudoinverse is obtained by taking the reciprocal of each non-zero element on the diagonal, leaving the zeros in place. If $\mathbf{\Sigma}$ is rectangular, then $\mathbf{\Sigma}^+$ has the size of $\mathbf{\Sigma}^\mathsf{T}$. Using this, one can derive, from eqn. (E.12), that

$$\mathbf{H}^+ = \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^\mathsf{T}\,, \tag{E.13}$$

which can be used to show, for example, Lemma 5.1.

E.3.2 The limit pseudoinverse

Another appealing avenue for unifying $\mathbf{H}^R$ and $\mathbf{H}^L$ is to solve the problem

$$\begin{aligned} &\underset{\boldsymbol{x}\in\mathbb{R}^m}{\mathsf{minimize}} \quad J(\boldsymbol{x}, \boldsymbol{q}, \boldsymbol{r}) = \|\boldsymbol{q}\|_2^2 + \|\boldsymbol{r}\|_2^2 \\ &\mathsf{sub.\,to} \quad \begin{cases} \boldsymbol{x} + \boldsymbol{q} = 0 \\ \mathbf{H}\boldsymbol{x} + \boldsymbol{r} = \boldsymbol{y}\,, \end{cases} \end{aligned} \tag{E.14}$$

as a "mix" of the underdetermined and the overdetermined problem. Equivalently,

$$\begin{aligned} &\underset{\boldsymbol{x}\in\mathbb{R}^m}{\mathsf{minimize}} \quad J(\boldsymbol{x},\boldsymbol{q},\boldsymbol{r}) = \|\boldsymbol{q}\|_2^2 + \|\boldsymbol{r}\|_2^2 \\ &\mathsf{sub.\,to} \qquad \begin{bmatrix}\mathbf{I}_m \\ \mathbf{H}\end{bmatrix}\boldsymbol{x} + \begin{bmatrix}\boldsymbol{q} \\ \boldsymbol{r}\end{bmatrix} = \begin{bmatrix}0 \\ \boldsymbol{y}\end{bmatrix}, \end{aligned} \tag{E.15}$$

which is in the form of problem (E.5). Therefore, eqn. (E.7) can be inserted, yielding

$$\boldsymbol{x}_\star = \left(\begin{bmatrix}\mathbf{I}_m & \mathbf{H}^\mathsf{T}\end{bmatrix}\begin{bmatrix}\mathbf{I}_m \\ \mathbf{H}\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{I}_m & \mathbf{H}^\mathsf{T}\end{bmatrix}\begin{bmatrix}0 \\ \boldsymbol{y}\end{bmatrix} \tag{E.16}$$

$$= \left(\mathbf{I}_m + \mathbf{H}^\mathsf{T}\mathbf{H}\right)^{-1}\mathbf{H}^\mathsf{T}\boldsymbol{y}\,. \tag{E.17}$$

Note that $\left(\mathbf{I}_m + \mathbf{H}^\mathsf{T}\mathbf{H}\right)^{-1}\mathbf{H}^\mathsf{T}$ always exists since $\left(\mathbf{I}_m + \mathbf{H}^\mathsf{T}\mathbf{H}\right)$ is SPD.

A modification of the cost function by a weighting, ε^2,

$$J(\boldsymbol{x},\boldsymbol{q},\boldsymbol{r}) = \varepsilon^2\|\boldsymbol{q}\|_2^2 + \|\boldsymbol{r}\|_2^2\,, \tag{E.18}$$

does not pose a new challenge, because it can be transferred to the constraints by a scaling change of variables, $\boldsymbol{q} = \varepsilon\boldsymbol{q}'$. Thus, the solution of eqn. (E.17) can be inserted:

$$\boldsymbol{x}_\star = \left(\varepsilon^2\mathbf{I}_m + \mathbf{H}^\mathsf{T}\mathbf{H}\right)^{-1}\mathbf{H}^\mathsf{T}\boldsymbol{y}\,. \tag{E.19}$$

Now, if $(\mathbf{H}^\mathsf{T}\mathbf{H})^{-1}$ or $(\mathbf{H}\mathbf{H}^\mathsf{T})^{-1}$ exists, then, respectively,

$$\lim_{\varepsilon\to 0} \quad \left(\varepsilon^2\mathbf{I}_m + \mathbf{H}^\mathsf{T}\mathbf{H}\right)^{-1}\mathbf{H}^\mathsf{T} = \left(\mathbf{H}^\mathsf{T}\mathbf{H}\right)^{-1}\mathbf{H}^\mathsf{T} \tag{E.20}$$

$$\text{or} \quad \lim_{\varepsilon\to 0} \quad \mathbf{H}^\mathsf{T}\left(\varepsilon^2\mathbf{I}_m + \mathbf{H}\mathbf{H}^\mathsf{T}\right)^{-1} = \mathbf{H}^\mathsf{T}\left(\mathbf{H}\mathbf{H}^\mathsf{T}\right)^{-1}. \tag{E.21}$$

But, as can be shown using Corollary A.2, or an SVD of $\mathbf{H}$,

$$\left(\varepsilon\mathbf{I}_m + \mathbf{H}^\mathsf{T}\mathbf{H}\right)^{-1}\mathbf{H}^\mathsf{T} = \mathbf{H}^\mathsf{T}\left(\varepsilon\mathbf{I}_p + \mathbf{H}\mathbf{H}^\mathsf{T}\right)^{-1} \tag{E.22}$$

for any $\varepsilon > 0$. Thus, for any $\varepsilon > 0$, the left hand sides of eqns. (E.20) and (E.21) are the same. On the other hand, for $\varepsilon = 0$, one of the right hand sides will not exist, since $p \neq m$. Remarkably, however, both *limits* exist, even if $(\mathbf{H}^\mathsf{T}\mathbf{H})^{-1}$ or $(\mathbf{H}\mathbf{H}^\mathsf{T})^{-1}$ does not. Moreover, the unique limit value is the pseudoinverse, $\mathbf{H}^+$, as defined in section E.3.1.

A proof of the above statements is provided by Ben-Israel and Greville [17, §3]. Heuristically, it can be argued that (i) the cost function formulation, eqn. (E.18), indicates that the limiting solution (E.19) should minimize the 2-norm of the residual, as is the first priority of $\mathbf{H}^+$, according to Theorem E.1. But, (ii) for any $\varepsilon > 0$, the solution is also trying to minimize $\|\boldsymbol{q}\|_2^2$. Thus, (iii) providing the limit exists, even if the minimum-residual solution is non-unique, one would expect the limit to pick out the solution that also minimizes $\|\boldsymbol{q}\|_2^2$.

E.3.3 The Moore-Penrose pseudoinverse

Lastly, the Moore-Penrose characterization of the pseudoinverse:

$$\mathbf{H}\mathbf{H}^+\mathbf{H} = \mathbf{H}\,, \tag{E.23}$$

$$\mathbf{H}^+\mathbf{H}\mathbf{H}^+ = \mathbf{H}^+\,, \tag{E.24}$$

$$(\mathbf{H}\mathbf{H}^+)^\mathsf{T} = \mathbf{H}\mathbf{H}^+\,, \tag{E.25}$$

$$(\mathbf{H}^+\mathbf{H})^\mathsf{T} = \mathbf{H}^+\mathbf{H}\,, \tag{E.26}$$

can also be shown to be equivalent to the SVD and the limit definitions [19, §1].

Appendix F

Tuned values of the inflation factor

A descriptions of the process of manually tuning the inflation factor is given in section 5.7. The tables read from left to right, corresponding to the abscissa of the data points on the plots. For example, Fig. (5.3a) has 10 data points (on each curve), each one corresponding to one of the 10 values in the row of Fig. (5.3a) in Table F.1.

Table F.1: Inflation factors used by all of the different EnKF noise treatment methods in the benchmark experiments of chapter 5.

Fig.	Inflation														
5.2	None														
5.3a	1.12	1.10	1.08	1.06	1.05	1.04	1.03	1.03	1.03	1.01					
5.3b	1.14	1.10	1.06	1.03	1.02	1.01	1.01	1.01	1.01	1.00					
5.4a	1.68	1.50	1.32	1.13	1.10	1.07	1.05	1.04	1.04	1.01					
5.4b	1.43	1.35	1.27	1.18	1.15	1.12	1.10	1.09	1.09	1.06					
5.5	1.01	1.02	1.02	1.04	1.05	1.07	1.09	1.17							
5.6	1.01	1.01	1.02	1.02	1.04	1.08	1.15	1.38	1.56	1.73	2.10				
5.7a	1.25	1.22	1.19	1.15	1.13	1.11	1.09	1.07	1.04	1.01					
5.7b	2.10	1.80	1.60	1.50	1.40	1.32	1.25	1.15	1.08	1.04					
5.8	1.13	1.25	1.30	1.35	1.43	1.50	1.57	1.65	1.70						
5.9	1.02	1.02	1.02	1.03	1.04	1.05	1.07	1.09	1.13	1.17	1.21	1.31			
5.10	1.00	1.00	1.00	1.05	1.07	1.10	1.12	1.13	1.14	1.16	1.19	1.21	1.22	1.25	1.27

Table F.2: Inflation factors used by the tuned EnKF in the benchmark experiments of chapter 6.

Fig.	Inflation															
6.4a	6.00	5.10	3.84	1.86	1.20	1.16	1.07	1.06	1.02	1.01						
6.4b	6.00	5.10	3.84	1.86	1.22	1.17	1.12	1.10	1.02	1.01						
6.6a	1.06	1.12	1.19	1.25	1.42	1.60	1.77	1.95	2.12							
6.6b	1.21	1.35	1.47	1.58	1.67	1.77	1.86	1.95	2.00							
6.10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.01	1.01	1.02	1.02	1.03	1.05	1.08	1.1

Table F.3: (i) Inflation factors used by all of the ensemble methods of Fig. 7.2. (ii) Lag length used by the tuned EnKS. (iii) Contraction factors used by the tuned EnRTS.

Type	Values										
Infl.	1.00	1.00	1.00	1.00	1.00	1.01	1.01	1.03	1.08	1.13	1.23
	...	1.35	2.00	2.65							
Lag	102	100	80	60	40	27.50	15.00	10.67	2.00	0.95	0.49
	...	0.30	0	0							
Contr.	0.9999	0.9999	0.99987	0.99983	0.9998	0.9989	0.998	0.995	0.990	0.985	0.977
	...	0.967	0.910	0.853							

Bibliography

[1] Sigurd Aanonsen, Geir Nævdal, Dean Oliver, Albert Reynolds, and Brice Vallès. The ensemble Kalman filter in reservoir engineering–a review. *SPE Journal*, 14 (3):393–412, 2009.

[2] Hassan Abdolhosseini and Ehsan Khamehchi. History matching using traditional and finite size ensemble Kalman filter. *Journal of Natural Gas Science and Engineering*, 27:1748–1757, 2015.

[3] Boujemaa Ait-El-Fquih and François Desbouvries. On Bayesian fixed-interval smoothing algorithms. *Automatic Control, IEEE Transactions on*, 53(10):2437–2442, 2008.

[4] M. U. Altaf, T. Butler, T. Mayo, X. Luo, C. Dawson, A. W. Heemink, and I. Hoteit. A comparison of ensemble Kalman filters for storm surge assimilation. *Monthly Weather Review*, 142(8):2899–2914, 2014.

[5] B. D. O. Anderson and J. B. Moore. *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, NJ, 1979.

[6] Jeffrey L. Anderson. An ensemble adjustment Kalman filter for data assimilation. *Monthly Weather Review*, 129(12):2884–2903, 2001.

[7] Jeffrey L. Anderson. A local least squares framework for ensemble filtering. *Monthly Weather Review*, 131(4):634–642, 2003.

[8] Jeffrey L. Anderson. An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus A*, 59(2):210–224, 2007.

[9] Jeffrey L. Anderson. Spatially and temporally varying adaptive covariance inflation for ensemble filters. *Tellus A*, 61(1):72–83, 2009.

[10] Jeffrey L. Anderson. A non-Gaussian ensemble filter update for data assimilation. *Monthly Weather Review*, 138(11):4186–4198, 2010.

[11] Jeffrey L. Anderson and Stephen L. Anderson. A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Monthly Weather Review*, 127(12):2741–2758, 1999.

[12] Jose Angulo, Hwa-Lung Yu, Andrea Langousis, Alexander Kolovos, Jinfeng Wang, A. Madrid, and George Christakos. Spatiotemporal infectious disease modeling: A BME-SIR approach. *PloS One*, 8(9), 2013.

[13] J. D. Annan, D. J. Lunt, J. C. Hargreaves, and P. J. Valdes. Parameter estimation in an atmospheric GCM using the ensemble Kalman filter. *Nonlinear processes in geophysics*, 12(3):363–371, 2005.

[14] M. Armstrong. *Basic Linear Geostatistics*. Springer Verlag, 1998.

[15] H. M. Arnold, I. M. Moroz, and T. N. Palmer. Stochastic parametrizations and model uncertainty in the Lorenz'96 system. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1991): 20110479, 2013.

[16] Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015.

[17] Adi Ben-Israel and Thomas N. E. Greville. *Generalized Inverses: Theory and Applications*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, 15. Springer-Verlag, New York, second edition, 2003.

[18] Riccardo Benedetti. Scoring rules for forecast verification. *Monthly Weather Review*, 138(1):203–211, 2010.

[19] Andrew F. Bennett. *Inverse Methods in Physical Oceanography*. Cambridge University Press, 1992.

[20] Tyrus Berry and John Harlim. Linear theory for filtering nonlinear multiscale systems with model error. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 470(2167):20140168, 2014.

[21] Laurent Bertino, Geir Evensen, and Hans Wackernagel. Sequential data assimilation techniques in oceanography. *International Statistical Review*, 71(2): 223–241, 2003.

[22] J. Eric Bickel. Some comparisons among quadratic, spherical, and logarithmic scoring rules. *Decision Analysis*, 4(2):49–65, 2007.

[23] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[24] Craig H. Bishop, Brian J. Etherton, and Sharanya J. Majumdar. Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Monthly Weather Review*, 129(3):420–436, 2001.

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

[26] Marc Bocquet and Pavel Sakov. An iterative ensemble Kalman smoother. *Quarterly Journal of the Royal Meteorological Society*, 140(682):1521–1535, 2014.

[27] Marc Bocquet, Carlos A. Pires, and Lin Wu. Beyond Gaussian statistical modeling in geophysical data assimilation. *Monthly Weather Review*, 138(8): 2997–3023, 2010.

[28] Marc Bocquet, Pavel Sakov, et al. Combining inflation-free and iterative ensemble Kalman filters for strongly nonlinear systems. *Nonlinear Processes in Geophysics*, 19(3):383–399, 2012.

[29] Marc Bocquet, Pavel Sakov, et al. Joint state and parameter estimation with an iterative ensemble Kalman smoother. *Nonlinear Processes in Geophysics*, 20 (5):803–818, 2013.

[30] Marc Bocquet, Patrick N. Raanes, and Alexis Hannart. Expanding the validity of the ensemble Kalman filter without the intrinsic need for inflation. *Nonlinear Processes in Geophysics*, 22(6):645–662, 2015.

[31] Jean-Michel Brankart, Clément Ubelmann, Charles-Emmanuel Testut, Emmanuel Cosme, Pierre Brasseur, and Jacques Verron. Efficient parameterization of the observation error covariance matrix for square root or ensemble Kalman filters: application to ocean altimetry. *Monthly Weather Review*, 137(6):1908–1927, 2009.

[32] Yoram Bresler. Two-filter formulae for discrete-time non-linear Bayesian smoothing. *International Journal of Control*, 43(2):629–641, 1986.

[33] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

[34] Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519, 2009.

[35] G. Burgers, P. Jan van Leeuwen, and G. Evensen. Analysis scheme in the ensemble Kalman filter. *Monthly Weather Review*, 126(6):1719–1724, 1998.

[36] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ci You Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal of Scientific Computing*, 16(5):1190–1208, 1995.

[37] Emmanuel J. Candès and Michael B. Wakin. An introduction to compressive sampling. *Signal Processing Magazine, IEEE*, 25(2):21–30, 2008.

[38] Alberto Carrassi and Stéphane Vannitsem. Accounting for model error in variational data assimilation: A deterministic formulation. *Monthly Weather Review*, 138(9):3369–3386, 2010.

[39] Alberto Carrassi, Michael Ghil, Anna Trevisan, and Francesco Uboldi. Data assimilation as a nonlinear dynamical systems problem: Stability and convergence of the prediction-assimilation system. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 18(2):023112, 2008.

[40] Alberto Carrassi, Stephane Vannitsem, Dusanka Zupanski, and Milija Zupanski. The maximum likelihood ensemble filter performances in chaotic systems. *Tellus A*, 61(5):587–600, 2009.

[41] J. Chandrasekar, I. S. Kim, D. S. Bernstein, and A. J. Ridley. Reduced-rank unscented Kalman filtering using Cholesky-based decomposition. *International Journal of Control*, 81(11):1779–1792, 2008.

[42] Yan Chen and Dean S. Oliver. Ensemble randomized maximum likelihood method as an iterative ensemble smoother. *Mathematical Geosciences*, 44(1): 1–26, 2012.

[43] Yan Chen and Dean S. Oliver. History matching of the Norne full field model using an iterative ensemble smoother-(SPE-164902). In *75th EAGE Conference & Exhibition incorporating SPE EUROPEC*, 2013.

[44] Yan Chen and Dean S. Oliver. Levenberg–Marquardt forms of the iterative ensemble smoother for efficient history matching and uncertainty quantification. *Computational Geosciences*, 17(4):689–703, 2013.

[45] Remi Chou, Yvo Boers, Martin Podt, and Matthieu Geist. Performance evaluation for particle filters. In *Information Fusion (FUSION), Proceedings of the 14th International Conference on*, pages 1–7. IEEE, 2011.

[46] H.M. Christensen. Decomposition of a new proper score for verification of ensemble forecasts. *Monthly Weather Review*, 143(5):1517–1532, 2015.

[47] Emmanuel Cosme, J.-M. Brankart, Jacques Verron, Pierre Brasseur, and Monika Krysta. Implementation of a reduced rank square-root smoother for high resolution ocean data assimilation. *Ocean Modelling*, 33(1):87–100, 2010.

[48] Emmanuel Cosme, Jacques Verron, Pierre Brasseur, Jacques Blum, and Didier Auroux. Smoothing problems in a Bayesian framework and their linear Gaussian solutions. *Monthly Weather Review*, 140(2):683–695, 2012.

[49] Roger Daley. Estimating model-error covariances for application to atmospheric data assimilation. *Monthly Weather Review*, 120(8):1735–1746, 1992.

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

[51] Dick P. Dee and Arlindo M. Da Silva. Maximum-likelihood estimation of forecast and observation error covariance parameters. Part I: Methodology. *Monthly Weather Review*, 127(8):1822–1834, 1999.

[52] Ziwang Deng, Youmin Tang, and Howard J. Freeland. Evaluation of several model error schemes in the EnKF assimilation: Applied to Argo profiles in the Pacific Ocean. *Journal of Geophysical Research: Oceans (1978–2012)*, 116(C9), 2011.

[53] François Desbouvries, Yohan Petetin, and Boujemaa Ait-El-Fquih. Direct, prediction-and smoothing-based Kalman and particle filter algorithms. *Signal Processing*, 91(8):2064–2077, 2011.

[54] Susan Dunne and Dara Entekhabi. An ensemble-based reanalysis approach to land data assimilation. *Water Resources Research*, 41(2), 2005.

[55] Alexandre A. Emerick and Albert C. Reynolds. Ensemble smoother with multiple data assimilation. *Computers & Geosciences*, 55:3–15, 2013.

[56] Alexandre A. Emerick and Albert C. Reynolds. Investigation of the sampling performance of ensemble-based methods with a simple reservoir model. *Computational Geosciences*, 17(2):325–350, 2013.

[57] G. Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, 99:10–10, 1994.

[58] G. Evensen. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, 53(4):343–367, 2003.

[59] G. Evensen. The ensemble Kalman filter for combined state and parameter estimation. *Control Systems, IEEE*, 29(3):83–104, 2009.

[60] G. Evensen and P. J. van Leeuwen. An ensemble Kalman smoother for nonlinear dynamics. *Monthly Weather Review*, 128(6):1852–1867, 2000.

[61] Geir Evensen. Advanced data assimilation for strongly nonlinear dynamics. *Monthly Weather Review*, 125(6):1342–1354, 1997.

[62] Geir Evensen. Sampling strategies and square root analysis schemes for the EnKF. *Ocean Dynamics*, 54(6):539–560, 2004.

[63] Geir Evensen. *Data Assimilation*. Springer, 2 edition, 2009.

[64] Geir Evensen and Peter Jan van Leeuwen. Assimilation of Geosat altimeter data for the Agulhas current using the ensemble Kalman filter with a quasigeostrophic model. *Monthly Weather Review*, 124(1):85–96, 1996.

[65] Geir Evensen, Joakim Hove, Hilde Meisingset, Edel Reiso, Knut Sponheim Seim, Øystein Espelid, et al. Using the EnKF for assisted history matching of a North Sea reservoir model. In *SPE Reservoir Simulation Symposium*. Society of Petroleum Engineers, 2007.

[66] C. Farmer. Bayesian field theory applied to scattered data interpolation and inverse problems. *Algorithms for Approximation*, pages 147–166, 2007.

[67] M. Fisher, M. Leutbecher, and G. A. Kelly. On the equivalence between Kalman smoothing and weak-constraint four-dimensional variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 131(613):3235–3246, 2005.

[68] R. Fitzgerald. Divergence of the Kalman filter. *Automatic Control, IEEE Transactions on*, 16(6):736–747, 1971.

[69] Marco Luca Flavio Frei. *Ensemble Kalman Filtering and Generalizations*. PhD thesis, ETH Zurich, Dept. of Mathematics, 2013.

[70] Reinhard Furrer and Thomas Bengtsson. Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis*, 98(2):227–255, 2007.

[71] C. W. Gardiner. *Handbook of Stochastic Methods*. Springer, 2004.

[72] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian data analysis*. Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL, second edition, 2004.

[73] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian data analysis*, volume 2. Taylor & Francis, 2014.

[74] M. E. Gharamti, B. Ait-El-Fquih, and Ibrahim Hoteit. An iterative ensemble Kalman filter with one-step-ahead smoothing for state-parameters estimation of contaminant transport models. *Journal of Hydrology*, 527:442–457, 2015.

[75] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477): 359–378, 2007.

[76] Gene H. Golub and Charles F. Van Loan. *Matrix Computations. 1996*. Johns Hopkins University, Press, Baltimore, MD, USA, third edition, 1996.

[77] Gene H. Golub and Charles F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013.

[78] Mohinder S. Grewal and Angus P. Andrews. Applications of Kalman filtering in aerospace 1960 to the present. *Control Systems, IEEE*, 30(3):69–78, 2010.

[79] A. K. Gupta and D. K. Nagar. *Matrix variate distributions*, volume 104 of *Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics*. Chapman & Hall/CRC, Boca Raton, FL, 2000.

[80] K. S. Gurumoorthy, C. Grudzien, A. Apte, A. Carrassi, and C. K. R. T. Jones. Rank deficiency of Kalman error covariance matrices in linear perfect model. *arXiv preprint arXiv:1108.0158*, mar 2015.

[81] William W. Hager. Updating the inverse of a matrix. *SIAM Review*, 31(2): 221–239, 1989.

[82] Thomas M. Hamill and Jeffrey S. Whitaker. Accounting for the error due to unresolved scales in ensemble data assimilation: A comparison of different approaches. *Monthly Weather Review*, 133(11):3132–3147, 2005.

[83] R. G. Hanea, G. J. M. Velders, A. J. Segers, M. Verlaan, and A. W. Heemink. A hybrid Kalman filter algorithm for large-scale atmospheric chemistry data assimilation. *Monthly Weather Review*, 135(1):140–151, 2007.

[84] Alexis Hannart, Alberto Carrassi, Marc Bocquet, Michael Ghil, Philippe Naveau, Manuel Pulido, Juan Ruiz, and Pierre Tandeo. DADA: Data assimilation for the detection and attribution of weather-and climate-related events. *arXiv preprint arXiv:1503.05236*, 2015.

[85] P. C. Hansen. Regularization tools: A Matlab package for analysis and solution of discrete ill-posed problems. *Numerical Algorithms*, 6(1):1–35, 1994.

[86] Arnold W. Heemink, Martin Verlaan, and Arjo J. Segers. Variance reduced ensemble Kalman filtering. *Monthly Weather Review*, 129(7):1718–1728, 2001.

[87] R. C. Hilborn and J. C. Sprott. *Chaos And Nonlinear Dynamics: An Introduction For Scientists And Engineers*. Oxford University Press, New York, 1994.

[88] H. S. Hoang, R. Baraille, and O. Talagrand. On an adaptive filter for altimetric data assimilation and its application to a primitive equation model, MICOM. *Tellus A*, 57(2):153–170, 2005.

[89] Matthew J. Hoffman, Steven J. Greybush, R. John Wilson, Gyorgyi Gyarmati, Ross N. Hoffman, Eugenia Kalnay, Kayo Ide, Eric J. Kostelich, Takemasa Miyoshi, and Istvan Szunyogh. An ensemble Kalman filter data assimilation system for the Martian atmosphere: Implementation and simulation experiments. *Icarus*, 209(2):470–481, 2010.

[90] Mevin B. Hooten, Jessica Anderson, and Lance A. Waller. Assessing North American influenza dynamics with a statistical SIRS model. *Spatial and spatio-temporal epidemiology*, 1(2):177–185, 2010.

[91] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, second edition, 2013.

[92] I. Hoteit, D.-T. Pham, M.E. Gharamti, and X. Luo. Mitigating observation perturbation sampling errors in the stochastic EnKF. *Monthly Weather Review*, 143(7):2918–2936, 2015.

[93] Ibrahim Hoteit, Dinh-Tuan Pham, and Jacques Blum. A simplified reduced order Kalman filtering and application to altimetric data assimilation in tropical pacific. *Journal of Marine systems*, 36(1):101–127, 2002.

[94] P. L. Houtekamer and Herschel L. Mitchell. Ensemble Kalman filtering. *Quarterly Journal of the Royal Meteorological Society*, 131(613):3269–3289, 2005.

[95] Peter L. Houtekamer and Herschel L. Mitchell. Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review*, 126(3):796–811, 1998.

[96] Peter L. Houtekamer, Herschel L. Mitchell, Gerard Pellerin, Mark Buehner, Martin Charron, Lubos Spacek, and Bjarne Hansen. Atmospheric data assimilation with an ensemble Kalman filter: Results with real observations. *Monthly Weather Review*, 133(3):604–620, 2005.

[97] B. R. Hunt, E. Kalnay, E. J. Kostelich, E. Ott, D. J. Patil, T. Sauer, I. Szunyogh, J. A. Yorke, and A. V. Zimin. Four-dimensional ensemble Kalman filtering. *Tellus A*, 56(4):273–277, 2004.

[98] Brian R. Hunt, Eric J. Kostelich, and Istvan Szunyogh. Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D: Nonlinear Phenomena*, 230(1):112–126, 2007.

[99] Randall J. Hunt, John Doherty, and Matthew J. Tonkin. Are models too simple? arguments for increased parameterization. *Groundwater*, 45(3):254–262, 2007.

[100] K. Ide, P. Courtier, M. Ghil, and A. Lorenc. Unified notation for data assimilation: Operational, sequential and variational. *Journal of the Meteorological Society of Japan*, 75(1B):181–189, 1997.

[101] Edwin T. Jaynes. *Probability theory: the logic of science*. Cambridge university press, 2003.

[102] A. H. Jazwinski. *Stochastic Processes and Filtering Theory*, volume 63. Academic Press, 1970.

[103] Ian T. Jolliffe and David B. Stephenson. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley, 2012.

[104] S. Julier, J. Uhlmann, and H. F. Durrant-Whyte. A new method for the nonlinear transformation of means and covariances in filters and estimators. *Automatic Control, IEEE Transactions on*, 45(3):477–482, 2000.

[105] R. E. Kalman et al. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

[106] James L. Kaplan and James A. Yorke. Chaotic behavior of multidimensional difference equations. In *Functional Differential equations and approximation of fixed points*, pages 204–227. Springer, 1979.

[107] Shree P. Khare, Jeffrey L. Anderson, Timothy J. Hoar, and Douglas Nychka. An investigation into the application of an ensemble Kalman smoother to high-dimensional geophysical systems. *Tellus A*, 60(1):97–112, 2008.

[108] Genshiro Kitagawa. Non-gaussian state-space modeling of nonstationary time series. *Journal of the American statistical association*, 82(400):1032–1041, 1987.

[109] Peter K. Kitanidis. Quasi-linear geostatistical theory for inversing. *Water resources research*, 31(10):2411–2419, 1995.

[110] Ioannis Kontoyiannis and Mokshay Madiman. Sumset and inverse sumset inequalities for differential entropy and mutual information. *Information Theory, IEEE Transactions on*, 60(8):4503–4514, 2014.

[111] Samuel Kotz and Saralees Nadarajah. *Multivariate t distributions and their applications*. Cambridge University Press, Cambridge, 2004.

[112] Duc H. Le, Alexandre A. Emerick, Albert C. Reynolds, et al. An adaptive ensemble smoother with multiple data assimilation for assisted history matching. In *SPE Reservoir Simulation Symposium*. Society of Petroleum Engineers, 2015.

[113] François Le Gland, Valerie Monbet, and Vu-Duc Tran. Large sample asymptotics for the ensemble Kalman filter. Research Report RR-7014, INRIA, 2009.

[114] Erich Leo Lehmann and George Casella. *Theory of Point Estimation*, volume 31. Springer, 1998.

[115] Jing Lei and Peter Bickel. A moment matching ensemble filter for nonlinear non-Gaussian data assimilation. *Monthly Weather Review*, 139(12):3964–3973, 2011.

[116] P. F. J. Lermusiaux, A. R. Robinson, P. J. H. Haley, and W. G. Leslie. Advanced interdisciplinary data assimilation: Filtering and smoothing via error subspace statistical estimation. In *OCEANS'02 MTS/IEEE*, volume 2, pages 795–802. IEEE, 2002.

[117] Pierre F. J. Lermusiaux and A. R. Robinson. Data assimilation via error subspace statistical estimation. Part I: Theory and schemes. *Monthly Weather Review*, 127(7):1385–1407, 1999.

[118] John M. Lewis, Sivaramakrishnan Lakshmivarahan, and Sudarshan Dhall. *Dynamic Data Assimilation: A Least Squares Approach*, volume 13. Cambridge University Press, 2006.

[119] H. Li, E. Kalnay, and T. Miyoshi. Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter. *Quarterly Journal of the Royal Meteorological Society*, 135(639):523–533, 2009.

[120] Hong Li, Eugenia Kalnay, Takemasa Miyoshi, and Christopher M. Danforth. Accounting for model errors in ensemble data assimilation. *Monthly Weather Review*, 137(10):3407–3419, 2009.

[121] David M. Livings, Sarah L. Dance, and Nancy K. Nichols. Unbiased ensemble square root filters. *Physica D: Nonlinear Phenomena*, 237(8):1021–1028, 2008.

[122] A. C. Lorenc. Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 112(474):1177–1194, 1986.

[123] Andrew C. Lorenc. The potential of the ensemble Kalman filter for NWP - a comparison with 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, 129(595):3183–3203, 2003.

[124] Andrew C. Lorenc, Neill E. Bowler, Adam M. Clayton, Stephen R. Pring, and David Fairbairn. Comparison of hybrid-4DEnVar and hybrid-4DVar data assimilation methods for global NWP. *Monthly Weather Review*, 143(1):212–229, 2014.

[125] Rolf J. Lorentzen and Geir Nævdal. An iterative ensemble Kalman filter. *Automatic Control, IEEE Transactions on*, 56(8):1990–1995, 2011.

[126] E. N. Lorenz. Atmospheric predictability experiments with a large numerical model. *Tellus*, 34(6):505–513, 1982.

[127] Edward N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2):130–141, 1963.

[128] Edward N. Lorenz. Predictability: A problem partly solved. In *Proc. ECMWF Seminar on Predictability*, volume 1, pages 1–18, Reading, UK, 1996.

[129] Edward. N. Lorenz. Designing chaotic models. *Journal of the Atmospheric Sciences*, 62(5):1574–1587, 2005.

[130] Edward N. Lorenz and Kerry A. Emanuel. Optimal sites for supplementary weather observations: Simulation with a small model. *Journal of the Atmospheric Sciences*, 55(3):399–414, 1998.

[131] Dana Mackenzie. Ensemble Kalman filters bring weather models up to date. *Siam news*, 36(8):10–03, 2003.

[132] Jan Mandel, Lynn S. Bennethum, Jonathan D. Beezley, Janice L. Coen, Craig C. Douglas, Minjeong Kim, and Anthony Vodacek. A wildland fire model with data assimilation. *Mathematics and Computers in Simulation*, 79(3):584–606, 2008.

[133] Jan Mandel, Loren Cobb, and Jonathan D. Beezley. On the convergence of the ensemble Kalman filter. *Applications of Mathematics*, 56(6):533–541, 2011.

[134] Jan Mandel, Elhoucine Bergou, and Serge Gratton. 4DVAR by ensemble Kalman smoother. *arXiv preprint arXiv:1304.5271*, 2013.

[135] Justin G. McLay, Craig H. Bishop, and Carolyn A. Reynolds. Evaluation of the ensemble transform analysis perturbation scheme at nrl. *Monthly Weather Review*, 136(3):1093–1108, 2008.

[136] Sammy Metref, E. Cosme, C. Snyder, and P. Brasseur. A non-Gaussian analysis scheme using rank histograms for ensemble data assimilation. *Nonlinear Processes in Geophysics*, 21(4):869–885, 2014.

[137] Herschel L. Mitchell, P. L. Houtekamer, and Gérard Pellerin. Ensemble size, balance, and model-error representation in an ensemble Kalman filter. *Monthly Weather Review*, 130(11):2791–2808, 2002.

[138] L. Mitchell and A. Carrassi. Accounting for model error due to unresolved scales within ensemble Kalman filtering. *Quarterly Journal of the Royal Meteorological Society*, 2014.

[139] Takemasa Miyoshi, Eugenia Kalnay, and Hong Li. Estimating and including observation-error correlations in data assimilation. *Inverse Problems in Science and Engineering*, 21(3):387–398, 2013.

[140] Klaus Mosegaard, Albert Tarantola, et al. Probabilistic approach to inverse problems. *International Geophysics Series*, 81(A):237–268, 2002.

[141] Robb J. Muirhead. *Aspects of multivariate statistical theory*. John Wiley & Sons, Inc., New York, 1982. Wiley Series in Probability and Mathematical Statistics.

[142] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[143] Geir Nævdal, Kristian Thulin, Hans Julius Skaug, Sigurd Ivar Aanonsen, et al. Quantifying Monte Carlo uncertainty in the ensemble Kalman filter. *SPE Journal*, 16(01):172–182, 2011.

[144] Shin'ya Nakano. A prediction algorithm with a limited number of particles for state estimation of high-dimensional systems. In *Information Fusion (FUSION), 16th International Conference on*, pages 1356–1363. IEEE, 2013.

[145] Shin'ya Nakano. Hybrid algorithm of ensemble transform and importance sampling for assimilation of non-Gaussian observations. *Tellus A*, 66, 2014.

[146] Lars Nerger, Svenja Schulte, and Angelika Bunse-Gerstner. On the influence of model nonlinearity and localization on ensemble Kalman smoothing. *Quarterly Journal of the Royal Meteorological Society*, 140(684):2249–2259, 2014.

[147] C. Nicolis. Dynamics of model error: The role of unresolved scales revisited. *Journal of the Atmospheric Sciences*, 61(14):1740–1753, 2004.

[148] Shuli Niu, Yiqi Luo, Michael C. Dietze, Trevor F. Keenan, Zheng Shi, Jianwei Li, and F. Stuart Chapin III. The role of data assimilation in predictive ecology. *Ecosphere*, 5(5):65, 2014.

[149] Met Office. Met office numerical weather prediction models, 2015. Online. Retreived from `http://www.metoffice.gov.uk/research/modelling-systems/unified-model/weather-forecasting`. December 2015.

[150] Dean S. Oliver. Minimization for conditional simulation: Relationship to optimal transport. *Journal of Computational Physics*, 265:1–15, 2014.

[151] Dean S. Oliver and Yan Chen. Recent progress on reservoir history matching: a review. *Computational Geosciences*, 15(1):185–221, 2011.

[152] Dean S. Oliver, Albert C. Reynolds, and Ning Liu. *Inverse Theory for Petroleum Reservoir Characterization and History Matching*. Cambridge University Press, 2008.

[153] Ingram Olkin and S. N. Roy. On multivariate distribution theory. *The Annals of Mathematical Statistics*, pages 329–339, 1954.

[154] Edward Ott, Brian R. Hunt, Istvan Szunyogh, Aleksey V. Zimin, Eric J. Kostelich, Matteo Corazza, Eugenia Kalnay, D. J. Patil, and James A. Yorke. A local ensemble Kalman filter for atmospheric data assimilation. *Tellus A*, 56 (5):415–428, 2004.

[155] Edward Ott, Brian R. Hunt, Istvan Szunyogh, Aleksey V. Zimin, Eric J. Kostelich, Matteo Corazza, Eugenia Kalnay, D. J. Patil, and James A. Yorke. A local ensemble Kalman filter for atmospheric data assimilation. *Tellus A*, 56 (5):415–428, 2004.

[156] Luigi Palatella, Alberto Carrassi, and Anna Trevisan. Lyapunov vectors and assimilation in the unstable subspace: theory and applications. *Journal of Physics A: Mathematical and Theoretical*, 46(25):254020, 2013.

[157] Lies Peters, R. J. Arts, G. K. Brouwer, C. R. Geel, Stan Cullick, Rolf J. Lorentzen, Yan Chen, K. N. B. Dunlop, Femke C. Vossepoel, Rong Xu, et al. Results of the Brugge benchmark study for flooding optimization and history matching. *SPE Reservoir Evaluation & Engineering*, 13(3):391–405, 2010.

[158] Dinh Tuan Pham. Stochastic methods for sequential data assimilation in strongly nonlinear systems. *Monthly Weather Review*, 129(5):1194–1207, 2001.

[159] James E. Potter and Robert Gottlieb Stern. Statistical filtering of space navigation measurements. *Massachusetts Institute of Technology, Experimental Astronomy Laboratory*, 1963.

[160] Patrick N. Raanes, Alberto Carrassi, and Laurent Bertino. Extending the square root method to account for model noise in the ensemble Kalman filter. *Monthly Weather Review*, 143(10):3857–3873, 2015.

[161] Patrick Nima Raanes. On the ensemble Rauch-Tung-Striebel smoother and its equivalence to the ensemble Kalman smoother. *Quarterly Journal of the Royal Meteorological Society*, 2015.

[162] C. Radhakrishna Rao. *Linear statistical inference and its applications*, volume 22. John Wiley & Sons, 2009.

[163] Herbert E. Rauch, C. T. Striebel, and F. Tung. Maximum likelihood estimates of linear dynamic systems. *AIAA journal*, 3(8):1445–1450, 1965.

[164] Sai Ravela and Dennis McLaughlin. Fast ensemble smoothing. *Ocean Dynamics*, 57(2):123–134, 2007.

[165] Sebastian Reich and Colin J. Cotter. Large scale inverse problems. computational methods and applications in the earth sciences: Ensemble filter techniques for intermittent data assimilation. *Radon Series on Computational and Applied Mathematics*, 13:91–134, 2013.

[166] Rolf H. Reichle, Dennis B. McLaughlin, and Dara Entekhabi. Hydrologic data assimilation with the ensemble Kalman filter. *Monthly Weather Review*, 130(1): 103–114, 2002.

[167] Albert Coburn Reynolds, Alexandre Anoze Emerick, et al. History-matching production and seismic data in a real field case using the ensemble smoother with multiple data assimilation. In *SPE Reservoir Simulation Symposium*. Society of Petroleum Engineers, 2013.

[168] Christian Robert. *The Bayesian Choice: From Decision-theoretic Foundations to Computational Implementation*. Springer Verlag, 2007.

[169] William Sacher and Peter Bartello. Sampling errors in ensemble Kalman filtering. Part II: Application to a barotropic model. *Monthly Weather Review*, 137(5): 1640–1654, 2009.

[170] Pavel Sakov and Laurent Bertino. Relation between two common localisation methods for the EnKF. *Computational Geosciences*, 15(2):225–237, 2011.

[171] Pavel Sakov and Peter R. Oke. A deterministic formulation of the ensemble Kalman filter: an alternative to ensemble square root filters. *Tellus A*, 60(2): 361–371, 2008.

[172] Pavel Sakov and Peter R. Oke. Implications of the form of the ensemble transformation in the ensemble square root filters. *Monthly Weather Review*, 136(3): 1042–1053, 2008.

[173] Pavel Sakov, Geir Evensen, and Laurent Bertino. Asynchronous data assimilation with the EnKF. *Tellus A*, 62(1):24–29, 2010.

[174] Pavel Sakov, F. Counillon, L. Bertino, K. A. Lisæter, P. R. Oke, and A. Korablev. TOPAZ4: an ocean-sea ice data assimilation system for the North Atlantic and Arctic. *Ocean Science*, 8(4):633, 2012.

[175] Pavel Sakov, Dean S. Oliver, and Laurent Bertino. An iterative EnKF for strongly nonlinear systems. *Monthly Weather Review*, 140(6):1988–2004, 2012.

[176] Barry Saltzman. Finite amplitude free convection as an initial value problem – I. *Journal of the Atmospheric Sciences*, 19(4):329–341, 1962.

[177] Simo Särkkä. *Recursive Bayesian inference on stochastic differential equations*. PhD thesis, Helsinki University of Technology, 2006.

[178] Simo Särkkä. Bayesian estimation of time-varying systems: Discrete-time systems, 2010. Unpublished. Retrieved from `http://www.lce.hut.fi/~ssarkka/course_k2010/full_course_booklet.pdf`. December 2015.

[179] Simo Särkkä. *Bayesian filtering and smoothing*, volume 3. Cambridge University Press, 2013.

[180] G. A. F. Seber. *Multivariate Observations*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1984.

[181] Jeffrey Shaman and Alicia Karspeck. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*, 109(50):20425–20430, 2012.

[182] S. Sherman. Non-mean-square error criteria. *Information Theory, IRE transactions on*, 4(3):125–126, 1958.

[183] Emir H. Shuford Jr., Arthur Albert, and H. Edward Massengill. Admissible probability measurement procedures. *Psychometrika*, 31(2):125–145, 1966.

[184] Glenn Shutts. A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quarterly Journal of the Royal Meteorological Society*, 131 (612):3079–3102, 2005.

[185] J. Skjervheim and G. Evensen. An ensemble smoother for assisted history matching. In *SPE Reservoir Simulation Symposium*, pages 1–15, 2011.

[186] Julia Slingo and Tim Palmer. Uncertainty in weather and climate prediction. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 369(1956):4751–4767, 2011.

[187] Chris Snyder. Particle filters, the "optimal" proposal and high-dimensional systems. In *Proceedings of the ECMWF Seminar on Data Assimilation for Atmosphere and Ocean*, 2011.

[188] A. S. Stordal, H. A. Karlsen, G. Nævdal, H. J. Skaug, and B. Vallès. Bridging the ensemble Kalman filter and particle filters: the adaptive Gaussian mixture filter. *Computational Geosciences*, 15(2):293–305, 2011.

[189] Jonathan R. Stroud, Michael L. Stein, Barry M. Lesht, David J. Schwab, and Dmitry Beletsky. An ensemble Kalman filter and smoother for satellite data assimilation. *Journal of the American Statistical Association*, 105(491):978–990, 2010.

[190] Michael K. Tippett, Jeffrey L. Anderson, Craig H. Bishop, Thomas M. Hamill, and Jeffrey S. Whitaker. Ensemble square root filters. *Monthly Weather Review*, 131(7):1485–1490, 2003.

[191] Ricardo Todling and Stephen E. Cohn. Suboptimal schemes for atmospheric data assimilation based on the Kalman filter. *Monthly Weather Review*, 122(11): 2530–2557, 1994.

[192] Lloyd N. Trefethen. *Finite Difference and Spectral Methods for Ordinary and Partial Differential Equations*. Unpublished, 1996. Retrieved from `http://people.maths.ox.ac.uk/trefethen/pdetext.html`. December 2015.

[193] Lloyd N. Trefethen. *Spectral methods in MATLAB*, volume 10 of *Software, Environments, and Tools*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000.

[194] Lloyd N. Trefethen and David Bau, III. *Numerical linear algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[195] Peter Jan van Leeuwen. Comment on "Data assimilation using an ensemble Kalman filter technique". *Monthly Weather Review*, 127(6):1374–1377, 1999.

[196] Peter Jan van Leeuwen. Particle filtering in geophysical systems. *Monthly Weather Review*, 137(12):4089–4114, 2009.

[197] M. Verlaan and A. W. Heemink. Tidal flow forecasting using reduced rank square root filters. *Stochastic Hydrology and Hydraulics*, 11(5):349–368, 1997.

[198] Qing Wang, Sanjeev R. Kulkarni, and Sergio Verdú. Divergence estimation for multidimensional densities via-nearest-neighbor distances. *Information Theory, IEEE Transactions on*, 55(5):2392–2405, 2009.

[199] Xuguang Wang and Craig H. Bishop. A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *Journal of the Atmospheric Sciences*, 60(9):1140–1158, 2003.

[200] Xuguang Wang and Ting Lei. GSI-based four dimensional ensemble-variational (4DEnsVar) data assimilation: formulation and single resolution experiments with real data for NCEP Global Forecast System. *Monthly Weather Review*, 142 (9):3303–3325, 2014.

[201] Xuguang Wang, Craig H. Bishop, and Simon J. Julier. Which is better, an ensemble of positive-negative pairs or a centered spherical simplex ensemble? *Monthly Weather Review*, 132(7):1590–1605, 2004.

[202] Y. Wang, F. Counillon, and L. Bertino. Alleviating the bias induced by the linear analysis update with an isopycnal ocean model. *Quarterly Journal of the Royal Meteorological Society*, 2015.

[203] Steven V. Weijs, Ronald Van Nooijen, and Nick Van De Giesen. Kullback-Leibler divergence as a forecast skill score with classic reliability-resolution-uncertainty decomposition. *Monthly Weather Review*, 138(9):3387–3399, 2010.

[204] Jeffrey S. Whitaker and Thomas M. Hamill. Ensemble data assimilation without perturbed observations. *Monthly Weather Review*, 130(7):1913–1924, 2002.

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

[206] Jeffrey S. Whitaker, Gilbert P. Compo, Xue Wei, and Thomas M. Hamill. Reanalysis without radiosondes using ensemble data assimilation. *Monthly Weather Review*, 132(5):1190–1200, 2004.

[207] Jeffrey S. Whitaker, Thomas M. Hamill, Xue Wei, Yucheng Song, and Zoltan Toth. Ensemble data assimilation with the NCEP global forecast system. *Monthly Weather Review*, 136(2):463–482, 2008.

[208] C. K. Wikle and L. M. Berliner. A Bayesian tutorial for data assimilation. *Physica D: Nonlinear Phenomena*, 230(1-2):1–16, 2007.

[209] Daniel S. Wilks. *Statistical Methods in the Atmospheric Sciences*, volume 100. Academic Press, 2011.

[210] C. Wunsch. *Discrete Inverse and State Estimation Problems: With Geophysical Fluid Applications*. Cambridge University Press, 2006.

[211] Xiaozhen Xiong, Ionel Michael Navon, and Bahri Uzunoglu. A note on the particle filter with posterior Gaussian resampling. *Tellus A*, 58(4):456–460, 2006.

[212] Benjamin F. Zaitchik, Matthew Rodell, and Rolf H. Reichle. Assimilation of GRACE terrestrial water storage data into a land surface model: Results for the Mississippi river basin. *Journal of Hydrometeorology*, 9(3):535–548, 2008.

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

[214] Dusanka Zupanski and Milija Zupanski. Model error estimation employing an ensemble data assimilation approach. *Monthly Weather Review*, 134(5):1337–1354, 2006.