ISBN 978-82-326-3598-6 (printed ver.)
ISBN 978-82-326-3599-3 (electronic ver.)
ISSN 1503-8181

NTNU
Norwegian University of
Science and Technology

Jacob Skauvold

Doctoral theses at NTNU, 2018:404

NTNU

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Information Technology and Electrical
Engineering
Department of Mathematical Sciences

Doctoral theses at NTNU, 2018:404

Jacob Skauvold

Ensemble-based Data Assimilation Methods Applied to Geological Process Modeling

NTNU
Norwegian University of
Science and Technology

Jacob Skauvold

Ensemble-based Data Assimilation Methods Applied to Geological Process Modeling

Thesis for the Degree of Philosophiae Doctor

Trondheim, December 2018

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

© Jacob Skauvold

ISBN 978-82-326-3598-6 (printed ver.)
ISBN 978-82-326-3599-3 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2018:404

Printed by NTNU Grafisk senter

Abstract

This thesis is a collection of three research papers about data assimilation, and its application to geological process modeling. Data assimilation is the task of bringing information from observations into a statistical model of the evolution of a dynamical system. Paper I applies the ensemble Kalman filter (EnKF), a popular data assimilation method, to the problem of conditioning GPM, a geological simulation model, to measurements of the real rock layers that were formed by the simulated sedimentation process. Paper II concerns the implicit equal-weights particle filter (IEWPF), a data assimilation method constructed to make the most of a small ensemble in a high-dimensional and strongly non-linear setting. The original formulation of this particle filter produces biased estimates because of a problem with the update step. Paper II proposes a revised version of the IEWPF which potentially eliminates the bias. Paper III focuses once again on the EnKF, which must estimate covariances between state variables in order to update its ensemble with respect to observations. Errors in covariance estimates can have a large effect on the update step, significantly reducing the quality of the estimates produced by the filter. Paper III explores the idea of introducing simple parametric covariance representations into this part of the EnKF in order to make the covariance estimation more robust.

Acknowledgements

First, I would like to thank my main supervisor, Jo Eidsvik, for excellent supervision, advice, direction, counselling, guidance and other synonyms. I would also like to thank my co-supervisor, Henning Omre, for frequent encouragement and occasional discussion, and for urging me to travel abroad during my studies. Thanks to Hilde Grude Borgos, Bjørn Harald Fotland and the others at Schlumberger, Tananger for providing access to data and software, for illuminating discussions, and for acting as de facto tech support. Thanks to the technical group at the Department of Mathematical Sciences for acting as actual tech support, and for being patient with me on the several occasions when my poorly crafted scripts made a mess on the department computing servers. Thanks to my friends, colleagues and fellow students for making the department a good place to study and work. Thanks to Peter Jan van Leeuwen and Javier Amezcua for making me feel welcome during my stay in Reading. To Kristoffer, thanks for lunch. To Xin, thanks for being an exemplary office mate and traveling companion. To my family, thanks for not giving me a hard time about not calling or writing more often. To Irene, thanks for being there.

Jacob Skauvold

Jacob Skauvold

Trondheim

August 2018

Contents

Introduction

Earth sciences rely increasingly on computer simulations—numerical models which mimic the physical processes that shape, sustain and evolve the natural environment. Since the middle of the twentieth century, steady improvements in computing power have led to the adoption of ever more sophisticated models capable of increasingly faithful representation of the geophysical processes and phenomena under study. Today, computer simulations are indispensable workhorses of oceanography, hydrology, climate science and meteorology where they continue to provide critical insights. In other applied sciences, numerical modeling is brought to bear on long standing problems with encouraging results, providing bountiful research opportunities for new generations of scientists.

Gaining knowledge about the world through the use of computer simulations requires that the simulations be anchored to reality. Suppose a numerical model is used to make a prediction about the future state of the atmosphere. If we are to take the prediction seriously, we need a reason to believe that the model resembles the actual atmosphere as it really is, not merely as it could have been. In other words, it's not enough that the simulated atmosphere obeys the laws of physics. It also has to match the actual atmosphere as we observe it.

Data assimilation is the task of establishing a connection between model and reality. The solutions are many and varied but they all work by bringing models closer to data. Adjusting the simulated system so that it matches measurements of the actual system grounds the simulation in reality and justifies its use for prediction and inference.

Data assimilation

This section gives a general introduction to data assimilation, placing it in the context of inverse problems, and providing an overview of the kinds of approaches that are used to solve data assimilation problems.

Inverse problems

Physical theories allow prediction. In most cases, given a sufficiently detailed description of a physical system, a theory can predict the outcome of a future measurement to some degree of precision. This kind of prediction task is sometimes referred to as a *forward problem* because of the sense of motion from cause to effect, i.e. from the latent system state to the measurement outcome (Tarantola, 2005). An *inverse problem*, by contrast, is the task of inferring the state or properties of a system on the basis of an observation or measurement. Figure 1 illustrates the complementary relation between the prediction and inversion tasks.

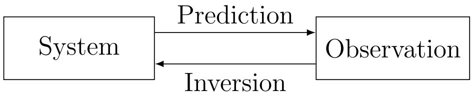


FIGURE 1: Forward problem and inverse problem.

Let $\mathbf{x} \in \mathbb{R}^{N_x}$ denote a representation of the unknown system state in the form of a vector. This could be a list of physical quantities describing the system or values of a spatial field discretized onto a grid or lattice. Further let $\mathbf{y} \in \mathbb{R}^{N_y}$ denote an observation vector containing values obtained by carrying out some measurement operation on the system. We also need a modeling operator $\mathcal{G}$ to represent the action of the measurement operation on the system state. Thus, $\mathcal{G}(\mathbf{x})$ is the expected value of $\mathbf{y}$ provided that $\mathbf{x}$ is the true state of the system and that $\mathcal{G}$ accurately represents the actual measurement procedure. This procedure could be a direct sampling of some or all the elements of $\mathbf{x}$, or the relationship could be more indirect. In this notation, the generic form of the inverse problem can be written as a measurement equation,

$$\mathcal{G}(\mathbf{x}) = \mathbf{y}, \tag{0.1}$$

where our objective is to solve for $\mathbf{x}$ given $\mathbf{y}$. The inverse problem (0.1) is called *well-posed* if *(i)* for every observation vector $\mathbf{y}$ there exists a solution $\mathbf{x}$ that satisfies (0.1), *(ii)* the solution $\mathbf{x}$ is unique, and *(iii)* the solution is stable with respect to perturbations of $\mathbf{y}$ in the sense that if $\mathcal{G}(\mathbf{x}) = \mathbf{y}$ and $\mathcal{G}(\mathbf{x}') = \mathbf{y}'$ then $\mathbf{x}$ will be close to $\mathbf{x}'$ as long as $\mathbf{y}$ is close to $\mathbf{y}'$. Problems which do not possess these three properties are called *ill-posed* (Vogel, 2002).

Inverse problems that arise in practical applications—such as the use of statistical estimation techniques to solve problems related to numerical modeling in geology, as described in the next section—tend to be ill-posed. When observations are subject to measurement uncertainty, there may not exist a solution which satisfies (0.1) exactly. Furthermore, measurement outcomes can be multiply realizable, meaning that several distinct system states yield the same measurement outcome when the modeling operator is applied to them. Consider for instance an averaging operator $\mathcal{G} : \mathbf{x} \mapsto \bar{\mathbf{x}}$ which computes the average of all the elements of $\mathbf{x}$. If $\mathbf{x}'$ is defined by taking $\mathbf{x}$ and adding a small quantity to one element, and subtracting the same quantity from some other element, then it is clear that the two averages are equal so that $\mathcal{G}(\mathbf{x}') = \mathcal{G}(\mathbf{x})$ even though $\mathbf{x}' \neq \mathbf{x}$. Stability poses a smaller challenge for practical applications, but if the modeling operator involves forward integration of a dynamical system exhibiting chaotic behaviour, small changes in $\mathbf{x}$ may lead to disproportionately large changes in $\mathbf{y}$, i.e. the observation vector may be unstable with respect to the system state. This does not entail that $\mathbf{x}$ is unstable with respect to $\mathbf{y}$, but it could still be harder to guarantee the stability of the inverse problem solution in such cases.

Ill-posed inverse problems do not admit exact and unique solutions. Nevertheless, useful solutions are available if we are willing to forgo one or both of these properties. Regularization replaces the ill-posed problem with a related well-posed problem, in the hope that the unique solution to the well-posed problem will be a good approximate solution of the original problem. Tikhonov regularization (Tikhonov et al., 2013) seeks the solution $\mathbf{x}$ which minimizes the *smoothing functional*

$$M_\Gamma[\mathbf{x}] = \|\mathcal{G}(\mathbf{x}) - \mathbf{y}\|^2 + \|\mathbf{\Gamma}\mathbf{x}\|^2. \tag{0.2}$$

The minimizing solution $\mathbf{x}_\Gamma = \arg\min_{\mathbf{x}} M_\Gamma[\mathbf{x}]$ depends on the regularization matrix $\mathbf{\Gamma}$. The second term of the smoothing functional introduces a preference for some solutions over others, thus resolving the ambiguity responsible for the ill-posedness of the original inverse problem. If $\mathbf{0}$ is not a reasonable solution, the second term of (0.2) can be changed to $\|\mathbf{\Gamma}(\mathbf{x} - \mathbf{x}_0)\|^2$. One then interprets $\mathbf{x}_0$ as a reasonable initial guess that the solution is expected to be close to. Then the data $\mathbf{y}$ is allowed to determine in which direction and by how much to deviate from the initial guess.

For a linear modeling operator, say $\mathcal{G}(\mathbf{x}) = \mathbf{g}_0 + \mathbf{G}(\mathbf{x} - \mathbf{x}_0)$, the smoothing functional is

$$M_\Gamma[\mathbf{x}] = \|\mathbf{g}_0 + \mathbf{G}(\mathbf{x} - \mathbf{x}_0) - \mathbf{y}\|^2 + \|\mathbf{\Gamma}(\mathbf{x} - \mathbf{x}_0)\|^2, \tag{0.3}$$

with a unique minimum at

$$\mathbf{x}_\Gamma = \mathbf{x}_0 + (\mathbf{G}^T\mathbf{G} + \mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\mathbf{G}^T(\mathbf{y} - \mathbf{g}_0). \tag{0.4}$$

Another approach to obtaining an approximate solution of an ill-posed inverse problem of the form (0.1) is *Bayesian inversion.* In a Bayesian setting, the unknown system state is regarded as a random vector $\mathbf{X}$ with an a priori probabilty density function $p(\mathbf{x})$. This prior distribution represents background information about $\mathbf{X}$, known in advance of experience. The observation vector is also considered a random vector $\mathbf{Y}$, with the actual measurement outcome $\mathbf{y}$ a realization of this random vector. The deterministic modeling operator $\mathcal{G}$ is generalized to the conditional probability distribution of $\mathbf{Y}$ given that $\mathbf{X} = \mathbf{x}$, with $\mathcal{G}(\mathbf{x})$ usually specifying the expected value $\mathrm{E}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$. When viewed as a function of $\mathbf{x}$ rather than $\mathbf{y}$ the conditional probability density function $p(\mathbf{y}|\mathbf{x})$ is referred to as the likelihood. The immediate goal of Bayesian inference is determining or approximating the posterior probability distribution of $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$, i.e. the conditional distribution of the unknown state vector given the available information. Once a representation of the posterior distribution is obtained, all inference—such as the construction of point and interval estimates—proceeds on the basis of that representation.

Bayes' theorem gives the posterior density in terms of the prior density and likelihood (Gelman et al., 1995),

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}. \tag{0.5}$$

Since $p(\mathbf{y})$ does not depend on $\mathbf{x}$, it can safely be omitted when $\mathbf{y}$ is fixed, yielding the simplified relationship

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{x})p(\mathbf{y}|\mathbf{x}). \tag{0.6}$$

Returning to the special case of a linear modeling operator, suppose we have a multivariate Gaussian prior on $\mathbf{X}$ with mean $\mathbf{x}_0$ and covariance matrix $\mathbf{\Sigma}$. Further suppose we have a likelihood specifying that the conditional distribution of $\mathbf{Y}$ given $\mathbf{X} = x$ is also multivariate Gaussian, with mean $\mathcal{G}(\mathbf{x}) = \mathbf{g}_0 + \mathbf{G}(\mathbf{x} - \mathbf{x}_0)$ and covariance matrix $\mathbf{C}$. In this case the prior density is

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^{N_x}|\mathbf{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathrm{x}_0)^T \mathbf{\Sigma}^{-1}(\mathbf{x} - \mathrm{x}_0)\right) \tag{0.7}$$

while the likelihood density is

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^{N_y}|\mathbf{C}|}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{g}_0 - \mathbf{G}(\mathbf{x} - \mathbf{x}_0))^T \mathbf{C}^{-1}(\mathbf{y} - \mathbf{g}_0 - \mathbf{G}(\mathbf{x} - \mathbf{x}_0))\right). \tag{0.8}$$

Multiplying the densities in (0.7) and (0.8) together, we find (Johnson and Wichern, 1998) that the posterior density is itself multivariate Gaussian, with mean

$$\begin{aligned} \hat{\mathbf{x}} &= \mathbf{x}_0 + \mathbf{\Sigma}\mathbf{G}^T(\mathbf{G}\mathbf{\Sigma}\mathbf{G}^T + \mathbf{C})^{-1}(\mathbf{y} - \mathbf{g}_0) \\ &= \mathbf{x}_0 + (\mathbf{G}^T\mathbf{C}^{-1}\mathbf{G} + \mathbf{\Sigma}^{-1})^{-1}\mathbf{G}^T\mathbf{C}^{-1}(\mathbf{y} - \mathbf{g}_0) \end{aligned} \tag{0.9}$$

and covariance matrix

$$\widehat{\mathbf{\Sigma}} = \mathbf{\Sigma} - \mathbf{\Sigma}\mathbf{G}^T(\mathbf{G}\mathbf{\Sigma}\mathbf{G}^T + \mathbf{C})^{-1}\mathbf{G}\mathbf{\Sigma}. \tag{0.10}$$

If we choose the covariance matrices in the prior and likelihood such that $\mathbf{C} = \mathbf{I}$ and $\mathbf{\Sigma}^{-1} = \mathbf{\Gamma}^T\mathbf{\Gamma}$, then the posterior mean $\hat{\mathbf{x}}$ in (0.9) is equal to the solution $\mathbf{x}_{\mathbf{\Gamma}}$ in (0.4) of the Tikhonov-regularized inverse problem. Moreover, with this choice of covariance matrices, the log-posterior density is

$$\begin{aligned} \log p(\mathbf{x}|\mathbf{y}) &= -\frac{1}{2}(\mathbf{y} - \mathbf{g}_0 - \mathbf{G}(\mathbf{x} - \mathbf{x}_0))^T(\mathbf{y} - \mathbf{g}_0 - \mathbf{G}(\mathbf{x} - \mathbf{x}_0)) \\ &\quad - \frac{1}{2}(\mathbf{x} - \mathrm{x}_0)^T\mathbf{\Gamma}^T\mathbf{\Gamma}(\mathbf{x} - \mathrm{x}_0) \\ &= -\frac{1}{2}\left(\|\mathbf{y} - \mathbf{g}_0 - \mathbf{G}(\mathbf{x} - \mathbf{x}_0)\|^2 + \|\mathbf{\Gamma}(\mathbf{x} - \mathbf{x}_0)\|^2\right) \\ &= -\frac{1}{2}M_{\mathbf{\Gamma}}[\mathrm{x}] \end{aligned} \tag{0.11}$$

where $M_{\mathbf{\Gamma}}[\mathbf{x}]$ is the smoothing functional in (0.3). In this example, therefore, minimizing the smoothing functional in the regularization formulation is equivalent to maximizing the posterior density in the Bayesian formulation. Hence, the Bayesian framework offers an interpretation of regularization whereby the introduction of a seemingly arbitrary regularization matrix $\mathbf{\Gamma}$ can be justified as imposing a specific prior distribution on $\mathbf{X}$.

Another way to view the equivalence between the regularization and Bayesian inference approaches to this example problem is that the imposition of a prior distribution on the unknown target variable has a regularizing effect on the ill-posed inverse problem. Crucially, this regularizing influence can also be relied upon in more general cases involving non-linear modeling operators. The main difference between the two approaches is that in the Bayesian case one incorporates explicit assumptions about the unknown system state via the probabilistic model, and then regularization follows as a consequence of these assumptions. In the regularization case, on the other hand, a regularization matrix and corresponding smoothing functional are introduced as a matter of necessity, usually in order to facilitate practical computation of the solution. The assumptions about properties of the target variable that go along with the regularization are implicit in this case. This difference in particular makes Bayesian inversion an attractive solution method for inverse problems.

Data assimilation as an inverse problem

When discussing the Bayesian approach to inversion in the previous section, we considered the problem of estimating a random vector $\mathbf{x}$. This section will concern Bayesian inference, or inversion, for processes that evolve over time. We begin by introducing stochastic processes, which are sets of related random variables, linked by an index variable. In applied modeling, the index variable often refers to time, space or both. We discuss indexing briefly before moving on to discretization of spatial domains and time intervals, and presenting a convenient vector notation for spatially discretized fields.

A *stochastic process* is a set $\{X_t : t \in \mathbb{T}\}$ where t is an index variable belonging to the index set $\mathbb{T}$, and for each value of the index variable, X_t is a random variable. The index variable can represent time, in which case the index set is some subset of $\mathbb{R}$. For models of spatiotemporal processes, the index variable can be a combination of spatial coordinates and time, with the index set defined as a cartesian product $\mathcal{D} \times \mathcal{T}$ of a spatial domain and a time interval.

The random variables X_t that belong to a stochastic process take values in state space, and statistical models using this formulation are termed state space models. A state space could be a space of continuous functions, but in an applied context it is common to work with a discretization of the spatial domain. A spatially continuous random field

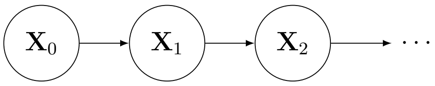


FIGURE 2: Directed acyclic graph representing the causal relationship between random variables belonging to a stochastic process, considered at different time points.

$X(\mathbf{s})$, indexed by the coordinates $\mathbf{s} \in \mathcal{D}$ is then represented by the random vector

$$\mathbf{X} = (X_1, X_2, \ldots, X_N)^T = (X(\mathbf{s}_1), X(\mathbf{s}_2), \ldots, X(\mathbf{s}_N))^T \tag{0.12}$$

where $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N\}$ is a set of discrete points in the spatial domain $\mathcal{D}$. These points could form a regular lattice, but could also be unstructured. The number of points N depends on the spatial resolution, which may differ between directions. For example, some three-dimensional grids have a higher vertical than lateral resolution.

Numerical models of temporal and spatiotemporal processes are usually discrete-time. This is true even when the target process is assumed to be continuous in time, and when the underlying mathematical model has a continuous-time formulation, such as a set of partial differential equations. To get a discrete-time representation of a spatiotemporal process $X(\mathbf{s}, t)$ with $\mathbf{s} \in \mathcal{D}$ and $t \in \mathcal{T}$ that is continuous in both space and time, we can discretize $\mathcal{T}$ into a sequence of time points $t_0 \leq t_1 \leq \ldots \leq t_K$ where t_0 and t_K are the first and last times in $\mathcal{T}$ respectively. The number of time points $K + 1$ depends on the temporal resolution, which is the duration $\Delta t = t_k - t_{k-1}$ between each pair of consecutive time points. Time points are usually equidistant, so that Δt is the same for all k, but they need not be.

With both space and time discretized, the spatiotemporal process $X(\mathbf{s}, t)$ is represented by a sequence of vectors

$$\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_K \tag{0.13}$$

each one being a spatial discretization $(X(\mathbf{s}_1, t_k), X(\mathbf{s}_2, t_k), \ldots, X(\mathbf{s}_N, t_k))$ of the "time slice" random field $\{X(\mathbf{s}, t_k) : \mathbf{s} \in \mathcal{D}\}$. In the remainder of this section, we will use the vector notation of (0.13) for the general state space model with a process evolving through time. This process could be, but does not have to be, a spatial field.

The causal structure of a state space model can be illustrated by a graph, as in Figure 2. Each node in the graph represents a random variable, or random vector. Here, the nodes correspond to states at different time points. A directed edge from one node to another, say from $\mathbf{X}_k$ to $\mathbf{X}_{k+1}$ means that there is a causal relationship between the variables

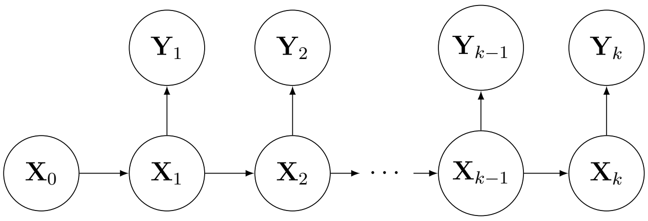


FIGURE 3: Graph representation of the hidden Markov model.

represented by these nodes and that the direction of causality is, unsurprisingly, forward in time. Statistically this means that we are able to specify the conditional probability distribution of $\mathbf{X}_{k+1}$ given $\mathbf{X}_k$. The graphical model of Figure 2 constitutes a first-order Markov chain, meaning that the conditional distribution of the current state $\mathbf{X}_k$ given the entire history of past states $\mathbf{X}_0$, $\mathbf{X}_1, \ldots, \mathbf{X}_{k-1}$ in fact depends only on the most recent past state. In probability density terms, we may therefore write

$$p(\mathbf{x}_k|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{k-1}) = p(\mathbf{x}_k|\mathbf{x}_{k-1}). \tag{0.14}$$

It is frequently assumed that the transition density $p(\mathbf{x}_{k+1}|\mathbf{x}_k)$ is Gaussian, i.e.

$$[\mathbf{X}_{k+1}|\mathbf{x}_k] \sim N(\mathcal{M}(\mathbf{x}_k), \mathbf{Q}), \tag{0.15}$$

where $\mathbf{Q}$ is a model error covariance matrix and $\mathcal{M}$ is a forward model operator. This is an example of additive model error, but the relationship between $\mathbf{x}_k$ and $\mathbf{X}_{k+1}$ may be expressed more generally as

$$\mathbf{X}_k = \mathcal{M}(\mathbf{x}_{k-1}, \mathbf{V}_k), \tag{0.16}$$

which allows the model error $\mathbf{V}_k$ to be combined with the previous state in non-additive ways, including multiplicative error.

The statistical model underlying data assimilation is the hidden Markov model (HMM), illustrated by the graph in Figure 3. The HMM links a sequence of states $\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, \ldots$ of a Markov process evolving according to a dynamical model equation of the form (0.16) to a sequence of observations $\mathbf{Y}_1, \mathbf{Y}_2, \ldots$ of the Markov process at different times. While the horizontal arrows in Figure 3 represent the causal structure imposed by the dynamical forward model $\mathcal{M}$, the vertical arrows represent the relationship between the state at time t_k and the observation $\mathbf{y}_k$ of the state at that time. Nodes that are not directly connected in the graph are conditionally independent given the intermediate nodes. This means, for example that since the causal link between $\mathbf{X}_k$ and $\mathbf{Y}_{k+1}$ goes through $\mathbf{X}_{k+1}$, observing the value of $\mathbf{Y}_{k+1}$ will not provide any new information about $\mathbf{X}_k$ if the value of $\mathbf{X}_{k+1}$ is already known. In many cases, the conditional distribution of $\mathbf{Y}_k$ given $\mathbf{x}_k$ is assumed to be Gaussian,

$$[\mathbf{Y}_k|\mathbf{x}_k] \sim N(\mathcal{H}(\mathbf{x}_k), \mathbf{R}), \tag{0.17}$$

where $\mathbf{R}$ is an observation error covariance matrix, and $\mathcal{H}$ is an observation operator. This is a special case of additive observation error, and a more general expression of the relationship $\mathbf{x}_k$ and $\mathbf{Y}_k$ is

$$\mathbf{Y}_k = \mathcal{H}(\mathbf{x}_k, \mathbf{W}_k), \tag{0.18}$$

where $\mathbf{W}_k$ is the observation error, which could be additive, multiplicative or have some other form, like the model error $\mathbf{V}_k$ in (0.16). Taken together, the model and observation equations (0.16) and (0.18) combined with suitable error probability densities $p(\mathbf{v}_k)$ and $p(\mathbf{w}_k)$ provide a complete description of the HMM graphical model of Figure 3.

The joint distribution of all variables from time t_0 up to and including time t_K can now be conveniently expressed as

$$p(\mathbf{x}_{0:K}, \mathbf{y}_{1:K}) = p(\mathbf{x}_0) \prod_{k=1}^{K} p(\mathbf{x}_k|\mathbf{x}_{k-1}) p(\mathbf{y}_k|\mathbf{x}_k). \tag{0.19}$$

A consequence of this product decomposition is that in principle the joint density can be computed sequentially, or recursively, by starting with $p(\mathbf{x}_0)$ and alternating between forecasting

$$p(\mathbf{x}_{k+1}|\mathbf{y}_{1:k}) = \int p(\mathbf{x}_{k+1}|\mathbf{x}_k) p(\mathbf{x}_k|\mathbf{y}_{1:k}) d\mathbf{x}_k \tag{0.20}$$

and updating

$$p(\mathbf{x}_{k+1}|\mathbf{y}_{1:k+1}) = \frac{p(\mathbf{x}_{k+1}|\mathbf{y}_{1:k}) p(\mathbf{y}_{k+1}|\mathbf{x}_{k+1})}{p(\mathbf{y}_{k+1})}. \tag{0.21}$$

In practice, however, the marginal density of the observation vector $\mathbf{y}_k$ cannot be computed exactly (except in special cases, see the discussion on the Kalman filter on page 15). This motivates the use of Monte Carlo methods and ensemble based inference methods to avoid having to compute $p(\mathbf{y}_k)$, and we will discuss these in the next section.

Data assimilation (DA) is used to refer to a range of inference procedures whereby observations are assimilated into a statistical model of a dynamical system. In HMM terms, DA methods compute or approximate the conditional distribution of some or all of the $\mathbf{X}_k$s given some or all of the $\mathbf{y}_k$s. Three commonly distinguished kinds of DA problem are *filtering*, *smoothing* and *prediction*.

Filtering entails inference of the state $\mathbf{X}_k$ at time t_k from only those observations $\mathbf{y}_{1:k}$ that are available at time t_k. Smoothing consists of inferring the entire trajectory $\mathbf{X}_{0:K}$ from all the observations $\mathbf{y}_{1:K}$. Prediction is estimation of the state $\mathbf{X}_l$ at time t_l from the observations $\mathbf{y}_{1:k}$ available at time t_k when $l > k$. The distinction between these problem types is not always sharp, and solving one of them might entail solving another as a part of the full solution. For example, an estimate of $\mathbf{X}_K$ given $\mathbf{y}_{1:K}$ is the solution of the filtering problem when $k = K$, but it is also a part of the smoothing solution. In fact, the full smoothing solution can be computed, at least in principle, by starting from $p(\mathbf{x}_K|\mathbf{y}_{1:K})$ and doing a *backwards pass* of sequential updates analogous to (0.21). Similarly, to predict $\mathbf{X}_l$ based on $\mathbf{y}_{1:k}$ when $l > k$, one can start from $p(\mathbf{x}_k|\mathbf{y}_{1:k})$ and forecast from time t_k to time t_l using an expression similar to (0.20). In the next section, we describe practical methods for obtaining approximate solutions of the filtering and smoothing problems.

Methods for data assimilation

There exist several kinds of methods for practical data assimilation. We begin by discussing methods that are used for smoothing, i.e. inference about the entire trajectory of states. This includes variational methods, which are based on optimization, and Monte Carlo methods, which make use of random sampling to explore the posterior distribution. We then move on to methods for filtering, i.e, estimating the current state given all currently available data. This includes the Kalman filter, the ensemble Kalman filter and the particle filter.

It should be noted that this way of dividing up data assimilation methods into smoothers and filters is somewhat arbitrary. Monte Carlo methods can be used for filtering as well as smoothing purposes, and smoothing problems can be solved sequentially. Nevertheless, emphasising typical areas of application for the various kinds of methods is a convenient way to give a brief overview, and to highlight important differences between methods.

Variational data assimilation

Consider a hidden Markov model specified by a model equation of the form (0.16) and an observation equation of the form (0.18). The forward model operator $\mathcal{M}$ determines how the state changes from one time step to the next, and the observation operator $\mathcal{H}$ describes the observation mechanism, and can be used to predict measurement outcomes. The assimilation time interval is divided into K time steps, and observations $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M$ are available at some or all of these time steps.

If the goal is to estimate the entire trajectory of states, $\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_K$, then we are solving the smoothing problem. If all we need is an estimate of the final state $\mathbf{X}_K$, then we are solving the filtering problem. Either way, variational data assimilation can be used.

The posterior density of the trajectory $\mathbf{X}_{0:K}$ is

$$p(\mathbf{x}_{0:K}|\mathbf{y}_{1:M}) = cp(\mathbf{x}_0) \prod_{k=1}^{K} p(\mathbf{x}_k|\mathbf{x}_{k-1}) \prod_{m=1}^{M} p(\mathbf{y}_m|\mathbf{x}_{k_m}), \tag{0.22}$$

which is similar to the HMM decomposition (0.19), except that the transition density and likelihood factors are indexed separately to account for observations not necessarily being available at every time step. The normalization constant c comes from the denominator in Bayes' theorem and is not required to locate the posterior mode or mean, or to determine the shape of the posterior density. The first product in (0.22) consists of transition densities and favors trajectories with small model errors. The second product contains the likelihood of every observation vector, and favors trajectories with small observation errors. Variational data analysis proceeds by locating the trajectory which maximizes the posterior density, usually by defining an objective function, or loss function,

$$\begin{aligned} J(\mathbf{x}_{0:K}) &= -\log p(\mathbf{x}_{0:K}|\mathbf{y}_{1:M}) + \log c \\ &= -\log p(\mathbf{x}_0) - \sum_{k=1}^{K} \log p(\mathbf{x}_k|\mathbf{x}_{k-1}) - \sum_{m=1}^{M} \log p(\mathbf{y}_m|\mathbf{x}_{k_m}), \end{aligned} \tag{0.23}$$

and then seeking the minimum of the objective. The maximum a posteriori estimate of the trajectory is then

$$\mathbf{x}_{0:K}^{\text{MAP}} = \underset{\mathbf{x}_{0:K}}{\arg\min}\, J(\mathbf{x}_{0:K}). \tag{0.24}$$

The objective function is similar in form to the log-posterior (0.11). One term leads to a preference for solutions that fit the data closely, while another term leads to a preference for solutions whose structure agrees with the modeling assumptions. To maximize the posterior density is to seek a good compromise between these two objectives.

Estimation of the full model trajectory in the presence of non-negligible model error is sometimes called the *weak-constraint* variational DA problem, to distinguish it from the *strong-constraint* problem. In the latter case, the dynamical model is assumed to be perfect in the sense of exactly matching the real process being simulated. A deterministic forward model operator $\mathcal{M}$ is then used, so that the entire trajectory $\mathbf{x}_{0:K}$ is fully determined by the initial state $\mathbf{x}_0$. This reduces the DA problem to locating the initial condition that produces the best fit with observations.

Uncertainty quantification is not an integral part of variational data assimilation, but the posterior variance can still be approximated by taking the inverse of the Hessian matrix of the log-posterior density $\nabla^2 \log p(\mathbf{x}_{0:K}|\mathbf{y}_{1:M}) = -\nabla^2 J(\mathbf{x}_{0:K})$ evaluated at the MAP estimate.

Monte Carlo methods

Monte Carlo methods aim to characterize the posterior distribution by generating samples from it. Like variational methods, this circumvents the need to compute normalization constants, such as c in (0.22), which are often intractable.

One of the simplest Monte Carlo methods is approximate Bayesian computation, or ABC (Tavaré et al., 1997; Beaumont et al., 2002). In each iteration of an ABC algorithm, a proposal realization is drawn from the prior distribution. Then the proposal is compared with all available observations, typically by predicting the measurements based on the realization and then computing the distance between the predicted measurements and the actual observations. If the match with data is sufficiently good according to a preset threshold, the proposal is accepted. Otherwise, it is rejected. This procedure is repeated until the desired number of accepted realizations has been generated.

ABC as described here is an example of a likelihood-free inference method, since the decision to accept or reject a proposed realization can be made without evaluating the likelihood, which is sometimes a desired property in an estimation method. In situations where the likelihood is easy to evaluate, one can use a more efficient rejection sampling algorithm with an acceptance probability proportional to the likelihood of the proposed realization.

To estimate the trajectory $\mathbf{X}_{0:K}$ in the HMM using likelihood-free ABC, we would draw a realization of the initial state $\mathbf{x}_0$ from a prior distribution. Then we would generate realizations of model error $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_K$, and use the forward model equation (0.16) to generate a realization of the trajectory. Predicted observations $\hat{\mathbf{y}}_{1:M} = \mathcal{H}(\mathbf{x}_{0:K})$ can be obtained using the observation operator as in (0.18) with no observation error. We then compute a summary statistic which measures the mismatch between the predicted observations and the actual observations, such as

$$S(\mathbf{x}_{0:K}, \mathbf{y}_{1:M}) = \sum_{m=1}^{M} \|\mathbf{y}_m - \hat{\mathbf{y}}_m\|^2 = \sum_{m=1}^{M} \left(\mathbf{y}_m - \mathcal{H}(\mathbf{x}_{k_m})\right)^T \left(\mathbf{y}_m - \mathcal{H}(\mathbf{x}_{k_m})\right). \tag{0.25}$$

If $S(\mathbf{x}_{0:K}, \mathbf{y}_{1:M}) < \epsilon$ the proposed trajectory $\mathbf{x}_{0:K}$ would be accepted. The efficiency of this sampling approach depends on the acceptance rate

$$A = P\left(S(\mathbf{X}_{0:K}, \mathbf{y}_{1:M}) < \epsilon\right), \tag{0.26}$$

which is the a priori probability that a proposal will be accepted. When iterations are carried out independently, one expects to perform N_s/A iterations in order to generate N_s accepted realizations. The acceptance rate itself depends on the width of the prior distribution of the initial state and model errors, as well as the tolerance parameter ϵ controlling the strictness of the acceptance criterion, and on the number of independent observations. If the number of observation vectors M and the number of elements in each observation vector N_y are large, the right hand side of (0.25) will tend to be large, making the acceptance rate small. Increasing ϵ to compensate is only viable up to a point, because making the acceptance criterion too loose will degrade the quality of the approximation. In that case, the accepted samples will not be a good representation of the true posterior distribution. Again, not having to evaluate the likelihood can be highly useful in certain situations, such as when the likelihood contains unknown model parameters that we are not interested in estimating. Still, due to the sensitivity to the number of observations—an instance of the curse of dimensionality—ABC is not suitable for data assimilation problems with high-dimensional data. Adopting a likelihood-based acceptance criterion can improve the overall acceptance rate to some extent, and so can importance sampling—drawing proposal realizations from a well-chosen importance function approximating the posterior distribution, combined with subsequent reweighting of accepted realizations. Neither modification will improve efficiency enough to make this kind of method suitable for high-dimensional data assimilation, however.

Markov chain Monte Carlo (MCMC) is another class of Monte Carlo methods. These do not generate independent realizations of the posterior distribution. Instead, they produce a Markov chain of correlated realizations. Because an accepted realization can be used as a starting point for generating the next one, this is more efficient than generating independent realizations. Nearly independent samples can still be obtained by thinning out the full Markov chain, since correlation lengths are usually not too long.

The Metropolis–Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970) is a particularly common MCMC method. In the following description, we will drop the subscript time indices from the model trajectory and observation variables $\mathbf{x}$ and $\mathbf{y}$. The MH algorithm, and the Markov chain it produces, are initialized by generating a realization from the prior distribution. Let $\mathbf{x}$ denote the current state of the Markov chain. A proposal $\mathbf{x}'$ for the next state is drawn from the proposal distribution $q(\mathbf{x}'|\mathbf{x})$. Next, the acceptance probability

$$\alpha(\mathbf{x}, \mathbf{x}') = \min\left(1, \frac{p(\mathbf{x}'|\mathbf{y})q(\mathbf{x}|\mathbf{x}')}{p(\mathbf{x}|\mathbf{y})q(\mathbf{x}'|\mathbf{x})}\right) \tag{0.27}$$

is computed. Here $p(\mathbf{x}|\mathbf{y})$ is the target posterior distribution (0.22), and since it appears in both the numerator and the posterior, the normalization constant cancels so that its value is not needed to compute the acceptance probability. The proposal is now accepted with probability $\alpha(\mathbf{x}, \mathbf{x}')$, in which case $\mathbf{x}'$ is used as the new state. If instead the proposal is rejected, $\mathbf{x}'$ is discarded and $\mathbf{x}$ is re-used as the new state.

The Metropolis–Hastings algorithm and other MCMC methods are less vulnerable to the curse of dimensionality than ABC, because acceptance of a realization is not conditional on a good match with data, i.e. a high likelihood, but rather depends on the likelihood ratio of the proposal relative to the current Markov chain state.

Variational inference, ABC and MCMC are well suited for solving smoothing problems where the target of inference is a trajectory of states. In the rest of this section, we will consider data assimilation methods which exploit the sequential structure of filtering problems, targeting the state at a single time. Whereas smoothing is usually carried out after a dataset has been collected, filtering is sometimes done *online*, meaning estimates are updated continuously, assimilating a stream of incoming observations. Online estimation places severe constraints on the computational cost of potential solutions.

Kalman filter

For HMMs that are Gauss-linear, i.e. that have linear forward model and observation operators and Gaussian model and observation errors along with a Gaussian prior distribution for the initial state, the sequential inference procedure comprised of equations (0.20) and (0.21) can be realized very efficiently. Under Gauss-linear conditions the filtering distribution, the conditional distribution of the current state given all observations available before or at the current time, will always be Gaussian. As such it is fully characterized by its mean vector and covariance matrix. Rather than constructing a new estimate of the filtering probability density at every time step, it is enough to update the estimated mean and covariance matrix of the filtering distribution. The Kalman filter (Kalman, 1960) does this efficiently using a set of forecast and update equations.

To keep the notation uncluttered, suppose there are observations at every time step $k = 1, 2, \ldots, K$. Extending the algorithm to the case with observations at only some times is straightforward. Write $\hat{\mathbf{x}}_{k|k-1}$ and $\mathbf{P}_{k|k-1}$ for the mean vector and covariance matrix of the one-step-ahead forecast distribution with density $p(\mathbf{x}_k|\mathbf{y}_{1:k-1})$. Write $\hat{\mathbf{x}}_{k|k}$ and $\mathbf{P}_{k|k}$ for the mean vector and covariance of the filtering distribution with density $p(\mathbf{x}_k|\mathbf{y}_{1:k})$. Denote the linear forward model and observation operators by $\mathbf{M}$ and $\mathbf{H}$, so that $\mathcal{M}(\mathbf{x}) = \mathbf{M}\mathbf{x}$ and $\mathcal{H}(\mathbf{x}) = \mathbf{H}\mathbf{x}$. Let the prior distribution of the initial state be Gaussian with mean $\hat{\mathbf{x}}_{0|0}$ and covariance matrix $\mathbf{P}_{0|0}$. Moreover, let the model error and observation error distributions be $N(\mathbf{0}, \mathbf{Q})$ and $N(\mathbf{0}, \mathbf{R})$ respectively. The forecast equations corresponding to (0.20) are then

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{M}\hat{\mathbf{x}}_{k-1|k-1} \tag{0.28}$$

for the forecast mean vector, and

$$\mathbf{P}_{k|k-1} = \mathbf{M}\mathbf{P}_{k-1|k-1}\mathbf{M}^T + \mathbf{Q} \tag{0.29}$$

for the forecast covariance matrix. Similarly, the update equations corresponding to (0.21) are

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k(\mathbf{y}_k - \mathbf{H}\hat{\mathbf{x}}_{k|k-1}) \tag{0.30}$$

for the updated mean vector and

$$\mathbf{P}_{k|k} = (\mathbf{I} - \mathbf{K}_k\mathbf{H})\mathbf{P}_{k|k-1} \tag{0.31}$$

for the updated covariance matrix. The gain matrix $\mathbf{K}_k$ appearing in (0.30) and (0.31) is the *optimal Kalman gain*, given by

$$\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{H}^T(\mathbf{R} + \mathbf{H}\mathbf{P}_{k|k-1}\mathbf{H}^T)^{-1}. \tag{0.32}$$

The update equation (0.30) is linear in the observation vector $\mathbf{y_k}$. If the model is a correct description of the dynamical model and data-generating process, then the Kalman filter is optimal in that no other linear filter gives a smaller mean square error $\mathrm{E}((\mathbf{X}_k - \hat{\mathbf{x}}_{k|k})^2)$. In the filtering equations (0.28) to (0.32), the matrices $\mathbf{M}$, $\mathbf{H}$, $\mathbf{Q}$ and $\mathbf{R}$ are identical for all time steps. This can be generalized so that some or all of these vary with k. If observations are not available at every time step, the update step is simply skipped when no observation is available, and the forecast estimates are used as updated estimates in the next time step.

Ensemble-based methods

Ensemble-based data assimilation methods use a collection of realizations, called an ensemble, to represent the forecast and filtering distributions. Each member of the ensemble is advanced individually from one observation time point to the next using the forward model. The ensemble is updated with respect to new observations either by moving the ensemble members to new positions in state space, or by shifting the weight of the ensemble onto the most representative realizations.

The ensemble Kalman filter, or EnKF (Evensen, 2009) is, as the name suggests, an ensemble-based variant of the Kalman filter. The advantage of the EnKF over the standard Kalman filter is the ability to handle mild non-linearity in the forward model and observation operators. The EnKF algorithm begins at time t_0 with an N_e-member initial ensemble $\{\mathbf{x}_0^{1,\mathrm{a}}, \mathbf{x}_0^{2,\mathrm{a}}, \ldots, \mathbf{x}_0^{N_e,\mathrm{a}}\}$. This could be a sample drawn from a prior distribution, but it could also be a set of equally probable realizations generated in some other way.

Not having to explicitly represent the prior distribution is an additional advantage of ensemble-based data assimilation methods. The superscript a in the ensemble members denotes *analysis*, while a superscript f is used to denote *forecast*.

The kth iteration proceeds by using the forward model operator to advance each ensemble member from t_{k-1} to time t_k as in (0.16),

$$\mathbf{x}_k^{i,\mathrm{f}} = \mathcal{M}(\mathbf{x}_{k-1}^{i,\mathrm{a}}, \mathbf{v}_k^i), \tag{0.33}$$

where $\mathbf{v}_k^i$ is drawn from the model error density $p(\mathbf{v}_k)$. The forecast ensemble is then $\{\mathbf{x}_k^{1,\mathrm{f}}, \mathbf{x}_k^{2,\mathrm{f}}, \ldots, \mathbf{x}_k^{N_e,\mathrm{f}}\}$. Next, a synthetic observation is created for each member of the forecast ensemble, using the observation operator as in (0.18),

$$\hat{\mathbf{y}}_k^i = \mathcal{H}(\mathbf{x}_k^{i,\mathrm{f}}, \mathbf{w}_k^i), \tag{0.34}$$

where $\mathbf{w}_k^i$ is drawn from the observation error density $p(\mathbf{w}_k)$. The combined forecast ensemble $\{(\mathbf{x}_k^{1,\mathrm{f}}, \hat{\mathbf{y}}_k^1), (\mathbf{x}_k^{2,\mathrm{f}}, \hat{\mathbf{y}}_k^2), \ldots, (\mathbf{x}_k^{N_e,\mathrm{f}}, \hat{\mathbf{y}}_k^{N_e})\}$ is now considered to be a sample from the joint conditional distribution of $\mathbf{X}_k$ and $\mathbf{Y}_k$ given $\mathbf{y}_{1:k-1}$. Based on the joint forecast ensemble, the cross-covariance matrix between $\mathbf{X}_k$ and $\mathbf{Y}_k$ and the covariance matrix of $\mathbf{Y}_k$ under this distribution are estimated,

$$\widehat{\mathbf{\Sigma}}_{\mathbf{X}_k,\mathbf{Y}_k} = \frac{1}{N_e - 1} \sum_{i=1}^{N_e} \left(\mathbf{x}_k^{i,\mathrm{f}} - \bar{\mathbf{x}}_k^{\mathrm{f}}\right) \left(\hat{\mathbf{y}}_k^i - \bar{\hat{\mathbf{y}}}_k\right)^T, \tag{0.35}$$

$$\widehat{\mathbf{\Sigma}}_{\mathbf{Y}_k} = \frac{1}{N_e - 1} \sum_{i=1}^{N_e} \left(\hat{\mathbf{y}}_k^i - \bar{\hat{\mathbf{y}}}_k\right) \left(\hat{\mathbf{y}}_k^i - \bar{\hat{\mathbf{y}}}_k\right)^T. \tag{0.36}$$

In the sample covariance expressions (0.35) and (0.36), $\bar{\mathbf{x}}_k^{\mathrm{f}}$ and $\bar{\hat{\mathbf{y}}}_k$ refer to averages taken over the ensemble members. Each forecast ensemble member is now updated according to a formula analogous to (0.30)

$$\mathbf{x}_k^{i,\mathrm{a}} = \mathbf{x}_k^{i,\mathrm{f}} + \widehat{\mathbf{K}}_k \left(\mathbf{y}_k - \hat{\mathbf{y}}_k^i\right), \tag{0.37}$$

where the estimated Kalman gain $\widehat{\mathbf{K}}_k$ is given by

$$\widehat{\mathbf{K}}_k = \widehat{\mathbf{\Sigma}}_{\mathbf{X}_k,\mathbf{Y}_k} \widehat{\mathbf{\Sigma}}_{\mathbf{Y}_k}^{-1}. \tag{0.38}$$

Unlike in the standard Kalman filter case, we are generally unable to compute the exact value of the optimal Kalman gain since the required covariance matrices are unknown, so we use the estimate (0.38) instead. With Gauss-linear modeling assumptions, $\widehat{\mathbf{K}}_k$ converges in probability to the optimal gain as $N_e \to \infty$, and the distribution of the analysis ensemble $\{\mathbf{x}_k^{1,\mathrm{a}}, \mathbf{x}_k^{2,\mathrm{a}}, \ldots, \mathbf{x}_k^{N_e,\mathrm{a}}\}$ converges to the correct posterior distribution. When Gauss-linear assumptions do not hold, there are no general convergence results. Still, the EnKF has been successfully applied to many data assimilation problems where either the forward model, the observation operator or both are weakly non-linear (Asch et al., 2016).

Particle filters (Gordon et al., 1993; Van Leeuwen et al., 2018), sometimes also called sequential Monte Carlo methods (Liu and Chen, 1998), are a class of Monte Carlo methods for filtering that represent the filtering distribution as a weighted ensemble. Like the EnKF, the particle filter starts with an initial ensemble, and alternates between forecast and update steps. The forecast in the kth iteration is more or less identical to the EnKF forecast (0.33), but the two methods differ in the update step. Instead of changing the states of the ensemble members, or particles, the particle filter updates their weights

$$w_k^i = \frac{w_{k-1}^i p(\mathbf{y}_k|\mathbf{x}_k^i)}{\sum_{j=1}^{N_e} w_{k-1}^j p(\mathbf{y}_k|\mathbf{x}_k^j)}. \tag{0.39}$$

The algorithm is initialized with equal weights $w_0^i = 1/N_e$ for $i = 1, \ldots, N_e$. After the update (0.39) has been applied to every weight, the resulting weighted ensemble $\{(\mathbf{x}_k^1, w_k^1), (\mathbf{x}_k^2, w_k^2), \ldots, (\mathbf{x}_k^{N_e}, w_k^{N_e})\}$ weakly represents the filtering distribution in the sense that for an arbitrary, suitably integrable function g, integrals of the form

$$I[g] = \int g(\mathbf{x}_k) p(\mathbf{x}_k|\mathbf{y}_{1:k}) d\mathbf{x}_k, \tag{0.40}$$

can be approximated, using the ensemble representation, by

$$\hat{I}_{N_e}[g] = \int g(\mathbf{x}_k) \sum_{i=1}^{N_e} w_k^i \delta(\mathbf{x}_k - \mathbf{x}_k^i) d\mathbf{x}_k = \sum_{i=1}^{N_e} w_k^i g(\mathbf{x}_k^i), \tag{0.41}$$

where $\delta(\mathbf{x} - \mathbf{x}_0)$ is the Dirac delta function centered at $\mathbf{x}_0$, with the approximation $\hat{I}_{N_e}[g]$ approaching the true integral $I[g]$ when $N_e \to \infty$.

Both the EnKF and the particle filter are affected by the curse of dimensionality. In the case of the EnKF this comes in the form of ensemble coupling between the ensemble members, introduced by sampling error in the estimated gain. This unwanted correlation between nominally independent ensemble members causes systematic underestimation of the variance of the state. Possible remedies include inflation of the estimated covariance, decoupling of the ensemble members after updating and, for spatial fields, localizing the updating scheme, so that observations are assimilated gradually by location. In the particle filter, the curse of dimensionality causes the distribution of weights $w_k^1, w_k^2, \ldots, w_k^{N_e}$ to degenerate and become progressively more concentrated with every iteration. The effect is exaggerated for high-dimensional observation vectors, but can be enough to render the naïve particle filter useless for $N_y \gtrsim 3$. When the weight distribution degenerates, the effective sample size

$$N_{\text{eff},k} = \frac{1}{\sum_{i=1}^{N_e} (w_k^i)^2} \tag{0.42}$$

decreases towards its minimum value of 1, as eventually the ensemble consists of a single particle whose weight is close to 1, and $N_e - 1$ particles whose weights are all close to 0. One common way to avoid the issue of degenerating weights is resampling, as in the sequential importance resampling (SIR) particle filter, due to Gordon et al. (1993). Resampling is done after computing updated weights, and before the next forecast step. A new ensemble is formed by drawing N_e particle indices from a multinomial distribution with the probability of choosing particle i equal to the weight w_k^i. Since the resampled ensemble is likely to contain multiple instances of some particles, small perturbation are added to every particle. The weights are then reset to $1/N_e$ before continuing to the next iteration.

Geological process modeling

This section reviews some of the ways computer models are used to study geological systems. In particular it presents the Geological Process Modeling simulation software, and its use in forward stratigraphic modeling of clastic sedimentation. The chapter closes by describing an important data assimilation problem arising in this setting, and sketching out how techniques from the previous chapter can be brought to bear on it.

Numerical models of geological systems

Computer models of subsurface geological structures are important in geology and hydrology as well as industrial applications like the mining and petroleum sectors. Although they answer the same kinds of questions, different modeling approaches differ widely in how they provide answers. The geological structures that exist today are the result of processes happening over long time periods. One important distinction is between models that try to capture the structure of the end result, and models that attempt to trace the process of formation.

Geostatistics began as the application of spatial statistics to estimation problems in mining engineering and related disciplines. Geostatistical models, once they have been fit to observations, describe the statistical properties of spatial or spatiotemporal phenomena. A typical application would be to assess the probability that a certain metal is present in high concentrations at a given location in the model region.

One kind of geostatistical model is Kriging (Krige, 1951; Cressie, 1990), also called Gaussian process regression, since it relies on the first two moments of the target distribution, and is a linear preciction or interpolation method. It models a spatial process $\{X(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$ as a Gaussian random field with mean $\mathrm{E}(X(\mathbf{s})) = \mu(\mathbf{s})$ and covariance

$$\mathrm{Cov}(X(\mathbf{s}), X(\mathbf{s}')) = \sigma(\mathbf{s})\sigma(\mathbf{s}')\rho(\mathbf{s}, \mathbf{s}') \tag{0.43}$$

where $\sigma(\mathbf{s})$ is the marginal standard deviation at location $\mathbf{s}$ and $\rho(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$ is a function specifying the correlation between the random field at locations $\mathbf{s}$ and $\mathbf{s}'$. Using suitable parmetrizations of $\mu(\mathbf{s})$, $\sigma(\mathbf{s})$ and $\rho(\mathbf{s}, \mathbf{s}')$, the Kriging model can be used for interpolation by computing the conditional expectation and variance of the field at unobserved locations, given observed values.

A Kriging model could also be fitted to realizations of a random field, such as a forecast ensemble produced by a forward model. In that case, the entire field is observed, and having several realizations improves estimation of marginal variance as well as pairwise covariances.

Another type of geostatistical model is multiple-point statistics, or MPS (Strebelle, 2002). This method takes as input one or more training images that are representative of the prediction target. The pixels of the training images represent grid cells, while the color or brightness of each pixel represents the value of the modeled process. MPS identifies patterns in the training images by looking for frequently occuring configurations of neighboring or nearby pixels. As the name suggests, it does not rely on two-point statistics, or pairwise covariances, like Kriging. Instead it analyses collections of more than two points simultaneosly. Found configurations, along with their observed frequencies, are then used to compute an empirical estimate of the conditional distribution of the value of a pixel given nearby pixel values that have already been determined. The simulation can be initialized by filling in pixels whose values are specified by observations, and conditionally simulating the other pixels one at a time. Sometimes it is assumed that the value of each pixel, or grid cell, belongs to a discrete set of classes. One then seeks estimates of posterior probability mass functions instead of densities.

If the covariance structure has the Markov property of localized dependence, then the spatial process can be represented as a Markov random field (MRF). An MRF defined on a regular grid or lattice can further be described as a Markov mesh model. This class of models enables parametric representations of the same kind of heterogeneous spatial structure that MPS is typically used to extract from training images.

Geostatistical models are geared towards representing observed present-day configurations of geological systems, and are comparatively easy to condition to measurements. Their main drawback is that they do not explicitly model the physics involved in the formation of the geological structures under study. To make use of such information, and to avoid physically implausible estimates, one must impose constraints through careful specification of the prior distribution, which can be difficult in practice.

Process models differ from geostatistical models in that they use representations of geological processes. These methods span a continuum from essentially geostatistical techniques with some physical constraints, via process mimicking simulation procedures involving objects that get stacked on top of each other, to purely physics based simulators (Pyrcz and Deutsch, 2014). Process models vary in terms of the scale of analysis, level of detail and intended use. For example, the gravity current simulations of Necker et al. (2002) are highly detailed, but are only applicable to a single phenomenon. Delft3D (Roelvink and Van Banning, 1995) is a less detailed but more comprehensive modeling system for studying sediment transport in fluvial systems and formation of deltas. Another example is BARSIM (Storms et al., 2002; Storms, 2003a) which is a two-dimensional simulator focussed on shallow marine systems dominated by wave action, and includes carbonate growth.

Generally, the more faithfully a modeling approach represents real processes, the harder it is to assimilate data into it. Simulation can create realizations of subsurface structures that are highly geologically and physically plausible, in and of themselves. They will not be useful for prediction purposes, however, unless relevant information from measurements is taken into account. Since most process models produce representations of intermediate states of the modeled system between the simulation start and end times, they should, in principle, fit into a HMM framework, as illustrated in Figure 3. We do not, however, have access to observations of the actual geological processes that have formed present-day geological structures. All available observations are of the current state, which does not quite fit the HMM causal structure. We will return to this issue when discussing application of data assimilation methods to geological process models at the end of this section.

Forward stratigraphic modeling

Stratigraphy is the part of the geology that concerns layered structures of rock. It deals with sedimentary processes, formation of sedimentary structures and the temporal and spatial variation in those structures. Sequence stratigraphy is specifically concerned with identifying sequences of sedimentary rock layers that share a common origin. Features such as subaerial unconformities and flooding surfaces are assumed to have been formed at one time. The structure of the rock units bounded above and below by such surfaces can be explained by variations in the conditions that the sedimentation process occurred under, including sea level, sediment supply and tectonic uplift or subsidence (Bryant, 1996).

Stratigraphic analyses can be based on data from seismic surveys, which show the locations of reflective layer boundaries, and well logs, which are observations of local physical properties taken by lowering a train of measurement tools attached to a wire into a borehole passing through the sedimentary rock. Measured properties can include porosity, levels of naturally occurring gamma radiation and the speed of sound waves passing through the rock. Based on this information, stratigraphic analysis can be used to predict the type and physical properties of the rock in an unobserved location (Catuneanu, 2002).

Stratigraphy in this sense has the structure of an inverse problem, as discussed in Section 4. Forward stratigraphic modeling (FSM) is the corresponding forward problem.

Given an initial state of the modeled region, and a specification of the external controls that affect the sedimentation process, i.e. sea level, sediment supply and tectonics, a simulation can be used to predict the system state at a later time. The state in this case includes both the spatial structure and the composition of the sedimentary layers. In other words, the forward stratigraphic model must keep track of where the layers are located, and what kinds of sediment they contain. Sediment composition can be characterized in terms of the distribution of grain sizes at each location.

Geological Process Modeling (Tetzlaff et al., 2014; Christ et al., 2016), hereafter referred to as GPM, is an example of an FSM simulator. GPM models the effect of erosion, transport and deposition of clastic sediment in areas spanning tens to hundreds of kilometers over time periods of millions of years. To enable simulation of long periods of geologic time, GPM avoids explicitly modeling repetitive phenomena, such as wave action, accounting instead for their aggregate effect over time. This abstraction allows the simulator to use time steps on the order of years. Several processes are simulated in parallel, the most important ones being sediment advection due to water flowing over the surface, and sediment diffusion. Modeling of the latter entails forward integration of a diffusion equation of the form

$$\frac{\partial z}{\partial t} = k_D \nabla^2 z \tag{0.44}$$

where z denotes the elevation of the current top surface of the deposited sediment, ∇^2 is the Laplacian operator and k_D is a constant controlling the strength of the diffusion process.

Inputs to GPM include the total duration of the simulated time interval, the temporal resolution of the simulation, the spatial dimensions and resolution of a two-dimensional grid covering the model region, the initial Bathymetry defined on the grid, a specification of the rate of tectonic uplift or subsidence at each grid point and the locations of sediment sources and sinks along with their intensities. For modeling surface flow of water, sources and sinks for the water, and the mixture of sediment types suspended in the water, must be specified as well. The direction and strength of ocean currents can also be specified. A sea level curve giving the changing sea level over the simulation time period is required. Sediment supply and tectonic uplift and subsidence rates can be constant over time, or can vary in both time and space. Together with the sea level curve, the overall rate at which sediment enters the model region at source locations and the change in this rate over time, is of special interest due to the important joint influence of these parameters on the architecture of the deposited layer package.

GPM outputs the state of the simulated system after each display time step, which are usually much longer than the internal simulation time steps. For example, a simulation covering 5 million years of geologic time might have 10, 50 or 100 display time steps, depending on the desired level of detail. The sediment deposited during one display time step gives rise to a single layer in the model representation of the complete layer package. All layers are bounded above and below by layer boundary surfaces defined on the two-dimensional model grid. A model run with K display time steps and a grid with $n_x \times n_y$ cells will therefore produce a three-dimensional grid of Kn_xn_y cells. Each cell is assumed to contain sediment of L discrete types. A proportion vector $\mathbf{p} = (p_1, p_2, \ldots, p_L)^T$ describes the sediment composition of a single cell. The model state after k display time steps can be represented by a set $S_{z,k}$ of layer boundary surface locations

$$\begin{aligned} S_{z,k} = \{z_{ijk} : i \in \{1, \ldots, n_x\}, j \in \{1, \ldots, n_y\}, \\ k \in \{0, \ldots, k\}\} \end{aligned} \tag{0.45}$$

and a set $S_{p,k}$ of sediment proportions

$$\begin{aligned} S_{p,k} = \{p_{ijkl} : i \in \{1, \ldots, n_x\}, j \in \{1, \ldots, n_y\}, \\ k \in \{1, 2, \ldots, k\}, l \in \{1, \ldots, L\}\} \,. \end{aligned} \tag{0.46}$$

A corresponding state vector $\mathbf{x}_k$ can be formed by arranging all the elements of S_k into a single column vector. The state dimension after k display time steps is

$$N_{x,k} = |S_{z,k} \cup S_{p,k}| = (k+1)n_xn_y + kLn_xn_y = [(L+1)k+1]n_xn_y. \tag{0.47}$$

Data assimilation for geological process modeling

Data assimilation for the GPM simulator fits into the HMM framework from Section 4 with the forward model operator $\mathcal{M}$ representing forward simulation from one display time step to the next. There are several different data assimilation problems to consider depending on which variables are targeted, and what kind of data are available.

Perhaps the most immediate estimation problem is to infer the present-day state of the system given all available information. Most of the modeled sedimentary layer structure will be buried and not directly observable. Available observations tend to have either high precision but limited coverage, such as borehole measurements, or to have good spatial coverage but poor precision, such as seismic data. An important advantage of process models is the ability to create geologically plausible realizations of the complete system. Observations, even if they are imperfect and incomplete, can constrain the process models so that realizations are not only plausible in and of themselves, but also probable in light of available information, and likely close to the actual, unobservable state of the real system. We assume that the observation operator $\mathcal{H}$ represents or approximates the data generating mechanism for whichever combination of observation types is used.

Another estimation problem is to infer one or several process-controlling parameters, such as the sea level and sediment supply curves. Parameters that are internal to the simulation, such as the diffusion strength k_D in (0.44), could also be targeted. In this sense, the model itself contributes to the uncertainty in the unobserved state of the system, along with natural sources of uncertainty like the unknown initial condition and environmental controls.

Suppose we are only interested in estimating some parameters $\boldsymbol{\theta}$, the initial state $\mathbf{x}_0$, or a combination of these. If the simulator is run through the entire simulation time interval to produce realizations of the state $\mathbf{x}$ at the final time point, and model equivalents $\hat{\mathbf{y}}$ of all observations $\mathbf{y}$ are generated from these realizations, then we can combine the forward model operator $\mathcal{M}$ and the observation operator $\mathcal{H}$ into a single modeling operator $\mathcal{G}$. We then have an inverse problem of the kind discussed in Section 4. In this situation, we assess the likelihood of the target variables $\boldsymbol{\theta}$ and $\mathbf{x}_0$ directly,

$$p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{x}_0) = \int p(\mathbf{x}|\mathbf{x}_0, \boldsymbol{\theta}) p(\mathbf{y}|\mathbf{x}) d\mathbf{x} \tag{0.48}$$

since we are not interested in the intermediate variable $\mathbf{x}$. In the strong-constraint case described in Section 4, the transition density $p(\mathbf{x}|\mathbf{x}_0, \boldsymbol{\theta})$ collapses to a Dirac delta function $\delta(\mathbf{x} - \mathcal{M}(\mathbf{x}_0, \boldsymbol{\theta}))$ because the dynamical model is deterministic. Equation (0.48) then simplifies to

$$p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{x}_0) = p\left(\mathbf{y}|\mathbf{x} = \mathcal{M}(\mathbf{x}_0, \boldsymbol{\theta})\right). \tag{0.49}$$

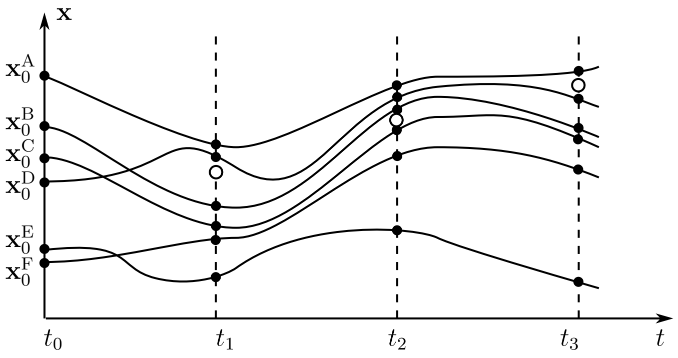

$\mathbf{x}$
$\mathbf{x}_0^{\mathrm{A}}$
$\mathbf{x}_0^{\mathrm{B}}$
$\mathbf{x}_0^{\mathrm{C}}$
$\mathbf{x}_0^{\mathrm{D}}$
$\mathbf{x}_0^{\mathrm{E}}$
$\mathbf{x}_0^{\mathrm{F}}$
t_0
t_1
t_2
t_3
t

FIGURE 4: Illustration of forward model runs. The initial ensemble members $\mathbf{x}_0^{\mathrm{A}}, \ldots, \mathbf{x}_0^{\mathrm{F}}$ are drawn from the prior distribution. Curves represent model trajectories. Model states are shown as filled circles and observations as empty circles. If the eventual poor fit of ensemble member E had been detected and adressed at t_1 or t_2, computational effort spent on integrating it to t_3 could have been put to better use.

In the weak-constraint case, the integral in (0.48) can be approximated using Monte Carlo methods where samples are generated by running the simulation several times. In practice, if the goal is to estimate the posterior distribution, it is not necessary to evaluate the likelihood at all combinations of $\mathbf{x}_0$ and $\boldsymbol{\theta}$. Instead, we could draw a sample $(\mathbf{x}_0^i, \boldsymbol{\theta}^i)$, $i = 1, \ldots, N_e$ from the prior distribution with density $p(\mathbf{x}_0, \boldsymbol{\theta})$ and evaluate the likelihood for each sampled realization. An approximation of the posterior distribution is then given by the resulting weighted ensemble $(\mathbf{x}^i, w^i)$, $i = 1, \ldots, N_e$ where $\mathbf{x}^i = \mathcal{M}(\mathbf{x}_0^i, \boldsymbol{\theta}^i)$ and $w^i \propto p(\mathbf{y}|\mathbf{x}^i)$.

Estimation of the model trajectory $\mathbf{x}_{0:K}$ or of the final state $\mathbf{x}_K$ only, could proceed along the same lines, since realizations of both of these are produced as a by-product of Monte Carlo based parameter estimation. However, if evaluation of the forward model is computationally intensive and time consuming, this approach will be not be preferred, as it is likely to cause considerable waste of computational effort. Figure 4 shows how continuing to run ensemble members that ends up fitting the data poorly is inefficient and diminishes the useful part of the ensemble. A drastic, but sometimes justifiable reaction to this is to forgo the uncertainty quantification capabilities of Monte Carlo methods, and move to optimization based solutions, like variational data assimilation methods. A more moderate compromise would be to use an ensemble based method which exploits the sequential nature of the modeled process. By halting the simulation mid-process and assessing how well each ensemble member fits the data, we can prevent

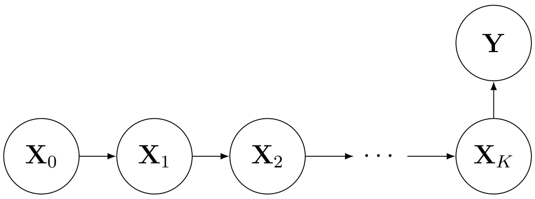


FIGURE 5: Graph representation of the causal relationship between states of the GPM simulator at different time points and the available observations.

the waste of computational effort either by adjusting the ensemble to make it more representative, as in the ensemble Kalman filter, or by resampling wayward ensemble members that match the data poorly, as in the particle filter.

An obstacle to applying sequential data assimilation to geological process models specifically is the structure of the causal relationship between the model trajectory and the available data. The sedimentation process was not observed while it unfolded, meaning that causal influence of earlier model states on actual observations is mediated by the final model state. That is, instead of the usual HMM graph of Figure 3, we have a structure more like the graph in Figure 5. To get around this one can use the available observations, perhaps in combination with other information, to construct pseudo-observations that can be used as if they were actual observations of the system at earlier time points.

Summary of papers

We end the introduction by briefly summarizing each of the three papers included in the thesis, discussing potential improvements and extensions of the work.

Paper I

Data assimilation for a geological process model using the ensemble Kalman filter

Co-author: Jo Eidsvik

Published in *Basin Research*, 30(4):730–745, 2018

The first paper describes an effort to apply sequential Bayesian inversion to stratigraphic forward modeling. GPM, introduced in the previous section, is a geological process-based simulation software, which produces three-dimensional models by simulating transport and deposition of sediments. Data derived from borehole measurements provide information about the sedimentation process, and can be used to constrain the simulation so that it outputs more realistic models. Considerable uncertainties in initial condition and environmental controls make it difficult to obtain a good match between simulation and observations.

Using Bayesian inversion to tackle the problem of constraining the simulation, we emphasise uncertainty quantification by putting prior distributions on unknown parameters. Moreover, imposing a state space model on the dynamical system exploits the sequential nature of the process-based simulation, where a layer package forms gradually over geological time. The variables which make up the system state describe the geometry and sediment type composition of the layer package.

Samples from the posterior distribution of the state variables are produced using a variant of the ensemble Kalman filter. Two test cases are considered. The first involves synthetic data and is used to assess the performance of the proposed estimation method, and to compare the filtering method with an ensemble-based smoothing algorithm that assimilates the data all at once rather than sequentially. The EnKF and the smoother are also compared with a multiple data assimilation (MDA) implementation of the smoother, whereby several updates are carried out iteratively on the same observations, using an inflated observation error covariance matrix when computing the gain matrix. The second and more realistic test case involves real well data from the Colville foreland basin, North Slope, Alaska. This is a difficult test case where it is necessary to make many assumptions about unknown parameters. To construct a reasonable observation operator for well log data, free runs of the simulation are used to learn the approximate relationship between deposition time and cumulative thickness under the prior distribution.

New contributions in this paper include the application of the EnKF to an inverse problem in forward stratigraphic modeling, the modification of the state variable representation to accommodate a spatially varying field of proportion vectors, and an observation operator using information from free runs of the dynamical model to determine which subset of the data to include in the conditioning.

Extending the model from using data from a single well to using data from multiple wells would be relatively straightforward, as one could essentially use the existing observation operator in parallel for each well. Generalizing to other data types such as seismic data, is also possible but requires more work. Since the lateral resolution, and the distance between neighboring traces, in a typical seismic survey are several orders of magnitude smaller than the model grid resolution in the real data test case, using the seismic data directly is likely not an acceptable solution.

It is possible to run the GPM simulator at different spatial and temporal resolutions, and to switch on or off various simulated processes. This makes it possible to trade off speed against fidelity, suggesting some sort of gradual model refinement approach where a large number of fast but rough model runs are used to update the prior distribution to a preliminary posterior distribution which then serves as the prior for the next iteration, where a smaller number of slower but more accurate simulation runs are used to further update the distribution, and so on.

Paper II

A revised implicit equal-weights particle filter

Co-authors: Peter Jan van Leeuwen, Javier Amezcua and Jo Eidsvik

Submitted to the Quarterly Journal of the Royal Meteorological Society, 2018

The second paper presents a revised version of the implicit equal-weights particle filter, a particle filter variant first introduced by Zhu, Van Leeuwen and Amezcua in 2016. As an ensemble-based filtering algorithm the IEWPF is an alternative to the EnKF, yet it avoids the assumption of a Gaussian filtering distribution. This makes it better suited for applications which involve highly non-linear, possibly chaotic dynamical systems. As model resolution increases and observation operators become more complex, the data assimilation problem becomes more non-linear. Hence there is a need for fully non-linear methods. The standard importance sampling particle filter, which involves a resampling step, becomes degenerate when used on high-dimensional systems. Recent developments point the way to new particle filters which are free of this limitation.

The IEWPF is an efficient particle filtering scheme which avoids filter degeneracy by forcing all particle weights to be equal by construction. This allows the filter to be used in very high-dimensional systems with a large number of independent observations. To achieve this, the method uses implicit sampling whereby perturbation vectors are drawn from a proposal distribution and transformed according to a certain mapping before being applied to each particle. The exact mapping used varies between particles and is determined by solving a non-linear scalar equation.

In the original formulation of the IEWPF, the proposal distribution has a gap causing all but one particle to have an inaccessible region in state space, leading to biased estimates. This paper describes a modification of the proposal distribution that eliminates the gap by adding an additional random perturbation. The variance of this extra perturbation is the same for every particle in the ensemble. We also discuss the properties of the new mapping, with a view to ensuring complete coverage of state space, and keeping in mind the aim of re-sampling as few particles as possible. The revised filtering algorithm is tested in synthetic experiments using a Gauss-linear dynamical system and the non-linear Lorenz96 model.

New contributions in this paper include an improved two-stage IEWPF proposal scheme which samples two perturbation vectors instead of one. This ensures that the proposal distribution has positive probability density at every point in state space. Furthermore, the scaling of the extra perturbation can be adjusted to fine-tune the ensemble variance, providing a means of reducing or eliminating bias.

The new tuning parameter controls the spread of the updated ensemble, so choosing the right value is necessary for the filter to be well calibrated. A procedure for identifying a good parameter value could be made part of the IEWPF algorithm. For example, one could generate multiple ensembles of synthetic observations, and compute the rank of the actual observation relative to each of these. The distribution of these ranks would depend on the value of the tuning parameter, and this procedure could be repeated with different parameter values until an approximately uniform rank distribution is achieved. To initialize the search, the optimal value from the previous assimilation time step could be used.

Paper III

Parametric spatial covariance models in the ensemble Kalman filter

Co-author: Jo Eidsvik

Submitted to Spatial Statistics, 2018

The third paper examines the potential for improving ensemble-based filtering algorithms by using a combination of simple parametric covariance models and maximum likelihood estimation instead of standard covariance estimators.

Ensemble-based data assimilation methods, like the ensemble Kalman filter, are widely used for prediction and parameter estimation in high-dimensional spatiotemporal applications. In methods like the EnKF, which require estimates of covariances between state variables to form the gain matrix, spurious correlations in these estimates can severely influence the updates applied to the ensemble members in the conditioning step of the algorithm. As a consequence, these methods are sensitive to the quality of observations and to the choice of the initial ensemble. To improve the robustness of estimation methods, we propose to replace the sample covariance estimator by a parametric estimate obtained by applying maximum likelihood estimation to a sparsely parametrized covariance model. We also consider semi-parametric estimation, where the variance is fit empirically from samples, while correlation parameters are determined by maximum likelihood estimation. Two covariance parametrizations for random fields are studied: An exponential covariance function and a Gaussian Markov random field (GMRF) specification with a parametrized precision matrix. For each of these, we demonstrate how parametric covariance estimation can work in the context of the ensemble Kalman filter and apply the methods to a test case involving the GPM simulator.

New contributions in this paper include an exploration of the advantages and disadvantages of parametric covariance estimation in a data assimilation setting, and an approach to incorporating available information on spatial non-stationarity induced by the dynamical model when estimating the precision matrix of a GMRF.

Estimation of non-stationary correlation structures with the GMRF specification relies on basis functions to represent spatial variation in precision matrix entries. These basis functions could be created by smoothing out sample variances, as in the GPM example. They could also be constructed using information about the kind of correlation structure the simulator tends to induce, perhaps based on free runs carried out in advance.

Many implementations of the EnKF and related methods never actually compute and store an estimate of the full covariance matrix of the state vector. Indeed, the update step only requires the cross-covariance matrix between the state vector and the observation vector, and the covariance matrix of the observation vector. If the observation vector has fewer elements than the state vector, it makes sense to avoid computing the full state covariance matrix. This applies to parametric covariance estimation too. We may want to directly estimate the actual covariances needed to perform the update step.

Paper I

Data assimilation for a geological process model using the ensemble Kalman filter

Jacob Skauvold and Jo Eidsvik

Published in *Basin Research*, 30(4):730–745, 2018

Data assimilation for a geological process model using the ensemble Kalman filter

Abstract

We consider the problem of conditioning a geological process-based computer simulation, which produces basin models by simulating transport and deposition of sediments, to data. Emphasising uncertainty quantification, we frame this as a Bayesian inverse problem, and propose to characterize the posterior probability distribution of the geological quantities of interest by using a variant of the ensemble Kalman filter, an estimation method which linearly and sequentially conditions realisations of the system state to data.

A test case involving synthetic data is used to assess the performance of the proposed estimation method, and to compare it with similar approaches. We further apply the method to a more realistic test case, involving real well data from the Colville foreland basin, North Slope, Alaska.

I.1 Introduction

Process-based geological models are important for exploring connections between geological variables in a theoretical setting. The potential predictive value of the process-based approach has begun to receive recognition, but effective prediction requires that the model can be conditioned to observations. Conditioning methods for process-based models are typically impractical relative to data conditioning in other modelling settings, such as more traditional geostatistical models. Hence, examples of successful predictive application of process-based models are rare (Pyrcz and Deutsch, 2014).

This paper considers the problem of data assimilation for a geological process computer simulation, referred to as the Geological Process Model (GPM), where we specifically use the simulator developed by Tetzlaff (2005), which produces basin models by simulating transport and deposition of sediments, and erosion of existing geological layers. By data assimilation we mean bringing together information from well or seismic data and from the geological model, in a consistent manner, such that the result correctly characterises our knowledge about the system state–the geological details of the area under study–as well as other relevant parameters describing the depositional environment.

In this paper, data assimilation for the GPM is carried out using the ensemble Kalman filter (EnKF). In this filter, realisations of the model state, referred to as ensemble members, represent a sample from the probability distribution of the geological state variables. When observations, such as measurements of part of the actual geological system are to be assimilated, the simulation is halted and each ensemble member is modified to better match these observations. Then the simulation is resumed on the basis of the updated ensemble. The end result of completing the simulation, and assimilating all data, is a final updated ensemble which represents the posterior probability distribution of the geological quantities of interest given all available data. This is the desired solution to the data assimilation problem (Evensen, 2009).

There has been recent interest in uncertainty quantification and data conditioning for complex geological models. Promising approaches include the ones due to Charvin et al. (2009a), who use an iterative Monte Carlo sampling scheme to condition a 2D simulation of a shallow-marine sedimentation process to observations of thickness and grain size, Bertoncello et al. (2013), who condition a surface-based model with iterative matching of sub-problems for a turbidite application, and Sacchi et al. (2015), who use a mismatch criterion for well log and seismic data from simulations. By assimilating data gradually, the approach taken in the current paper exploits the way that the simulated sedimentation process forms layers in sequence. In cases where it is applicable, it has the potential to be considerably more efficient than other methods.

The next section provides a more detailed description of the GPM simulator, its inputs and outputs, and how the model state is represented. The subsequent *Methodology* section gives an overview of the EnKF, and how it is implemented to work with the GPM (details are in the Appendix). In the *Numerical experiments* section, the EnKF/GPM combination is tested on two different data sets: One synthetic case created using the GPM, and one real case with well data from North Slope, Alaska. We close with a discussion section, reviewing the strengths and weaknesses of the proposed data assimilation scheme in light of the results from the two test cases, and pointing out possible directions for further development.

I.2 Geological Process Model

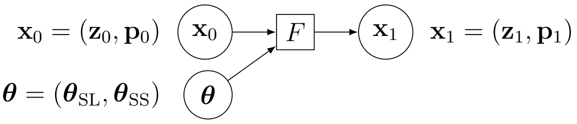
During the last three decades, the field of stratigraphic and sedimentological process modelling has seen much development, with simulation efforts including SEDSIM (Tetzlaff and Harbaugh, 1989), SEDFLUX (Syvitski and Hutton, 2001; Hutton and Syvitski, 2008), BARSIM (Storms, 2003b) and FLUMY (Lopez et al., 2009). See Paola (2000) or Tetzlaff and Priddy (2001) for details.

Process-based geological models differ from other geological and geostatistical models in that they seek to capture not only the nature of geology existing today, but also the processes which formed it. Process-based models, sometimes called process-response models, are powerful tools for establishing relationships between processes and results, especially when the processes in question cannot be simulated by a physical experiment in a laboratory. On the other hand, we require validation using field measurements or experimental observations in order to have confidence in process-based models, as well as any inferences drawn on the basis of their output.

One rather indirect way of using process-based models for prediction is to use the process simulation output as training data for some geostatistical prediction method, like multiple point statistics (Edwards et al., 2016; Strebelle, 2002). In doing this, one assumes that the process realisations used as training data are representative of the spatial structure of the geological features of interest so that, for instance, the variability in shape and size of features produced by simulation matches the variability observed in nature. One further assumes that the geostatistical method is able to extract the relevant structural information from the training data, and that this information generalises well enough to the geology of the prediction target. In other words, this “digital analogue” way of using process-based models to inform prediction requires essentially the same fundamental assumptions as do traditional geostatistical methods (Pyrcz and Deutsch, 2014). By

Table I.1: Input and output variables of the GPM simulator. The dimensions n_x and n_y define the size of the horizontal grid, and n_t is the number of discrete time steps used, starting at geological time t_0 and moving forward until time t_{n_t}. Some input variables have no symbol as they are not modelled explicitly in this paper.

	Variable	Symbol	Dimensions, Type
Input	Initial bathymetry	$\mathbf{z}_0$	$n_x \times n_y$ matrix
	Sea level curve	$\boldsymbol{\theta}_{\mathrm{SL}}$	$n_t \times 1$ vector
	Sediment supply rate	$\boldsymbol{\theta}_{\mathrm{SS}}$	$n_t \times 1$ vector
	Sediment source locations	-	$n_x \times n_y$ matrix
	Tectonic uplift/subsidence rate	-	$n_x \times n_y$ matrix
	Initial surface sediment proportions	-	4×1 vector
Output	Surface elevation (kth layer, $k = 1, \ldots, n_t$)	$\mathbf{z}_k$	$n_x \times n_y$ matrix
	Sediment proportion (kth layer, lth sediment type)	$\mathbf{p}_{k,l}$	$n_x \times n_y$ matrix


$\mathbf{x}_0 = (\mathbf{z}_0, \mathbf{p}_0)$
$\mathbf{x}_0$
F
$\mathbf{x}_1$
$\mathbf{x}_1 = (\mathbf{z}_1, \mathbf{p}_1)$
$\boldsymbol{\theta} = (\boldsymbol{\theta}_{\mathrm{SL}}, \boldsymbol{\theta}_{\mathrm{SS}})$
$\boldsymbol{\theta}$

constraining the simulation one avoids these assumptions, introducing in their stead the assumption that the process-based model is valid.

The GPM considered in this paper produces basin models by simulating transport and deposition of sediments, and erosion of existing geological layers (Tetzlaff, 2005; Christ et al., 2016). The same software is also capable of simulating other processes, such as carbonate growth, though that is not discussed in this paper. The basin is filled by sediments entering at a defined source location. In this case there is no sink in the model, and a gradual basin-filling process takes place, where the layer composition is defined by the sediment supply at the source and the sea level. The composition of particles in the sediment supply is kept fixed in our case, but since it can be modified in the software, it could be included in the statistical model as a set of uncertain parameters to be estimated from data. An overview of the simulator's input and output variables is given in Table I.1. The graph just below the table illustrates the relation between input and output. We next discuss each element in more detail. Figure I.6 shows an example of model output from the GPM conditioned to data (the context will be clarified in the examples).

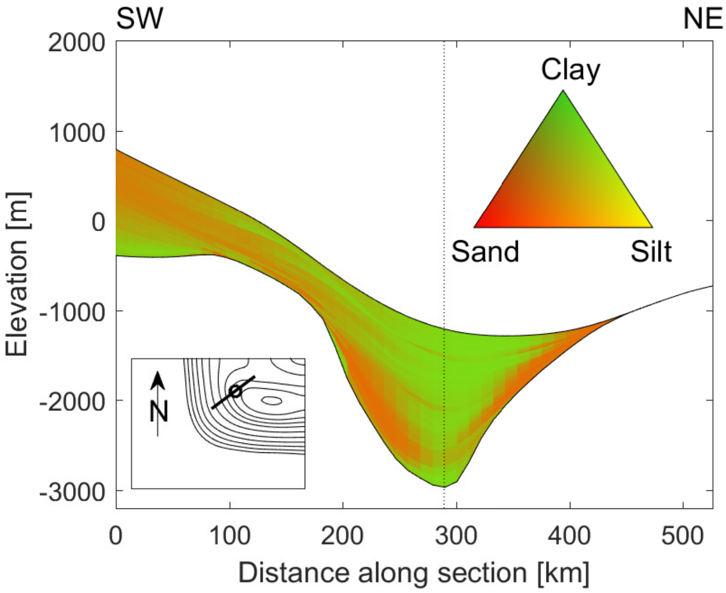


FIGURE I.6: Cross section of a simulated layer package with colours indicating proportions of sediment types (sand, silt, clay) in each position. The elevations and sediment proportions shown here are the ensemble mean of the final analysis ensemble in the North Slope, Alaska test case. The dotted vertical line near the centre indicates the location of the Tunalik 1 well. Inset map shows locations of section and well in modelled basin.

The forward model F represents the GPM simulation software. We treat F as a black box which accepts as input the system state $\mathbf{x}_k$ at time t_k, and the vector $\boldsymbol{\theta}$ of environmental parameters, and returns the system state $\mathbf{x}_{k'}$ at a later time $t_{k'} > t_k$. We are free to choose the time interval $t_{k'} - t_k$, but can only go forward in time. The forward model F is deterministic. Given the same input, it will always produce the same output. We thus make the important assumption that F is a correct representation of reality in the sense that there is no stochastic model error associated with it. The state uncertainty will be represented by an ensemble; multiple input realisations at time t_k are propagated through F to give multiple output representations at time $t_{k'}$.

The parameter vector $\boldsymbol{\theta}$ consists of the sea level curve $\boldsymbol{\theta}_{\mathrm{SL}}$, which describes the evolution of the sea level over simulated geological time, and the sediment supply curve $\boldsymbol{\theta}_{\mathrm{SS}}$ which specifies, as a function of time, the rate at which sediment enters the model area from outside (sediment source locations must also be specified, but the details of this will not be considered here). The sea level and sediment supply curves are represented as piecewise linear functions over time, with the vectors $\boldsymbol{\theta}_{\mathrm{SL}}$ and $\boldsymbol{\theta}_{\mathrm{SS}}$ containing function values at a shared set $\{t_i : i = 0, \ldots, n_t\}$ of discrete time points. Other quantities could have been included as parameters, such as the intensity of erosion as a function of water depth, or the rate of tectonic uplift and subsidence as a function of time and horizontal location. In the interest of a limited scope, however, we have chosen to focus on the sea level and sediment supply curves in this study. Other parameters which are required input for the simulator, are treated as known quantities, and kept fixed throughout.

The state vector $\mathbf{x}$ represents the physical configuration of the modelled system at a given moment in time. Together with the parameter vector $\boldsymbol{\theta}$, it contains all the information necessary to compute the system state at a later time. The rationale for treating $\mathbf{x}$ and $\boldsymbol{\theta}$ as separate entities is the asymmetry of the causal relationship between them; namely that the parameters influence how the state evolves over time, but the state does not influence the parameters.

There are two components of the state vector $\mathbf{x}$: The elevation component $\mathbf{z}$ and the sediment proportion component $\mathbf{p}$, which specifies how much sediment belongs to each of the categories coarse sand, fine sand, silt, and clay. Both are defined over a two-dimensional grid of discrete locations. (Additional details about the state vector are given in the Appendix.)

The elevation component $\mathbf{z}$ is a set of surfaces corresponding to the boundaries between layers of sediment deposited during successive time steps. The initial state $\mathbf{x}_0$ has only one elevation surface, $\mathbf{z}_0$, referred to as the initial bathymetry. After running the simulator F from time t_0 to time t_1, there will be two elevation surfaces, $\mathbf{z}_0$ and $\mathbf{z}_1$, corresponding to the bottom and top of the layer formed during the time interval (t_0, t_1). Due to erosion and tectonics, the new bottom surface $\mathbf{z}_0$ at time t_1 will generally be different from the original bottom surface $\mathbf{z}_0$ at time t_0.

The sediment proportion component $\mathbf{p}$ is a field of proportions characterising the type of sediment present in each location. To each three-dimensional grid cell c—defined horizontally by the two-dimensional model grid, and bounded vertically by two successive elevation surfaces in $\mathbf{z}$—is associated a vector $\mathbf{p}(c) = (p_1(c), p_2(c), p_3(c), p_4(c))$ specifying the proportion of each of four discrete sediment types (coarse and, fine sand, silt, and clay) present in the cell. In the following, the cell index c will be suppressed from the notation when the meaning is clear from the context.

I.3 Methodology

Uncertainty quantification

To meaningfully characterise quantitative geological variables of interest, it is necessary to assess the uncertainty associated with predictions and parameter estimates. For non-linear systems it is natural to quantify uncertainty with a sample or an ensemble of Monte Carlo realisations. As mentioned above, the GPM forward model F is assumed to be deterministic, and realisations of the geological system state are obtained by propagating sampled initial conditions forward in time under different versions of the sea-level and sediment supply parameters, which are also drawn from their prior distributions. The resulting output of GPM is an ensemble of state variables at relevant geological times. As this ensemble is purely model-driven and has not been conditioned to data, we refer to it as a prior ensemble.

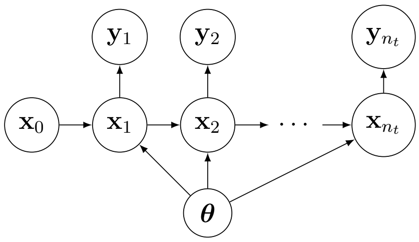
In its Bayesian flavour, the approach described in Charvin et al. (2009a) is quite similar to that of the current paper. But unlike their Markov chain Monte Carlo sampling procedure for assessing the posterior probability density function (pdf), the EnKF approach described here performs sequential updating of the state variables.

EnKF conditioning approach

The graph in Table I.2 illustrates the geological variables as a latent process. The variables are coupled in time according to the GPM forward model F, which is assumed to be Markovian in the sense that only the current state is relevant to the future evolution of the system, not the history leading to the current state. The top nodes in the graph illustrate data on which the process simulation is conditioned. Here, we assume that the data are well log observations of elevations and sediment proportions at discrete intervals in geological time, although other sources of information could also be used, such as attributes derived from seismic data. We assume that the observations are made in one well, at grid coordinates $(i^{\text{obs}}, j^{\text{obs}})$, for layers deposited at n_t different time points; $t_1 < \ldots < t_{n_t}$. The synthetic data used in the simulation study have the form $\mathbf{y}_k = (z_k(i^{\text{obs}}, j^{\text{obs}}), \mathbf{p}_k(i^{\text{obs}}, j^{\text{obs}}))$, while the data in the Alaska North Slope case consist of thickness and gamma ray observations linked to sediment proportions.

Table I.2: Generic variables and model components in the hidden Markov model (HMM) view of dynamical processes, and corresponding entities in the GPM setting. The graph illustrating the HMM dependence structure of the state and observation variables at discrete time points $t_0, t_1, \ldots, t_{n_t}$ as well as the parameter vector $\boldsymbol{\theta}$. Arrows between nodes indicate statistical dependence.

Generic	GPM-specific
System state $\mathbf{x}$	Layer elevation $\mathbf{z}$ Layer sediment composition $\mathbf{p}$
Initial state $\mathbf{x_0}$	Initial bathymetry $\mathbf{z}_0$ Base layer sediment composition $\mathbf{p}_0$
Parameters $\boldsymbol{\theta}$	Sea level $\boldsymbol{\theta}_{\text{SL}}$ Sediment supply $\boldsymbol{\theta}_{\text{SS}}$
Dynamic model F	Geological process simulator
Observations $\mathbf{y}$	Well logs
Observation model $\mathbf{h}$	Synthetic well logs



The well log data are modelled by a likelihood function. This means that a conditional pdf for the data is specified given the geological variables. Conditional on the geological elevation and sediment proportion at a time t_k, the measurement has an expectation defined by a functional relationship $\mathbf{h}(\mathbf{x}_k)$ and an additive, zero-mean Gaussian noise ϵ_y with covariance $\text{Cov}(\epsilon_y, \epsilon_y)$. The observation operator $\mathbf{h}$ could simply select values of the state variables (here elevation and sediment proportions) at the observation site, i.e. the location of the well in the model grid. In realistic settings however, $\mathbf{h}$ will typically have a more complex form, such as a local spatial average or a nonlinear function of one or more state variables. Such operators may involve parameters which require tuning to provide an adequate likelihood model for a specific application. The observation operator for gamma ray measurements used in the North Slope, Alaska case is an example of this.

In real applications the likelihood model will also require some form of matching between simulated depth at the time of deposition and measured depth at the time of observation.

The data $\mathbf{y}_1, \dots, \mathbf{y}_{n_t}$ shown in the graph in Table I.2 are assumed to be informative of the system state at the time of deposition. In the synthetic simulation study, this matching problem is avoided altogether by recording synthetic observations during, rather than after, the simulation. For the North Slope, Alaska case, a time-to-thickness relationship is established ahead of time as a part of model calibration. This involves sampling initial states and parameters from the prior distribution, and running the simulation based on these without assimilating data. The resulting unconditional model runs are used to construct a time-to-depth curve.

The goal of data assimilation is to characterise the posterior pdf of the system state, given all data by the current time step: $(\mathbf{y}_1, \dots, \mathbf{y}_k)$. In the EnKF, the solution is constructed sequentially; forecasting one step ahead using the GPM model F, and then conditioning on one more part $\mathbf{y}_k$ of the data, at every time step k. It is convenient for conditioning purposes to build an augmented state vector that includes geological layer variables for all previous geological times. The sea level and sediment supply parameters are also part of this augmented state vector. These parameters are distinct from the geological layer variables only in the sense that they are not changed by the GPM forward model. Hence, the geological state variables change in both the forecast and update steps, while the parameters change only in the update step.

The details of the EnKF implementation are provided in the Appendix, but the important elements are summarised here. To apply the EnKF to the GPM data assimilation problem, n_e samples from the initial state $\mathbf{x}_0$ and parameters $\boldsymbol{\theta}$ are generated from the prior pdf. Then the GPM runs from time t_0 until t_1 for all n_e ensemble members, giving an n_e-member forecast ensemble at time t_1. Using a generic notation where $\mathbf{v}_1$ denotes the state vector after one time step, the forecast ensemble at that time is

$$\mathbf{v}_1^{1,\mathrm{f}}, \mathbf{v}_1^{2,\mathrm{f}}, \dots, \mathbf{v}_1^{n_e,\mathrm{f}}.$$

Next, for each ensemble member b, pseudo-data are created by evaluating the expectation in the likelihood $\mathbf{h}_1(\mathbf{v}_1^{b,\mathrm{f}})$ and adding a random Gaussian perturbation $\boldsymbol{\epsilon}_{y,1}^b$ with the likelihood covariance. Thus, the pseudo-data are

$$\mathbf{y}_1^b = \mathbf{h}_1(\mathbf{v}_1^{b,\mathrm{f}}) + \boldsymbol{\epsilon}_{y,1}^b. \tag{I.50}$$

Finally, the Kalman filter update is applied to each ensemble member $b = 1, \ldots, n_e$,

$$\begin{aligned}
\mathbf{v}_1^{b,\mathrm{a}} &= \mathbf{v}_1^{b,\mathrm{f}} + \hat{\mathbf{K}}_1 \left(\mathbf{y}_1 - \mathbf{y}_1^b\right), \\
\hat{\mathbf{K}}_1 &= \widehat{\mathrm{Cov}}[\mathbf{v}_1^\mathrm{f}, \mathbf{h}_1(\mathbf{v}_1^\mathrm{f})] \left(\widehat{\mathrm{Cov}}[\mathbf{h}_1(\mathbf{v}_1^\mathrm{f}), \mathbf{h}_1(\mathbf{v}_1^\mathrm{f})] + \mathrm{Cov}(\boldsymbol{\epsilon}_y, \boldsymbol{\epsilon}_y)\right)^{-1}.
\end{aligned} \tag{I.51}$$

The covariances are estimated empirically from the forecast ensemble (see Appendix). Once all ensemble members have received their respective updates, an analysis ensemble

$$\mathbf{v}_1^{1,\mathrm{a}}, \mathbf{v}_1^{2,\mathrm{a}}, \ldots, \mathbf{v}_1^{n_e,\mathrm{a}}.$$

of size n_e is available after time step 1.

This forecast-update cycle is then repeated, using the newly formed analysis ensemble instead of the prior ensemble used initially, producing first a t_2-forecast ensemble, then a t_2-analysis ensemble, and so on. With each update, data from one observation vector is integrated into the ensemble. In probability density terms, the conditional pdf of $\mathbf{v}_2$ given $\mathbf{y}_1$ and $\mathbf{y}_2$ is $p(\mathbf{v}_2|\mathbf{y}_1, \mathbf{y}_2) \propto p(\mathbf{y}_2|\mathbf{v}_2)p(\mathbf{v}_2|\mathbf{y}_1)$, where the first term on the right hand side is the likelihood model of $\mathbf{y}_2$, which is conditionally independent of the other variables given $\mathbf{v}_2$ (see dependence structure in Table I.2). The second term is the forecast pdf which is represented by taking each ensemble member from the previous time step forward one step using the GPM. When all data have been assimilated into the analysis ensemble at the last time point t_{n_t}, the ensemble is representative of the posterior pdf of all geological variables, given all the data.

In the simulation study below, this EnKF approach is compared with two alternative methods. The first is often called the Ensemble Kalman Smoother (EnS), see e.g. Evensen (2009). It runs the ensemble members forward through all time steps before updating. The benefit of this approach in the geological process setting is that data are compared at the same geological time, which makes the likelihood model easier to specify. The downside is that it is very difficult to match present-day geology directly. In contrast, a filtering approach which integrates data sequentially is guided towards more realistic solutions as it steps forward through geological time.

The second alternative approach is EnS with multiple data assimilation (MDA), as described by Emerick and Reynolds (2013). The MDA approach relies on the following relation between pdfs: $p(\mathbf{y}_k|\mathbf{v}_k) = \prod_{r=1}^R p(\mathbf{y}_k|\mathbf{v}_k)^{1/R}$. Just like EnS, the ensemble members are now run all the way through the geological time interval, and updating is done at the end. But the MDA approach runs forward R times, with each update using an inflated likelihood covariance, R times larger than the actual observation covariance. A larger covariance means that the linear updates are smaller than the ones in the EnS. It can be difficult to tune R in practice, and if an application calls for a large number of iterations, the computational cost will be high.

I.4 Numerical experiments

Synthetic data

To demonstrate how the data conditioning works in practice, we apply it to an artificial test case. Our case is inspired by, but distinct from, the case considered by Charvin et al. (2009a). A reference realisation has been created by simulating sediment diffusion over 20 000 years. We use a grid consisting of $n_x = 72$ by $n_y = 16$ cells. Each cell has a horizontal size of 100 by 100 meters, so that the modelled region is a rectangle, 7.2 kilometers long in the cross-shore direction and 1.6 kilometers wide in the along-shore direction. The initial surface is roughly planar, with a downward slope of approximately $0.4°$ in the positive x-direction, but it also has smoothly varying deviations from this trend, drawn from a Gaussian random field with a correlation range of 10 grid cells, or 1 kilometer. Sediments enter the modelled area along the landward edge of the grid, at the top of the slope. New sediments appear here at a rate controlled by the sediment supply parameter $\boldsymbol{\theta}_{\mathrm{SS}}$. Over time they diffuse downhill, and are deposited at various distances down slope, depending on grain size. Figure I.7 shows an example of a realisation, at the final time of the simulation.

We assume that data are layer elevations and sediment proportions from a vertical well. The location of this conditioning well is shown in Fig. I.8. The goal is then to recover the layer package of the reference realisation by conditioning new GPM simulations to the data from this well. The initial bathymetry, sediment supply rate and sea level curve are all considered unknown, and must be estimated. The setting is an ideal case for the filtering method in the sense that the ensemble members are generated by the same process as the reference. Thus, model misspecification is essentially eliminated as a source of error. Any observed mismatch between the reference case and the filtering prediction will be due to limitations of the methodology, and not the simulation model itself. This is not the case when working with real data.

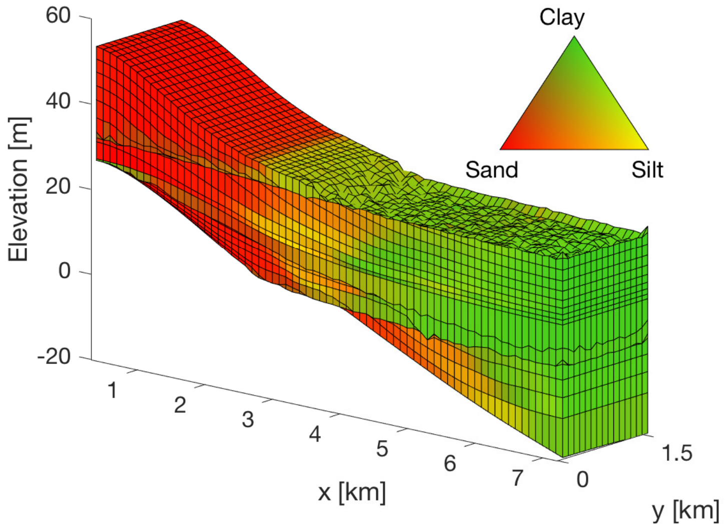


FIGURE I.7: An example of a reference realisation created with GPM.

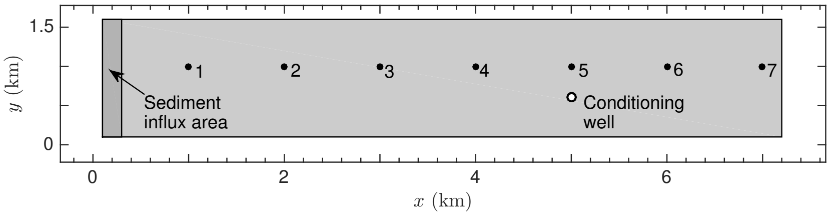


FIGURE I.8: Locations of the conditioning well and the 7 blind wells in the modelled area. The source area at the top of the slope, where new sediment is introduced, is also indicated.

We refer to the experiment of generating a GPM reference realisation, assigning well log data, and predicting the model state from this data, as one trial. To assess the performance of the filtering method, 100 independent trials were performed. Results of each trial, including both the reference realisation and the prediction, are stored for seven different "blind wells" placed at regular intervals down the length of the modelled area. The positions of the blind wells in relation to the conditioning well are indicated in Fig. I.8.

Table I.3: EnKF trial results for z and $\mathbf{p}$.

		Well number						
		1	2	3	4	5	6	7
z	MSE	18.96	13.55	14.37	9.62	3.62	6.80	15.36
	CRPS	1.48	1.13	1.01	0.80	0.53	0.74	0.99
	Cov.Pr.(80)	0.67	0.70	0.70	0.70	0.73	0.72	0.72
$\mathbf{p}$	MSE	49.98	51.74	37.58	30.42	12.45	23.96	16.59
	CRPS	1.29	1.34	1.08	0.98	0.52	0.85	0.79
	Cov.Pr.(80)	0.77	0.73	0.74	0.76	0.84	0.76	0.73

From the 100 trial results, we compute the following statistics to gauge the quality of the predictions obtained from the ensemble representation in the EnKF:

- Mean square error (MSE) which measures the average square difference between reference realisation and prediction. Smaller values of MSE mean better prediction.
- Continuously ranked probability score (CRPS) which measures the accuracy of the predictive distribution represented by the ensemble, relative to the reference blind well data (Gneiting and Raftery, 2007). Smaller CRPS is preferred since it means more precise predictions.
- Empirical coverage probability (Cov.Pr.) of 80% confidence intervals, which is the empirically observed proportion of trials producing confidence intervals which actually cover the corresponding value of the reference realisation. A probability near 80% means the ensemble members correctly quantify the uncertainty associated with the prediction. With an ensemble size of 100 it is convenient to form an 80 percent confidence interval by trimming 10 ensemble members from each tail of the distribution.

The results are given in Table I.3, where we averaged over all 20 layers. The smallest values of MSE and CRPS are the ones for blind well number 5, which is closest to the conditioning well. This holds both for the layer depths and for the proportions. It is more difficult to predict far from the conditioning well. The coverage probabilities tend to be a little below 80%, but no spatial trends are apparent. Nor is there any dramatic underestimation of the uncertainty.

Table I.4: Trial results for EnKF, EnS and MDA compared in terms of MSE, CRPS and coverage probability.

		EnKF	EnS	MDA
z	MSE	11.76	978.50	996.69
	CRPS	0.95	22.98	24.93
	Cov. Pr.	0.70	0.07	0.01
$\mathbf{p}$	MSE	31.82	177.08	134.04
	CRPS	0.98	3.88	2.76
	Cov. Pr.	0.76	0.31	0.48
θ_{SL}	MSE	125.78	528.62	335.72
	CRPS	4.56	12.43	12.00
	Cov. Pr.	0.58	0.37	0.23
θ_{SS}	MSE	13.72	181.19	153.15
	CRPS	1.42	8.57	9.66
	Cov. Pr.	0.66	0.23	0.08

Next, we compare the EnKF approach with EnS and MDA. Summary statistics over 100 trials are given in the Table I.4. In this case, both the EnS and MDA are clearly underestimating the uncertainty, which lessens the quality of the predictions significantly. The MDA used $R = 4$ iterations, which is not necessarily optimal, but it is not obvious how the number of iterations R should be tuned.

The statistics reported for $\mathbf{p}$ in Table I.3 and Table I.4 are not computed directly from the proportion vector itself, but a different variable $\mathbf{s}$ related to $\mathbf{p}$ via a logistic transformation. The reason for using a transformation is that while the elements of $\mathbf{p}$ must be valid probabilities, the elements of $\mathbf{s}$ can take any real value. See the Appendix for details.

Figure I.9 shows the posterior distributions for sea level and sediment supply as a function of geological time for a single trial. The EnKF results (left) are clearly better at covering the reference values. EnS (middle) and MDA (right) tend to give biased results. The sea level prediction obtained by EnKF only covers the larger geological time trends. Based on only one conditioning well, it appears difficult to capture the smaller fluctuations giving coarsening and fining upwards trends in Fig. I.7.

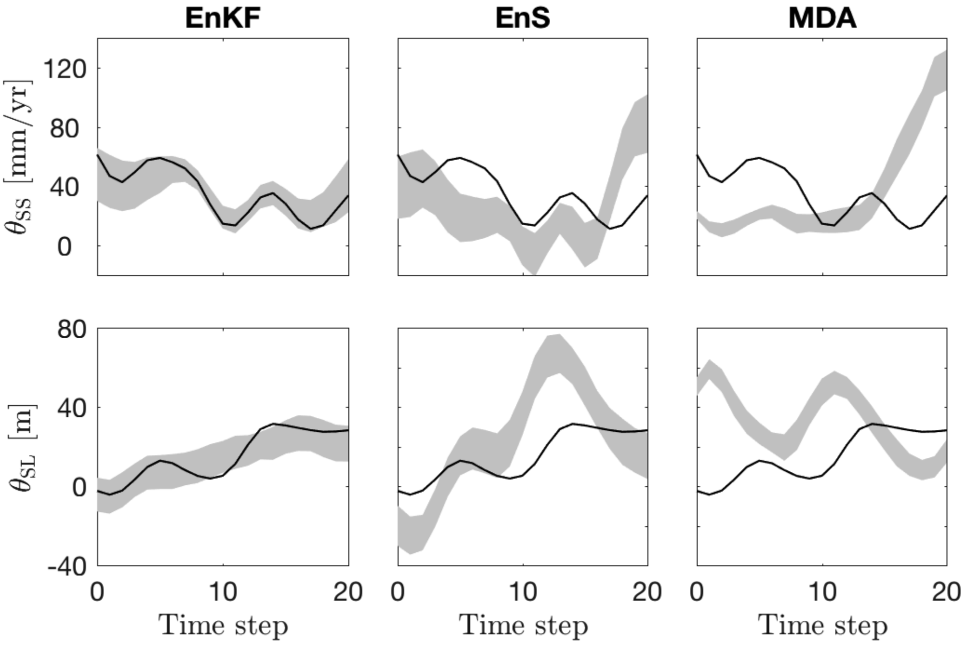


FIGURE I.9: Example trial results for the sediment supply (top) and sea level (bottom) parameters. The shaded areas indicate empirical 80% confidence intervals constructed from ensemble members obtained from the EnKF (left), the EnS (middle) and MDA (right). The true parameter values for the trial in question are shown as solid lines.

Real data case: North Slope, Alaska

The northern part of Alaska is an important oil and gas region, with much available data in the form of well logs and seismic surveys. In this section, we use GPM to model the Colville foreland basin. The area is indicated in Fig. I.10. This is the area studied by Schenk et al. (2012). Starting with an initial bathymetry corresponding to the top surface at 120 Ma, we use GPM to simulate deposition in the basin until 115 Ma. The simulation is conditioned on gamma ray well log data from the Tunalik 1 well, located at $70.20°$ N, $161.07°$ W, indicated by a circle on the map in Fig. I.10. The gamma ray (GR) data are measurements of natural gamma radiation taken at regular depth intervals along the trajectory of the borehole. It is measured in API (American Petroleum Institute) units, defined as a scaling of the observed radioactivity count rate by that recorded with the same logging tool in a reference depth zone (Killeen, 1982; Keys, 1996).

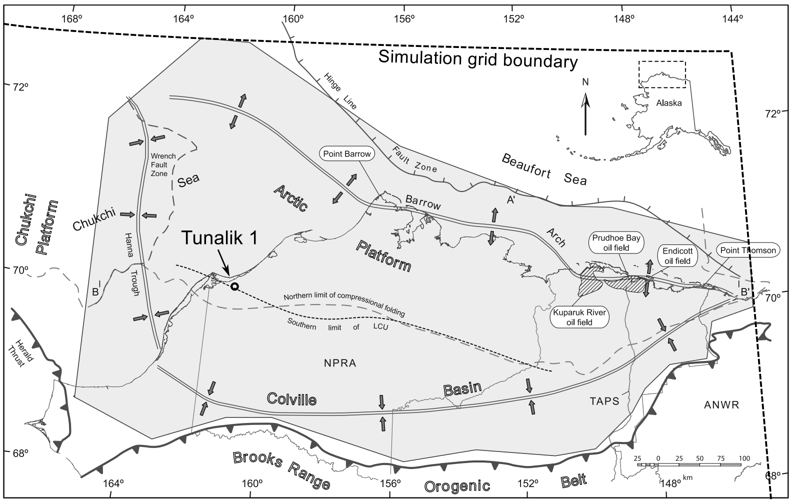


FIGURE I.10: Location of Tunalik 1 well relative to the Colville foreland basin study area (shaded region), and to the present-day coastline of northern Alaska. Parts of the northern and eastern boundaries of the simulation grid are shown as dashed lines. The southern and western boundaries are located outside the area covered by the figure. The map shown here is an adaptation of Fig. 1 in Schenk et al. (2012)

The lateral model grid covers a rectangular geographical area measuring approximately 1600 km in the east-west direction and 1300 km in the north-south direction, which is discretised into 110 × 87 grid cells, yielding a lateral resolution of ∼15 km in either direction. The Colville foreland basin is located in the northeast corner of the rectangle (see Fig. I.10). The southern and western parts of the rectangle are included in order to properly model sediment entering the actual basin region.

The Tunalik 1 well is approximately 6 km deep and, given the coarse horizontal spatial resolution of the model grid, can safely be assumed to be vertical. Based on the existing conceptual model of the study area by Christ et al. (2016), the part of the well log relevant to the modelled time interval is believed to be the ∼1900 m depth interval between 1300 m and 3200 m of depth relative to the present-day surface. The right panel of Fig. I.11 shows this part of the Tunalik 1 gamma ray log. The Tunalik 1 well data is available in LAS-format online (U.S. Geological Survey, 1981).

The time interval between 120 Ma and 115 Ma is discretised into 50 time steps of 100 000 years each. Consequently, the completed model output will consist of 50 distinct layers. We choose to update the model once every fifth time step. This means that the forecast

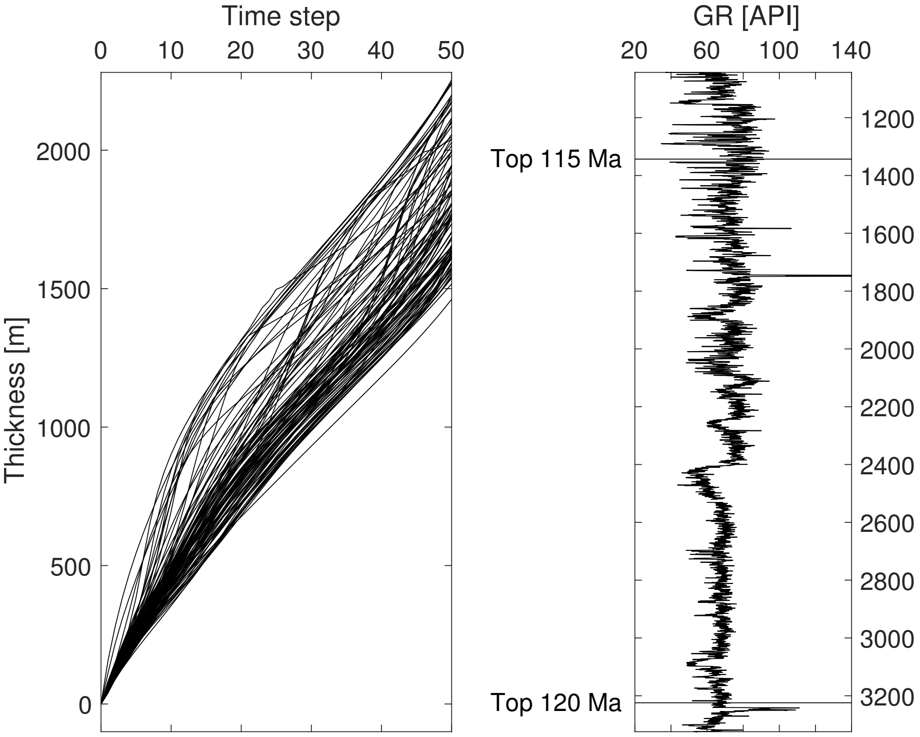


FIGURE I.11: Left: Prior realisations of cumulative thickness of deposited sediment at well location over simulated geological time. Right: The part of the Tunalik 1 gamma ray well log believed to be informative of the sedimentation happening between 120 Ma and 115 Ma.

ensemble after 5 time steps consists of only the first 5 layers, and is updated based on the deepest part of the gamma ray log shown in the right panel of Fig. I.11. Similarly, after 10 time steps, each member of the forecast ensemble contains 10 layers, and they are all updated based on the next segment of the well log, and so on, at times $15, 20, \dots, 50$.

Time steps of 100 000 years were chosen as a compromise between model resolution, both temporal and spatial, on the one hand, and computational efficiency on the other hand. Using shorter time steps would produce a larger number of thinner layers, which would allow us to resolve smaller details. At the same time, limiting the number of layers by using longer time steps reduces the amount of information that has to be passed to and from the simulator during data assimilation, making the procedure more efficient. Based on our experience with the simulator, we believe that time steps of 100 000 years give us the resolution necessary to capture relevant changes in grain size over time, such as the progradational Brookian sequences we are trying to model in this case (Christ et al., 2016).

The reason for assimilating data only once every five time steps, and observing blocks of five layers at a time, is that early experiments showed that updating on every time step tended to cause overfitting. That is, observations would be matched closely by the estimated system state, but the latter would have changed so much to accommodate the observations as to be unrealistic. At the other extreme, updating every tenth time step, and observing larger blocks of ten layers tended to produce very smooth estimates of the system state.

In order to carry out updates, it is necessary to identify which segment of the well log corresponds to the most recent five-layer block in the state forecast. This matching relies on a mapping between simulated time and thickness of the part of the well log relevant to the layers which have been simulated after that amount of time. In other words, a curve specifying cumulative present day thickness of the deposited layer package as a function of time. Rather than resorting to traditional back-stripping and decompaction methods, we obtain an estimate of this time-to-thickness map using results of unconditional simulations (i.e. model runs without data assimilation) carried out in advance, with initial state and environmental parameters drawn from the prior pdf. The left panel of Fig. I.11 shows 100 realisations of the time-to-thickness relationship at the location of the Tunalik 1 well. The curves shown in the figure give the thickness at 115 Ma. Dividing each curve by its final thickness and multiplying by the thickness of the relevant well log depth interval gives standardised thickness curves. The specific map used for conditioning is a single sequence of depths $\{\Delta z_0, \Delta z_1, \Delta z_2, \ldots, \Delta z_{50}\}$ chosen for being representative of the ensemble of curves shown.

After $k \in \{5, 10, 15, \ldots, 50\}$ time steps, we want to update the forecast state vector $\mathbf{x}_k^{\mathrm{f}}$ with respect to the observation vector $\mathbf{y}_k$, given by

$$\mathbf{y}_k = (\Delta z_k,\ \bar{\gamma}_k)^T,$$

where Δz_k is the standardised cumulative thickness after k time steps, and $\bar{\gamma}_k$ is a harmonic mean of gamma ray values,

$$\bar{\gamma}_k = \left(\frac{1}{n_{\gamma,k}} \sum_{z_i \in I_k} \frac{1}{\gamma(z_i)} \right)^{-1}.$$

The average is taken over the depth interval I_k, corresponding to the newest five-layer block, and $n_{\gamma,k}$ is the number of gamma ray measurements in the well log which belong to this interval. The depth interval of each block begins where the previous one ended, so that after 50 time steps, the ensemble will have been conditioned to all the well log data in the depth interval shown in the right panel of Fig. I.11.

The term $\mathbf{h}_k(\mathbf{x}_k^{\mathrm{f}})$ in the update equation (I.50) entering in equation (I.51) corresponds to the expected value of $\mathbf{y}_k$ given the state forecast $\mathbf{x}_k^{\mathrm{f}}$ To compute

$$\mathbf{h}_k(\mathbf{x}_k^f) = (\Delta z_k^{\mathrm{f}},\ \bar{\gamma}_k^{\mathrm{f}})^T$$

based on $\mathbf{x}_k^f$, we first extract the current thickness Δz_k^{f} of the simulated layer package at the well location by taking the difference in elevation between the top and bottom surfaces. The synthetic gamma ray value for grid cell i is given by

$$\gamma_i^{\mathrm{f}} = \sum_{\ell=1}^{4} p_{\ell,i}^{\mathrm{f}} \tilde{\gamma}_\ell.$$

where $p_{\ell,i}^{\mathrm{f}}$ is the forecasted proportion of sediment type $\ell \in \{1,2,3,4\}$ in grid cell i, and $\tilde{\gamma}_\ell$ is the expected gamma ray measurement for a grid cell containing only sediment of type ℓ. If, for instance, grid cell i' contains pure clay, then we will have $\gamma_{i'}^{\mathrm{f}} = \tilde{\gamma}_4$. The expected gamma ray values are parameters which must be chosen in advance to calibrate the model. Here, they were chosen so that the distribution of gamma ray values obtained by simulating from the prior distribution matches the marginal, depth-averaged distribution of gamma ray measurements in the relevant part of the Tunalik 1 well log.

Once the cell-wise gamma ray values have been synthesised, we average them over the most recent five-layer block. Let the 5 top grid cells in the well, after k time steps, be numbered $i_{1,k}, i_{2,k}, \ldots, i_{5,k}$. Then the synthetic gamma ray block average is

$$\bar{\gamma}_k^{\mathrm{f}} = \left(\frac{1}{5} \sum_{j=1}^{5} \frac{1}{\gamma_{i_{j,k}}^{\mathrm{f}}} \right)^{-1}.$$

The covariance of the measurement noise terms, denoted $\mathrm{Cov}(\boldsymbol{\epsilon}_y, \boldsymbol{\epsilon}_y)$ in equation (I.51), must be specified. Here, we used a very large standard deviation of $\sigma_{\Delta z} = 3$ km for the cumulative thickness observations, and a standard deviation of $\sigma_{\bar{\gamma}} = 3$ API for the local gamma ray averages. The thickness and gamma ray noise terms are assumed to be independent of each other. Choosing a very large standard deviation for the thickness means we are modelling the thickness as highly uncertain. As a result, observations of thickness will not contribute much to the shape of the likelihood function, and will have a limited influence on the posterior ensemble. As the effect of compaction between the end of the simulation time period and the time of observation is not explicitly accounted for in the model, it is reasonable that the uncertainty associated with observations of depth will be much larger than the uncertainty associated with gamma ray measurements. To illustrate the effect of $\sigma_{\Delta z}$ on estimates, we also run the EnKF with $\sigma_{\Delta z} = 30$ m.

When assessing the estimates of the system state and parameters in the North Slope case, we do not know the true state of the system, neither today nor at 115 Ma. Unlike for the simulation study in the previous section, there are no reference values of the estimated quantities to be used for judging the quality of the estimates. Still, we may get some insight by looking at the evolution of the system state estimate over time, and by comparing the synthetic data associated with the final system state estimate with the observations used for conditioning.

The top left panel of Fig. I.12 shows, for the case where $\sigma_{\Delta z} = 3$ km, the evolution over the simulated time interval of total thickness at the location of the Tunalik 1 well, represented by all 100 members of every forecast and analysis ensemble. Updates occur every 5 time steps, as can be seen by the vertical shifts. The bottom left panel shows the same thickness as represented by the final analysis ensemble. Values extracted from unconditional simulations are included for reference (the same prior thickness curves are shown in the left panel of Fig. I.11). The right panels show the same for the case where $\sigma_{\Delta z} = 30$ m.

Figure I.13 shows the match between estimated and observed well log measurements. The left panel shows layer-wise gamma ray values synthesised from prior simulation results, while the middle and right panels shows gamma ray values synthesised from posterior ensemble members with large and small observation uncertainty for cumulative thickness, respectively. All three panels also include the layer-wise ensemble mean and the observed block-wise average. The latter being identical in all panels.

Figure I.14 shows a comparison of the prior and posterior distributions of the sediment supply (top) and sea level (bottom) parameters in the form of point-wise, empirical 90% confidence intervals computed from ensemble members. Estimates for both large (left) and small (right) $\sigma_{\Delta z}$ are shown.

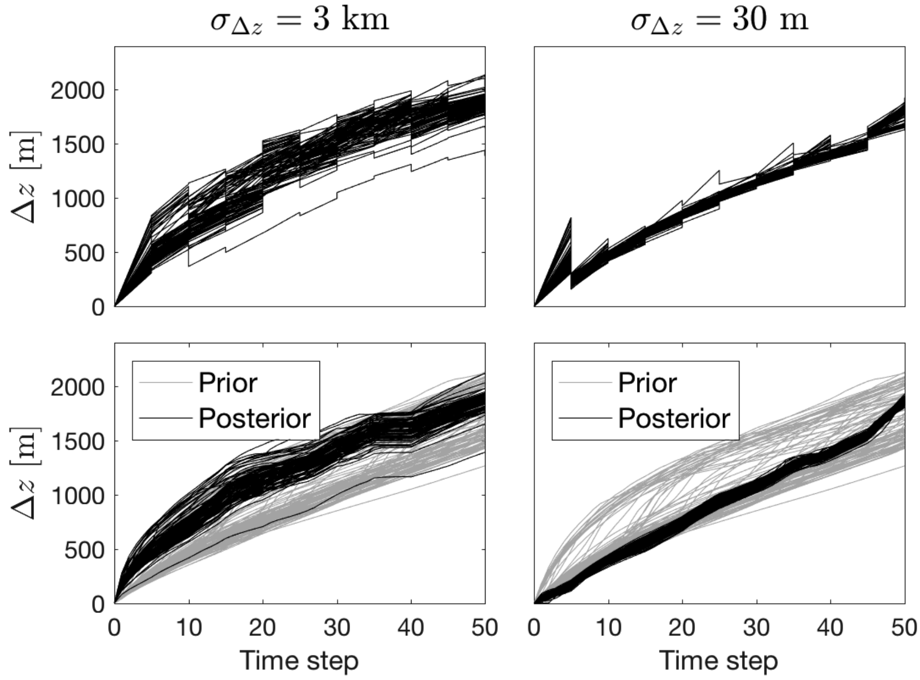


Figure I.12: Top: Evolution of estimate of total thickness of deposited sediment at Tunalik 1 well location. Bottom: Final analysis (posterior) ensemble compared with prior (unconditional) realisations. Results shown correspond to observing cumulative thickness with high (left) and relatively low (right) uncertainty.

I.5 Discussion

The conditioning problem

Loosely speaking, the data assimilation task considered in this paper consists of inferring causes from partially observed results. Since the measurable outcomes of the simulated geological processes are multiply realisable in terms of the various inputs, this inference problem lacks a unique solution. In this regard, it is a typical inverse problem.

Taking a Bayesian approach is natural for two reasons. First, introducing a prior pdf for the unknown quantities to be estimated provides necessary regularisation of the solution of the inverse problem. Second, Bayesian inference is a consistent and principled way of

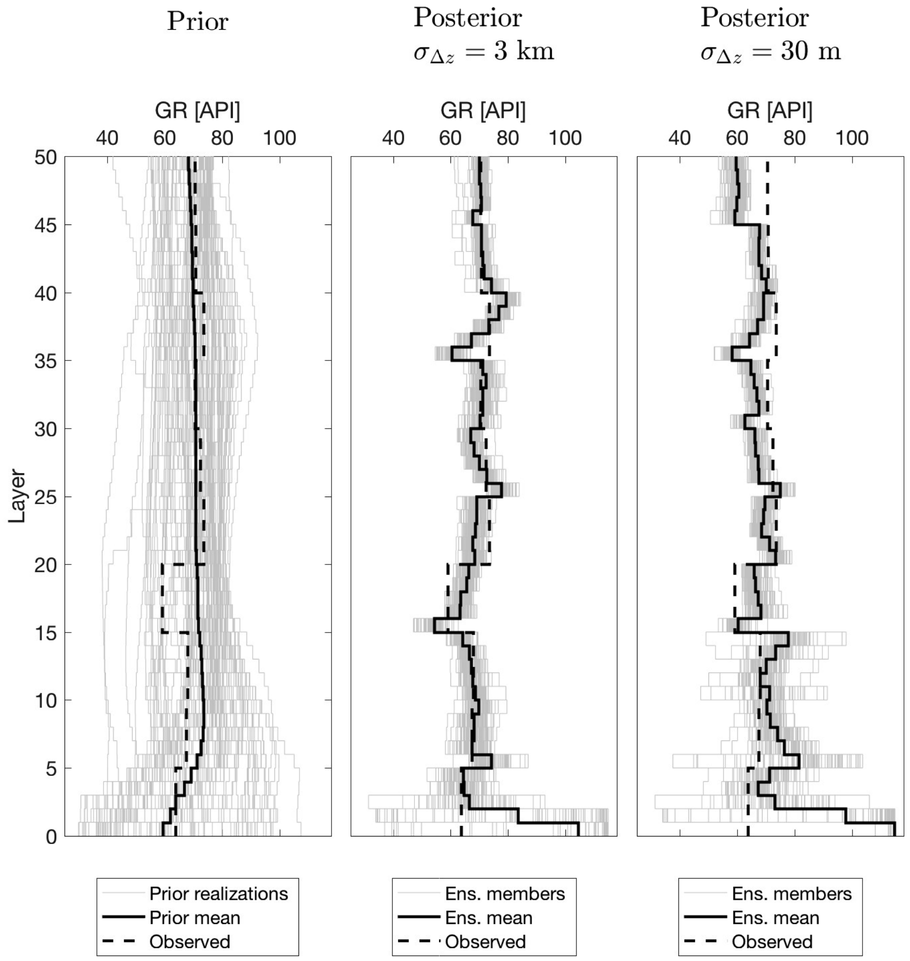


FIGURE I.13: Synthetic and observed gamma ray measurements in the Tunalik 1 well. Left: Realisations of the prior distribution, obtained by simulating without conditioning. Middle: Ensemble members after final update at time t_{50} when thickness observations are highly uncertain. Right: Posterior ensemble when thickness observations are informative. Observed gamma ray values are block-wise averages.

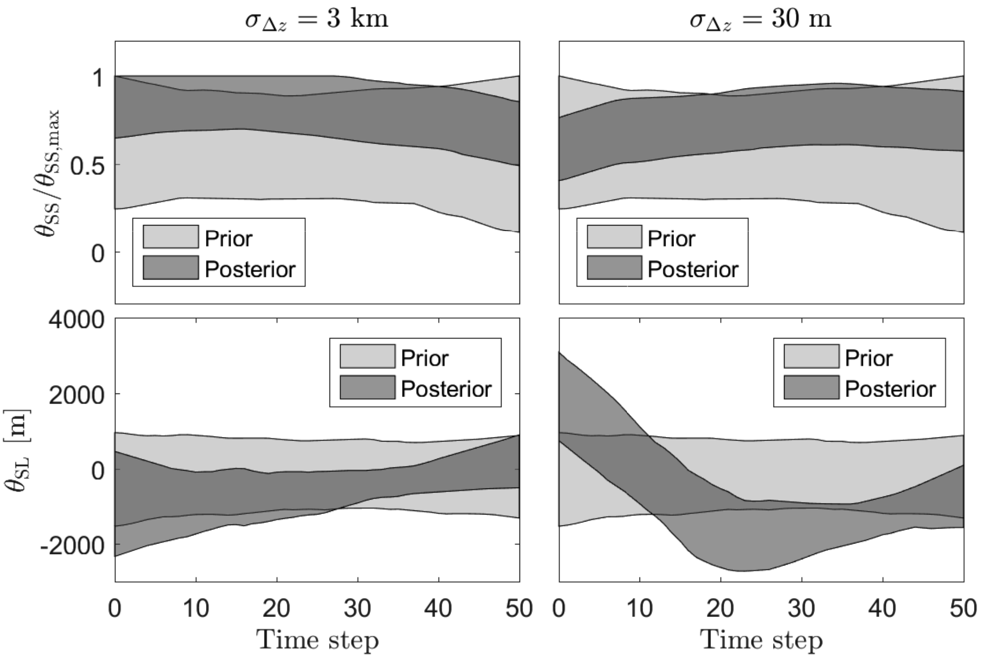


Figure I.14: Prior and posterior distributions of sediment supply (top) and sea level (bottom) parameters represented by empirical point-wise 90 percent credible intervals computed from 100 realisations of each distribution. As in Fig. I.12, estimates correspond to the two cases where cumulative thickness is observed with high uncertainty (left) and relatively low uncertainty (right).

combining quantitative observations of a physical system with relevant subject matter knowledge, while taking into account varying degrees of uncertainty associated with different sources of information (Tarantola, 2005; Evensen, 2009).

The EnKF is an appealing method for assimilating data to process-based numerical models like the GPM, as the sequential fashion in which the ensemble members are updated is well suited for exploiting the temporally ordered nature of the simulated sedimentation process. For situations beyond purely accumulative basin filling scenarios, however, the suitability of the EnKF is not assured. In scenarios characterised by more complex dynamical environments, for instance significant erosion events or faulting [see e.g. Chap 4.5 in (Pyrcz and Deutsch, 2014)], comparing a layer forming early in the simulated time period with present-day observations may be ill-advised, since the layer could be partially eliminated or moved in a later stage of the simulation. We outline possible extensions of the data assimilation method in the discussion at the end of this paper.

Synthetic test case

In the *Numerical experiments* section, to provide a context for assessing the performance of the EnKF on the synthetic test case, two additional estimation methods were tested: EnS and MDA. These methods update the ensemble only at the final geological time point.

With regard to the performance measures reported in Table I.3 and Table I.4, it is worth keeping in mind that both MSE and CRPS depend on the scale of the variable estimated. Hence, comparing values between methods for the same variable is always valid, while comparisons between estimated variables, whether within-method or between-method, are not necessarily meaningful.

The overall conclusion to be drawn from the results of the synthetic test case is that the EnKF performs reasonably well on this problem, which has the important property that the reference realisation, or ground truth, was generated using the same simulator which was used in the estimation. Furthermore, the observations used in the conditioning were generated from the reference realisation according to the specified likelihood model. This guarantees that the prediction target is realisable by the simulator, and that the likelihood model accurately represents the data generating process. This test case, therefore, represents an ideal case, and the filter's performance here should not be expected to generalise to cases without these properties. Nevertheless, comparing the EnKF with other estimation methods on an idealised, synthetic test case, is informative of relative performance between the methods in question, at least when applied to cases with a similar structure.

Compared with the EnKF, both the EnS and the MDA perform poorly on the synthetic test case, with larger MSE and CRPS for all variables. Empirical confidence interval coverage probabilities, while a little below the mark for EnKF, are surprisingly small for both of the other estimation methods, indicating that the large linear updates that they apply to the ensemble result in underestimation of posterior uncertainty.

Real data test case

In the second, more realistic test case, we model a piece of the Colville foreland basin in North Slope, Alaska, by conditioning a simulation of five million years of sedimentation on gamma ray well log data from the Tunalik 1 well, located within the basin. This is an example of the kind of basin filling scenario that we expect sequential data assimilation to be applicable to.

The depths in the Tunalik 1 well corresponding to the top and bottom of the simulated layer package were picked based on a conceptual model of the same region. Since our a priori confidence in this model is high relative to the level of uncertainty associated with the initial state and parameters, we treat the two depth markers as known constants. Nor do we attempt to explicitly model how the studied layer package changes between the end of the simulated time interval and the present day. The EnKF implementation used on the North Slope test case generally conditions on both locally averaged gamma ray measurements and observations of cumulative thickness. When the thickness observations are treated as very imprecise, by letting $\sigma_{\Delta z} = 3$ km, the system state and parameters are, in effect, being conditioned on gamma ray measurements only.

An alternative approach would be to explicitly model changes happening after the studied layer package was deposited, either by extending the simulated time period beyond the time interval of primary interest, or by using a less computationally expensive model to account for these changes. This could be a proxy model, based on a more coarse grained representation of the same processes as in the full model, or it could be a surrogate model, built by identifying regularities in the relationship between inputs and outputs of the full model [see e.g. (Frolov et al., 2009)]. In either case, an estimate of the present-day system state would be produced, and synthetic observations would be created by applying the likelihood model to this intermediate estimate.

The results of the real data test case are harder to interpret than the synthetic case results. Lacking a reference realisation to compare the estimates to, we resort to comparing the observations used in conditioning to predictions of the same observed quantities, synthesised from the estimated system state. In the North Slope, Alaska case, this means producing a synthetic gamma ray log from the estimated sediment proportions at the location of the Tunalik 1 well, and comparing this with the corresponding observed gamma ray measurements.

Although the data match for locally averaged gamma ray measurements in the Tunalik 1 well does leave something to be desired, it is clear, from comparing the panels of Fig. I.13, that both posterior ensembles fit the well log better than the prior ensemble does, with the $\sigma_{\Delta z} = 3$ km estimate achieving the closest match. For the sea level and sediment parameters, we see a marked tightening of the confidence intervals in Fig. I.14 going from the prior ensemble to the posterior ensembles, yet in both cases the posterior is still quite diffuse, suggesting that conditioning on gamma ray measurements from a single well yields only a modest reduction in uncertainty. The estimates of $\boldsymbol{\theta}_{\mathrm{SS}}$ and $\boldsymbol{\theta}_{\mathrm{SL}}$ obtained with $\sigma_{\Delta z} = 3$ km and $\sigma_{\Delta z} = 30$ m are broadly similar, the main difference being that the 30 m estimate (bottom right panel of Fig. I.14) detects a sea level decrease in the first half of the simulated time interval, which is less apparent in the 3 km estimate (bottom left).

Assumptions and limitations

When developing our problem-adapted version of the EnKF, we have assumed that the system dynamics are deterministic, so that identical inputs at one time will always produce identical outputs in the next time step. A consequence of this is that all the stochastic variation in a forecast ensemble is derived from variation in the updated ensemble one time step earlier. Adding a stochastic element to the time-evolution of the system state could be a way to account for possible model error, that is a possible discrepancy between the simulation and the actual physical processes being simulated.

As mentioned at the start of this section, we do not expect sequential data assimilation to be practical for geological scenarios deviating significantly from the kind of accumulative or additive behaviour which dominates the two test cases in this paper. Effectively conditioning simulations of more general geological processes likely requires a different approach.

Potential for further development

The modifications made to the standard EnKF in this paper concern only the observation likelihood and the representation of the system state and parameters. Other modifications, affecting how covariances are estimated, and how updates are applied to ensemble members, are possible. For example, updates could be localised in time, so that layers formed recently receive a more substantial update than older layers. In many applications, localisation can have a stabilising effect on the posterior ensemble (Nychka and Anderson, 2010; Sakov and Bertino, 2011). Another possible modification is to inflate the variance of the ensemble for a more accurate representation of uncertainty (Sætrom and Omre, 2013).

With respect to extending the ensemble-based simulation conditioning approach to make it more widely applicable, two directions of extension seem especially pertinent. First, one might wish to condition a simulation to several different kinds of data at the same time. In the North Slope case, for instance, we could imagine using not just gamma ray observations, but also observations of porosity or electric potential in the same well, or we could condition the simulation to well log data from several distinct wells. We may also want to combine information from well logs with data from seismic or other geophysical surveys. What is required in either case, is a likelihood model describing how the measurable quantities relate to the unobserved system state. Given the relatively coarse lateral resolution of the North Slope case, assuming conditional independence between observations at different sites might be reasonable. If so, the task of specifying the likelihood model for a seismic survey is effectively reduced to the problem of synthesising a seismic trace at a given grid location given the system state at only that location.

The other notable direction to expand the approach in is to try and get beyond accumulative basin filling. For this to work, data cannot be assimilated sequentially, as in the straightforward implementation of the EnKF. One alternative sampling approach is Markov chain Monte Carlo sampling, see e.g. Charvin et al. (2009a) or Laloy et al. (2017) for applications on similar problems. It is not always clear, however how to guide such samplers to give reasonable results for complex high dimensional problems. Another possibility is to move to the particle filter (PF) or similar conditioning methods. The PF has the distinct advantage over the EnKF that it never manipulates simulator outputs directly, instead performing conditioning by updating a set of weights associated with the ensemble of model realisations. On the other hand, the PF is typically less efficient than the EnKF at sampling the space of possible parameters and states, so that a relatively large number of realisations may be needed to obtain useful estimates and to prevent weights from collapsing (Chen, 2003). Whether the accompanying computational cost is prohibitive or not seems a worthwhile question to pursue.

Acknowledgments

We are grateful to Hilde G. Borgos, Bjørn Harald Fotland, Per Salomonsen, Lars Sonneland and Jan Tveiten at Schlumberger, Stavanger, Norway for access to the simulation software, technical support and helpful advice, to Oliver Schenk at Schlumberger, Aachen, Germany for much needed data and background information on the North Slope, Alaska case, and to Henning Omre at NTNU, Trondheim, Norway for useful discussions. We also thank the partners of the Uncertainty in reservoir evaluation (URE) project at NTNU.

Code Availability

Matlab scripts used to carry out computations, analyse results, and create some of the figures in this article are available online at github.com/Skauvold/DA-GPM (DOI: 10.5281/zenodo.1012346).

Conflict of Interest

No conflict of interest declared.

Appendix: Implementation details of the EnKF

In our situation, the EnKF is used to build a sequential approximation to the conditional probability density function (pdf) of the geological state variables, given information from the well log.

Let $\mathbf{v}_k$ be the state vector of variables at geological time t_k, $k \in \{1, \ldots, n_t\}$. This state is constructed from two distinct parts: a) the layer elevations $\mathbf{z}_k$ and the layer sediment compositions $\mathbf{p}_k$, b) the sediment supply $\boldsymbol{\theta}_{\mathrm{SS}}$ and sea level $\boldsymbol{\theta}_{\mathrm{SL}}$. Parts in a) are layer variables represented on a grid of lateral coordinates, while part b) variables do not vary with location. The elevations and sediment compositions in part a) are connected over

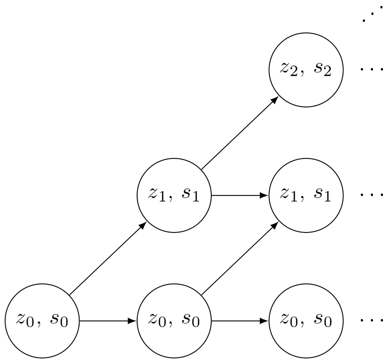


FIGURE I.15: Graph illustrating how the dimension of the state variable increases with each time step.

geological time by the GPM forecast model, so that with $k' < k$, we have

$$(\mathbf{z}_k, \mathbf{p}_k) = F(\mathbf{z}_{k'}, \mathbf{p}_{k'}, \boldsymbol{\theta}_{\mathrm{SS}}, \boldsymbol{\theta}_{\mathrm{SL}}). \tag{I.1}$$

The sediment proportions $\mathbf{p}_k$ have a one-to-one relation with another variable $\mathbf{s}_k$, in the form of a logistic transformation (Dobson and Barnett, 2008). The transformation is identical for all grid cells. Considering one time step and one grid cell only, and the four sediment types, we have $s_j = \log\left(\frac{p_j}{p_4}\right)$, $j = 1, 2, 3$, with inverse transformation $p_j = \frac{e^{s_j}}{1+\sum_{j=1}^{3} e^{s_j}}$ and $p_4 = \frac{1}{1+\sum_{j=1}^{3} e^{s_j}}$. Here, $p_j \geq 0$, $j = 1, 2, 3, 4$, and $\sum_{j=1}^{4} p_j = 1$, while $s_j \in (-\infty, \infty)$, $j = 1, 2, 3$, which makes it more robust to the linear updating in the EnKF. Layers are built up over geological time according to equation (I.1), and we include in the state vector all layers generated up to the current time (Fig. I.15). Part b) variables are represented as curves indexed by time, and the entire curve (for the whole simulated geological time interval) is included in every state vector. There is hence no change in part b) variables in the forecast step. Altogether, the state vector at time k is then

$$\mathbf{v}_k = (\mathbf{z}_0, \mathbf{s}_0, \ldots, \mathbf{z}_k, \mathbf{s}_k, \boldsymbol{\theta}_{\mathrm{SS}}, \boldsymbol{\theta}_{\mathrm{SL}}).$$

The EnKF is based on Monte Carlo sampling. At the initial time, n_e ensemble members are sampled independently from the prior pdf $p(\mathbf{z}_0, \mathbf{s}_0, \boldsymbol{\theta}_{\mathrm{SS}}, \boldsymbol{\theta}_{\mathrm{SL}})$. For later time steps k, the EnKF consists of two steps: i) the forecast step and ii) the analysis step (also referred to as the update step). In i) the state vector is propagated forward in geological time as described above. We denote the n_e members of the forecast ensemble by $\mathbf{v}_k^{1,\mathrm{f}}$,

..., $\mathbf{v}_k^{n_e,\mathrm{f}}$. The assimilation in ii) is done by building a regression model between the state variables and the data, and then using a linear update formula:

$$\mathbf{v}_k^{b,\mathrm{a}} = \mathbf{v}_k^{b,\mathrm{f}} + \hat{\mathbf{K}}_k(\mathbf{y}_k - \mathbf{y}_k^b), \quad \hat{\mathbf{K}}_k = \hat{\mathbf{\Sigma}}_{vy,k}\hat{\mathbf{\Sigma}}_{y,k}^{-1}.$$

Here, the $\mathbf{y}_k^b$ are pseudo-data obtained from $\mathbf{v}^{b,\mathrm{f}}$ and the likelihood model, while $\hat{\mathbf{\Sigma}}_{vy,k}$ and $\hat{\mathbf{\Sigma}}_{y,k}$ are the empirical cross-covariance and covariance matrices of the data:

$$\begin{aligned} \hat{\mathbf{\Sigma}}_{y,k} &= \frac{1}{n_e}\sum_{b=1}^{n_e}(\mathbf{y}_k^b - \bar{\mathbf{y}}_k)(\mathbf{y}_k^b - \bar{\mathbf{y}}_k)^T, & \bar{\mathbf{y}}_t &= \frac{1}{n_e}\sum_{b=1}^{n_e}\mathbf{y}^b, \\ \hat{\mathbf{\Sigma}}_{vy,k} &= \frac{1}{n_e}\sum_{b=1}^{n_e}(\mathbf{v}_k^{b,\mathrm{f}} - \bar{\mathbf{v}}_k^{\mathrm{f}})(\mathbf{y}_k^b - \bar{\mathbf{y}}_k)^T, & \bar{\mathbf{v}}_k^{\mathrm{f}} &= \frac{1}{n_e}\sum_{b=1}^{n_e}\mathbf{v}^{b,\mathrm{f}}. \end{aligned}$$

In our context, the likelihood model is $\mathbf{y}_k = \mathbf{h}_k(\mathbf{v}_k) + \boldsymbol{\epsilon}_{y,k}$, where $\boldsymbol{\epsilon}_{y,k}$ is a zero-mean Gaussian vector with covariance matrix $\mathrm{Cov}(\boldsymbol{\epsilon}_y, \boldsymbol{\epsilon}_y)$, which gives the form described in the main body of this paper [equation (I.51)], with

$$\begin{aligned} \widehat{\mathrm{Cov}}[\mathbf{v}_k^{\mathrm{f}}, \mathbf{h}_k(\mathbf{v}_k^{\mathrm{f}})] &= \frac{1}{n_e}\sum_{b=1}^{n_e}[\mathbf{v}_k^{b,\mathrm{f}} - \bar{\mathbf{v}}_k^{\mathrm{f}}][\mathbf{h}_k(\mathbf{v}_k^{b,\mathrm{f}}) - \bar{\mathbf{h}}_k(\mathbf{v}_k^{\mathrm{f}})]^T, \quad \bar{\mathbf{h}}_k(\mathbf{v}_k^{\mathrm{f}}) = \frac{1}{n_e}\sum_{b=1}^{n_e}\mathbf{h}_k(\mathbf{v}^{b,\mathrm{f}}), \\ \widehat{\mathrm{Cov}}[\mathbf{h}_k(\mathbf{v}_k^{\mathrm{f}}), \mathbf{h}_k(\mathbf{v}_k^{\mathrm{f}})] &= \frac{1}{n_e}\sum_{b=1}^{n_e}[\mathbf{h}_k(\mathbf{v}_k^{b,\mathrm{f}}) - \bar{\mathbf{h}}_k(\mathbf{v}_k^{\mathrm{f}})][\mathbf{h}_k(\mathbf{v}_k^{b,\mathrm{f}}) - \bar{\mathbf{h}}_k(\mathbf{v}_k^{\mathrm{f}})]^T. \end{aligned}$$

After the final analysis step the ensemble represents an approximation of the posterior pdf of the geological variables, given all data. For Gauss-linear dynamical systems, the EnKF is asymptotically correct, and the approximation becomes exact in the limit as $n_e \to \infty$. For other situations, there are no theoretical results regarding the quality of the approximation, but it has shown very useful in many practical applications.

Paper II

A revised implicit equal-weights particle filter

Jacob Skauvold, Peter Jan van Leeuwen, Javier Amezcua and Jo Eidsvik

Submitted to the *Quarterly Journal of the Royal Meteorological Society*, 2018

This article is awaiting publication not included in NTNU Open

Paper III

Parametric spatial covariance models in the ensemble Kalman filter

Jacob Skauvold and Jo Eidsvik

Submitted to *Spatial statistics*, 2018

Parametric spatial covariance models in the ensemble Kalman filter

Abstract

Several applications rely on data assimilation methods for complex spatio-temporal problems. The focus of this paper is on ensemble-based methods, where some approaches require estimation of covariances between state variables and observations in the assimilation step. Spurious correlations present a challenge in such cases as they can degrade the quality of the ensemble representation of probability distributions. In particular, prediction variability is often underestimated. We propose to replace the sample covariance estimate by a parametric approach using maximum likelihood estimation for a small number of parameters in a spatial covariance model. Parametric covariance and precision estimation is employed in the context of the ensemble Kalman filter, and applied to a Gauss-linear autoregressive model and a geological process model. We learn that parametric approaches reduce the underestimation in prediction variability. Furthermore rich, non-stationary models do not seem to add much over simpler models with fewer parameters.

III.1 Introduction

The ensemble Kalman filter (EnKF) is a popular Monte Carlo method for sequential data assimilation in complex systems (Evensen, 2009). At each step of this approach, Monte

Carlo samples, also called ensemble members, are first forecasted using the forward model and then updated with repect to data. The update step of the EnKF is based on covariances between forecast variables and data, the updated ensemble members being linear combinations of the forecast ensemble members with weights determined by estimated covariances. Empirical covariance matrices are typically used to specify the Kalman gain, i.e. the matrix of update weights. Although this empirical approach gives unbiased estimates of covariances, the formulation tends to produce inaccurate state estimates, especially when the number of state variables is much larger than the Monte Carlo sample size. The effect is undesired overfitting, and ensemble representations produced by the standard EnKF typically underestimate variability.

Localization and inflation of the covariance are common remedies for reducing the underestimation of variance in the EnKF (Furrer and Bengtsson, 2007; Asch et al., 2016). Hierarchical Bayes formulations have also been considered as a means of stabilizing the EnKF matrix expressions (Myrseth and Omre, 2010; Ueno and Nakamura, 2016; Tsyrulnikov and Rakitko, 2017; Stroud et al., 2018). In a similar vein, penalization of the inverse covariance matrix has been used in various ways, for instance by imposing a sparse neighborhood structure (Ueno and Tsuchiya, 2009) or by an ℓ_1 norm penalty to get a sparse graph structure (Hou et al., 2016).

Albeit promising in many applications of the EnKF, none of the mentioned approaches make explicit use of the spatial elements seen in many application domains (Cressie and Wikle, 2011; Katzfuss et al., 2016). In this paper we advocate stronger links between spatial statistics and EnKF approaches to improve the properties of the analysis ensemble. Our focus is to use Gaussian random field models and spatial covariance functions in the specification of covariances entering in the Kalman gain. Within this framework we apply maximum likelihood estimation to specify covariance parameters. This geostatistical approach means that only a small number of covariance parameters must be estimated on the basis of the ensemble, reducing the risk of overfitting and giving less underestimation of prediction variability. Ueno et al. (2010) used likelihood analysis within the EnKF for estimation of parameters in the measurement model. Similarly, Ueno and Nakamura (2016) and Stroud et al. (2018) used a Bayesian formulation for parameter estimation. Our approach is different in that we embed the forecast ensemble in a Gaussian process framework, and estimate the parameters of that approximation to the forecast distribution.

Many applications for which the EnKF has turned out to be useful are characterized by complex dynamical behavior giving rise to non-stationarity. Irregular data sampling design can also lead to non-stationarity because some regions are densely sampled while others are hardly informed by data at all. A parametric approach must accommodate these aspects in a realistic manner, and we explore how a trade off between model flexibility and complexity is sought.

In Section III.2 we describe the ideas underlying linear updating of an ensemble in a static situation. In Section III.3 we extend this to a dynamic state-space model, using parametric covariance or precision matrices in the EnKF update. In Section III.4 and III.5 we study the performance of the suggested approaches on a linear model and on an example from geology.

III.2 Approximate linear posterior sampling

Here we describe the underlying idea of posterior sampling with linear conditioning to data, considering a static situation. The time-dependent case is studied in Section III.3.

III.2.1 Notation and assumptions

Let $\boldsymbol{x} = (x_1, \ldots, x_n)'$ denote the uncertain variables of interest, and $p(\boldsymbol{x})$ the prior probability density function of $\boldsymbol{x}$. The size n of the target vector is in our case equal to the number of grid cells in a discretized spatial domain, typically in the order of 10^5 or higher. We assume, as is often the case for numerical simulations of physical systems, that it is comparatively easy to generate samples from $p(\boldsymbol{x})$, but that density evaluation is difficult or infeasible. The prior ensemble consists of B independent, equally likely realizations

$$\{\boldsymbol{x}^{1,f}, \ldots, \boldsymbol{x}^{B,f}\}. \tag{III.1}$$

The superscript f denotes *forecast* in this context. In applications involving computer-intensive numerical simulations, the ensemble size B is usually on the order of 10 to 100 because of limitations in computing resources (processing and memory).

The data are denoted by $\boldsymbol{y} = (y_1, \ldots, y_m)'$. In our context we use the common linear and Gaussian likelihood model; $p(\boldsymbol{y}|\boldsymbol{x}) = \text{Normal}(\boldsymbol{G}\boldsymbol{x}, \boldsymbol{T})$. The posterior distribution of the target vector $\boldsymbol{x}$, given the data $\boldsymbol{y}$, is defined by $p(\boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})$. We consider approaches that construct a posterior ensemble

$$\{\boldsymbol{x}^{1,a}, \dots, \boldsymbol{x}^{B,a}\}, \tag{III.2}$$

of equally weighted realizations, approximately representative of the posterior $p(\boldsymbol{x}|\boldsymbol{y})$. The superscript a denotes *analysis* or *assimilated.*

III.2.2 Simulation and linear updating

When the prior is represented by a forecast ensemble (III.1) it is possible, in principle, to use the likelihood $p(\boldsymbol{y}|\boldsymbol{x}^{b,f})$ to re-weight the prior samples and get a posterior representation. However, methods going in this direction, such as the particle filter (Doucet et al., 2000), tend to place all weight on one ensemble member in high-dimensional settings (Snyder et al., 2008). Hence, practical use of these approaches is limited. One can try to reduce the data dimension in various ways, for instance by conditioning only on some summary of the data as in approximate Bayesian computation (Beaumont, 2010), but methods of this type typically require that a large number of proposed realizations be generated, to the point of having a prohibitive computational cost in the kind of setting we are envisioning. Nor is it clear how to construct viable summary statistics or acceptance criteria for large spatial models.

We focus on approaches that use linear updating to get construct the analysis ensemble (III.2). This class of updating schemes correctly approximates the posterior distribution when the prior distribution and likelihood are both Gaussian, and when the ensemble size tends to infinity. While no performance guarantees can be given in the general case, for instance when assumptions of Gaussianity cannot be justified, this approach has shown itself to be very useful in several applications (Asch et al., 2016).

The linear update means that approximate posterior samples are generated by

$$\boldsymbol{x}^{b,a} = \boldsymbol{x}^{b,f} + \hat{\boldsymbol{K}}(\boldsymbol{y} - \boldsymbol{y}^{b}), \tag{III.3}$$

where $\hat{\boldsymbol{K}}$ is a weight matrix or gain that must be specified, and $\boldsymbol{y}^b$ is a synthetic observation or perturbed model equivalent given by

$$\boldsymbol{y}^{b} = \boldsymbol{G}\boldsymbol{x}^{b,f} + \boldsymbol{\epsilon}^{b}, \tag{III.4}$$

and the observation error realization $\boldsymbol{\epsilon}^b$ is drawn from a zero-mean multivariate normal with covariance matrix $\boldsymbol{T}$. Underlying the update in (III.3) is the joint covariance

$$\mathrm{Cov}\left(\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}\right) = \begin{bmatrix}\boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{x,y}\\ \boldsymbol{\Sigma}_{y,x} & \boldsymbol{\Sigma}_y\end{bmatrix}, \tag{III.5}$$

from which the gain matrix $\boldsymbol{K}$ is defined as

$$\boldsymbol{K} = \boldsymbol{\Sigma}_{x,y}\boldsymbol{\Sigma}_y^{-1}. \tag{III.6}$$

When the model is correctly specified, the gain matrix in (III.6) is the optimal linear regression weight for regressing the forecast state ensemble in (III.3) on the ensemble of synthetic observations in (III.4). In practice, the optimal gain is unknown, and an estimated gain matrix $\hat{\boldsymbol{K}}$ is obtained from samples $\boldsymbol{x}^{b,f}$ and $\boldsymbol{y}^b$, $b = 1, \ldots, B$. Since we assume that the likelihood model, including $\boldsymbol{G}$ and $\boldsymbol{T}$, is known, the ensemble is only used to estimate the prior covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_x$. The estimated gain then becomes

$$\hat{\boldsymbol{K}} = \hat{\boldsymbol{\Sigma}}_{x,y}\hat{\boldsymbol{\Sigma}}_y^{-1} = \hat{\boldsymbol{\Sigma}}\boldsymbol{G}'(\boldsymbol{G}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}' + \boldsymbol{T})^{-1}. \tag{III.7}$$

We will also use a formulation with the precision matrix $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$, which sometimes has a sparse (Markovian) structure. Moreover, for some models one can incorporate non-stationarity directly through the precision structure (Section III.3.3). An algebraically equivalent formulation of the assimilation step specified by (III.3), (III.4) and (III.6) is then

$$\left[\hat{\boldsymbol{Q}} + \boldsymbol{G}'\boldsymbol{T}^{-1}\boldsymbol{G}\right](\boldsymbol{x}^{b,a} - \boldsymbol{x}^{b,f}) = \boldsymbol{G}'\boldsymbol{T}^{-1}(\boldsymbol{y} - \boldsymbol{y}^b). \tag{III.8}$$

Computing the updated ensemble using this expression requires the solution of a system of linear equations with coefficient matrices that are sparse in most cases.

III.2.3 Empirical and parametric covariance specification

An empirical or non-parametric estimate of the covariance based on the prior ensemble is

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{B}\sum_{b=1}^{B}(\boldsymbol{x}^{b,f} - \bar{\boldsymbol{x}}^f)(\boldsymbol{x}^{b,f} - \bar{\boldsymbol{x}}^f)', \quad \bar{\boldsymbol{x}}^f = \frac{1}{B}\sum_{b=1}^{B}\boldsymbol{x}^{b,f}. \tag{III.9}$$

When the dimension of $\boldsymbol{x}$ is large, compared with the sample size B, the empirical estimates of the sample covariances are prone to large Monte Carlo errors (Furrer and Bengtsson, 2007; Sætrom and Omre, 2011).

In a parametric approach, the covariance is defined by a few model parameters $\boldsymbol{\theta}$, and we use $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ to denote the covariance matrix controlled by this parameter vector. The parameters must be chosen so that the resulting covariance matrix describes the simulation results well. Using a likelihood function

$$l(\boldsymbol{\theta}) = l(\boldsymbol{\theta}; \boldsymbol{x}^{1,f}, \ldots, \boldsymbol{x}^{B,f}) \tag{III.10}$$

for this purpose, the parameter estimate is

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}). \tag{III.11}$$

We assume that the likelihood is representative of a Gaussian process, where the mean is computed directly from the ensemble. Moreover, the B ensemble members are assumed to be independent and identically distributed, so that the likelihood is given by

$$l(\boldsymbol{\theta}) = -\frac{B}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}| - \frac{1}{2}\sum_{b=1}^{B}(\boldsymbol{x}^{b,f} - \bar{\boldsymbol{x}}^{f})'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{x}^{b,f} - \bar{\boldsymbol{x}}^{f}). \tag{III.12}$$

The parametric estimate of the covariance matrix $\boldsymbol{\Sigma}$ is then $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}$.

For common parametrizations of spatial dependence in $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, there are closed form expressions for the derivatives of the likelihood (III.12), see e.g. Petersen et al. (2008). These are calculated at every iteration of the optimization procedure. Parameter estimates typically converge after no more than 5 to 10 Fisher-scoring iterations,

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} + \left[E\left(\frac{d^2 l(\hat{\boldsymbol{\theta}})}{d\boldsymbol{\theta}^2}\right)\right]^{-1}\frac{dl(\hat{\boldsymbol{\theta}})}{d\boldsymbol{\theta}}. \tag{III.13}$$

If derivatives are not available, other optimization schemes must be used, such as Nelder-Mead search (Lagarias et al., 1998).

Table III.5: Performance of different covariance specifications. Matrix norms are means of 100 replicates, and parentheses represent the standard deviation of these.

$B = 100$	Kullback–Leibler	Bhattacharyya	Frobenius
emp	96 000 (741)	558 (0.212)	63.3 (2.6)
emp, loc	113 (1.6)	26.2 (0.3)	88.1 (0.2)
par	0.01 (0.01)	0.002 (0.002)	4.54 (3.2)
semi-par	11.5 (0.9)	2.78 (0.2)	31.1 (1.4)
$B = 1000$	Kullback–Leibler	Bhattacharyya	Frobenius
emp	370 (2.7)	40.3 (0.2)	20.2 (0.8)
emp, loc	52.1 (0.07)	16 (0.07)	87 (0.02)
par	0.001 (0.001)	0.0004 (0.0003)	1.8 (1.3)
semi-par	5.6 (0.5)	1.36 (0.1)	13.6 (1.6)

III.2.4 Illustrative spatial example

We compare the results of empirical and parametric covariance estimates for different sample sizes ($B = 100$ and $B = 1000$). The spatial variable $\boldsymbol{x}$ is here represented on a regular 25×25 grid, and entry x_i represents the variable in grid cell $i \in \{1, \ldots, 625\}$. The prior mean values are 0, and the covariance model is stationary with variance $\sigma^2 = 1$ in all grid cells and an exponential correlation function. Defining $\boldsymbol{D}$ to be the 625×625 matrix of distances between all grid cells, the true covariance matrix is $\boldsymbol{\Sigma} = \sigma^2 \exp(-3\boldsymbol{D}/\eta)$, where $\eta = 10$ indicates an effective correlation range of 10 grid cells.

We study covariance estimation performance using three criteria: Kullback-Leibler divergence, Bhattacharyya distance and the Frobenius norm. For all these measures we compare the specified covariance $\hat{\boldsymbol{\Sigma}}$ with the true covariance matrix $\boldsymbol{\Sigma}$. For zero-mean multivariate Gaussian vectors, the Kullback-Leibler divergence D_{KL}, Bhattacharyya distance D_{B} and Frobenius norm distance D_{F} between $\hat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}$ are

$$D_{\mathrm{KL}}(\hat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}) = \frac{1}{2}[\mathrm{trace}(\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}) - n + \log|\hat{\boldsymbol{\Sigma}}| - \log|\boldsymbol{\Sigma}|], \tag{III.14}$$

$$D_{\mathrm{B}}(\hat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}) = \frac{1}{2}\log|\tilde{\boldsymbol{\Sigma}}| - \frac{1}{4}|\log|\hat{\boldsymbol{\Sigma}}| - \frac{1}{4}\log|\boldsymbol{\Sigma}|, \quad \tilde{\boldsymbol{\Sigma}} = [\boldsymbol{\Sigma} + \hat{\boldsymbol{\Sigma}}]/2, \tag{III.15}$$

$$D_{\mathrm{F}}(\hat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}) = \sqrt{\mathrm{trace}[(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})']}. \tag{III.16}$$

Results of the covariance estimation are presented in Table III.5. The empirical estimate

Table III.6: Performance of linear updating using different covariance specifications. The results are means of 500 replicates, and the parentheses represent the standard deviation of these.

$B = 100$	MSPE	CovPr(80)	CRPS
emp	0.372 (0.035)	32.3 (2.1)	0.241 (0.007)
emp, loc	0.263 (0.009)	76.7 (1.8)	0.209 (0.007)
par	0.245 (0.008)	79.2 (1.9)	0.197 (0.006)
semi-par	0.246 (0.008)	79.1 (1.9)	0.198 (0.006)
$B = 1000$	MSPE	CovPr(80)	CRPS
emp	0.252 (0.011)	72.7 (1.9)	0.216 (0.007)
emp, loc	0.2622 (0.008)	79.7 (1.7)	0.205 (0.006)
par	0.2470 (0.008)	80.0 (1.6)	0.198 (0.006)
semi-par	0.2471 (0.007)	80.0 (1.6)	0.198 (0.006)

(emp) is poor for all measures, even though the norms decrease when the sample size B is increased. When using a localized version of the empirical approach (emp, loc), the performance is clearly improved from the straightforward empirical covariance specification method, except for the Frobenius norm which might carry some localization artifacts. Localization is here done by elementwise multiplication of the estimated covariance matrix with a tapering matrix setting covariance entries beyond a 10-cell range to 0. In Table III.5 we further see that the parametric approach (par), which has the same form as the generating mechanism in this case, is clearly the best for all norms. For a semi-parametric approach (semi-par), the norms are smaller than for the localized empirical approach. In the semi-parametric approach we set the diagonal entries of the covariance matrix from sample variances, while a single exponential correlation decay parameter is estimated by maximizing the likelihood, given the assigned variances. (See Section III.3.) In summary, the results indicate that the empirical approaches have very large Monte Carlo errors. They do not estimate $\boldsymbol{\Sigma}$ very well.

We next simulate data to study properties of the different covariance specification approaches under linear data updating. Data are collected at all locations in the 25 × 25 grid, according to $y(\boldsymbol{s}_i) = x(\boldsymbol{s}_i) + N(0, \tau^2)$, $i = 1, \dots, m = n = 625$, with $\tau = 0.5$. We study the performance in terms of posterior mean square prediction error (MSPE), ensemble coverage probabilities (CovPr) at the 80 % nominal level and continuously ranked probability score (CRPS), see e.g. Gneiting et al. (2007).

Table III.6 summarizes the results. For all prediction measures, the fully parametric and

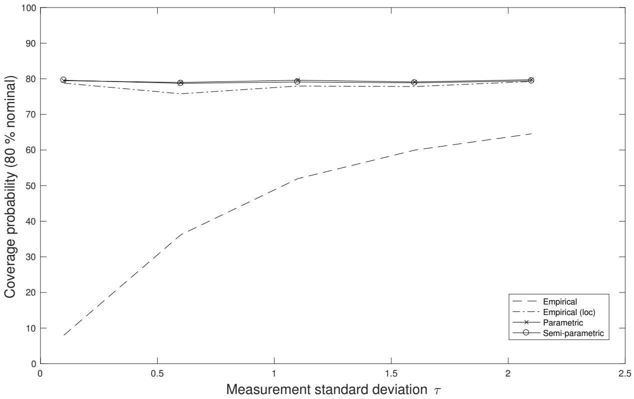


FIGURE III.30: Coverage probabilities for different measurement noise levels in the data.

semi-parametric approaches give the best results. The localized empirical method is a little worse, but much better than the straightforward empirical estimate. When the sample size increases, the latter improves markedly, but for $B = 1000$ it is still not at the performance levels of the other approaches. The poor performance of the empirical approach is largely due to sampling variability causing erroneous covariances which are again influencing the linear updating. In particular, the coverage probabilities of this empirical approach are very small at the 80 % nominal level.

Figure III.30 shows the relationship between the coverage probabilities and the measurement noise standard deviation τ. The two parametric approaches are close to the nominal level of 80 % for all noise levels, for $B = 100$. The localized empirical specification also performs well, while the straightforward empirical approach is very poor for small noise levels and only gradually goes towards the nominal level for larger noise levels. It is not surprising that the deviation from the nominal coverage level gets smaller for large measurement errors since the data has little influence on the update in that case. The very low coverage (10%) for noise standard deviation 0.1 is more surprising. In fact, one might expect the gain $\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \tau^2 \boldsymbol{I})^{-1}$ to be close to the identity matrix, because the addition of τ^2 is negligible. In this case, however, the Monte Carlo errors in the sample covariance matrix are too large relative to τ^2.

Since the computing time is larger for the parametric approaches one could argue that for a fair comparison a larger sample size should be used for the empirical approaches. Then again, typical applications of linear updating have long evaluation times for the mechanism providing $\boldsymbol{x}^{b,f}$, $b = 1, \ldots, B$, so the additional time spent on covariance estimation is negligible in comparison.

III.3 Parametric covariance estimation in the EnKF

Here we extend the parametric covariance estimation approach to dynamical systems. We begin by presenting some modeling and methodological assumptions. Then we describe the updating scheme incorporating parametric covariance estimation. Finally we show how this scheme can be applied to non-stationary models.

III.3.1 Assumptions

The state vector is denoted $\boldsymbol{x}_t = (x_{t,1}, \ldots, x_{t,n})'$, for time $t = 0, 1, \ldots, N$. Assuming a prior density $p(\boldsymbol{x}_0)$ at the initial step, the state evolves according to a dynamic model

$$\boldsymbol{x}_t = \boldsymbol{f}_t(\boldsymbol{x}_{t-1}, \boldsymbol{\delta}_t), \quad t = 1, \ldots, N, \tag{III.17}$$

where $\boldsymbol{\delta}_t$ is a noise term and the functional relationship defined by $\boldsymbol{f}_t$ is known. In realistic situations, this relationship is often obtained by forward integration of a system of differential equations. Moreover, data at time $t = 1, \ldots, N$ is denoted by $\boldsymbol{y}_t = (y_{t,1}, \ldots, y_{t,m})'$ and the likelihood model is defined by

$$\boldsymbol{y}_t = \boldsymbol{G}_t \boldsymbol{x}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \text{Normal}(\boldsymbol{0}, \boldsymbol{T}_t), \quad t = 1, \ldots, N, \tag{III.18}$$

where the design matrix $\boldsymbol{G}_t$ and covariance matrix $\boldsymbol{T}_t$ are known.

The goal of filtering is assessing the conditional density $p(\boldsymbol{x}_t | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_t)$, for times $t = 1, \ldots, N$. Because of the non-linear relationship in (III.17), there is no closed form expression for the filtering density. The EnKF sequentially computes and maintains an ensemble representation of the filtering distribution at all times. Assimilation is effected by linear updates of ensemble members with respect to observations. Starting from an analysis ensemble representation $\boldsymbol{x}_{t-1}^{b,a}$, $b = 1, \ldots, B$, at the previous time step, the EnKF iteration proceeds in two steps:

i) **Forecasting** by advancing each ensemble member through time by forward integration of the dynamical model,

$$\boldsymbol{x}_t^{b,f} = \boldsymbol{f}(\boldsymbol{x}_{t-1}^{b,a}, \boldsymbol{\delta}_t^b), \quad b = 1, \ldots, B. \tag{III.19}$$

ii) **Updating** the ensemble members with respect to data, based on a linear relationship between the two,

$$\boldsymbol{x}_t^{b,a} = \boldsymbol{x}_t^{b,f} + \hat{\boldsymbol{K}}_t^{-1}(\boldsymbol{y}_t - \boldsymbol{y}_t^{b,f}) \tag{III.20}$$

$$\hat{\boldsymbol{K}}_t = \hat{\boldsymbol{\Sigma}}_t \boldsymbol{G}_t'(\boldsymbol{G}_t \hat{\boldsymbol{\Sigma}}_t \boldsymbol{G}_t' + \boldsymbol{T}_t)^{-1}. \tag{III.21}$$

As in the static case described in Section III.2, the Kalman gain relies on an estimate of the forecast covariance matrix $\boldsymbol{\Sigma}_t = \mathrm{Cov}(\boldsymbol{x}_t | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1})$. The standard formulation of the EnKF uses the empirical or non-parametric covariance matrix of the forecast ensemble for this purpose,

$$\hat{\boldsymbol{\Sigma}}_t = \frac{1}{B}\sum_{b=1}^{B}(\boldsymbol{x}_t^{b,f} - \bar{\boldsymbol{x}}_t^f)(\boldsymbol{x}_t^{b,f} - \bar{\boldsymbol{x}}_t^f)', \quad \bar{\boldsymbol{x}}_t^f = \frac{1}{B}\sum_{b=1}^{B}\boldsymbol{x}_t^{b,f}. \tag{III.22}$$

However, as was discussed in the previous section, this direct empirical estimate is prone to large Monte Carlo errors. We proceed instead by describing parametric estimates of this covariance matrix, or the associated precision matrix.

III.3.2 Parametric EnKF update

Denote a parametric specification of the forecast covariance by $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_{t,\boldsymbol{\theta}_t}$. As suggested in (III.11) and (III.12), a parametric specification of the covariance is defined by $\boldsymbol{\Sigma}_{t,\hat{\boldsymbol{\theta}}_t}$ where

$$\hat{\boldsymbol{\theta}}_t = \mathrm{argmax}_{\boldsymbol{\theta}_t} l(\boldsymbol{\theta}_t; \boldsymbol{x}_t^{1,f}, \ldots, \boldsymbol{x}_t^{B,f}). \tag{III.23}$$

Again we assume this likelihood is that of a Gaussian process and, although there will be coupling over time because the Kalman gain is formed from ensemble-based covariance estimates, we proceed as if the B ensemble members are independent and identically distributed. This means that the likelihood is

$$l(\boldsymbol{\theta}_t) = -\frac{B}{2}\log|\boldsymbol{\Sigma}_{t,\boldsymbol{\theta}_t}| - \frac{1}{2}\sum_{b=1}^{B}(\boldsymbol{x}_t^{b,f} - \bar{\boldsymbol{x}}_t^f)'\boldsymbol{\Sigma}_{t,\boldsymbol{\theta}_t}^{-1}(\boldsymbol{x}_t^{b,f} - \bar{\boldsymbol{x}}_t^f). \tag{III.24}$$

In our formulation the parameters will vary over time. However, the change from one time point to the next tends to be small, so we start the optimization of the likelihood using the estimate from the previous time step. The actual maximization will depend on the functional form of the parametric covariance model, and whether derivatives are available (see (III.13)).

Again, it can sometimes be useful to fit parameters of the precision matrix $\boldsymbol{Q}_{t,\boldsymbol{\theta}_t}$, rather than working with the covariance matrix (see Section III.3.3).

III.3.3 Choice of parametric models

Common spatial covariance functions include the spherical and Matern-type with the exponential and the Gaussian as extreme cases (Cressie and Wikle, 2011). The exponential covariance function was used in Section III.2.4.

In the simplest, stationary model, the forecast variances $\text{Diag}(\boldsymbol{\Sigma}_{t,\boldsymbol{\theta}_t})$, are the same at all spatial locations, and pairwise correlation depend only on distance and not on specific locations. This is attractive from a computational point of view because there are only a few model parameters to estimate. For instance, the exponential covariance function has one variance parameter and one correlation decay parameter. Assuming that the target random field is stationary might be unrealistic in situations where the dynamical model affects various parts of the domain differently. Also, sparse data would lead to much smaller variance near data locations than far away, and possibly to a non-stationary correlation decay.

Non-stationary models are more flexible and, in the context of data assimilation for non-linear dynamical systems, arguably better suited at capturing relevant features of the spatio-temporal process. The main challenge of building a non-stationary model is that there are several kinds of non-stationarity. Which kind is useful for our situation? We discuss some approaches, and then pursue a couple of them in more detail.

There are several popular non-stationary covariance models, e.g. Paciorek and Schervish (2006) and Jun and Stein (2008). Various attempts have been made to impose structure in the spatial domain or via spatially varying covariates (Neto et al., 2014; Parker et al., 2016). However, non-stationary spatial models have been found to be relatively difficult to parameterize, mainly due to the requirement that the fitted model must give positive definite covariance matrices for any configuration of spatial sites. Another approach involves non-stationary modeling of the precision matrix or inverse covariance matrix: Fuglstad et al. (2015a) used spatially dependent basis functions to represent the precision structure. However, it has been difficult to estimate model parameters in such rich model formulations, particularly when many basis functions are involved, and sometimes much simpler parsimonious models perform equally well in practice (Fuglstad et al., 2015b).

The first non-stationary model we consider here is a semi-parametric approach where marginal variances can differ between spatial locations, while the correlation structure is the same everywhere (semi-par in Section III.2.4), see also Asfaw and Omre (2016). This model entails that the diagonal entries $\hat{\sigma}^2_{1,t}, \ldots, \hat{\sigma}^2_{n,t}$ of the forecast covariance matrix $\hat{\boldsymbol{\Sigma}}_t$ are specified empirically from the data,

$$\hat{\sigma}^2_{i,t} = \frac{1}{B}\sum_{b=1}^{B}(x^{f,b}_{t,i} - \bar{x}^f_{t,i})^2. \tag{III.25}$$

Assuming a parametric spatial correlation function, the likelihood is maximized using fixed variances as calculated in (III.25), meaning the likelihood is

$$\begin{aligned} l(\boldsymbol{\theta}_t) &= -\frac{B}{2}\log|\boldsymbol{\Sigma}_{t,\boldsymbol{\theta}_t}| - \frac{1}{2}\sum_{b=1}^{B}(\boldsymbol{x}^{b,f}_t - \bar{\boldsymbol{x}}^f_t)'\boldsymbol{\Sigma}^{-1}_{t,\boldsymbol{\theta}_t}(\boldsymbol{x}^{b,f}_t - \bar{\boldsymbol{x}}^f_t), \\ \boldsymbol{\Sigma}_{t,\boldsymbol{\theta}_t} &= \text{diag}(\hat{\sigma}_{1,t}, \ldots, \hat{\sigma}_{n,t})\boldsymbol{R}_{t,\boldsymbol{\theta}_t}\text{diag}(\hat{\sigma}_{1,t}, \ldots, \hat{\sigma}_{n,t}), \end{aligned} \tag{III.26}$$

where $\boldsymbol{R}_{t,\boldsymbol{\theta}_t}$ is the correlation matrix with unknown parameters, and diag$(\cdot)$ forms a diagonal matrix of the vector input.

Another parametric model we consider here is based on a stochastic partial differential equation (SPDE) formulation (Lindgren et al., 2011). Let Δ be the Laplacian operator, κ a model parameter and $\mathcal{W}(\mathbf{s})$ a white noise spatial process, and define the process $x(\boldsymbol{s})$ by

$$(\kappa^2 - \Delta)^{\alpha/2}x(\mathbf{s}) = \mathcal{W}(\mathbf{s}), \quad \mathbf{s} \in \mathbb{R}^2. \tag{III.27}$$

Lindgren et al. (2011) showed how spatial discretization naturally connects the SPDE in (III.27) to the precision matrix coefficients of the Gaussian Markov random field representation. The model parameters κ and α are not directly interpretable like the marginal variance or correlation decay, but Lindgren et al. (2011) showed closed-form relations between these parameters and variance, correlation range and smoothness parameters for the Matern covariance function. Fuglstad et al. (2015a) extended the SPDE in (III.27) to a non-stationary formulation,

$$(\kappa^2(\boldsymbol{s}) - \nabla \cdot \boldsymbol{H}(\boldsymbol{s})\nabla)^{\alpha/2} x(\mathbf{s}) = \mathcal{W}(\mathbf{s}). \tag{III.28}$$

Here, ∇ is the differential operator and the 2×2 matrix $\boldsymbol{H}(\boldsymbol{s})$ contains basis functions with location-dependent covariates, giving non-stationarity. We present a particular parameterization for the geological example in Section III.5, where the shape of a basis function is set on the basis of information from the forecast ensemble.

With precision matrix $\boldsymbol{Q}_{t,\boldsymbol{\theta}_t}$, the likelihood equals

$$\ell(\boldsymbol{\theta}_t) = \frac{B}{2} \log |\boldsymbol{Q}_{t,\boldsymbol{\theta}_t}| - \frac{1}{2} \sum_{b=1}^{B} (\boldsymbol{x}_t^{b,f} - \bar{\boldsymbol{x}}_t^f)^T \boldsymbol{Q}_{t,\boldsymbol{\theta}_t} (\boldsymbol{x}_t^{b,f} - \bar{\boldsymbol{x}}_t^f). \tag{III.29}$$

The maximum likelihood estimate of parameter $\boldsymbol{\theta}_t$ is computed by optimizing (III.29). As stated in Fuglstad et al. (2015a), the parametrization should not be too rich, as the optimization procedure can be hampered by a difficult likelihood surface. Analytical expressions for log-likelihood derivatives are available in some cases. However, for stability reasons, the numerical experiments of this paper use the simplex method for derivative-free optimization.

III.4 Simulation study for linear dynamic model

In this part we extend the 25×25 grid example to an autoregressive case in the spatio-temporal domain (Cressie and Wikle, 2011). As in Section III.2.4, the Gaussian initial distribution has mean 0, a covariance matrix $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}$ specifying a variance of 1 and an exponential covariance function with effective correlation range $\eta = 10$ cells. The dynamic model is

$$x_t(\mathbf{s}) = \phi x_{t-1}(\mathbf{s}) + \delta_t(\mathbf{s}), \quad \boldsymbol{\delta}_t \sim N(\mathbf{0}, (1-\phi^2)\boldsymbol{\Sigma}), \quad t = 1, \ldots, 10,$$

for all grid cells $\boldsymbol{s}$. With this covariance structure for the additive noise terms, $\boldsymbol{x}_t$ is a stationary spatial process over time. In the experiments we set $\phi = 0.9$.

The data gathering scheme is defined by sampling at $m = 15$ irregular sites, which are the same at all 10 time steps. This sparse sampling scheme will induce non-stationarity

Table III.7: Performance of EnKF variants using $B = 100$ ensemble members. Results at time step 10 in the autoregressive process.

	Far from data - Cell (2,13)			Near data - Cell (18,13)		
	MSPE	CovPr(80)	CRPS	MSPE	CovPr(80)	CRPS
emp	1.93	65	0.62	0.23	74	0.20
emp, loc	1.94	79	0.55	0.23	78	0.19
par	1.89	80	0.54	0.27	78	0.21
semi-par	1.89	80	0.54	0.23	78	0.19

in the covariance over time. The measurement noise terms are independent with variance 0.5^2.

Estimation approaches are again compared in terms of MSPE, coverage probability and CRPS. We consider two locations: Grid cell (2,13), which is far from data locations, and grid cell (18,13) which is near data locations. Results for ensemble size $B = 100$ are shown in Table III.7. These are averages over 500 replicates.

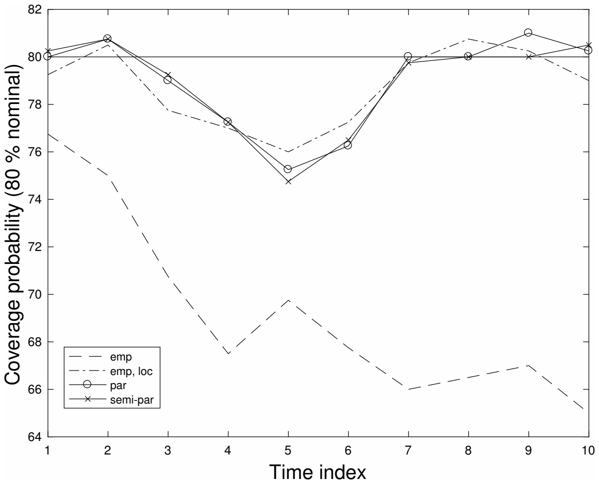


Figure III.31: Coverage probabilities for the grid cell far from data, plotted over time indices.

We notice that the straightforward empirical or non-parametric EnKF approach (emp) underestimates the variability in the prediction, especially for the cell far from data. Localization (emp, loc) improves this coverage problem, but it seems to give larger MSPE than the other approaches for sites far from data. The performance of the

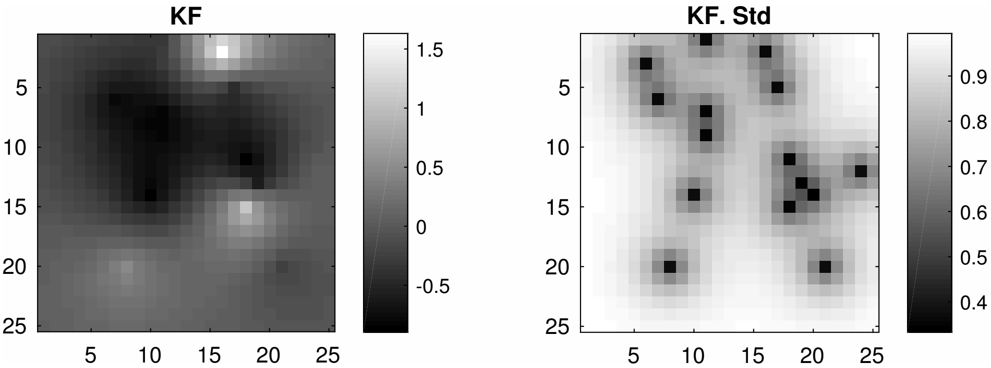


Figure III.32: Prediction (left) and prediction standard deviation (right) of the Kalman filter, at time step 10.

localized approach might be improved by tuning the tapering matrix, but considering the results of a particular taper still gives a basis for comparison with the other estimation methods. Perhaps surprisingly, the results of the simple parametric approach (par) are rather good even for this sparse design, where the true covariance is non-stationary. For the cell near data, however, its MSPE and CRPS are larger than those of the semi-parametric (semi-par) approach, which has the overall best performance.

Figure III.31 shows the coverage probabilities (at the 80 % nominal level) plotted against time indices. All approaches are shown for the grid cell far from data. The probabilities are roughly constant over time, except for the empirical approach where probabilities decline, likely due to the coupling of the ensemble members in the estimation of the Kalman gain (Sætrom and Omre, 2013). This effect is much smaller with localization and with parametric specification of the covariance parameters.

For this linear and Gaussian dynamical model, the optimal solution is provided by the Kalman filter, and we next compare the filtering results at time step 10 with this solution. The Kalman filter results are shown in Figure III.32, empirical EnKF results in Figure III.33, and semi-parametric EnKF results in Figure III.34.

The prediction obtained by the semi-parametric approach shows less small-scale variability than the straightforward empirical solution, because of smaller Monte Carlo errors in the covariance estimates. This smoothness makes the semi-parametric prediction more similar to the Kalman filter result. Moreover, the empirical approach has smaller standard deviations on average, leading to the low coverage probability in Table III.7.

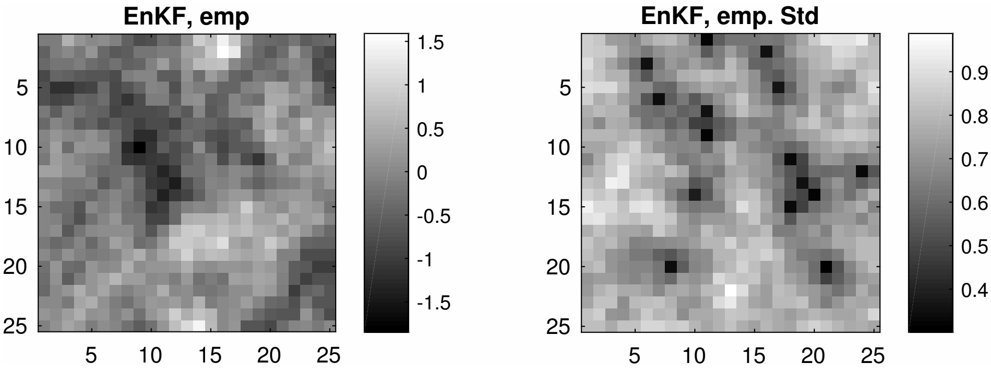


FIGURE III.33: Prediction (left) and prediction standard deviation (right) of the standard empirical EnKF, at time step 10.

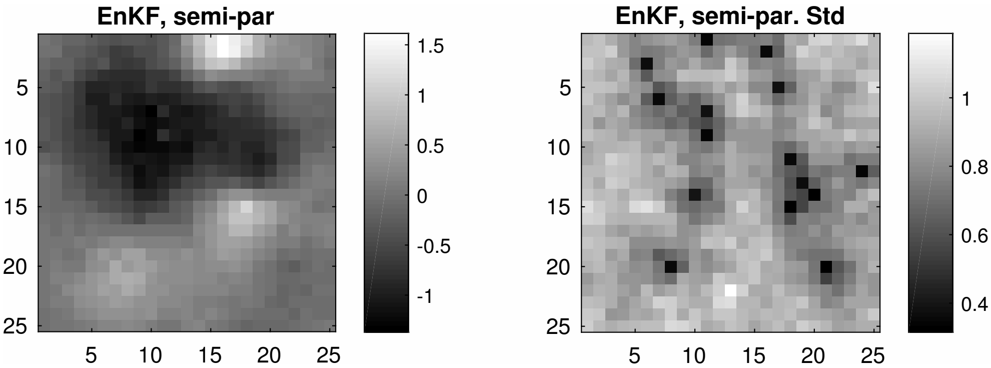


FIGURE III.34: Prediction (left) and prediction standard deviation (right) of the semi-parametric EnKF approach, at time step 10.

Figure III.35 shows the exact prediction covariance (solid) of this model at time step 10. This is plotted for the grid cells near and far from data, with the horizontal axis giving distance measured towards the south from each starting point. The covariance at distance 0 is much higher for the cell far from data, but it is more difficult to detect differences in the rate of decay with distance. Along with the exact calculation, the display shows the fitted prediction covariance using semi-parametric estimation (dashed). There are random variations caused by the empirical variance estimates, but the covariance decay appears similar to the Kalman filter results, indicating that non-stationarity in correlation is moderate for this sampling design. This reasonably good fit in terms of covariance seems to account for the good prediction efficiency of the semi-parametric approach.

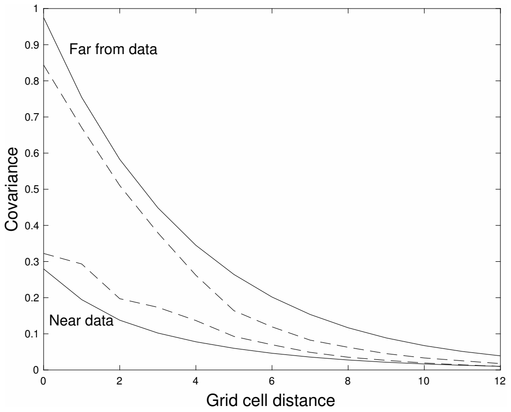


FIGURE III.35: Prediction covariances in the south direction from two grid cells near and far from data locations, at time step 10. Exact covariance (solid) and the fitted by the semi-parametric approach (dashed).

III.5 Example: Geological process model

We now apply the EnKF with parametric covariance estimation to a non-linear data assimilation problem from geology. The Geological Process Model (GPM) simulates erosion, transport and deposition of clastic sediment on length scales of tens to hundreds of kilometres over millions of years (Tetzlaff, 2005). In this case it is used to simulate sedimentation taking place in block F3 of the Dutch sector of the North Sea between 5 and 3.5 million years ago, during the early to mid Pliocene (Ogg et al., 2016).

III.5.1 Problem description and setup

Figure III.36 shows the rectangular model region, which measures 66 kilometres in the East-West direction and 37 kilometers in the North-South direction, discretized into a regular two-dimensional grid with a 0.5 km × 0.5 km resolution, for a total of $n = 9900$ cells.

The 1.5 million year time interval covered by the simulation is discretized into 15 time steps of 100 000 years each, indexed by $t = 0, 1, \dots, 15$. The simulator takes as input the initial elevation of the model region, i.e the surface shown in Figure III.36. Sediment then accumulates on top of this surface over time, producing a layered structure whose

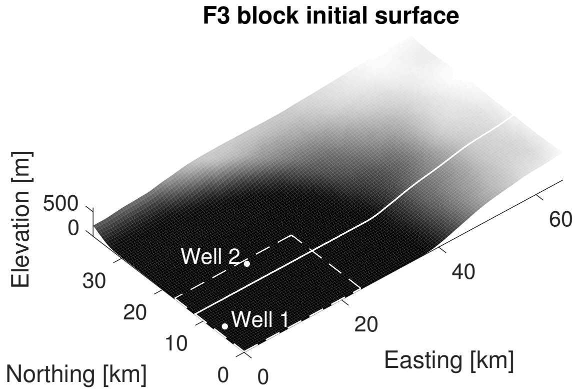


FIGURE III.36: Overview of the modeled region. The rectangular boundary of the F3 block is indicated by the dashed, white lines. To facilitate simulation of sediment influx the domain has been extended towards the north and east. Also included in the figure are two well locations where the thickness of the accumulating layer package is observed, and the location of the 2D section in Figure III.37, shown as a solid, white line.

thickness tends to increase more or less monotonically over time. Figure III.37 shows a 2D section through a simulated stack of 15 layers.

Statistical inversion of geological process models has been studied by e.g. Charvin et al. (2009b) who used Markov chain Monte Carlo sampling and Skauvold and Eidsvik (2018) who used EnKF and ensemble smoother approaches. The example given here is one part of a larger data assimilation problem. A more complete analysis would also track the sediment type composition or grain size distribution in the layer structure, and might additionally estimate changing environmental controls on the sedimentation process, such as sea level and sediment supply.

At time t the representation of the simulated layer package consists of $t+1$ layer boundary surfaces $\mathbf{z}_{kt} \in \mathbb{R}^n$, $k = 0, 1, \ldots, t$. These correspond to the black curves in Figure III.37. The cumulative thickness of all deposited sediment at time t is

$$\mathbf{x}_t = \mathbf{z}_{tt} - \mathbf{z}_{0t}, \tag{III.30}$$

and this will be the variable of interest here.

Noisy observations of cumulative thickness are available at every time step at two sites in the F3 block domain: Well 1 near the western boundary of the block and Well 2 near the northern boundary. These locations are shown in Figure III.36. For this example,

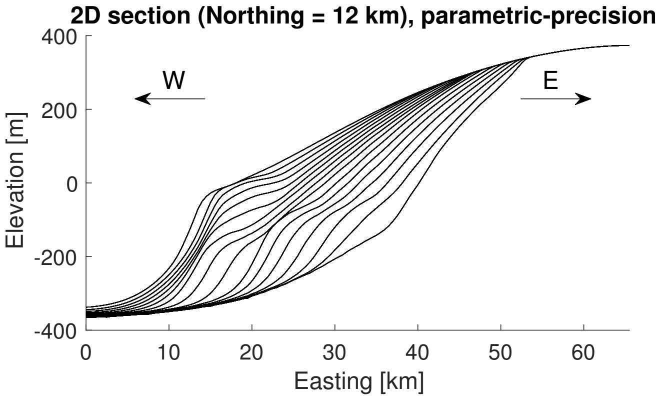


Figure III.37: Two-dimensional section through estimated layer package at $t = 15$, showing internal structure. One new layer is produced at the top of the stack every time step, leading to a strict chronological ordering with the oldest layers at the bottom. The location of the section is shown in Figure III.36 as a solid, white line.

the observations have been generated by running the simulator with known input to produce a reference realization of $\mathbf{z}_{kt}$ for $k = 1, 2, \ldots, t$ and $t = 1, 2, \ldots, N$ from which thicknesses were computed. Finally, Gaussian noise was added to the reference thickness values.

The goal of the data assimilation exercise is now to estimate the entire thickness field $\boldsymbol{x}_t$ at each time t, given the noisy measurements $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t$ at the well locations. We solve this filtering problem using three different versions of the EnKF: a) EnKF using a semi-parametric covariance model with empirical variances and likelihood estimation of a single correlation decay parameter. b) EnKF with a parametric representation of the precision matrix in the SPDE representation described in Section III.3.3. c) Empirical or nonparametric standard EnKF approach. Each EnKF variant is run once with $B = 50$ ensemble members.

In the precision parameterization, a three-element parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^T$ is used to specify $\boldsymbol{H}(\boldsymbol{s})$ and $\kappa(\boldsymbol{s})$ in the non-stationary SPDE (III.28),

$$\boldsymbol{H}(\mathrm{s}) \equiv \exp(\theta_1)\boldsymbol{I}, \quad \kappa(\mathrm{s}) = \exp(\theta_2)\left[1 + \theta_3\psi(\mathrm{s})\right], \tag{III.31}$$

where $\psi(\mathbf{s})$ is obtained at every time step t by smoothing out the ensemble variance of $\boldsymbol{x}_t$ and normalizing the smoothed variance estimate so that $\max_{\mathbf{s}}|\psi(\mathbf{s})|$ is equal to 1. In practice this leads to basis functions with larger values near the shoreline, i.e.

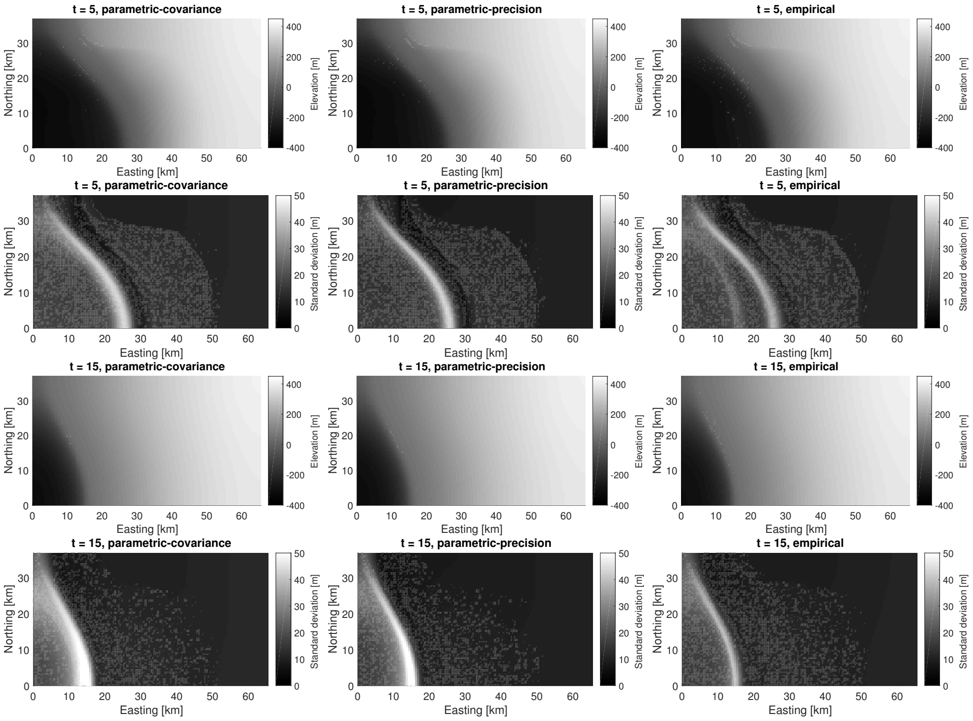


FIGURE III.38: Ensemble mean and standard deviation of top surface $\mathbf{z}_{tt}$ at time $t = 5$ and $t = 15$ for three EnKF variants.

the intersection between the top layer and the sea surface, as this is where the largest variances in thickness are found. An alternative way to create such basis functions would be to compute the location of the shoreline explicitly, and calculate the distance from every grid cell to the closest point on the shoreline.

This is a parsimonious parametrization of the precision structure, with only 3 parameters. Tests with more complex basis functions either led to difficulties in the likelihood optimization, or yielded prediction results that were very similar to those obtained using the simpler parametrization. This is in line with the findings of Fuglstad et al. (2015b).

III.5.2 Filtering results

Figures III.38 shows estimates of the top surface $\mathbf{z}_{tt}$ at $t = 5$ and $t = 15$ for the three EnKF variants. While the estimated fields are rather similar between the three ap-

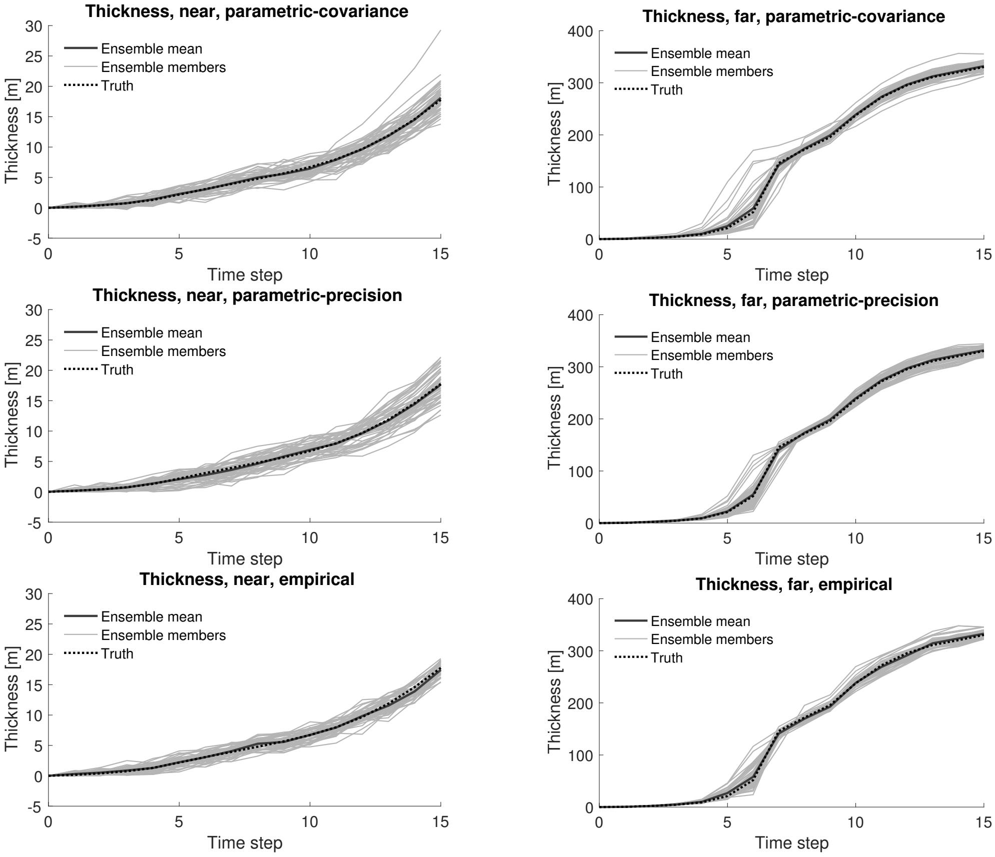


FIGURE III.39: Ensemble estimates of cumulative thickness x_t at time $t = 0, 1, \ldots, 15$ at two locations in block F3. Left: Thickness at location (15,5) which is close to Well 1. Right: Thickness at location (46,3) which in the southeast corner of the F3 block region, moderately far away from both wells.

proaches,there are differences in standard deviation. Relative to the empirical approach, the parametric estimates have both a higher overall variance level, and a sharper transition between the high and low variability regions.

Figure III.39 shows how the filtering ensemble evolves from $t = 1$ to $t = 15$ in each case. This display shows that the ensemble tracks the reference realization well in each case. The variability appears to be smaller at all times for the empirical approach.

Figure III.40 shows the location-wise rank of the reference thickness relative to EnKF ensemble thicknesses at time $t = 5$ and $t = 15$. The brighter a cell is, the larger the

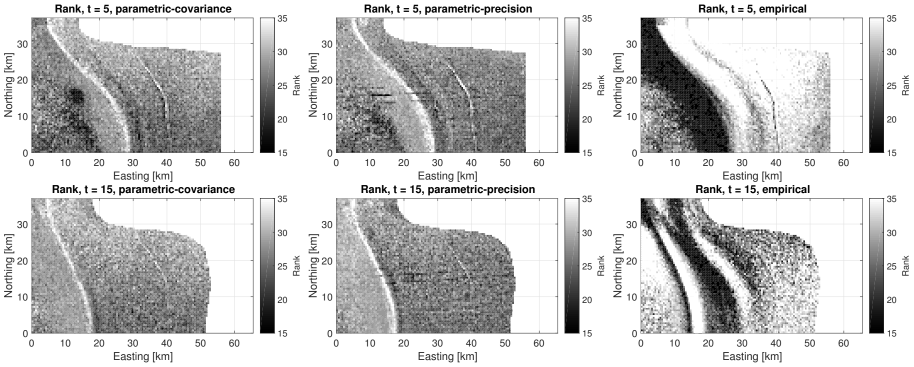


FIGURE III.40: Rank of reference realization total thickness relative to ensemble total thickness at $t = 5$ and $t = 15$ for parametric covariance, parametric precision and standard empirical EnKF variants. The strip of missing values along the northern and eastern boundaries is due to all ensemble members being equal on this region.

reference value is relative to the ensemble members at the location in question. Since the modeled field is spatially correlated, aggregating ranks over the domain may be misleading. Still, for a well-calibrated filter one expects a uniform distribution of ranks over the integers $\{1, 2, \ldots, B, B+1\}$. In the present case, however, we find that all three filter variants are overdispersive, with almost no ranks below 15 or over 35, which are the limits of the color ranges in Figure III.40. The rank plots of the two parametric filters have similar patterns, with the highest ranks concentrated near the shoreline, in the region where the estimated standard deviation is largest. This indicates that both parametric EnKF variants overestimate the variance in the high-variability region. In the rank plots of the empirical filter we see a different pattern, featuring large, contiguous, bright and dark regions of over- and underestimation. Furthermore, the rank plot at $t = 15$ appear to retain some features of the earlier pattern. The rank patterns of the parametric filters at $t = 15$ do not show clear traces of the earlier pattern.

III.6 Closing remarks

In this paper we have suggested some approaches for integrating more spatial statistical modeling in ensemble-based filtering methods. By parametrizing the covariance going into the updating step of the ensemble Kalman filter, we got results with improved prediction properties. In particular, the underestimation of variance that is often seen in ensemble Kalman filter predictions was reduced.

The parametric models used include simple stationary covariance models, and semi-parametric models for the covariance or precision matrix structure. In judging the results, we noted that parsimonious covariance structures did surprisingly well.

Applying the EnKF with parametric covariance estimation to a synthetic data assimilation test case, we found little difference between the results of using an exponential covariance function specification and a GMRF precision matrix specification of the correlation structure of the target random field. However, comparing these parametric filters with a standard stochastic EnKF whose gain matrix is based on empirical sample covariance estimators, we found clearer differences. In our non-linear test case, going from empirical to parametric covariance estimation gave no obvious improvement in estimates of the overall level of variance, but did appear to produce a less systematic pattern of bias in the estimated random field.

While introducing parametric estimation into a larger workflow can improve the quality of estimates, we find that very simple parametrizations tend to be preferable to even slightly more complex ones, as the flexibility gained by adding an extra parameter seldom makes up for the added difficulty in estimation. Using non-stationary variance entries and single parameter correlation sometimes improved results, while having complex precision structures led to difficult likelihood surfaces, without always improving predictive power. This means that finding a useful parametric model to embed in the ensemble Kalman filter update can be relatively easy, as one can bet on simplicity by choosing an uncomplicated model. Even if increased flexibility gives a better description of the random field being modeled, it does not follow that the estimates obtained under the more flexible model will be better than ones obtained under the simpler model in terms of predictive ability.

None of the parametric approaches studied here allow for anisotropy. This extension would give a few extra model parameters to estimate, and could be interesting for some applications. We conducted maximum likelihood estimation separately at every time step. The procedure could be extended to have coupling of parameters at different time steps. One could also apply Bayesian hierarchical models in this context.

Bibliography

Ades, M. and Van Leeuwen, P. J. (2013). An exploration of the equivalent weights particle filter. *Quarterly Journal of the Royal Meteorological Society*, 139(672):820–840.

Asch, M., Bocquet, M., and Nodet, M. (2016). *Data assimilation: methods, algorithms, and applications*. SIAM.

Asfaw, Z. G. and Omre, H. (2016). Localized/shrinkage kriging predictors. *Mathematical Geosciences*, 48(5):595–618.

Beaumont, M. A. (2010). Approximate Bayesian computation in evolution and ecology. *Annual review of ecology, evolution, and systematics*, 41:379–406.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

Bertoncello, A., Sun, T., Li, H., Mariethoz, G., and Caers, J. (2013). Conditioning surface-based geological models to well and thickness data. *Mathematical Geosciences*, 45(7):873–893.

Brookes, M. (2011). The matrix reference manual. *http://www.ee.imperial.ac.uk*.

Bryant, I. D. (1996). The application of physical measurements to constrain reservoir-scale sequence stratigraphic models. *Geological Society, London, Special Publications*, 104(1):51–63.

Catuneanu, O. (2002). Sequence stratigraphy of clastic systems: concepts, merits, and pitfalls. *Journal of African Earth Sciences*, 35(1):1–43.

Charvin, K., Gallagher, K., Hampson, G. L., and Labourdette, R. (2009a). A Bayesian approach to inverse modelling of stratigraphy, part 1: Method. *Basin Research*, 21(1):5–25.

Charvin, K., Gallagher, K., Hampson, G. L., and Labourdette, R. (2009b). A Bayesian approach to inverse modelling of stratigraphy, part 1: Method. *Basin Research*, 21(1):5–25.

Chen, Z. (2003). Bayesian filtering: From Kalman filters to particle filters, and beyond. *Statistics*, 182(1):1–69.

Chorin, A. J., Morzfeld, M., and Tu, X. (2013). A survey of implicit particle filters for data assimilation. In *State-Space Models*, pages 63–88. Springer.

Christ, A., Schenk, O., and Salomonsen, P. (2016). Using stratigraphic forward modeling to model the brookian sequence of the alaska north slope. In *Geostatistical and Geospatial Approaches for the Characterization of Natural Resources in the Environment*, pages 623–626. Springer.

Cressie, N. (1990). The origins of kriging. *Mathematical geology*, 22(3):239–252.

Cressie, N. and Wikle, C. K. (2011). *Statistics for spatio-temporal data*. John Wiley & Sons.

Dobson, A. J. and Barnett, A. (2008). *An introduction to generalized linear models*. CRC press.

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential monte carlo sampling methods for Bayesian filtering. *Statistics and computing*, 10(3):197–208.

Edwards, J., Lallier, F., and Caumon, G. (2016). Using a forward model as training model for 3D stochastic multi-well correlation. In *Second Conference on Forward Modelling of Sedimentary Systems*. EAGE.

Emerick, A. A. and Reynolds, A. C. (2013). Ensemble smoother with multiple data assimilation. *Computers & Geosciences*, 55:3–15.

Evensen, G. (2009). *Data assimilation: the ensemble Kalman filter*. Springer.

Frei, M. and Künsch, H. R. (2013). Bridging the ensemble Kalman and particle filters. *Biometrika*, 100(4):781–800.

Frolov, S., Baptista, A. M., Leen, T. K., Lu, Z., and van der Merwe, R. (2009). Fast data assimilation using a nonlinear Kalman filter and a model surrogate: An application to the Columbia River estuary. *Dynamics of Atmospheres and Oceans*, 48(1):16–45.

Fuglstad, G.-A., Lindgren, F., Simpson, D., and Rue, H. (2015a). Exploring a new class of non-stationary spatial Gaussian random fields with varying local anisotropy. *Statistica Sinica*, pages 115–133.

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2015b). Does non-stationary spatial data always require non-stationary random fields? *Spatial Statistics*, 14:505–531.

Furrer, R. and Bengtsson, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis*, 98(2):227–255.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gordon, N. J., Salmond, D. J., and Smith, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proceedings F (Radar and Signal Processing)*, volume 140(2), pages 107–113. IET.

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(3):550–560.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Hou, E., Lawrence, E., and Hero, A. O. (2016). Penalized ensemble Kalman filters for high dimensional non-linear systems. *arXiv preprint arXiv:1610.00195*.

Hutton, E. W. and Syvitski, J. P. (2008). Sedflux 2.0: An advanced process-response model that generates three-dimensional stratigraphy. *Computers & Geosciences*, 34(10):1319–1337.

Johnson, R. A. and Wichern, D. W. (1998). Applied multivariate statistical analysis, printice-hall. *Inc., Upper Saddle River, NJ.*

Jun, M. and Stein, M. L. (2008). Nonstationary covariance models for global data. *The Annals of Applied Statistics*, 2(4):1271–1289.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45.

Katzfuss, M., Stroud, J. R., and Wikle, C. K. (2016). Understanding the ensemble Kalman filter. *The American Statistician*, 70(4):350–357.

Keys, W. S. (1996). *A practical guide to borehole geophysics in environmental investigations*. CRC Press.

Killeen, P. (1982). Gamma-ray logging and interpretation. *Developments in geophysical exploration methods*, 3:95–150.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139.

Lagarias, J. C., Reeds, J. A., Wright, M. H., and Wright, P. E. (1998). Convergence properties of the nelder–mead simplex method in low dimensions. *SIAM Journal on optimization*, 9(1):112–147.

Laloy, E., Beerten, K., Vanacker, V., Christl, M., Rogiers, B., and Wouters, L. (2017). Bayesian inversion of a CRN depth profile to infer Quaternary erosion of the northwestern Campine Plateau (NE Belgium). *Earth Surface Dynamics*, 5(3):331.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Liu, J. S. and Chen, R. (1998). Sequential monte carlo methods for dynamic systems. *Journal of the American statistical association*, 93(443):1032–1044.

Lopez, S., Cojan, I., Rivoirard, J., and Galli, A. (2009). Process-based stochastic modelling: Meandering channelized reservoirs. *Analogue and Numerical Modelling of Sedimentary Systems: From Understanding to Prediction*, 40:139–144.

Lorenz, E. N. (1995). Predictability: A problem partly solved, vol. i. In *Proceedings of Seminar at ECMWF*.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Myrseth, I. and Omre, H. (2010). Hierarchical ensemble Kalman filter. *Spe Journal*, 15(02):569–580.

Necker, F., Härtel, C., Kleiser, L., and Meiburg, E. (2002). High-resolution simulations of particle-driven gravity currents. *International Journal of Multiphase Flow*, 28(2):279–300.

Neto, J. H. V., Schmidt, A. M., and Guttorp, P. (2014). Accounting for spatially varying directional effects in spatial covariance structures. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, 63(1):103–122.

Nychka, D. and Anderson, J. L. (2010). Data assimilation. *Handbook of Spatial Statistics, edited by: Gelfand, A., Diggle, P., Guttorp, P., and Fuentes, M., Chapman & Hall/CRC, New York*.

Ogg, J. G., Ogg, G. M., and Gradstein, F. M. (2016). *The Concise Geologic Time Scale*. Elsevier.

Paciorek, C. J. and Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5):483–506.

Paola, C. (2000). Quantitative models of sedimentary basin filling. *Sedimentology*, 47(s1):121–178.

Parker, R. J., Reich, B. J., and Eidsvik, J. (2016). A fused lasso approach to nonstationary spatial covariance estimation. *Journal of agricultural, biological, and environmental statistics*, 21(3):569–587.

Petersen, K. B., Pedersen, M. S., et al. (2008). The matrix cookbook. *Technical University of Denmark*, 7:15.

Pyrcz, M. J. and Deutsch, C. V. (2014). *Geostatistical reservoir modeling*. Oxford University Press.

Rezaie, J. and Eidsvik, J. (2012). Shrinked (1-α) ensemble Kalman filter and α Gaussian mixture filter. *Computational Geosciences*, 16(3):837–852.

Roelvink, J. and Van Banning, G. (1995). Design and development of Delft3D and application to coastal morphodynamics. *Oceanographic Literature Review*, 11(42):925.

Sacchi, Q., Weltje, G. J., and Verga, F. (2015). Towards process-based geological reservoir modelling: Obtaining basin-scale constraints from seismic and well data. *Marine and Petroleum Geology*, 61:56–68.

Sætrom, J. and Omre, H. (2011). Ensemble Kalman filtering with shrinkage regression techniques. *Computational Geosciences*, 15(2):271–292.

Sætrom, J. and Omre, H. (2013). Uncertainty quantification in the ensemble Kalman filter. *Scandinavian Journal of Statistics*, 40(4):868–885.

Sakov, P. and Bertino, L. (2011). Relation between two common localisation methods for the EnKF. *Computational Geosciences*, 15(2):225–237.

Särkkä, S. (2013). *Bayesian filtering and smoothing*, volume 3. Cambridge University Press.

Schenk, O., Magoon, L. B., Bird, K. J., and Peters, K. E. (2012). *Petroleum system modeling of northern Alaska*. AAPG Special Volumes.

Skauvold, J. and Eidsvik, J. (2018). Data assimilation for a geological process model using the ensemble Kalman filter. *Basin Research*, 30(4):730–745.

Snyder, C., Bengtsson, T., Bickel, P., and Anderson, J. (2008). Obstacles to high-dimensional particle filtering. *Monthly Weather Review*, 136(12):4629–4640.

Stordal, A. S., Karlsen, H. A., Nævdal, G., Skaug, H. J., and Vallès, B. (2011). Bridging the ensemble Kalman filter and particle filters: the adaptive Gaussian mixture filter. *Computational Geosciences*, 15(2):293–305.

Storms, J. (2003a). Simulating event-based shallow marine deposition over geological timescales. *Marine Geology*, 199:83–100.

Storms, J. E. (2003b). Event-based stratigraphic simulation of wave-dominated shallow-marine environments. *Marine Geology*, 199(1):83–100.

Storms, J. E., Weltje, G. J., Van Dijke, J., Geel, C., and Kroonenberg, S. (2002). Process-response modeling of wave-dominated coastal systems: simulating evolution and stratigraphy on geological timescales. *Journal of Sedimentary Research*, 72(2):226–239.

Strebelle, S. (2002). Conditional simulation of complex geological structures using multiple-point statistics. *Mathematical Geology*, 34(1):1–21.

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

Syvitski, J. P. and Hutton, E. W. (2001). 2D SEDFLUX 1.0 C: an advanced process-response numerical model for the fill of marine sedimentary basins. *Computers & Geosciences*, 27(6):731–753.

Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation*. SIAM.

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring coalescence times from dna sequence data. *Genetics*, 145(2):505–518.

Tetzlaff, D. and Priddy, G. (2001). Sedimentary process modeling: from academia to industry. In *Geologic Modeling and Simulation*, pages 45–69. Springer.

Tetzlaff, D., Tveiten, J., Salomonsen, P., Christ, A., Athmer, W., Borgos, H. G., Sonneland, L., Martinez, C., and Raggio, M. F. (2014). Geologic process modeling. In *IX Congreso de Exploración y Desarrollo de Hidrocarburos*. IAPG.

Tetzlaff, D. M. (2005). Modelling coastal sedimentation through geologic time. *Journal of coastal research*, 21(3):610–617.

Tetzlaff, D. M. and Harbaugh, J. W. (1989). *Simulating clastic sedimentation*. Computer Methods in the Geosciences, New York, NY, van Nostrand Reinholt.

Tikhonov, A. N., Goncharsky, A., Stepanov, V., and Yagola, A. G. (2013). *Numerical methods for the solution of ill-posed problems*, volume 328. Springer Science & Business Media.

Tsyrulnikov, M. and Rakitko, A. (2017). A hierarchical Bayes ensemble Kalman filter. *Physica D: Nonlinear Phenomena*, 338:1–16.

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

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

Ueno, G. and Tsuchiya, T. (2009). Covariance regularization in inverse space. *Quarterly Journal of the Royal Meteorological Society*, 135(642):1133–1156.

U.S. Geological Survey (1981). Wildcat well Tunalik 1, LAS format well log data. *https://certmapper.cr.usgs.gov/data/PubArchives/of00-200/wells/TUNALIK1/LAS/TU1LAS.HTM. Last accessed September 2017.*

Van Leeuwen, P. J. (2009). Particle filtering in geophysical systems. *Monthly Weather Review*, 137(12):4089–4114.

Van Leeuwen, P. J. (2010). Nonlinear data assimilation in geosciences: an extremely efficient particle filter. *Quarterly Journal of the Royal Meteorological Society*, 136(653):1991–1999.

Van Leeuwen, P. J., Künsch, H. R., Nerger, L., Potthast, R., and Reich, S. (2018). Particle filters for applications in geosciences. *arXiv preprint arXiv:1807.10434*.

Vogel, C. R. (2002). *Computational methods for inverse problems*, volume 23. SIAM.

Weisstein, E. W. (2002). Lambert W-function. From MathWorld—A Wolfram Web Resource. *http://mathworld.wolfram.com/LambertW-Function.html Last accessed on 1 August 2018.*

Zhu, M., Van Leeuwen, P. J., and Amezcua, J. (2016). Implicit equal-weights particle filter. *Quarterly Journal of the Royal Meteorological Society*, 142(698):1904–1919.