Nonlin. Processes Geophys., 18, 49–70, 2011
www.nonlin-processes-geophys.net/18/49/2011/
doi:10.5194/npg-18-49-2011

Nonlinear Processes in Geophysics

Earthquake forecasting based on data assimilation: sequential Monte Carlo methods for renewal point processes

M. J. Werner[1,*], **K. Ide**[2], **and D. Sornette**[3]

[1]Swiss Seismological Service, Institute of Geophysics, ETH Zurich, Switzerland
[2]Department of Atmospheric and Oceanic Science, Center for Scientific Computation and Mathematical Modeling, Institute for Physical and Scientific Technology, University of Maryland, College Park, USA
[3]Department of Management, Technology and Economics, and Department of Earth Sciences, ETH Zurich, Switzerland
[*]now at: Department of Geosciences, Princeton University, USA

Received: 27 January 2010 – Revised: 10 December 2010 – Accepted: 13 January 2011 – Published: 3 February 2011

Abstract. Data assimilation is routinely employed in meteorology, engineering and computer sciences to optimally combine noisy observations with prior model information for obtaining better estimates of a state, and thus better forecasts, than achieved by ignoring data uncertainties. Earthquake forecasting, too, suffers from measurement errors and partial model information and may thus gain significantly from data assimilation. We present perhaps the first fully implementable data assimilation method for earthquake forecasts generated by a point-process model of seismicity. We test the method on a synthetic and pedagogical example of a renewal process observed in noise, which is relevant for the seismic gap hypothesis, models of characteristic earthquakes and recurrence statistics of large quakes inferred from paleoseismic data records. To address the non-Gaussian statistics of earthquakes, we use sequential Monte Carlo methods, a set of flexible simulation-based methods for recursively estimating arbitrary posterior distributions. We perform extensive numerical simulations to demonstrate the feasibility and benefits of forecasting earthquakes based on data assimilation.

1 Introduction

In dynamical meteorology, the primary purpose of data assimilation has been to estimate and forecast as accurately as possible the state of atmospheric flow, using all available appropriate information (Talagrand, 1997). Recent advanced methods of data assimilation attempt to include the effects of uncertainties explicitly in the estimation by taking probabilistic approaches. Kalnay (2003) defines data assimilation as a statistical combination of observations and short-range forecasts. According to Wikle and Berliner (2007), data assimilation is an approach for fusing data (observations) with prior knowledge (e.g., mathematical representations of physical laws or model output) to obtain an estimate of the distribution of the true state of a process. To perform data assimilation, three components are required: (i) a statistical model for observations (i.e., a data or measurement model), (ii) an a priori statistical model for the state process (i.e., a state or process model), which may be obtained through a physical model of the time-evolving system, and (iii) a method to effectively merge the information from (i) and (ii).

Both data and model are affected by uncertainty, due to measurement and model errors and/or stochastic model elements, leading to uncertain state estimates that can be described by probability distributions. Data assimilation is therefore a Bayesian estimation problem: the prior is given by model output (a forecast from the past) and the likelihood by the measurement error distribution of the data. The posterior provides the best estimate of the true state and serves as initial condition for a new forecast. The essence of data assimilation is to inform uncertain data through the model, or, equivalently, to correct the model using the data. The cycle of predicting the next state and updating, or correcting this forecast given the next observation, constitutes sequential data assimilation (see Daley, 1991; Ghil and Malanotte-Rizzoli, 1991; Ide et al., 1997; Talagrand, 1997; Kalnay, 2003, for introductions to data assimilation and Tarantola, 1987; Miller et al., 1999; Pham, 2001, and Wikle and Berliner, 2007, for a Bayesian perspective).

Correspondence to: M. J. Werner
(mwerner@princeton.edu)

Although data assimilation is increasingly popular in meteorology, climatology, oceanography, computer sciences, engineering and finance, only a few partial attempts, reviewed in Sect. 2.1, have been made within the statistical seismology community to use the concept for seismic and fault activity forecasts. But earthquake forecasting suffers from the same issues encountered in other areas of forecasting: measurement uncertainties in the observed data and incomplete, partial prior information from model forecasts. Thus, basing earthquake forecasting on data assimilation may provide significant benefits, some of which we discuss in Sect. 2.2.

There are perhaps two major challenges for developing data assimilation methods for earthquake forecasts: seismicity models differ from standard models in data assimilation, and earthquake statistics are non-Gaussian. We briefly discuss each of the two issues.

First, seismicity models that are capable of modeling entire earthquake catalogs (i.e., occurrence times, locations and magnitude) generally belong to the class of stochastic point processes, which, loosely speaking, are probabilistic rules for generating a random collection of points (see Daley and Vere-Jones, 2003, for formal definitions). Examples of these seismicity models can be found in the works of Vere-Jones (1970, 1995), Kagan and Knopoff (1987), Ogata (1998), Kagan and Jackson (2000), Helmstetter and Sornette (2002), Rhoades and Evison (2004), and Werner et al. (2010a,b). This class of models is different from the class that is usually assumed in data assimilation, which is often cast in terms of discrete-time state-space models, or Hidden Markov models (HMMs), reflecting the underlying physics-based stochastic differential equations (Daley, 1991; Kalnay, 2003; Künsch, 2001; Cappé et al., 2005; Doucet et al., 2001). An HMM is, loosely speaking, a Markov chain observed in noise (Doucet et al., 2001; Durbin and Koopman, 2001; Künsch, 2001; Robert and Casella, 2004; Cappé et al., 2005, 2007): an HMM consists of an unobserved Markov (state) process and an associated, conditionally independent observation process (both processes being potentially nonlinear/non-Gaussian; see Sect. 3.1 for precise definitions). The Kalman filter is an archetypical assimilation method for such a model (Kalman, 1960; Kalman and Bucy, 1961). In contrast, earthquake catalogs have many features which make them uniquely distinct from the forecast targets in other disciplines and hence the models are completely different from the noisy differential or finite difference equations decorated by noise of standard data assimilation methods. There seems to exist little statistical work that extends the idea of data assimilation or state filtering to point processes, which model the stochastic pointwise space-time occurrence of events along with their marks.

The second challenge, that of non-Gaussian probability distributions, has been solved to some extent by recent Monte Carlo methods, at least for models with a small number of dimensions (Evensen, 1994; Liu, 2001; Doucet et al., 2001; Robert and Casella, 2004). In particular, Sequential Monte Carlo (SMC) methods, a set of simulation-based methods for recursively estimating arbitrary posterior distributions, provide a flexible, convenient and (relatively) computationally-inexpensive method for assimilating non-Gaussian data distributions into nonlinear/non-Gaussian models (Doucet et al., 2001; Durbin and Koopman, 2001; Künsch, 2001; Robert and Casella, 2004; Cappé et al., 2005, 2007). Also called particle filters, SMC filters have been particularly successful at low-dimensional filtering problems for the family of HMMs or state-space models. The Kalman-Lévy filter (Sornette and Ide, 2001) provides an analytic solution extending the Kalman filter for Lévy-law and power-law distributed model errors and data uncertainties. We present an overview of SMC methods in Sects. 3.3 and 3.4.

The main purpose of this article is to develop an implementable method for forecasting earthquakes based on data assimilation. We test this sequential method on a pedagogical and synthetic example of a simulated catalog of "observed" occurrence times of earthquakes, which are not the "true" event times because of observational errors. We specifically use a point-process as our model of seismicity. To estimate arbitrary posterior distributions of the "true" event times, we use the SMC methods we just mentioned. To benchmark their performance, we compare the results against those obtained by a simple Kalman filter and an ensemble Kalman filter.

Our technique offers a step towards the goal of developing a "brick-by-brick" approach to earthquake predictability (Jordan, 2006; Jackson, 1996; Kagan, 1999), given the enormous difficulties in identifying reliable precursors to impending large earthquakes (Geller, 1997; Geller et al., 1997; Kagan, 1997). With suitable adaptations and extensions, our approach should find its natural habitat in the general testing framework developed within the Regional Earthquake Likelihood Models (RELM) Working Group (Field, 2007a; Schorlemmer et al., 2007, 2010) and the international Collaboratory for the Study of Earthquake Predictability (CSEP) (Jordan, 2006; Werner et al., 2010c; Zechar et al., 2010), in which forecast-generating models are tested in a transparent, controlled, reproducible and fully prospective manner.

The importance of data uncertainties in earthquake predictability experiments was highlighted by several recent studies. Werner and Sornette (2008) showed that measurement errors in the magnitudes of earthquakes have serious, adverse effects on short-term forecasts that are generated from a general class of models of clustered seismicity, including two of the most popular models, the Short Term Earthquake Probabilities (STEP) model (Gerstenberger et al., 2005) and the Epidemic-Type Aftershock Sequence (ETAS) model (Ogata, 1988). Moreover, Werner and Sornette (2008) showed that the RELM evaluation tests are not appropriate for the broadened forecast distributions that arise from taking into account uncertainties in data and recommended that forecasts should be replaced by a full distribution. Schorlemmer et al. (2010) confirmed and supported this

recommendation after examining first results from the five-year RELM forecast competition. The methods used in this article for evaluating point-process forecasts when the observations are noisy provide an alternative to the current forecast evaluation method used in RELM and CSEP.

Data and parameter uncertainties also play a crucial role in the ongoing debate about the relevance of the seismic gap hypothesis (McCann et al., 1979; Nishenko, 1991; Kagan and Jackson, 1991, 1995; Rong et al., 2003; McGuire, 2008), of models of characteristic earthquakes (Wesnousky, 1994; Bakun et al., 2005; Scholz, 2002; Kagan, 1993) and of recurrence statistics of earthquakes on a particular fault segment inferred from paleoseismic data records (Biasi et al., 2002; Bakun et al., 2005; Davis et al., 1989; Rhoades et al., 1994; Ogata, 1999, 2002; Sykes and Menke, 2006; Parsons, 2008). The data are often modeled using renewal processes, and studies investigating data and parameter uncertainty confirmed that any model inference or forecast must take into account uncertainties (Davis et al., 1989; Rhoades et al., 1994; Ogata, 1999, 2002; Sykes and Menke, 2006; Parsons, 2008).

In this article, we focus on the class of renewal processes as models of seismicity. On the one hand, renewal processes are extensively used to model paleoseismic data records, characteristic earthquakes, seismic gaps and seismic hazard, as mentioned above. On the other hand, renewal processes are the point-process analog of Markov chains, thereby enabling us to use sequential Monte Carlo methods developed for state-space models. In other words, renewal processes are the simplest class of point process models relevant to statistical seismology. By developing rigorously a data assimilation procedure for renewal processes, we aim at providing the building blocks for more complicated models. In addition to the obvious relevance to earthquake forecasts, we hope to generate interest among statisticians to tackle the general problem of state filtering for point processes, for which the Markovian state-space model framework seems too restrictive.

The article is structured as follows. Section 2 provides a brief literature review of data assimilation in connection with statistical seismology and points out potential benefits of data assimilation to earthquake forecasting. Section 3 introduces the methods we believe are relevant in the seismicity context. Section 3.1 provides the notation and basic Bayesian estimation problem we propose to solve for renewal processes. Section 3.2 defines renewal processes, which serve as our forecast models. Section 3.3 explains the basics of Sequential Monte Carlo methods. In Sect. 3.4, we describe a particular SMC filter. To perform model inference, we must estimate parameters, which is described in Sect. 3.5. In Sect. 3.6, we describe two more filters that will serve as benchmarks for the particle filter: a simple deterministic Kalman filter and an Ensemble Kalman filter. Section 4 describes numerical experiments to demonstrate how earthquake forecasting based on data assimilation can be implemented for a particular renewal process, where inter-event times are lognormally distributed. Section 4.1 describes the set-up of the simulations: we use a lognormal renewal process of which only noisy occurrence times can be observed. In Sect. 4.2 we use the particle and Kalman filters to estimate the actual occurrence times, demonstrating that the filters improve substantially on a forecasting method that ignores the presence of data uncertainties. In Sect. 4.3, we show that parameter estimation via maximum (marginal) likelihood is feasible. We conclude in Sect. 5.

2 Data assimilation and probabilistic earthquake forecasting

2.1 Literature on probabilistic earthquake forecasting and data assimilation

The general concepts of data assimilation or Hidden Markov models (HMMs) state inference are relatively new to statistical earthquake modeling. The few studies that are related can be separated into three categories. (i) Varini (2005, 2008) studied a HMM of seismicity, in which the (unobserved) state could be in one of three different states (a Poisson process state, an ETAS process state and a stress-release process state) and the observational data were modeled according to one of the three processes. Varini did not consider measurement uncertainties of the data. (ii) Grant and Gould (2004) proposed data formats and standards for the assimilation of uncertain paleoseismic data into earthquake simulators. Van Aalsburg et al. (2007) assimilated uncertain paleoseismic data into "Virtual California", a fixed-geometry earthquake simulator of large earthquakes: model runs are accepted or rejected depending on whether simulated earthquakes agree with the paleoseismic record. (iii) Rhoades et al. (1994) calculated seismic hazard on single fault segments by averaging the hazard function of a renewal process over parameter and data uncertainties, achieved by sampling over many parameter and data samples. Ogata (1999) presented a Bayesian approach to parameter and model inference on uncertain paleoseismic records, closely related to our approach. Data uncertainties were represented with either a uniform or a triangular distribution. To compute the integrals, Ogata seems to have used numerical integration, a process that becomes increasingly difficult as the number of events increases, in contrast to the particle filters that we use below. Sykes and Menke (2006) assumed Gaussian data errors and uncorrelated recurrence intervals, also providing a maximum likelihood estimation procedure for the parameters of a lognormal process based on a Monte Carlo integration approach. Parsons (2008) provided a simple but inefficient Monte Carlo method for estimating parameters of renewal processes from paleoseismic catalogs.

2.2 Why base earthquake forecasting on data assimilation?

Data assimilation can be used as a framework for likelihood-based model inference and development, fully accounting for uncertainties. The current surge in earthquake predictability experiments (Field, 2007a; Jordan, 2006; Schorlemmer et al., 2007, 2010; Werner et al., 2010c; Zechar et al., 2010) provides strong motivational grounds for developing earthquake forecasting methods that are robust with respect to observational uncertainties in earthquake catalogs. Dealing with observational errors is particularly important for operational earthquake forecasts (e.g., Jordan and Jones, 2010), as observations are poorer and scarcer in real-time. Data assimilation provides a vehicle for correcting an existing forecast without having to re-calibrate and re-initialize the model on the entire data set. In its general formulation as a state and parameter estimation problem, data assimilation may also be viewed as a method for estimating physical quantities ("states") and model parameters, directly related to physics-based models, such as rate-and-state friction and Coulomb stress-change models (see, e.g., Hainzl et al., 2009). In the future, the coupled integration of several types of different data to constrain estimates of physical states is highly desirable. Numerical weather prediction has a long history of integrating different types of data – statistical seismology may be able to adapt these methods. Finally, the theory of point processes has so far largely focused on exact data (e.g., Daley and Vere-Jones, 2003). The development of the statistical theory and practical methodology for taking into account noisy observations is therefore interesting for applications beyond earthquake forecasting.

3 Method: sequential Monte Carlo methods for renewal processes

3.1 Bayesian data assimilation of state-space or Hidden Markov Models (HMMs)

In this section, we state the general problem of Bayesian data assimilation that will be solved for specific model and observation assumptions in Sect. 4. The presentation borrows from Doucet et al. (2000, 2001) and Arulampalam et al. (2002) (see also Künsch, 2001; Robert and Casella, 2004; Cappé et al., 2005, 2007; Wikle and Berliner, 2007, and references therein).

We use the class of Hidden Markov Models (HMMs), i.e. Markovian, nonlinear, non-Gaussian state-space models. The unobserved signal (the hidden states) $\{x_t\}_{t\geq 1}$ is modeled as a Markov process (in this article, x_t is a scalar). The initial state x_0 has initial distribution $p(x_0)$. The transition from x_t to x_{t+1} is governed by a Markov transition probability distribution $p(x_{t+1}|x_t)$. The observations $\{y_t\}_{t\geq 1}$ are assumed to be conditionally independent given the process $\{x_t\}_{t\geq 1}$ and of conditional distribution $p(y_t|x_t)$ (the observations may also be vectors, in general of different dimension than the state). The model can be summarized by

$$\text{Initial condition:} \quad p(x_0) \tag{1}$$

$$\text{Model forecast:} \quad p(x_{t+1}|x_t) \qquad t \geq 0 \tag{2}$$

$$\text{Conditional data likelihood:} \quad p(y_t|x_t) \qquad t \geq 1 \tag{3}$$

We denote $x_{0:t} = \{x_0, \ldots, x_t\}$ and $y_{1:t} = \{y_1, \ldots, y_t\}$. The problem statement is then as follows: the aim is to estimate sequentially in time the posterior distribution $p(x_{0:t}|y_{1:t})$. We may also be interested in estimating the marginal distribution $p(x_t|y_{1:t})$, also known as the filtering distribution, and the marginal complete data likelihood $p(y_{1:t})$, which we will use for parameter estimation.

At any time t, the posterior distribution is given by Bayes' theorem

$$p(x_{0:t}|y_{1:t}) = \frac{p(y_{1:t}|x_{0:t})\ p(x_{0:t})}{\int p(y_{1:t}|x_{0:t})\ p(x_{0:t})dx_{0:t}} \tag{4}$$

A recursive or sequential formula can be derived from (i) the Markov property of the state process and (ii) the independence of observations given the state:

$$p(x_{0:t+1}|y_{1:t+1}) = p(x_{0:t}|y_{1:t})\ \frac{p(y_{t+1}|x_{t+1})\ p(x_{t+1}|x_t)}{p(y_{t+1}|y_{1:t})} \tag{5}$$

where $p(y_{t+1}|y_{1:t})$ is given by

$$p(y_{t+1}|y_{1:t}) = \int p(y_{t+1}|x_{t+1})p(x_{t+1}|x_t)p(x_{0:t}|y_{1:t})dx_{0:t+1} \tag{6}$$

The marginal distribution $p(x_t|y_{1:t-1})$ also satisfies the following recursion:

$$p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})dx_{t-1} \tag{7}$$

$$p(x_t|y_{1:t}) = \frac{p(y_t|x_t)\ p(x_t|y_{1:t-1})}{\int p(y_t|x_t)\ p(x_t|y_{1:t-1})dx_t} \tag{8}$$

Expressions (7) and (8) are the essential steps in sequential data assimilation. Using the last update (the posterior, also often called analysis) as initial condition, the Chapman-Kolmogorov (prediction) Eq. (7) is used to forecast the state at the next time step. When observations y_t become available, they are assimilated into the model forecast by the update Eq. (8). This cycle constitutes sequential data assimilation of state-space models. The problem appears in other research fields under different guises, e.g. Bayesian, optimal, nonlinear or stochastic filtering, or online inference and learning (Doucet et al., 2001; Cappé et al., 2005).

In general, there may be unknown parameters in the model forecast distribution that need to be estimated. We assume that the parameters of the conditional data likelihood are

known, since they should be characterized by the measurement process and its associated uncertainties. Several parameter estimation techniques exist; we will focus on maximizing the marginal complete data likelihood, the denominator in Bayes' theorem:

$$p(y_{1:t}) = \int p(y_{1:t}|x_{0:t})\ p(x_{0:t})dx_{0:t} \tag{9}$$

Equation (9) provides a measure of how successfully a particular model is explaining the data. The marginal complete data likelihood is the analog of the traditional likelihood function, but generalized to noisy observational data. This, in turn, implies that different models may be compared and tested for their consistency with observed data, while explicitly acknowledging data uncertainties. In other words, (earthquake) forecasts may be evaluated based on this measure.

Only in very special cases are the prediction and update Eqs. (7) and (8) amenable to analytical solutions. In the case of a linear Gaussian state-space model, the widespread Kalman filter (Kalman, 1960; Kalman and Bucy, 1961) calculates exactly the posterior distributions. Much of filtering theory and data assimilation has been concerned with identifying useful, suitable and computationally inexpensive filters for a variety of particular problems. For instance, the extended Kalman filter performs a local tangent linearization of nonlinear model and observation operators for nonlinear problems. The Kalman-Lévy filter (Sornette and Ide, 2001) generalizes the Kalman filter to Lévy-law and power-law distributed model and data uncertainties. In other cases, numerical integration may be possible, or approximate grid-based methods, e.g. HMM filters, may be convenient. The ensemble Kalman filter (Evensen, 1994; Tippett et al., 2003) is a Monte Carlo approach to the nonlinear extension of the Kalman filter by introducing an ensemble of particles with equal weights, each evolved individually, to approximate distributions. The general, nonlinear, non-Gaussian, sequential Bayesian estimation problem, however, seems best solved with sequential Monte Carlo methods whenever the model's dimensionality is small (usually less than several dozen according to Snyder et al., 2008).

3.2 Renewal processes as forecast models

Data assimilation is an iterative method that involves two steps, forecast (7) and analysis (8), in each cycle. To formulate the data assimilation problem for earthquakes, we use a renewal point process as the model in the forecast. Renewal point processes are characterized by intervals between successive events that are identically and independently distributed according to a probability density function that defines the process (Daley and Vere-Jones, 2003). Examples of such a probability density function (pdf) include the lognormal, exponential, gamma, Brownian passage time and Weibull pdf. The time of the next event in a renewal process depends solely on the time of the last event:

$$p(t_k|t_{k-1}) = p(t_k - t_{k-1}) = p(\tau) \tag{10}$$

where τ is the interval between events. The time of the event t_k corresponds to the model state x_k in data assimilation. Renewal point processes provide prior information for the analysis, which we will discuss in the next section.

The class of renewal processes is widely used in seismology and seismic hazard analysis. For example, Field (2007a) summarized how the Working Group on California Earthquake Probabilities (WGCEP), mandated to provide the official California seismic hazard map, estimates the occurrence probability of large earthquakes on major faults in the region by using various recurrence models, including the lognormal pdf. While physics-based models of seismicity certainly exist, the models are non-unique, the physics is far from fully understood, and we lack basic measurements (e.g. of the state of stress) to properly calibrate such models. As a result, most seismic hazard analyses are either entirely time-independent (i.e., they use an exponential pdf), an approach pioneered by Cornell (1968) that remains state-of-the-art in many regions. Or alternatively, only the probabilities of large earthquakes on major fault segments are estimated with renewal models calibrated with paleoseismological and more recent instrumental data. To infer the most appropriate pdf, seismologists use likelihood-based inference of renewal models.

Renewal models can also be motivated by the elastic rebound theory proposed by Reid (1910). According to the theory, large earthquakes release the elastic strain that has built up since the last large earthquake. Some seismologists deduce that the longer it has been since the last earthquake, the more probable is an imminent event (e.g. Nishenko, 1991; Sykes and Menke, 2006), while others contend that the data contradict this view (e.g. Davis et al., 1989; Sornette et al., 1996; Kagan and Jackson, 1995). Renewal models are often used to quantitatively demonstrate that earthquakes either cluster or occur quasi-periodically.

3.3 Sequential Monte Carlo methods

Earthquake statistics often violate Gaussian approximations in terms of their temporal, spatial and magnitude occurrences, so much so that approximate algorithms based on Gaussian approximations (e.g. the traditional Kalman filter) are unlikely to produce good results. Furthermore, the continuous state space of seismicity rules out methods in which that space is assumed to be discrete (such as grid-based methods). This leaves us with numerical integration techniques and Monte Carlo methods. The former are numerically accurate but computationally expensive in problems with medium to high dimensionality.

Sequential Monte Carlo (SMC) methods bridge the gap between these cost-intensive methods and the methods based on Gaussian approximations. They are a set of simulation-based methods that provide a flexible alternative to computing posterior distributions. They are applicable in very

general settings, parallelisable and often relatively easy to implement. SMC methods have been applied in target tracking, financial analysis, diagnostic measures of fit, missing data problems, communications and audio engineering, population biology, neuroscience, and many more. Good introductions were provided by Arulampalam et al. (2002), Cappé et al. (2005, 2007), Doucet et al. (2000, 2001), Künsch (2001), Liu (2001), Liu and Chen (1998) and de Freitas (1999, Chapter 6).

Sequential Monte Carlo filters use the techniques of Monte Carlo sampling, of (sequential) importance sampling and of resampling, which we describe briefly below before defining a particular particle filter which we will use for our numerical experiments.

3.3.1 Monte Carlo sampling

In Monte Carlo (MC) simulation (Liu, 2001; Robert and Casella, 2004), a set of N weighted "particles" (or samples) $x_{0:t}^{(i)}$ are drawn identically and independently from a distribution, say, a posterior $p(x_{0:t}|y_{1:t})$. Then, an empirical estimate of the distribution is given by

$$\hat{p}_N(x_{0:t}|y_{1:t}) = \frac{1}{N}\sum_i^N \delta_{x_{0:t}^{(i)}}(x_{0:t}) \tag{11}$$

where $\delta_{x_{0:t}^{(i)}}(x_{0:t})$ denotes the Dirac mass located at $x_{0:t}^{(i)}$. The essential idea of Monte Carlo sampling is to convert an integral into a discrete sum. One is often interested in some function of the posterior distributions, say, its expectation, covariance, marginal or another distribution. Estimates of such functions $I(f_t)$ can be obtained from

$$I_N(f_t) = \int f_t(x_{0:t})\hat{p}_N(x_{0:t}|y_{1:t})dx_{0:t} = \frac{1}{N}\sum_i^N f_t(x_{0:t}^{(i)}) \tag{12}$$

This estimate is unbiased. If the posterior variance of $f_t(x_{0:t})$ is finite, say $\sigma_{f_t}^2$, then the variance of $I_N(f_t)$ is equal to $\sigma_{f_t}^2/N$. From the law of large numbers,

$$I_N(f_t) \xrightarrow[N\to\infty]{a.s.} I(f_t) \tag{13}$$

where a.s. denotes almost sure convergence. That is, the probability that the estimate $I_N(f_t)$ converges to the "true" value $I(f_t)$ equals one in the limit of infinite number of particles. Furthermore, if the posterior variance $\sigma_{f_t}^2 < \infty$, then a central limit theorem holds:

$$\sqrt{N}(I_N(f_t) - I(f_t)) \xrightarrow[N\to\infty]{\Delta} \mathcal{N}(0, \sigma_{f_t}^2) \tag{14}$$

where $\xrightarrow[N\to\infty]{\Delta}$ denotes convergence in distribution and $\mathcal{N}(0,\sigma_{f_t}^2)$ is the normal (Gaussian) distribution with mean zero and variance $\sigma_{f_t}^2$. The advantage of this perfect Monte Carlo method is therefore that the rate of convergence of the MC estimate is independent of the dimension of the integrand. This stands in contrast to any deterministic numerical integration method, whose rate of convergence decreases with the dimensionality of the integrand.

Unfortunately, because the posterior distribution is usually highly complex, multi-dimensional and only known up to a normalizing constant, it is often impossible to sample directly from the posterior. One very successful solution for generating samples from such distributions is Markov Chain Monte Carlo (MCMC). Its key idea is to generate samples from a proposal distribution, different from the posterior, and then to cause the proposal samples to migrate, so that their final distribution is the target distribution. The migration of the samples is caused by the transition probabilities of a Markov chain (see, e.g., Appendix D of de Freitas, 1999). However, MCMC are iterative algorithms unsuited to sequential estimation problems and will not be pursued here. Rather, SMC methods primarily rely on a sequential version of importance sampling.

3.3.2 Importance Sampling (IS)

Importance Sampling (IS) introduced the idea of generating samples from a known, easy-to-sample probability density function (pdf) $q(x)$, called the importance density or proposal density, and then "correcting" the weights of each sample so that the weighted samples approximate the desired density. As long as the support of the proposal density includes the support of the target density, one can make use of the substitution

$$p(x_{0:t}|y_{1:t}) = \frac{p(x_{0:t}|y_{1:t})}{q(x_{0:t}|y_{1:t})}q(x_{0:t}|y_{1:t}) \tag{15}$$

to obtain the identity

$$I(f_t) = \frac{\int f_t(x_{0:t})w(x_{0:t})q(x_{0:t}|y_{1:t})dx_{0:t}}{\int w(x_{0:t})q(x_{0:t}|y_{1:t})dx_{0:t}} \tag{16}$$

where $w(x_{0:t})$ is known as the importance weight

$$w(x_{0:t}) = \frac{p(x_{0:t}|y_{1:t})}{q(x_{0:t}|y_{1:t})} \tag{17}$$

Therefore, if one can generate N independently and identically distributed samples $x_{0:t}^{(i)}$ from the importance density $q(x_{0:t}|y_{0:t})$, a Monte Carlo estimate of $I(f_t)$ is given by

$$\hat{I}_N(f_t) = \frac{\frac{1}{N}\sum_i^N f_t(x_{0:t}^{(i)})w(x_{0:t}^{(i)})}{\frac{1}{N}\sum_j^N w(x_{0:t}^{(j)})} = \sum_i^N f_t(x_{0:t}^{(i)})\tilde{w}_t^{(i)} \tag{18}$$

where the normalized importance weights $\tilde{w}_t^{(i)}$ are given by

$$\tilde{w}_t^{(i)} = \frac{w(x_{0:t}^{(i)})}{\sum_{j=1}^N w(x_{0:t}^{(j)})} \tag{19}$$

Thus, the posterior density function can be approximated arbitrarily well by the point-mass estimate

$$\hat{p}(x_{0:t}|y_{1:t}) = \sum_{i}^{N} \tilde{w}_t^{(i)} \delta_{x_{0:t}^{(i)}}(x_{0:t}) \tag{20}$$

In summary, the advantage that IS introduces lies, firstly, in being able to easily generate samples from the importance density rather than a potentially complex target density, and, secondly, in only needing to correct the weights of the samples from the ratio of the target and importance densities, eliminating the need to calculate normalization constants of the target density.

3.3.3 Sequential Importance Sampling (SIS)

In its simplest form, IS is not adequate for sequential estimation. Whenever new data y_t become available, one needs to recompute the importance weights over the entire state sequence. Sequential Importance Sampling (SIS) modifies IS so that it becomes possible to compute an estimate of the posterior without modifying the past simulated trajectories. It requires that the importance density $q(x_{0:t}|y_{1:t})$ at time t admits as marginal distribution at time $t-1$ the importance function $q(x_{0:t-1}|y_{1:t-1})$:

$$q(x_{0:t}|y_{1:t}) = q(x_{0:t-1}|y_{1:t-1}) q(x_t|x_{0:t-1}, y_{1:t}) \tag{21}$$

After iterating, one obtains:

$$q(x_{0:t}|y_{1:t}) = q(x_0) \prod_{k=1}^{t} q(x_k|x_{0:k-1}, y_{1:k}) \tag{22}$$

Assuming that the state evolves according to a Markov process and that the observations are conditionally independent given the states, one can obtain

$$p(x_{0:t}) = p(x_0) \prod_{k=1}^{t} p(x_k|x_{k-1}) \tag{23}$$

and

$$p(y_{1:t}|x_{0:t}) = \prod_{k=1}^{t} p(y_k|x_k) \tag{24}$$

Substituting Eqs. (22), (23) and (24) into Eq. (19) and using Bayes' theorem, we arrive at a recursive estimate of the importance weights

$$\tilde{w}_t^{(i)} \propto \tilde{w}_{t-1}^{(i)} \frac{p(y_t|x_t^{(i)}) p(x_t^{(i)}|x_{t-1}^{(i)})}{q(x_t^{(i)}|x_{0:t-1}^{(i)}, y_{1:t})} \tag{25}$$

where the normalization is provided by $\sum_{j=1}^{N} \tilde{w}_t^{(j)}$. Equation (25) provides a mechanism for sequentially updating the importance weights. In summary, SIS provides a method to approximate the posterior density function (20) (or some function thereof) sequentially in time without having to draw samples directly from the posterior. All that is required is (i) sampling from the importance density and evaluating it up to some constant, (ii) evaluating the likelihood $p(y_t|x_t^{(i)})$ up to some proportionality constant, (iii) evaluating the forecast $p(x_t^{(i)}|x_{t-1}^{(i)})$ up to some constant, and (iv) normalizing the importance weights via $\sum_{j=1}^{N} \tilde{w}_t^{(j)}$. The SIS thus makes sequential Bayesian estimation feasible.

3.3.4 Choice of the importance density and resampling

The problem encountered by the SIS method is that, as t increases, the distribution of the importance weights becomes more and more skewed. For instance, if the support of the importance density is broader than the posterior density, then some particles will have their weights set to zero in the update stage. But even if the supports coincide exactly, many particles will over time decrease in weight so that after a few time steps, only a few lucky survivors have significant weights, while a large computational effort is spent on propagating unimportant particles. It has been shown that the variance of the weights can only increase over time, thus it is impossible to overcome the degeneracy problem (Kong et al., 1994). Two solutions exist to minimize this problem: (i) a good choice of the importance density and (ii) resampling.

– **Importance density**: The optimal importance density is given by:

$$\begin{aligned} q_{opt}(x_t|x_{0:t-1}, y_{1:t}) &= p(x_t|x_{0:t-1}, y_{1:t}) \\ &= \frac{p(y_t|x_t, x_{t-1}^{(i)}) p(x_t|x_{t-1}^{(i)})}{p(y_t|x_{t-1}^{(i)})} \end{aligned} \tag{26}$$

because it can be proven to minimize the variance of the importance weights (see Kong et al., 1994, and Chapter 6 of de Freitas, 1999). However, using the optimal importance density requires the ability to sample from $p(x_t|x_{t-1}^{(i)}, y_t)$ and to evaluate the integral over the new state $p(y_t|x_{t-1}^{(i)})$ (Arulampalam et al., 2002; Doucet et al., 2001; de Freitas, 1999). In many situations, this is impossible or very difficult, prompting the use of other importance densities. Perhaps the simplest and most common choice for the importance density is given by the prior:

$$q(x_t|x_{0:t-1}, y_{1:t}) = p(x_t|x_{t-1}) \tag{27}$$

which, although resulting in a higher variance of the Monte Carlo estimator, is usually easy to implement. Many other choices are possible (Arulampalam et al., 2002; Doucet et al., 2001; Liu, 2001).

– **Resampling**: Even the optimal importance density will lead to this "degeneracy" of the particles (few important

ones and many unimportant ones). One therefore introduces an additional selection or resampling step, in which particles with little weight are eliminated and new particles are sampled in the important regions of the posterior. De Freitas (1999) and Arulampalam et al. (2002) provide an overview of different resampling methods.

Resampling introduces its own problems. Since particles are sampled from discrete approximations to density functions, the particles with high weights are statistically selected many times. This leads to a loss of diversity among the particles as the resultant sample will contain many repeated points. This is known as "sample impoverishment" (Arulampalam et al., 2002) and is severe when the model forecast is very narrow or deterministic. The various methods that exist to deal with this problem will not be necessary here because of the broad and highly stochastic model forecast.

Because of the additional problems introduced by resampling, it makes sense to resample only when the variance of the weights has decreased appreciably. A suitable measure of degeneracy of an algorithm is the effective sample size N_{eff} introduced by Liu and Chen (1998) and defined by

$$N_{\text{eff}} = \frac{N}{1+\text{var}(w_t^{\star i})} \tag{28}$$

where $w_t^{\star i} = p(x_t^{(i)}|y_{1:t})/q(x_t^{(i)}|x_{t-1}^{(i)}, y_t)$ is referred to as the true weight. This may not be available, but an estimate $\hat{N}_{\text{eff}}$ can be obtained as the inverse of the so-called Participation Ratio (Mézard et al., 1987) (or Herfindahl index (Polakoff and Durkin, 1981; Lovett, 1988)):

$$\hat{N}_{\text{eff}} = \frac{1}{\sum_{i=1}^{N}(w_t^{(i)})^2} \tag{29}$$

Thus, resampling can be applied when $\hat{N}_{\text{eff}}$ falls below a certain threshold N_{thres}.

3.4 Numerical algorithms of the Sequential Importance Resampling (SIR) filter

In this section, we define the Sequential Importance Resampling (SIR) particle filter, which uses the prior given by Eq. (27) as the (sub-optimal) importance density and includes a resampling step to counteract the degeneracy of particles. The prior is obtained by random draw for individual particles using the forecast model, i.e. the renewal point process defined by Eq. (10). The presentation and the pseudo-codes in this section closely follow Arulampalam et al. (2002). More information on other particle filters can be found in Arulampalam et al. (2002), de Freitas (1999), Doucet et al. (2000, 2001), Liu (2001), and Cappé et al. (2005, 2007).

The SIR particle filter is characterized by choosing the prior $p(x_t|x_{t-1})$ as the importance density:

$$q(x_t|x_{0:t-1}, y_{1:t}) = p(x_t|x_{t-1}) \tag{30}$$

It can be shown (Arulampalam et al., 2002) that the SIR can be reduced to the pseudo-code given by Algorithm 1, where the weights are given by:

$$w_t^{(i)} \propto w_{t-1}^{(i)} p(y_t|x_t^{(i)}) \tag{31}$$

where $p(y_t|x_t^{(i)})$ is simply the likelihood and the weights are normalized by

$$\tilde{w}_t^{(i)} = \frac{w_t^{(i)}}{\sum_{j=1}^{N} w_t^{(j)}} \tag{32}$$

This filter, called the "bootstrap" filter by Doucet et al. (2001), is simple and easy to implement. If the likelihood has a much narrower support than the importance density, then the weights of many particles will be set to zero so that only few active particles are left to approximate the posterior. To counteract this particle death, a resampling step is included.

Algorithm 1 SIR particle filter.

$[\{x_t^{(i)}, w_t^{(i)}\}_{i=1}^N] = \text{SIR}[\{x_{t-1}^{(i)}, w_{t-1}^{(i)}\}_{i=1}^N, y_t]$

for i=1 to N **do**
  Draw $x_t^{(i)} \sim p(x_t|x_{t-1}^i)$
  Assign the particle a weight, $w_t^{(i)}$, according to Eq. (31)
end for
Calculate total weight: $W = \text{SUM}[\{w_t^{(i)}\}_{i=1}^N]$
for i=1 to N **do**
  Normalize: $w_t^{(i)} = W^{-1} w_t^{(i)}$
end for
Calculate $\hat{N}_{\text{eff}} = \frac{1}{\sum_{i=1}^{N}(w_t^{(i)})^2}$
if $\hat{N}_{\text{eff}} < N_{\text{thres}}$ **then**
  Resample using Algorithm 2:
  $[\{x_t^{(i)}, w_t^{(i)}\}_{i=1}^N] = \text{RESAMPLE}[\{x_t^{(i)}, w_t^{(i)}\}_{i=1}^N]$
end if

There are many methods to resample from the posterior (Doucet et al., 2001; de Freitas, 1999; Arulampalam et al., 2002). The basic idea is to eliminate particles that have small weights and to concentrate on particles with large weights. It involves generating a new set of particles and associated weights by resampling (with replacement) N times from an approximate discrete representation of the posterior. The resulting sample is an independently and identically distributed sample so that the weights are reset to $1/N$. The method of choice of Arulampalam et al. (2002) is systematic resampling

since "it is easy to implement, takes $O(N)$ time and minimizes the Monte Carlo variation." Its operation is described in Algorithm 2, where $U[a,b]$ is the uniform distribution on the interval $[a,b]$.

Algorithm 2 Systematic resampling.

$[\{x_t^{(j\star)}, w_t^{(j)}\}_{j=1}^N] = \text{RESAMPLE}[\{x_t^{(i)}, w_t^{(i)}\}_{i=1}^N]$

Initialize the CDF: $c_1 = 0$
for i=2 to N **do**
  Construct CDF: $c_i = c_{i-1} + w_t^{(i)}$
end for
Start at the bottom of the CDF: $i = 1$
Draw a starting point: $u_1 \sim U[0, N^{-1}]$
for j=1 to N **do**
  Move along the CDF: $u_j = u_1 + N^{-1}(j-1)$
  while $u_j > c_i$ **do**
    $i = i + 1$
  end while
  Assign sample: $x_t^{(j\star)} = x_t^{(i)}$
  Assign weight: $w_t^{(j)} = N^{-1}$
end for

3.5 Parameter estimation

Parameter estimation techniques within sequential Monte Carlo methods are discussed by, e.g., Doucet et al. (2001), Künsch (2001), Andrieu et al. (2004) and Cappé et al. (2005, 2007). The methods are either online-sequential or offline-batch methods. For simplicity, we will restrict this section to one particular technique, based on the offline or batch technique of maximizing (an MC estimate of) the complete marginal data likelihood defined in Eq. (9). The presentation follows Doucet et al. (2001).

We assume that the Markov transition kernel, defined by Eq. (2), depends on an unknown, static parameter vector θ. Moreover, we assume the marginal likelihood $L(\theta|y_{1:t}) = p_\theta(y_{1:t})$ admits a sequential formulation:

$$L(\theta|y_{1:t}) = p_\theta(y_{1:t}) = p_\theta(y_0)\prod_{k=1}^{t} p_\theta(y_k|y_{0:k-1}) \tag{33}$$

where the individual predictive likelihoods are defined as

$$p_\theta(y_k|y_{0:k-1}) = \int p_\theta(y_k, x_k|y_{0:k-1})dx_k \tag{34}$$

These can be estimated from the weighted particles $\{(x_k^{(i,\theta)}, w_k^{(i,\theta)})\}_{1\le i\le N}$ as

$$\begin{aligned} &p_\theta(y_k|y_{0:k-1}) \\ &= \int\int p_\theta(y_k|x_k)\, p_\theta(x_k|x_{k-1})\, p_\theta(x_{k-1}|y_{0:k-1})dx_{k-1}dx_k \end{aligned} \tag{35}$$

$$\approx \sum_{i=1}^{N} w_{k-1}^{(i,\theta)} \int p_\theta(y_k|x_k)\, p_\theta(x_k|x_{k-1}^{(i,\theta)})dx_k \tag{36}$$

$$\approx \sum_{i=1}^{N} w_k^{(i,\theta)} \tag{37}$$

where $w_k^{(i,\theta)}$ are the unnormalized weights at the k^{th} time step. Expression (25) is used to go from the second to the third approximate equality.

The log-likelihood $\ell(\theta)$ is therefore given by

$$\begin{aligned} \ell(\theta) &= \log(L(\theta|y_{1:t})) = \log\left[\prod_{k=1}^{t} p_\theta(y_k|y_{0:k-1})\right] \\ &= \sum_{k=1}^{t} \log\left[p_\theta(y_k|y_{0:k-1})\right] \\ &\approx \sum_{k=1}^{t} \log\left[\sum_{i=1}^{N} w_k^{(i,\theta)}\right] \end{aligned} \tag{38}$$

Maximizing the sum of the unnormalized weights given by expression (38) with respect to the parameter set θ results in the maximum likelihood estimator $\hat{\theta}$:

$$\hat{\theta} = \arg\max\left[\sum_{k=1}^{t} \log\left(\sum_{i=1}^{N} w_k^{(i,\theta)}\right)\right] \tag{39}$$

Doucet et al. (2001), Andrieu et al. (2004), Cappé et al. (2005, 2007) and Olsson and Rydén (2008) consider the estimator's statistical properties. To find the maximum of the log-likelihood in Eq. (39), one may use the standard optimization algorithms, such as gradient-based approaches, the expectation-maximization algorithm, or random search algorithms such as simulated annealing, genetic algorithms, etc. (see, e.g., Sambridge and Mosegaard, 2002). In our parameter estimation experiments (see Sect. 4.3), we chose a combination of a coarse direct grid-search method and a pattern search method to refine the coarse estimate (Hooke and Jeeves, 1961; Torczon, 1997; Lewis and Torczon, 1999).

3.6 Kalman filters

To provide benchmarks for the SIR particle filter, we use two Kalman filters. The first is a very simple, deterministic Kalman filter (DKF) based on the approximation that all distributions are Gaussian (Kalman, 1960). The second is the Ensemble Square Root Filter (EnSRF) proposed by Tippett et al. (2003), a popular instance of the ensemble Kalman filter (Evensen, 1994). The EnSRF approximates priors and posteriors with an ensemble of unweighted particles and assumes the measurement errors are Gaussian. This section defines the filters and derives the relevant equations that we implemented numerically.

3.6.1 Deterministic Kalman Filter (DKF)

The forecast of the deterministic Kalman filter for time step t is given by a Gaussian distribution with mean $\langle x_t^f \rangle$ and variance P_t^f which are determined by

$$\begin{aligned} \langle x_t^f \rangle &= \langle x_{t-1}^a \rangle + \langle dx^f \rangle \\ P_t^f &= P_{t-1}^a + Q^f \end{aligned} \tag{40}$$

where $\langle x_{t-1}^a \rangle$ and P_{t-1}^a are the mean and variance, respectively, of the (Gaussian) posterior from the previous time step, and $\langle dx^f \rangle$ and Q^f are the mean and variance, respectively, of the forecast model, which in our case is the renewal process defined in Eq. (10).

The analysis of the DKF is also given by a Gaussian, with mean $\langle x_t^a \rangle$ and variance P_t^a determined by

$$\begin{aligned} \langle x_t^a \rangle &= \langle x_t^f \rangle + K_t(y_t - \langle x_t^f \rangle) \\ P_t^a &= (1 - K_t) P_t^f \end{aligned} \tag{41}$$

where $y_t - \langle x_t^f \rangle$ is often called the innovation or measurement residual and the Kalman gain K_t is determined by

$$K_t = \frac{P_t^f}{P_t^f + R^{\mathrm{o}}} \tag{42}$$

where R^{o} is the variance of the observation error distribution.

As for the particle filter, we will use the marginal complete likelihood function (9), i.e. the denominator of Bayes' theorem, to estimate the parameters of the forecast model. The likelihood is given by

$$\begin{aligned} p_\theta(y_t|y_{0:t-1}) &= \int p(y_t|x_t)\, p(x_t|y_{0:t-1}) dx_t \\ &= \int \mathcal{N}(y_t, R^{\mathrm{o}}) \mathcal{N}(\langle x_t^f \rangle, P_t^f) dx_t \end{aligned} \tag{43}$$

where $\mathcal{N}(a,b)$ denotes the normal distribution with mean a and variance b. Using expressions (40) and (41), Eq. (43) reduces to

$$\begin{aligned} p_\theta(y_t|y_{0:t-1}) \;=\; & \frac{1}{\sqrt{2\pi}\sqrt{P_{t-1}^a + Q^f + R^{\mathrm{o}}}} \times \\ & \exp\left[-\frac{\left(\langle x_{t-1}^a \rangle + \langle dx^f \rangle - y_t\right)^2}{2\left(P_{t-1}^a + Q^f + R^{\mathrm{o}}\right)} \right] \end{aligned} \tag{44}$$

where $\langle dx^f \rangle$ and Q^f are explicit functions of any parameters θ of the renewal process. As above, we sum over the logarithms of each individual likelihood and maximize the joint log-likelihood to estimate parameters.

The DKF only requires two parameters to be tracked, the forecast mean and variance, making it a very simple and cheap filter. However, all distributions are assumed to be Gaussian.

3.6.2 Ensemble Square Root Filter (EnSRF)

Tippett et al. (2003) discussed the Ensemble Square Root Filter (EnSRF), a particular instance of the Ensemble Kalman Filter (EnKF) invented by Evensen (1994). The EnKF uses ensemble representations for the forecast and analysis error covariances. Starting with an unweighted ensemble $\{x_{t-1}^{a,(i)}\}_{i=1}^m$ of m members that represent the analysis of the previous time step, the (potentially) non-linear and non-Gaussian dynamics of the model $p(x_t|x_{t-1}^{(i)})$ is applied to each member to produce the forecast ensemble $\{x_t^{f,(i)}\}_{i=1}^m$. The ensemble representation of the forecast produces any required statistics such as mean $\langle x_t^f \rangle = 1/m \sum_i^m x_t^{f,(i)}$, covariance P_t^f or the full pdf of the forecast can be obtained from a kernel density estimate. The forecast ensemble $x_k^{f,(i)}$ is thus obtained from

$$x_t^{f,(i)} = x_{t-1}^{a,(i)} + dx^{f,(i)} \tag{45}$$

$$X_t^{f,(i)} = x_t^{f,(i)} - \langle x_t^f \rangle \tag{46}$$

$$P_t^f = \frac{1}{m-1} \sum_i^m \left(X_t^{f,(i)} \right)^2 \tag{47}$$

Once an observation is available, the mean $\langle x_t^a \rangle$ of the analysis is obtained from

$$\langle x_t^a \rangle = \langle x_t^f \rangle + K_t(y_t - \langle x_t^f \rangle) \tag{48}$$

where the Kalman gain K_t is obtained as in the classical Kalman filter from

$$K_t = \frac{P_t^f}{P_t^f + R^{\mathrm{o}}} \tag{49}$$

where R^{o} is the covariance of the observation error distribution.

To obtain the full pdf of the analysis rather than just the mean, and in the case of observations being assimilated one by one serially, Tippett et al. (2003) show (their Eq. 10) that the perturbations $X_t^{a,(i)}$ of the analysis ensemble about the analysis mean $\langle x_t^a \rangle$ are given by

$$X_t^{a,(i)} = X_t^{f,(i)} (1 - \beta_t P_t^f) \tag{50}$$

where $\beta_t = \left(D_t + \sqrt{R^{\mathrm{o}} D_t} \right)^{-1}$, and $D_t = P_t^f + R^{\mathrm{o}}$ is the innovation (co-)variance.

As before, we also derive the expression for the complete marginal joint log-likelihood. The observational error distribution is characterized solely by the covariance R^{o}, i.e. the EnSRF implicitly assumes Gaussian measurement errors. In contrast to the Gaussian model forecast of the DKF, the EnSRF approximates the model forecast with the m member ensemble or Monte Carlo representation $p_t^f(x_t|y_{0:t-1}) \approx$

$\hat{p}_m(x_t) = (1/m)^{-1} \sum_i^m \delta_{x_t^{(i)}}(x_t)$. The likelihood is given by

$$p_\theta(y_t|y_{0:t-1}) = \int p(y_t|x_t)\, p(x_t|y_{0:t-1}) dx_t$$
$$= \int \mathcal{N}(y_t, R^o)\, p_t^f(x_t|y_{0:t-1}) dx_t \quad (51)$$

To evaluate Eq. (51), we use Monte Carlo integration (expression 12) to obtain:

$$p_\theta(y_t|y_{0:t-1}) \approx \int \mathcal{N}(y_t, R^o)\, \hat{p}_m(x_t) dx_t$$
$$\approx \frac{1}{m} \sum_i^m p(y_t - x_t^{(i)}) \quad (52)$$

where $p(y_t - x_t^{(i)})$ is the normal distribution $N(y_t, R)$ evaluated at $x_t^{(i)}$. As before, we maximize the marginal joint log-likelihood, i.e. the sum over the logarithms of the individual likelihood functions, to estimate the parameter set θ.

The EnSRF is thus a Monte Carlo method that allows for non-Gaussian and non-linear model dynamics to produce arbitrary forecast pdfs. However, unlike the SIR, it is only concerned with the variance of the observational error distribution during the analysis.

4 Numerical experiments and results

In this section, we present a simple, pedagogical example of earthquake forecasting based on data assimilation. Our model is the one-dimensional, temporal lognormal renewal process (Sect. 4.1.1): the simplest point process, which nevertheless draws much interest in earthquake seismology and seismic hazard, as mentioned above. We assume the process is observed in noise, i.e., the unobservable true occurrence times are perturbed by (additive) identically and independently distributed noise (Sect. 4.1.2). The aim of this section is to show an example of how data assimilation provides better forecasts, as measured by the likelihood gain, than a forecast ("the benchmark") which ignores the data errors (assumes the observed times are the true times). We will compare the performance of the SIR particle filter with that of the deterministic and ensemble Kalman filters and measure their skills against the benchmark (Sect. 4.2). Finally, in Sect. 4.3 we will use maximum likelihood estimation to obtain parameter estimates using both the filters and the benchmark. The results in this section thereby demonstrate that data assimilation can help make earthquake forecasting and forecast validation robust with respect to observational data errors.

4.1 Experiment design

4.1.1 The forecast model: lognormal renewal process

Motivated by its relevance to paleoseismology, seismic hazard and the characteristic earthquake debate, we use a lognormal renewal process as our forecast model. The lognormal renewal process has a long tradition in modeling the recurrences of earthquakes (see, e.g., Nishenko and Buland, 1987; Ogata, 1999; Biasi et al., 2002; Sykes and Menke, 2006; Field, 2007b). According to the lognormal process, the intervals τ between subsequent earthquakes are distributed according to:

$$f_{\text{logn}}(\tau; \mu, \sigma) = \frac{1}{\tau\sqrt{2\pi}\sigma} \exp(-(\log\tau - \mu)^2/2\sigma^2) \quad (53)$$

where the parameters μ and σ may need to be estimated. In the notation of Sect. 3, using a physically meaningful t_k for the state variable instead of x_k, the lognormal distribution of the intervals is the transition kernel defined in Eq. (2):

$$p(t_k|t_{k-1}; \mu, \sigma) = \frac{1}{(t_k - t_{k-1})\sqrt{2\pi}\sigma} \times$$
$$\exp\left(-(\log(t_k - t_{k-1}) - \mu)^2/2\sigma^2\right) \quad (54)$$

To mimic a realistic process, we use parameters taken from the study by Biasi et al. (2002), who fit the lognormal process to a paleoseismic data set from the San Andreas fault in California:

$$\mu = -0.245 \quad \text{and} \quad \sigma = 0.7 \quad (55)$$

where we obtained μ by normalizing the average recurrence interval to one, without loss of generality. Figure 1 shows the lognormal distribution (solid black curve) with these parameter values.

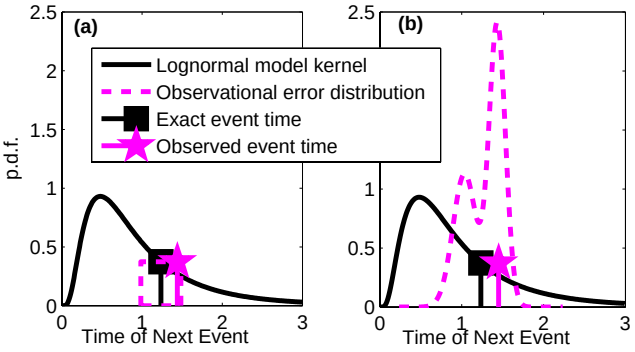

Fig. 1. Visual comparison of the model transition kernel of the (unobservable) true occurrence times (solid black curves) and the conditional likelihood functions of the noisy observed occurrence time given the true occurrence time (dashed magenta line). **(a)** Uniform error distribution. **(b)** Gaussian mixture error distribution. Also shown are a sample true occurrence time (black square) and a sample observation (magenta star).

4.1.2 The observations: noisy occurrence times

We suppose that the k-th observed occurrence time t_k^o is a noisy perturbation of the "true" occurrence time t_k^t:

$$t_k^o = t_k^t + \epsilon_k \quad (56)$$

where ϵ is an additive noise term distributed according to some distribution $p_\epsilon(\epsilon)$. For our numerical experiments below, we choose two different distributions: a uniform distribution and a Gaussian mixture model. The uniform distribution was chosen to mimic measurement errors that are poorly constrained, so that only a fixed interval is provided without knowledge of the distribution (e.g. Ogata, 1999). The Gaussian mixture model (GMM), on the other hand, is an illustrative example of better-constrained uncertainties that give rise to more complex distributions with bi- or multi-modal structures (e.g., Biasi et al., 2002).

The uniform distribution is given by:

$$p_{\text{uni}}(\epsilon) = \frac{1}{\Delta} H(\epsilon + \frac{\Delta}{2}) H(\frac{\Delta}{2} - \epsilon) = \begin{cases} \frac{1}{\Delta} & -\frac{\Delta}{2} \le \epsilon \le +\frac{\Delta}{2} \\ 0 & \text{otherwise} \end{cases} \quad (57)$$

where $H(\cdot)$ is the Heaviside step function. Substituting $\epsilon = t^{\text{o}} - t^t$ gives the density (conditional likelihood) of the data given the true occurrence time, defined by Eq. (3):

$$p^{\text{o}}_{\text{uni}}(\epsilon_k) = p(t^{\text{o}}_k | t^t_k) = p(t^{\text{o}}_k - t^t_k) = \begin{cases} \frac{1}{\Delta} & t^t_k - \frac{\Delta}{2} \le t^{\text{o}}_k \le t^t_k + \frac{\Delta}{2} \\ 0 & \text{otherwise} \end{cases} \quad (58)$$

We set the parameter to

$$\Delta = 0.5 \quad (59)$$

so that the uncertainty in the measurement is roughly half of the expected reoccurrence time, mimicking paleoseismic data sets (Ogata, 1999, 2002). In Fig. 1a), we show the lognormal model kernel and the uniform error distribution with our choices of parameters.

The Gaussian mixture model $p_{\text{GM}}(\epsilon)$, on the other hand, consists for our purposes of two one-dimensional, weighted and uncorrelated Gaussian distributions

$$p_{\text{GM}}(\epsilon) = p_1 \mathcal{N}(\eta_1, \rho_1) + p_2 \mathcal{N}(\eta_2, \rho_2) \quad (60)$$

where the weights $p_1 = 0.4$ and $p_2 = 0.6$ sum to one and the normal distributions $\mathcal{N}(\cdot,\cdot)$ are each characterized by their averages $\eta_1 = -0.2$ and $\eta_1 = +0.2$ and their standard deviations $\sqrt{\rho_1} = 0.02$ and $\sqrt{\rho_2} = 0.01$. These values were chosen to provide a simple bi-modal distribution that mimics certain well-constrained uncertainties on earthquake occurrences in paleoseismic datasets. In Fig. 1b), we compare the lognormal model kernel with the GMM error distribution with the present parameter values.

4.1.3 Initial condition and observation period

We assume for simplicity that the period $T = [a, b]$ over which the point process is observed begins with an event at $t_0 = 0 = a$. We further assume that the true and observed occurrence times of this first event coincide, so that our initial condition $p(x_0)$ is a delta function $p(x_0) = \delta(t_0 = 0)$, and that the observation period ends with the last observed event $t^{\text{o}}_n = b$. This assumption can be relaxed: Ogata (1999) provided the relevant equations.

4.1.4 Simulation procedure

In this entirely simulated example, we begin by generating the "true" (unobservable) process. We generate n random samples from the lognormal distribution given by Eq. (54) to obtain the sequence of true event times $\{t^t_k\}_{0 \le k \le n}$. Next, we simulate the observed process by generating n random samples from either the uniform or the Gaussian mixture conditional likelihood given by Eqs. (58) and (60) to obtain the sequence of observed event times $\{t^{\text{o}}_k\}_{0 \le k \le n}$.

To perform the particle filtering, we initialize $N = 10\,000$ particles at the exactly known $t_0 = 0$. To forecast t_1, we propagate each particle through the model kernel (54). Given the observation t^{o}_1 and the model forecast, we use the SIR particle filter described in Sect. 3.4 to obtain the analysis of t_1. The approximation of the posterior is then used to forecast t_2 according to Eq. (7). This cycle is repeated until the posteriors of all n occurrence times are computed.

The Kalman filters are implemented similarly. Like the SIR, we initialized the EnSRF with $m = 10\,000$ ensemble members to adequately represent the forecast and analysis distributions. The DKF requires only the mean $\langle dx^f \rangle$ and variance Q^f of the forecast model given by the lognormal distribution (53):

$$\langle dx^f \rangle = \exp\left(\mu + \frac{1}{2}\sigma^2\right) \quad (61)$$

$$Q^f = \left(\exp(\sigma^2) - 1\right) \exp\left[2\left(\mu + \frac{1}{2}\sigma^2\right)\right] \quad (62)$$

Both filters require the variance of the observational error distribution to assimilate observations. The variance of the uniform error distribution is given by

$$R^{\text{o}} = \Delta^2 / 12 \quad (63)$$

We computed the variance of the Gaussian mixture distributed errors empirically from numerous samples.

4.2 Data assimilation

This section presents examples of the forecast and posterior distributions using a large number of particles ($N = 10\,000$). We compare the SIR particle filter, defined in Sect. 3.4, with the Kalman filters and the benchmark, which entirely neglects the presence of data uncertainties. We assume in this section that the parameters are known.

4.2.1 Forecast and analysis (priors and posteriors)

In Fig. 2, we present the forecast and analysis distributions of the filters and of the benchmark for a particular example event (number 11). Panels (a) and (b) show the forecasts in

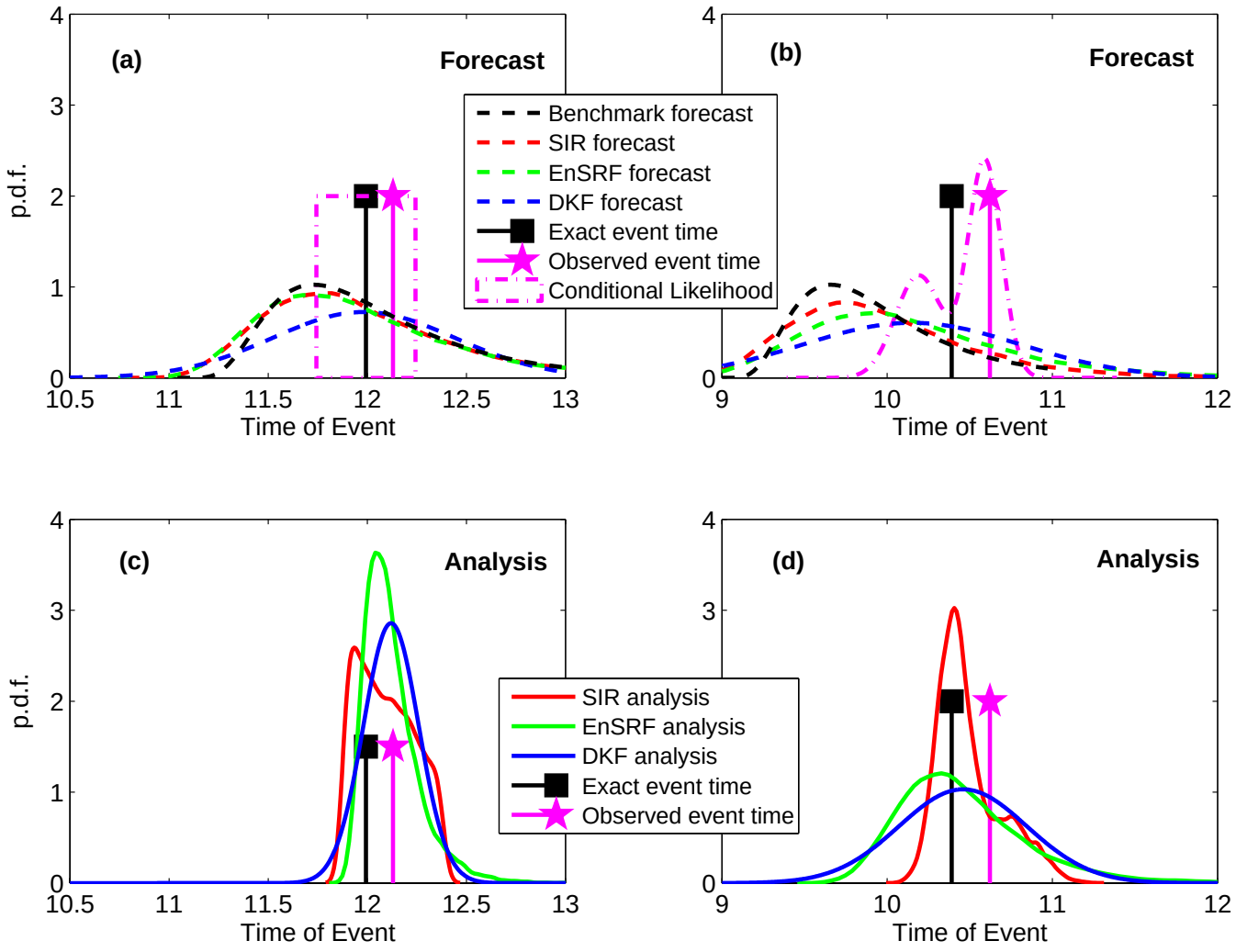

Fig. 2. Illustration of the data assimilation cycle for event 11. Forecasts of the SIR, EnSRF and DKF filters, and the benchmark, which ignores measurement errors, are shown in **(a)** and **(b)** for the case of uniformly and Gaussian mixture distributed measurement errors, respectively. Analyses (posteriors) are shown in **(c)** and **(d)** on the same scale.

the case of uniformly and Gaussian mixture distributed measurement errors, respectively, while (c) and (d) show the corresponding posteriors. Concentrating first on the forecasts, the benchmark assumes that the observed events correspond to the previous "true" events without data errors. Therefore, the forecast distribution of the next occurrence time is simply the lognormal distribution. In contrast, the filter forecasts are broadened because of the uncertainty in the last occurrence time. As a result, the benchmark forecast is artificially sharper than the filter forecasts. In some cases, the sharper peak may lead to higher likelihoods – but the benchmark will pay a price when the observed event is in the tails of its overly optimistic forecast. In those cases, the broader filter forecasts will more than recuperate. Section 4.2.2 compares the likelihood scores and gains of the benchmark and the particle filters.

The SIR forecast is broader than the benchmark forecast because of the uncertain last occurrence time, but the lognormal shape can still be identified. The EnSRF forecast is almost identical to the SIR forecast for the case of uniform errors, while differences are clearly visible in the case of Gaussian mixture distributed noise. Recall that the EnSRF assumes that measurement noise is Gaussian distributed. Because the lognormal model kernel is so strongly stochastic, this approximation to the uniform noise seems acceptable, while the differences are more apparent if the measurement error is bi-modal, as in panel (b). The DKF forecast is Gaussian by construction, presenting a poor approximation to the asymmetric lognormal model kernel. Moreover, irrespective of the measurement errors, the forecast displays little skill compared to the SIR and EnSRF forecasts, which are more strongly peaked.

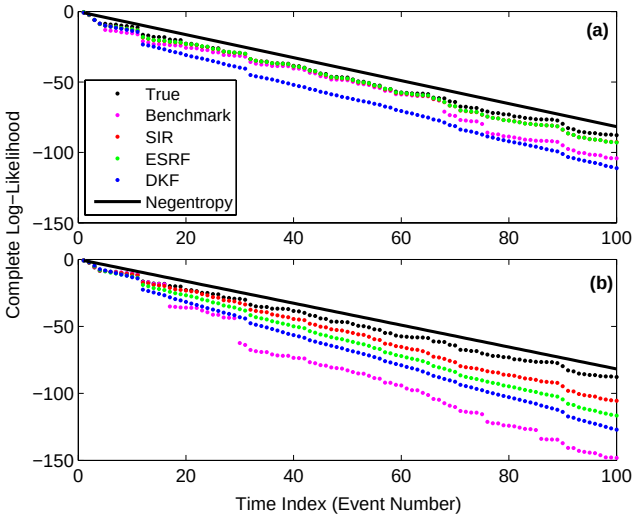

Fig. 3. Evolution of the cumulative complete marginal log-likelihood of the particle and Kalman filters and the benchmark for the case of **(a)** uniformly distributed and **(b)** Gaussian mixture distributed observational errors. Also shown are the log-likelihoods of the "true" (unobservable) times using the "true" event times (black) and the average log-likelihood given by the negative entropy of the lognormal distribution (solid black line). In **(a)**, the scores of the SIR and EnSRF are indistinguishable.

After applying Bayes' theorem, the resulting posteriors are shown in panels (c) and (d). While the benchmark simply assumes that the observed event time is "true", increasingly better approximations to the actual posterior are obtained by the DKF, EnSRF and SIR posteriors. While the DKF presents a simple Gaussian approximation, the EnSRF still displays the asymmetry of its forecast. Only the SIR does not assume the conditional likelihood to be Gaussian and can therefore recover the posterior more accurately than the other filters. This is particularly visible in the case of Gaussian mixture noise.

Having obtained the best possible estimate of the true occurrence time using this Bayesian approach, the data assimilation cycle is closed by using the posteriors as initial conditions for the forecast of the next event.

4.2.2 Comparison: likelihood scores and gains

To measure the improvement of the "earthquake" forecasts based on data assimilation over the naive benchmark, which ignores data uncertainties, we use the log-likelihood score and gain. Both are common measures of earthquake forecasts based on point processes (Daley and Vere-Jones, 2004; Harte and Vere-Jones, 2005). However, we extend the measures by taking into account uncertainties (see also Ogata, 1999), as suggested by Doucet et al. (2001), Andrieu et al. (2004) and Cappé et al. (2005). In particular, we employ the marginal log-likelihood of the data, defined by Eq. (9), which reflects both the model forecast and the conditional likelihood function (the measurement process). This marginal likelihood is nothing but the denominator in Bayes' theorem (4), which judges how well the data are explained, assuming both a model forecast and a measurement process.

For the SIR particle filter, the marginal log-likelihood of the data can be approximated by Eq. (38). The marginal log-likelihoods of the DKF and EnSRF are determined by Eqs. (44) and (52), respectively. The benchmark effectively assumes that the measurement process is perfect, such that any conditional likelihood is replaced by a Dirac function $p(t_k^o|t_k^t) = \delta(t_k^o - t_k^t)$. The benchmark log-likelihood score is thus simply obtained by using the lognormal density function and plugging in the observed occurrence times. Since this is a stochastic prediction problem, it is also of interest to compare these forecasts to the ideal case of having access to the "true" occurrence times. For the "true" process, the log-likelihood score is obtained by using the lognormal distribution and plugging in the "true" event times, again replacing the conditional likelihoods by Dirac functions. Since this will only give the score for one particular realization, we also calculate the average log-likelihood score per event, given by the negative entropy of the lognormal distribution, which is available analytically.

In Fig. 3, we show the evolution of the cumulative (complete) marginal log-likelihood of the data using the particle and Kalman filters and the benchmark for a simulation of a 100-event point process. The cases of uniformly and Gaussian mixture distributed measurement noise are shown in panels (a) and (b), respectively. The DKF does not perform well in this experiment. Its Gaussian forecast is too far from the actual distribution, resulting in overly broad and unskillful forecasts. For these particular parameters, the DKF provides no better scores than the benchmark. Not surprisingly, the SIR obtains higher scores than the Gaussian DKF and the benchmark. However, the EnSRF can compete with the SIR in the case of uniformly distributed measurement errors: the likelihood scores are nearly identical. As already mentioned above, the Gaussian approximation to the uniform distribution appears sufficient so as not to degrade the scores. This is no longer true for bi-modally distributed noise (panel b). Here, the SIR displays the full power of the particle-based approximation to arbitrary distributions and surpasses the EnSRF.

To investigate the average performance improvement, we simulated 100 realizations of a 100-event point process. We calculated the mean of the log-likelihood scores at each event index, as shown in Fig. 4. Fluctuations are now mostly smoothed out. The mean "true" likelihood scores now match the negative entropy predictions exactly. When the observational errors are uniformly distributed (panel a), the SIR and EnSRF have essentially identical mean likelihood scores. In contrast, the SIR's scores are much higher when the errors are Gaussian mixture distributed (panel b). The DKF performs worse than the benchmark because its Gaussian approximation of the forecast is worse than the benchmark's

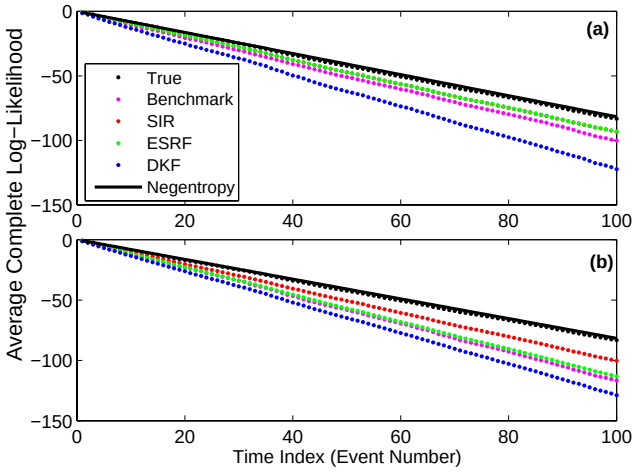

Fig. 4. Evolution of the sample mean of the cumulative complete log-likelihood, averaged over 100 realizations of a 100-event point process. Same explanation as for Fig. 3.

lognormal forecast that neglects data errors. However, in rare cases, the benchmark obtains scores of negative infinity, i.e. certain event times are impossible according to the benchmark. Having excluded these values from the calculations, we need to interpret the benchmark's likelihood scores in Fig. 4 as conditional on "survival".

To measure the quality of a point process forecast with respect to a reference forecast, we employ several common measures. The individual probability gain $G_k^{(1)}$ measures the ratio of the likelihood $p_1(t_k^o)$ of the k-th observed event under a specific forecast over the likelihood of the same event under a reference forecast $p_0(t_k^o)$:

$$G_k^{(1)} = \frac{p_1(t_k^o)}{p_0(t_k^o)} \tag{64}$$

The individual probability gain $G_k^{(1)}$ measures how much better the event is explained by a particle or EnSRF filter forecast over the naive benchmark forecast (for the remainder of this section, we do not consider the poorly performing DKF). $G_k^{(1)} = 1$ corresponds to no improvement. Since usually log-likelihood scores are used rather than likelihood values, it is common to use the (individual) log-likelihood ratio, defined by:

$$\begin{aligned} LR_k^{(1)} &= \log G_k^{(1)} = \log\left(\frac{p_1(t_k^o)}{p_0(t_k^o)}\right) \\ &= \log p_1(t_k^o) - \log p_0(t_k^o) = LL_1(t_k^o) - LL_0(t_k^o) \end{aligned} \tag{65}$$

where $LL(t_k^o)$ is the marginal log-likelihood of event t_k^o and $LR_k^{(1)} = 0$ corresponds to no improvement.

The (cumulative) probability gain $G^{(n)}$ per earthquake of the proposed forecast with respect to a reference forecast is defined as (Daley and Vere-Jones, 2004; Harte and Vere-Jones, 2005):

$$G^{(n)} = \exp\left(\frac{LL_1(n) - LL_0(n)}{n}\right) \tag{66}$$

where $LL_1(n)$ and $LL_0(n)$ are the cumulative marginal log-likelihood scores of the proposed model and a reference model, respectively, for the n considered events. This measure quantifies the cumulative improvement due to the proposed forecast over a reference forecast. The measure is motivated by its expression as the geometric average of the individual conditionally independent probability gains:

$$\begin{aligned} G^{(n)} &= \left[\prod_{k=1}^{n} G_k^{(1)}\right]^{\frac{1}{n}} = \left[\prod_{k=1}^{n} \frac{p_1(t_k^o)}{p_0(t_k^o)}\right]^{\frac{1}{n}} \\ &= \left[\frac{\prod_{k=1}^{n} p_1(t_k^o)}{\prod_{k=1}^{n} p_0(t_k^o)}\right]^{\frac{1}{n}} \end{aligned} \tag{67}$$

where the product over all $k = 1, \ldots, n$ events specifies the joint probability density of the entire process under a specific model. In our experiments, the benchmark is the reference forecast, i.e. we directly measure any improvement of the SIR and EnSRF over the benchmark.

For a 100-event point-process simulation, we calculated the individual probability gains $G_k^{(1)}$ for each event t_k^o for the SIR and EnSRF, as shown in Fig. 5. The individual gains $G_k^{(1)}$ fluctuate wildly, from about 0.5 to 10^6 (to display the variability near $G_k^{(1)} = 1$, we set an upper limit to the ordinate axes). There are many events that are better forecast by the benchmark than by the filters ($G_k^{(1)} < 1$), but there are some events for which the filters outperform the benchmark by several orders of magnitude. For this particular simulation, the average probability gains $G^{(100)}$ per earthquake of the SIR filter and EnSRF filters were 1.26 and 1.25, respectively, in the case of uniform noise. The gains were 1.23 and 1.12 in the case of Gaussian mixture distributed errors.

The seemingly surprising occurrence of $G_k^{(1)} < 1$ forecasts can be explained by the fact that the benchmark forecasts are sharper than the particle filter forecasts, since the benchmark does not take into account the uncertainty in the last occurrence time (compare the forecasts in Fig. 2). As a result, if the next observed event actually falls near the peak of the benchmark forecast, the likelihood of the data is higher under the benchmark forecast than under the broadened filter forecasts. Thus, frequently, the benchmark produces higher likelihood scores than the filters. However, precisely because the benchmark forecast does not take into account data errors, the forecasts are overly optimistic. When the observed events fall outside of this artificially narrow window, the filters performs better than the benchmark, and sometimes immensely better. Such surprises for the benchmark are reflected in the very large individual likelihood gains of up to 10^6 and in outright "death", i.e., scores of negative infinity (not shown).

To illuminate the performance of the SIR and EnSRF against the benchmark further, we simulated a 10 000-event

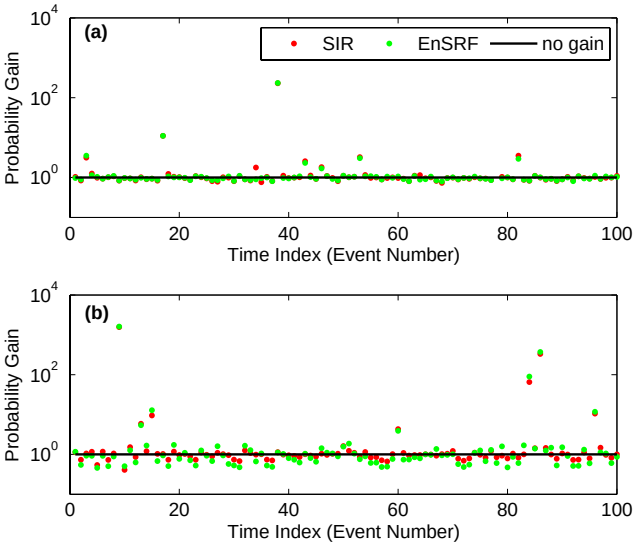

Fig. 5. Probability gains of the SIR particle filter and the EnSRF over the benchmark for each individual earthquake for the case of **(a)** uniformly distributed and **(b)** Gaussian mixture distributed observational errors. In **(a)**, the gains of the SIR and EnSRF are often nearly identical.

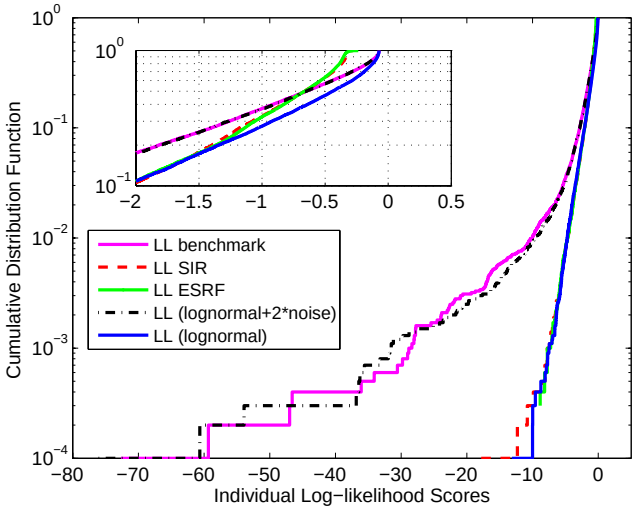

Fig. 6. Empirical cumulative distribution functions of the log-likelihood scores of each event obtained by the SIR particle filter, the EnSRF, the benchmark and the "true" (lognormal) process in a 10 000-event point-process simulation. Also shown is a distribution explained in the text. Inset: magnification of the same data.

point-process using uniformly distributed errors and calculated the log-likelihood scores and ratios $LR_k^{(1)}$ against the benchmark for each event. Calculations with Gaussian mixture distributed noise give the same qualitative results. The empirical cumulative distribution functions of the log-likelihood scores are shown in Fig. 6. For comparison, we also show the distribution of log-likelihood scores obtained by the using the "true" process (lognormal), and by another distribution, explained below. The log-likelihood distribution of the "true" process has consistently the highest scores, up to statistical fluctuations, as expected. The log-likelihood scores of the SIR and EnSRF, however, are not consistently better than the benchmark (as already seen in Fig. 5). Rather, the highest scores of the benchmark are higher than those of the filters (see the inset of Fig. 6). These values correspond to those events that occur near the peak of the overly optimistic and sharp forecast of the benchmark, thus resulting in a higher score compared with the broadened filter forecasts. However, the scores of the benchmark quickly become worse than the filters', and indeed the lowest scores are orders of magnitude smaller. The body and tail of the distributions show the filters' advantage: the benchmark sometimes produces terrible forecasts, for which it pays with a poor score. At the same time, the individual filters' scores remain relatively close to the scores of the "true" process.

We found it helpful to include another distribution of log-likelihood scores in Fig. 6, labeled LL(lognormal $+2^*$noise). To produce it, we simulated lognormally distributed samples and then perturbed each sample twice by an additive, uniformly distributed error with the same distribution as the observational uncertainty that perturbs the lognormal point process. We calculated their likelihood scores using the original lognormal function. The point is to show that the log-likelihood scores of the benchmark naturally come from the fact that we assume a lognormal function in the calculation of the likelihood scores, but that the random variables we observe are not actually lognormally distributed. In fact, the benchmark makes two mistakes: (i) the origin point (start of the interval) is a perturbed version of the last true occurrence time, and (ii) the observed next event is again a perturbed version of the next true occurrence time. The simulated LL(lognormal +2*noise) thus corresponds exactly to the log-likelihood distribution of the benchmark (up to statistical fluctuations).

Figure 7 displays the kernel density estimate of the individual log-likelihood ratios $LR_k^{(1)}$, a direct comparison of the SIR and the benchmark for each event. The vertical black line at $LR^{(1)} = 0$ separates the region in which the benchmark performs better ($LR_k^{(1)} < 0$) from the one in which the particle filter performs better ($LR_k^{(1)} > 0$). The statistics of this distribution are particularly illuminating: the median is $LR^{(1)} = -0.02$, and 55% of the time, the benchmark outperforms the particle filter. However, the amount by which the benchmark outperforms the SIR particle filter is never very much, since the SIR forecast is never much broader than the benchmark forecast. Thus the potential loss of the SIR particle filter is limited, as seen by the truncation of the distribution for low log-likelihood ratios. At the same time, the tail of the distribution towards large log-likelihood ratios decays much more slowly. Inset 1 of Fig. 7 shows the survivor function in semi-logarithmic axes to emphasize the

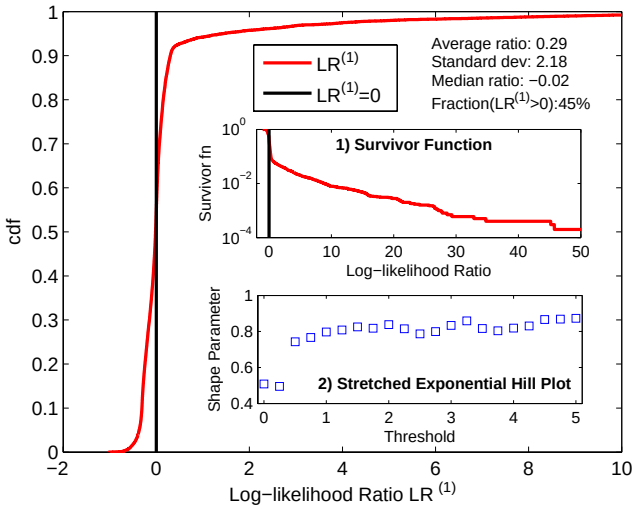

Fig. 7. Cumulative distribution function (cdf) of the individual log-likelihood ratios $LR_k^{(1)}$ between log-likelihood scores of the SIR particle filter and the benchmark. Inset 1: Survivor function. Inset 2: Hill plot of the maximum likelihood estimates of the shape parameter of a stretched exponential distribution as a function of the threshold above which the parameter is estimated.

slow decay. We found that a slightly stretched exponential distribution (Laherrère and Sornette, 1998; Sornette, 2004) fits the tail adequately, with a shape parameter (exponent) of about 0.8 ± 0.3 (see inset 2 of Fig. 7). As a result of the stretched exponential tail of the log-likelihood ratio, the potential benefit of the SIR particle filter can be enormous, while its potential disadvantage is limited. As a result, the average log-likelihood ratio is $\langle LR^{(1)} \rangle = 0.29$, despite the negative median.

4.3 Parameter estimation

So far, we have assumed that the parameters of the lognormal distribution are known. In reality, one would like to estimate the parameters from the observed occurrence times. As stated in Sect. 3.5, in this article we perform offline maximum likelihood estimation of a batch of data at a time. In particular, we maximize the complete marginal data likelihood, approximated by Eq. (38), to estimate parameters. To find the maximum, we first perform a coarse grid-search over the parameter space and then use a pattern search algorithm (Hooke and Jeeves, 1961; Torczon, 1997; Lewis and Torczon, 1999). In this section, we first describe the estimation and compare the parameter estimates of the particle and Kalman filters with those of the benchmark for single simulations of a 200-event and a 10-event point process assuming Gaussian mixture distributed errors (uniform errors give qualitatively similar results). We then show results for a large number of 100-event point process simulations to test for statistical bias in the estimates for both uniform and Gaussian mixture distributed errors.

In Fig. 8, we show approximate contour levels of the log-likelihood as a function of the two parameters μ and σ for a single 200-event point process simulation assuming Gaussian mixture distributed noise. For reference, we include the "true" constants used for the simulation and the maximum likelihood estimates based on the "true" occurrence times along with 95% confidence bounds. The likelihood contours, all plotted on the same scale in Fig. 8, reveal several interesting features. The SIR achieves the highest likelihood values, and its maximum is well constrained. The likelihood function of the EnSRF does not attain the scores of the SIR, and its structure reveals stronger correlations between the parameters and a flatter maximum. Nonetheless, its maximum likelihood parameter estimates are close to the "true" ones. The likelihood function of the DKF highlights stronger correlations between μ and σ and substantially smaller likelihood scores than the SIR or EnSRF. Moreover, its parameter estimates indicate a slight bias. Finally, the benchmark's likelihood function demonstrates a clear bias towards larger σ, which provides the benchmark with better insurance against unexpected occurrence times. Interestingly, the biased parameter estimates allow the benchmark to achieve likelihood scores higher than those of the EnSRF and the DKF. However, the higher scores come at the cost of obtaining the wrong values.

We also estimated parameters in simulations of few events, reflecting the typical size of good paleoseismic records of earthquakes (e.g., Biasi et al., 2002). Figure 9 shows the approximate log-likelihood contours of the SIR particle filter for a simulation of a 10-event point process. The benchmark estimate is now much closer to the "true" constant, but the benchmark's 95% confidence interval still does not include the target, namely the maximum likelihood estimate based on the "true" occurrence times. The SIR maximum likelihood estimate, on the other hand, is very close to the "true" estimate.

To establish that the SIR and Kalman filters consistently obtain better parameter estimates than the benchmark, we investigated the statistics of repeated estimates from different realizations of the point process. Desirable properties of parameter estimators include unbiasedness, consistency and (asymptotic) normality. Olsson and Rydén (2008) treat theoretical properties of the maximum likelihood estimator of particle filters. Here, we concentrate on simulations.

We simulated 500 replicas of a 100-event point process with the usual parameters and using both uniform and Gaussian mixture distributed observational errors. We then estimated the parameters of each catalog by maximizing the likelihood function of the SIR filter, the Kalman filters, the benchmark and the "true" lognormal process. Because of the stochastic nature of the estimation problem, this resulted in a distribution of parameters for each method. In Fig. 10, we compare the distributions of the parameter estimates resulting from uniform noise (a and b) and from bi-modal noise (c and d).

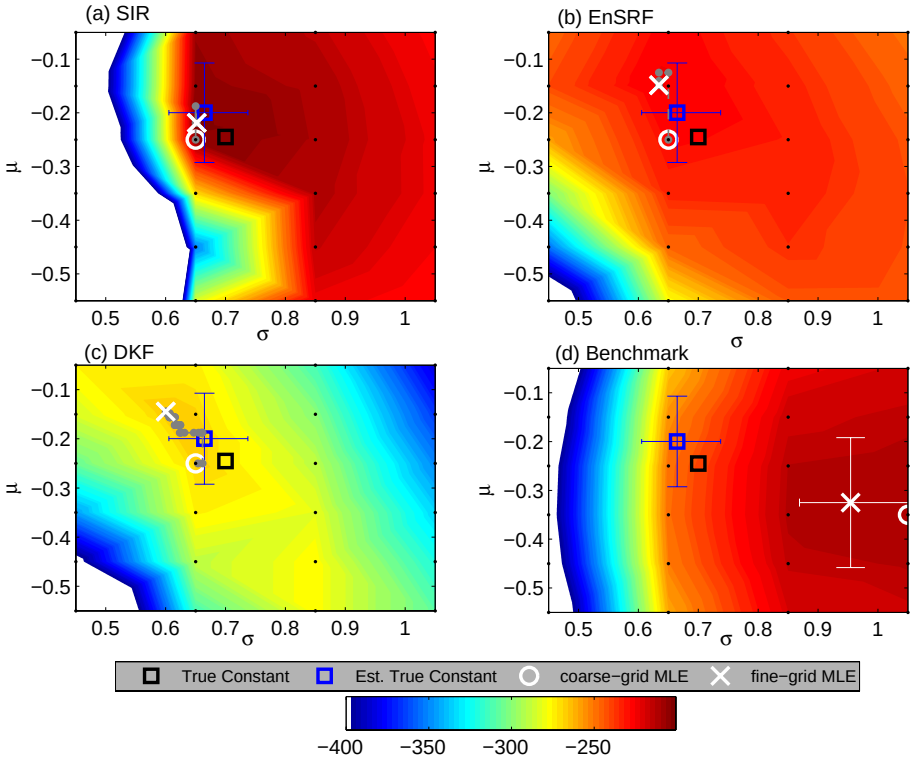

Fig. 8. Complete marginal log-likelihood contour levels of a single 200-event point process simulation as a function of the parameters $\{\mu,\sigma\}$, approximated by **(a)** the SIR particle filter, **(b)** the EnSRF, **(c)** the DKF and **(d)** the benchmark. We also show the "true" constants (black square), the maximum likelihood estimates (and 95% confidence bounds) using the (unobservable) "true" occurrence times (blue square). The grid of the coarse-grid search is indicated by black dots. The path of the pattern search optimization is shown in grey leading to the final maximum likelihood estimate.

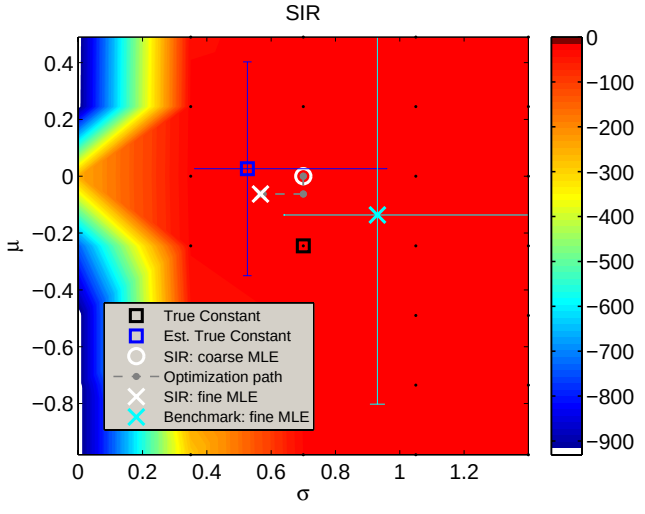

Fig. 9. Log-likelihood contour levels of the SIR filter for a simulation of a 10-event point process. Same explanation as for Fig. 8.

In the case of uniformly distributed observational errors, all three filters provide unbiased estimates of μ, but only the SIR and the EnSRF on average recover the "true" value of σ. The benchmark strongly underestimates μ and overestimates σ, as we had already observed for the single simulations discussed above.

The bi-modally distributed noise makes the estimation problem harder for all filters. The SIR filter provides the best estimates, providing an essentially unbiased estimate of μ and a slightly underestimated value of σ. The EnSRF and DKF both overestimate μ and underestimate σ as a result of the approximations to the likelihood function that the filters make. As in the case of uniformly distributed errors, the benchmark provides the worst estimates. It is also interesting to note that the variance of the benchmark estimators is larger than the filter estimators' variances.

We also tested how the SIR, EnSRF and DKF compare to the benchmark in terms of log-likelihood scores when the parameters are being estimated from an observed (but synthetic) data set, rather than using the exact parameters as in Sect. 4.2.2. Using again the 500 replicas of a 100-event point process from which we estimated maximum likelihood parameters, we calculated the log-likelihood ratios between the filters and the benchmark using the estimated parameters. We found that in 93% of the 500 replicas, the log-likelihood score of the SIR particle filter was larger than that of the

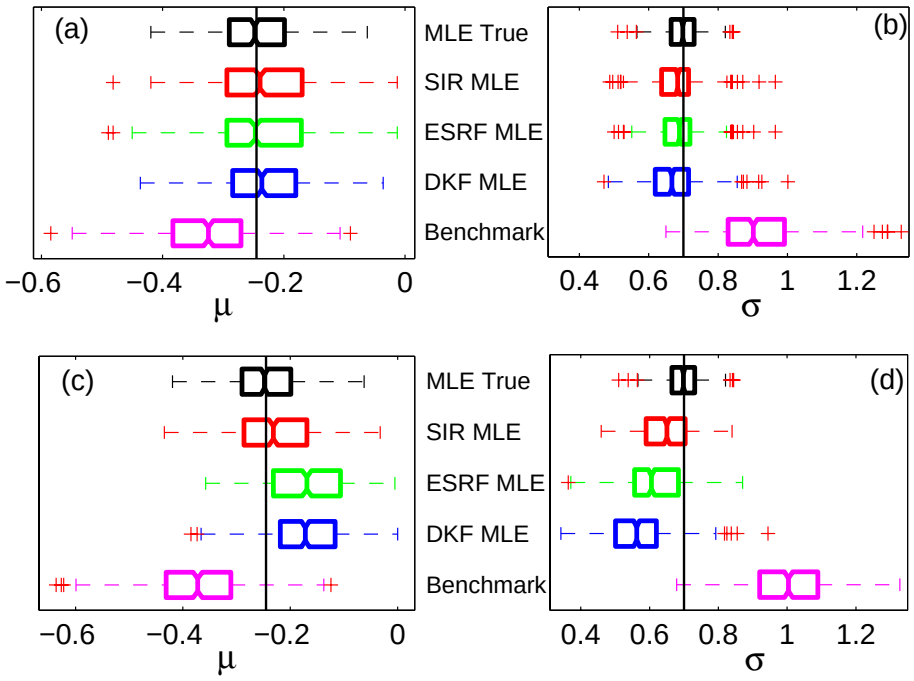

Fig. 10. Boxplot of the distributions of the parameters μ (left) and σ (right) estimated using the "true" event times, the SIR filter, the EnSRF, the DKF and the benchmark. **(a)** and **(b)** uniformly distributed observational errors. **(c)** and **(d)** Gaussian mixture distributed observational errors. Each boxplot shows the median and its 95% confidence intervals (notches). The boxes mark the 25th and 75th percentile (the interquartile range) of the empirical distribution. The whiskers mark 1.5 times the interquartile range and the red pluses mark outliers beyond this range.

benchmark, when the observational errors were uniformly distributed. That number was 94% for the EnSRF and only 11% for the DKF. In the case of Gaussian mixture distributed errors, the SIR filter generated higher likelihood scores than the benchmark in 89% of the simulations. Strikingly, the EnSRF only outperformed the benchmark in 33% of the simulations, while the DKF beat the benchmark in only 8% of the cases. Thus, on average, the SIR retrieves better parameter estimates than both the benchmark and the Kalman filters in the case of bi-modal measurement errors, and it achieves higher likelihood scores.

5 Conclusions

In this article, we have shown the potential benefits and the feasibility of data assimilation-based earthquake forecasting for a simple renewal process observed in noise. We used sequential Monte Carlo methods, a flexible set of simulation-based methods for sampling from arbitrary distributions, and both simple and more advanced, ensemble-based Kalman filters to represent the posterior distributions of the exact event times given noisy observed event times. We showed that a particular particle filter, the Sampling Importance Resampling (SIR) filter, which uses the prior as the importance density to sample the posterior and includes a resampling step to rejuvenate the particles, can solve this particular pedagogical example for an arbitrary number of events and even for complex, bi-modal distributions of the observational error distribution. The SIR may thus be useful for realistic problems such as likelihood-based inference of competing recurrence models on paleoseismic data sets, fully accounting for complex, multi-modal observational error distributions. In contrast, the simple, deterministic Kalman filter (DKF) retrieved biased parameters and low likelihood scores as a result of its inadequate Gaussian approximations to the forecast and analysis. The Ensemble Square Root Filter (EnSRF), which approximates the forecast and analysis with an ensemble of particles but assumes Gaussian observational errors, was able to compete with the particle filter in the case of uniform noise, but it failed to adequately solve the problem for more complex measurement noise. Thus the EnSRF may be a viable alternative to the SIR whenever the measurement error distribution is close to Gaussian, but the problem with arbitrary data uncertainties is best solved with particle filters.

We measured the improvement of the data assimilation-based methods over the uncertainty-ignoring benchmark method by using the marginal complete data likelihood, the denominator in Bayes' theorem. The marginal likelihood generalizes the traditional likelihood by accounting for the presence of observational errors when judging the quality of a forecast or estimating parameters. The marginal likelihood

function explicitly accounts for data uncertainties, and this desired property makes it a powerful and currently underutilized tool in the analysis of earthquake forecasts and model inference. In particular, the marginal likelihood could help in the framework of earthquake predictability experiments such as RELM and CSEP.

The pedagogical example we presented in this article illustrated the power of data assimilation and suggests many avenues of future research. As discussed, one application lies in model inference from paleoseismic data sets with complex conditional likelihood functions that truly capture uncertainties in the dating process. Another interesting possibility is to extend the present renewal process forecast model to more advanced, multi-dimensional point-process models of seismicity. For example, the poorly constrained magnitudes of earthquakes and the resulting amount of slip on the causative fault can be incorporated into renewal processes, whether for the purposes of modeling paleo-earthquakes (e.g. Ogata, 2002) or present-day small repeating earthquakes (Nadeau and Johnson, 1998). Furthermore, formulating and implementing data-assimilation-based schemes for spatio-temporal short-term models of clustered seismicity could reduce their sensitivity to magnitude uncertainties (Werner and Sornette, 2008), errors in locating quakes or uncertain stress calculations (Hainzl et al., 2009). Finally, physics-based seismicity models, such as models based on static stress transfer or other earthquake simulators that require estimates of otherwise unobservable quantities, are particularly likely to benefit from methods of data assimilation. We hope this article, by illustrating some of the potential methods, stimulates some interest in this area.

Acknowledgements. MJW and DS were supported by the EXTREMES project of ETH's Competence Center Environment and Sustainability (CCES). KI was supported by Office of Naval Research grants N00014040191 and N000140910418.

Edited by: O. Talagrand
Reviewed by: D. Rhoades and another anonymous referee

References

Andrieu, C., Doucet, A., Singh, S., and Tadic, V.: Particle methods for change detection, system identification, and control, Proceedings of the IEEE, 92, 423–438, doi:10.1109/JPROC.2003.823142, 2004.

Arulampalam, M., Maskell, S., Gordon, N., and Clapp, T.: A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking, Signal Processing, IEEE Transactions on, 50, 174–188, doi:10.1109/78.978374, 2002.

Bakun, W. H., Aagaard, B., Dost, B., Ellsworth, W. L., Hardebeck, J. L., Harris, R. A., Ji, C., Johnston, M. J. S., Langbein, J., Lienkaemper, J. J., Michael, A. J., Murray, J. R., Nadeau, R. M., Reasenberg, P. A., Reichle, M. S., Roeloffs, E. A., Shakal, A., Simpson, R. W., and Waldhauser, F.: Implications for prediction and hazard assessment from the 2004 Parkfield earthquake, Nature, 437, 969–974, doi:10.1038/nature04067, 2005.

Biasi, G., Weldon, R., Fumal, T., and Seitz, G.: Paleoseismic event dating and the conditional probability of large earthquakes on the southern San Andreas fault, California, B. Seismol. Soc. Am., 92, 2761–2781, 2002.

Cappé, O., Moulines, E., and Ryden, T.: Inference in Hidden Markov Models (Springer Series in Statistics), Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

Cappé, O., Godsill, S., and Moulines, E.: An Overview of Existing Methods and Recent Advances in Sequential Monte Carlo, Proceedings of the IEEE, 95, 899–924, doi:10.1109/JPROC.2007.893250, 2007.

Cornell, C. A.: Engineering seismic risk analysis, Bull. Seismol. Soc. Am., 58, 1583–1606, http://www.bssaonline.org/cgi/content/abstract/58/5/1583, 1968.

Daley, R.: Atmospheric Data Analysis, Cambridge University Press, Cambridge, UK, 1991.

Daley, D. J. and Vere-Jones, D.: An Introduction to the Theory of Point Processes, vol. I, Springer, New York, USA, 2003.

Daley, D. J. and Vere-Jones, D.: Scoring probability forecasts for point processes: The entropy score and information gain, J. Appl. Prob., 41A, 297–312, 2004.

Davis, P. M., Jackson, D. D., and Kagan, Y. Y.: The longer it has been since the last earthquake, the longer the expected time till the next?, B. Seismol. Soc. Am., 79, 1439–1456, 1989.

de Freitas, N.: Bayesian Methods for Neural Networks, PhD thesis, Cambridge University, http://www.cs.ubc.ca/~nando/publications.php, 1999.

Doucet, A., Godsill, S., and Andrieu, C.: On Sequential Monte Carlo Sampling methods for Bayesian filtering, Stat. Comput., 10, 197–208, 2000.

Doucet, A., de Freitas, N., and Gordon, N. (Eds.): Sequential Monte Carlo Methods in Practice, Springer-Verlag, New York, 2001.

Durbin, J. and Koopman, S. J.: Time Series Analysis by State Space Methods, Oxford University Press, Oxford, UK, 2001.

Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, J. Geophys. Res., 99, 10143–10162, doi:10.1029/94JC00572, 1994.

Field, E. H.: A Summary of Previous Working Groups on California Earthquake Probabilities, B. Seismol. Soc. Am., 97, 1033–1053, doi:10.1785/0120060048, 2007a.

Field, E. H.: Overview of the Working Group for the Development of Regional Earthquake Likelihood Models (RELM), Seismol. Res. Lett., 78, 7–16, doi:10.1785/gssrl.78.1.7, 2007b.

Geller, R. J.: Earthquake prediction: A critical review, Geophys. J. Int., 131, 425–450, doi:10.1111/j.1365-246X.1997.tb06588.x, 1997.

Geller, R. J., Jackson, D. D., Kagan, Y. Y., and Mulargia, F.: Earthquakes cannot be predicted, Science, 275, 1616–1617, 1997.

Gerstenberger, M. C., Wiemer, S., Jones, L. M., and Reasenberg, P. A.: Real-time forecasts of tomorrow's earthquakes in California, Nature, 435, 328–331, 2005.

Ghil, M. and Malanotte-Rizzoli, P.: Data Assimilation in Meteorology and Oceanography, Adv. Geophys., 33, 141–266, doi:10.1016/S0065-2687(08)60442-2, 1991.

Grant, L. B. and Gould, M. M.: Assimilation of Paleoseismic Data for Earthquake Simulation, Pure and Applied Geophysics, 161,

2295–2306, doi:10.1007/s00024-003-2564-8, 2004.
Hainzl, S., Enescu, B., Cocco, M., Woessner, J., Catalli, F., Wang, R., and Roth, F.: Aftershock modeling based on uncertain stress calculations, Journal of Geophysical Research Solid Earth, 114, B05309, doi:10.1029/2008JB006011, 2009.
Harte, D. and Vere-Jones, D.: The Entropy Score and its Uses in Earthquake Forecasting, Pure and Applied Geophysics, 162, 1229–1253, doi:10.1007/s00024-004-2667-2, 2005.
Helmstetter, A. and Sornette, D.: Subcritical and supercritical regimes in epidemic models of earthquake aftershocks, J. Geophys. Res., 107, 2237, doi:10.1029/2001JB001580, 2002.
Hooke, R. and Jeeves, T. A.: "Direct Search" Solution of Numerical and Statistical Problems, J. ACM, 8, 212–229, doi:10.1145/321062.321069, 1961.
Ide, K., Bennett, A., Courtier, P., Ghil, M., and Lorenc, A.: Unified notation for data assimilation: Operational, sequential and variational, J. Meteor. Soc. Japan, 75, 71–79, 1997.
Jackson, D. D.: Hypothesis Testing and Earthquake Prediction, Proc. Natl. Acad. Sci. USA, 93, 3772–3775, 1996.
Jordan, T. H.: Earthquake Predictability: Brick by Brick, Seismol. Res. Lett., 77, 3–6, 2006.
Jordan, T. and Jones, L.: Operational Earthquake Forecasting: Some Thoughts on Why and How, Seismol. Res. Lett., 81, 571–574, doi:10.1785/gssrl.81.4.571, 2010.
Kagan, Y. Y.: Statistics of characteristic earthquakes, B. Seismol. Soc. Am., 83, 7–24, 1993.
Kagan, Y. Y.: Are earthquakes predictable?, Geophys. J. Int., 131, 505–525, 1997.
Kagan, Y. Y.: Universality of the seismic moment-frequency relation, Pure and Appl. Geophys., 155, 537–573, 1999.
Kagan, Y. Y. and Jackson, D. D.: Seismic gap hypothesis: Ten years after, J. Geophys. Res., 96, 21419–21431, 1991.
Kagan, Y. Y. and Jackson, D. D.: New Seismic Gap Hypothesis: Five Years After, J. Geophys. Res., 100, 3943–3959, 1995.
Kagan, Y. Y. and Jackson, D. D.: Probabilistic forecasting of earthquakes, Geophys. J. Int., 143, 483–453, 2000.
Kagan, Y. Y. and Knopoff, L.: Statistical Short-Term Earthquake Prediction, Science, 236, 1563–1567, 1987.
Kalman, R. E.: A new Approach to Linear Filtering and Predictive Problems, Transactions ASME, J. Basic Eng., 82D, 35–45, 1960.
Kalman, R. E. and Bucy, R. S.: New Results in Linear Filtering and Prediction Theory, Transactions ASME, J. Basic Eng., 83, 95–108, 1961.
Kalnay, E.: Atmospheric Modeling, Data Assimilation and Predictability, Cambridge Univ. Press, Cambridge, UK, 2003.
Kong, A., Liu, J., and Wong, W.: Sequential Imputations and Bayesian Missing Data Problems, J. Am. Stat. Assoc., 89, 278–288, 1994.
Künsch, H. R.: State space and hidden Markov models, in: Complex stochastic systems (Eindhoven, 1999), vol. 87 of Monogr. Statist. Appl. Probab., pp. 109–173, Chapman & Hall/CRC, Boca Raton, FL, 2001.
Laherrère, J. and Sornette, D.: Stretched exponential distributions in nature and economy: "Fat tails" with characteristic scales, European Physical Journal B, 2, 525–539, doi:10.1007/s100510050276, 1998.
Lewis, R. M. and Torczon, V.: Pattern Search Algorithms for Bound Constrained Minimization, SIAM J. Optimization, 9, 1082–1099, doi:http://dx.doi.org/10.1137/S1052623496300507, 1999.
Liu, J. S.: Monte Carlo Strategies in Scientific Computing, Springer-Verlag, New York, 2001.
Liu, J. S. and Chen, R.: Sequential Monte Carlo Methods for Dynamic Systems, J. Am. Stat. Assoc., 93, 1032–1044, 1998.
Lovett, W.: Banking and Financial Institutions Law in a Nutshell, West Publishing Co., second edn., 1988.
McCann, W., Nishenko, S., Sykes, L., and Krause, J.: Seismic Gaps and plate tectonics: Seismic Potential for major boundaries, Pure Appl. Geophys., 117, 1082–1147, 1979.
McGuire, J. J.: Seismic Cycles and Earthquake Predictability on East Pacific Rise Transform Faults, B. Seismol. Soc. Am., 98, 1067–1084, doi:10.1785/0120070154, 2008.
Mézard, M., Parisi, G., and Virasoro, M.: Spin glass theory and beyond, World Scientific Lecture Notes in Physics Vol. 9, Cambridge Univ. Press, Cambridge, UK, 1987.
Miller, R. N., Carter, E. F., and Blue, S. T.: Data Assimilation into Nonlinear Stochastic Models, Tellus, 51A, 167–194, 1999.
Nadeau, R. and Johnson, L.: Seismological studies at Parkfield VI: Moment release rates and estimates of source parameters for small repeating earthquakes, B. Seismol. Soc. Am., 88, 790–814, 1998.
Nishenko, S. P.: Circum-Pacific seismic potential 1989–1999, Pure Appl. Geophys., 135, 169–259, 1991.
Nishenko, S. P. and Buland, R.: A generic recurrence interval distribution for earthquake forecasting, B. Seismol. Soc. Am., 77, 1382–1399, 1987.
Ogata, Y.: Statistical models for earthquake occurrence and residual analysis for point processes, J. Am. Stat. Assoc., 83, 9–27, 1988.
Ogata, Y.: Space-time Point-process Models for Earthquake Occurrences, Ann. Inst. Stat. Math., 5, 379–402, 1998.
Ogata, Y.: Estimating the Hazard of Rupture Using Uncertain Occurrence Times of Paleoearthquakes, J. Geophys. Res., 104, 17995–18014, 1999.
Ogata, Y.: Slip-Size-Dependent Renewal Processes and Bayesian Inferences for Uncertainties, J. Geophys. Res., 107, 2268, doi:10.1029/2001JB000668, 2002.
Olsson, J. and Rydén, T.: Asymptotic properties of particle filter-based maximum likelihood estimators for state space models, Stoc. Proc. Appl., 118, 649–680, doi:10.1016/j.spa.2007.05.007, 2008.
Parsons, T.: Monte Carlo method for determining earthquake recurrence parameters from short paleoseismic catalogs: Example calculations for California, J. Geophys. Res. (Solid Earth), 113, B03302, doi:10.1029/2007JB004998, 2008.
Pham, D. T.: Stochastic Methods for Sequential Data Assimilation in Strongly Nonlinear Systems, Mon. Weather Rev., 129, 1194–1207, 2001.
Polakoff, M. and Durkin, T.: Financial institutions and markets, Houghton Mifflin Boston, second edn., 1981.
Reid, H.: The Mechanics of the Earthquake, The California Earthquake of April 18, 1906, Report of the State Investigation Commission, Vol. 2, Carnegie Institution of Washington, Washington, D.C., pp. 16–28, 1910.
Rhoades, D. A. and Evison, F. F.: Long-range Earthquake Forecasting with Every Earthquake a Precursor According to Scale, Pure Appl. Geophys., 161, 47–72, doi:10.1007/s00024-003-2434-9, 2004.
Rhoades, D. A., Dissen, R. V., and Dowrick, D.: On the Handling of Uncertainties in Estimating the Hazard of Rupture on a Fault

Segment, J. Geophys. Res., 99, 13701–13712, 1994.

Robert, C. P. and Casella, G.: Monte Carlo Statistical Methods (Springer Texts in Statistics), Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2004.

Rong, Y., Jackson, D. D., and Kagan, Y. Y.: Seismic gaps and earthquakes, J. Geophys. Res. (Solid Earth), 108, 2471, doi: 10.1029/2002JB002334, 2003.

Sambridge, M. and Mosegaard, K.: Monte Carlo Methods in Geophysical Inverse Problems, Rev. Geophys., 40, 1009, doi: 10.1029/2000RG000089, 2002.

Scholz, C. H.: The Mechanics of Earthquakes and Faulting, Cambridge University Press, Cambridge, 2nd edn., 2002.

Schorlemmer, D., Gerstenberger, M. C., Wiemer, S., Jackson, D. D., and Rhoades, D. A.: Earthquake Likelihood Model Testing, Seismol. Res. Lett., 78, 17, 2007.

Schorlemmer, D., Zechar, J. D., Werner, M. J., Field, E., Jackson, D. D., and Jordan, T. H.: First Results of the Regional Earthquake Likelihood Models Experiment, Pure and Appl. Geophys.: The Frank Evison Volume, 167(8/9), doi:10.1007/s00024-010-0081-5, 2010.

Snyder, C., Bengtsson, T., Bickel, P., and Anderson, J.: Obstacles to High-Dimensional Particle Filtering, Mon. Weather Rev., 136, 4629–4640, 2008.

Sornette, D.: Critical Phenomena in Natural Sciences: Chaos, Fractals, Selforganization and Disorder: Concepts and Tools, Springer, Berlin, 2nd edn., 2004.

Sornette, D. and Ide, K.: The Kalman-Lévy filter, Physica D Nonlinear Phenomena, 151, 142–174, 2001.

Sornette, D., Knopoff, L., Kagan, Y., and Vanneste, C.: Rank-ordering statistics of extreme events: application to the distribution of large earthquakes, J. Geophys. Res., 101, 13883–13893, 1996.

Sykes, L. R. and Menke, W.: Repeat Times of Large Earthquakes: Implications for Earthquake Mechanics and Long-Term Prediction, B. Seismol. Soc. Am., 96, 1569–1596, doi:10.1785/0120050083, 2006.

Talagrand, O.: Assimilation of Observations, J. Meteorol. Soc. Japan, 75, 191–209, 1997.

Tarantola, A.: Inverse Problem Theory, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, 1987.

Tippett, M. K., Anderson, J. L., Bishop, C. H., Hamill, T. M., and Whitaker, J. S.: Ensemble Square Root Filters, Mon. Weather Rev., 131, 1485–1490, doi:10.1175/1520-0493(2003)131⟨1485:ESRF⟩2.0.CO;2, http://journals.ametsoc.org/doi/abs/10.1175/1520-0493282003291313C14853AESRF3E2.0.CO3B2, 2003.

Torczon, V.: On the Convergence of Pattern Search Algorithms, SIAM J. on Optimization, 7, 1–25, doi:http://dx.doi.org/10.1137/S1052623493250780, 1997.

Van Aalsburg, J., Grant, L. B., Yakovlev, G., Rundle, P. B., Rundle, J. B., Turcotte, D. L., and Donnellan, A.: A feasibility study of data assimilation in numerical simulations of earthquake fault systems, Phys. Earth Planet. In., 163, 149–162, doi:10.1016/j.pepi.2007.04.020, 2007.

Varini, E.: Sequential Estimation Methods in Continuous-Time State Space Models, PhD thesis, Università Commerciale "Luigi Bocconi" – Milano, 2005.

Varini, E.: A Monte Carlo method for filtering a marked doubly stochastic Poisson process, Stat. Method Appl., 17, 183–193, 2008.

Vere-Jones, D.: Stochastic Models for Earthquake Occurrence, J. Roy. Stat. Soc. Series B (Methodological), 32, 1–62 (with discussion), 1970.

Vere-Jones, D.: Forecasting Earthquakes and earthquake risk, Intern. J. Forecasting, 11, 503–538, 1995.

Werner, M. J. and Sornette, D.: Magnitude Uncertainties Impact Seismic Rate Estimates, Forecasts and Predictability Experiments, J. Geophys. Res. Solid Earth, 113, B08302, doi: 10.1029/2007JB005427, 2008.

Werner, M. J., Helmstetter, A., Jackson, D. D., and Kagan, Y. Y.: High Resolution Long- and Short-Term Earthquake Forecasts for California, B. Seismol. Soc. Am., accepted, preprint available at http://arxiv.org/abs/0910.4981, 2010a.

Werner, M. J., Helmstetter, A., Jackson, D. D., Kagan, Y. Y., and Wiemer, S.: Adaptively Smoothed Seismicity Earthquake Forecasts for Italy, Annals of Geophysics, 53, 107–116, doi: 10.4401/ag-4839, 2010b.

Werner, M. J., Zechar, J. D., Marzocchi, W., and Wiemer, S.: Retrospective Evaluation of the Five-year and Ten-year CSEP-Italy Earthquake Forecasts, Annals of Geophysics, 53, 11–30, doi: 10.4401/ag-4840, 2010c.

Wesnousky, S. G.: The Gutenberg-Richter or characteristic earthquake distribution, which is it?, B. Seismol. Soc. Am., 84, 1940–1959, 1994.

Wikle, C. K. and Berliner, L. M.: A Bayesian tutorial for data assimilation, Physica D Nonlinear Phenomena, 230, 1–2, doi: 10.1016/j.physd.2006.09.017, 2007.

Zechar, J. D., Schorlemmer, D., Liukis, M., Yu, J., Euchner, F., Maechling, P. J., and Jordan, T. H.: The Collaboratory for the Study of Earthquake Predictability perspective on computational earthquake science, Concurrency and Computation: Practice and Experience, 22(12), 1836–1847, doi:10.1002/cpe.1519, 2010.