Statistics and Computing (2020) 30:1497–1522
https://doi.org/10.1007/s11222-020-09957-3

Inference on high-dimensional implicit dynamic models using a guided intermediate resampling filter

Joonha Park[1] · **Edward L. Ionides**[2]

Received: 27 August 2019 / Accepted: 4 June 2020 / Published online: 26 June 2020

Abstract

We propose a method for inference on moderately high-dimensional, nonlinear, non-Gaussian, partially observed Markov process models for which the transition density is not analytically tractable. Markov processes with intractable transition densities arise in models defined implicitly by simulation algorithms. Widely used particle filter methods are applicable to nonlinear, non-Gaussian models but suffer from the curse of dimensionality. Improved scalability is provided by ensemble Kalman filter methods, but these are inappropriate for highly nonlinear and non-Gaussian models. We propose a particle filter method having improved practical and theoretical scalability with respect to the model dimension. This method is applicable to implicitly defined models having analytically intractable transition densities. Our method is developed based on the assumption that the latent process is defined in continuous time and that a simulator of this latent process is available. In this method, particles are propagated at intermediate time intervals between observations and are resampled based on a forecast likelihood of future observations. We combine this particle filter with parameter estimation methodology to enable likelihood-based inference for highly nonlinear spatiotemporal systems. We demonstrate our methodology on a stochastic Lorenz 96 model and a model for the population dynamics of infectious diseases in a network of linked regions.

Keywords Sequential Monte Carlo · Particle filter · Spatiotemporal inference · Curse of dimensionality · Implicit models · Plug-and-play property

1 Introduction

In this paper, we consider inference on highly nonlinear, moderately high-dimensional Markov process models for which evaluation of the transition density is not available. Data are modeled as partial or noisy measurements of the latent Markov process. We will first introduce in turn the three model aspects we are concerned with, namely intractable transition densities, high-dimensionality, and nonlinearity.

A model that is defined using a simulator, instead of an analytically tractable characterization, of an underlying process is said to be implicitly defined (Diggle and Gratton 1984). Mechanistic models for complex dynamic systems are sometimes defined implicitly by a computer simulation algorithm, and such models often lack analytically tractable transition densities. Inference methods that can be used on implicitly defined models are said to possess the *plug-and-play* property (Bretó et al. 2009; He et al. 2009), or alternatively called *equation-free* (Kevrekidis et al. 2004; Xiu et al. 2005) or *likelihood-free* (Marjoram et al. 2003; Sisson et al. 2007).

Inference on dynamic systems sometimes requires fitting models with high-dimensional latent processes to high-dimensional data. For example, population dynamics in geographically linked locations are sometimes studied in ecology or epidemiology using partially observed Markov process (POMP) models for which the dimension of both the latent process and the measurement process scale linearly with the number of spatial locations. In systems biology, models for networks of reactions may add stochasticity to collections of deterministic differential equations (Kitano

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11222-020-09957-3) contains supplementary material, which is available to authorized users.

✉ Joonha Park
joonhap@bu.edu

Edward L. Ionides
ionides@umich.edu

[1] Department of Mathematics and Statistics, Boston University, 111 Cummington Mall, Boston, MA 02215, USA

[2] Department of Statistics, University of Michigan, Ann Arbor, MI, USA

2002). The model dimension typically increases with the number of system components, but even state-of-the-art inference methods are often not suitable for application beyond small systems (Owen et al. 2015).

Ensemble Kalman filter (EnKF) methods have been used for geophysical models in data assimilation due to their good scalability to high dimensions (Houtekamer and Mitchell 2001; Evensen 1994). However, these methods can be ineffective for highly nonlinear and non-Gaussian systems, because they rely on locally linear and Gaussian approximations (Ades and Van Leeuwen 2015; Lei et al. 2010; Miller et al. 1999). For example, host-pathogen population dynamics of infectious diseases in geographically coupled regions often exhibit a long period of epidemic trough followed by a sharp peak, which may be sparked by an invasion of the pathogen from a different region. Strong nonlinearity of the dynamics, as well as the inferential relevance of a small, discrete, number of initial infections, makes the use of ensemble Kalman filter methods unsuitable for these models.

We propose an approach for inference on a class of implicitly defined nonlinear partially observed Markov process (POMP) models of moderately high dimensions. A POMP model, otherwise known as a state space model or a hidden Markov model (HMM), consists of a latent Markov process representing the time evolution of a system and measurement processes describing the randomness of observations at specified time points. Each obervation Y_n provides a partial or noisy information about the latent process state X_n at time t_n. We denote the density of Y_n given $X_n = x_n$ by $g_n(\cdot \mid x_n)$. A sequence of observations $y_{1:N}$ made at time $t_{1:N}$ are assumed to be given as fixed data. Sequential Monte Carlo (SMC) methods are recursive algorithms that enable estimation of the likelihood of observed data and the conditional distribution of the latent process given data from a POMP model (Doucet et al. 2001; Cappé et al. 2007; Doucet and Johansen 2011). In the context of POMP models, SMC algorithms are known as particle filters (PFs), and the simulated random variables used by SMC to represent conditional latent processes are called particles. Particle filter methods are capable of handling highly nonlinear latent processes, but they suffer from rapid deterioration in performance as the model dimension increases (Bengtsson et al. 2008; Snyder et al. 2008).

In order to introduce our method, we briefly review some particle filter methods. In inference on POMP models, the conditional distribution of X_n given observations $y_{1:n}$, called the filtering distribution at time t_n, often makes a distribution of interest. Particle filters recursively represent the filtering distributions using particle ensembles. A collection of particles $X^{1:J} := \{X^j ; 1 \le j \le J\}$ is said to represent a distribution with density p if the sample average $\frac{1}{J}\sum_{j=1}^{J} f(X^j)$ for a class of functions f gives an estimate of $\int f(x)p(x)\mathrm{d}x$. Suppose that an ensemble $\tilde{X}_n^{1:J}$ represent the filtering density $p(x_n \mid y_{1:n})$. The next filtering density can be expressed as

$$p(x_{n+1} \mid y_{1:n+1}) = \frac{\int p(x_n \mid y_{1:n}) p(x_{n+1} \mid x_n) g_{n+1}(y_{n+1} \mid x_{n+1}) \mathrm{d}x_n}{\int p(x_n \mid y_{1:n}) p(x_{n+1} \mid x_n) g_{n+1}(y_{n+1} \mid x_{n+1}) \mathrm{d}x_n \mathrm{d}x_{n+1}}. \quad (1)$$

Based on (1), a particle representation $\tilde{X}_{n+1}^{1:J}$ for the next filtering density $p(x_{n+1} \mid y_{1:n+1})$ can be obtained as follows. The particle ensemble $\tilde{X}_n^{1:J}$ can be propagated using the transition kernel $p(x_{n+1} \mid x_n)$. The propagated particles, denoted by $X_{n+1}^{1:J}$, can then be resampled according to weights proportional to $\{g_{n+1}(y_{n+1} \mid X_{n+1}^j) ; j \in 1:J\}$. The resampled particles $\tilde{X}_{n+1}^{1:J}$ represent $p(x_{n+1} \mid y_{1:n+1})$. The method of recursively updating the particle ensemble in this way is called the bootstrap particle filter (Gordon et al. 1993). The resampling can be carried out by, for example, taking $\tilde{X}_{n+1}^k := X_{n+1}^{a_k}$, where a_k, $k \in 1:J$, are independent and $\mathbb{P}(a_k = i) = \frac{g_{n+1}(y_{n+1} \mid X_{n+1}^i)}{\sum_{j=1}^{J} g_{n+1}(y_{n+1} \mid X_{n+1}^j)}$. Alternative methods of resampling may be preferable (Douc et al. 2005).

Another method for obtaining a particle representation of the next filtering density is based on the equation

$$p(x_{n+1} \mid y_{1:n+1}) = \frac{\int p(x_n \mid y_{1:n}) p(x_{n+1} \mid x_n, y_{n+1}) p(y_{n+1} \mid x_n) \mathrm{d}x_n}{\int p(x_n \mid y_{1:n}) p(x_{n+1} \mid x_n, y_{n+1}) p(y_{n+1} \mid x_n) \mathrm{d}x_n \mathrm{d}x_{n+1}}.$$

Suppose now that $X_n^{1:J}$ represent the distribution $p(x_n \mid y_{1:n})$. Resampling according to weights proportional to $p_{Y_{n+1}|X_n}(y_{n+1} \mid X_n^j)$ leads to particles denoted by $\tilde{X}_n^{1:J}$, representing $p(x_n \mid y_{1:n+1})$. Propagating $\tilde{X}_n^{1:J}$ with the kernel $p(x_{n+1} \mid x_n, y_{n+1})$, one obtains a particle representation $X_{n+1}^{1:J}$ of $p(x_{n+1} \mid y_{1:n+1})$. This method corresponds to the fully adapted auxiliary particle filter (APF) (Pitt and Shephard 1999). The propagation kernel $p(x_{n+1} \mid x_n, y_{n+1})$ in this context is called *adapted* to y_{n+1}, because it uses the information in the next observation. A method equivalent to the fully adapted APF, in which particles are resampled according to weights proportional to $p(y_{n+1} \mid x_n)$ and propagated with the adapted kernel $p(x_{n+1} \mid x_n, y_{n+1})$ has been considered by Kong et al. (1994), Liu and Chen (1995), and Chen et al. (2000). The auxiliary particle filter by Pitt and Shephard (1999) uses $g_{n+1}(y_{n+1} \mid \bar{\xi}_{n+1}(x_n))$ as an approximation to $p(y_{n+1} \mid x_n)$, where $\bar{\xi}_{n+1}(x_n)$ is a point that can represent the conditional distribution $p_{X_{n+1}|X_n}(\cdot \mid x_n)$, such as the mean of the distribution $p_{X_{n+1}|X_n}(\cdot \mid x_n)$ or an approximation thereof. Doucet et al. (2000) called the propagation of $\tilde{X}_n^{1:J}$ by the adapted kernel $p(x_{n+1} \mid x_n, y_{n+1})$ optimal when only the next observation is available, since the particles $X_{n+1}^{1:J}$ having *equal* weights represent $p(x_{n+1} \mid y_{1:n+1})$. An advantage of the fully adapted APF compared to the bootstrap PF is that the coefficient of variation of the resampling weights is smaller: we have

$$\mathrm{Var}\big[p_{Y_{n+1}|X_n}(y_{n+1} \mid X_n^{\mathrm{APF}})\big] \le \mathrm{Var}\big[g_{n+1}(y_{n+1} \mid X_{n+1}^{\mathrm{BPF}})\big],$$

and

$$\mathbb{E}\big[p_{Y_{n+1}|X_n}(y_{n+1} \mid X_n^{\mathrm{APF}})\big] = \mathbb{E}\big[g_{n+1}(y_{n+1} \mid X_{n+1}^{\mathrm{BPF}})\big],$$

if X_n^{APF} is a draw from $p(x_n \mid y_{1:n})$ and X_{n+1}^{BPF} is a draw from $p(x_{n+1} \mid y_{1:n})$. However, as Snyder et al. (2015) showed using counterexamples, even the fully adapted APF suffers from rapid deterioration of performance as the latent process and measurement dimension increases, because $\mathrm{Var}\big[p_{Y_{n+1}|X_n}(y_{n+1} \mid X_n^{\mathrm{APF}})\big]$ scales exponentially.

The fully adapted APF can be viewed as a particle filter operating on a twisted POMP model (Whiteley and Lee 2014; Guarniero et al. 2017). A twisted POMP model is defined based on a given POMP model $\{(X_n, Y_n)\,; n \in 1:N\}$ and a sequence of functions $\psi := \{\psi_n\,; n \in 1:N\}$. Denoting $\tilde{\psi}_n(x_n) := \int \psi_{n+1}(x_{n+1}) p(x_{n+1} \mid x_n) \mathrm{d}x_{n+1}$, we consider a sequence of densities

$$p^{\psi}(x_n\,; y_{1:n}) := \frac{p(x_n \mid y_{1:n})\tilde{\psi}_n(x_n)}{\int p(x_n \mid y_{1:n})\tilde{\psi}_n(x_n)\mathrm{d}x_n}, \quad n \in 1:N,$$

where $\tilde{\psi}_N \equiv 1$. Algorithm 1 is motivated by the recursive relation

$$\begin{aligned} p^{\psi}(x_{n+1}\,; y_{1:n+1}) &\propto \int p^{\psi}(x_n\,; y_{1:n}) \\ &\cdot \frac{p(x_{n+1} \mid x_n)\psi_{n+1}(x_{n+1})}{\tilde{\psi}_n(x_n)} \\ &\cdot g_{n+1}(y_{n+1} \mid x_{n+1}) \frac{\tilde{\psi}_{n+1}(x_{n+1})}{\psi_{n+1}(x_{n+1})} \mathrm{d}x_n. \end{aligned} \tag{2}$$

Algorithm 1: Particle filter on a twisted model

For a particle ensemble $\tilde{X}_n^{1:J}$ that represents $p^{\psi}(x_n\,; y_{1:n})$,

(a) propagate $\tilde{X}_n^{1:J}$ using a kernel that is adapted with respect to ψ_{n+1},

$$f_{n+1}^{\psi}(x_{n+1}\,; x_n) := \frac{p(x_{n+1} \mid x_n)\psi_{n+1}(x_{n+1})}{\tilde{\psi}_n(x_n)},$$

(b) resample the propagated particles $X_{n+1}^{1:J}$ according to weights proportional to $g_{n+1}^{\psi}(X_{n+1}^j)$, where

$$g_{n+1}^{\psi}(x_{n+1}) := g_{n+1}(y_{n+1} \mid x_{n+1}) \frac{\tilde{\psi}_{n+1}(x_{n+1})}{\psi_{n+1}(x_{n+1})}.$$

The case of $\psi_n(x_n) \equiv g_n(y_n \mid x_n)$ corresponds to the fully adapted APF. The variances of the resampling weights are minimized when $\psi_n(x_n) = p(y_{n:N} \mid x_n)$, the forecast likelihood of all future observations (Guarniero et al. 2017). In this case, no resampling is necessary, because $g_n(y_n \mid x_n)\frac{\tilde{\psi}(x_n)}{\psi(x_n)} = g_n(y_n \mid x_n)\frac{p(y_{n+1:N} \mid x_n)}{p(y_{n:N} \mid x_n)} \equiv 1$ for all n. This ideal case targets the smoothing distribution, that is

$$p^{\psi}(x_n\,; y_{1:n}) = p(x_n \mid y_{1:N}).$$

More accessible than the ideal case is the choice $\psi_n(x_n) = p(y_{n:n+L-1} \mid x_n)$ for some $L \ge 2$. The particle filter corresponding to this case looks ahead to L observations in the future. Looking ahead for the information in future observations can lead to robust filtering estimates with regard to outliers in observed data (Lin et al. 2013). The target density for this ψ is given by

$$p^{\psi}(x_n\,; y_{1:n}) = p(x_n \mid y_{1:n+L}),$$

which is known as the fixed lag smoothing distribution (Clapp and Godsill 1999). In this L-lookahead approach, particles are propagated according to

$$\begin{aligned} &f_{n+1}^{\psi}(x_{n+1}\,; x_n) \\ &= \frac{p(x_{n+1} \mid x_n)p(y_{n+1:n+L} \mid x_{n+1})}{p(y_{n+1:n+L} \mid x_n)} \\ &= p(x_{n+1} \mid x_n, y_{n+1:n+L}), \end{aligned}$$

and the resampling weights for the propagated particles $X_{n+1}^{1:J}$ are given by

$$\begin{aligned} &g_{n+1}(y_{n+1} \mid X_{n+1}^j)\frac{\tilde{\psi}_{n+1}(X_{n+1}^j)}{\psi_{n+1}(X_{n+1}^j)} \\ &= g_{n+1}(y_{n+1} \mid X_{n+1}^j)\frac{p(y_{n+2:n+L+1} \mid X_{n+1}^j)}{p(y_{n+1:n+L} \mid X_{n+1}^j)} \\ &= p(y_{n+L+1} \mid X_{n+1}^j, y_{n+2:n+L}). \end{aligned} \tag{3}$$

Particle filtering on this twisted model corresponds to an adapted version of the block sampling method by Doucet et al. (2006) when we look marginally at X_{n+1}. The block sampling method updates a block of latent process states $X_{n+1:n+L}$ based on the observations $y_{n+1:n+L}$.

The coefficient of variation of resampling weights (3) decreases as L increases. If we denote $v_L := \mathrm{Var}\,[p(y_n \mid X_{n-L}^j, y_{n-L+1:n-1})]$ and $e_L := \mathbb{E}\left[p(y_n | X_{n-L}^j, y_{n-L+1:n-1})\right]$ where $X_{n-L}^j \sim p(x_{n-L} \mid y_{1:n-1})$ for $1 \le L \le n-1$, then we have $v_{L_1} \le v_{L_2}$ and $e_{L_1} = e_{L_2}$ for $L_1 > L_2$. Doucet et al. (2006) considered an example where the variance v_L decreases exponentially with L.

The above L-lookahead approach requires that one can evaluate $\psi_n(x_n) = p(y_{n:n+L-1} \mid x_n)$ and $\tilde{\psi}_n(x_n) = p(y_{n+1:n+L} \mid x_n)$ and can sample from the adapted kernel $p(x_{n+1} \mid x_n, y_{n+1:n+L})$. When these requirements cannot be

met, an approximate method can be used. We denote an approximation to $p(y_{n:n+L-1} \mid x_n)$ by $\psi_n(x_n)$, an approximation to $\tilde{\psi}_n(x_n) = \int \psi_{n+1}(x_{n+1}) p(x_{n+1}|x_n) \mathrm{d}x_{n+1}$ by $r_n(x_n)$, and a kernel approximating $f^{\psi}_{n+1}(x_{n+1}; x_n)$ by $q_{n+1}(x_{n+1}; x_n)$. A recursive relation

$$
\begin{aligned}
&p(x_{n+1} \mid y_{1:n+1}) r_{n+1}(x_{n+1}) \\
&\quad \propto \int p(x_n \mid y_{1:n}) r_n(x_n) \cdot q_{n+1}(x_{n+1}; x_n) \\
&\quad \cdot g_{n+1}(y_{n+1} \mid x_{n+1}) \frac{r_{n+1}(x_{n+1})}{r_n(x_n)} \\
&\quad \times \frac{p(x_{n+1} \mid x_n)}{q_{n+1}(x_{n+1}; x_n)} \mathrm{d}x_n, \qquad (4)
\end{aligned}
$$

which is analogous to (2), motivates an approximate L-lookahead filter, shown in Algorithm 2.

Algorithm 2: An approximate L-lookahead filter

For $\tilde{X}^{1:J}_n$ that represent the density proportional to $p(x_n \mid y_{1:n}) r_n(x_n)$,

(a) propagate $\tilde{X}^{1:J}_n$ using the kernel $q_{n+1}(x_{n+1}; x_n)$, and
(b) resample the propagated particles $X^{1:J}_{n+1}$ according to weights proportional to

$$
g_{n+1}(y_{n+1} \mid X^j_{n+1}) \frac{r_{n+1}(X^j_{n+1})}{r_n(\tilde{X}^j_n)} \frac{p(X^j_{n+1} \mid \tilde{X}^j_n)}{q_{n+1}(X^j_{n+1}; \tilde{X}^j_n)}. \qquad (5)
$$

The resampled particles $\tilde{X}^{1:J}_{n+1}$ represent density proportional to $p(x_{n+1} \mid y_{1:n+1}) r_{n+1}(x_{n+1})$.

In this approximate L-lookahead approach, the function ψ_n indirectly affects the algorithm via r_n and q_n. The propagation kernel $q_{n+1}(x_{n+1}; x_n)$ approximates

$$
\begin{aligned}
q_{n+1}(x_{n+1}; x_n) &\approx \frac{p(x_{n+1} \mid x_n) p(y_{n+1:n+L} \mid x_{n+1})}{p(y_{n+1:n+L} \mid x_n)} \\
&= \frac{p(x_{n+1}, y_{n+1:n+L} \mid x_n)}{p(y_{n+1:n+L} \mid x_n)} = p(x_{n+1} \mid x_n, y_{n+1:n+L}),
\end{aligned}
$$

and the resampling weights (5) approximates (3). We note that the resampling weights (5) depend on the density $p(x_{n+1} \mid x_n)$, so Algorithm 2 cannot be used when the transition density of the latent Markov process is not evaluable. An exception is when we use $q_{n+1}(x_{n+1}; x_n) = p(x_{n+1} \mid x_n)$. In this case, the resampling weights are given by

$$
\begin{aligned}
&\frac{g_{n+1}(y_{n+1} \mid X^j_{n+1}) r_{n+1}(X^j_{n+1})}{r_n(\tilde{X}^j_n)} \\
&\quad \approx \frac{g_{n+1}(y_{n+1} \mid X^j_{n+1}) p(y_{n+2:n+L+1} \mid X^j_{n+1})}{p(y_{n+1:n+L} \mid \tilde{X}^j_n)} \\
&\quad = \frac{p(y_{n+1:n+L+1} \mid X^j_{n+1})}{p(y_{n+1:n+L} \mid \tilde{X}^j_n)}. \qquad (6)
\end{aligned}
$$

However, in this case the variance of the resampling weights becomes too large even for moderate dimensional models, because the weights are lower bounded by

$$
\begin{aligned}
&\mathrm{Var}\left(\frac{p(y_{n+1:n+L+1} \mid X^j_{n+1})}{p(y_{n+1:n+L} \mid \tilde{X}^j_n)} \right) \\
&\quad \geq \mathbb{E}\left[\frac{\mathrm{Var}\left[p(y_{n+1:n+L+1} \mid X^j_{n+1}) \,\middle|\, \tilde{X}^j_n \right]}{p(y_{n+1:n+L} \mid \tilde{X}^j_n)^2} \right],
\end{aligned}
$$

and $\mathrm{Var}\left[p(y_{n+1:n+L+1} \mid X^j_{n+1}) \,\middle|\, \tilde{X}^j_n \right]$ grows exponentially quickly with increasing latent process and measurement dimension.

In this paper, we propose a method that (a) uses the simulator of the latent process for particle propagation and (b) has favorable scaling with respect to increasing dimension. For this method, $X^{1:J}_n$ represents the density proportional to $p(x_n \mid y_{1:n-1}) \psi_n(x_n)$ and the subsequent particle ensemble $X^{1:J}_{n+1}$ represents the density proportional to $p(x_{n+1} \mid y_{1:n}) \psi_{n+1}(x_{n+1})$. Here, $\psi_n(x_n)$ is an approximation to $p(y_{n:n+L-1} \mid x_n)$. This method uses intermediate propagation and resampling steps, as described below. We assume that the latent Markov process is defined in continuous time. We further assume that the latent process, denoted by $\{X(t)\}$, can be simulated for any length of time. A connection to the discrete time process $\{X_n\}$ can be made by understanding $X_n := X(t_n)$ where t_n, $n \in 1:N$, denote the observations times. In the continuous-time context, dummy variables for the latent process will be indexed by time (e.g., x_{t_n}). We consider intermediate time points $t_{n,1} < t_{n,2} < \cdots < t_{n,S-1}$ between t_n and t_{n+1}. We will denote $t_{n,0} := t_n$ and $t_{n,S} := t_{n+1}$. For each intermediate time point $t_{n,s}$, $s \in 1:S$, we define $\psi_{t_{n,s}}(x_{t_{n,s}})$ to be an approximation to $p(y_{n+1:n+L} \mid x_{t_{n,s}})$. We call $\psi_{t_{n,s}}$ the guide function at $t_{n,s}$. At $t_n = t_{n-1,S}$, $\psi_{t_n}(x_{t_n})$ approximates $p(y_{n:n+L-1} \mid x_{t_n})$. We call our method a guided intermediate resampling filter (GIRF), and it works as follows. For $s \in 1:S-1$, suppose that $\tilde{X}^{1:J}_{t_{n,s}}$ represent the density proportional to $p(x_{t_{n,s}} \mid y_{1:n}) \psi_{t_{n,s}}(x_{t_{n,s}})$. These particles are propagated using the simulator of $p(x_{t_{n,s+1}} \mid x_{t_{n,s}})$. The propagated particles, denoted by $X^{1:J}_{t_{n,s+1}}$, are resampled according to weights proportional to

$$
\frac{\psi_{t_{n,s+1}}(X^j_{t_{n,s+1}})}{\psi_{t_{n,s}}(\tilde{X}^j_{t_{n,s}})}.
$$

The resampled particles represent the density proportional to $p(x_{t_{n,s+1}} \mid y_{1:n}) \psi_{t_{n,s+1}}(x_{t_{n,s+1}})$ and are denoted by $\tilde{X}^{1:J}_{t_{n,s+1}}$. For $s = 0$, the particles $\tilde{X}^{1:J}_{t_n}$ representing $p(x_{t_n} \mid y_{1:n-1}) \psi_{t_n}(x_{t_n})$ are propagated with $p(x_{t_{n,1}} \mid x_{t_n})$, and the propagated particles $X^j_{t_{n,1}}$ are resampled according to weights proportional to

$$g_n(y_n \mid \tilde{X}^j_{t_n}) \cdot \frac{\psi_{t_{n,1}}(X^j_{t_{n,1}})}{\psi_{t_n}(\tilde{X}^j_{t_n})}.$$

The resampled particles $\tilde{X}^{1:J}_{t_{n,1}}$ represent $p(x_{t_{n,1}} \mid y_{1:n})\psi_{t_{n,1}}(x_{t_{n,1}})$.

This method combines the L-lookahead approach with the intermediate propagation and resampling approach of Del Moral and Murray (2015). The intermediate propagation and resampling method was considered by Del Moral and Murray (2015) mostly in the context of Markov processes with highly informative observations, such as precisely observed diffusion processes. Both precisely observed diffusion processes and high-dimensional Markov processes with high-dimensional measurements share the property that each observation carries a lot of information, in the sense that the variance of the measurement density $g_n(y_n \mid X_{t_n})$ with respect to $X_{t_n} \sim p(x_{t_n} \mid y_{1:n-1})$ is large. However, in the case of precisely observed low dimensional diffusion processes, the observation y_n can sufficiently localize X_{t_n}, whereas it is often not the case for high-dimensional Markov processes with high-dimensional measurements. The L-lookahead strategy in the GIRF helps localize particles. Also, compared to the L-lookahead method without the intermediate propagation and resampling (i.e., the case of $S = 1$ in the method above), the variance of the resampling weights in the combined approach tends to be much smaller. Intermediate resampling enables the method to use particles more efficiently by focusing on regions in the state space of the latent process that are consistent with future observations.

We consider the case where the dimension of the latent process and the measurement dimension grow linearly with each other. In this case, the number of intermediate steps S can also be chosen to scale linearly with the increasing dimension for good performance. We show in Theorem 2 that, when we can take $\psi_{t_{n,s}}(x_{t_{n,s}})$, $n \in 0:N-1$, $s \in 1:S$ to be the exact forecast likelihood $p(y_{n+1:N} \mid x_{t_{n,s}})$, a bound on the Monte Carlo (MC) error in the estimate $\frac{1}{J}\sum_{j=1}^{J} f(\tilde{X}^{1:J}_{t_N})$ of $\mathbb{E}[f(X_{t_N}) \mid y_{1:N}]$ scales at a polynomial rate with respect to the dimension d under certain circumstances, if we take $S = d$. Due to the $\frac{1}{\sqrt{J}}$ scaling rate of the MC error with respect to the particle size J, the number of particles required to obtain filtering results of desired accuracy also scales at a polynomial rate. In contrast, if intermediate propagation and resampling is not carried out, the number of particles required for a given accuracy typically scales at an exponential rate with respect to d even when the exact forecast likelihoods are available (Snyder et al. 2008, 2015).

Theorem 3 explains a relationship between the quality of the approximation $\psi_{t_{n,s}}(x_{t_{n,s}})$ of $p(y_{n+1:N} \mid x_{t_n,s})$ and the magnitude of the MC error. When the approximation of the forecast likelihood is not exact, a multiplicative factor in our bound on the MC error in Theorem 3 scales exponentially with d. This exponentially scaling factor explains a fundamental limitation in high-dimensional filtering. If there are only a few particles that are consistent with future observations and they are lost in earlier time steps, the accuracies of the sequential particle representations are damaged, until the effect of the lost particles are diluted due to the mixing of the latent process conditional on observed data. Nevertheless, reasonably chosen approximation $\psi_{t_{n,s}}$ can make the multiplicative factor on the bound on the MC error due to the inaccurate approximation of the forecast likelihoods increase at a slow exponential rate. Together with another multiplicative factor that scales polynomially when the number of intermediate propagation and resampling steps S is set equal to d, the number of particles required for a desired accuracy can scale at a slow exponential rate. Our empirical results support that the combination of intermediate propagation and resampling approach and reasonable approximation to forecast likelihoods can significantly extend the dimensionality of the model that is practically accessible. Our theoretical results give a practical suggestion that the number of intermediate steps for propagation and resampling can be chosen equal to the latent process and measurement dimension d.

Based on this guided intermediate resampling approach, we also propose a parameter inference method for implicitly defined, moderately high-dimensional POMP models. Particle filter methods can give unbiased likelihood estimates of the observed data (Del Moral 2004). In high dimensions, the MC errors in the likelihood estimates are typically very high. Nonetheless, the noisy estimates can contain useful information about the parameter. We use the noisy estimates of profile log likelihoods to make an approximate inference for the parameter of interest by taking into account the MC errors in those estimates.

1.1 Related work

We review some of earlier works related to the problem we consider and the method we propose. Since our approach uses lookahead methods for high-dimensional filtering, we describe the separate literatures on these two topics.

1. *High-dimensional filtering*. High-dimensional filtering problems naturally occur when POMP models for spatiotemporal systems are considered. There is a class of local particle filter approaches that use an approximation based on the assumption that the correlation between spatial units decay as the distance between them increases (Farchi and Bocquet 2018). These approaches build upon partitioning of the latent variables into blocks and approximating the one step transitions of the latent process as being independent between the blocks. Rebeschini and Van Handel (2015) developed a theoretical bound for the filtering error, which only depends on the size of the largest block but not on the entire space dimension. Despite this very desirable scaling property, this

approach has some practical limitations, because it is not applicable to highly interdependent spatial models, and the filter estimates are not reliable near the boundaries of the blocks. We note that our GIRF approach does not rely on any assumption on spatial structure and does not suffer from any boundary effects.

Localized data assimilation has also been used with ensemble Kalman filter methods (Hunt et al. 2007). Local ensemble Kalman filter methods use only the observations made near a spatial unit when updating the latent state distribution for that unit. Local implementation can be crucial for the numerical stability of the EnKF, where the number of the particles used J is smaller than the dimension of the model d. We note that particle filters are used for nonlinear, non-Gaussian models that are much lower-dimensional than the locally near-Gaussian models for which local Ensemble Kalman filter methods are used. In this case, the challenge is given by the fact that $J \ll e^d$ but not $J < d$. Local EnKF methods often inflate the one-step forecast variance or the measurement variance or both, as an additional means to ensure numerical stability and to guard against model misspecification (Hunt et al. 2007). The localized filtering approach has also been combined with the ensemble transport method by Reich (2013), Cheng and Reich (2015) and Acevedo et al. (2017).

Beskos et al. (2014a, b) theoretically investigated the approach of using the annealed importance sampling method by Neal (2001) for particle filtering in high dimensions. The annealed importance sampling method introduces a series of bridging distributions between observations. These bridging densities are usually set proportional to a fractional power of the desired target density. Between two adjacent importance resampling, the particles are transformed according to a transition kernel whose stationary distribution equals the target bridging distribution. These transition kernels provide mixing that helps maintain the stability of the particle approximations. The authors gave stability results for the case where the original high-dimensional latent process is composed of many copies of independent and identically distributed (IID) one dimensional processes and the number of bridging steps is equal to the space dimension. In particular, Beskos et al. (2014b) showed that the annealed importance weights are non-degenerate as the dimension goes to infinity even with fixed particle size. Beskos et al. (2014a) showed that both the L^2 error of the filter estimates and the variance of the corresponding likelihood estimates are bounded uniformly in the space dimension. Their approach and our GIRF method have a similarity in the use of intermediate propagation and resampling steps and in the fact that the number of intermediate steps is equal to the space dimension. However, their approach is not applicable to implicitly defined models because analytically tractable transition densities are required.

Beskos et al. (2017) studied a high-dimensional filtering algorithm in the case where the spatial structure of the model can be hierarchically factorized. Specifically, they assumed that the one step transition density is given, or can be well approximated, by a product of functions of increasing collections of latent variables. The theoretical results they obtained by considering a few simple IID cases show that filtering can be stable when the number of particles increases linearly with the space dimension. These promising results provide insights into what might be achieved in more general cases.

2. *The lookahead approach*. A number of particle filter methods proposed in the literature use the information provided by future observations in order to obtain stable filtering estimates. These methods include the auxiliary particle filter by Pitt and Shephard (1999) and the block sampling method by Doucet et al. (2006). Lin et al. (2013) reviewed various lookahead strategies. Johansen (2015) proposed a method based on both the block sampling idea and the annealed importance sampling approach.

The resampling weights in lookahead methods are closely related to approximations to forecast likelihoods. Lin et al. (2010) proposed a method for estimating the optimal resampling weights using backward pilots, in an intermediate propagation and resampling approach for perfectly observed diffusion processes. Guarniero et al. (2017) proposed a method for estimating the exact guide function $\psi_n^*(x_n) = p(y_{n:N} \mid x_n)$ in a backward direction $n = N, N-1, \ldots, 1$, using parametric fitting to mixtures of normals. The filtering on the ψ-twisted models can be iterated to obtain ψ_n functions that gradually approach ψ_n^*. Both of these backward approaches for estimating forecast likelihoods require analytically tractable transition densities of the latent Markov process. In the current paper, we propose a forward-simulation method for approximating the guide function that does not require transition densities to be evaluated.

Serveral works in the literature developed concrete methods for propagating particles according to the adapted kernel $p(x_{n+1} \mid y_{n+1}, x_n)$ in an approximate manner. The implicit particle filter approximates the optimal kernel by directly sampling particles at the vicinity of the maximum of the optimal kernel density (Chorin and Tu 2009; Morzfeld et al. 2012; Chorin et al. 2013). The equivalent-weights particle filter nudges particles toward the next observation over intermediate time steps (Van Leeuwen 2010; Ades and Van Leeuwen 2015); it was developed for applications in geosciences and is based on local Gaussianity of the transition density and the Gaussian measurement density. Papadakis et al. (2010) proposed the use of the ensemble Kalman filter updates as a propagation kernel within a particle filter. Bunch and Godsill (2016) proposed an algorithm that moves particles according to a Gaussian flow that approximates the optimal kernel density. The aforementioned methods assume that the transition density is either known or locally Gaussian.

Table 1 Notation used in the paper and the section each symbol is defined

Notation	Description	Section
$X_t, t \geq t_0$	Continuous-time latent process	2
$y_n, n \in 1:N$	Partial observation at time t_n	2
$g_n(y_n \mid X_{t_n}), n \in 1:N$	Measurement density at y_n given X_{t_n}	2
$t_{n,s}, s \in 1:S-1$	Intermediate time points between t_n and t_{n+1}	2
$\tilde{X}^j_{t_{n,s}}, j \in 1:J$	Filtered particles at $t_{n,s}$	2
$X^j_{t_{n,s}}, j \in 1:J$	Propagated particles at $t_{n,s}$	2
$\hat{\ell}$	Likelihood estimate from Algorithm 3	2
$\psi_{t_{n,s}}$	Guide function at $t_{n,s}$	2
$w_{t_{n,s}}(X^j_{t_{n,s}}, \tilde{X}^j_{t_{n,s-1}})$	Resampling weight for $X^j_{t_{n,s}}$	2
L	Number of lookahead observations used by the guide function	2
$\psi(x; t_{n,s} \to t_{n+b})$	An approximation to the forecast likelihood $p(y_{n+b} \mid X_{t_{n,s}} = x)$	2.1
$\xi(x; t_{n,s} \to t_{n+b})$	Guide simulation from $X_{t_{n,s}} = x$ to t_{n+b}	2.1
$\bar{\xi}(x; t_{n,s} \to t_{n+b})$	Deterministic simulation from $X_{t_{n,s}} = x$ to t_{n+b}	2.1
$\Theta^{m,j}_{t_{n,s}}, j \in 1:J$	Perturbed parameters at $t_{n,s}$ in the m-th iteration of Algorithm 4	5
$\tilde{\Theta}^{m,j}_{t_{n,s}}, j \in 1:J$	Filtered parameters at $t_{n,s}$ in the m-th iteration of Algorithm 4	5

1.2 Summary of contributions

We summarize the contributions of our paper as follows.

- We develop a particle filtering algorithm for moderately high-dimensional, nonlinear, non-Gaussian, implicitly defined, partially observed Markov process models. In particular, the algorithm can be used for models where the latent Markov process has intractable transition density. We demonstrate that our guided intermediate resampling filter (GIRF, Algorithm 3) can be used to enable likelihood-based inference in this class of models. As an example, we make inference for the spatiotemporal coupling parameter in a mechanistic, coupled Markov jump process model describing the metapopulation dynamics of infectious disease (Fig. 3).
- We propose approaches to constructing a guide function using forward simulations (Sect. 2.1). A guide function approximates the forecast likelihood of future observations. The choice of the guide function does not affect the asymptotic consistency of the GIRF algorithm, but does influence its scaling rate as the model dimension increases.
- We develop theoretical results for the GIRF algorithm, including a finite-sample bound on the Monte Carlo filtering error (Theorem 3). These results explain how the Monte Carlo filtering error is influenced by various factors such as the model dimension, the guide function, and the temporal mixing of the latent process conditioned on the observed data. Our results offer insights into why our GIRF algorithm scales more favorably than other particle filter methods that do not emply intermediate propagation and resampling.

1.3 Organization of the paper

This paper is organized as follows. Section 2 explains our intermediate propagation and resampling approach (see Table 1). Section 3 gives empirical results on scaling of the algorithm. Section 4 gives theoretical results. Section 5 describes a parameter estimation procedure that combines the guided intermediate resampling approach with the iterated filtering scheme of Ionides et al. (2015). Section 6 is a concluding discussion.

2 The guided intermediate resampling filter (GIRF)

We denote the latent continuous-time Markov process model by $\{X_t\,; t \geq t_0\}$, where each random variable X_t takes value in a measurable space $(\mathbb{X}, \mathcal{X})$. The measurement process is defined at discrete time points $t_n > t_0$, $n \in 1:N$ and yields an observation $Y_n \in \mathbb{Y}$ that is a noisy or incomplete measurement of X_{t_n}. The measurement Y_n is independent of other observations $Y_m, m \neq n$, and of the latent process $\{X_t\}$, given the current state X_{t_n}. The measurement process for Y_n conditioned on $X_{t_n} = x_{t_n}$ is assumed to have density $g_n(\cdot \mid x_{t_n})$. We will assume that the latent process space and the measurement space are d-dimensional, $\mathbb{X} = \prod_{i=1}^d \mathbb{X}^{[i]}$, $\mathbb{Y} = \prod_{i=1}^d \mathbb{Y}^{[i]}$. We will study the scaling property of our method with respect to d. The observations $Y_n = y_n$ for

$n \in 1:N$ are assumed to be fixed data. In what follows, we assume that the transition kernel of the latent process can be simulated, but we do not require its density to be evaluated.

Algorithm 3: A guided intermediate resampling filter (GIRF)

Input : data, $y_{1:N}$; simulator for $p_{X_{t_0}}$; simulator for $p_{X_{t_{n,s}} \mid X_{t_{n,s-1}}}$; evaluator for the measurement density, $g_n(y_n \mid x_{t_n})$; evaluator for the guide function, $\psi_{t_{n,s}}(x_{t_{n,s}})$; number of particles, J

Output: filtered particle swarm, $\tilde{X}_{t_N}^{1:J}$; likelihood estimate, $\hat{\ell}$

Initialize: $\hat{\ell} \leftarrow 1$, $\tilde{X}_{t_0}^{j} \sim p_{X_{t_0}}(\cdot)$ for $j \in 1:J$
for $n \leftarrow 0:N-1$ **do**
  for $s \leftarrow 1:S$ **do**
    $X_{t_{n,s}}^{j} \sim p_{X_{t_{n,s}} \mid X_{t_{n,s-1}}}(\cdot \mid \tilde{X}_{t_{n,s-1}}^{j})$ for $j \in 1:J$
    $w^j \leftarrow w_{t_{n,s}}(X_{t_{n,s}}^{j}, \tilde{X}_{t_{n,s-1}}^{j})$ given by equation (8) for $j \in 1:J$
    $\hat{\ell} \leftarrow \hat{\ell} \times (\sum_{j=1}^{J} w^j)/J$
    Draw a^j with $\mathbb{P}(a^j = i) = w^i / \sum_{i'=1}^{J} w^{i'}$ for $j \in 1:J$
    Set $\tilde{X}_{t_{n,s}}^{j} = X_{t_{n,s}}^{a^j}$
  end
end

Pseudocode for our guided intermediate resampling filter (GIRF) is given in Algorithm 3. The intermediate time points between t_n and t_{n+1} will be denoted by $t_{n,s}$, $s \in 1:S-1$, and we write $t_{n,0} = t_n$ and $t_{n,S} = t_{n+1}$. The collection of *filtered particles*, $\tilde{X}_{t_{n,s}}^{1:J}$, provide a Monte Carlo representation of a *guided filter distribution* $P_{t_{n,s}}^{\psi}$ given by

$$\frac{\mathrm{d}P_{t_{n,s}}^{\psi}}{\mathrm{d}x_{t_{n,s}}} \propto \psi_{t_{n,s}}(x_{t_{n,s}}) \cdot p(x_{t_{n,s}} \mid y_{1:n}). \tag{7}$$

The filtered particles are moved according to the law of the latent process to construct the *propagated particles*, $X_{t_{n,s+1}}^{1:J}$. The collection of propagated particles is resampled recursively to obtain the next generation of filtered particles. The weighting of the propagated particles is based on the *guide function* $\psi_{t_{n,s}} : \mathbb{X} \to \mathbb{R}^+$ that approximates the forecast likelihood $p(y_{n+1:(n+L\wedge N)} \mid x_{t_{n,s}})$ for some $L \geq 1$, where $n + L \wedge N = \min(n+L, N)$. We require that $\psi_{t_0}(x) = 1$ and $\psi_{t_N}(x) = g_N(y_N \mid x)$ for all $x \in \mathbb{X}$. The assigned importance weight for $X_{t_{n,s}}^{j}$ is given by:

$$w^j \leftarrow w_{t_{n,s}}(X_{t_{n,s}}^{j}, \tilde{X}_{t_{n,s-1}}^{j}) := \begin{cases} \dfrac{\psi_{t_{n,s}}(X_{t_{n,s}}^{j})}{\psi_{t_{n,s-1}}(\tilde{X}_{t_{n,s-1}}^{j})} & \text{if } s \neq 1 \text{ or } n = 0 \\ \dfrac{\psi_{t_{n,s}}(X_{t_{n,s}}^{j}) g_n(y_n \mid \tilde{X}_{t_{n,s-1}}^{j})}{\psi_{t_{n,s-1}}(\tilde{X}_{t_{n,s-1}}^{j})} & \text{otherwise.} \end{cases} \tag{8}$$

If $s = 1$ and $n \geq 1$, that is if $t_{n,s-1}$ is an observation time, we effectively divide the denominator $\psi_{t_{n,s-1}}(\tilde{X}_{t_{n,s-1}}^{j})$ in (8) by $g_n(y_n \mid \tilde{X}_{t_n}^{j})$, because at time $t_{n,1} > t_n$, the past observation y_n should no longer be considered in assessing the fitness of the particle. Particles $X_{t_{n,s}}^{1:J}$ are resampled with probability proportional to these weights. We used systematic resampling for our numerical implementation (Douc et al. 2005). The likelihood of data is defined as

$$\ell_{1:N}(y_{1:N}) = \mathbb{E}\left[\prod_{n=1}^{N} g_n(y_n \mid X_{t_n})\right],$$

where the expectation is taken with respect to the law of $\{X_t ; t \geq t_0\}$. In common with standard particle filters, Algorithm 3 computes a likelihood estimate denoted by $\hat{\ell}$:

$$\hat{\ell} = \prod_{n=0}^{N-1} \prod_{s=1}^{S} \frac{1}{J} \sum_{j=1}^{J} w_{t_{n,s}}(X_{t_{n,s}}^{j}, \tilde{X}_{t_{n,s-1}}^{j}).$$

The GIRF defined by Algorithm 3 is equivalent to the bootstrap particle filter if we take $S = 1$ and $\psi_{t_n}(x_{t_n}) = g_n(y_n \mid x_{t_n})$. Algorithm 3 becomes an instance of APF in the special case where $S = 1$ and $\psi_{t_n}(x_{t_n}) = g_n(y_n \mid x_{t_n}) \cdot g_{n+1}\{y_{n+1} \mid \bar{\xi}_{t_{n+1}}(x_{t_n})\}$, where $\bar{\xi}_{t_{n+1}}(x_{t_n})$ denotes a forecast value for $X_{t_{n+1}}$ given $X_{t_n} = x_{t_n}$. Since APF does not include intermediate resampling, we will find that it does not have the favorable scaling properties that GIRF methodology can enjoy when $S \approx d$.

The computational cost of Algorithm 3 typically scales as $O(JSd)$. The storage cost is $O(Jd)$ since only the current latent process and guide function values need to be saved for each particle during the filtering and propagation recursions. Our implementation of Algorithm 3 is available at https://github.com/joonhap/GIRF.git. A critical scaling question is the rate at which J has to grow with d in order to obtain satisfactory Monte Carlo performance. Numerical results in Sect. 3 show that the MC error in the likelihood estimate and the filtering estimates is reasonably small for moderately large dimensions with feasible number of particles. Our theoretical results in Sect. 4 supports the empirically observed scaling.

2.1 Constructing a guide function

An ideal guide function is the forecast likelihood of all future observations (Whiteley and Lee 2014). Theorem 3 in Sect. 4 will show that a bound on the MC error in filtering estimates is minimized with this guide function. In practice, one can consider approximations to the forecast likelihood of a certain number of future observations for the guide function: for $n \in 0:N-1$, $s \in 1:S$,

$$\psi_{t_{n,s}}(x) \approx p(y_{n+1:n+L} \mid X_{t_{n,s}} = x). \tag{9}$$

Model dependent constructions of the guide function have been proposed for specific latent processes, such as perfectly observed diffusion processes (Lin et al. 2010) or stochastically generated graph models (Bloem-Reddy and Orbanz 2018). Del Moral and Murray (2015) discussed construction of guide function using Gaussian processes. A general, iterative method to construct guide functions that lead to progressively more balanced resampling weights has been proposed in Guarniero et al. (2017). However, supplementary regularization in the construction of the guide function will be necessary for application to high-dimensional models, because methods discussed in Guarniero et al. (2017) rely on approximations using mixtures of normal densities, which become problematic for high-dimensional distributions.

We propose simulation-based approaches for constructing the guide function. These approaches can be used for implicit models for which only the simulator of the latent process is available. The method described below is used in our numerical studies (Sects. 3 and 5.1) for non-Gaussian examples. We will use an approximation to forecast likelihood $p(y_{n+1:n+L} \mid X_{t_{n,s}} = x)$ of the form

$$\psi_{t_{n,s}}(x) = \prod_{b=1}^{\min(L, N-n)} \psi\left(x; t_{n,s} \to t_{n+b}\right)^{\eta(t_{n,s} \to t_{n+b})}, \tag{10}$$

where

$$\psi(x; t_{n,s} \to t_{n+b}) \approx p_{Y_{n+b} \mid X_{t_{n,s}}}\left(y_{n+b} \mid x\right)$$

and $0 \le \eta(t_{n,s} \to t_{n+b}) \le 1$ denotes fractional powers that are non-decreasing as $t_{n,s}$ increases. If $s = S$ and $b = 1$, we set $\psi(x_{t_{n,S}}, t_{n,S} \to t_{n+1}) := g_{n+1}(y_{n+1} \mid x_{t_{n,S}})$ and $\eta(t_{n,S} \to t_{n+1}) = 1$, because the measurement density at $t_{n,S} = t_{n+1}$ can be exactly evaluated. The fact that the powers $\eta(t_{n,s} \to t_{n+b})$ are non-decreasing as $t_{n,s}$ increases may reflect the algorithm user's increasing confidence in the accuracy of the approximated forecast likelihood as the forecast interval becomes shorter. Increasing powers $\eta(t_{n,s} \to t_{n+b})$ as time progresses can also be understood as gradually introducing the information provided by y_{n+L} to the filtering algorithm over the time interval $[t_{n,1}, t_{n+L}]$. We propose a sequence of powers defined as

$$\eta(t_{n,s} \to t_{n+b}) := 1 - \frac{t_{n+b} - t_{n,s}}{\max\{t_{n+b} - t_{\max(n+b-L,0)},\ 2(t_{n+1} - t_n)\}}. \tag{11}$$

The variance of the resampling weights (8) under (10) and (11) can be $\mathcal{O}(1)$ in d, the model dimension. In cases where $n + b - L \ge 0$ and $t_{n+b} - t_{n+b-L} \ge 2(t_{n+1} - t_n)$, the resampling weight for $s \ne 1$ is given by

$$\frac{\psi_{t_{n,s+1}}(X^j_{t_{n,s+1}})}{\psi_{t_{n,s}}(\tilde{X}^j_{t_{n,s}})} = \prod_{b=1}^{L} \frac{\psi(X^j_{t_{n,s+1}}; t_{n,s+1} \to t_{n+b})^{1 - \frac{t_{n+b} - t_{n,s+1}}{t_{n+b} - t_{n+b-L}}}}{\psi(\tilde{X}^j_{t_{n,s}}; t_{n,s} \to t_{n+b})^{1 - \frac{t_{n+b} - t_{n,s}}{t_{n+b} - t_{n+b-L}}}}. \tag{12}$$

If the difference between $\psi(\tilde{X}^j_{t_{n,s}}; t_{n,s} \to t_{n+b})$ and $\psi(X^j_{t_{n,s+1}}; t_{n,s+1} \to t_{n+b})$ is small, (12) is close to

$$\prod_{b=1}^{L} \frac{\psi(\tilde{X}^j_{t_{n,s}}; t_{n,s} \to t_{n+b})^{1 - \frac{t_{n+b} - t_{n,s+1}}{t_{n+b} - t_{n+b-L}}}}{\psi(\tilde{X}^j_{t_{n,s}}; t_{n,s} \to t_{n+b})^{1 - \frac{t_{n+b} - t_{n,s}}{t_{n+b} - t_{n+b-L}}}} = \prod_{b=1}^{L} \psi(\tilde{X}^j_{t_{n,s}}; t_{n,s} \to t_{n+b})^{\frac{t_{n,s+1} - t_{n,s}}{t_{n+b} - t_{n+b-L}}}. \tag{13}$$

If all observation interval is of equal length and the intermediate time points $t_{n,s}$ for $s \in 1 : S-1$ are equally spaced between t_n and t_{n+1}, then (13) is equal to

$$\prod_{b=1}^{L} \psi(\tilde{X}^j_{t_{n,s}}; t_{n,s} \to t_{n+b})^{\frac{1}{LS}},$$

which is approximately on the order of

$$\left[\left\{ \prod_{b=1}^{L} p_{Y_{n+b} \mid X_{t_{n,s}}}(y_{n+b} \mid \tilde{X}^j_{t_{n,s}}) \right\}^{\frac{1}{L}} \right]^{\frac{1}{d}}$$

if $S = d$. Since the predictive likelihoods $p_{Y_{n+b} \mid X_{t_{n,s}}}$ typically scale exponentially in d, raising them to a power of $\frac{1}{d}$ can make the resampling weights (12), and consequently their variance, $\mathcal{O}(1)$ in d. We also found that the powers given by (11) led to good numerical performance of GIRF on the examples we considered. On the contrary, if we set $\eta(t_{n,s} \to t_{n+b}) = 1$ for all $t_{n,s} \le t_{n+b}$, the variance of the resampling weights at $s = 1$ can be noticeably larger than at other intermediate time points because a new term $\psi(x; t_{n,1} \to t_{n+L})$ is suddenly multiplied to the resampling weights at $t_{n,1}$. Setting the denominator in (11) to be at least twice the observation interval, $2(t_{n+1} - t_n)$, ensures that for $L = 1$ and s small, the power $\eta(t_{n,s} \to t_{n+1})$ is at least $\frac{1}{2}$. Otherwise, if $\eta(t_{n,s} \to t_{n+1})$ is too small and $L = 1$, the guide function $\psi_{t_{n,s}}(x) = \psi(x; t_{n,s} \to t_{n+1})^{\eta(t_{n,s} \to t_{n+1})}$ can become too uninformative to guide particles to the regions of the sample space that are consistent with the future observation. In this case, the particles that are not properly guided may have large resampling weight variance at later time steps.

2.1.1 Approximating the forecast likelihood using guide simulations

We propose two ways of obtaining an approximate forecast likelihood $\psi(x\,;t_{n,s}\to t_{n+b})$ in the absence of a closed-form transition density for the latent process.

(i) A moment matching method. We will assume that the measurement density $g_{n+b}(\cdot\,|\,X_{t_{n+b}})$ belongs to a family of densities $\{\breve{g}(\cdot\,|\,\mu,\,\Sigma)\,;\mu,\,\Sigma\}$ that are parameterized by the mean μ and the variance Σ. We denote the mean and the variance by $\mu_{n+b}(X_{t_{n+b}})$ and $\Sigma_{n+b}(X_{t_{n+b}})$:

$$g_{n+b}(\cdot\,|\,X_{t_{n+b}})\equiv\breve{g}\big[\,\cdot\,|\,\mu_{n+b}(X_{t_{n+b}}),\,\Sigma_{n+b}(X_{t_{n+b}})\big].$$

We make a forecast from the current state $X_{t_{n,s}}=x$ to time t_{n+b} using a deterministic skeleton of $\{X_t\}$. A deterministic skeleton is a deterministic process that approximates the conditional mean of the latent process $\{X_t\,;t\geq t_{n,s}\}$ given $X_{t_{n,s}}=x$. This deterministic forecast will be denoted by $\bar{\xi}(x;\,t_{n,s}\to t_{n+b})$. We next approximate the forecast variance of Y_{n+b} given $X_{t_{n,s}}=x$, which can be expressed as

$$\begin{aligned}\mathrm{Var}(Y_{n+b}|X_{t_{n,s}}=x)&=\mathrm{Var}\big(\mathbb{E}[Y_{n+b}|X_{t_{n+b}}]\big|X_{t_{n,s}}=x\big)\\&\quad+\mathbb{E}\big[\mathrm{Var}(Y_{n+b}|X_{t_{n+b}})\big|X_{t_{n,s}}=x\big],\end{aligned}\tag{14}$$

using a collection of J_{G} random forecast simulations for $X_{t_{n+b}}$ from $X_{t_{n,s}}=x$, which we call guide simulations and denote by $\xi_{j_{\mathrm{G}}}(x;\,t_{n,s}\to t_{n+b})$, $j_{\mathrm{G}}\in 1:J_{\mathrm{G}}$. The sample variance of $\mathbb{E}[Y_{n+b}|X_{t_{n+b}}]=\mu_{n+b}(X_{t_{n+b}})$ evaluated at these guide simulations approximates the first term on the right hand side of (14). We denote this sample variance by $\Xi(x;\,t_{n,s}\to t_{n+b})$. The second term on the right of (14) can be approximated by $\Sigma_{n+b}\big(\bar{\xi}(x;\,t_{n,s}\to t_{n+b})\big)$. We then approximate the forecast likelihood of $Y_{n+b}=y_{n+b}$ given $X_{t_{n,s}}=x$ by

$$\begin{aligned}\psi(x;\,t_{n,s}\to t_{n+b})&=\breve{g}\,\big[y_{n+b}\,\big|\,\mu_{n+b}\big(\bar{\xi}(x;\,t_{n,s}\to t_{n+b})\big),\\&\quad\Sigma_{n+b}\big(\bar{\xi}(x;\,t_{n,s}\to t_{n+b})\big)+\Xi(x;\,t_{n,s}\to t_{n+b})\big].\end{aligned}\tag{15}$$

One may use (15) for measurement processes without well-defined first and second moments, if the measurement noise is additive and the measurement process belongs to a family that is closed under independent sums, such as the Cauchy distribution. We view the parameters μ and Σ of the family $\{\breve{g}(\cdot\,|\,\mu,\,\Sigma)\}$ as representing the center and the variability of the distributions respectively. For two independent random variables X_1 and X_2 with densities $\breve{g}(\cdot\,|\,\mu,\,\Sigma)$ and $\breve{g}(\cdot\,|\,0,\,\Sigma')$ respectively, we suppose that X_1+X_2 has density $\breve{g}(\cdot\,|\,\mu,\,\Sigma+\Sigma')$. The forecast variability $\Xi(x;\,t_{n,s}\to t_{n+b})$ may be approximated by, for example, a value for which the distribution with density $\breve{g}\,\big(\cdot\,|\,0,\,\Xi(x;\,t_{n,s}\to t_{n+b})\big)$ has the same inter-quantile distance as the sample inter-quantile distance of the random forecasts.

Often times, the measurement process of a spatiotemporal POMP model is local, in the sense that the measurement in the i-th spatial unit depends only on the state of the same unit. In such cases, the measurement density can be expressed as

$$g_n(y_n\,|\,x_{t_n})=\prod_{i=1}^{d}g_n^{[i]}(y_n^{[i]}\,|\,x_{t_n}^{[i]}).\tag{16}$$

If each local measurement density $g_n^{[i]}$ belongs to a family $\{\breve{g}^{[i]}(\cdot\,|\,\mu^{[i]},\,\Sigma^{[i]})\}$, we may take

$$\begin{aligned}&\psi(x;\,t_{n,s}\to t_{n+b})\\&\quad=\prod_{i=1}^{d}\breve{g}^{[i]}\Big[y_{n+b}^{[i]}\,\Big|\,\mu_{n+b}^{[i]}\big\{\bar{\xi}^{[i]}(x;\,t_{n,s}\to t_{n+b})\big\},\\&\qquad\Sigma_{n+b}^{[i]}\big\{\bar{\xi}(x;\,t_{n,s}\to t_{n+b})\big\}+\Xi^{[i]}(x;\,t_{n,s}\to t_{n+b})\Big],\end{aligned}\tag{17}$$

where $\bar{\xi}^{[i]}(x;\,t_{n,s}\to t_{n+b})$ is the i-th component of the deterministic forecast and $\Xi^{[i]}(x;\,t_{n,s}\to t_{n+b})$ is the sample variance of $\mu_{n+b}^{[i]}$ evaluated at the guide simulations. We note that $\bar{\xi}^{[i]}(x;\,t_{n,s}\to t_{n+b})$ is obtained by simulating the deterministic skeleton jointly for all dimensions, and also $\Xi^{[i]}(x;\,t_{n,s}\to t_{n+b})$ by simulating the joint random latent process. Thus $\psi(x;\,t_{n,s}\to t_{n+b})$ constructed by (17) makes some allowance for the correlation of the latent process between dimensions. The forecast likelihood approximated this way can be reasonably accurate when the variances of the independent measurement processes in (16) are larger than the covariance of the guide simulations between the spatial components.

We note that one can save computational effort by using locally linear approximations for the forecast variability. Suppose that for $t\in(t_{n,s},\,t_{n+b})$ the ancestor of a particle X_t^j is $X_{t_{n,s}}^{j'}$. One may approximate the forecast variability from t to t_{n+b} for particle X_t^j as

$$\begin{aligned}&\Xi(X_t^j;\,t\to t_{n+b})\\&\quad\approx\Xi(X_{t_{n,s}}^{j'};\,t_{n,s}\to t_{n+b})\cdot\frac{t_{n+b}-t}{t_{n+b}-t_{n,s}}.\end{aligned}\tag{18}$$

The forecast variability can be re-estimated using new random forecasts at each $t_{n,1}$, $n\in 1:N-1$, or more often if the locally linear approximation becomes unreliable.

(ii) A quantile-based method. The second method uses the sample quantiles of the guide simulations. For some $K>1$ and for $k\in 1:K$, let $\hat{q}_k^{[i]}(x;\,t_{n,s}\to t_{n+b})$ denote the sample quantile corresponding to the cumulative probability of $\frac{k-0.5}{K}$

for the i-th component of the guide simulations for $X_{t_{n+b}}$ given $X_{t_{n,s}} = x$. We then define the guide function as

$$\psi(x; t_{n,s} \to t_{n+b}) = \prod_{i=1}^{d} \frac{1}{K} \sum_{k=1}^{K} g_{n+b}^{[i]}\left[y_{n+b}^{[i]} | \hat{q}_k^{[i]}(x; t_{n,s} \to t_{n+b})\right]. \tag{19}$$

The number K of sample quantile values can be chosen such that at least one of the sample quantiles belong to the effective support of the measurement likelihood function $g_{n+b}^{[i]}(y_{n+b}^{[i]} | \cdot)$. Similarly to the moment-matching method, the guide simulations can be made only at a small fraction of the intermediate time points. Suppose again that for $t \in (t_{n,s}, t_{n+b})$ the ancestor of X_t^j is $X_{t_{n,s}}^{j'}$. Under the same assumption that the forecast variance increases approximately linearly in the forecast time length, we can approximate the k-th quantile of the forecast distribution of $X_{t_{n+b}}^{[i]}$ given X_t^j as

$$\hat{q}_k^{[i]}(X_t^j; t \to t_{n+b}) \approx \bar{\xi}^{[i]}(X_t^j; t \to t_{n+b}) + \left(\hat{q}_k^{[i]}(X_{t_{n,s}}^{j'}; t_{n,s} \to t_{n+b}) - \bar{\xi}^{[i]}(X_{t_{n,s}}^{j'}; t_{n,s} \to t_{n+b})\right) \cdot \sqrt{\frac{t_{n+b} - t}{t_{n+b} - t_{n,s}}}, \tag{20}$$

where $\bar{\xi}^{[i]}(x; t \to t_{n+b})$ is the i-th component of the deterministic forecast for $X_{t_{n+b}}$ given $X_t = x$. We point out the case of using all guide simulations, that is, letting K equal to the number of guide simulations J_{G} and replacing $\hat{q}_k^{[i]}(x; t_{n,s} \to t_{n+b})$ in (19) and (20) by

$$\tilde{\xi}_{j_{\mathrm{G}}}(X_t^j; t \to t_{n+b}) = \bar{\xi}(X_t^j; t \to t_{n+b}) + \left(\xi_{j_{\mathrm{G}}}(X_{t_{n,s}}^{j'}; t_{n,s} \to t_{n+b}) - \bar{\xi}(X_{t_{n,s}}^{j'}; t_{n,s} \to t_{n+b})\right) \cdot \sqrt{\frac{t_{n+b} - t}{t_{n+b} - t_{n,s}}} \tag{21}$$

for $j_{\mathrm{G}} \in 1 : J_{\mathrm{G}}$. This can be particularly useful in the case where each local latent process $\{X_t^{[i]}\}$ is multi-dimensional. In this case, ordering the vectors $\xi_{j_{\mathrm{G}}}^{[i]}(X_{t_{n,s}}^j; t_{n,s} \to t_{n+b})$, $j_{\mathrm{G}} \in 1 : J_{\mathrm{G}}$, to compute sample quantiles may not be straightforward, but using all guide simulations in (19) removes the need for ordering.

2.1.2 Dealing with the correlation between spatial units

The two approaches discussed in Sect. 2.1.1 approximates the forecast likelihood $p_{Y_{n+b}|X_{t_{n,s}}}(y_{n+b} \mid x)$ by the product of terms approximating $p_{Y_{n+b}^{[i]}|X_{t_{n,s}}}(y_{n+b}^{[i]} \mid x)$ for $i \in 1 : d$ under the assumption of spatially local, independent measurements (16). We now consider the case where (16) is not satisfied. Specifically, we address two sources of correlation between $\{Y_{n+b}^{[i]}; , i \in 1 : d\}$ conditional on $X_{t_{n,s}}$. First, $Y_{n+b}^{[i]}$ may not depend only on $X_{t_{n+b}}^{[i]}$ but also the other components $X_{t_{n+b}}^{[i']}$, $i' \neq i$. Second, the measurement processes for $Y_{n+b}^{[i]}, i \in 1 : d$, conditional on $X_{t_{n+b}}$ may not be independent of each other. We propose a Monte Carlo approximation of the forecast likelihood using guide simulations in the the case where the measurement density can be expressed as

$$g_{n+b}(y_{n+b} \mid X_{t_{n+b}}) = \mathbb{E}_Z \prod_{i=1}^{d} \tilde{g}_{n+b}^{[i]}\left[y_{n+b}^{[i]}; h(X_{t_{n+b}}, Z), X_{t_{n+b}}^{[i]}\right], \tag{22}$$

where Z is a random variable that induces correlation between local measurement processes and h and $\tilde{g}_{n+b}^{[i]}$, $i \in 1 : d$ are some functions. We assume that the random variable Z can be simulated and that it is independent of $\{X_t\}$. Given $X_{t_{n,s}}^j$, we make J_{G} guide simulations $\xi_{j_{\mathrm{G}}}(X_{t_{n,s}}^j; t_{n,s} \to t_{n+b})$ for $j_{\mathrm{G}} \in 1 : J_{\mathrm{G}}$ and simulate Z_{j_Z} for $j_Z \in 1 : J_Z$ according to the law of Z. We order the values $h(\xi_{j_{\mathrm{G}}}, Z_{j_Z})$ for $j_{\mathrm{G}} \in 1 : J_{\mathrm{G}}$ and $j_Z \in 1 : J_Z$ and partition $(1 : J_{\mathrm{G}}) \times (1 : J_Z)$ into $\mathcal{K}_k$, $k \in 1 : K$, such that each $\mathcal{K}_k$ has the same size and that $h(\xi_{j_{\mathrm{G}}}, Z_{j_Z}) \leq h(\xi_{j'_{\mathrm{G}}}, Z_{j'_Z})$ whenever $(j_{\mathrm{G}}, j_Z) \in \mathcal{K}_k$, $(j'_{\mathrm{G}}, j'_Z) \in \mathcal{K}_{k'}$, and $k < k'$. We can then approximate the forecast likelihood $p_{y_{n+b}|X_{t_{n,s}}}(y_{n+b} \mid X_{t_{n,s}}^j)$ by

$$\frac{1}{K} \sum_{k=1}^{K} \prod_{i=1}^{d} \left\{ \frac{1}{|\mathcal{K}_k|} \sum_{(j_{\mathrm{G}}, j_Z) \in \mathcal{K}_k} \tilde{g}_{n+b}^{[i]}\left[y_{n+b}^{[i]}; h(\xi_{j_{\mathrm{G}}}, Z_{j_Z}), \xi_{j_{\mathrm{G}}}^{[i]}\right] \right\}, \tag{23}$$

where $|\mathcal{K}_k|$ denotes the size of $\mathcal{K}_k$. We note that at the intermediate time points where the guide simulations are not made, the $\xi_{j_{\mathrm{G}}}$ in (23) can be replaced by the approximations $\tilde{\xi}_{j_{\mathrm{G}}}$ in (21). The approximation (23) is motivated by the expression

$$p_{Y_{n+b}|X_{t_{n,s}}}(y_{n+b} \mid X_{t_{n,s}}^j) = \mathbb{E}\left[\mathbb{E}\left\{\prod_{i=1}^{d} \tilde{g}_{n+b}^{[i]}\left[y_{n+b}^{[i]}; h(X_{t_{n+b}}, Z), X_{t_{n+b}}^{[i]}\right] | h(X_{t_{n+b}}, Z)\right\} | X_{t_{n,s}} = X_{t_{n,s}}^j\right]. \tag{24}$$

There is a bias-variance tradeoff associated with the choice of K. Since (23) is an average of products of d terms, its value will likely be determined by one of the partitions giving the largest product. Therefore the Monte Carlo variance of (23) can scale linearly with K, because effectively only $\frac{1}{K}$ of the simulations are used. On the other hand, if K is small, the values of $h(\xi_{j_{\mathrm{G}}}, Z_{j_Z})$ within each partition can have a large range, and the average over the partition can have a large bias with respect to inner conditional expectation in (24).

We show two examples that belong to the class of measurement models described in (22).

(i) Correlated measurement noise. The first example is a measurement model with correlated noise given by

$$Y_{n+b}^{[i]} = X_{t_{n+b}}^{[i]} + Z + \epsilon^{[i]}, \quad i \in 1:d,$$

where Z is a common noise term and $\epsilon^{[i]}$ are independent measurement noises specific to the i-th spatial unit. This corresponds to the case of $h(X_{t_{n+b}}, Z) = Z$ in (22). In this case, each partition $\mathcal{K}_k$ consists of the values of Z_{j_Z} within a certain range paired with all guide simulations.

(ii) A global latent process parameterizing the measurement process. The second example concerns the case where there is a component in the latent process, $\{X_t^{[i_0]}\}$, which affects all local measurement processes that are independent of one another:

$$g_n(y_n \mid X_{t_n}) = \prod_{i=1}^{d} \tilde{g}_n^{[i]}(y_n^{[i]}; X_{t_n}^{[i_0]}, X_{t_n}^{[i]}).$$

This corresponds to the case where $h(X_{t_{n+b}}, Z) = X_{t_{n+b}}^{[i_0]}$ in (22). Being a global process parameterizing all local measurement processes, $X_t^{[i_0]}$ may have no local measurement process for itself, but we may formally write the measurement density for the i_0-th component as $g_n^{[i_0]}(y_* \mid X_{t_n}) \equiv 1$ for an arbitrary observation value y_*. The approximation of forecast likelihood by (23) involves the partitioning of $\xi_{j_G}^{[i_0]}$, with no Z component.

3 Numerical examples

In this section, we apply the GIRF to two examples. We investigate the empirical scaling properties of an implementation of GIRF compared to alternative methods. More numerical results that demonstrate the practical utility of the GIRF approach in parameter estimation are given in Sect. 5.1. In all our examples, the number of intermediate sub-intervals S is set equal to the space dimension d.

3.1 Correlated Brownian motion

We first applied our algorithm to a multi-dimensional correlated Brownian motion. Each component of the Brownian motion was identically distributed with increments per unit time having mean zero and unit variance. The correlation coefficient matrix A for the increments was chosen such that its all off-diagonal entries equaled α. The initial latent distribution at time $t_0 = 0$ was given by the point mass at the origin of $\mathbb{R}^d$. Measurements were made at positive integer time points $t_{1:50} = 1:50$, with independent Gaussian noises of mean zero and unit variance. The POMP model can be expressed as follows, where I denotes the d dimensional identity matrix:

$$X_{t+\delta} = X_t + \mathcal{N}(0, \delta A),$$
$$Y_n = X_{t_n} + \mathcal{N}(0, I).$$

The guide function $\psi_{t_{n,s}}$ was defined as in (10), where $L = 2$ or 3, and $\eta_{t_{n,s},t_{n+b}}$ were taken as in (11). Since the process had zero drift, the forward state projection by the deterministic mean process was given by $\mu_{t_{n+b}}(x_{t_{n,s}}) = x_{t_{n,s}}$. The variance of $X_{t_{n+b}}$ conditioned on $X_{t_{n,s}} = x_{t_{n,s}}$ was equal to $(t_{n+b} - t_{n,s}) \cdot A$, so the guide function was defined as

$$\psi_{t_{n,s}}(x_{t_{n,s}}) = \prod_{b=1}^{L} \phi_d\left[y_{t_{n+b}}; x_{t_{n,s}}, (t_{n+b} - t_{n,s}) \cdot A + I\right]^{\eta(t_{n,s} \to t_{n+b})}, \tag{25}$$

where $\phi_d(\cdot; \mu, \Sigma)$ denotes the density of the d-dimensional Gaussian distribution with mean μ and variance Σ. Evaluating (25) typically requires procedures such as the Cholesky decomposition and takes $O\left(d^3\right)$ computations. Since this could be demanding for large d, we also used an approximation of (25) obtained by ignoring the off-diagonal elements of A,

$$\psi_{t_{n,s}}(x_{t_{n,s}}) = \prod_{b=1}^{L} \phi_d\left[y_{t_{n+b}}; x_{t_{n,s}}, (t_{n+b} - t_{n,s}) \cdot I + I\right]^{\eta(t_{n,s} \to t_{n+b})}. \tag{26}$$

We first compared the filtering performance of the auxiliary particle filter (APF), 2-lookahead filter, and the GIRF with $L = 2$ and 3 for varying dimensions $d = 5, 20, 50, 100, 200$, and 500. The correlation coefficient was fixed at $\alpha = 0$. The APF was implemented by setting $S = 1$ and $L = 2$ in Algorithm 3, and the 2-lookahead filter by setting $S = 1$ and $L = 3$. The GIRF method used two thousand particles for all models. The APF and the 2-lookahead filter used d times as many particles, so that the computation time would be similar for all methods. For the APF and the 2-lookahead filter, we used a parallelized version of Algorithm 3, following the island particle filter approach of Vergé et al. (2015), for models with $d \geq 50$ in order to avoid memory deficiency. In these cases, the particles were divided into $d/10$ islands. For $d = 500$, we could not run the APF and the 2-lookahead filter with $2000d$ particles even after parallelization, due to insufficient memory. Each experiment was independently repeated

for twenty times. All experiments were carried out using our C++ implementation. The computational resources used and the numerical results averaged over twenty repetitions are shown in Table 2. The exact likelihood of the data and the exact filtering distributions were computed using the Kalman filter. We compared the log likelihood estimates ($\log \hat{\ell}$) and the mean squared errors of the estimated filter means at the terminal time t_{50} averaged over all d components (MSFE). All experiments in Sect. 3.1 were conducted on the Boston University Shared Computing Cluster.

The numerical results showed that the performance of the methods that did not use intermediate propagation and resampling steps, namely the APF and the 2-lookahead filter, decayed rapidly with dimension beyond $d = 20$. In contrast, GIRF produced relatively accurate estimates of the likelihoods and the filter means in much higher dimensions. In particular, the error in the Monte Carlo likelihood estimate for $d = 200$ was only 23 log units for $L = 3$. The mean squared filter errors by GIRF were also relatively small compared to the marginal variance of the filtering distribution at the terminal time, $\mathrm{Var}(X_{t_{50}}^{[1]} \mid y_{1:50})$, which was equal to 0.62 for all models with different dimensions. The mean squared filter errors by GIRF scaled roughly at a polynomial rate up to $d = 500$. In contrast, the mean squared filter errors by the APF and the 2-lookahead filter were much greater than the filter variances beyond $d = 20$. Snyder et al. (2008) reported that a standard bootstrap particle filter would require at least 10^{11} particles for the same filtering problem in two hundred dimension in order to obtain filter mean estimates that are even less accurate than our GIRF estimates shown in Table 2. In contrast, our GIRF estimates were obtained using only two thousand particles. We remark that we also tried taking all $\eta_{t_{n,s},t_{n+b}}$ in (10) equal to the unity regardless of b; in this case the GIRF also scaled substantially better than the APF, but the performance of the GIRF was somewhat worse than when we took $\eta_{t_{n,s},t_{n+b}}$ as in (11) (results not shown).

Next, we investigated varying the correlation coefficient α of the Brownian motion. The dimension was fixed at $d = 100$, and the correlation coefficient varied from 0 to 0.5. We used GIRF with $S = 100$, $L = 3$, and two thousand particles. Twenty independent filter runs were carried out for each value of correlation coefficient. Table 3 shows the errors in the log of the estimated likelihoods and the mean squared filter errors at the terminal time averaged over d components. All results were averaged over twenty independent filter runs. We used both the guide function with the exact covariance as in (25) and the guide function with diagonal covariance as in (26). When the exact covariance was used, the Monte Carlo errors in both the likelihood estimates and the filter means were relatively constant or slowly increased as the correlation coefficient α increased. When the diagonal covariance was used, the Monte Carlo errors increased more rapidly as α increased, due to inaccurate approximation of the forecast likelihood by the guide function. However, the GIRF runs still produced reasonable MC estimates using only two thousand particles in one hundred dimension even when the diagonal covariance approximation was used for $\alpha = 0.5$: the mean squared filter errors were about 0.14, which was less than the marginal variance of the filtering distribution at the terminal time, which was 0.50. Considering that the diagonal covariance approximation differs significantly from the exact covariance in the case of $\alpha = 0.5$ and the fact that a one hundred dimensional model is well beyond the practically accessible range by the standard particle filters, we see that the GIRF method can be relatively robust with respect to inaccurate approximation to forecast likelihoods even in moderately high dimensions.

3.2 Stochastic Lorenz 96 model

The Lorenz 96 model is a nonlinear chaotic system which provides a simplified representation of global atmospheric circulation (Lorenz 1996). Stochastic versions of this model have been used to support the increased use of non-deterministic models for atmospheric science (Wilks 2005; Palmer 2012). We considered a stochastic Lorenz 96 model with added Gaussian process noise, defined as follows:

$$\begin{aligned} \mathrm{d}X_t^{[i]} = \{&(X_t^{[i+1]} - X_t^{[i-2]}) \cdot X_t^{[i-1]} \\ &- X_t^{[i]} + F\}\mathrm{d}t + \sigma_p \mathrm{d}B_t^{[i]}, \qquad i \in 1:d. \end{aligned} \tag{27}$$

In the equation above, we understand that $X^{[0]} = X^{[d]}$, $X^{[-1]} = X^{[d-1]}$, and $X^{[d+1]} = X^{[1]}$. The terms $\{B_t^{[i]}; i \in 1:d\}$ denote d independent standard Brownian motions, and σ_p the process noise magnitude. F is a forcing constant, with $F = 8$ considered by Lorenz (1996) to induce chaotic behavior. The system is started at the initial state $X_0^{[i]} = 0$ for $i \in 1:d-1$ and $X_0^{[d]} = 0.01$. Observations are independently made for each spatial unit at $t_n = \Delta_{\mathrm{obs}} \cdot n$ for $n \in 1:200$, where Δ_{obs} is either 0.1 or 0.5. The measurement noise is normally distributed with mean zero and standard deviation σ_m. We generated data for $d = 4$ and $d = 50$ with $F = 8$ and $\sigma_p = \sigma_m = 1$, using the Euler-Maruyama method for numerical approximation of the sample paths of X_t with time increments of 0.01.

We compared our implementation of GIRF with an ensemble Kalman filter (EnKF) for the generated data. Our GIRF implementation used the guide function constructed via forty guide simulations, according to the quantile-based method (19) and (20) with $L = 2$ and $K = 8$. The guide simulations were made at every observation time when Δ_{obs} was 0.1 and at every time interval of 0.25 when Δ_{obs} was 0.5. The likelihood of data was also estimated from EnKF using the Gaussian approximation to the empirical distribution of the particle swarm using the sample mean and the sample vari-

Table 2 : Comparison between the auxiliary particle filter, 2-lookahead method, and the GIRF with $L = 2$ and $L = 3$ for the correlated Brownian motion

Method	Total no. of particles	S	L	CPU time (sec)					
				$d = 5$	$d = 20$	$d = 50$	$d = 100$	$d = 200$	$d = 500$
(a) Computational costs									
APF	$2000 \times d$	1	2	1	13	102	382	1397	–
2-lookahead	$2000 \times d$	1	3	1	15	139	474	1874	–
GIRF ($L = 2$)	2000	d	2	1	10	55	206	814	4990
GIRF ($L = 3$)	2000	d	3	1	12	68	294	1060	6416

	APF ($S = 1,\ L = 2$)	2-lookahead ($S = 1,\ L = 3$)	GIRF ($S = d,\ L = 2$)	GIRF ($S = d,\ L = 3$)	Kalman filter $\log \ell$
(b) Difference between the log of the average of twenty likelihood estimates and the exact log likelihood ($\log \hat{\ell} - \log \ell$), the standard deviation of twenty log likelihood estimates ($s.d.(\log \hat{\ell})$), and the mean squared filter error (MSFE) calculated as the squared error of the estimated filter means at terminal time averaged over d components and over twenty repetitions. The exact log likelihoods ($\log \ell$) and filter means were computed using the Kalman filter.					
$d = 5$ $\log \hat{\ell} - \log \ell$	-0.001	-0.07	-0.32	-0.06	-485.6
$s.d.(\log \hat{\ell})$	(0.53)	(0.46)	(0.49)	(0.62)	
MSFE	0.0003	0.0002	0.0008	0.0008	
$d = 20$ $\log \hat{\ell} - \log \ell$	-37.3	-24.8	-1.1	+0.26	-1904.0
$s.d.(\log \hat{\ell})$	(9.1)	(8.6)	(1.1)	(0.86)	
MSFE	0.15	0.17	0.007	0.006	
$d = 50$ $\log \hat{\ell} - \log \ell$	-1366	-1146	-5.6	-0.6	-4790.2
$s.d.(\log \hat{\ell})$	(144)	(119)	(5.4)	(1.8)	
MSFE	1.9	1.7	0.033	0.018	
$d = 100$ $\log \hat{\ell} - \log \ell$	-7096	-6717	-73	-7.7	-9499.1
$s.d.(\log \hat{\ell})$	(424)	(366)	(10)	(3.4)	
MSFE	4.0	3.8	0.08	0.04	
$d = 200$ $\log \hat{\ell} - \log \ell$	-30688	-29544	-277	-23	-18909
$s.d.(\log \hat{\ell})$	(1323)	(1333)	(27)	(7.2)	
MSFE	8.8	8.2	0.15	0.10	
$d = 500$ $\log \hat{\ell} - \log \ell$	–	–	-1282	-162	-47415
$s.d.(\log \hat{\ell})$	–	–	(56)	(16)	
MSFE	–	–	0.31	0.22	

Table 3 Difference between the log of averaged likelihood estimates and the exact likelihood, the standard deviation of log likelihood estimates, and the mean squared filter errors for $d = 100$ dimensional models with varying degrees of correlation

Correlation coefficient		0.0	0.1	0.2	0.3	0.4	0.5
Kalman	$\log \ell$	−9499	−9431	−9322	−9198	−9059	−8905
GIRF	$\log \hat{\ell} - \log \ell$	−7.7	−1.8	−4.0	−7.7	−15	−20
[exact covariance]	$s.d.(\log \hat{\ell})$	(3.4)	(4.7)	(6.0)	(5.3)	(6.1)	(6.6)
	MSFE	0.04	0.03	0.03	0.03	0.04	0.04
GIRF	$\log \hat{\ell} - \log \ell$	−7.7	−36	−99	−183	−273	−373
[diag covariance]	$s.d.(\log \hat{\ell})$	(3.4)	(6.4)	(9.2)	(12)	(16)	(28)
	MSFE	0.04	0.05	0.08	0.13	0.13	0.14

Exact log likelihoods of data are shown in the first row. Results for both the guide function using the exact covariance (25) and that using the diagonal covariance (26) are shown

ance. For a model with $d = 4$, we also ran the bootstrap particle filter (BPF). We ran each method using 400, 2000, and 10,000 particles. The experiments with 10,000 particles for the BPF and for the GIRF ran five particle islands each comprising 2000 particles. We used our C++ implementation for GIRF, and the BPF was implemented as a GIRF with $S = 1$ and $L = 1$. For the EnKF, we used the `enkf` function in R package **pomp**, which speeds up computations using C snippet declarations (King et al. 2016, 2019).

Figure 1 shows the log likelihood estimates by each method. When the observations were made at intervals of $\Delta_{\text{obs}} = 0.5$, the likelihood estimates by GIRF were higher than those by the EnKF. This was due to the fact that the EnKF made Gaussian approximations to one-step forecast distributions $p(X_{t_n}|y_{1:n-1})$, which were moderately non-Gaussian. The likelihood estimates for the $d = 50$ dimensional model by GIRF using 400 particles, which took 27 min, was higher than those by the EnKF using 10,000 particles, which took 5.5 min. The likelihood estimates by the EnKF showed a bias that did not go away as the number of particles increased. For $d = 4$, the likelihood estimate by GIRF agreed with those by the BPF, which may be considered as a benchmark when filtering for low dimensional models. The results for $\Delta_{\text{obs}} = 0.5$ show that the GIRF can give better numerical results than the EnKF for nonlinear, non-Gaussian models for which one can construct a guide function that reasonably approximates forecast likelihoods.

When Δ_{obs} was 0.1 instead, the EnKF produced good results relative to the GIRF. This was because our stochastic Lorenz 96 model behaved like a linear Gaussian model for this shorter observation time interval and the one-step forecast distribution $p(x_{t_n}|y_{1:n-1})$ could be well approximated by a Gaussian distribution. For $d = 4$, the GIRF, the BPF, and the EnKF gave likelihood estimates that were close to each other, but the EnKF scaled better to $d = 50$ than the GIRF. We remark that a longer observation time interval posed difference challenges for GIRF and the EnKF. For the GIRF, the deterministic forecast simulations became less reliable as the forecast time length increased due to the chaotic property of the Lorenz 96 model. For the EnKF, the one-step forecast distribution became increasingly non-Gaussian as the observation time interval increased due to the nonlinearity of the model. This made local data assimilation, which was based on the assumption that the one-step forecast distribution was Gaussian, less accurate.

For the $d = 50$ dimensional model, we also ran a local ensemble Kalman filter (LEnKF) (Hunt et al. 2007). Our implementation of the LEnKF used the observations at three neighboring spatial units on each side for a total of seven observations $y_n^{[i-3:i+3]}$ to update the i-th coordinate of the particles. It also inflated the sample variance of the proposed particles by a factor of 1.1 by linearly perturbing the particles away from their sample mean. Local implementation of the EnKF is commonly used to improve the numerical results in geophysical models where the dimension can be much higher than the number of particles. In our relatively low-dimensional examples, however, both local implementation and variance inflation did not improve the numerical results.

4 Theoretical results

We first show that the standard results for SMC apply to the GIRF defined in Algorithm 3. GIRF can be cast into the standard framework of particle filters by extending the latent space to $\mathbb{X}^2$ where the new latent variable is the pair $(X_{t_{n,s}}, X_{t_{n,s-1}})$. This extension is necessary because the resampling weights (8) depend on both $X^j_{t_{n,s}}$ and $\tilde{X}^j_{t_{n,s-1}}$. Likelihood estimates obtained from the standard particle filter are unbiased (Del Moral and Jacod 2001). It follows that the likelihood estimates from GIRF are also unbiased for any guide function $\psi_{t_{n,s}} : \mathbb{X} \to \mathbb{R}^+$.

Theorem 1 *The likelihood estimate $\hat{\ell}$ of Algorithm 3 is unbiased for $\ell_{1:N}(y_{1:N})$.*

Proof See Section S1 in the supplementary material. □

The consistency and the asymptotic normality of the filter estimates from GIRF also follow naturally from the stan-

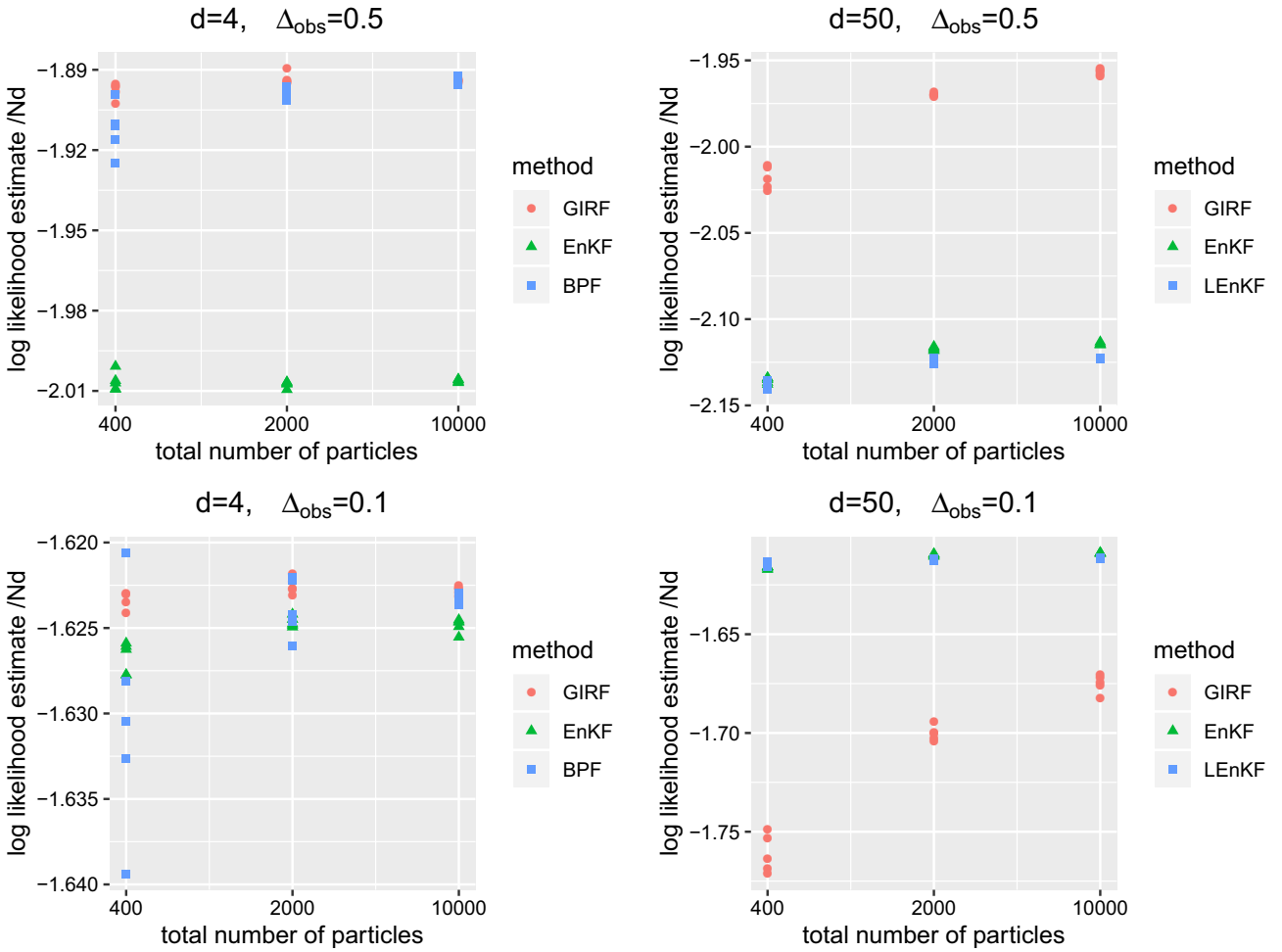

Fig. 1 : Log likelihood estimates per spatial unit per time ($\frac{\hat{\ell}}{N \times d}$) by GIRF, EnKF, a local ensemble Kalman filter (LEnKF), and the bootstrap particle filter (BPF) for stochastic Lorenz 96 examples with various dimensions and observation time intervals

dard particle filter theory (Chopin 2004; Del Moral 2004). The results of the unbiasedness of likelihood estimates and the consistency of filtering distribution have been given for methods with intermediate resampling, but the scaling properties with respect to increasing dimension have not been established (Del Moral and Murray 2015; Bloem-Reddy and Orbanz 2018). In what follows, we examine the scaling properties of GIRF.

GIRF converts a filtering problem with highly informative observations into one that deals with a slower rate of incoming information, at the expense of operating on a refined time scale. There are many results in the literature which concern the stability of particle filters (see for example, Del Moral and Guionnet 2001; Del Moral 2004; Le Gland and Oudjane 2004; Whiteley 2013; Giraud and Del Moral 2017). However, these results do not directly address the scaling with respect to increasing dimension. Another major issue in applying these results to the "infill" scenario we study in which the number of intermediate time steps S is increasing is that the number of time steps needed for the mixing of the latent process conditional on data increases proportionally with S. We provide a novel theoretical analysis of the scaling rate when the number of intermediate time steps grow linearly with the amount of information each observation carries, which in turn increases with the model dimension. In particular, we provide a finite sample bound on the filtering error (Theorem 3) and asymptotic bounds on the variance of the likelihood estimate (Theorem 4) and filter estimates (Theorem 5) for GIRF. These bounds show how intermediate propagation and resampling and the guide function can remedy the otherwise problematic dimensional scaling properties of particle filters.

4.1 Scaling properties when the guide function is exact

Given the observations $y_{1:N}$ and for $t_n < t \le t_{n+1}$, we initially consider the situation where the guide function matches the forecast likelihood of all future observations:

$$\psi_t^*(x_t) := p_{Y_{n+1:N}|X_t}(y_{n+1:N} \mid x_t). \tag{28}$$

This is called the exact guide function. We will show that the number of particles required for accurate filtering can scale polynomially in dimension d under some assumptions if the exact guide function is used and $S = d$. Since the exact guide function is not generally computationally tractable, a theory for inexact guide functions will be developed in Sect. 4.2.

Assumption 1* *There exists $C_1^* \ge 1$ such that for every $s \in 1:S$, $n \in 0:N-1$, and $x \in \mathbb{X}$,*

$$\frac{\mathrm{Var}\left[p(y_{n+1:N} \mid X_{t_{n,s}}) \,\middle|\, X_{t_{n,s-1}} = x\right]}{p(y_{n+1:N} \mid X_{t_{n,s-1}} = x)^2} \le C_1^{*2} - 1. \tag{29}$$

In (29), the distribution of $X_{t_{n,s}}$ given $X_{t_{n,s-1}} = x$ is understood as given by the law of the latent Markov process, unconditional on data. Assumption 1 asserts that the forecast likelihood of all future observations given $X_{t_{n,s}}$ does not deviate too much from the forecast likelihood given the value $X_{t_{n,s-1}} = x$. Note that C_1^* depends on the length of the time interval $[t_{n,s-1}, t_{n,s}]$, and thus on the number of intermediate steps S. Assumption 1* is related to the rate at which the information provided by future observations are processed by the filtering algorithm.*

In what follows, we will assume that multinomial resampling is used. Under multinomial resampling, the indices a^j in Algorithm 3 are drawn independently of each other, given $\{w^j ; j \in 1:J\}$.

Theorem 2 *Suppose multinomial resampling and the exact guide function (28) are used in Algorithm 3. Also suppose that Assumption 1* holds. If f is a measurable function such that $\|f\|_\infty \le 1$ and $a > 1$ is an arbitrary constant, then we have*

$$\left| \frac{1}{J} \sum_{j=1}^{J} f(\tilde{X}_{t_N}^j) - \mathbb{E}[f(X_{t_N}) | Y_{1:N} = y_{1:N}] \right| \le \frac{4a(C_1^* + 1)}{\sqrt{J}} (NS + 1) \tag{30}$$

with probability at least $1 - \frac{(2NS+1)(NS+1)}{a^2}$, given that $\sqrt{J} \ge 8a(C_1^* + 1)NS$.

Proof See Section S2 in the supplementary material. □

Theorem 2 gives a bound on the MC error in filtering estimates when the GIRF approach is used with the exact guide function. If we are to keep the probability $\frac{(2NS+1)(NS+1)}{a^2}$ with which the bound is violated at a fixed level, the number a needs to increase linearly with S, and thus the error bound increases at a rate of at most $O(S^2)$. We will show below that if we take $S = d$, C_1^* can be uniformly bounded (i.e., $\mathcal{O}(1)$) as the dimension d increases, under certain circumstances. Theorem 2 implies that if $S = d$ and $C_1^* = \mathcal{O}(1)$ in d, the MC error will scale at most polynomially in d. We note that if there are no intermediate propagation and resampling steps, that is if $S = 1$, C_1^* typically scales exponentially in d.

Proposition 1 *Consider a POMP model consisting of d independent one dimensional latent process $\{X_t\} = \{X_t^{[1:d]}\}$ and measurement processes $\{Y_n\} = \{Y_n^{[1:d]}\}$. Let each observation be denoted by $y_n = y_n^{[1:d]}$. Suppose that there exists d positive real numbers $\zeta^{[1:d]}$ such that for every $i \in 1:d$, $s \in 1:S$, $n \in 0:N-1$, $\tau \in [t_{n,s-1}, t_{n,s}]$, and $x \in \mathbb{X}$,*

$$\frac{d}{d\tau} \log \mathrm{Var}\left[p(y_{n+1:N}^{[i]} \mid X_\tau^{[i]}) \,\middle|\, X_{t_{n,s-1}}^{[i]} = x^{[i]}\right] \le 2\zeta^{[i]}. \tag{31}$$

Suppose further that

$$t_{n,s} - t_{n,s-1} \le \frac{\Delta}{d}$$

for some $\Delta > 0$ and all $s \in 1:S$ and $n \in 0:N-1$. Then in Assumption 1, we can set*

$$C_1^* = \exp\left\{ \frac{1}{d} \sum_{i=1}^{d} \zeta^{[i]} \cdot \Delta \right\}. \tag{32}$$

Thus if $\sum_{i=1}^d \zeta^{[i]} = \mathcal{O}(d)$, C_1^* in (32) is $\mathcal{O}(1)$.

Proof See supplementary section S4. □

If we set $S = d$ such that $|t_{n,s} - t_{n,s-1}| = \mathcal{O}(\frac{1}{d})$, Proposition 1 says that the MC error bound in Theorem 2 scales polynomially in d for independent models. However, we note that the independence assumption is not crucial; see a correlated Brownian motion example in the supplemenary section S5.

Assumption 1* takes explicit advantage of the requirement for the GIRF method that the latent process operates in continuous time. The latent process transition kernel that is non-deterministic over intermediate time intervals provides the randomness necessary for gradually guiding the particles to the next guided filter distribution. As a counterexample, consider a case where the latent process is deterministic except for making random jumps at fixed observation times $t_{1:N}$. We suppose that the sample paths are right-continuous at $t_{1:N}$. Due to the deterministic evolution of the latent process in the interval $[t_n, t_{n+1})$, we have

$$\mathrm{Var}\left[p(y_{n+1:N} \mid X_{t_{n,s}}) \,\middle|\, X_{t_{n,s-1}}\right] = 0$$

for $s \in 1:S-1$. However, for a POMP model consisting of d independent processes and for $s = S$, we have

$$\begin{aligned}&\frac{\mathrm{Var}\left[p(y_{n+1:N} \mid X_{t_{n+1}}) \,\middle|\, X_{t_{n,S-1}} = x_{t_{n,S-1}}\right]}{p(y_{n+1:N} \mid X_{t_{n,S-1}} = x_{t_{n,S-1}})^2}\\ &= \frac{\mathrm{Var}\left[p(y_{n+1:N} \mid X_{t_{n+1}}) \,\middle|\, X_{t_n} = x_{t_n}\right]}{p(y_{n+1:N} \mid X_{t_n} = x_{t_n})^2}\\ &= \prod_{i=1}^{d} \frac{\mathbb{E}\left[p(y^{[i]}_{n+1:N} \mid X^{[i]}_{t_{n+1}})^2 \,\middle|\, X^{[i]}_{t_n} = x^{[i]}_{t_n}\right]}{p(y^{[i]}_{n+1:N} \mid X^{[i]}_{t_n} = x^{[i]}_{t_n})^2} - 1,\end{aligned}$$

where x_{t_n} is a value of the latent process at t_n from which the deterministic evolution leads to $x_{t_{n,S-1}}$ at $t_{n,S-1}$. Since the product of d terms in the right hand side generally scales exponentially in d, the bound C_1^{*2} also scales exponentially. We see that the continuously random property of the latent process is necessary for Algorithm 3 to be able to scale favorably.

4.2 Scaling properties when the guide function is not exact

We now consider the case when ψ_t is not exact. The MC error is affected by the inaccurate approximation of the forecast likelihoods $p(y_{n+1:N} \mid x_{t_{n,s}})$ by $\psi_{t_{n,s}}(x_{t_{n,s}})$. In order to derive a bound on the MC error similar to that in Theorem 2, we introduce two technical assumptions. The first assumption is analogous to Assumption 1* in Sect. 4.1.

Assumption 1 There exists $C_1 \geq 1$ such that for all $s, s' \in 1:S$ and $n, n' \in 0:N-1$ such that $t_{n,s} \leq t_{n',s'}$ and for every $x \in \mathbb{X}$,

$$\frac{\mathrm{Var}\left[\mathbb{E}\left(\psi_{t_{n',s'}}(X_{t_{n',s'}}) \prod_{m=n+1}^{n'} g_m(y_m \mid X_{t_m}) \,\middle|\, X_{t_{n,s}}\right) \,\middle|\, X_{t_{n,s-1}} = x\right]}{\mathbb{E}\left(\psi_{t_{n',s'}}(X_{t_{n',s'}}) \prod_{m=n+1}^{n'} g_m(y_m \mid X_{t_m}) \,\middle|\, X_{t_{n,s-1}} = x\right)^2} \leq C_1^2 - 1.$$

If $\psi = \psi^*$, we have

$$\begin{aligned}&\mathbb{E}\left[\psi_{t_{n',s'}}(X_{t_{n',s'}}) \prod_{m=n+1}^{n'} g_m(y_m \mid X_{t_m}) \,\middle|\, X_{t_{n,s}} = x\right]\\ &= p(y_{n+1:N} \mid X_{t_{n,s}} = x) = \psi^*_{t_{n,s}}(x),\end{aligned} \tag{33}$$

and Assumption 1 simplifies to Assumption 1*. If $\psi_{t_{n',s'}}(x)$ approximates the forecast likelihood $p(y_{n'+1:n'+L} \mid X_{t_{n',s'}} = x)$, the quantity $\mathbb{E}\left[\psi_{t_{n',s'}}(X_{t_{n',s'}}) \prod_{m=n+1}^{n'} g_m(y_m \mid X_{t_m}) \,\middle|\, X_{t_{n,s}} = x\right]$ in turn gives an approximation to $p(y_{n+1:n'+L} \mid X_{t_{n,s}} = x)$.

The second assumption concerns how closely the guide function ψ_t approximates the forecast likelihood of future observations. For a constant $c \geq 1$ and a subset $\mathcal{C}$ of $\mathbb{X}$, we define $\mathrm{Osc}(c\,;\mathcal{C})$ to be a class of positive functions f on $\mathbb{X}$ such that

$$c \cdot \inf_{x \in \mathcal{C}} f(x) \geq \sup_{x \in \mathbb{X}} f(x).$$

Assumption 2 There exist constants $C_2 \geq 1$, $\rho \in (0, 1]$, and a collection of regions $\mathcal{C}_{t_{n,s}} \in \mathcal{X}$ for $s \in 1:S$ and $n \in 0:N-1$ such that the following hold:

(i) For all $s \in 1:S$ and $n \in 0:N-1$,

$$P^{\psi}_{t_{n,s}}(\mathcal{C}_{t_{n,s}}) \geq \rho.$$

(ii) For all $s, s' \in 1:S$ and $n, n' \in 0:N-1$,

$$\frac{\mathbb{E}\left[\psi_{t_{n',s'}}(X_{t_{n',s'}}) \prod_{m=n+1}^{n'} g_m(y_m \mid X_{t_m}) \,\middle|\, X_{t_{n,s}} = x\right]}{\psi_{t_{n,s}}(x)} \in \mathrm{Osc}(C_2\,;\mathcal{C}_{t_{n,s}}). \tag{34}$$

A value of C_2 that is close to unity indicates that $\psi_{t_{n,s}}(x)$ is approximately proportional to the quantity $\mathbb{E}\left[\psi_{t_{n',s'}}(X_{t_{n',s'}}) \prod_{m=n+1}^{n'} g_m(y_m \mid X_{t_m}) \,\middle|\, X_{t_{n,s}} = x\right]$. If $\psi = \psi^*$, due to (33), one can take $C_2 = 1$, $\mathcal{C}_{t_{n,s}} = \mathbb{X}$ for all $t_{n,s}$, and $\rho = 1$. The fact that a constant multiple of the infimum of the ratio in (34) over $\mathcal{C}_{t_{n,s}}$ is lower bounded by the global supremum indicates that the guide function $\psi_{t_{n,s}}$ can overestimate the forecast likelihood outside $\mathcal{C}_{t_{n,s}}$. For instance, if we consider the case where the region $\mathcal{C}_{t_{n,s}}$ is defined as $\{x \in \mathbb{X}\,;\, \psi_{t_{n,s}} > c\}$ for some value $c > 0$, then Assumption 2 (ii) might be interpreted as that $\psi_{t_{n,s}}$ has tails at least as thick as those of the numerator in (34). Assumption 2 (i) says that the region $\mathcal{C}_{t_{n,s}}$ has to carry a probability mass of at least ρ with respect to $P^{\psi}_{t_{n,s}}$.

Under Assumptions 1 and 2, the MC error in filtering estimates can be bounded as follows.

Theorem 3 *Suppose multinomial resampling is used in Algorithm 3. Also suppose that Assumptions 1 and 2 hold. If f is a measurable function such that $\|f\|_\infty \leq 1$ and $a > 1$ is an arbitrary constant, then we have*

$$\begin{aligned}&\left|\frac{1}{J}\sum_{j=1}^{J} f(\tilde{X}^j_{t_N}) - \mathbb{E}[f(X_{t_N}) | Y_{1:N} = y_{1:N}]\right|\\ &\leq \frac{4aC_2(C_1+1)}{\rho\sqrt{J}}(NS+1)\end{aligned} \tag{35}$$

with probability at least $1 - \frac{(2NS+1)(NS+1)}{a^2}$, given that $\sqrt{J} \geq 8\rho^{-2}aC_2(C_1+1)NS$.

Proof See supplementary section S2. □

When $\psi = \psi^*$, Theorem 3 reduces to Theorem 2. When $\psi \neq \psi^*$, it is possible to show a result similar to Proposition 1 and claim that C_1 is uniformly bounded in d for independent

models under certain conditions, provided that $S = d$. Unfortunately, C_2 scales exponentially in d. Nevertheless, taking $\psi_{t_{n,s}}$ to be an approximation to $p(y_{n+1:n+L} \mid X_{t_{n,s}} = x)$ can greatly reduce the rate of exponential growth of C_2 compared to the case

$$\psi_{t_{n,s}}(x) = \begin{cases} 1 & \text{for } s \in 1:S-1 \\ g_{n+1}(y_{n+1} \mid X_{t_{n,S}} = x) & \text{for } s = S, \end{cases}$$

which corresponds to the bootstrap particle filter. As shown in Sect. 3, even rough approximations for ψ, such as those made by ignoring the correlation between components (Table 3 for the correlated Brownian motion example) or by simulation-based moment matching (stochastic Lorenz 96 example), can extend the dimensionality of the models for which reasonably good filtering estimates can be obtained.

A sufficient condition for Assumption 2 can be obtained based on the mixing property of the latent process conditional on data. We say that the latent process mixes well over the interval $[t_{n,s}, t_{n+L}]$ conditional on data if the conditional expectation $\mathbb{E}\left[f(X_{t'}) \middle| y_{n+1:n+L}, X_{t_{n,s}} = x\right]$ for $t' \geq t_{n+L}$ does not vary substantially across the space as a function of x. Loosely speaking, this condition implies that the state $X_{t_{n,s}}$ does not influence the future state $X_{t'}$ much, given the observations $y_{n+1:n+L}$. This condition is related to the φ-mixing of the conditional law of the latent process $\{X_t\}$ given $y_{n+1:n+L}$ between the two σ-algebras generated by $\{X_t\,;\, t \geq t_{n+L}\}$ and $\{X_t\,;\, t \leq t_{n,s}\}$ (Billingsley 1999, p. 260). The following proposition supports taking $\psi_{t_{n,s}}$ as an approximation to $p(y_{n+1:n+L})$, provided that the latent process mixes over $[t_{n,s}, t_{n+L}]$ conditional on data.

Proposition 2 *Let $s, s' \in 1:S$ and $n, n' \in 0:N-1$ be such that $n' \geq n+L$ and let $\mathcal{C}_{t_{n,s}} \in \mathcal{X}$ be given. Suppose that the following two conditions hold for some constants $C_{2,a}, C_{2,b} \geq 1$:*

$$\mathbb{E}\left[\psi_{t_{n',s'}}(X_{t_{n',s'}}) \prod_{m=n+L+1}^{n'} g_m(y_m \mid X_{t_m}) \,\middle|\, y_{n+1:n+L}, X_{t_{n,s}} = x\right] \in \mathrm{Osc}(C_{2,a}\,;\, \mathcal{C}_{t_{n,s}}), \tag{36}$$

$$\frac{p(y_{n+1:n+L} \mid X_{t_{n,s}} = x)}{\psi_{t_{n,s}}(x)} \in \mathrm{Osc}(C_{2,b}\,;\, \mathcal{C}_{t_{n,s}}). \tag{37}$$

Then we have

$$\frac{\mathbb{E}\left[\psi_{t_{n',s'}}(X_{t_{n',s'}}) \prod_{m=n+1}^{n'} g_m(y_m \mid X_{t_m}) \,\middle|\, X_{t_{n,s}} = x\right]}{\psi_{t_{n,s}}(x)} \in \mathrm{Osc}(C_{2,a}C_{2,b}\,;\, \mathcal{C}_{t_{n,s}}). \tag{38}$$

Proof See supplementary section S4. □

The condition (37) states that the latent process mixes over $[t_{n,s}, t_{n+L}]$ conditional on data, with respect to a specific function $\psi_{t_{n',s'}}(X_{t_{n',s'}}) \prod_{m=n+L+1}^{n'} g_m(y_m \mid X_{t_m})$ of future states. The condition (37) says $\psi_{t_{n,s}}$ approximates the forecast likelihood of L future observations. Provided these two conditions, (38) says the condition (34) in Assumption 2 (ii) holds for $C_2 = C_{2,a}C_{2,b}$. Proposition 2 implies that if the latent process mixes slowly conditional on data, the guide function will need to approximate the forecast likelihood of a large number of future observations. Since the approximation of the forecast likelihood of a large number of future observations can be practically difficult, the MC error in filtering estimates is likely to increase. This situation can be intuitively understood as that if the latent process has long memory, it is difficult to know early enough which particles will be consistent with distant future observations.

The implications of the theoretical results in this section may be summarized as follows. Assumption 1 concerns the source of filtering error coming from the MC randomness in propagation steps. This source of error can be controlled by carrying out intermediate propagation and resampling with $S = d$. By contrast, the auxiliary particle filter, which is equivalent to Algorithm 3 with $S = 1$, scales poorly even when equipped with a good guide function, as indicated by both theory and practice (Snyder et al. 2015). Assumption 2 bounds the source of filtering error originating from targeting the guided filter distribution P_t^{ψ} instead of the smoothing distribution $p(x_t \mid y_{1:N})$. The filtering error can be reduced by making accurate approximations to forecast likelihoods, reducing C_2. If mixing of the latent process conditional on data happens fast, it may be practically feasible to use ψ that approximates the forecast likelihood of a few number of future observations.

We present two results on the asymptotic normality of the MC error in the likelihood estimate (Theorem 4) and the filtering estimates (Theorem 5). Under Assumptions 1 and 2, we derive upper bounds on the asymptotic variances of these quantities. The connection with Assumptions 1 and 2 is the novel contribution of these results, since the asymptotic normality itself follows directly from existing results in the literature (e.g., Section 9 in, Del Moral 2004). The proofs are given in supplementary section S3.

Theorem 4 *In the limit where the particle size J tends to infinity, the likelihood estimate $\hat{\ell}$ from GIRF (Algorithm 3) converges in distribution to a normal distribution:*

$$\sqrt{J}\left(\frac{\hat{\ell}}{\ell_{1:N}(y_{1:N})} - 1\right) \Longrightarrow \mathcal{N}(0, \mathcal{V}).$$

Under Assumptions 1 and 2, the asymptotic variance is bounded above by

$$\mathcal{V} < NS\left(\frac{C_1^2 C_2^2}{\rho^2} - 1\right).$$

An application of the delta method leads to the asymptotic normality of the log likelihood estimate (Bickel and Doksum 2015):

$$\sqrt{J}\left(\log\hat{\ell} - \log\ell_{1:N}(y_{1:N})\right) \Longrightarrow \mathcal{N}(0,\mathcal{V}).$$

Theorem 5 *In the limit where the particle size J tends to infinity, the following asymptotic normality holds for every measurable function $f : \mathbb{X} \to \mathbb{R}$ such that $\|f\|_\infty \le 1$:*

$$\sqrt{J}\left(\frac{1}{J}\sum_{j=1}^{J} f(\tilde{X}^j_{t_N}) - \mathbb{E}[f \mid Y_{1:N} = y_{1:N}]\right) \Longrightarrow \mathcal{N}(0,\mathcal{W}(f)).$$

Under Assumptions 1 and 2, the asymptotic variance is bounded above by

$$\mathcal{W}(f) < 1 + 4NS\frac{C_1^2C_2^2}{\rho^2}.$$

5 Parameter inference using the GIRF

Being a Monte Carlo algorithm that yields unbiased estimates of the likelihood of data, GIRF (Algorithm 3) can be easily combined with existing parameter inference methods that build upon the particle filter. These parameter estimation methods include particle Markov chain Monte Carlo (PMCMC) (Andrieu et al. 2010), SMC^2 (Chopin et al. 2013), and iterated filtering (Ionides et al. 2015). For high-dimensional POMP models, likelihood estimates often have large amount of Monte Carlo error, for any feasible amount of Monte Carlo effort, even when filtering is successful. This prevents the use of PMCMC, which requires a standard deviation order of 1 log unit (Doucet et al. 2015). In this paper, we will focus on parameter estimation carried out by iterated filtering. We will show that iterated filtering, together with Monte Carlo adjusted profile methodology by Ionides et al. (2017), is able to operate successfully in the presence of relatively high levels of Monte Carlo error.

The iterated filtering approach of Ionides et al. (2015) is a plug-and-play parameter estimation algorithm that finds the maximum likelihood estimate (MLE) of multi-dimensional parameters via an SMC approximation to an iterated, perturbed Bayes map. This algorithm, when implemented via a plug-and-play SMC filtering approach, provides plug-and-play inference on unknown model parameters. Iterated filtering runs a sequence of particle filter on the augmented space comprising the latent variable and the parameter, where the parameters are subject to random perturbations at each time point. The size of perturbations decrease over iterations to induce convergence. In the limit where the perturbation size approaches zero, Ionides et al. (2015) showed that the distribution of filtered parameters approaches a point mass at the MLE under regularity conditions.

Algorithm 4: An iterated guided intermediate resampling filter (iGIRF)

Input : data, $y_{1:N}$; simulator for $p(x_{t_0};\theta)$; simulator for $p(x_{t_{n,s}} \mid x_{t_{n,s-1}};\theta)$; evaluator for $g_n(y_n \mid x_{t_n},\theta)$; evaluator for $\psi_{t_{n,s}}(x_{t_{n,s}},\theta)$; number of particles, J; initial parameter swarm, $\Theta^{0,1:J}$; perturbation kernel for initial value parameter, $\kappa_0(d\theta;\phi,\sigma)$; perturbation kernel, $\kappa_{n,s}(d\theta;\phi,\sigma)$; number of iterations, M; sequence of perturbation sizes, $\sigma_{1:M}$

Output: final parameter swarm $\Theta^{M,1:J}$

for $m \leftarrow 1:M$ **do**
  Run Algorithm 3 on the extended latent space $(X_{t_{n,s}}, \Theta^m_{t_{n,s}})$ with initial draws from (39) and subsequent draws from (40)
  Set $\Theta^{m,j} = \tilde{\Theta}^{m,j}_{t_N}$ for $j \in 1:J$
end

Algorithm 4 presents an iterated guided intermediate resampling filter (iGIRF). The algorithm starts with an initial set of parameters $\{\Theta^{0,j};\, j \in 1:J\}$. At the beginning of the m-th iteration, the parameter component of each particle is perturbed from its current position $\Theta^{m-1,j}$ with kernel κ_0 independently for each $j \in 1:J$. A pre-set decreasing sequence $(\sigma_m)_{m=1:M}$ determines the size of perturbation. The initial latent variables $\tilde{X}^{1:J}_{t_0}$ are drawn from the initial latent distributions parameterized by the perturbed parameters $\tilde{\Theta}^{m,1:J}_{t_0}$, as follows:

$$\begin{aligned}\tilde{\Theta}^{m,j}_{t_0} &\sim \kappa_0(\cdot;\Theta^{m-1,j},\sigma_m),\\ \tilde{X}^j_{t_0} &\sim p_{X_{t_0}}(\cdot;\tilde{\Theta}^{m,j}_{t_0}), \quad j \in 1:J.\end{aligned} \tag{39}$$

Parameters are perturbed at each intermediate time $t_{n,s}$ with kernel $\kappa_{n,s}$, and the states are then drawn from the parameterized transition kernel:

$$\begin{aligned}\Theta^{m,j}_{t_{n,s}} &\sim \kappa_{n,s}(\cdot;\tilde{\Theta}^{m,j}_{t_{n,s-1}},\sigma_m),\\ X^j_{t_{n,s}} &\sim p_{X_{t_{n,s}}\mid X_{t_{n,s-1}}}(\cdot \mid \tilde{X}^j_{t_{n,s-1}},\Theta^{m,j}_{t_{n,s}}), \quad j \in 1:J.\end{aligned} \tag{40}$$

These perturbations define an extended POMP model for $(X_{t_{n,s}}, \Theta^m_{t_{n,s}})$, and the weighting and resampling steps are carried out on this extended model following GIRF (Algorithm 3). At the end of filtering, the parameter swarm $\tilde{\Theta}^{m,j}_{t_N}$ are set as $\Theta^{m,j}$. After M iterations, the final parameter swarm $\Theta^{M,j}$ is considered to be a collection of numerical approximations of the MLE.

Our implementation of iGIRF uses Gaussian parameter perturbations. For parameters with interval constraints, we apply certain transformations beforehand such as taking the logarithm for positive parameters to ensure that Gaussian

perturbations do not violate the constraints. Our examples require us to consider two forms for the kernel $\kappa_{n,s}$. *Initial value parameters* (IVPs) are perturbed only by κ_0, and all other $\kappa_{n,s}$ have a point mass at the identity for the IVPs. IVPs are parameters which encode the value of X_{t_0} but play no subsequent role in the dynamics of the system. For our examples of iGIRF, all parameters other than IVPs use a non-singular kernel which does not depend on n and s, and we call these *regular parameters*. Intuitively, treating parameters as regular is appropriate in iGIRF if information about the parameters arrives at a steady rate through the time series.

5.1 Numerical results

5.1.1 Stochastic Lorenz 96 model

In order to test the parameter estimation capability of iGIRF, we made inference on F with or without the knowledge of σ_p and σ_m from the data for the fifty dimensional stochastic Lorenz 96 model considered in Sect. 3.2. The likelihoods of data were estimated at values of F between 6.0 and 10.0 with intervals of 0.5 (Fig. 2). The guide function was constructed according to (17) and (18) using forty guide simulations. The likelihoods estimated at $\sigma_p = \sigma_m = 1$ were used to estimate the slice likelihood curve. We also estimated the MLEs for σ_p and σ_m using iGIRF (Algorithm 4) and estimated the likelihoods at the obtained Monte Carlo MLE using Algorithm 3 to approximate the profile likelihood curve. The Monte Carlo MLE was taken to be the mean value of the parameter swarm at the end of the twentieth iteration in Algorithm 4 (i.e., $M = 20$). The estimation at each value of F was repeated twice independently. Five particle islands with two thousand particles each were used to estimate the slice and the profile likelihood estimates. We used $S = 50$ intermediate steps per observation interval and $L = 2$ future observations for the guide function.

We fit smooth curves through the estimated likelihoods using a non-parametric local regression procedure. We further constructed approximate 95% confidence intervals for F based on locally quadratic fits through the likelihood estimates around the maximum of the smoothed fits, following the procedure proposed in Ionides et al. (2017). This procedure further developed methods proposed by Diggle and Gratton (1984) that enable parameter inference from models that are implicitly defined by simulation algorithms. We give more details here in order to make the explanation of this procedure self-contained. When the likelihood of data from a one-parameter model can be exactly evaluated, the 95%-confidence interval for the maximum likelihood estimate of the parameter can be obtained by a cut-off on the likelihood curve at $\frac{z^2_{0.975}}{2} = 1.92$, where $z_{0.975}$ is the 0.975 quantile of the standard normal distribution. In large and complex models where the likelihoods of data are estimated with Monte Carlo methods with non-negligible amount of error, the uncertainty in the likelihood estimates has to be taken into account in computing the cut-off. The procedure for constructing the Monte Carlo adjusted profile (MCAP) confidence intervals are as follows. We assume that the Monte Carlo profile points $\check{\ell}^{\mathrm{P}}_{1:K}$ are evaluated at $\vartheta_{1:K}$. We fit a smooth curve $\check{\ell}^{\mathrm{S}}(\vartheta)$ through the profile points using a local smoother, such as the R function `loess` (Cleveland et al. (1992), implemented in R-3.4.1). The `loess` function locally fits polynomial curves by giving less weights to points farther away from the point being estimated. The point $\check{\vartheta}$ at which the maximum of the smoothed curve $\check{\ell}^{\mathrm{S}}$ is attained can be taken as the MLE of the parameter ϑ. In order to quantify the Monte Carlo error in the estimated maximum likelihood $\check{\ell}^{\mathrm{S}}(\check{\vartheta})$, we make a local quadratic fit near the maximum, using the weights that were used in evaluating the smoothed curve $\check{\ell}^{\mathrm{S}}$ at $\check{\vartheta}$. Write the fitted quadratic equation as $-\check{a}\vartheta^2 + \check{b}\vartheta + \check{c}$. The variance and covariance of the coefficients $\check{\mathrm{V}}\mathrm{ar}[\check{a}]$, $\check{\mathrm{V}}\mathrm{ar}[\check{b}]$, and $\check{\mathrm{C}}\mathrm{ov}[\check{a}, \check{b}]$ can be obtained as usual. Using the delta method, the standard error of the maximum $\frac{\check{b}}{2\check{a}}$ can be estimated as

$$\mathrm{SE}^2_{\mathrm{mc}} = \frac{1}{4\check{a}^2}\left(\check{\mathrm{V}}\mathrm{ar}[\check{b}] - \frac{2\check{b}}{\check{a}}\check{\mathrm{C}}\mathrm{ov}[\check{a}, \check{b}] + \frac{\check{b}^2}{\check{a}^2}\check{\mathrm{V}}\mathrm{ar}[\check{a}]\right)$$

(Bickel and Doksum 2015). On the other hand, the statistical error originating from the randomness in data can be estimated with the usual formula

$$\mathrm{SE}_{\mathrm{stat}} = \frac{1}{\sqrt{2\check{a}}}.$$

Assuming that the size of the Monte Carlo error is roughly the same across the possible realizations of the data, we can reasonably approximate the total standard error of the Monte Carlo maximum likelihood estimate as

$$\mathrm{SE}_{\mathrm{total}} = \sqrt{\mathrm{SE}^2_{\mathrm{stat}} + \mathrm{SE}^2_{\mathrm{mc}}}.$$

It follows that the cut-off for an approximate $(1-\alpha)$ confidence interval can be obtained as

$$\delta = \check{a} \cdot \mathrm{SE}^2_{\mathrm{total}} \cdot \chi_\alpha = \left(\check{a} \cdot \mathrm{SE}^2_{\mathrm{mc}} + \frac{1}{2}\right) \cdot \chi_\alpha,$$

where χ_α is the $(1-\alpha)$ quantile of the χ-square distribution on one degree of freedom.

The estimated Monte Carlo adjusted confidence intervals from the slice and the profile likelihood estimates, indicated by two blue and red vertical lines in Fig. 2a, were given by (7.90, 7.99) and (7.85, 8.01) respectively. The upper ends of both confidence intervals were located near the true value of $F = 8$. We remark that the log likelihood estimates with

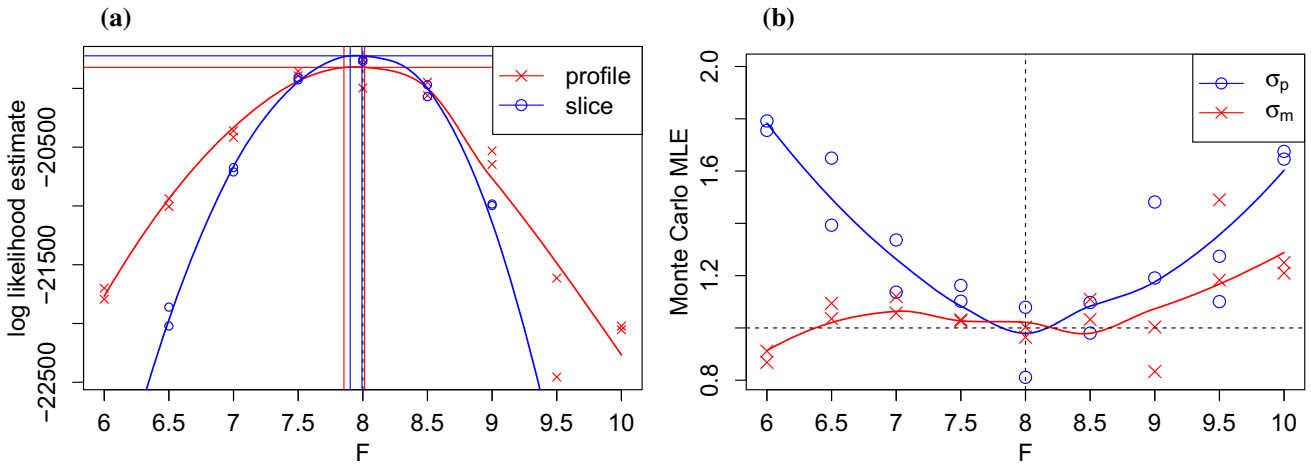

Fig. 2 : Inference on the fifty dimensional stochastic Lorenz 96 model. **a** Estimated slice and profile likelihood curves and Monte Carlo confidence intervals for F. **b** Monte Carlo MLE for σ_p and σ_m

known σ_p and σ_m dropped rapidly to around -4.7×10^4 at $F = 10$, and for this reason the log likelihood estimates at this value of F was excluded from fitting a locally quadratic slice likelihood curve to compute the Monte Carlo adjusted confidence interval. In contrast, the profile likelihood estimates at $F = 10$ did not drop suddenly, thanks to the inflated Monte Carlo MLE for the process noise σ_p (Fig. 2b). Inaccurate values of the forcing constant F were compensated by the process noise estimates larger than the truth. The Monte Carlo MLE for the process noise tended to increase as the value of F deviated from the truth.

5.1.2 Coupled spatiotemporal measles epidemics model

Spatiotemporal inference for epidemiological and ecological systems is arguably the last remaining open problem from the six challenges in time series analysis of nonlinear systems posed by Bjørnstad and Grenfell (2001). Plug-and-play SMC techniques have been central to solving the other five challenges of Bjørnstad and Grenfell (2001), all of which can be represented in the framework of inference for low-dimensional nonlinear non-Gaussian POMP models. Population dynamics of ecological and epidemiological systems can exhibit highly nonlinear stochastic behavior, leading to computational challenges even in low dimensions. Likelihood maximization via iterated filtering has emerged as a practical inference tool for such systems (e.g., Blackwood et al. 2013; Blake et al. 2014; Bakker et al. 2016; Becker et al. 2016; Ranjeva et al. 2017; Pons-Salort and Grassly 2018).

We demonstrate that the GIRF methodology can enable likelihood-based inference on a spatiotemporal mechanistic model addressing a scientific application. We studied the epidemic dynamics of measles, which is well understood compared to other infectious diseases and is characterized by patterns that are closely replicable using a mechanistic model. The study of measles has motivated previous statistical methodology for spatiotemporal population dynamics based on a log-linearization as in Xia et al. (2004) and other approximations as in Eggo et al. (2010), but full likelihood-based fitting using spatially coupled versions of city-level measles transmission models has not previously been carried out. We built on the model of He et al. (2009), adding spatial interaction between multiple cities. We implemented our algorithms with the parameter estimation approach described in Sect. 5 to make inference on the spatial coupling parameter. We used the data collated and studied by Dalziel et al. (2016). The data consisted of biweekly reported case counts in the prevaccination era from year 1949 to 1964 for forty largest cities in England and Wales. Likelihood-based inference for the nonlinear coupled stochastic dynamics of infectious disease in forty cities has not previously been demonstrated and opens the possiblity of various scientific investigations in epidemiological systems and beyond.

The model compartmentalized the population of each city into susceptible (S), exposed (E), infectious (I), and recovered/removed (R) categories. Their sizes for the k-city were denoted by S_k, E_k, I_k, and R_k. The population dynamics was described by the following set of stochastic differential equations:

$$\begin{aligned} \mathrm{d}S_k(t) &= r_k(t)\mathrm{d}t - \mathrm{d}N_{SE,k}(t) - \mu S_k(t)\mathrm{d}t \\ \mathrm{d}E_k(t) &= \mathrm{d}N_{SE,k}(t) - \mathrm{d}N_{EI,k}(t) - \mu E_k(t)\mathrm{d}t \qquad k \in 1:d. \\ \mathrm{d}I_k(t) &= \mathrm{d}N_{EI,k}(t) - \mathrm{d}N_{IR,k}(t) - \mu I_k(t)\mathrm{d}t \end{aligned}$$

Here, $N_{SE,k}(t)$, $N_{EI,k}(t)$, and $N_{IR,k}(t)$ denote the cumulative number of transitions between the corresponding compartments up to time t in city k, μ denotes per-capita mortality rate, and r_k the recruitment rate of susceptible population. The total population $P_k(t)$ was assumed known and

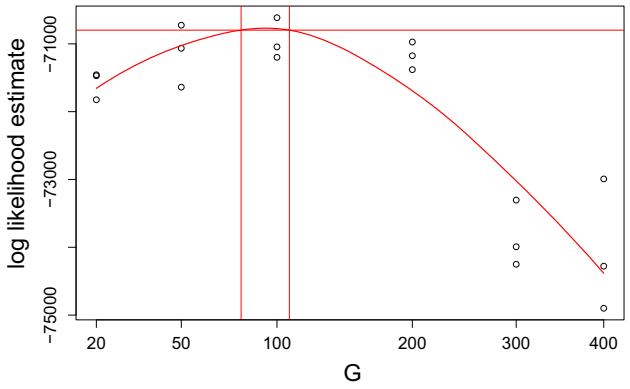

Fig. 3 : Estimated profile likelihood points for various values of G in our spatiotemporal measles dynamics model and the estimated approximate 95% confidence interval (between red vertical lines)

we let $R_k(t) = P_k(t) - S_k(t) - E_k(t) - I_k(t)$. The cumulative transitions were modelled as counting processes with overdispersion relative to Poisson processes, following the construction of Bretó et al. (2009). The term $N_{SE,k}(t)$, representing the cumulative number of infections in the k-th city, has the expected increment of

$$\mathbb{E}\left[N_{SE,k}(t+\mathrm{d}t) - N_{SE,k}(t)\right] = \beta(t)\cdot S_k(t)\cdot\left[\left(\frac{I_k}{P_k}\right)^{\alpha} + \sum_{l\neq k}\frac{v_{kl}}{P_k}\left\{\left(\frac{I_l}{P_l}\right)^{\alpha} - \left(\frac{I_k}{P_k}\right)^{\alpha}\right\}\right]\mathrm{d}t + o(\mathrm{d}t),$$

where $\beta(t)$ denotes the seasonal transmission coefficient and α the mixing exponent (He et al. 2009). The population of city k was denoted by P_k, and the number of travelers from city k to l by v_{kl}. We used the gravity model of Xia et al. (2004) that describes the number of travelers by

$$v_{kl} = G\cdot\frac{\bar{d}}{\bar{P}^2}\cdot\frac{P_k\cdot P_l}{d_{kl}}, \tag{41}$$

where d_{kl} denotes the distance between city k and city l. The gravitation constant G in (41) was scaled with respect to the average population of all forty cities $\bar{P}$ and their average distance $\bar{d}$. The data consisted of the biweekly reported case numbers in each city. The model assumed that a certain fraction ρ_k, called the reporting probability, of the transitions from the infectious compartment to the recovered compartment were, on average, counted as reported cases. The measurement model was chosen to allow for overdispersion relative to the binomial distribution with success probability ρ_k. More details on the model and the inference procedure are given in the supplementary text S6.

We made inference on the gravitation constant G, based on an estimated profile likelihood curve. Ability to infer about the spatial coupling parameter G implies that the filter can recover the full joint distribution for all spatial locations. We fixed G at various levels and estimated other parameters using Algorithm 4. The reporting probabilities ρ_k were estimated by dividing the total case reports by the total births for the corresponding periods in each city, due to the modelling assumption that individuals who once contracted to measles attain lifelong immunity. The estimated ρ_k closely matched the values estimated in He et al. (2009) separately for each city using a mechanistic model. We evaluated the guide function ψ_t using the approach described in Eqs. (10)–(18) in Sect. 2.1 to approximate the forecast likelihood of $L = 3$ future data points. The forecast variability was estimated by making forty random forecasts at every first intermediate time point after observation time (i.e., $t_{n,1}$).

All parameters except G and ρ_k were estimated using iGIRF (Algorithm 4). The IVPs and the regular parameters were estimated alternatingly. For IVP estimation we only used the first three data points, because the information about the initial states was concentrated on the early data points. We iterated fifty times the filtering over the three data points using fifty particle islands comprising sixty particles each. Since the IVPs were only perturbed at the start of each filtering, the particle swarm tended to quickly collapse to a single point. Using many particle islands helped maintain diversity among particles. The regular parameters were estimated by filtering through the whole data once starting from the estimated IVP values. Five islands of six hundred particles each were used for regular parameter estimation. The estimation of IVPs and regular parameters in total took about thirty hours on average using 5 cores. We iterated the alternating estimation ten times. The parameter perturbation size decreased at a geometric factor of 0.92 for each subsequent iteration. The mean of the final swarm of regular parameters was taken as the Monte Carlo MLE. We estimated the IVP corresponding to the Monte Carlo MLE and estimated the likelihood of data using Algorithm 3 with ten islands of one thousand particles each. The obtained likelihood estimate was considered a Monte Carlo profile likelihood for the specified G value. We independently repeated the estimation of profile likelihood six times for each value of G.

We constructed an approximate 95% confidence interval based on the obtained profile likelihood estimates. We used three points of highest profile likelihood estimates among six points for each value of G. Figure 3 shows the estimates of profile log likelihoods and the approximate 95% confidence interval for G. The procedure for obtaining the Monte Carlo adjusted confidence interval was carried out on a transformed scale of $\sqrt{G}$ for a better quadratic fit. The approximate confidence interval was found to be (79, 108), indicated by two vertical lines, using a Monte Carlo adjusted profile cut-off of 35.1 log units. All experiments in Sects. 3.2 and 5.1.2 were

conducted on the Olympus cluster at the Pittsburgh Supercomputing Center.

6 Discussion

Our guided intermediate resampling filter (GIRF) approach enables likelihood-based inference on relatively high-dimensional, nonlinear, implicitly defined dynamic models. Alternative approaches based on information reduction, such as approximate Bayesian computation (ABC), can fail to capture full complexities in the model or result in inaccurate parameter estimates (Fasiolo et al. 2016). There is also a risk of subconscious bias when the scientist's expert knowledge is used to select criteria used to fit a model. In comparison, inference based on the likelihood of data can add to the reliability of scientific conclusions, since the likelihood of data, uniquely defined by the model, provides a common measure of fit. In addition, the statistical efficiency of likelihood-based inference leads to inferences that might be unobtainable for methods requiring information reduction.

Our intermediate propagation and resampling approach can be used when the transition density of the latent process is not evaluable, provided that the latent process is defined in continuous time and a simulator for the process is available. Empirically, we have demonstrated that GIRF can scale up to dimensions substantially larger than the capabilities of alternative algorithms such as the APF or a L-lookahead filter. GIRF can be successfully applied to highly nonlinear models for which the ensemble Kalman filter fails. We also showed that the method enables inference on a scientifically challenging spatiotemporal epidemiological model. Further potential applications may be found in areas such as ecology, behavioral sciences, or epidemiology, when the data are collected at linked spatial locations or structured into many categories. An R package **spatPomp** (a pre-release version available at https://github.com/kidusasfaw/spatPomp) provides a general realization of spatiotemporal POMP models, where the user can define a model by specifying the latent and the measurement processes and analyze data using the GIRF and other algorithms. Many scientific and statistical challenges remain involving analysis of partially observed, highly nonlinear, coupled stochastic systems, and we have shown that the GIRF approach can provide a framework for progress in this enterprise.

Acknowledgements The authors thank Aaron King for the discussions motivating this research and for insightful feedback. Comments on the manuscript by Kidus Asfaw, Yves Atchadé, and two anonymous referees have led to improvements. This work was supported by National Science Foundation Grants DMS-1308919, DMS-1761603, and DMS-1513040, and National Institutes of Health Grants 1-U54-GM111274 and 1-U01-GM110712.

References

Acevedo, W., de Wiljes, J., Reich, S.: Second-order accurate ensemble transform particle filters. SIAM J. Sci. Comput. **39**, A1834–A1850 (2017)

Ades, M., Van Leeuwen, P.J.: The equivalent-weights particle filter in a high-dimensional system. Q. J. R. Meteorol. Soc. **141**, 484–503 (2015)

Andrieu, C., Doucet, A., Holenstein, R.: Particle Markov chain Monte Carlo methods. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **72**, 269–342 (2010)

Bakker, K.M., Martinez-Bakker, M.E., Helm, B., Stevenson, T.J.: Digital epidemiology reveals global childhood disease seasonality and the effects of immunization. Proc. Natl. Acad. Sci. **113**, 6689–6694 (2016)

Becker, A.D., Birger, R.B., Teillant, A., Gastanaduy, P.A., Wallace, G.S., Grenfell, B.T.: Estimating enhanced prevaccination measles transmission hotspots in the context of cross-scale dynamics. Proc. Natl. Acad. Sci. **113**, 14595–14600 (2016)

Bengtsson, T., Bickel, P., Li, B.: Curse-of-dimensionality revisited: collapse of the particle filter in very large scale systems. In: Probability and Statistics: Essays in Honor of David A. Freedman, pp. 316–334. Institute of Mathematical Statistics (2008)

Beskos, A., Crisan, D.O., Jasra, A., Whiteley, N.: Error bounds and normalising constants for sequential Monte Carlo samplers in high dimensions. Adv. Appl. Probab. **46**, 279–306 (2014a)

Beskos, A., Crisan, D., Jasra, A.: On the stability of sequential Monte Carlo methods in high dimensions. Ann. Appl. Probab. **24**, 1396–1445 (2014b)

Beskos, A., Crisan, D., Jasra, A., Kamatani, K., Zhou, Y.: A stable particle filter for a class of high-dimensional state-space models. Adv. Appl. Probab. **49**, 24–48 (2017)

Bickel, P.J., Doksum, K.A.: Mathematical Statistics: Basic Ideas and Selected Topics, vol. 117. CRC Press, Boca Raton (2015)

Billingsley, P.: Convergence of Probability Measures. Wiley Series in Probability and Statistics. Wiley, New York (1999)

Bjørnstad, O.N., Grenfell, B.T.: Noisy clockwork: time series analysis of population fluctuations in animals. Science **293**, 638–643 (2001)

Blackwood, J.C., Streicker, D.G., Altizer, S., Rohani, P.: Resolving the roles of immunity, pathogenesis, and immigration for rabies persistence in vampire bats. Proc. Natl. Acad. Sci. **110**, 20837–20842 (2013)

Blake, I.M., Martin, R., Goel, A., Khetsuriani, N., Everts, J., Wolff, C., Wassilak, S., Aylward, R.B., Grassly, N.C.: The role of older children and adults in wild poliovirus transmission. Proc. Natl. Acad. Sci. **111**, 10604–10609 (2014)

Bloem-Reddy, B., Orbanz, P.: Random-walk models of network formation and sequential Monte Carlo methods for graphs. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **80**, 871–898 (2018)

Bretó, C., He, D., Ionides, E.L., King, A.A.: Time series analysis via mechanistic models. Ann. Appl. Stat. **3**, 319–348 (2009)

Bunch, P., Godsill, S.: Approximations of the optimal importance density using Gaussian particle flow importance sampling. J. Am. Stat. Assoc. **111**, 748–762 (2016)

Cappé, O., Godsill, S.J., Moulines, E.: An overview of existing methods and recent advances in sequential Monte Carlo. Proc. IEEE **95**, 899–924 (2007)

Chen, R., Wang, X., Liu, J.S.: Adaptive joint detection and decoding in flat-fading channels via mixture Kalman filtering. IEEE Trans. Inf. Theory **46**, 2079–2094 (2000)

Cheng, Y., Reich, S.: Assimilating data into scientific models: an optimal coupling perspective. In: Van Leeuwen, P.J., Cheng, Y., Reich, S. (eds.) Nonlinear Data Assimilation, Frontiers in Applied Dynamical Systems: Reviews and Tutorials, vol. 2, pp. 75–118. Springer, Cham (2015)

Chopin, N.: Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. Ann. Stat. **32**, 2385–2411 (2004)

Chopin, N., Jacob, P.E., Papaspiliopoulos, O.: SMC^2: an efficient algorithm for sequential analysis of state space models. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **75**, 397–426 (2013)

Chorin, A.J., Tu, X.: Implicit sampling for particle filters. Proc. Natl. Acad. Sci. **106**, 17249–17254 (2009)

Chorin, A.J., Morzfeld, M., Tu, X.: A survey of implicit particle filters for data assimilation. In: Zeng, Y., Wu, S. (eds.) State-Space Models, pp. 63–88. Springer, Berlin (2013)

Clapp, T., Godsill, S.: Fixed-lag smoothing using sequential importance sampling. In: Bayesian statistics 6: Proceeding of the Sixth Valencia International Meeting, vol. 6, pp. 743–752 (1999)

Cleveland, W.S., Grosse, E., Shyu, W.M.: Local regression models. In: Chambers, J., Hastie, T. (eds.) Statistical Models in S, pp. 309–376. Chapman and Hall, London (1992)

Dalziel, B.D., Bjørnstad, O.N., van Panhuis, W.G., Burke, D.S., Metcalf, C.J.E., Grenfell, B.T.: Persistent chaos of measles epidemics in the prevaccination United States caused by a small change in seasonal transmission patterns. PLoS Comput. Biol. **12**, e1004655 (2016)

Del Moral, P.: Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications. Springer, New York (2004)

Del Moral, P., Guionnet, A.: On the stability of interacting processes with applications to filtering and genetic algorithms. Ann. l'Inst. Henri Poincaré (B) Probab. Stat. **37**, 155–194 (2001)

Del Moral, P., Jacod, J.: Interacting particle filtering with discrete observations. In: Doucet, A., de Freitas, N., Gordon, N. (eds.) Sequential Monte Carlo Methods in Practice, pp. 43–75. Springer, Berlin (2001)

Del Moral, P., Murray, L.M.: Sequential Monte Carlo with highly informative observations. SIAM/ASA J. Uncertain. Quantif. **3**, 969–997 (2015)

Diggle, P.J., Gratton, R.J.: Monte Carlo methods of inference for implicit statistical models. J. R. Stat. Soc. Ser. B (Methodol.) **46**, 193–212 (1984)

Douc, R., Cappé, O., Moulines, E.: Comparison of resampling schemes for particle filtering. In: Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005, pp. 64–69. IEEE (2005)

Doucet, A., Johansen, A.M.: A tutorial on particle filtering and smoothing: Fifteen years later. In: Crisan, D., Rozovskii, B. (eds.) Oxford Handbook of Nonlinear Filtering. Oxford University Press, Oxford (2011)

Doucet, A., Godsill, S., Andrieu, C.: On sequential Monte Carlo sampling methods for Bayesian filtering. Stat. Comput. **10**, 197–208 (2000)

Doucet, A., De Freitas, N., Gordon, N.: Sequential Monte Carlo Methods in Practice. Springer, Berlin (2001)

Doucet, A., Briers, M., Sénécal, S.: Efficient block sampling strategies for sequential Monte Carlo methods. J. Comput. Graph. Stat. **15**, 693–711 (2006)

Doucet, A., Pitt, M., Deligiannidis, G., Kohn, R.: Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. Biometrika **102**, 295–313 (2015)

Eggo, R.M., Cauchemez, S., Ferguson, N.M.: Spatial dynamics of the 1918 influenza pandemic in England, Wales and the United States. J. R. Soc. Interface **8**, 233–243 (2010)

Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res. Oceans **99**, 10143–10162 (1994)

Farchi, A., Bocquet, M.: Comparison of local particle filters and new implementations. Nonlinear Process. Geophys. **25**, 765–807 (2018)

Fasiolo, M., Pya, N., Wood, S.N.: A comparison of inferential methods for highly nonlinear state space models in ecology and epidemiology. Stat. Sci. **31**, 96–118 (2016)

Giraud, F., Del Moral, P.: Nonasymptotic analysis of adaptive and annealed Feynman-Kac particle models. Bernoulli **23**, 670–709 (2017)

Gordon, N.J., Salmond, D.J., Smith, A.F.: Novel approach to nonlinear/non-Gaussian Bayesian state estimation. IEE Proc. F (Radar Signal Process.) **140**, 107–113 (1993)

Guarniero, P., Johansen, A.M., Lee, A.: The iterated auxiliary particle filter. J. Am. Stat. Assoc. **112**, 1636–1647 (2017)

He, D., Ionides, E.L., King, A.A.: Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. J. R. Soc. Interface **7**, 271–283 (2009)

Houtekamer, P.L., Mitchell, H.L.: A sequential ensemble Kalman filter for atmospheric data assimilation. Mon. Weather Rev. **129**, 123–137 (2001)

Hunt, B.R., Kostelich, E.J., Szunyogh, I.: Efficient data assimilation for spatiotemporal chaos: a local ensemble transform Kalman filter. Phys. D Nonlinear Phenomena **230**, 112–126 (2007)

Ionides, E.L., Nguyen, D., Atchadé, Y., Stoev, S., King, A.A.: Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. Proc. Natl. Acad. Sci. **112**, 719–724 (2015)

Ionides, E.L., Breto, C., Park, J., Smith, R.A., King, A.A.: Monte Carlo profile confidence intervals for dynamic systems. J. R. Soc. Interface **14**, 20170126 (2017)

Johansen, A.M.: On blocks, tempering and particle MCMC for systems identification. In: Proceedings of 17th IFAC Symposium on System Identification, pp. 969–974 (2015)

Kevrekidis, I.G., Gear, C.W., Hummer, G.: Equation-free: the computer-aided analysis of complex multiscale systems. AIChE J. **50**, 1346–1355 (2004)

King, A.A., Nguyen, D., Ionides, E.L.: Statistical inference for partially observed Markov processes via the R package pomp. J. Stat. Softw. **69**, 1–43 (2016)

King, A.A., Ionides, E.L., Bretó, C.M., Ellner, S.P., Ferrari, M.J., Kendall, B.E., Lavine, M., Nguyen, D., Reuman, D.C., Wearing, H., Wood, S.N.: pomp: Statistical inference for partially observed Markov processes. R package, version 2.4 (2019). https://kingaa.github.io/pomp/. Accessed 10 Mar 2020

Kitano, H.: Computational systems biology. Nature **420**, 206 (2002)

Kong, A., Liu, J.S., Wong, W.H.: Sequential imputations and Bayesian missing data problems. J. Am. Stat. Assoc. **89**, 278–288 (1994)

Le Gland, F., Oudjane, N.: Stability and uniform approximation of nonlinear filters using the Hilbert metric and application to particle filters. Ann. Appl. Probab. **14**, 144–187 (2004)

Lei, J., Bickel, P., Snyder, C.: Comparison of ensemble Kalman filters under non-Gaussianity. Mon. Weather Rev. **138**, 1293–1306 (2010)

Lin, M., Chen, R., Mykland, P.: On generating Monte Carlo samples of continuous diffusion bridges. J. Am. Stat. Assoc. **105**, 820–838 (2010)

Lin, M., Chen, R., Liu, J.S.: Lookahead strategies for sequential Monte Carlo. Stat. Sci. **28**, 69–94 (2013)

Liu, J.S., Chen, R.: Blind deconvolution via sequential imputations. J. Am. Stat. Assoc. **90**, 567–576 (1995)

Lorenz, E.N.: Predictability: a problem partly solved. Proc. Seminar Predict. **1**, 1–18 (1996)

Marjoram, P., Molitor, J., Plagnol, V., Tavaré, S.: Markov chain Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. **100**, 15324–15328 (2003)

Miller, R.N., Carter, E.F., Blue, S.T.: Data assimilation into nonlinear stochastic models. Tellus A Dyn. Meteorol. Oceanogr. **51**, 167–194 (1999)

Morzfeld, M., Tu, X., Atkins, E., Chorin, A.J.: A random map implementation of implicit filters. J. Comput. Phys. **231**, 2049–2066 (2012)

Neal, R.M.: Annealed importance sampling. Stat. Comput. **11**, 125–139 (2001)

Owen, J., Wilkinson, D.J., Gillespie, C.S.: Scalable inference for Markov processes with intractable likelihoods. Stat. Comput. **25**, 145–156 (2015)

Palmer, T.N.: Towards the probabilistic Earth-system simulator: a vision for the future of climate and weather prediction. Q. J. R. Meteorol. Soc. **138**, 841–861 (2012)

Papadakis, N., Mémin, É., Cuzol, A., Gengembre, N.: Data assimilation with the weighted ensemble Kalman filter. Tellus A Dyn. Meteorol. Oceanogr. **62**, 673–697 (2010)

Pitt, M.K., Shephard, N.: Filtering via simulation: auxiliary particle filters. J. Am. Stat. Assoc. **94**, 590–599 (1999)

Pons-Salort, M., Grassly, N.C.: Serotype-specific immunity explains the incidence of diseases caused by human enteroviruses. Science **361**, 800–803 (2018)

Ranjeva, S.L., Baskerville, E.B., Dukic, V., Villa, L.L., Lazcano-Ponce, E., Giuliano, A.R., Dwyer, G., Cobey, S.: Recurring infection with ecologically distinct HPV types can explain high prevalence and diversity. Proc. Natl. Acad. Sci. **114**, 13573–13578 (2017)

Rebeschini, P., Van Handel, R.: Can local particle filters beat the curse of dimensionality? Ann. Appl. Probab. **25**, 2809–2866 (2015)

Reich, S.: A nonparametric ensemble transform method for Bayesian inference. SIAM J. Sci. Comput. **35**, A2013–A2024 (2013)

Sisson, S.A., Fan, Y., Tanaka, M.M.: Sequential Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. **104**, 1760–1765 (2007)

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

Snyder, C., Bengtsson, T., Morzfeld, M.: Performance bounds for particle filters using the optimal proposal. Mon. Weather Rev. **143**, 4750–4761 (2015)

Van Leeuwen, P.J.: Nonlinear data assimilation in geosciences: an extremely efficient particle filter. Q. J. R. Meteorol. Soc. **136**, 1991–1999 (2010)

Vergé, C., Dubarry, C., Del Moral, P., Moulines, E.: On parallel implementation of sequential Monte Carlo methods: the island particle model. Stat. Comput. **25**, 243–260 (2015)

Whiteley, N.: Stability properties of some particle filters. Ann. Appl. Probab. **23**, 2500–2537 (2013)

Whiteley, N., Lee, A.: Twisted particle filters. Ann. Stat. **42**, 115–141 (2014)

Wilks, D.S.: Effects of stochastic parametrizations in the Lorenz 96 system. Q. J. R. Meteorol. Soc. **131**, 389–407 (2005)

Xia, Y., Bjørnstad, O.N., Grenfell, B.T.: Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics. Am. Nat. **164**, 267–281 (2004)

Xiu, D., Kevrekidis, I.G., Ghanem, R.: An equation-free, multiscale approach to uncertainty quantification. Comput. Sci. Eng. **7**, 16 (2005)