Statistical Science
2016, Vol. 31, No. 1, 96–118
DOI: 10.1214/15-STS534

A Comparison of Inferential Methods for Highly Nonlinear State Space Models in Ecology and Epidemiology

Matteo Fasiolo, Natalya Pya and Simon N. Wood

Abstract. Highly nonlinear, chaotic or near chaotic, dynamic models are important in fields such as ecology and epidemiology: for example, pest species and diseases often display highly nonlinear dynamics. However, such models are problematic from the point of view of statistical inference. The defining feature of chaotic and near chaotic systems is extreme sensitivity to small changes in system states and parameters, and this can interfere with inference. There are two main classes of methods for circumventing these difficulties: information reduction approaches, such as Approximate Bayesian Computation or Synthetic Likelihood, and state space methods, such as Particle Markov chain Monte Carlo, Iterated Filtering or Parameter Cascading. The purpose of this article is to compare the methods in order to reach conclusions about how to approach inference with such models in practice. We show that neither class of methods is universally superior to the other. We show that state space methods can suffer multimodality problems in settings with low process noise or model misspecification, leading to bias toward stable dynamics and high process noise. Information reduction methods avoid this problem, but, under the correct model and with sufficient process noise, state space methods lead to substantially sharper inference than information reduction methods. More practically, there are also differences in the tuning requirements of different methods. Our overall conclusion is that model development and checking should probably be performed using an information reduction method with low tuning requirements, while for final inference it is likely to be better to switch to a state space method, checking results against the information reduction approach.

Key words and phrases: Nonlinear dynamics, state space models, particle filters, approximate Bayesian computation, statistical ecology.

1. INTRODUCTION

Nonlinear or near-chaotic dynamical systems represent a challenging setting for statistical inference. The chaotic nature of such systems implies that small variations in model parameters can lead to very different observed dynamics. This characteristic alone is enough to invalidate many conventional statistical methods, but in most cases additional complications are present. First, the process under study is generally observed with errors. In addition, many models include a further layer of uncertainty, which we call process stochasticity. In ecology this is often environmental noise, driving the

Matteo Fasiolo is Research Assistant, School of Mathematics, University of Bristol, Bristol BS81TW, United Kingdom (e-mail: matteo.fasiolo@bristol.ac.uk). Natalya Pya is Assistant Professor, School of Science and Technology, Nazarbayev University, Astana 010000, Republic of Kazakhstan (e-mail: natalya.pya@nu.edu.kz). Simon N. Wood is Professor, School of Mathematics, University of Bristol, Bristol BS81TW, United Kingdom (e-mail: simon.wood@bristol.ac.uk).

system dynamics. Process stochasticity increases the complexity of the model in a nontrivial way: apart from being unobservable, its presence makes every realized trajectory of the system essentially unique. This is particularly true for chaotic models where any amount of process noise will cause rapid divergence of two paths generated using identical parameters and initial conditions, in sharp contrast to the situation in which dynamics lie on a stable attractor.

Developing statistical methods that can deal effectively with highly nonlinear systems is not simply a matter of theoretical interest, since examples of nonlinear or near-chaotic behavior in ecological systems abound: lemmings (Kausrud et al., 2008), voles (Turchin and Ellner, 2000), mosquitos (Yang et al., 2008), moths (Kendall et al., 2005) and fish (Anderson et al., 2008). Similar degrees of nonlinearity have been observed in experimental settings, for example, blowflies (Nicholson, 1957) and flour beetles (Desharnais et al., 2001).

The focus of epidemiologists often differs from that of ecologists. Both groups are concerned with explaining the persistence of the species under study, but epidemiologists and ecologists are often aiming respectively at causing and avoiding its extinction (Earn, Rohani and Grenfell, 1998). Despite this divergence in objectives, the mathematical structures used to study population dynamics are often very similar. Hence, the role of nonlinearities in the population dynamics of infectious diseases has attracted much attention in epidemiology as well. In the context of measles, Grenfell (1992) and Grenfell et al. (1995) describe how the interaction between seasonal forcing and observed heterogeneities, such as age structure or spatial coupling, can result in chaotic or stable dynamics, while Grenfell, Bjørnstad and Finkenstädt (2002) address the issue of predictability under a time-series Susceptible Infected Recovered model. More recently, King et al. (2008), Lavine et al. (2013) and Bhadra et al. (2011) use nonlinear stochastic models with multiple compartments to analyze cholera, pertussis and malaria epidemics, respectively.

The relation between chaos, statistics and probability theory has been discussed by Berliner (1992) and Chan and Tong (2001), among others. We have a quite different focus, which is to review and compare the main statistical methods for highly nonlinear dynamic models in ecology and epidemiology, investigating the difficulties involved in their use, and attempting to establish the best approach to take in practical applications.

The paper is organized as follows: in Section 2 we show that the likelihood function of simple dynamic models can be intractable in certain areas of the parameter space, while in Section 3 we briefly review the set of statistical methods most useful in the context of nonlinear dynamic systems. How these methods deal with the issue discussed in Section 2 is the subject of Section 4. In Section 5 we compare the relative performance of these methodologies on a sequence of increasingly realistic (and hence complex) ecological and epidemiological models. We conclude with a discussion.

2. CHAOS AND THE LIKELIHOOD FUNCTION

To provide a simple example illustrating how the dynamics of an ecological model can challenge conventional statistical approaches, let us consider the noisily observed Ricker map

$$y_t \sim \text{Pois}(\phi n_t), \tag{2.1}$$

$$n_{t+1} = r n_t e^{-n_t + z_{t+1}}, \quad z_t \sim N(0, \sigma^2), \tag{2.2}$$

which can be used to describe the evolution in time t of a population n_t. Parameter r is the intrinsic growth rate of the population, controlling the dynamics of the system; ϕ is a scale parameter. The process noise z_t can be interpreted as environmental noise.

Denote with $\mathbf{y}_{1:T} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T\}$ and $\mathbf{n}_{1:T} = \{\mathbf{n}_1, \mathbf{n}_2, \dots, \mathbf{n}_T\}$ the observations and hidden state sequence up to time T, where $\mathbf{y}_t \in \mathbb{R}^{d_y}$ and $\mathbf{n}_t \in \mathbb{R}^{d_n}$ for $t \in \{1, \dots, T\}$. Equations (2.1) and (2.2) define a simple state space model (SSM), for which parameter inference is nontrivial: defining $\boldsymbol{\theta} = \{r, \phi, \sigma\}^T$, the likelihood $p(\mathbf{y}_{1:T}|\boldsymbol{\theta})$ is intractable in certain areas of the parameter space. For example, when $\sigma = 0$, the likelihood is analytically available, but extremely irregular for high values of r. The plot on the top left of Figure 1 shows a transect of the log-likelihood w.r.t. $\log(r)$, obtained using 50 observations, y_t, simulated using parameters $\log(r) = 3.8$, $\sigma = 0$ and $\phi = 10$. Given the ragged shape of the log-likelihood, estimating the parameters by maximum likelihood would be very challenging computationally, while having only limited theoretical motivation. Similarly, any standard MCMC algorithm targeting the parameter posterior distributions would hardly mix at all. This behavior is generic to highly nonlinear dynamic systems: Figure 1 shows likelihood transects for three more dynamic models, defined in Table 1, any of which could be used to make the same points made using the Ricker map, below.

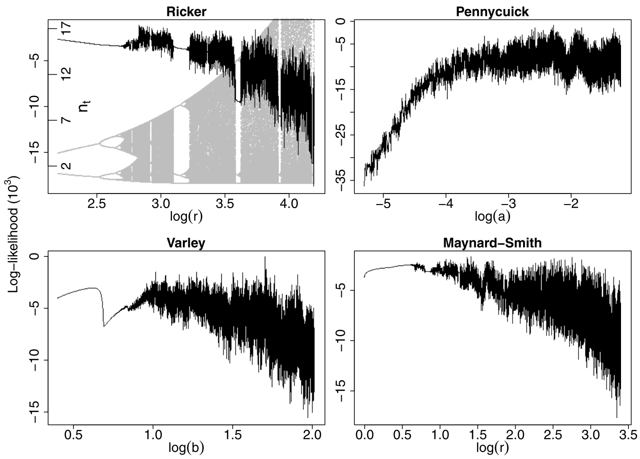

FIG. 1. *Slices of the log-likelihoods of four simple models w.r.t. different parameters (black). In each case $\sigma = 0$, hence, the likelihoods are analytically available. For the Ricker map a bifurcation diagram is included (gray).*

Figure 1 reflects the extreme sensitivity of the likelihood of chaotic models to minuscule changes in parameters or process noise. The bifurcation diagram of the Ricker map (grey) shows the possible long term values n_t of the map, as a function of $\log(r)$. While the trajectories oscillate between two values for $\log(r) \approx 2$, increasing $\log(r)$ above 2.5 leads to a sequence of closely spaced bifurcations, each doubling the periodicity of the map. This period-doubling cascade has a direct effect on the likelihood. Notice that this function is smooth again for values of $\log(r)$ where stable periodic oscillations are recovered. Further increasing $\log(r)$ leads to more period-doubling phases and eventually to chaos.

Figure 2 illustrates the origin of this extreme multimodality. We generated two state paths, $\mathbf{n}_{1:50}$, using $\sigma = 0$ and the same initial value $n_1 = 7$, but different values of $\log(r)$: 3.8 (solid) and 3.799 (dashed). The two paths are close to each other for the first steps, but the mismatch between them increases with time, and by $t = 15$ the peaks and troughs of the paths do not coincide any more. This sort of divergence of neighboring trajectories is the defining feature of chaotic dynamics (measured formally in terms of Lyapunov exponents).

The choice $\sigma = 0$ is quite peculiar. What does the likelihood look like when the process dynamics are stochastic?

TABLE 1
Five simple maps that can show chaotic dynamics. In each case $y_t \sim \text{Pois}(\phi n_t)$ and $z_t \sim N(0, \sigma^2)$

Model name	**Process equation**
Generalized Ricker	$n_{t+1} = r n_t e^{-n_t^\theta + z_t}$
Pennycuick	$n_{t+1} = \frac{r n_t}{1+e^{-a(1-n_t)}} e^{z_t}$
Maynard–Smith	$n_{t+1} = \frac{r n_t}{(1+n_t^b)} e^{z_t}$
Varley	$n_{t+1} = \begin{cases} r n_t e^{z_t}, & \text{if } n_t \le c \\ r n_t^{1-b} e^{z_t} & \text{if } n_t > c \end{cases}$

Box 1
Sequential Importance Resampling (SIR) for likelihood estimation

This algorithm, originally proposed by Gordon, Salmond and Smith (1993), exploits the Markov property to approximate integral (2.3) in T sequential steps. Let $\mathbf{n}_0^{1:M}$ be a sample of particles from the prior distribution $p(\mathbf{n}_0)$. Then $p(\mathbf{y}_{1:T}|\boldsymbol{\theta})$ is estimated as follows.
For $t = 1$ to T:

1. For $i = 1, \dots, M$:
 propagate the ith particle forward
$$\mathbf{n}_t^i \sim p(\mathbf{n}_t^i|\mathbf{n}_{t-1}^i, \boldsymbol{\theta}),$$
 and weight it using the tth observation
$$w^i = p(\mathbf{y}_t|\mathbf{n}_t^i, \boldsymbol{\theta}).$$
2. Estimate the tth likelihood component
$$\hat{p}(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\theta}) = \frac{1}{M}\sum_{i=1}^{M} w^i.$$
3. Resample $\mathbf{n}_t^{1:M}$ with replacement, using probabilities proportional to $\mathbf{w}^{1:M}$.

Finally, estimate the likelihood by using
$$\hat{p}(\mathbf{y}_{1:T}|\boldsymbol{\theta}) = \hat{p}(\mathbf{y}_1|\boldsymbol{\theta}) \prod_{t=2}^{T} \hat{p}(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\theta}).$$

In this case the likelihood, $p(\mathbf{y}_{1:T}|\boldsymbol{\theta})$, must be evaluated by integration:

$$p(\mathbf{y}_{1:T}|\boldsymbol{\theta}) = \int p(\mathbf{y}_{1:T}, \mathbf{z}_{1:T}|\boldsymbol{\theta})\, d\mathbf{z}_{1:T} \tag{2.3}$$
$$= \int p(\mathbf{y}_{1:T}, \mathbf{n}_{1:T}|\boldsymbol{\theta})\, d\mathbf{n}_{1:T},$$

where the second integral is generally the more computationally tractable version. The plot on the right of Figure 2 shows a transect of the estimated log-likelihood of the Ricker map w.r.t. parameter $\log(r)$,

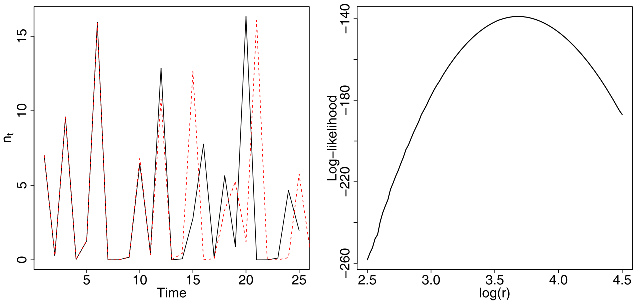

FIG. 2. *Left*: *two trajectories* $\mathbf{n}_{1:T}$ *of the hidden state*, *generated using the same initialization*, *but slightly different values of* $\log(r)$. *Right*: *transect w.r.t.* $\log(r)$ *of the log-likelihood of the Ricker map with* $\sigma = 0.3$, *estimated using the SIR particle filter. The irregularities at* $\log(r) \approx 2.6$ *are due to Monte Carlo noise*.

obtained using the Sequential Importance Resampling (SIR) particle filter with 5×10^5 particles. Box 1 details the main steps of this algorithm, while we refer to Doucet and Johansen (2009) for a more detailed introduction to particle filters. The observed path $\mathbf{y}_{1:50}$ has been simulated using $\log(r) = 3.8$, $\sigma = 0.3$ and $\phi = 10$. In sharp contrast with the deterministic case (Figure 1), it appears that the injection of process noise ($\sigma > 0$) into the system has made the likelihood smooth and unimodal. At this point several questions arise: is the likelihood really smooth, as Figure 2 suggests, or is it possible that the particle filter is hiding the extreme multimodality of Figure 1, so that what we observe in Figure 2 is an artefact of Monte Carlo integration? If the likelihood is indeed smooth, how did the transition from Figure 1 to Figure 2 occur? How much noise σ should be present in order to obtain a smooth likelihood?

Checking the reliability of the estimates provided by a particle filter is difficult because, for nonlinear and/or non-Gaussian models, Monte Carlo or numerical integration are the only ways to get an approximation to (2.3). To obtain a benchmark against which to compare the estimates of the likelihood provided by the filter, we have therefore discretized the state space of the Ricker map in 500 intervals. In this way we can calculate the likelihood exactly, since the integrations are replaced by efficiently computable summations over all the possible values of the states, as detailed in the supplementary material (Fasiolo, Pya and Wood, 2016). Obviously, we do not propose discretization as a viable alternative to particle filters, but we want to use a discretized SSM to compare the performance of a particle filter with the true likelihood. It is interesting to check whether the injection of any amount of noise is sufficient to smooth the likelihood, or whether there is a

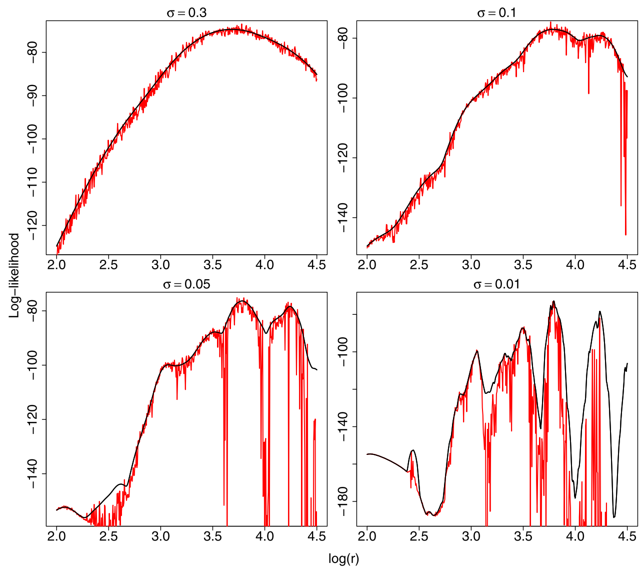

FIG. 3. *Transects of the true log-likelihood* (*black*) *of the discrete Ricker map w.r.t.* $\log(r)$ *for decreasing values of* σ. *The red lines are SIR's estimates*, *using* 1000 *particles*.

slow transition from the intractable likelihood shown in Figure 1 to the unimodal case of Figure 2. Perhaps unsurprisingly, Figure 3 shows that the latter is the case, since as we reduce the process noise the likelihood becomes first multimodal and then (for any practical purpose) nondifferentiable for very low σ. Notice that the SIR estimate of the likelihood deteriorates as multimodality sets in: we will investigate this more fully in Section 4.

This suggests that there is an area of the parameter space, corresponding to high $\log(r)$ and low σ, where the likelihood is essentially intractable. For practical purposes, it is therefore important to compare the robustness of alternative statistical methods across the parameter space, and to understand how alternative methods behave in the face of this difficulty. In particular, we need to avoid the possibility of concluding that a system's dynamics are relatively stable and noisy, not because they really are, but because that is the only case in which the likelihood is numerically tractable.

3. AVAILABLE STATISTICAL METHODS

The literature contains two main classes of statistical methods for nonlinear dynamical systems:

1. Information reduction: methods that discard the information in the data that is most sensitive to extreme divergence of trajectories, so that fitting objectives become more regular. Two methodologies belonging to this group will be described in Section 3.1.

2. State space: these work on the hidden states ($\mathbf{n}_{1:t}$ in Section 2 notation) in order to estimate model parameters and/or the hidden states themselves. Some of these approaches work without modifying the model or the data in any way, by using advanced computational techniques based on particle filtering. We describe two members of this family in Section 3.2.

Given that the main purpose of this work is to consider the applicability and relative performance of these methods in the context of near-chaotic dynamic systems, we will skip over the technical detail whenever they are not essential for the discussion. Obviously our analysis is by no means exhaustive, as we do not examine all the approaches that could be applied in this context. In Section 3.3 we briefly describe some of the alternatives to the methods included in this work.

3.1 Approaches Based on Information Reduction

Since the trajectories of near chaotic systems are extremely sensitive to perturbations of parameters or system state, statistical methods that rely on recovering the true system state face a difficult task. At the same time it is often the case that the true state itself is only a nuisance for parameter estimation, and discarding some information regarding the particular observed trajectory might ease the inferential process.

To make this point clearer, consider again the Ricker paths in Figure 2. Even though the two trajectories, which we indicate with $\mathbf{y}_{1:T}$ and $\mathbf{x}_{1:T}$, are very different in terms of Euclidean distance $\|\mathbf{y}_{1:T} - \mathbf{x}_{1:T}\|$, it is clear that they share some common features. A way around the impossibility of replicating the observed path, even when the simulations use the true or "best-fitting" parameters and initial value, is focusing on the relationship between some characteristic features of the data and the unknown parameters. One way of doing this is to transform the observed and simulated data into a set of summary statistics and to base subsequent inferences on these.

In the following we denote by $\mathbf{y}^0_{1:T}$ the observed path, and with $\mathbf{s}^0 = S(\mathbf{y}^0_{1:T})$ the vector of observed summary statistic. Often methods based on summary statistics involve two main approximations of the likelihood function. The first is implied by the use of $p(\mathbf{s}^0|\boldsymbol{\theta})$ as a proxy for $p(\mathbf{y}^0_{1:T}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are the model parameters. The second approximation arises from the fact that $p(\mathbf{s}^0|\boldsymbol{\theta})$ itself is generally not available analytically, and hence it has to be approximated or estimated by simulation.

We will focus on two approaches based on information reduction: Approximate Bayesian Computation (ABC) (Beaumont, Zhang and Balding, 2002; Fearnhead and Prangle, 2012) and Synthetic Likelihood (SL) (Wood, 2010). These methods will be outlined in Sections 3.1.1 and 3.1.2, respectively.

3.1.1 *Approximate Bayesian computation*. The main purpose of ABC algorithms is approximating the posterior density $p(\boldsymbol{\theta}|\mathbf{y}^0_{1:T}) \propto p(\mathbf{y}^0_{1:T}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, where $p(\boldsymbol{\theta})$ is the prior distribution of the model parameters, when the likelihood $p(\mathbf{y}^0_{1:T}|\boldsymbol{\theta})$ is unavailable or intractable. Given that the data is often transformed into a vector of summary statistics, these methods are generally aiming at sampling from $p(\boldsymbol{\theta}|\mathbf{s}^0)$ rather than $p(\boldsymbol{\theta}|\mathbf{y}^0_{1:T})$.

An elementary ABC algorithm iterates the following rejection procedure (Toni et al., 2009):

1. Sample a vector of parameters $\boldsymbol{\theta}^i$ from $p(\boldsymbol{\theta})$.
2. Simulate a path $\mathbf{y}^i_{1:T}$ from the model $p(\mathbf{y}_{1:T}|\boldsymbol{\theta}^i)$.
3. Transform $\mathbf{y}^i_{1:T}$ to a vector of summary statistics $\mathbf{s}^i = S(\mathbf{y}^i_{1:T})$.
4. Compare $\mathbf{s}^i$ to the observed statistics $\mathbf{s}^0$ using a prespecified distance measure $d(\cdot,\cdot)$. If $d(\mathbf{s}^i, \mathbf{s}^0) \le \varepsilon$, where $\varepsilon \ge 0$, accept $\boldsymbol{\theta}^*$, otherwise reject it.

The output of this algorithm will be distributed according to

$$p(\boldsymbol{\theta})p\{d(\mathbf{s},\mathbf{s}^0) < \varepsilon|\boldsymbol{\theta}\} \propto p\{\boldsymbol{\theta}|d(\mathbf{s},\mathbf{s}^0) < \varepsilon\},$$

which approximates the posterior density, $p(\boldsymbol{\theta}|\mathbf{s}_0)$, for sufficiently small ε. In practice, simple rejection ABC is replaced with MCMC or Sequential Monte Carlo (SMC) algorithms.

3.1.2 *Synthetic likelihood*. Similarly to ABC, this method can be used for problems where the likelihood is intractable, but it is still possible to simulate from the model. The main difference between ABC and SL is how $p(\mathbf{s}^0|\boldsymbol{\theta})$ is approximated. ABC does not rely on any distributional assumption on $\mathbf{s}$, while SL assumes that, approximately,

$$S(\mathbf{y}) \sim N(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta).$$

Briefly, a pointwise estimate of the synthetic likelihood at $\boldsymbol{\theta}$ can be obtained as follows:

1. Simulate N data sets $\mathbf{y}^1_{1:T}, \dots, \mathbf{y}^N_{1:T}$ from the model $p(\mathbf{y}_{1:T}|\boldsymbol{\theta})$.
2. Transform each data set $\mathbf{y}^i_{1:T}$ into a d-dimensional vector of summary statistics $S(\mathbf{y}^i_{1:T})$.

3. Calculate the sample mean $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$ of the statistics (often robustly).
4. Estimate the synthetic likelihood

$$\hat{p}(\mathbf{s}^0|\boldsymbol{\theta}) = (2\pi)^{-d/2}|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}|^{-1/2} \cdot \exp\{-\tfrac{1}{2}(\mathbf{s}^0 - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}})^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{-1}(\mathbf{s}^0 - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}})\}.$$

Hence, SL explicitly provides point estimates of $p(\mathbf{s}^0|\boldsymbol{\theta})$. This estimator can be used within Markov chain Monte Carlo (MCMC) algorithms approximately targeting $p(\boldsymbol{\theta}|\mathbf{s}^0)$ or within an optimizer aiming at maximizing the synthetic likelihood.

3.2 State Space Methods

If discarding information through the use of summary statistics is not desirable, then it is necessary to deal with the hidden states explicitly. As previously stated, calculating the likelihood of SSMs involves integrating the hidden states $\mathbf{n}_{1:T}$ out of the joint density $p(\mathbf{y}^0_{1:T}, \mathbf{n}_{1:T}|\boldsymbol{\theta})$. The SIR particle filter can be used to obtain a Monte Carlo estimate of the likelihood by employing a sequential integration scheme. The use of a sequential approach allows filters to direct the simulated trajectories of the hidden states toward values that are consistent with the observations. This feature is particularly attractive in the context of near-chaotic models, where simulated paths diverge rapidly (recall Figure 2). In this work we mainly focus on algorithms based on the SIR scheme, but many other approaches are available. For example, it is possible to use algorithms that sample directly from the joint posterior density of parameters and hidden states, thus circumventing the estimation of the likelihood. For detailed overviews see Andrieu, Doucet and Holenstein (2010) and Doucet, Godsill and Andrieu (2000).

Here we consider three state space approaches, two of which are based on particle filtering. In Section 3.2.1 we describe a sampler belonging to the family of Particle Markov chain Monte Carlo (PMCMC) methods (Andrieu, Doucet and Holenstein, 2010), while in Section 3.2.2 we introduce the Iterated Filtering (IF) algorithm (Ionides et al., 2011). We consider the Parameter Cascading approach proposed by Ramsay et al. (2007) in Section 3.2.3.

3.2.1 *Particle marginal Metropolis–Hastings sampler.* Filters such as the SIR algorithm can provide point estimates $\hat{p}(\mathbf{y}^0_{1:T}|\boldsymbol{\theta})$ of the likelihood, which ideally converge to the true likelihood as the number of simulations increases. Andrieu, Doucet and Holenstein (2010) proposed to use these estimates of the likelihood to set up a Particle Marginal Metropolis–Hastings (PMMH) algorithm, which can be used to sample from the posterior distribution of the parameters. The algorithm is formed by the following steps:

- *Step* 1: Initialization $i = 0$.
  Given an estimate or a guess of the parameters $\boldsymbol{\theta}_0$, estimate the likelihood $p(\mathbf{y}^0_{1:T}|\boldsymbol{\theta}_0)$ using a particle filter.
- Iteration $i \geq 1$:
  1. Sample a new vector of parameters $\boldsymbol{\theta}^*$ from a transition kernel $K(\boldsymbol{\theta}^*|\boldsymbol{\theta}_{i-1})$.
  2. Using a particle filter, estimate the likelihood $\hat{p}(\mathbf{y}^0_{1:T}|\boldsymbol{\theta}^*)$.
  3. With probability
  $$\min\left\{1, \frac{\hat{p}(\mathbf{y}^0_{1:T}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}{\hat{p}(\mathbf{y}^0_{1:T}|\boldsymbol{\theta}_{i-1})p(\boldsymbol{\theta}_{i-1})} \frac{K(\boldsymbol{\theta}_{i-1}|\boldsymbol{\theta}^*)}{K(\boldsymbol{\theta}^*|\boldsymbol{\theta}_{i-1})}\right\},$$
  set $\boldsymbol{\theta}_i = \boldsymbol{\theta}^*$, otherwise set $\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1}$.

This algorithm is exact in the sense that, despite the use of noisy estimates of $p(\mathbf{y}^0_{1:T}|\boldsymbol{\theta})$ in the acceptance step, it will generate a dependent sample from $p(\boldsymbol{\theta}|\mathbf{y}^0_{1:T})$. The conditions under which this occurs are detailed in Andrieu and Roberts (2009).

3.2.2 *Iterated filtering.* The IF algorithm uses particle filters to provide approximate Maximum Likelihood estimates of the unknown parameters. As shown by Ionides, Bretó and King (2006), by including the unknown parameters in the state space and running a filtering operation, it is possible to estimate the gradient of the likelihood function, which can then be used within an optimization routine. In more detail, Ionides, Bretó and King (2006) treat the parameters as if they were following a multivariate random walk

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{\psi}_t \quad \text{with } \boldsymbol{\psi}_t \sim N(\mathbf{0}, \sigma^2\boldsymbol{\Sigma}). \tag{3.1}$$

With this choice we have that

$$E(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) = \boldsymbol{\theta}_{t-1}, \quad \mathrm{Var}(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) = \sigma^2\boldsymbol{\Sigma},$$
$$E(\boldsymbol{\theta}_0) = \hat{\boldsymbol{\theta}} \quad \text{and} \quad \mathrm{Var}(\boldsymbol{\theta}_0) = c^2\sigma^2\boldsymbol{\Sigma},$$

where σ and c^2 are two variance multipliers, $\hat{\boldsymbol{\theta}}$ is an initial estimate, while $\boldsymbol{\Sigma}$ is typically a diagonal matrix, giving the respective scale of the parameters.

The main result underlying the IF algorithm is

$$\lim_{\sigma^2 \to 0} \sum_{t=1}^{T} \mathbf{V}_t^{-1}(\hat{\boldsymbol{\theta}}_t - \hat{\boldsymbol{\theta}}_{t-1}) = \nabla \log p(\mathbf{y}^0_{1:T}|\boldsymbol{\theta}), \tag{3.2}$$

where

$$\hat{\boldsymbol{\theta}}_t = E(\boldsymbol{\theta}_t|\mathbf{y}^0_{1:t}) \quad \text{and} \quad \mathbf{V}_t = \mathrm{Var}(\boldsymbol{\theta}_t|\mathbf{y}^0_{1:t}),$$

can be estimated using the SIR particle filter. The IF algorithm is composed of the following steps:

- Choose initial value $\hat{\boldsymbol{\theta}}_0^{(0)}$, parameters σ^2, c^2, $\boldsymbol{\Sigma}$, $\alpha \in (0, 1)$ and number of iterations M.
- Iterate for j in $1, \dots, M$:

  1. Set $\sigma_j = \alpha^{j-1}$. Estimate $\hat{\boldsymbol{\theta}}_t^{(j)}$ and $\mathbf{V}_t^{(j)}$, for $t = 1, \dots, T$, using a particle filter.
  2. Update the parameter estimate

$$\hat{\boldsymbol{\theta}}_0^{(j+1)} = \hat{\boldsymbol{\theta}}_0^{(j)} + \mathbf{V}_1^{(j)} \sum_{t=1}^{T} (\mathbf{V}_t^{(j)})^{-1} (\hat{\boldsymbol{\theta}}_t^{(j)} - \hat{\boldsymbol{\theta}}_{t-1}^{(j)}).$$

- Then $\hat{\boldsymbol{\theta}}_0^{(M+1)}$ is an approximate Maximum Likelihood estimate of the parameters.

Notice that, as long as $\sigma > 0$, IF will not be fitting the original model, which will be recovered as $\sigma \to 0$. Ionides et al. (2011) give results concerning the theoretical foundation of IF and describe how slowly σ has to decrease to assure convergence.

3.2.3 *Parameter cascading*. In the context of Ordinary Differential Equations (ODEs), Ramsay et al. (2007) proposed an approach to parameter estimation which can be adapted to the discrete-time models, such as the Ricker map. The estimation procedure is a nested optimization problem with three levels. Given λ and a current estimate $\hat{\boldsymbol{\theta}}$, the hidden states are estimated by minimizing an inner criterion

$$\begin{aligned}
\mathbf{n}_{1:T}^{\hat{\boldsymbol{\theta}}} &= \underset{\mathbf{n}_{1:T}}{\operatorname{argmin}}\, J(\mathbf{n}_{1:t} | \hat{\boldsymbol{\theta}}, \lambda) \\
&= \underset{\mathbf{n}_{1:T}}{\operatorname{argmin}} \left\{ - \sum_{t=1}^{T} \log p(\mathbf{y}_t^0 | \mathbf{n}_t, \hat{\boldsymbol{\theta}}) + \lambda \psi(\mathbf{n}_{1:T} | \hat{\boldsymbol{\theta}}) \right\},
\end{aligned}$$

where

$$\psi(\mathbf{n}_{1:T} | \hat{\boldsymbol{\theta}}) = \sum_{t=1}^{T} \{\mathbf{n}_t - E(\mathbf{n}_t | \mathbf{n}_{t-1}, \hat{\boldsymbol{\theta}})\}^2$$

quantifies deviations of the estimated state from the model, while λ determines the trade-off between data fitting and model compliance. The parameters are estimated using the higher level criterion

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, H(\boldsymbol{\theta} | \mathbf{n}_{1:T}^{\hat{\boldsymbol{\theta}}}, \lambda) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left\{ - \sum_{t=1}^{T} \log p(\mathbf{y}_t^0 | \mathbf{n}_t^{\hat{\boldsymbol{\theta}}}, \boldsymbol{\theta}) \right\}.
\end{aligned}$$

A further level can be added in which an outer grid search is used to select λ. This method is especially useful for exploring multimodality problems in Section 4.

3.3 Alternative Approaches

The methods described in the preceding sections represent a subset of those that could be used in the context of parameter estimation for nonlinear state space models. Here we discuss some of the alternatives, describe their relation with the methods described above and detail our reasons for not including them in this work.

There exist a large variety of particle-filtering-based methods that can be used to obtain approximate Maximum Likelihood (ML) estimates of the static parameters, such as Andrieu, Doucet and Tadic (2005), Andrieu and Doucet (2003), Malik and Pitt (2011), Poyiadjis, Doucet and Singh (2011) and Nemeth, Fearnhead and Mihaylova (2013). IF belongs to this class of methods, and we chose to include it, rather than some of the alternatives, in this work because (i) it is theoretically justified, as detailed in Ionides et al. (2011), (ii) it is has been tested on a variety of complex models, such as those described in King et al. (2008), He, Ionides and King (2010) and Bhadra et al. (2011), which are of direct interest to applied researchers in ecology and epidemiology, and (iii) the computational cost of a score function estimate is $O(M)$ in the number of particles, which, to the best of our knowledge, is state of the art. Hence, we argue that, by including IF, this work should adequately cover this class of methods.

Notably, this work does not include MCMC methods for parameter identification, such as those proposed by Carlin, Polson and Stoffer (1992), Geweke and Tanizaki (2001), Polson, Stroud and Müller (2008) and Niemi and West (2010). One reason for this is that highly nonlinear models, such as those considered here, are often characterized by strong dependencies between states and static parameters. Under such circumstances, implementing an efficient MCMC sampler requires the design of adequate conditional proposal densities, which is not trivial for nonlinear non-Gaussian models (Andrieu, Doucet and Holenstein, 2010; Kantas et al., 2014). In addition, the model presented in Section 5.3 is a discretized version of a continuous time model, where the discretization error was limited by using a large number of intermediate states between each pair of observations. Sampling this enlarged state space using standard MCMC methods would be challenging because the convergence rate of such schemes can be arbitrarily slow if the amount of augmentation is large (Roberts and Stramer, 2001). With the exception of Parameter Cascading, all the methods described in our work are less affected by this problem because the intermediate states are simply simulated forward using $p(\mathbf{n}_t|\mathbf{n}_{t-1}, \boldsymbol{\theta})$. This "plug-and-play" property is one of the reasons behind the popularity of these methods (Ionides et al., 2011).

Apart from PMCMC and MCMC algorithms, the methods proposed by Kitagawa (1998) and Liu and West (2001) could also be used to sample the posterior distribution of $\boldsymbol{\theta}$. Analogously to IF, these filters include the parameters in the state space, and perturb them using an artificial noise process. Even though Liu and West (2001) counteract the resulting overdispersion of the posterior by shrinking the perturbed parameters toward their mean, this does not entirely eliminate the information loss, if the posterior is far from Gaussian. Hence, in this work we preferred to target $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ using PMMH because of the convergence guarantees detailed in Andrieu and Roberts (2009). However, the computational cost of PMMH is fairly high, and the filter of Liu and West (2001) might be able to sample a close approximation to $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$, using far fewer filtering operations.

Finally, the versions of IF and PMMH used here are based on the SIR algorithm, as described in Gordon, Salmond and Smith (1993) and Doucet, Godsill and Andrieu (2000). More sophisticated filters, such as those proposed by Pitt and Shephard (1999) and Klaas, De Freitas and Doucet (2012), might provide more accurate estimates of the likelihood or of $\nabla p(\mathbf{y}_{1:T}|\boldsymbol{\theta})$ in the context of IF. Similarly, it might be possible to improve upon the MCMC implementation of ABC and SL used in Section 5 by using more sophisticated SMC samplers (Toni et al., 2009) or Gaussian Processes (Meeds and Welling, 2014), respectively. We do not explore these possibilities here, because doing so would increase the complexity of this work, without adding much to its main results.

4. MULTIMODALITY AND STATE SPACE METHODS

If the presence of process noise smooths the likelihood sufficiently, then methods that discard information should be outperformed by those that retain it. However, we cannot generally prove that the likelihood for any particular model is smoothed and, as shown in Section 2, there exist models for which smoothing is only partial, and may be inadequate, when process noise is low. In this section we further investigate the impact of multimodality on state space methods and show that information reduction methods can reduce the associated problems.

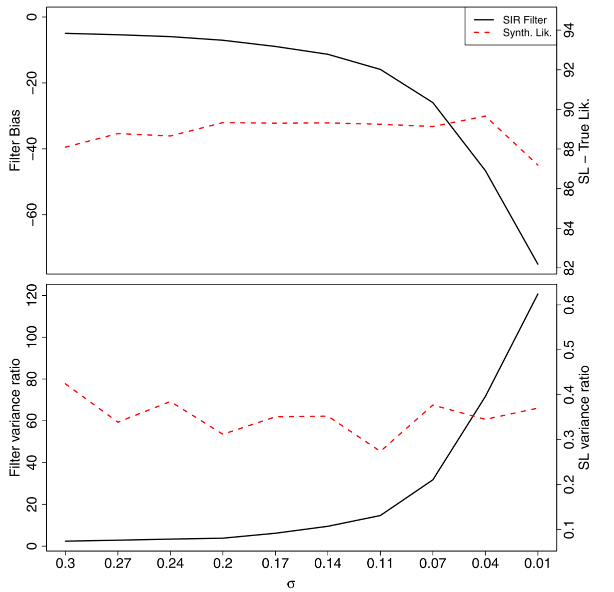

FIG. 4. *Top: average difference between the full likelihood and the estimated full (solid) or synthetic likelihood (dashed) as a function of* σ*, obtained using respectively the SIR filter and SL. Bottom: ratio between the sample variance of estimated full (solid) or synthetic (dashed) likelihoods and the true likelihood for several values of* σ*.*

In order to evaluate the accuracy of the likelihood estimates given by the SIR algorithm for different levels of noise, we used the discretized SSM described in Section 2 and in the supplementary material (Fasiolo, Pya and Wood, 2016). We chose ten levels of process noise in the interval $\sigma \in [0.01, 0.3]$. For each level we simulated 1000 paths using the Ricker map, with $\log(r) = 3.8$, $\phi = 0.5$, and evaluated the likelihood of each of them at the true parameters. Figure 4 shows the results.

The plot on the top shows that, as the process noise decreases, the average bias of the likelihood estimated by the filter (solid) increases in absolute value. Indeed, while the true log-likelihood (not shown) is roughly constant (≈ -70) for different levels of σ, the mean filter's estimates drop from -65 for $\sigma = 0.3$ to -140 for $\sigma = 0.01$. The strong dependence between likelihood bias and σ suggests that a sampler using these likelihood estimates will never explore areas of the parameter space where σ is low. In addition, any model comparison criterion based on the biased likelihood estimates is unreliable.

On the bottom of Figure 4 we plotted the ratios between sample variance of the likelihood estimated by the filter and the sample variance of the true likelihood for each value of σ, that is,

$$\frac{\hat{\mathrm{Var}}\{\log \hat{p}(\mathbf{y}_{1:50}|\boldsymbol{\theta})\}}{\hat{\mathrm{Var}}\{\log p(\mathbf{y}_{1:50}|\boldsymbol{\theta})\}}.$$

From the plot we see that the variance of the estimated log-likelihood increases exponentially as σ decreases, suggesting that Monte Carlo variability of the integration procedure dwarfs sampling variation for low σ. This has implications for algorithms based on particle filters: with such noisy likelihood estimates, the PMMH algorithm will have an extremely low acceptance rate (Doucet et al., 2012), while the IF procedure will become quite unstable, due to the high variability of the estimated gradients.

The broken lines in Figure 4 show corresponding quantities for the synthetic likelihood, obtained using the set of 13 summary statistics proposed by Wood (2010) and reported in the supplementary material (Fasiolo, Pya and Wood, 2016). Interestingly, both the average and the variance of the synthetic likelihood estimates remain roughly constant for different degrees of process noise. This suggests that the SL approach is quite robust to the level of process noise in the system, as it gives stable estimates also when the process dynamics are near-deterministic. On the other hand, the variance of the synthetic likelihood is lower than that of the true likelihood for any σ, which might be a consequence of the information loss.

Note that to use a synthetic likelihood when the system is (close to) deterministic, the initial values of the simulated paths have to be randomized [$N_1 \sim \mathrm{Unif}(0.1, 5)$], otherwise the variances of the summary statistics can be close to zero for very low process noise. Random initial values are consistent with the information reduction philosophy: inference should be robust to the particular values of the hidden states. In this context, we are confident that ABC, being based on summary statistics, would perform similarly to SL.

Figure 5 shows why the SIR algorithm is struggling to estimate the log-likelihood when σ is very low. Each of the 20 columns in the top image represents the true filtering density $p(n_t|\mathbf{y}_{1:t}, \boldsymbol{\theta})$ at each time step, when $\sigma = 0.3$. Areas of high density are represented in yellow, while areas of lower density are colored in red. With this level of process noise, the filtering densities are smooth and unimodal, so the filter places the particles around each mode, thus providing a reliable estimate of the likelihood. In contrast, the image on the bottom of Figure 5 shows that for very low process noise the filtering densities are unimodal in the first couple of time steps, but then they break into narrow

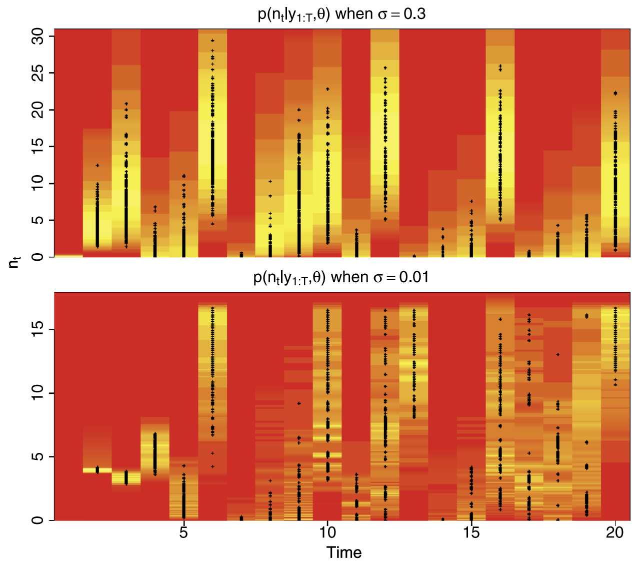

FIG. 5. *Filtering densities $p(n_t|\mathbf{y}_{1:t}, \boldsymbol{\theta})$ for a single Ricker path generated using* $\log(r) = 3.8$, $\phi = 10$ *and* $\sigma = 0.3$ (*top*) *or* $\sigma = 0.01$ (*bottom*).

multiple modes. Because of the irregularity of the filtering densities, the quality of the particle approximation is poor in this case (see time 19 in particular). The filter struggles to explore all the important modes of the filtering distributions, and hence the resulting estimates of the log-likelihood are very variable.

So Figure 5 helps to explain the variability in performance of the particle filter approach seen in Figures 3 and 4 as the process noise level changes. For models capable of showing chaotic or near-chaotic dynamics, there will be areas of the parameter space where the likelihood is highly multimodal. In these areas particle filtering methods will struggle to estimate the likelihood. In such situations most of the likelihood-based asymptotic theory will not be applicable, and even if it was possible to sample the corresponding parameter posterior exactly, it would not be obvious how the results should be interpreted. Hence, we argue that in such situations the use of approaches based on information reduction, which can provide a smooth proxy to likelihood, might be preferable from both a methodological and practical point of view.

To emphasize that the issue of multimodality is generic to the state space approach, rather than being specific to filtering, or a particular filtering implementation, or our discretized state space example, we illustrate how Parameter Cascading can encounter similar problems on the unmodified Ricker model. Figure 6 shows transects of the parameter fitting objective function, $H(\boldsymbol{\theta}|\mathbf{n}^{\boldsymbol{\theta}}_{1:T}, \lambda)$ (see Section 3.2.3), with respect to

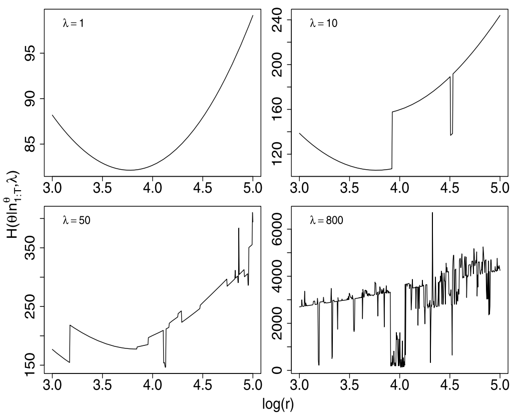

FIG. 6. *Transects of* $H(\boldsymbol{\theta}|\mathbf{n}_{1:T}, \lambda)$ *w.r.t.* $\log(r)$, *as* λ *increases.*

$\log(r)$ for four values of λ, and shows that this function becomes more irregular as λ increases. For large λ, which is appropriate when σ is low, this hinders the optimization and makes estimating $\boldsymbol{\theta}$ problematic. In the following we illustrate that jumps in the objective function correspond to transitions between modes of the objective function for the state, $J(\mathbf{n}_{1:T}|\boldsymbol{\theta}, \lambda)$.

The upper plot of Figure 7 shows other transects of $H(\boldsymbol{\theta}|\mathbf{n}^{\boldsymbol{\theta}}_{1:T}, \lambda)$, for $\lambda = 65$. The solid line was obtained using the same initial value $\mathbf{n}^{\boldsymbol{\theta}}_{1:T} = \mathbf{y}_{1:T}/\phi$ for each value of $\log(r)$. The dashed lines show the $H(\boldsymbol{\theta}|\mathbf{n}^{\boldsymbol{\theta}}_{1:T}, \lambda)$ curves corresponding to two different modes of $J(\mathbf{n}_{1:t}|\boldsymbol{\theta}, \lambda)$ and have been obtained by carefully tracking the modes. We refer to these modes as A and B. The plots on the bottom of Figure 7 represent the estimated hidden states $\mathbf{n}^{\theta}_{1:T}$ corresponding to two values of $\log(r)$ and to each mode. This shows that the same value of $\log(r)$ leads to two different modes

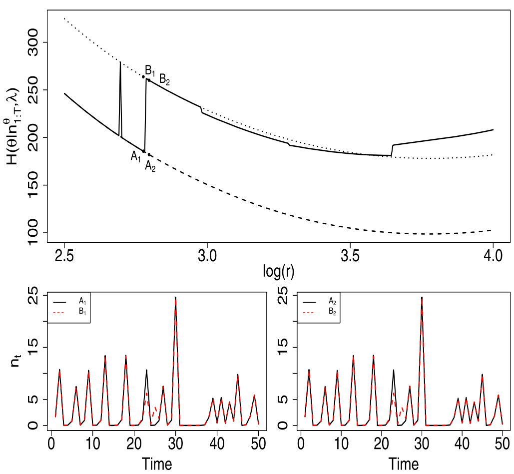

FIG. 7. *Top*: *transects of* $H(\boldsymbol{\theta}|\lambda, \mathbf{n}_t)$ *with respect to* $\log(r)$. *Bottom*: *paths corresponding to two points* 1 *or* 2 *along the* $\log(r)$ *axis and to modes A or B in the state space.*

in the state space, depending on the initialization. The similarity between the pairs A1–A2 and B1–B2 shows that these initialization-dependent modes are persistent along $\log(r)$.

5. PERFORMANCE COMPARISON

In the last section we saw that state space methods for highly nonlinear dynamic models can encounter difficulties in some regions of parameter space. Information reduction approaches might then be preferable, if they show little practical reduction in inferential performance when the dynamics are less problematic. This section therefore compares the relative performance of the statistical approaches presented by employing them to fit several models, using both simulated and real data sets.

5.1 Example 1: Simple Chaotic Maps with Sufficient Noise

Here we consider the models summarized in Table 1, in addition to the Ricker map. The parameter values of each model, reported in the supplementary material (Fasiolo, Pya and Wood, 2016), have been chosen so that the simulated paths show similar chaotic dynamics (Figure 8).

The data consist of 50 simulated paths $\mathbf{y}_{1:T}$, where $T = 50$, from each model. All paths were used to estimate the parameters using each method. For SL and for the ABC-MCMC algorithm of Marjoram et al. (2003) we have used 3×10^4 iterations to sample the posterior of each path. The PMMH algorithm had an extremely low acceptance rate unless the likelihood of the latest accepted position was re-estimated at each MCMC

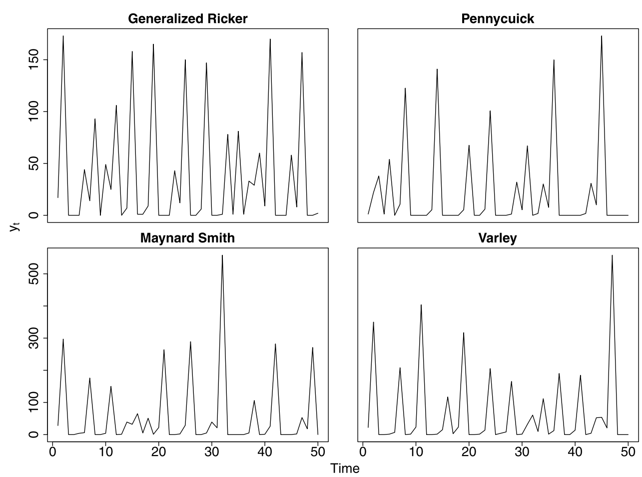

FIG. 8. *Trajectories simulated using the four models described in Table 1.*

step. This doubled the computational effort, and hence we used only 1.5×10^4 iterations for this method. To check if recomputing the likelihood was biasing the results in favor of PMMH, we have implemented a version of SL (labeled SL-R) that uses the same approach. For SL and ABC we have discarded 5000 iterations as burn-in, while for PMMH and SL-R 2500 iterations were discarded. For IF we have used 3000 optimization steps.

At each MCMC step, SL and PMMH estimated the (synthetic) likelihood by using 500 simulations from the model, while IF used 5000 simulations at each step of the optimization step. ABC simulates only one sample at each step, but we stored an iteration every 500. Notice that, with this setup, SL, SL-R, PMMH and ABC used the same number of simulations (1.5×10^7) from the model in order to fit each of the 250 simulated data sets. Given that the methods have very different implementation, basing the comparison on the number of simulations from the model, rather than CPU time, ensures fairness.

We used proper uniform priors for all parameters. IF does not support the use of priors, so we interpreted the priors as box constraints for the optimization. All methods were initialized at the same starting values which, together with the priors and other details, are included in the supplementary material (Fasiolo, Pya and Wood, 2016).

To choose the tolerance and the distance measure used by ABC-MCMC, we employed the following approach. For each model, we simulated $L = 10^5$ parameter vectors, $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_L$, from $p(\boldsymbol{\theta})$ and the corresponding statistics vectors, $\mathbf{s}_1, \dots, \mathbf{s}_L$, from $p(\mathbf{s}|\boldsymbol{\theta})$. As distance measure $d(\mathbf{s}, \mathbf{s}^0)$ we used $(\mathbf{s} - \mathbf{s}^0)^T \mathbf{Q}^{-1} (\mathbf{s} - \mathbf{s}^0)$, where $\mathbf{Q} = \text{diag}(\hat{\boldsymbol{\Sigma}})$, with $\hat{\boldsymbol{\Sigma}}$ being the empirical covariance matrix of the simulated statistics. We then calculated the distances $d(\mathbf{s}_i, \mathbf{s}^0)$, for $i = 1, \dots, L$, and we chose ε so that only 0.1% of the distances fell below this threshold.

We evaluated the accuracy of different approaches in term of squared errors between point estimates and the true parameters. While IF provided point estimates directly, ABC, SL and PMMH give dependent samples from the (approximate) parameter posteriors. Hence, for the latter group of methods we have used the posterior means as point estimates.

The supplementary material (Fasiolo, Pya and Wood, 2016) reports the median squared errors for each model-method-parameter combination. Here we have summarized the results in Figure 9 which represents, for each model and method, the median and Inter-Quartile Range of the squared errors, averaged geometrically across the parameters. Letting m, k, j and i be the indexes of model, method, data set and parameter respectively, the average squared errors are then given by

$$\bar{e}_j^{m,k} = \left\{ \prod_{i=1}^{p_m} (\hat{\theta}_{j,i}^{m,k} - \theta_i^m)^2 \right\}^{1/p_m},$$

where p_m is the parameter count for model m.

Figure 9 shows that, on this set of simple models, methods based on particle filtering consistently outperform methods based on information reduction. The performance of IF and PMMH is quite similar, and the differences in average squared errors between these two methods might be due to the different type of point estimates used. ABC-MCMC seems to perform better that either SL or SL-R for all models. This performance gap might be attributable to the normal approximation used by SL, to the bias entailed by estimating $p(\mathbf{s}_0|\boldsymbol{\theta})$ using a finite sample or simply to the particular setup we have used for the experiment.

Tuning the tolerance and the scaling matrix of ABC-MCMC required little extra effort for the simple models used here. However, the tuning tends to be much more laborious under more complex models, such as described in the following sections. In particular, when the number of unknown parameters is high, training ε and $\mathbf{Q}$ using simulations from the prior can be very inefficient, especially if the prior contains little information. Hence, for complex models, tuning ε and $\mathbf{Q}$ might require a more sophisticated approach, possibly involving some degree of manual intervention. From this practical perspective, SL is at an advantage because the summary statistics are scaled automatically using $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$, while no tolerance needs to be chosen.

The clear result here is that, given sufficient noise, the information reduction methods have noticeably worse performance than the state space methods for these simple toy models. In the next subsections we turn to more realistic examples. In order to limit the computational and programming effort, we will restrict our attention to PMMH and SL, that is, one method from each of the two inferential philosophies. We chose SL rather than ABC because the former method requires much less tuning, as discussed above. We selected PMMH over IF because PMMH and SL have very similar MCMC implementations, which should limit the influence of other implementational confounders on the results of the comparison.

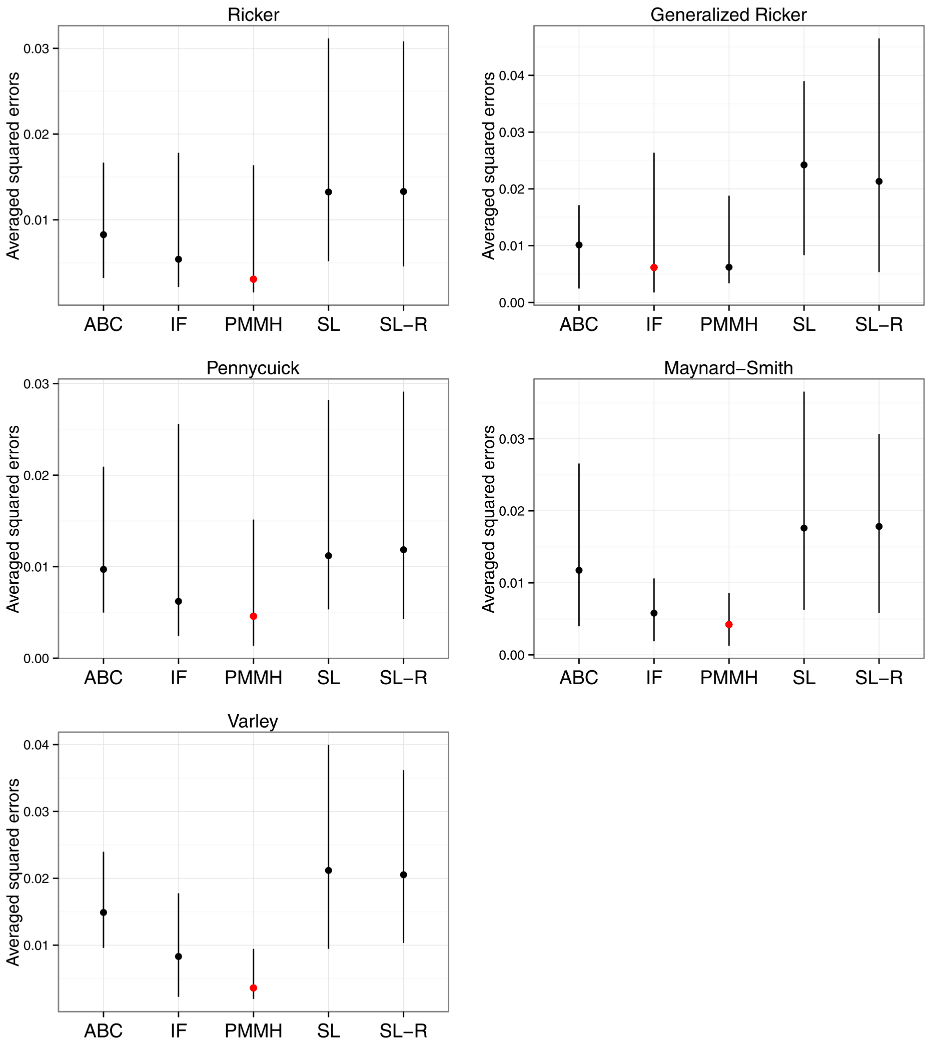

FIG. 9. *Medians and Inter-Quartile Ranges of the averaged squared errors for each model and method.*

5.2 Example 2: Nicholson's Blowflies

In this section we consider the results, reported by Nicholson (1954) and Nicholson (1957), of a series of laboratory experiments meant to elucidate the population dynamics of sheep blowfly *Lucilia cuprina* under resource limitation. Blowflies develop in four successive stages: eggs, larvae, pupae and adults. Feeding occurs only in the larval and adult stages. In two of the experiments (E1 and E2) the larvae had unlimited resources, while the adults had unlimited access to sugar and water, but were provided with a limited amount of protein, which is required for egg production. In another two experiments (E3 and E4) the larvae were supplied respectively with a moderately and severely restricted amount of food, while adults had unlimited resources. The resulting population dynamics are shown in the left column of Figure 10.

5.2.1 *The model*. A model potentially capable of explaining the observed dynamics of this population was proposed by Gurney, Blythe and Nisbet (1980), and it is represented by the following delayed differential equation:

$$\frac{dn(t)}{dt} = Pn(t-\tau)e^{-n(t-\tau)/n_0} - \delta n(t), \tag{5.1}$$

where n represents the adult population, while P, τ, n_0 and δ are parameters. In order to fit the model to the available data sets, Wood (2010) proposed a discretized version of equation (5.1) and added a stochastic component to its deterministic structure. More precisely, he proposed the following model:

$$n_t = r_t + s_t, \tag{5.2}$$

where

$$r_t \sim \text{Pois}\big(Pn_{t-\tau}e^{-n_{t-\tau}/n_0}e_t\big)$$

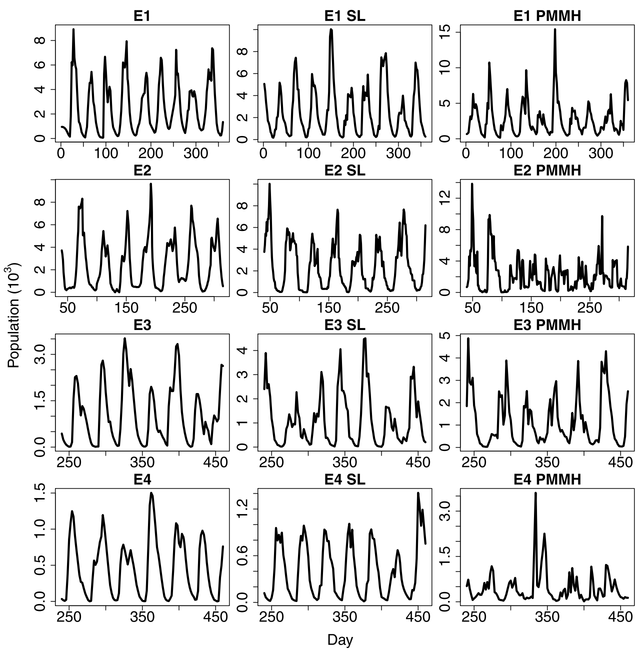

FIG. 10. *Left column: the data sets reported by* Nicholson (1954) *and* Nicholson (1957). *Central and right columns: paths simulated from model* (5.1) *using parameters equal to the posterior means obtained by fitting the four data sets using SL and PMMH.*

represents delayed recruitment process, while

$$s_t \sim \mathrm{binom}\big(e^{-\delta\varepsilon_t}, n_{t-1}\big)$$

denotes the adult survival process. Finally, e_t and ε_t are independent gamma distributed random variables, with unit means and variances equal to σ_p^2 and σ_d^2 respectively.

5.2.2 *Comparison using simulated data.* In order to verify the accuracy of SL and PMMH for the blowfly model, we have tested them on simulated data. Before moving to the results, notice that model (5.2) does not include any measurement noise: the number of blowflies n_t is assumed to be perfectly observed. This means that the model is not a SSM, hence, it cannot be fitted using methods based on particle filtering directly. Our solution has been to introduce an artificial measurement process when fitting the model using PMMH. More precisely, we use the following log-normal observational process:

$$\log y_t \sim \mathrm{N}\big(\log n_t, \sigma_o^2\big),$$

where the value of σ_o was predetermined, not estimated. Notice that, because of this modification, PMMH is fitting the wrong model and this procedure can be seen as an importance sampling ABC procedure, where σ_o plays the role of the tolerance. See Dean et al. (2011) for more details about the use of ABC procedures in the context of SSMs with intractable observational processes. Despite having introduced an artificial measurement process, we have decided to avoid estimating the initial values $n_1, \ldots, n_\tau$ when using PMMH, but we have fixed their values to that of the first τ observations.

For the comparison we have simulated 24 data sets of length $T = 200$, using parameter values $\delta = 0.16$, $P = 6.5$, $n_0 = 400$, $\sigma_p^2 = 0.1$, $\tau = 14$, $\sigma_d^2 = 0.1$. We have then estimated the parameters with both methods, using 2×10^4 MCMC iteration and 1000 simulations from the model at each step. The choice of σ_o was critical for the performance of PMMH. Obviously we would like σ_o to be as small as possible, but lowering it increases the variance of the importance weights and, in turn, of the estimated likelihood. In particular, if PMMH was initialized far from the true parameters, σ_o had to be increased in order to avoid particle depletion. Hence, we decided to include the results (PMMH0 and SL0) obtained using a realistic initialization ($\delta = 0.1$, $P = 4$, $n_0 = 200$, $\sigma_p^2 = 0.2$, $\tau = 10$, $\sigma_d^2 = 0.2$) and the results obtained by initializing the chains at the true parameters. In the first case σ_o was fixed to 0.05, while in the second to 0.01. For all parameters we used flat priors and for SL we used the set of 16 summary statistics proposed by Wood (2010) for this model. We report these details in the supplementary material (Fasiolo, Pya and Wood, 2016).

The running time of the two algorithms was very similar. In particular, when computed on one core of a 3.60 GHz i7-3820 CPU, single estimates of $p(\mathbf{y}^0|\boldsymbol{\theta})$ and $p(\mathbf{s}^0|\boldsymbol{\theta})$ took around 0.25 and 0.29 seconds, respectively.

The resulting Mean Squared Errors (MSEs) of the log-parameters are reported in Table 2. The table includes the p-values for differences in MSEs, which clearly show that PMMH is more accurate when the lower value of σ_o is used. On the other hand, in the more realistic setting the performance of the two procedures is more comparable, as PMMH underestimates both σ_p^2 and σ_d^2, while SL performs slightly worse than PMMH on the remaining parameters.

5.2.3 *Results using Nicholson's data sets.* Fitting Nicholson's data sets was relatively straightforward with SL, and we used the same initial values ($\delta = 0.16$, $P = 6.5$, $n_0 = 400$, $\sigma_p^2 = 0.1$, $\tau = 14$, $\sigma_d^2 = 0.1$) for each data set. Using this initialization was not possible for PMMH, as we would be forced to use values of σ_o as high as 0.2 in order to avoid failures in the Monte Carlo

TABLE 2

MSEs (coverage) of the log-parameters for SL and PMMH for the blowflies model for realistic (0) and optimistic (1) starting values. The p-values for the differences in log-absolute errors have been calculated using t-tests

	δ	P	n_0	σ_p^2	τ	σ_d^2
SL0	0.00598 (0.83)	0.01686 (0.83)	0.01032 (0.79)	0.05845 (1)	0.00123 (0.92)	0.18568 (0.96)
PMMH0	0.004 (0.67)	0.01176 (0.88)	0.00509 (0.88)	0.30579 (0.58)	0.00042 (0.92)	1.73206 (0.17)
p-value	0.414	0.197	0.01	0.359	0.03	<0.001
Best	PMMH0	PMMH0	PMMH0	SL0	PMMH0	SL0
SL1	0.00286 (0.83)	0.01929 (0.75)	0.00836 (0.88)	0.0634 (1)	0.00088 (0.96)	0.18419 (1)
PMMH1	0.00165 (0.88)	0.00416 (0.92)	0.00069 (0.92)	0.03322 (1)	1e–05 (1)	0.02965 (0.96)
p-value	0.123	0.006	< 0.001	0.058	0.006	<0.001
Best	PMMH1	PMMH1	PMMH1	PMMH1	PMMH1	PMMH1

integration step (i.e., all importance weights were going to zero). Hence, we initialized PMMH using values obtained through preliminary runs of SL on the four data sets. Still, we were forced to use values of σ_o equal to 0.1 for the second data set and 0.05 for the others. For each data set we used 3×10^4 MCMC iterations, of which the first 5000 were discarded as burn-in. The (synthetic) likelihood was estimated using 1000 particles or simulated paths at each step.

Figure 11 shows the stability diagrams for model (5.2), for each combination of data set and fitting procedure. These plots show how the stability properties of the system depend on the parameter combinations $P\tau$ and $\delta\tau$. All posterior samples obtained through SL lay strictly in the cyclic region of the parameter space, indicating that observed oscillations of the blowfly population are due to intrinsic blowfly biology, rather than stochastic perturbation of the system (Wood, 2010). On the other hand, the posteriors samples given by PMMH, in particular, those corresponding to data sets E2 and E4, are closer to the underdamped region, where the oscillations are driven by the stochasticity rather than intrinsic effects. With the exception of E1, the PMMH posteriors are more dispersed, which is attributable to the high estimates of noise parameters σ_d^2 and σ_p^2, as shown in Table 3.

Figure 10 compares the observed trajectories with those simulated from the model, using parameter values equal to the posterior means estimated by SL and PMMH. While using parameter values estimated through SL gives trajectories that are qualitatively similar to the observed ones in all cases, using the parameters estimated through PMMH gives a poor match for data sets E2 and E4.

To understand what happened, we have run a filtering operation using data set E2, 10^4 particles and parameters equal to the posterior mean given by SL and PMMH. Figure 12 shows the dynamics of the Effective Sample Size (ESS) using either parameter set. From the top plot we see the ESS drops to practically zero around the 25th, 95th and 250th observation, if SL estimates are used. On the other hand, PMMH gives much higher estimates of σ_p and σ_d and this keeps the ESS from dropping to zero in those occasions. This suggests that few idiosyncrasies or outliers in data sets E2 and E4 might be pushing PMMH toward the underdamped region. This is supported by the fact that, if PMMH is run using a log Student's t-distribution for the observational process

$$\frac{\log y_t - \log n_t}{\sigma_o} \sim \text{Student}(\nu = 2),$$

the resulting posterior estimates for E2 and E4 lay strictly inside the cyclic region, as shown in Figure 13. We comment on these results in Section 6.

5.3 Example 3: Cholera Epidemics in the Bay of Bengal

As a final example we consider a modified version of the Susceptible-Infected-Recovered-Susceptible (SIRS) model used by King et al. (2008) to explain cholera epidemics in the regions north of the Bay of Bengal. The data set considered here corresponds to cholera-related mortality records in the former Dacca district of the British East Indian province of Bengal, which is available within the *pomp* R-package (King, Nguyen and Ionides, 2015). The data, depicted in Figure 14, consists of monthly deaths count occurring between 1891 and 1941. See King et al. (2008) for additional details regarding the data.

5.3.1 *The model*. The model proposed by King et al. (2008) is composed of several classes, all of

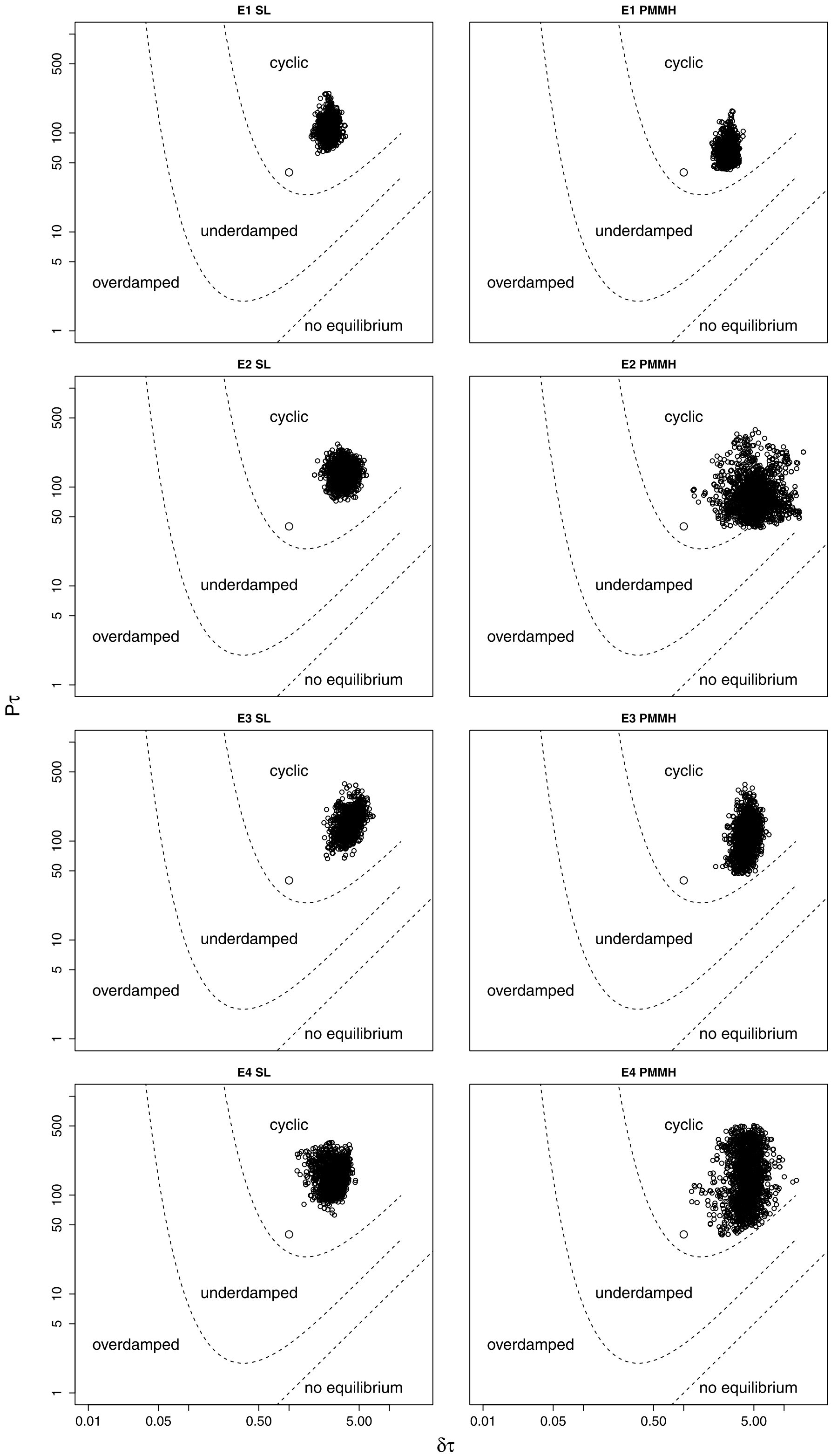

FIG. 11. *Stability plots for the blowfly model, obtained by fitting Nicholson's data sets using SL and PMMH. The black dots are* 2000 *values of the $P\tau$ and $\delta\tau$ randomly sampled from each MCMC chain. The white circle represents the initial value used for SL.*

TABLE 3
Posterior means for model (5.2), obtained by fitting each of Nicholson's data sets using either SL or PMMH

	δ	P	n_0	σ_p^2	τ	σ_d^2
E1 SL	0.17	7.57	395.30	0.70	14.44	0.47
E1 PMMH	0.19	4.45	653.93	1.54	14.82	0.30
E2 SL	0.22	8.70	407.61	0.21	15.95	1.77
E2 PMMH	0.37	6.26	576.30	2.35	15.02	3.47
E3 SL	0.29	10.48	184.38	0.64	14.62	0.55
E3 PMMH	0.28	7.71	229.32	1.56	15.18	0.53
E4 SL	0.22	12.81	59.16	0.71	12.91	0.55
E4 PMMH	0.30	12.10	88.33	2.42	14.46	1.23

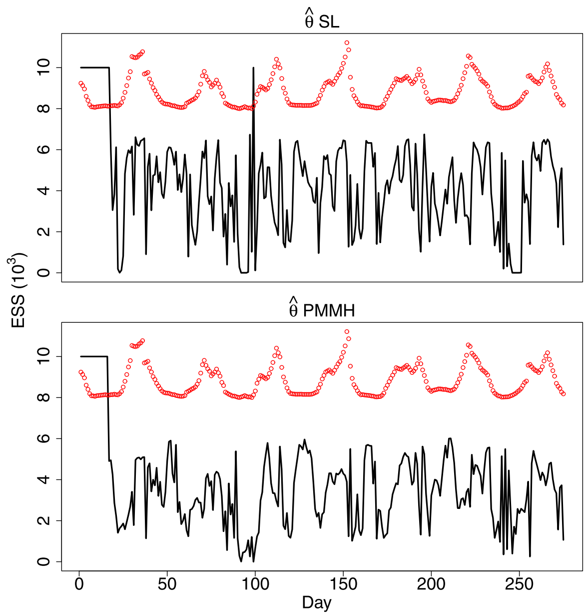

FIG. 12. *Dynamics of the ESS (lines) for the E2 data set (circles), using the parameter equal to the posterior means given by SL (top) and PMMH (bottom). For the first τ steps the ESS is equal to the number of particles because we have set $n_i = y_i$, for $i = 1, \dots, \tau$, as stated in the main text.*

which are completely unobserved apart from the infected class, which is observed indirectly through the deaths count. In King et al. (2008) the model was represented by a system of differential equations, which was solved numerically using a Euler–Maruyama scheme. The main issue with their formulation is that the positivity of the states is not guaranteed. To address this problem, we propose an alternative model formulation, to be justified later, which results in the following system of difference equations:

$$s_{t+1} = s_t - s_t^o + \frac{r_{kt}^o k\varepsilon}{k\varepsilon + \delta} + \frac{y_t^o \rho}{\rho + \delta} + b_{t+1},$$
$$i_{t+1} = i_t - i_t^o + c\frac{s_t^o \lambda_t}{\lambda_t + \delta},$$
$$y_{t+1} = y_t - y_t^o + (1 - c)\frac{s_t^o \lambda_t}{\lambda_t + \delta},$$
$$r_{1t+1} = r_{1t} - r_{1t}^o + \frac{i_t^o \gamma}{m + \gamma + \delta},$$
$$r_{it+1} = r_{it} - r_{it}^o + \frac{r_{i-1t}^o k\varepsilon}{k\varepsilon + \delta} \quad \text{for } i = 2, \dots, k,$$

where

$$b_{t+1} = p_{t+1} - p_t + \frac{s_t^o \delta}{\lambda_t + \delta} + \frac{i_t^o \delta}{m + \gamma + \delta} + \frac{y_t^o \delta}{\rho + \delta} + \sum_{i=1}^{k} \frac{r_{it}^o \delta}{k\varepsilon + \delta},$$

$$s_t^o = s_t\left(1 - e^{-(\lambda_t + \delta)\Delta t}\right), \qquad (5.3)$$
$$i_t^o = i_t\left(1 - e^{-(m+\gamma+\delta)\Delta t}\right),$$

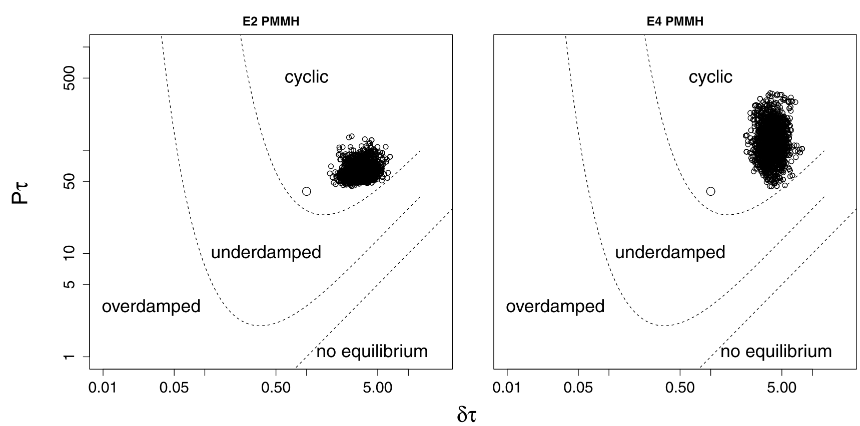

FIG. 13. *Stability plots for data sets E2 and E4 using PMMH with log Student's t observational error.*

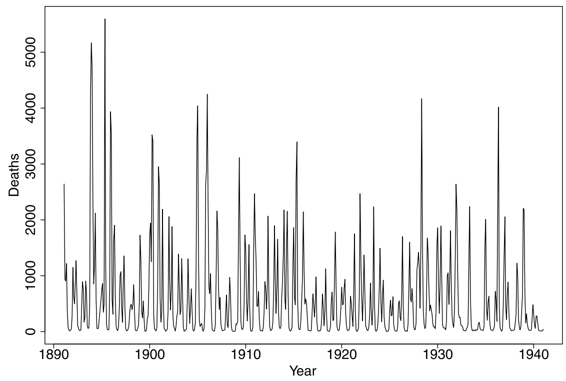

FIG. 14. *Cholera-related monthly deaths count in the Dacca district between* 1891 *and* 1941.

$$y_t^o = y_t\left(1 - e^{-(\rho+\delta)\Delta t}\right),$$
$$r_{it}^o = r_{it}\left(1 - e^{-(k\varepsilon+\delta)\Delta t}\right), \quad \text{for } i = 1, \dots, k.$$

Here b_{t+1} represents the number of births between time t and $t+1$, while p_t is the total population of the Dacca district at time t, characterized by a constant birth–death rate δ. Susceptible individuals s are infected by cholera at a time-varying rate λ_t, which will be explained in detail later. Parameter c determines the fraction of infected individuals that will undergo a full-blown infection, represented by class i, rather than an asymptomatic infection, represented by class y. Individuals in i suffer from an excess death rate m and transition to the first Recovered class r_1 with rate γ. On the other hand, individuals in y have the same death rate as susceptible individuals and do not acquire any long-term immunity, as they rejoin the s class directly at rate ρ. The duration of immunity is gamma distributed, with mean $1/\varepsilon$ and variance k/ε^2.

The rationale behind our discretized model needs to be clarified. Consider, for instance, y_t. To obtain y_{t+1} we model inputs and outputs involving y in turn, rather than simultaneously. Firstly, we obtain the number of individuals, y_t^o, leaving the asymptomatic infected class by solving

$$dy_s = -(\rho + \delta) y_s \, ds,$$

between t and $t+1$. The resulting solution is an exponential decay, which ensures the positivity of y_{t+1}. Then y_t^o is divided between b_{t+1} and s_{t+1}, with proportions determined by the output rates δ and ρ. This solution preserves the positivity of all classes and mass-balance, both of which are essential for a realistic model. In addition, our formulation becomes equivalent to the Euler–Maruyama scheme of King et al. (2008), as $\Delta t \to 0$.

The force of infection λ_t is given by

$$\lambda_t = \omega_t + e^{\beta t} \beta_t \frac{i_t}{p_t} \frac{\Delta w}{\Delta t}, \tag{5.4}$$

where $\Delta w \sim \Gamma(\Delta t/\sigma^2, 1/\sigma^2)$, so that $\Delta w/\Delta t$ represents multiplicative gamma noise with unit mean and variance equal to σ^2. We preferred this choice to the additive Gaussian noise originally used by King et al. (2008) because the multiplicative version assures the positivity of λ_t.

In (5.4), ω_t and β_t represent respectively the environmental and human feedback components of the force of infection

$$\omega_t = \exp\left(\sum_{i=1}^{6} \omega_i g_i(t)\right),$$
$$\beta_t = \exp\left(\sum_{i=1}^{6} \beta_i g_i(t)\right),$$

where $g_i(t)$, for $i = 1, \dots, 6$, are a periodic B-spline basis. Parameter β is the long-term trend in human-to-human transmission.

The observed number of deaths registered during the nth month is assumed to follow a negative binomial distribution

$$e_n \sim \mathrm{NB}\left(q_n, \frac{1}{\tau^2}\right),$$

with mean q_n and variance $q_n + q_n^2/\tau^2$, where q_n is the accumulated number of cholera-related deaths between the previous and the current month

$$q_n = \sum_{s=t_{n-1}}^{t_n} m i_s.$$

In the original model e_n was normally distributed around q_n, but that choice often produces negative death counts when the model is simulated. See King et al. (2008) for further model details.

5.3.2 *Setup and results using the dacca data set.* Similarly to King et al. (2008), we do not fit the full model, but we consider the following:

- a seasonal model where the y class is not included ($c = 1$);
- a two-path model were the environmental force of infection is constant [$\omega_s(t) = \omega_s$];
- a basic SIRS model where $c = 1$, $\omega_s(t) = \omega_s$ and $\beta_s(t) = \beta_s$.

TABLE 4
Estimated AICs and CPU times (sec) for each model, using SL and PMMH

Method	Seasonal	Two-paths	SIRS
AIC_{SL}	−38.4	−31.6	−34.6
AIC_{PMMH}	7458	7532.6	7528.2
CPU_{SL}	10	10.3	9.8
CPU_{PMMH}	9.6	10.1	9.4

We fitted each model to the Dacca data set using SL and PMMH. For both methods we used 1.4×10^6 MCMC iterations, the first half of which was discarded as a burn-in period, and 2000 simulations to estimate the (synthetic) likelihood at each step. We used uniform or diffuse priors for all parameters. We report them, together with the 26 summary statistics used by SL, in the supplementary material (Fasiolo, Pya and Wood, 2016).

Table 4 reports the estimated Akaike Information Criterion (AIC) and the time needed to obtain a single estimate of $p(\mathbf{y}^0|\boldsymbol{\theta})$ or $p(\mathbf{s}^0|\boldsymbol{\theta})$, on a single core of a 3.60 GHz i7-3820 CPU, for each model and method. SL and PMMH agree in selecting the seasonal reservoir model, while the two-paths mechanisms do not improve the fit enough, relatively to the SIRS model, to justify the additional complexity. This is in contrast with the results of King et al. (2008), whose second-order AIC estimate was lower for the two paths than for the SIRS model.

Almost all the marginal posterior variances were higher when SL was used, with a median increase equal to 7.2, 2.6 and 2.2 for the seasonal, two-paths and SIRS model, respectively. The variance increases were highest for the seasonal coefficients, $\omega_{1:6}$, of the force of infection, which suggest that the amount of information lost through the use of summary statistics is sizeable.

One important hypothesis examined by King et al. (2008) was that the mean duration of immunity, $d_L := 1/\varepsilon$, might be much shorter than previously thought. Our analysis partially supports this conclusion, as shown by Figure 15. The plots in the top row show the marginal densities of d_L under each model. Under the seasonal model, most of the posterior mass lies close to the lower prior boundary, corresponding to unrealistically low periods of immunity (shorter than one week). The posterior given by SL under the SIRS model is slightly less extreme, but it still suggests period of immunity of one to three months, which is much shorter than the 3 to 10 years timescale suggested by several sources (Cash et al., 1974; Glass et al., 1982; Koelle et al., 2005). One surprising result is that, under the two-paths model, d_L is still estimated to be lower than one month. This is in contrast with the results of King et al. (2008), who estimates d_L to be around 1.4 years, under the same model and data set. The mean duration of immunity after mild infections $d_S = 1/\rho$ is estimated to be shorter than three weeks under PMMH, while SL seems to have lost information regarding d_S, as the corresponding marginal posterior is bimodal and highly dispersed.

Figure 15 shows also the marginal distributions of the cholera-related death probability $f = m/(\delta + \gamma + m)$. Under the seasonal and the SIRS models our estimates roughly agree with those of King et al. (2008), but our fatality estimate is much higher than theirs when asymptomatic infections are included in the model. Similarly to King et al. (2008), we estimate the fraction of infection that are symptomatic to be very low under the two-path model.

Our results suggest that including asymptomatic infections does not improve the fit and does not provide more realistic estimates of immunity duration, following full-blown infections. In addition, this model is difficult to identify because there is a trade-off between parameters c, d_S and m, which is captured by Figure 16. The correlations observed in the PMMH joint posterior sample are explained by the fact that an increase in the fraction of individuals with full infection can be compensated by decreasing their mortality rate or by increasing the duration of long short-term immunity (thus delaying individuals with mild infection from rejoining the susceptible). Under SL this identifiability issue is more severe, and the corresponding posteriors are bimodal and more dispersed.

Another question addressed by King et al. (2008) is the relative importance of the environmental reservoir and of the human habitat for *V. Cholerae* persistence. They found that the basic reproductive number, R_0, which quantifies the strength of human-to-human transmission, was consistently low (around 1.5) across model and geographic area. Figure 15 shows that our estimates of R_0 are very low under all models and methods, thus supporting the hypothesis that humans might be only a marginal habitat for *V. Cholerae*.

6. DISCUSSION

We have described some of the difficulties that can be encountered when working with highly nonlinear

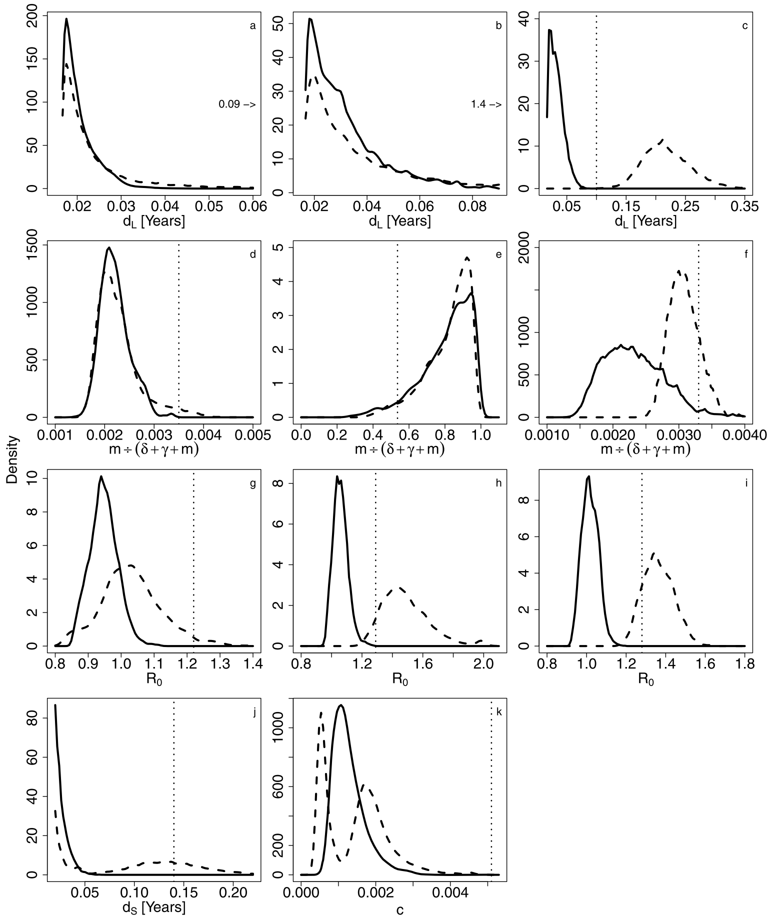

FIG. 15. *Posterior marginal distributions from PMMH* (*solid*) *and SL* (*dashed*). *The estimates of* King et al. (2008) *correspond to the vertical dotted lines*, *substituted by annotations when out of range. The first three rows contain the marginals of immunity duration after full-blown infections*, *fatality and basic reproductive number for the seasonal* (a), (d), (g), *two-paths* (b), (e), (h) *and SIRS* (c), (f), (i) *model. The last row shows the marginals of immunity duration after mild infections* (j) *and of the fraction of severe infections* (k) *for the two-paths model*.

dynamical models, and we have shown how these issues influence the performance of some popular inferential approaches. In particular, in Section 4 we have provided strong experimental evidence suggesting that, when the dynamics of the system are chaotic or near-chaotic, the likelihood function becomes increasingly multimodal as the process noise is reduced. While this directly undermines the performance of state space methods aiming at estimating the full likelihood, as in PMMH, or its derivatives, as in IF, approaches based on information reduction are less affected. This has practical implications because, in an applied setting, it is generally not known whether the best fitting parameters lay in an area of the parameter space where the stochasticity is too low for state space methods to work adequately. Hence, the ability of approaches based on information reduction to smooth the likelihood function, brought about by focusing on features of the data that are phase-independent, is appealing.

The blowflies example in Section 5.2 highlights the robustness of information reduction methods from a different perspective. Indeed, careless application of PMMH would have classified the dynamics of the system as nearly-underdamped under two of Nicholson's data sets, with the corresponding simulations from the model being clearly inconsistent with the data (see Figure 10). On the contrary, SL reliably classifies the dynamics as cyclic. In this example using a fat-tailed observation density mitigated the problem, but we argue that these results have deeper practical implications. Model (5.2) has sufficient flexibility to reproduce the main features (quantified by the summary statistics) of Nicholson's data sets, as demonstrated by Figure 10. On the other hand, the model struggles to explain certain nuances of Nicholson's data sets, and this is de-

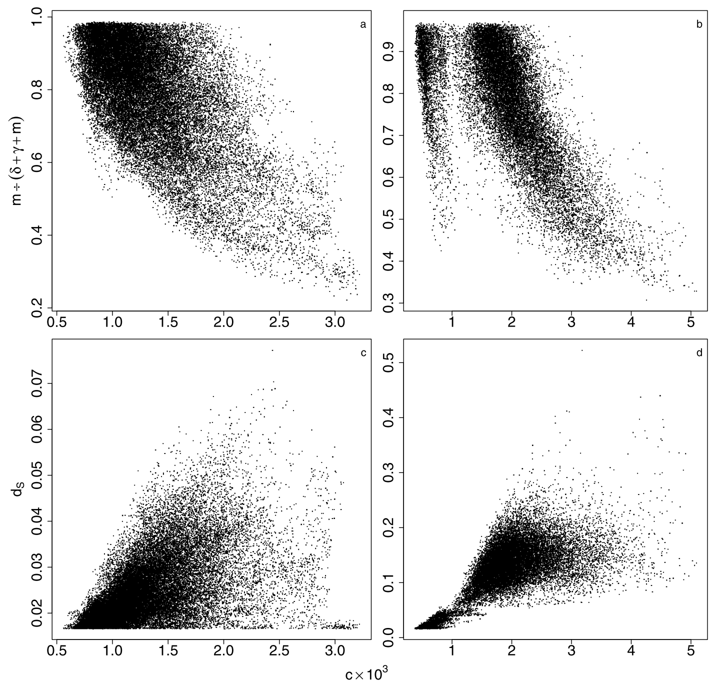

FIG. 16. *Joint posterior samples for fraction of symptomatic infections vs fatality and duration of short-term immunity under PMMH*(a), (c) *and SL*(b), (d).

tected by the particle filter, but overlooked by SL. This suggests that, in situations in which the model has a clear scientific interpretation but lacks the ability to explain the observed dynamics in all their complexity, focusing on some salient features of the data might be a reasonable approach. Conversely, if the model is believed to be an accurate description of the system under study, or if it is meant to be used for the purpose of state estimation or forecasting, then it is compelling to fit it using the full data.

Another lesson learned from the blowflies example is that, for particle-filtering-based methods to work properly, a good initialization is often indispensable. This is because these methods are generally based on some form of importance sampling, hence, when the initial estimates are far from the best fitting parameters most of the importance weights go to zero (particle depletion). In this context, methods based on information reduction can be useful because they are robust to bad initializations. Methods that can provide reliable initial estimates, to be fed to more accurate but less robust methods, are of high practical value, but often underrepresented in the literature. Exceptions are Lavine et al. (2013), who, in the context of pertussis epidemics, use SL to initialize an IF algorithm, and Owen, Wilkinson and Gillespie (2014), who propose to initialize PMMH using the output of preliminary ABC runs.

One recurrent theme in our examples is that reducing the data to a set of summary statistics generally entails a loss of accuracy in parameter estimation. This is particularly clear in Section 5.1, where SL and ABC are consistently outperformed by PMMH and IF in terms of MSEs. Mild losses of accuracy are often acceptable when parameter estimation is not the main focus of analysis, but the aim is, for example, to determine whether the dynamics of the system are stable or oscillatory, as in the blowflies example. On the other hand, when dealing with models that are weakly identified even under the full data, as in Section 5.3, any further loss of information can lead to unreliable estimates. Hence, an important drawback of information reduction methods is that, in the absence of a benchmark, quantifying inferential inaccuracies requires running simulation studies, which can be prohibitively expensive for complex models, such as those presented in Section 5.3. While in all the examples presented in this study one or more benchmarks were available, this is not always the case.

All the methods described in this work, with the exception of Parameter Cascading, are computationally intensive. In particular, obtaining pointwise estimates of $p(\mathbf{y}^0_{1:T}|\boldsymbol{\theta})$ or $\nabla p(\mathbf{y}^0_{1:T}|\boldsymbol{\theta})$ requires MT simulations, where M is the number of particles, from $p(\mathbf{n}_t|\mathbf{n}_{t-1}, \theta)$ under SIR and IF respectively. Similarly, SL uses N simulations from $p(\mathbf{y}_{1:T}|\boldsymbol{\theta})$ to estimate $p(\mathbf{s}^0|\boldsymbol{\theta})$. Within PMMH and the MCMC implementation of SL, this price has to be paid at each iteration and the efficiency of the sampler will depend on the trade-off between the variance of likelihood estimates and the number of simulations used to obtain them (Sherlock et al., 2015). Similar considerations hold for IF, but the optimizer generally needs much fewer iterations to reach convergence. On the other hand, IF does not directly provide parameter uncertainty estimates, which have to be obtained through an expensive likelihood profiling procedure (see Ionides, Bretó and King, 2006). On first sight, ABC samplers seem more efficient than the above approaches because they target $p(\boldsymbol{\theta}|\mathbf{s}^0)$ directly by simulating a single statistics vector at the time. However, ABC samplers generally have a very low acceptance rate because the latter increases with the tolerance ε, while their accuracy is inversely proportional to it.

These computational issues are aggravated by the curse of dimensionality. In particular, the number of particles in a particle filter need to increase super-exponentially with the number of hidden states in order to avoid particle-depletion (Snyder et al., 2008). This result applies directly to PMMH and IF. Analogously, the computational cost of a method based on information reduction typically increases with the number of summary statistics used (d). In ABC methods, the MSE of the posterior moments estimate decreases at rate $O(e^{-4/d+5})$, due to the nonparametric approximation used by such methods (Blum, 2010). SL scales better with d because it requires a number of simulations sufficient to estimate the $O(d^2)$ entries of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$. However, its Gaussian assumption might hold only approximately.

Summary statistics selection is, in our opinion, an open problem, as many approaches proposed in the literature require the user to specify an initial set of summary statistics which can then be refined upon (see, e.g., Blum et al., 2013; Fearnhead and Prangle, 2012 or Nunes and Balding, 2010). While some fairly general approaches exist (Drovandi, Pettitt and Lee, 2015), finding a set of initial statistics under which the model is identifiable is, at the time of writing, a time consuming, problem dependent and largely nonautomated process. In the context of models with several hidden states, devising summary statistics is particularly difficult because these have to capture the relation between all the states, while being based only on (noisy proxies of) a subset of them. The two-path cholera model is a perfect example of this problem: out of seven state variables, only one, the number of infected, is observed with noise.

Taken together, our results lead us to some very practical conclusions. When faced with a real nonlinear dynamic system for which good models are available, one should ideally use a state space method for final parameter estimation, combined with a minimum tuning information reduction approach for exploration of alternative model structures, initialization and checking of conclusions. Using state space methods alone may bias conclusions toward noise-driven stable dynamics, while using information reduction alone may lead to inference that is less precise than it could be. If the model is only attempting to explain some features of the system, and not every detail of the data, then information reduction is probably essential.

ACKNOWLEDGMENTS.

The authors would like to thank two anonymous reviewers for comments that helped us to improve this paper, Aaron King and Ed Ionides for useful discussion, and Chris Jennison for commenting on earlier versions of this work and for the suggestion that led to the discretized Ricker model.

Most of this work was undertaken at the University of Bath, where M.F. was a Ph.D. student, and it was supported in part by EPSRC Grants EP/I000917 and EP/K005251/1.

SUPPLEMENTARY MATERIAL

Supplement to "A Comparison of Inferential Methods for Highly Nonlinear State Space Models in Ecology and Epidemiology" (DOI: 10.1214/15-STS534SUPP; .pdf). The supplement describes how the likelihood of a discrete SSM can be computed exactly and it contains additional details regarding the examples considered in Section 5.

REFERENCES

ANDERSON, C. N. K., HSIEH, C.-H., SANDIN, S. A., HEWITT, R., HOLLOWED, A., BEDDINGTON, J., MAY, R. M. and SUGIHARA, G. (2008). Why fishing magnifies fluctuations in fish abundance. *Nature* **452** 835–839.

ANDRIEU, C. and DOUCET, A. (2003). Online expectation-maximization type algorithms for parameter estimation in general state space models. In *IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP '03*) **6** 69–72. IEEE, New York.

ANDRIEU, C., DOUCET, A. and HOLENSTEIN, R. (2010). Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 269–342. MR2758115

ANDRIEU, C., DOUCET, A. and TADIC, V. B. (2005). On-line parameter estimation in general state–space models. In *Decision and Control*, 2005 *and* 2005 *European Control Conference. CDC-ECC'*05. 44*th IEEE Conference on* 332–337. IEEE, Piscataway, NJ.

ANDRIEU, C. and ROBERTS, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.* **37** 697–725. MR2502648

BEAUMONT, M. A., ZHANG, W. and BALDING, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162** 2025–2035.

BERLINER, L. M. (1992). Statistics, probability and chaos. *Statist. Sci.* **7** 69–122. MR1173418

BHADRA, A., IONIDES, E. L., LANERI, K., PASCUAL, M., BOUMA, M. and DHIMAN, R. C. (2011). Malaria in Northwest India: Data analysis via partially observed stochastic differential equation models driven by Lévy noise. *J. Amer. Statist. Assoc.* **106** 440–451. MR2866974

BLUM, M. G. B. (2010). Approximate Bayesian computation: A nonparametric perspective. *J. Amer. Statist. Assoc.* **105** 1178–1187. MR2752613

BLUM, M. G. B., NUNES, M. A., PRANGLE, D. and SISSON, S. A. (2013). A comparative review of dimension reduction methods in approximate Bayesian computation. *Statist. Sci.* **28** 189–208. MR3112405

CARLIN, B. P., POLSON, N. G. and STOFFER, D. S. (1992). A Monte Carlo approach to nonnormal and nonlinear state–space modeling. *J. Amer. Statist. Assoc.* **87** 493–500.

CASH, R. A., MUSIC, S. I., LIBONATI, J. P., CRAIG, J. P., PIERCE, N. F. and HORNICK, R. B. (1974). Response of man to infection with Vibrio cholerae. II. Protection from illness afforded by previous disease and vaccine. *J. Infectous Diseases* **130** 325–333.

CHAN, K.-S. and TONG, H. (2001). *Chaos*: *A Statistical Perspective*. Springer, New York. MR1851668

DEAN, T. A., SINGH, S. S., JASRA, A. and PETERS, G. W. (2011). Parameter estimation for hidden Markov models with intractable likelihoods. Preprint. Available at arXiv:1103.5399.

DESHARNAIS, R. A., COSTANTINO, R. F., CUSHING, J. M., HENSON, S. M. and DENNIS, B. (2001). Chaos and population control of insect outbreaks. *Ecology Letters* **4** 229–235.

DOUCET, A., GODSILL, S. and ANDRIEU, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Stat. Comput.* **10** 197–208.

DOUCET, A. and JOHANSEN, A. M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. In *The Oxford Handbook of Nonlinear Filtering* 656–704. Oxford Univ. Press, Oxford. MR2884612

DOUCET, A., PITT, M., DELIGIANNIDIS, G. and KOHN, R. (2012). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. Preprint. Available at arXiv:1210.1871.

DROVANDI, C. C., PETTITT, A. N. and LEE, A. (2015). Bayesian indirect inference using a parametric auxiliary model. *Statist. Sci.* **30** 72–95. MR3317755

EARN, D. J., ROHANI, P. and GRENFELL, B. T. (1998). Persistence, chaos and synchrony in ecology and epidemiology. *Proc. R. Soc. Lond. Ser. B*: *Biol. Sci.* **265** 7–10.

FASIOLO, M., PYA, N. and WOOD, S. N. (2016). Supplement to "A Comparison of Inferential Methods for Highly Nonlinear State Space Models in Ecology and Epidemiology." DOI:10.1214/15-STS534SUPP.

FEARNHEAD, P. and PRANGLE, D. (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 419–474. MR2925370

GEWEKE, J. and TANIZAKI, H. (2001). Bayesian estimation of state–space models using the Metropolis–Hastings algorithm within Gibbs sampling. *Comput. Statist. Data Anal.* **37** 151–170. MR1856209

GLASS, R. I., BECKER, S., HUQ, M. I., STOLL, B. J., KHAN, M., MERSON, M. H., LEE, J. V. and BLACK, R. E. (1982). Endemic cholera in rural Bangladesh, 1966–1980. *American. J. Epidemiology.* **116** 959–970.

GORDON, N. J., SALMOND, D. J. and SMITH, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proc. F* (*Radar and Signal Process.*) **140** 107–113. IET, Stevenage.

GRENFELL, B. T. (1992). Chance and chaos in measles dynamics. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 383–398.

GRENFELL, B. T., BJØRNSTAD, O. N. and FINKENSTÄDT, B. F. (2002). Dynamics of measles epidemics: Scaling noise, determinism, and predictability with the TSIR model. *Ecological Monograph* **72** 185–202.

GRENFELL, B. T., KLECZKOWSKI, A., GILLIGAN, C. A. and BOLKER, B. M. (1995). Spatial heterogeneity, nonlinear dynamics and chaos in infectious diseases. *Stat. Methods Med. Res.* **4** 160–183.

GURNEY, W. S. C., BLYTHE, S. P. and NISBET, R. M. (1980). Nicholson's blowflies revisited. *Nature* **287** 17–21.

HE, D., IONIDES, E. L. and KING, A. A. (2010). Plug-and-play inference for disease dynamics: Measles in large and small populations as a case study. *Journal of the Royal Society Interface* **7** 271–283.

IONIDES, E. L., BRETÓ, C. and KING, A. A. (2006). Inference for nonlinear dynamical systems. *Proc. Natl. Acad. Sci. USA* **103** 18438–18443.

IONIDES, E. L., BHADRA, A., ATCHADÉ, Y. and KING, A. (2011). Iterated filtering. *Ann. Statist.* **39** 1776–1802. MR2850220

KANTAS, N., DOUCET, A., SINGH, S. S., MACIEJOWSKI, J. M. and CHOPIN, N. (2014). On particle methods for parameter estimation in state–space models. Preprint. Available at arXiv:1412.8695.

KAUSRUD, K. L., MYSTERUD, A., STEEN, H., VIK, J. O., ØSTBYE, E., CAZELLES, B., FRAMSTAD, E., EIKESET, A. M., MYSTERUD, I., SOLHØY, T. et al. (2008). Linking climate change to lemming cycles. *Nature* **456** 93–97.

KENDALL, B. E., ELLNER, S. P., MCCAULEY, E., WOOD, S. N., BRIGGS, C. J., MURDOCH, W. W. and TURCHIN, P. (2005). Population cycles in the pine looper moth: Dynamical tests of mechanistic hypotheses. *Ecological Monograph* **75** 259–276.

KING, A. A., IONIDES, E. L., PASCUAL, M. and BOUMA, M. J. (2008). Inapparent infections and cholera dynamics. *Nature* **454** 877–880.

KING, A. A., NGUYEN, D. and IONIDES, E. L. (2015). Statistical Inference for Partially Observed Markov Processes via the R Package pomp. Preprint. Available at arXiv:1509.00503.

KITAGAWA, G. (1998). A self-organising state–space model. *J. Amer. Statist. Assoc.* **93** 1203–1215.

KLAAS, M., DE FREITAS, N. and DOUCET, A. (2012). Toward practical N2 Monte Carlo: The marginal particle filter. Preprint. Available at arXiv:1207.1396.

KOELLE, K., RODÓ, X., PASCUAL, M., YUNUS, M. and MOSTAFA, G. (2005). Refractory periods and climate forcing in cholera dynamics. *Nature* **436** 696–700.

LAVINE, J. S., KING, A. A., ANDREASEN, V. and BJØRNSTAD, O. N. (2013). Immune boosting explains regime-shifts in prevaccine-era pertussis dynamics. *PloS One* **8** e72086.

LIU, J. and WEST, M. (2001). Combined parameter and state estimation in simulation-based filtering. In *Sequential Monte Carlo Methods in Practice* 197–223. Springer, New York. MR1847793

MALIK, S. and PITT, M. K. (2011). Particle filters for continuous likelihood evaluation and maximisation. *J. Econometrics* **165** 190–209. MR2846644

MARJORAM, P., MOLITOR, J., PLAGNOL, V. and TAVARE, S. (2003). Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **100** 15324–15328.

MEEDS, E. and WELLING, M. (2014). GPS-ABC: Gaussian process surrogate approximate Bayesian computation. Preprint. Available at arXiv:1401.2838.

NEMETH, C., FEARNHEAD, P. and MIHAYLOVA, L. (2013). Particle approximations of the score and observed information matrix for parameter estimation in state space models with linear computational cost. Preprint. Available at arXiv:1306.0735.

NICHOLSON, A. J. (1954). An outline of the dynamics of animal populations. *Aust. J. Zoology* **2** 9–65.

NICHOLSON, A. J. (1957). The self-adjustment of populations to change. In *Cold Spring Harbor Symposia on Quantitative Biology* **22** 153–173. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.

NIEMI, J. and WEST, M. (2010). Adaptive mixture modeling Metropolis methods for Bayesian analysis of nonlinear state–space models. *J. Comput. Graph. Statist.* **19** 260–280. MR2758305

NUNES, M. A. and BALDING, D. J. (2010). On optimal selection of summary statistics for approximate Bayesian computation. *Stat. Appl. Genet. Mol. Biol.* **9** Art. 34, 16. MR2721714

OWEN, J., WILKINSON, D. J. and GILLESPIE, C. S. (2014). Likelihood free inference for Markov processes: A comparison. Preprint. Available at arXiv:1410.0524.

PITT, M. K. and SHEPHARD, N. (1999). Filtering via simulation: Auxiliary particle filters. *J. Amer. Statist. Assoc.* **94** 590–599. MR1702328

POLSON, N. G., STROUD, J. R. and MÜLLER, P. (2008). Practical filtering with sequential parameter learning. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 413–428. MR2424760

POYIADJIS, G., DOUCET, A. and SINGH, S. S. (2011). Particle approximations of the score and observed information matrix in state space models with application to parameter estimation. *Biometrika* **98** 65–80. MR2804210

RAMSAY, J. O., HOOKER, G., CAMPBELL, D. and CAO, J. (2007). Parameter estimation for differential equations: A generalized smoothing approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 741–796. MR2368570

ROBERTS, G. O. and STRAMER, O. (2001). On inference for partially observed nonlinear diffusion models using the Metropolis–Hastings algorithm. *Biometrika* **88** 603–621. MR1859397

SHERLOCK, C., THIERY, A. H., ROBERTS, G. O. and ROSENTHAL, J. S. (2015). On the efficiency of pseudo-marginal random walk Metropolis algorithms. *Ann. Statist.* **43** 238–275. MR3285606

SNYDER, C., BENGTSSON, T., BICKEL, P. and ANDERSON, J. (2008). Obstacles to high-dimensional particle filtering. *Monthly Weather Review* **136** 4629–4640.

TONI, T., WELCH, D., STRELKOWA, N., IPSEN, A. and STUMPF, M. P. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Inter.* **6** 187–202.

TURCHIN, P. and ELLNER, S. P. (2000). Living on the edge of chaos: Population dynamics of Fennoscandian voles. *Ecology* **81** 3099–3116.

WOOD, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* **466** 1102–1104.

YANG, G.-J., BRADSHAW, C. J., WHELAN, P. I. and BROOK, B. W. (2008). Importance of endogenous feedback controlling the long-term abundance of tropical mosquito species. *Population Ecology* **50** 293–305.