Supporting Information

Ionides et al. 10.1073/pnas.1410597112

SI Text

Weak Convergence for Occupation Measures

We study the convergence of the processes $\{W_\sigma(t), 0 \le t \le 1\}$ toward $\{W(t), 0 \le t \le 1\}$ as $\sigma \to 0$ for Theorem 2. We are interested in showing that the fraction of time $\{W_\sigma(t)\}$ spends in a set $\Theta_0 \subset \Theta$ over the discrete set of times $\{k\sigma^2, k = 1, \ldots, 1/\sigma^2\}$ converges in distribution to the fraction of time $\{W(t)\}$ spends in Θ_0. We choose $\{W_\sigma(t)\}$ to be a right-continuous step function approximation to a diffusion to simplify the relationship between the occupancy fraction over the discrete set of times and over the continuous interval. However, this simplification requires us to work with convergence to $\{W(t)\}$ in a space of processes with discontinuous sample paths, leading us to work with a Skorokhod topology.

Let $D_p[0,1]$ be the space of $\mathbb{R}^p$-valued functions on $[0,1]$ which are right continuous with left limits. Let $X = \{X(t)\}_{t\in[0,1]}$ and $\{X_n(t)\}_{t\in[0,1]}$, $n \ge 1$, be stochastic processes with paths in $D_p[0,1]$. Let $\Rightarrow$ denote weak convergence, and suppose that $X_n \Rightarrow X$ as $n \to \infty$ in $D_p[0,1]$ equipped with the strong Skorokhod J_1 topology (1).

Proposition S1 (Proposition VI.1.17 of ref. 1). *If X has continuous paths, then $X_n \Rightarrow X$ as $n \to \infty$ in the space $D_p[0,1]$ equipped with the uniform metric.*

Suppose that $f : \mathbb{R}^p \to \mathbb{R}$ is Borel measurable function and define the map $T_f : D_p[0,1] \to \mathbb{R}$

$$T_f(x) := \int_0^1 f(x(t))dt, \quad x \in D_p[0,1].$$

Now, let $\mathrm{Disc}(T_f)$ denote the set of discontinuity points of T_f, let $C_p[0,1]$ be the space of $\mathbb{R}^p$-valued continuous functions on $[0,1]$, and write Leb for Lebesgue measure.

Proposition S2. *Suppose that f is bounded. We have that*

$$\mathrm{Disc}(T_f) \cap C_p[0,1] \subset \{x \in C[0,1] : \mathrm{Leb}(\{t \in [0,1] : x(t) \in \mathrm{Disc}(f)\}) > 0\} =: D_f. \quad \textbf{[S1]}$$

Proof. Suppose that $x \in C_p[0,1]$ does not belong to the right–hand side of Eq. **S1** and let $x_n \to x$ in J_1. Then, according to a standard property of the Skorokhod J_1 topology (1), we also have $\sup_{t\in[0,1]} |x_n(t) - x(t)| \to 0$, as $n \to \infty$. Now, since $x \notin D_f$, we have that for almost all $t \in [0,1]$, the point $x(t)$ is a continuity point of f. Therefore, $f(x_n(t)) \to f(x(t))$, $n \to \infty$, for almost all $t \in [0,1]$. Since f is bounded, the Lebesgue dominated convergence theorem then yields

$$T_f(x_n) \equiv \int_0^1 f(x_n(t))dt \to \int_0^1 f(x(t))dt \equiv T_f(x), \quad \text{as } n \to \infty.$$

This completes the proof.

In the context of stochastic processes, by the Continuous Mapping Theorem, we have convergence in distribution,

$$T_f(X_n) \xrightarrow{d} T_f(X), \quad \text{as } n \to \infty,$$

provided X has continuous paths and $\mathbb{P}(X \in \mathrm{Disc}(f)) = 0$. In the case when $f(x) = 1_A(x)$, the latter translates to

$$\mathbb{P}\{\text{The measure of the time } X \text{ spends on the boundary of } A \text{ is zero}\} = 1. \quad \textbf{[S2]}$$

If the stochastic process has continuous marginal distribution and the set A has zero boundary, the Fubini's theorem readily implies Eq. **S2**. Indeed, the probability in Eq. **S2** equals

$$\int_\Omega \int_0^1 1_{\partial A}(X(t,\omega))dt\, \mathbb{P}(d\omega) = \int_0^1 \mathbb{P}(X(t) \in \partial A)dt = 0,$$

provided that $\mathrm{Leb}(\partial A) = 0$ and if $X(t)$ has a marginal density for each $t \in (0,1)$. The above arguments lead to the proof of the following result.

Lemma S1. *Suppose that $X_n \Rightarrow X$ in $D_p[0,1]$, equipped with the uniform convergence topology. If the process X takes values in $C_p[0,1]$ and has continuous marginal distributions, then for all bounded Borel functions $f : \mathbb{R}^p \to \mathbb{R}$, that are continuous almost everywhere, i.e., such that $\mathrm{Leb}(\mathrm{Disc}(f)) = 0$, we have*

$$\int_0^1 f(X_n(t))dt \xrightarrow{d} \int_0^1 f(X(t))dt, \quad \textit{as } n \to \infty.$$

Iterated Importance Sampling

When $N = 1$ in IF2, we obtain a general latent variable algorithm in which each iteration involves importance sampling but not filtering. This situation is called iterated importance sampling (2) and we call this special case of our algorithm IIS2. Iterated importance sampling has previously been used to provide a route into proving convergence of iterated filtering (2, 3). However, in this article, we found it more convenient to prove the full result for iterated filtering directly. Although IIS2 may have some independent value as a practical algorithm, our only use of IIS2 in this article is to provide a convenient environment for explicit computations for Gaussian models in *Gaussian and Near-Gaussian Analysis* and non-Gaussian models in *A Class of Exact Non-Gaussian Limits*.

Algorithm IIS2. Iterated importance sampling

```
input:
Simulator for $f_X(x;\theta)$                              Evaluator for $f_{Y|X}(y|x;\theta)$
Data, $y*$                                                 Number of iterations, $M$
Initial parameter swarm, $\{\Theta_j^0, j \text{ in } 1:J\}$     Number of particles, $J$
Perturbation density, $h(\theta|\varphi;\sigma)$                 Perturbation sequence, $\sigma_{1:M}$
output: Final parameter swarm, $\{\Theta_j^M, j \text{ in } 1:J\}$
For $m$ in $1:M$
   $\Phi_j^m \sim h(\theta|\Theta_j^{m-1};\sigma_m)$ for $j$ in $1:J$
   $X_j^m \sim f_X(x;\Phi_j^m)$ for $j$ in $1:J$
   $w_j^m = f_{Y|X}(y^*|X_j^m;\Phi_j^m)$ for $j$ in $1:J$
   Draw $k_{1:J}$ with $\mathbb{P}(k_j = i) = w_{n,i}^m / \sum_{u=1}^J w_{n,u}^m$
   $\Theta_j^m = \Phi_{k_j}^m$ for $j$ in $1:J$
End For
```

A general latent variable model can be specified by a joint density $f_{XY}(x,y;\theta)$, with X taking values in $\mathbb{X} \subset \mathbb{R}^{\dim(\mathbb{X})}$, Y taking values in $\mathbb{Y} \subset \mathbb{R}^{\dim(\mathbb{Y})}$, and θ taking values in $\Theta \subset \mathbb{R}^{\dim(\Theta)}$. The data consist of a single observation, $y^* \in \mathbb{Y}$. The likelihood function is

$$\ell(\theta) = f_Y(y^*;\theta) = \int f_{XY}(x,y^*;\theta)dx,$$

and we look for a maximum likelihood estimate (MLE), i.e., a value $\hat{\theta}$ maximizing $\ell(\theta)$. The parameter perturbation step of Algorithm IIS2 is a Monte Carlo approximation to a perturbation map H_σ where

$$H_\sigma g(\theta) = \int g(\varphi)h(\theta|\varphi\ ;\sigma)d\varphi. \qquad \textbf{[S3]}$$

A natural choice for $h(\cdot|\varphi;\sigma)$ is the multivariate normal density with mean φ and variance $\sigma^2\Sigma$ for some covariance matrix Σ, but in general, h could be any condition density parameterized by σ. The resampling step of Algorithm IIS2 is a Monte Carlo approximation to a Bayes map, B, given by

$$Bf(\theta) = f(\theta)\ell(\theta)\left\{\int f(\varphi)\ell(\varphi)d\varphi\right\}^{-1}. \qquad \textbf{[S4]}$$

When the SD of the parameter perturbations is held fixed at $\sigma_m = \sigma > 0$, Algorithm IIS2 is a Monte Carlo approximation to $T_\sigma^M f(\theta)$ where

$$T_\sigma f(\theta) = BH_\sigma f(\theta) = \frac{\int f(\varphi)\ell(\theta)h(\theta|\varphi;\sigma)d\varphi}{\iint f(\varphi)\ell(\xi)h(\xi|\varphi;\sigma)d\varphi\, d\xi}. \qquad \textbf{[S5]}$$

Gaussian and Near-Gaussian Analysis of Iterated Importance Sampling

The convergence results of Theorems 1 and 2 in *Convergence of IF2* are not precise about the rate of convergence, either toward the MLE as $\sigma \to 0$ or toward the stationary distribution as $M \to \infty$. Explicit results are available in the Gaussian case and are also relevant to near-Gaussian situations. The near-Gaussian situation may arise in practice, since the parameter perturbations can be constructed to follow a Gaussian distribution and the log likelihood surface may be approximately quadratic due to asymptotic behavior of the likelihood for large sample sizes. The near-Gaussian situation for a POMP model does not require that the POMP itself is near Gaussian, only that the log likelihood surface is near quadratic. Here, we consider only the univariate case, and only for iterated importance sampling. We offer this simplified case as an illustrative example, rather than an alternative justification for the use of our algorithm. In principle, these results can be generalized, but such results do not add much to the general convergence guarantees already obtained.

We investigate the eigenvalues and eigenfunctions for a Gaussian system, and then we appeal to continuity of the eigenvalues to study systems that are close to Gaussian. Here, we consider the case of a scalar parameter, $\dim(\Theta = 1)$, and an additive perturbation given by

$$h(\theta|\varphi;\sigma) = \kappa(\theta - \varphi). \qquad \textbf{[S6]}$$

We first study the unnormalized version of Eq. **S5** defined as

$$Sf(\theta) = [f(\theta)\ell(\theta)] * \kappa(\theta) = \int [f(\theta-\varphi)\ell(\theta-\varphi)]\kappa(\varphi)d\varphi. \qquad \textbf{[S7]}$$

This is a linear map, and we obtain the eigenvalues and eigenfunctions when ℓ and h are Gaussian in Proposition S3. Iterations of the corresponding normalized map, T_σ, converge to the normalized eigenfunction corresponding to the largest eigenvalue of S, which can be seen by postponing normalization until having carried out a large number of iterations of the unnormalized map. Suppose, without loss of generality, that the maximum of the likelihood is at $\theta = 0$. Let $\phi(\theta;\sigma)$ be the normal density with mean zero and variance σ^2.

Proposition S3. *Let S_0 be the map constructed as in Eq. **S7** with the choices $\ell(\theta) = \phi(\theta;\tau)$ and $\kappa(\theta) = \phi(\theta;\sigma)$. Let*

$$u^2 = \left(\sigma^2 + \sqrt{\sigma^4 + 4\sigma^2\tau^2}\right)\Big/2 = \sigma\tau + o(\sigma). \qquad \textbf{[S8]}$$

The eigenvalues of S_0 are

$$\lambda_n = \sigma\tau\sqrt{2\pi}\left(\frac{u^2 - \sigma^2}{u^2}\right)^{(n+1)/2},$$

for $n = 0, 1, 2, \ldots$, and the corresponding eigenfunctions have the form

$$e_n = p_n(\theta)\phi(\theta;u), \qquad \textbf{[S9]}$$

where p_n is a polynomial of degree n.

Proof. Let P_n be the subspace of functions of the form $q(\theta)\phi(\theta;u)$ where q is a polynomial of degree less than or equal to n. We show that S_0 maps P_n into itself, and look at what happens to terms of degree n. Let H_n be the Hermite polynomial of degree n, defined by $(d/d\theta)^n\phi(\theta;1) = (-1)^n H_n(\theta)\phi(\theta;1)$. Let $\alpha = (1/u^2 + 1/\tau^2)^{-1/2}$, and set

$$f(\theta) = \alpha^{-2n}H_n(\theta/\alpha)\phi(\theta;u). \qquad \textbf{[S10]}$$

Then,

$$f(\theta)\ell(\theta) = \frac{\alpha}{\sigma\tau\sqrt{2\pi}}\alpha^{-2n}H_n(\theta/\alpha)\phi(\theta;\alpha) = \frac{\alpha}{\sigma\tau\sqrt{2\pi}}(-1)^n\frac{d^n}{d\theta^n}\phi(\theta;\alpha). \qquad \textbf{[S11]}$$

Since $[(d/d\theta)^n f\ell] * \kappa = (d/d\theta)^n[(f\ell) * \kappa]$, we get

$$(f\ell) * \kappa = \frac{\alpha}{\sigma\tau\sqrt{2\pi}}(-1)^n\frac{d^n}{d\theta^n}\phi(\theta;u) = \frac{\alpha}{\sigma\tau\sqrt{2\pi}}u^{-2n}H_n(\theta/u)\phi(\theta;u). \qquad \textbf{[S12]}$$

Writing $H_n(\theta) = h_0 + h_1\theta + \ldots + h_n\theta^n$, we see that the coefficient of the term in θ^n in Eq. **S10** is $\alpha^{-n}h_n$, whereas in Eq. **S12**, it is $\frac{\alpha}{\sigma\tau\sqrt{2\pi}}u^{-n}$. We have shown that S_0 operating on P_n multiplies the coefficient of degree n by a factor of λ_n. Letting L_n be the matrix representing S_0 on P_n with the basis $b_0, \ldots, b_m$ given by $b_m(\theta) = \theta^m\phi(\theta;u)$, we see that L_n is lower triangular with diagonal entries $\lambda_0, \ldots, \lambda_n$. Therefore, the eigenvalues are $\lambda_0, \ldots, \lambda_n$, and the eigenfunction corresponding to λ_m is in P_m.

The case where $\log\ell(\theta)$ is close to quadratic is relevant due to asymptotic log quadratic properties of the likelihood function. Choosing $\kappa(\theta)$ to be Gaussian, as in Proposition S3, we have the following approximation result.

Proposition S4. *Let S_ϵ be a map as in Eq. **S7**, with ℓ satisfying $\sup_\theta|\ell(\theta) - \phi(\theta;\tau)| < \epsilon$ and $\kappa(\theta) = \phi(\theta;\sigma)$. For ε small, the largest eigenvalue of S_ϵ is close to λ_0 and the corresponding eigenfunction is close to $\phi(\theta;u)$.*

Proof. Write $\ell(\theta) = \phi(\theta;\tau) + \eta(\theta)$, with $\sup_\theta|\eta(\theta)| < \epsilon$. Then,

$$||S_\epsilon f - S_0 f|| = ||(f\eta) * \kappa|| \le ||f\eta|| \le \epsilon||f||. \qquad \textbf{[S13]}$$

Here, $\|\cdot\|$ is the L^2 norm of a function or the corresponding operator norm (largest absolute eigenvalue). Convolution with κ is a contraction in L^2, which is apparent by taking Fourier transforms and making use of Parseval's relationship, since all frequencies are shrunk by multiplying with the Fourier transform of κ. From Eq. **S13**, we have $||S_0 - S_\epsilon|| < \epsilon$. This implies that S_ϵ has a largest eigenvalue μ_0 with $|\mu_0 - \lambda_0| < \epsilon$, based on the representation that

$$|\mu_0| = ||S|| = \sup_f \frac{||S_\epsilon f||}{||f||}. \qquad \textbf{[S14]}$$

Writing the corresponding unit eigenfunction as w_0, we have

$$w_0 = (1/\mu_0)S_\epsilon w_0 = (1/\mu_0)[S_0 w_0 + \eta], \qquad \textbf{[S15]}$$

where $||\eta(\theta)|| < \epsilon$. Writing $w_0 = \sum_{i=1}^{\infty}\alpha_i e_i$, in terms of $\{e_i\}$ from Eq. **S9**, Eq. **S15** gives

$$\sum_{i=1}^{\infty}\alpha_i e_i = \sum_{i=1}^{\infty}\alpha_i \frac{\lambda_i}{\mu_0} e_i + \eta = \sum_{i=1}^{\infty}\alpha_i \frac{\lambda_i}{\lambda_0} e_i + \tilde{\eta}, \qquad \textbf{[S16]}$$

where $||\tilde{\eta}|| < \epsilon \ (1 + [\lambda_0(\lambda_0 - \epsilon)]^{-1})$. Comparing terms in e_i, we see that all terms $\alpha_1, \alpha_2, \ldots$ must be of order ε.

A Class of Exact Non-Gaussian Limits for Iterated Importance Sampling

We look for exact solutions to the equation $Tf = f$ where $T = BH$, as specified in Eq. **S5** with $h(\theta|\varphi;\sigma) = \kappa(\theta - \varphi)$. This situation corresponds to iterated importance sampling with additive parameter perturbations that have no dependence on σ, as in Eq. **S6**. Now, for $g(x)$ being a probability density on Θ, define

$$\ell_g(x) = c\,\frac{g(x)}{\kappa * g(x)}, \qquad \textbf{[S17]}$$

where c is a nonnegative constant. For likelihood functions of the form of Eq. **S17**, supposing that ℓ_g is integrable, we obtain an eigenfunction $e(x) = \kappa * g(x)$ for the unnormalized map S defined in Eq. **S7** via the following calculation:

$$\begin{aligned} Se(x) &= c\int \frac{g(x-u)}{(g*\kappa)(x-u)}(g*\kappa)(x-u)\kappa(u)du \\ &= c\int g(x-u)\kappa(u)du \\ &= c[g*\kappa(x)] = c\ e(x). \end{aligned}$$

Under conditions such as Theorem 1, it follows that $\kappa * g$ is the unique eigenfunction for T, up to a scale factor, and that $\lim_{M\to\infty} T^M f = e$. We do not anticipate practical applications for the conjugacy relationship we have established between the pair (ℓ_g, κ) since we see no reason why the likelihood should have the form of Eq. **S17**. However, this situation does serve to identify a range of possible limiting behaviors for T^M.

Applying PMCMC to the Cholera Model

We carried out PMCMC for the cholera model, with the prior being uniform on the hyperrectangle specified by θ_{low} and θ_{high} in Table S1. Thus, the IF1 and IF2 searches were conducted starting with random draws from this prior. Since PMCMC is known to be computationally demanding, we investigated a simplified challenge: investigating the posterior distribution starting at the MLE. This would be appropriate, for example, if one aimed to obtain Bayesian inferences using PMCMC but giving it a helping hand by first finding a good starting value obtained by a maximization procedure. We used the PMMH implementation of PMCMC in pomp (4) with parameter proposals following a Gaussian random walk with SDs given by $(\theta_{\text{high}} - \theta_{\text{low}})/100$. We started 100 independent chains at the estimated MLE in Table S1. Each PMCMC chain, with $J = 1{,}500$ particles at each of $M = 2\times10^4$ likelihood evaluations, took around 30 h to run on a single core of the University of Michigan Flux cluster. Writing $V_{m,d}$ for the sample variance of variable $d \in \{1, \ldots, \dim(\Theta)\}$ among the 100 chains at time $m \in \{1\ldots, M\}$, and τ_d for the Gaussian random walk SD for parameter d, we tracked the quantity

$$V_m = \sum_{d=1}^{\dim(\Theta)} \frac{V_{m,d}}{\tau_d^2}. \qquad \textbf{[S18]}$$

Supposing the posterior variance is finite, a necessary requirement for convergence to stationarity as m increased is for V_m to approach its asymptotic limit. Since all of the chains start at the same place, one expects V_m to increase toward this limit. The number of iterations required for V_m to stabilize therefore provides a lower bound on the time taken for convergence of the chain. This test assesses the capability of the chain to explore the region of parameter space with high posterior probability density, rather than the capability to search for this region from a remote starting point. We also tested PMCMC on a harder challenge, investigating convergence of the MCMC chain to its stationary distribution from overdispersed starting values. We repeated the computation described above, with 100 chains initialized at draws from the prior distribution. The results are shown in Fig. S1. From Fig. S1*A*, we see that the stationary distribution has not yet been approached for the chains starting at the MLE, since the variance of independent chains continues to increase up to $M = 2\times10^4$. As a harder test, the variance for the initially overdispersed independent chains should approach that for the initially underdispersed chains, but we see in Fig. S1*B* that much more computation would be required to achieve this with the algorithmic settings used.

The PMCMC chains used here, for the cholera data with $N = 6\times10^2$ data points, involved $JMN = (1.5\times10^3)\times(2\times10^4)\times(6\times10^2) = 1.8\times10^{10}$ calls to the dynamic process simulator (the dominating computational expense), and yet failed to converge. By contrast, IF2 with $JMN = (10^4)\times10^2\times(6\times10^2) = \ 6\times10^8$ calls to the dynamic process simulator was shown to be an effective tool for global investigation of the likelihood surface. As with all numerical comparisons, it is hard to assess whether poor performance is a consequence of poor algorithmic choices. Conceptually, a major difference between iterated filtering and PMCMC is that the filtering particles in IF2 investigate the parameter space and latent dynamic variable space simultaneously, whereas, in PMCMC, each filtering iteration is used only to provide a single noisy likelihood evaluation. It may not be surprising that algorithms such as PMCMC struggle in situations where filtering is a substantial computational expense and the likelihood surface is sufficiently complex that many thousands of Monte Carlo steps are required to explore it. Indeed, IF1 and IF2 remain the only algorithms that have currently been demonstrated computationally capable of efficient likelihood-based inference for situations of comparable difficulty to our example.

Applying Liu and West's Method to the Toy Example

Bayesian parameter estimation for POMP models using sequential Monte Carlo with perturbed parameters was proposed by ref. 5. Similar approaches using alternative nonlinear filters have also been widely used (6, 7). Liu and West (8) proposed a development on the approach of ref. 5 that combines parameter perturbations with a contraction that is designed to counterbalance the variation added by the perturbations, thereby approximating the posterior distribution of the parameters for the fixed parameter model of interest. Liu and West (8) also included an auxiliary particle filter procedure in their algorithm (9). The auxiliary particle filter is a version of sequential Monte Carlo that looks ahead to a future observation when deciding which particles to propagate.

Generally, auxiliary particle filter algorithms do not have the plug-and-play property (10, 11) since they involve constructing weights that require evaluation of the transition density for the latent process. In addition, the auxiliary particle filter does not necessarily have superior performance over a basic sequential Monte Carlo filter (12). To compare with IF2 and PMCMC on our toy example, we therefore use a version of the Liu and West algorithm, which we call LW, that omits the auxiliary particle filter procedure. LW carries out the key innovation of parameter perturbation and contraction (Steps 3 and 4 in section 10.4 of ref. 8) while omitting the auxiliary particle filter (Steps 1 and 2, and the denominator in Step 5, in section 10.4 of ref. 8). LW was implemented via the bsmc2 function of the pomp package (4). If an effective auxiliary particle filter were available for a specific computation, it could also be used to enhance other sequential Monte Carlo based inference procedures such as IF1, IF2, and PMCMC.

For the numerical results reported in Fig. S2, we used $J=10^4$ particles for LW. This awards the same computational resources to LW that we gave IF1 and IF2 for the results in Fig. 1. The magnitude of the perturbations in LW is controlled by a discount factor (δ in the notation of ref. 8), and we considered three values, $\delta \in \{0.99, 0.999, 0.9999\}$. Liu and West (8) suggested that δ should take values in the range $\delta \in [0.95, 0.99]$, with smaller values of δ reducing Monte Carlo variability while increasing bias in the approximation to the target posterior distribution. For our toy example, we see from Fig. S2*A* that the choice $\delta = 0.99$ results in a stable Monte Carlo computation (since all eight realizations are close). However, Fig. S2*A* also reveals a large amount of bias. Increasing δ to 0.999, Fig. S2*B* shows some increase in the Monte Carlo variability and some decrease in the bias. Further increasing δ to 0.9999, Fig. S2*C* shows the bias becomes small while the Monte Carlo variability continues to increase. Values of δ very close to 1 are numerically tractable for this toy model, but not in most applications. As δ approaches 1, the ensuing numerical instability exemplifies the principal reason why Bayesian and likelihood-based inference for POMP models is challenging despite the development of modern nonlinear filtering techniques.

The justification provided by ref. 8 for their algorithm is based on a Gaussian approximation to the posterior distribution. Specifically, ref. 8 argued that the posterior distribution should be approximately unchanged by carrying out a linear contraction toward its mean followed by adding an appropriate perturbation. Therefore, it may be unsurprising that LW performs poorly in the presence of nonlinear ridges in the likelihood surface. Other authors have reported poor numerical performance for the algorithm of ref. 8, e.g., figure 2 of ref. 13 and figure 2 of ref. 14. Our results are consistent with these findings, and we conclude that the approach of ref. 8 should be used with considerable caution when the posterior distribution is not close to Gaussian.

Consequences of Perturbing Parameters for the Numerical Stability of SMC

The IF2 algorithm applies sequential Monte Carlo (SMC) to an extended POMP model in which the time-varying parameters are treated as dynamic state variables. This procedure increases the dimension of the state space by the number of time-varying parameters. Empirically, SMC has been found effective in many low-dimensional systems, but its numerical performance can degrade in larger systems. A natural concern, therefore, is the extent to which the extension of the state variable in IF2 increases the numerical challenge of carrying out SMC effectively. Two rival heuristics suggest different answers. One intuitive (but not universally correct) argument is that adding variability to the system stabilizes numerically unstable filtering problems, since it gives each particle at least a slim chance of following a trajectory compatible with the data. An opposing intuition, that SMC breaks down rapidly as the dimension increases, has theoretical support (15). However, the theoretical arguments of ref. 15 may be driven more by increasing the observation dimension than increasing the state dimension, so their relevance in the present situation is not entirely clear.

We investigated numerical stability of SMC, in the context of our cholera example, by measuring the effective sample size (ESS) (16). We investigated the ESS for two parameter vectors, the MLE and an alternative value for which SMC is more numerically challenging. We carried out particle filtering with and without random walk perturbations to the parameters, obtaining the results presented in Fig. S3. We found that the random walk perturbations led to a 5% decrease in the average ESS at the MLE, but a 13% increase in the average ESS at the alternative parameter vector. This example demonstrates that the random walk perturbations can have both a cost and a benefit for numerical stability, with the benefit outweighing the cost as the filtering problem becomes more challenging.

Checking Conditions B1 and B2

We check B1 and B2 when Θ is a rectangular region in $\mathbb{R}^{\dim(\Theta)}$, with $h_n(\theta|\phi\,;\sigma)$ describing a Gaussian random walk having as a limit a reflected Brownian motion on Θ. A more general study of the limit of reflected random walks to reflected Brownian motions (in particular, including limits where the random walk step distribution satisfies B5) was presented by Bossy et al. (17). The specific examples of the IF2 algorithm given in our paper all use Gaussian random walk perturbations for the parameters. The examples did not use boundary conditions to constrain the parameter to a bounded set. While such conditions could be used to ensure practical stability of the algorithm, we view the conditions primarily as a theoretical device to assist the mathematical analysis of the algorithm.

Suppose that $\Theta = [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_{\dim(\Theta)}, b_{\dim(\Theta)}]$. For each coordinate direction $d = 1, \ldots, \dim(\Theta)$, let $R_d : \mathbb{R} \to [a_d, b_d]$ be the reflection map defined recursively by

$$R_d(x) = \begin{cases} x & \text{if} \quad x \in [a_d, b_d] \\ R_d(2b_d - x) & \text{if} \quad x > b_d \\ R_d(2a_d - x) & \text{if} \quad x < a_d \end{cases}.$$

Let $h_{n,d}(\theta_d|\phi_d\,;\sigma)$ be the density of $R_d(\phi_d + \sigma Z)$ where Z is a standard Normal random variable. Let $h_n(\theta|\phi\,;\sigma)$ be the joint density corresponding to the product of $h_{n,1}, \ldots, h_{n,\dim(\Theta)}$. This choice of h_n corresponds to a perturbation process for the parameter vector in the IF2 algorithm following a Gaussian random walk on Θ with reflective boundary conditions, independently in each coordinate direction. By construction, the finite dimensional distributions of $W_\sigma(t)$ at the set of times

$$\{k\sigma^2 : k = 0, 1, 2, \ldots \text{and } \ k\sigma^2 \le 1\}$$

exactly match the corresponding finite dimensional distributions of a reflected Brownian motion $\{W(t)\}$ taking values in Θ. This $\{W(t)\}$ gives a construction of the limiting process whose existence is assumed in B1. For $A \subset \Theta$, we see from this construction of $\{W(t)\}$ that the probability $\{W(t)\}$ is in A for all $\epsilon \le t \le 1]$ is greater than the corresponding probability for an unreflected Brownian motion, $\{W_{(u)}(t)\}$ with the same intensity parameter. It is routine to check that $\{W_{(u)}(t)\}$ has a positive probability of remaining in any open set A for all $\epsilon \le t \le 1$ uniformly over all values of $W_{(u)}(0) \in \Theta$. Thus, we have completed the check of condition B1.

To check B2, the positivity of the marginal density of $W(t)$ on Θ, uniformly over the value of $W(0)$, again follows since this density is larger than the known density for $W_{(u)}(t)$.

Additional Details for the Proof of Theorem 1

In *Convergence of IF2*, a condensed proof of Theorem 1 is provided to describe the key steps in the argument. Here, we restate Theorem 1 and provide a more detailed proof. The reader is referred back to the main text for the notation and statement of conditions B2 and B4. Let $L^1\Theta$ denote the space of integrable real-valued functions on Θ with norm $||f||_1 = \int|f(\theta)|d\theta$. For nonnegative measures μ and ν on Θ, let $||\mu-\nu||_{\mathbf{tv}}$ denote the total variation distance and let $H(\mu,\nu)$ denote the Hilbert metric distance (18, 19). The measures μ and ν are said to be comparable if they are both nonzero and there exist constants $0 < a \le b$ such that $a\ \nu(A) \le \mu(A) \le b\ \nu(A)$ for all measurable subsets $A \subset \Theta$. For comparable measures, $H(\mu,\nu)$ is defined by

$$H(\mu,\nu) = \log\frac{\sup_A\ \mu(A)/\nu(A)}{\inf_A\ \mu(A)/\nu(A)}, \qquad \textbf{[S19]}$$

with the supremum and infimum taken over measurable subsets $A \subset \Theta$ having $\nu(A) > 0$. For noncomparable measures, the Hilbert metric is defined by $H(0,0) = 0$, and otherwise $H(\mu,\nu) = \infty$. The Hilbert metric is invariant to multiplication by a positive scalar, $H(a\mu,\nu) = H(\mu,\nu)$. This projective property makes the Hilbert metric convenient to investigate the Bayes map: In the context of the following proof, the projective property lets us analyze the linear map S_σ to study the nonlinear map T_σ.

Theorem 1. *Let T_σ be the map defined by Eq.* ***1*** *in the main text, and suppose B2 and B4. There exists a unique probability density f_σ such that for any probability density f on Θ,*

$$\lim_{m\to\infty} ||T_\sigma^m f - f_\sigma||_1 = 0, \qquad \textbf{[S20]}$$

where $||f||_1$ is the L^1 norm of f. Let $\{\Theta_j^M,\ j = 1,\ldots,J\}$ be the output of IF2, with $\sigma_m = \sigma > 0$. There exists a finite constant C such that

$$\lim_{M\to\infty}\sup\mathbb{E}\left[\left|\frac{1}{J}\sum_{j=1}^{J}\phi\left(\Theta_j^M\right) - \int\phi(\theta)f_\sigma(\theta)d\theta\right|\right] \le \frac{C\sup_\theta|\phi(\theta)|}{\sqrt{J}}. \qquad \textbf{[S21]}$$

Proof. For $\theta_{0:N} \in \Theta^{N+1}$, we single out the last component of $\theta_{0:N}$ by writing $\breve{\ell}(\theta_{0:N}) = \breve{\ell}(\theta_{0:N-1},\theta_N)$ and $h(\theta_{0:N}|\phi) = h(\theta_{0:N-1},\theta_N|\phi)$. Then, for ϕ and θ in Θ, we define

$$s_\sigma(\phi,\theta) = \int h(\theta_{0:N-1},\theta|\phi,\sigma)\breve{\ell}(\theta_{0:N-1},\theta)d\theta_{0:N-1}. \qquad \textbf{[S22]}$$

The function s_σ in Eq. **S22** defines a linear operator $S_\sigma f(\theta) = \int s_\sigma(\phi,\theta)f(\phi)\mathrm{d}\phi$ that maps $L^1(\Theta)$ into itself. Notice that $T_\sigma f(\theta) = S_\sigma f(\theta)/||S_\sigma f||_1$. More generally, if μ is a probability measure on Θ, $S_\sigma\mu$ denotes the function $S_\sigma\mu(\theta) = \int s_\sigma(\phi,\theta)\mu(d\phi)$. Notice also that $S_\sigma^m f$, the m-th iterate of S_σ, can be written as $S_\sigma^m f(\theta) = \int s_\sigma^{(m)}(\phi,\theta)f(\phi)d\phi$, where $s_\sigma^{(1)}(\phi,\theta) = s_\sigma(\phi,\theta)$, and for $m \ge 2$, $s_\sigma^{(m)}(\phi,\theta) = \int s_\sigma(\phi,u)s_\sigma^{(m-1)}(u,\theta)du$. Using the definition of $\breve{\ell}$ and B4,

$$s_\sigma(\phi,\theta) = \int h(\theta_{0:N-1},\theta|\phi,\sigma)\int f_X(x_{0:N}|\theta_{0:N-1},\theta)f_{Y|X}(y_{1:N}^*|x_{0:N})\, dx_{0:N}\, d\theta_{0:N-1} \ge \epsilon^N\int h(\theta_{0:N-1},\theta|\phi,\sigma)d\theta_{0:N-1}, \qquad \textbf{[S23]}$$

and, similarly,

$$s_\sigma(\phi,\theta) \le \epsilon^{-N}\int h(\theta_{0:N-1},\theta|\phi,\sigma)d\theta_{0:N-1}. \qquad \textbf{[S24]}$$

By iterating the Inequalities **S23** and **S24**, assumption B2 implies that there exists $m_0 \ge 1$ such that for any $m \ge m_0$, there exist $0 < \delta_m < \infty$, a probability measure λ_m on Θ such that for all measurable subsets $A \subset \Theta$ and all $\theta \in \Theta$,

$$\delta_m\lambda_m(A) \le \int_A s^{(m)}(\theta,\phi)d\phi \le \delta_m^{-1}\lambda_m(A). \qquad \textbf{[S25]}$$

In other words, $S_\sigma^{m_0}$ is mixing in the sense of ref. 19. In the terminology of ref. 18, this means that for each $m \ge m_0$, S^m has finite projective diameter (see lemma 2.6.2 of ref. 18). Therefore, by theorem 2.5.1 of ref. 18, we conclude that S_σ has a unique nonnegative eigenfunction f_σ with $||f_\sigma||_1 = 1$, and for any density f on Θ, as $q \to \infty$,

$$\left|\left|\frac{\left[S_\sigma^{m_0}\right]^q f}{||\left[S_\sigma^{m_0}\right]^q f||_1} - f_\sigma\right|\right|_1 = ||T_\sigma^{m_0 q}f - f_\sigma||_1 \to 0.$$

This implies the Statement **S20**, by writing for any $m \ge 1$, $m = qm_0 + r$, for $0 \le r \le m_0 - 1$, and $T_\sigma^m f = [T_\sigma^{qm_0}]T_\sigma^r f$.

Let the initial particle swarm $\{\Theta_j^0,\ 1 \le j \le J\}$ consist of independent draws from the density f. To prove Eq. **S21**, we decompose $M = qm_0 + r$, for some $r \in \{0,\ldots,m_0 - 1\}$, and we introduce the empirical measures $\mu^{(0)} = J^{-1}\sum_{j=1}^J\delta_{\Theta_j^{(r)}}$, and for $k = 1,\ldots,q$, $\mu^{(k)} = J^{-1}\sum_{j=1}^J\delta_{\Theta_j^{(r+m_0k)}}$, so that $\mu^{(q)} = J^{-1}\sum_{j=1}^J\delta_{\Theta_j^{(M)}}$. We then write, for any bounded measurable function ϕ,

$$\begin{aligned}\mu^{(q)}(\phi) - \left[T_\sigma^M f\right](\phi) &= \mu^{(q)}(\phi) - \left[T_\sigma^{m_0q}\mu^{(0)}\right](\phi) \\ &\quad + \left[T_\sigma^{m_0q}\mu^{(0)}\right](\phi) - \left[T_\sigma^{m_0q}T_\sigma^r f\right](\phi) \\ &= \sum_{i=1}^{q}\left\{\left[T_\sigma^{m_0(i-1)}\mu^{(q-i+1)}\right](\phi) - \left[T_\sigma^{m_0i}\mu^{(q-i)}\right](\phi)\right\} \\ &\quad + \left[T_\sigma^{m_0q}\mu^{(0)}\right](\phi) - \left[T_\sigma^{m_0q}T_\sigma^r f\right](\phi).\end{aligned}$$

Using theorem 2 of ref. 20, we can find a finite constant C_3 such that for all $k \ge 1$, and writing $||\phi||_\infty = \sup_\theta|\phi(\theta)|$,

$$\rho = \sup_{\phi:||\phi||_\infty = 1}\mathbb{E}\left[\left|\mu^{(k)}(\phi) - \left[T_\sigma^{m_0}\mu^{(k-1)}\right](\phi)\right|\right] \le \frac{C_3}{\sqrt{J}}, \qquad \textbf{[S26]}$$

with B4 implying that the constant C_3 constructed by ref. 20 does not depend on $\mu^{(k-1)}$. Since $S_\sigma^{m_0}$ is mixing and Eq. **S25** holds, using lemma 3.4, lemma 3.5, lemma 3.8, and equation 7 of ref. 19, we have

$$\begin{aligned}&\mathbb{E}\left[\left|\left[T_\sigma^{m_0q}\mu^{(0)}\right](\phi) - \left[T_\sigma^{m_0q}T_\sigma^r f\right](\phi)\right|\right] \\ &\le ||\phi||_\infty\mathbb{E}\left[||T_\sigma^{m_0q}\mu^{(0)} - T_\sigma^{m_0q}T_\sigma^r f||_{\mathbf{tv}}\right] \\ &\le \frac{2||\phi||_\infty}{\log 3}\mathbb{E}\left[H\left(S_\sigma^{m_0q}\mu^{(0)}, S_\sigma^{m_0q}T_\sigma^r f\right)\right] \\ &\le \frac{2||\phi||_\infty}{\log 3}\left(\frac{1-\delta_{m_0}^2}{1+\delta_{m_0}^2}\right)^{q-2}\frac{1}{\delta_{m_0}^2}\mathbb{E}\left[||T_\sigma^{m_0}\mu^{(0)} - T_\sigma^{m_0}T_\sigma^r f||_{\mathbf{tv}}\right] \\ &\le \frac{4||\phi||_\infty}{\log 3}\left(\frac{1-\delta_{m_0}^2}{1+\delta_{m_0}^2}\right)^{q-2}\frac{1}{\delta_{m_0}^2}\frac{\rho}{\delta_{m_0}^2}.\end{aligned}$$

For $i=3,\dots,q$, a similar calculation gives

$$\mathbb{E}\left[\left|T_\sigma^{m_0(i-1)}\mu^{(q-i+1)}(\phi)-T_\sigma^{m_0 i}\mu^{(q-i)}(\phi)\right|\right]=\mathbb{E}\left[\left|T_\sigma^{m_0(i-1)}\mu^{(q-i+1)}(\phi)\right.\right.$$
$$\left.\left.-T_\sigma^{m_0(i-1)}T_\sigma^{m_0}\mu^{(q-i)}(\phi)\right|\right]\le\frac{4||\phi||_\infty}{\log 3}\left(\frac{1-\delta_{m_0}^2}{1+\delta_{m_0}^2}\right)^{i-3}\frac{1}{\delta_{m_0}^2}\frac{\rho}{\delta_{m_0}^2}.$$

The case $i=1$ boils down to Eq. **S26**, where the case $i=2$ gives, by similar calculations:

$$\mathbb{E}\left[\left|T_\sigma^{m_0}\mu^{(q-1)}(\phi)-T_\sigma^{2m_0}\mu^{(q-2)}(\phi)\right|\right]\le 2||\phi||_\infty\frac{\rho}{\delta_{m_0}^2}.$$

Hence, using Eq. **S26**,

$$\mathbb{E}\left[\left|\mu^{(q)}(\phi)-\left[T_\sigma^M f\right](\phi)\right|\right]$$
$$\le\frac{C_3||\phi||_\infty}{\sqrt{J}}\left(1+\frac{2}{\delta_{m_0}^2}+\frac{4}{\log 3}\left(\frac{1}{\delta_{m_0}^2}\right)^2\sum_{j=0}^{q-2}\left(\frac{1-\delta_{m_0}^2}{1+\delta_{m_0}^2}\right)^j\right).$$

We conclude that there exists a finite constant C_4 such that

$$\mathbb{E}\left[\left|\frac{1}{J}\sum_{j=1}^{J}\phi\left(\Theta_j^M\right)-\int\phi(\theta)\left[T_\sigma^M f\right](\theta)d\theta\right|\right]\le\frac{C_4||\phi||_\infty}{\sqrt{J}}.\qquad\textbf{[S27]}$$

Eq. **S21** follows by combining Eq. **S27** with Eq. **S20**.

1. Jacod J, Shiryaev AN (1987) *Limit Theorems for Stochastic Processes* (Springer, Berlin).
2. Ionides EL, Bhadra A, Atchadé Y, King AA (2011) Iterated filtering. *Ann Stat* 39: 1776–1802.
3. Doucet A, Jacob PE, Rubenthaler S (2013) Derivative-free estimation of the score vector and observed information matrix with application to state-space models. *arxiv.1304.5768*.
4. King AA, Ionides EL, Bretó CM, Ellner S, Kendall B (2009) pomp: Statistical inference for partially observed Markov processes (R package). Available at cran.r-project.org/web/packages/pomp.
5. Kitagawa G (1998) A self-organising state-space model. *J Am Stat Assoc* 93:1203–1215.
6. Anderson BD, Moore JB (1979) *Optimal Filtering* (Prentice-Hall, Piscataway, NJ).
7. Wan E, van der Merwe R (2000) The unscented Kalman filter for nonlinear estimation. *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000.* (IEEE, Piscataway, NJ), pp 153–158.
8. Liu J, West M (2001) Combined parameter and state estimation in simulation-based filtering. *Sequential Monte Carlo Methods in Practice*, eds Doucet A, de Freitas N, Gordon NJ (Springer, New York), pp 197–224.
9. Pitt MK, Shepard N (1999) Filtering via simulation: Auxillary particle filters. *J Am Stat Assoc* 94:590–599.
10. Bretó C, He D, Ionides EL, King AA (2009) Time series analysis via mechanistic models. *Ann Appl Stat* 3:319–348.
11. He D, Ionides EL, King AA (2010) Plug-and-play inference for disease dynamics: Measles in large and small populations as a case study. *J R Soc Interface* 7(43):271–283.
12. Johansen AM, Doucet A (2008) A note on the auxiliary particle filter. *Stat Probab Lett* 78:1498–1504.
13. Storvik G (2002) Particle filters for state-space models with the presence of unknown static parameters. *IEEE Trans Signal Process* 50:281–289.
14. Chopin N, Jacob PE, Papaspiliopoulos O (2013) SMC^2: An efficient algorithm for sequential analysis of state space models. *J R Stat Soc Ser B* 75:397–426.
15. Bengtsson T, Bickel P, Li B (2008) Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. *Probability and Statistics: Essays in Honor of David A. Freedman*, eds Speed T, Nolan D (Inst Math Stat, Beachwood, OH), pp 316–334.
16. Liu JS (2001) *Monte Carlo Strategies in Scientific Computing* (Springer, New York).
17. Bossy M, Gobet E, Talay D (2004) A symmetrized Euler scheme for an efficient approximation of reflected diffusions. *J Appl Probab* 41:877–889.
18. Eveson SP (1995) Hilbert's projective metric and the spectral properties of positive linear operators. *Proc London Math Soc* 3:411–440.
19. Le Gland F, Oudjane N (2004) Stability and uniform approximation of nonlinear filters using the Hilbert metric and application to particle filters. *Ann Appl Probab* 14:144–187.
20. Crisan D, Doucet A (2002) A survey of convergence results on particle filtering methods for practitioners. *IEEE Trans Signal Process* 50:736–746.

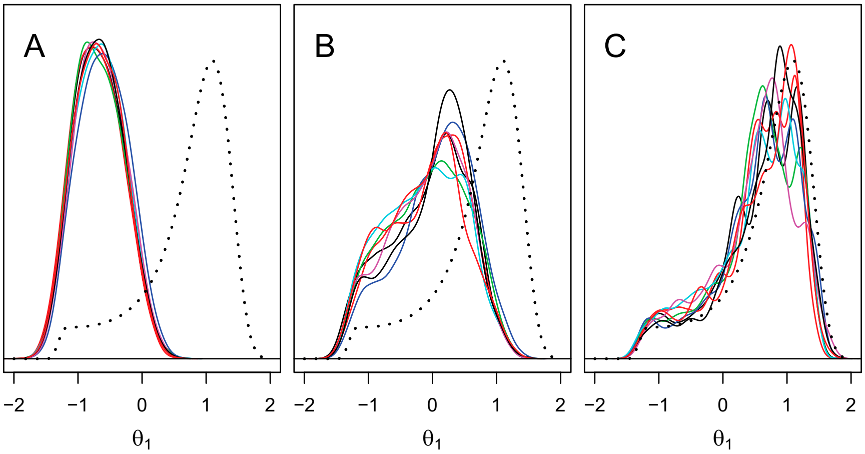

Fig. S1. The Liu and West algorithm (8) applied to the toy example with varying values of the discount factor: (*A*) $\delta=0.99$; (*B*) $\delta=0.999$; (*C*) $\delta=0.9999$. Solid lines show eight independent estimates of the marginal posterior density of θ_1. The black dotted line shows the true posterior density.

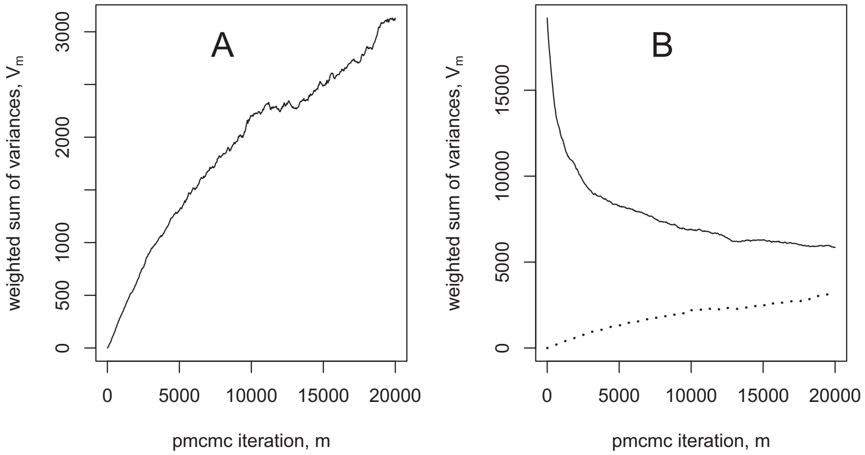

Fig. S2. PMCMC convergence assessment, using the diagnostic quantity in Eq. **S18**. (*A*) Underdispersed chains, all started at the MLE. (*B*) Overdispersed chains, started with draws from the prior (solid line), and underdispersed chains (dotted line). The average acceptance probability was 0.04238, with Monte Carlo SE 0.00072, calculated from iterations 5,000 through 20,000 for the 100 underdispersed PMCMC chains. For the overdispersed chains, the average acceptance probability was 0.04243 with SE 0.00100.

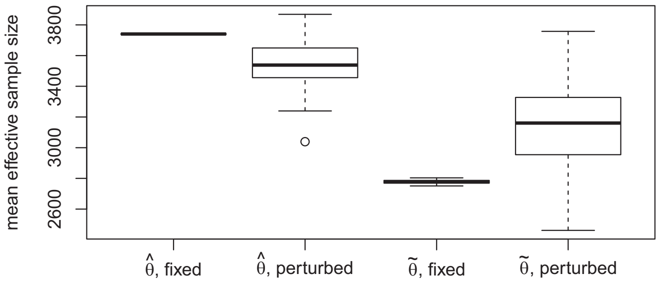

Fig. S3. Effective sample size (ESS) for SMC with fixed parameters and with perturbed parameters. We ran SMC for the cholera model with the parameter vector set at the MLE, $\hat{\theta}$, and at an alternative parameter vector $\tilde{\theta}$ for which the first 18 parameters in Table S1 were multiplied by a factor of 0.8. We defined the ESS at each time point by the reciprocal of the sum of squares of the normalized weights of the particles. The mean ESS was calculated as the average of these ESS values over the 600 time points. Repeating this computation 100 times, using $J = 10^4$ particles, gave 100 mean ESS values shown in the "fixed" columns of the box-and-whisker plot. Repeating the computation with additional parameter perturbations having random walk SD of 0.01 gave the 100 mean ESS values shown in the "perturbed" column. For both parameter vectors, the perturbations greatly increase the spread of the mean ESS. At $\hat{\theta}$, the perturbations decreased the mean ESS value by 5% on average, whereas at $\tilde{\theta}$ the perturbations increased the mean ESS value by 13% on average. The MLE may be expected to be a favorable parameter value for stable filtering, and our interpretation is that the parameter perturbations have some chance of moving the SMC particles away from this favorable region. When started away from the MLE, the numerical stability of the IF2 algorithm benefits from the converse effect that the parameter perturbations will move the SMC particles preferentially toward this favorable region. For parameter values even further from the MLE than $\tilde{\theta}$, SMC may fail numerically for a fixed parameter value yet be feasible with perturbed parameters.

Table S1. Parameters for the cholera model

	$\hat{\theta}$	θ_{low}	θ_{high}
γ	20.80	10.00	40.00
ε	19.10	0.20	30.00
m	0.06	0.03	0.60
$\beta_{\text{trend}} \times 10^2$	−0.50	−1.00	0.00
β_1	0.75	−4.00	4.00
β_2	6.38	0.00	8.00
β_3	−3.44	−4.00	4.00
β_4	4.23	0.00	8.00
β_5	3.33	0.00	8.00
β_6	4.55	0.00	8.00
ω_1	−1.69	−10.00	0.00
ω_2	−2.54	−10.00	0.00
ω_3	−2.84	−10.00	0.00
ω_4	−4.69	−10.00	0.00
ω_5	−8.48	−10.00	0.00
ω_6	−4.39	−10.00	0.00
σ	3.13	1.00	5.00
τ	0.23	0.10	0.50
S_0	0.62	0.00	1.00
I_0	0.38	0.00	1.00
$R_{1,0}$	0.00	0.00	1.00
$R_{2,0}$	0.00	0.00	1.00
$R_{3,0}$	0.00	0.00	1.00

$\hat{\theta}$ is the MLE reported by ref. 1. Three parameters were fixed ($\delta = 0.02$, $N_s = 6$, and $k = 3$) following ref. 1. Units are per year for γ, ε, m, β_{trend}, and δ; all other parameters are dimensionless. The θ_{low} and θ_{high} are the lower and upper bounds for a hyperrectangle used to generate starting points for the search. Nonnegative parameters (γ, ε, m, σ, τ) were logarithmically transformed for optimization. Unit scale parameters (S_0, I_0, $R_{1,0}$, $R_{2,0}$, $R_{3,0}$) were optimized on a logistic scale. These parameters were rescaled using the known population size to give the initial state variables, e.g., $S(t_0) = S_0\{S_0 + I_0 + R_{1,0} + R_{2,0} + R_{3,0}\}^{-1}P(t_0)$.

1. King AA, Ionides EL, Pascual M, Bouma MJ (2008) Inapparent infections and cholera dynamics. *Nature* 454(7206):877–880.

Other Supporting Information Files

Dataset S1 (TXT)
Dataset S2 (TXT)