Statistical Science
2018, Vol. 33, No. 1, 4–18
https://doi.org/10.1214/17-STS618

Approximate Bayesian Computation and Simulation-Based Inference for Complex Stochastic Epidemic Models

Trevelyan J. McKinley, Ian Vernon, Ioannis Andrianakis, Nicky McCreesh, Jeremy E. Oakley, Rebecca N. Nsubuga, Michael Goldstein and Richard G. White

Abstract. Approximate Bayesian Computation (ABC) and other simulation-based inference methods are becoming increasingly used for inference in complex systems, due to their relative ease-of-implementation. We briefly review some of the more popular variants of ABC and their application in epidemiology, before using a real-world model of HIV transmission to illustrate some of challenges when applying ABC methods to high-dimensional, computationally intensive models. We then discuss an alternative approach—history matching—that aims to address some of these issues, and conclude with a comparison between these different methodologies.

Key words and phrases: Approximate Bayesian Computation, history matching, emulation, Bayesian inference, infectious disease models.

Trevelyan J. McKinley is Lecturer in Mathematical Biology, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Penryn, United Kingdom (e-mail: t.mckinley@exeter.ac.uk). Ian Vernon is Associate Professor in Statistics, Department of Mathematical Sciences, Durham University, Durham, United Kingdom (e-mail: i.r.vernon@durham.ac.uk). Ioannis Andrianakis is Research Fellow, London School of Hygiene and Tropical Medicine, Keppel Street, London, United Kingdom (e-mail: ioannis.andrianakis@lshtm.ac.uk). Nicky McCreesh is Assistant Professor in Infectious Disease Modelling, London School of Hygiene and Tropical Medicine, Keppel Street, London, United Kingdom (e-mail: nicky.mccreesh@lshtm.ac.uk). Jeremy E. Oakley is Professor of Statistics, School of Mathematics and Statistics, University of Sheffield, Sheffield, United Kingdom (e-mail: j.oakley@sheffield.ac.uk). Rebecca N. Nsubuga is Senior Scientist (Statistician/Modeller), MRC/UVRI Research Unit on AIDS, Entebbe, Uganda (e-mail: Rebecca.Nsubuga@mrcuganda.org). Michael Goldstein is Professor of Statistics, Department of Mathematical Sciences, Durham University, Durham, United Kingdom (e-mail: michael.goldstein@durham.ac.uk). Richard G. White is Professor of Infectious Disease Modelling, London School of Hygiene and Tropical Medicine, Keppel Street, London, United Kingdom (e-mail: richard.white@lshtm.ac.uk).

1. INTRODUCTION

Complex mathematical models can provide important insights into the behaviour of dynamic epidemiological systems. However, to understand how well the model represents reality, and therefore how useful the model is for inference, regarding the actual system under study, it is necessary to fit it to observed data. This task can be challenging, partly due to the complexity of the model itself, but also because there is often a paucity of available data.

Common features of models used to study real-world epidemiological processes are that they are large-scale, dynamic, nonlinear and auto-correlated. Furthermore, information such as infection times are almost impossible to measure or record, and so the observed data often correspond to proxies such as medical reports, test results and mortality rates, and even these are frequently incomplete. These challenges have driven the development of a suite of statistical methodologies for model fitting, the most widespread of which are based around the use of a likelihood function. Here we focus on Bayesian methods, where we wish to estimate the posterior distribution for the parameters ($\boldsymbol{\theta}$), given the data ($\boldsymbol{y}$), which can be written as

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \pi(\boldsymbol{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}), \tag{1}$$

where $\pi(\boldsymbol{y} \mid \boldsymbol{\theta})$ is the likelihood function and $\pi(\boldsymbol{\theta})$ is the prior distribution (representing our beliefs in the values of the parameters in the absence of data). Usually the normalising constant is analytically intractable, requiring the use of numerical methods to generate empirical estimates of $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$.

In many cases the likelihood is also intractable, due to the presence of hidden variables (or missing data), and some form of imputation method is usually required in which the missing information is inferred. Data-augmentation (DA) methods (e.g., Gibson and Renshaw, 1998, O'Neill and Roberts, 1999, Jewell et al., 2009) provide a flexible and powerful framework for inference, where the parameter space is augmented to include the hidden variables ($\boldsymbol{x}$). The marginal posterior distribution of interest is then given by

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) = \int_{\mathcal{X}} \pi(\boldsymbol{\theta}, \boldsymbol{x} \mid \boldsymbol{y})\, d\boldsymbol{x}, \tag{2}$$

where $\mathcal{X}$ corresponds to the (multidimensional) parameter space for the hidden variables. The integrand in (2) can be written as

$$\pi(\boldsymbol{\theta}, \boldsymbol{x} \mid \boldsymbol{y}) \propto \pi(\boldsymbol{y}, \boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}), \tag{3}$$

where the joint likelihood function based on the observed data and the hidden variables, $\pi(\boldsymbol{y}, \boldsymbol{x} \mid \boldsymbol{\theta})$, is now tractable. If joint samples from $\pi(\boldsymbol{\theta}, \boldsymbol{x} \mid \boldsymbol{y})$ can then be produced (using numerical sampling algorithms such as Markov chain Monte Carlo—MCMC), then the integral in (2) is straightforward to evaluate numerically. The uncertainties due to the hidden variables are intrinsically incorporated into the resulting marginal posterior distribution. Despite their flexibility, these methods can quickly become computationally infeasible as the number of hidden variables, and the size and complexity of the system increases; not only because the additional variables must be stored, but also because designing and implementing efficient update schemes for the augmented variables in high dimensions can be very challenging (both methodologically and computationally).

An alternative approach is to consider that the marginal posterior (2) can also be written as

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \left[\int_{\mathcal{X}} \pi(\boldsymbol{y}, \boldsymbol{x} \mid \boldsymbol{\theta})\, d\boldsymbol{x}\right]\pi(\boldsymbol{\theta}), \tag{4}$$

and the integral in (4) can be approximated using importance sampling as

$$\hat{\pi}(\boldsymbol{y} \mid \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(\boldsymbol{y}, \boldsymbol{x}_i \mid \boldsymbol{\theta})}{q_X(\boldsymbol{x}_i \mid \boldsymbol{\theta})}, \tag{5}$$

where $\boldsymbol{x}_i \sim q_X(\cdot \mid \boldsymbol{\theta})$ and $q_X(\cdot \mid \boldsymbol{\theta})$ is a proposal distribution for the hidden variables $\boldsymbol{x}$. Some powerful theoretical results follow from this estimator. Indeed, Beaumont (2003) proved that using the estimator (5) in an MCMC algorithm of the form given in Algorithm 1(a) [with (5) replacing $\hat{\pi}(\boldsymbol{y} \mid \boldsymbol{\theta})$], gave exact posterior samples in probability. This result was

Algorithm 1 The (a) ABC-MCMC algorithm of Marjoram et al. (2003) (left-panel) and the (b) ABC-SMC algorithm of Toni et al. (2009) (right-panel).

A1. Initialise the tolerance ε, the number of iterations n_{iter}.
A2. Sample an initial set of parameters $\theta^{(0)} \sim \pi(\theta)$.
A3. Generate n data sets $\boldsymbol{z}_i^{(0)} \sim \pi(\cdot \mid \theta^{(0)})$ and calculate $\hat{\pi}(\boldsymbol{y} \mid \boldsymbol{z}_i^{(0)}) = (1/n)\sum_{i=1}^{n} \mathbb{1}(\rho(\boldsymbol{y}, \boldsymbol{z}_i^{(0)}) < \varepsilon)$.
A4. If $\hat{\pi}(\boldsymbol{y} \mid \boldsymbol{z}_i^{(0)}) = 0$ go to step A2.
A5. Set iteration indicator $j = 1$.
A6. Sample a candidate value $\theta' \sim Q(\cdot \mid \theta^{(j)})$ from some Markov transition kernel $Q(\cdot)$.
A7. Generate n data sets $\boldsymbol{z}_i' \sim \pi(\cdot \mid \theta')$ and calculate $\hat{\pi}(\boldsymbol{y} \mid \boldsymbol{z}') = (1/n)\sum_{i=1}^{n} \mathbb{1}(\rho(\boldsymbol{y}, \boldsymbol{z}_i') < \varepsilon)$.
A8. Set $\theta^{(j)} = \theta'$ and $\hat{\pi}(\boldsymbol{y} \mid \boldsymbol{z}^{(j)}) = \hat{\pi}(\boldsymbol{y} \mid \boldsymbol{z}')$ with probability

$$\alpha = \min\left(1, \frac{\hat{\pi}(\boldsymbol{y} \mid \boldsymbol{z}')}{\hat{\pi}(\boldsymbol{y} \mid \boldsymbol{z}^{(j-1)})} \times \frac{\pi(\theta')}{\pi(\theta^{(j-1)})} \times \frac{Q(\theta^{(j-1)} \mid \theta')}{Q(\theta' \mid \theta^{(j-1)})}\right),$$

else set $\theta^{(j)} = \theta^{(j-1)}$ and $\hat{\pi}(\boldsymbol{y} \mid \boldsymbol{z}^{(j)}) = \hat{\pi}(\boldsymbol{y} \mid \boldsymbol{z}^{(j-1)})$.
A9. If $j < n_{\text{iter}}$, increment $j = j + 1$ and go to step A6.

B1. Set the number of generations T, and the number of particles n_{part}.
B2. Initialise the tolerances $\varepsilon_1, \ldots, \varepsilon_T$. Set population indicator $t = 1$.
B3. Set particle indicator $j = 1$.
B4. If $t = 1$, sample θ'' independently from $\pi(\theta)$. If $t > 1$, sample θ' from the previous population $\{\theta_{t-1}\}$ with weights $\{W_{t-1}\}$, and perturb the particle to $\theta'' \sim Q_t(\cdot \mid \theta')$ according to a Markov transition kernel $Q_t(\cdot)$.
B5. If $\pi(\theta'') = 0$, return to B4.
B6. Generate n data sets $\boldsymbol{z}_i'' \sim \pi(\cdot \mid \theta'')$, and calculate $\hat{\pi}(\boldsymbol{y} \mid \boldsymbol{z}'') = (1/n)\sum_{i=1}^{n} \mathbb{1}(\rho(\boldsymbol{y}, \boldsymbol{z}_i'') < \varepsilon_t)$.
B7. If $\hat{\pi}(\boldsymbol{y} \mid \boldsymbol{z}'') = 0$, then go to B4.
B8. Set $\theta_t^{(j)} = \theta''$ and

$$W_t^{(j)} = \begin{cases} \hat{\pi}(\boldsymbol{y} \mid \boldsymbol{z}'') & \text{if } t = 1, \\ \dfrac{\hat{\pi}(\boldsymbol{y} \mid \boldsymbol{z}'')\pi(\theta_t^{(j)})}{\sum_{j=1}^{n_{\text{part}}} W_{t-1}^{(j)} Q_t(\theta_t^{(j)} \mid \theta_{t-1}^{(j)})} & \text{if } t > 1. \end{cases}$$

B9. If $j < n_{\text{part}}$, increment $j = j + 1$ and go to step B4.
B10. Normalise the weights so that $\sum_{j=1}^{n_{\text{part}}} W_t^{(j)} = 1$.
B11. If $t < T$, increment $t = t + 1$ and go to B3.

later generalised by Andrieu and Roberts (2009), who showed that this holds for any nonnegative unbiased estimator of $\pi(\boldsymbol{y} \mid \boldsymbol{\theta})$. In practice, the key challenge is finding efficient proposal distributions, $q_X(\cdot \mid \boldsymbol{\theta})$, for the hidden variables, which can be difficult for complex nonlinear models (see, e.g., Andrieu, Doucet and Holenstein, 2010, McKinley et al., 2014, Drovandi, Pettitt and McCutchan, 2016).

1.1 Direct Simulation from the Underlying Model

In certain situations, it may be possible to simulate outputs directly from an underlying statistical model, which can then be mapped to the observed data in an appropriate manner. In this case, equation (4) becomes

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \left[\int_{\mathcal{Z}} \pi(\boldsymbol{y} \mid z, \boldsymbol{\theta})\pi(z \mid \boldsymbol{\theta})dz\right]\pi(\boldsymbol{\theta}), \tag{6}$$

where $\pi(z \mid \boldsymbol{\theta})$ is the likelihood function for a single realisation of the underlying process, z, and $\pi(\boldsymbol{y} \mid z, \boldsymbol{\theta})$ is a probabilistic mapping between the realisation z and the observed data $\boldsymbol{y}$. ($\mathcal{Z}$ corresponds to the space of all possible realisations of z.) We can then write equation (5) as

$$\hat{\pi}(\boldsymbol{y} \mid \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \pi(\boldsymbol{y} \mid z_i, \boldsymbol{\theta}), \tag{7}$$

where $z_i \sim \pi(\cdot \mid \boldsymbol{\theta})$ corresponds to a single simulation from the underlying model. In the special case that we require exact matching between the simulated data and the observed data, then

$$\pi(\boldsymbol{y} \mid z_i, \boldsymbol{\theta}) = \begin{cases} 1 & \text{if } z_i = \boldsymbol{y}, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

In other cases, we could define the mapping between z_i and $\boldsymbol{y}$ to have a specific probabilistic form (for example, if the observed data were derived from an imperfect diagnostic test). From now on, we will focus on systems that can be written in the form described by (6). (Note that there are also promising methods that involve recoding or reparameterising the simulation model [e.g., Neal, 2012, McKinley et al., 2014, Kypraios, Neal and Prangle, 2017]. However, these approaches are not feasible for all models, and can be difficult to scale to very complex systems, so here we focus on methods that simulate directly from the underlying model.)

These ideas, coupled with the fact that it is often far easier to code a simulation model than reconstruct a likelihood function based around a large number of hidden variables, have facilitated the development of various 'simulation-based' methods for inference, where calculation of the likelihood is replaced by an estimate derived from simulations from the underlying model, an idea that goes back at least as far as Diggle and Gratton (1984) and Rubin (1984). The key bottleneck in the implementation of these methods is that even for small-scale systems, the probability of generating simulations that match all observed data points exactly is often very small. This often precludes direct implementation of these approaches, and instead motivated the development of a suite of techniques now known colloquially as *Approximate Bayesian Computation* (ABC) (e.g., Tavaré et al., 1997). In recent years, these techniques have exploded in popularity, since these ideas can be readily incorporated into existing numerical algorithms, such as rejection sampling (e.g., Tavaré et al., 1997, Beaumont, Zhang and Balding, 2002); MCMC (e.g., Marjoram et al., 2003, Ratmann et al., 2009, Wood, 2010); or sequential Monte Carlo (SMC) (e.g., Sisson, Fan and Tanaka, 2007, Toni et al., 2009, Beaumont et al., 2009, Del Moral, Doucet and Jasra, 2012, Drovandi and Pettitt, 2011, Lenormand, Jabot and Deffuant, 2013). Due to their relative ease-of-implementation, simulation-based methods are being increasingly adopted in stochastic epidemic modelling (e.g., O'Neill et al., 2000, Toni et al., 2009, McKinley, Cook and Deardon, 2009, McKinley et al., 2014, Neal, 2012, Conlan et al., 2012, Brooks Pollock, Roberts and Keeling, 2014, Kypraios, Neal and Prangle, 2017).

Good reviews of ABC can be found in Csilléry et al. (2010) and Beaumont (2010), and a more recent and technical review can be found in Marin et al. (2012). In many applications, vanilla rejection sampling approaches are hard to implement efficiently, and so most ABC routines in the literature are based around either MCMC or SMC methods; two popular examples are shown in Algorithms 1(a) and 1(b). We assume the reader is familiar with both SMC and MCMC methods (see, e.g., Marjoram et al., 2003, Toni et al., 2009). A recent tutorial for implementing ABC-MCMC methods for temporal stochastic epidemic models can be found in Kypraios, Neal and Prangle (2017). The fundamental challenge for implementation of ABC is that in many circumstances the probability of getting an exact match between the simulations and the data is vanishingly small, and there have been myriad innovations to try to alleviate this problem. Here we briefly introduce some of the more common ABC-type approaches, before focusing on key challenges when applying these methods to high-dimensional and computationally intensive models.

2. 'CLASSIC' ABC

Instead of requiring the simulations to match the data exactly, a distance metric, $\rho(\cdot,\cdot)$, can be introduced, and thus (7) can be approximated by

$$\hat{\pi}(\boldsymbol{y} \mid \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\big(\rho(\boldsymbol{y}, \boldsymbol{z}_i) < \varepsilon\big), \tag{9}$$

with ε defining some *tolerance* for matching. Using the estimator (9) as an estimate of the likelihood in standard numerical procedures will produce samples from the *approximate* posterior $\pi(\boldsymbol{\theta} \mid \rho(\boldsymbol{y}, \cdot) < \varepsilon)$. Generally speaking, the metric is set-up such that $\rho(\boldsymbol{y}, \boldsymbol{z}_i) \to 0$ as $\boldsymbol{z}_i \to \boldsymbol{y}$, and hence as $\varepsilon \to 0$ the approximate posterior will tend to the true posterior, but at a greater computational cost.

2.1 The Impact of the Tolerance

A key consideration is fixing (or reducing) the tolerance levels to be as small as possible (in order to minimise information loss in the approximate posterior), whilst retaining a reasonable acceptance rate. SMC methods are well-suited to the ABC framework, since they allow initial generations to use less restrictive tolerances than subsequent generations, which often makes them more efficient at exploring the parameter space than ABC-MCMC, provided a good set of initial particles can be found (see, e.g., Toni et al., 2009, McKinley, Cook and Deardon, 2009). Adaptive schemes are often used (Beaumont et al., 2009, Del Moral, Doucet and Jasra, 2012, Drovandi and Pettitt, 2011, Lenormand, Jabot and Deffuant, 2013), in which the choice of tolerance at each generation is determined as a function of the simulated metric distances at the previous generation (see Silk, Filippi and Stumpf, 2012 for some critique of these approaches).

ABC-MCMC methods tend to use a fixed tolerance for the entire chain, with a few notable exceptions: for example, Ratmann et al. (2007) use a tempering method to reduce the tolerance during the burn-in phase, before fixing the tolerance to collect the final samples, and Bortot, Coles and Sisson (2007) introduce a data-augmentation approach, in which they place a shrinkage pseudo-prior on the tolerance and estimate this as part of the model fitting.

2.2 Matching to Multiple Outputs

Acceptance rates are affected further when matching to multiple outputs. Here there are two main options: the first, the so-called *intersection* approach, sets a separate distance metric around each of the K outputs, each with its own tolerance. A simulation is then accepted if

$$\prod_{k=1}^{K} \mathbb{1}\big(\rho_k(\boldsymbol{y}, \boldsymbol{z}) < \varepsilon_k\big) = 1, \tag{10}$$

where $\boldsymbol{z}$ corresponds to the simulated data. The simulation must therefore match *each* output simultaneously in order to be accepted. An alternative is to create a single metric, $\rho^*(\cdot,\cdot)$, and accept a simulation if:

$$\mathbb{1}\big(\rho^*(\boldsymbol{y}, \boldsymbol{z}) < \varepsilon\big) = 1, \tag{11}$$

where $\rho^*(\boldsymbol{y}, \boldsymbol{z}) = f(\rho_1(\boldsymbol{y}, \boldsymbol{z}), \ldots, \rho_K(\boldsymbol{y}, \boldsymbol{z}))$, and $f(\cdot)$ is some function of the K outputs (e.g., Conlan et al., 2012). This is termed a *union* metric (see also Ratmann et al., 2014).

The trade-off between the two choices varies according to the particular system being modelled, but heuristically one can think of the union metric as smoothing out some of the patterns in the data, that is, the models are allowed to fit certain outputs less well than others, provided that the overall fit is reasonable. Combining metrics in a sensible manner is sometimes challenging, especially if they are defined on different scales (see, e.g., Conlan et al., 2012). Union metrics can sometimes lead to simulations being regularly accepted when they do not fit certain outputs very well at all, whereas intersection metrics can penalise misfitting simulations more, but at a cost of reduced acceptance rates. In the case of ABC, we expect the probability of rejecting a simulation to scale with K (although of course the exact relationship is harder to quantify, since some of the metrics may be correlated).

2.3 The Use of Summary Statistics

Based on the previous discussion, if the dimensionality of the observed data is large then it can be challenging to design a computationally efficient algorithm with minimal information loss in the approximate posterior. In a handful of cases, it may be possible to reduce the data to a set of lower-dimensional *sufficient* statistics, that contain the same amount of information as the full data. More often that not, sufficient statistics are unknown (or are equal to the data), and so often a set of lower-dimensional summary measures, $S_1(\boldsymbol{y}), \ldots, S_L(\boldsymbol{y})$, are used in their place (where $L < K$). The key questions are then: how well do the summary statistics capture the information in the data, and how do any biases introduced manifest in any inferences that we make from the model? Increasing research effort has been placed into deriving *approximately* sufficient summary measures (e.g., Joyce and

Marjoram, 2008, Nunes and Balding, 2010, Barnes et al., 2012, Fearnhead and Prangle, 2012, Ratmann et al., 2014). The use of summary statistics can also be extended to *indirect inference* methods, where the auxiliary models describe the distributions of the summary statistics (see Section 2.7).

2.4 Increasing the Number of Replicates ($n > 1$)

Interestingly, theoretical convergence of Algorithms 1(a) and 1(b) do not depend on the number of simulations, n, used in the estimator (9) (Andrieu and Roberts, 2009, Del Moral, Doucet and Jasra, 2012). For more general classes of pseudo-marginal algorithms, it has been shown that increasing n can improve the efficiency of the algorithms by reducing the variance of the estimator (5) (see, e.g., Pitt et al., 2012, Sherlock et al., 2015, Doucet et al., 2015). In the specific case of ABC-MCMC with uniform matching, Bornn et al. (2017) show that setting $n = 1$ results in run times that are at most a factor of 2 away from the optimum choice (obtained for some $n > 1$). However, their results also make the assumption that simulation run times are approximately constant, which is often not true for epidemic systems, where run times for individual simulations can often vary greatly even for fixed parameter inputs. Also, in the case of ABC-MCMC and more general pseudo-marginal algorithms, chains using low values of n can often get 'stuck' and fail to mix practically at all (see, e.g., McKinley, Cook and Deardon, 2009, Andrieu and Roberts, 2009). Mixing can generally be improved by increasing the tolerance(s), but at the cost of further information loss in the approximate posterior. Under the same assumptions as above, Bornn et al. (2017) show that for a simple rejection sampling ABC algorithm, $n = 1$ is indeed optimal. In practice ABC-SMC samplers, such as described in Algorithm 1(b) seem to perform better for low n (see, e.g., McKinley, Cook and Deardon, 2009), and it is for this reason that we choose to use ABC-SMC instead of ABC-MCMC for tackling the model in this paper.

Another option to alleviate the mixing issues in pseudo-marginal MCMC algorithms for low n is to refresh the $\hat{\pi}(\boldsymbol{y} \mid \theta)$ estimates for both the candidate and current parameters at each iteration of the chain (see, e.g., O'Neill et al., 2000, Andrieu and Roberts, 2009, McKinley et al., 2014). This exhibits substantially better mixing, at the cost of producing biased samples, with the bias decreasing as $n \to \infty$. It also doubles the number of simulations required per iteration of the chain, though this is often mediated by requiring shorter chains due to the improvement in mixing.

2.5 Interpretation of ABC Posterior

The term ABC derives from the fact that these methods were originally developed to obtain an approximation to the 'true' posterior (Tavaré et al., 1997). Wilkinson (2013) showed that in certain circumstances, for a fixed metric and (final) tolerance, ABC can be interpreted as giving the *exact* posterior under the assumption of model error. For example, if $\rho(\cdot, \cdot)$ is based on Euclidean distances, then up to some normalising constant, (9) corresponds to assuming uniform error around the observed data $\boldsymbol{y}$ [these normalising constants then cancel in the accept-reject steps of Algorithms 1(a) and (b)].

Wilkinson (2013) cites two possibilities for the interpretation of the error term: observation error or model discrepancy. The former is generally well understood, and it is often possible to either build this directly into the simulation code, or define this in terms of a probabilistic function mapping the hidden states to the observed data. The idea of model discrepancy (MD) is less familiar, and more difficult to define, but relates to the disparity between the model and reality. It has been argued to be an important source of uncertainty that should be incorporated into calibration routines to prevent over interpretation due to the choice/assumptions of the model, and hence increase robustness (e.g., Goldstein and Rougier, 2009, Oakley and Youngman, 2017). When viewed in this way, ABC ceases to be approximate. In practice, for this interpretation to be meaningful requires that the form and magnitude of MD is considered in advance and specified in epidemiologically relevant terms (see, e.g., Section 2.6). When we discuss 'classic' ABC in this paper, we do so in its original paradigm, that of an approximation to the posterior in the absence of MD.

Another important contribution is given in Fearnhead and Prangle (2012), in which they reframe ABC inference in terms of a set of desired properties (defined as *accuracy* and *calibration*), and provide methods for selecting summary statistics to optimise these desired characteristics.

2.6 Generalised ABC and Post-Processing

Beaumont, Zhang and Balding (2002) suggested improving the posterior approximation by *post-processing* the final set of parameters; reweighting each according to the distance between the simulated outputs and the data using localised linear-regression (see also Blum and François, 2010). An alternative is to choose some nonuniform discrepancy distribution for use directly

within the ABC estimate of the likelihood. Hence, (9) becomes

$$\hat{\pi}(\boldsymbol{y} \mid \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \pi(\boldsymbol{y} \mid \boldsymbol{z}_i, \varepsilon), \tag{12}$$

where $\varepsilon > 0$ now defines some variance controlling the discrepancy between the simulated and observed data sets. Wilkinson (2014) terms this *generalised* ABC (GABC). A natural choice of discrepancy distribution is one that has a single mode, centred around the observed data, such that $\pi(\boldsymbol{y} \mid \boldsymbol{z}_i, \varepsilon) \to 0$ as the distance between $\boldsymbol{y}$ and $\boldsymbol{z}_i$ increases. As an example, we could place independent Gaussian distributions with variance ε around each data point, or even used different variances for different data points. [Note also that (12) also includes the uniform discrepancy distribution discussed earlier as a special case.] The lack of hard-bound on the discrepancy distribution removes the problem of 'matching', but at the potential cost of (12) having a high Monte Carlo variance unless a large number of replicates is used. This could lead to mixing issues in ABC-MCMC and particle degradation in ABC-SMC. A truncated, but nonuniform, error term could alleviate the high uncertainty when simulating in the tails of the discrepancy kernel (see also the ideas in Bortot, Coles and Sisson, 2007 and Beaumont, Zhang and Balding, 2002).

2.7 Indirect Inference

There are also a series of approaches that are akin to the methods of *indirect inference* (Gouriéroux, Monfort and Renault, 1993), whereby an auxiliary model is introduced to describe the distribution of the data, and inference is based on comparison of the parameters of the auxiliary model as estimated through repeated simulations from the model-of-interest (e.g., Wood, 2010, Drovandi, Pettitt and Faddy, 2011, Ratmann et al., 2014). The *synthetic likelihood* approach of Wood (2010) assumes a parametric form (e.g., multivariate normal) for the distribution of outputs arising from repeated model simulations. The parameters of this auxiliary model are estimated from the simulations, and a synthetic likelihood can be constructed by estimating the likelihood that the observed data come from the auxiliary model. A huge advantage of this method is that there is no need to choose tolerance levels for the matching, though a suitable auxiliary model must be found (which is sometimes challenging), and replicate simulations per parameter set are necessary.

An insightful paper by Drovandi, Pettitt and Lee (2015), showed that classical ABC and the synthetic likelihood approaches are both special cases of a more general class of models, which they call Bayesian indirect likelihood (BIL) models. They show that in general convergence of the synthetic likelihood approach to the true posterior is not guaranteed, however the method often performs well if the auxiliary model is flexible enough to match the simulations to the data well in the region of nonnegligible posterior mass.

3. CHALLENGES FOR COMPUTATIONALLY INTENSIVE MODELS

In the previous section, we briefly reviewed various recent advances in simulation-based inference for statistical models. There are many possible choices of approach, with trade-offs in terms of computational complexity, accuracy, bias, interpretation and ease-of-implementation. The ability to plug a simulation algorithm into existing routines have made ABC-type methods attractive as a potential tool for statistical inference in large-scale, complex systems, such as those frequently studied in epidemiology (see also Ionides, Bretó and King, 2006, Ionides et al., 2011, 2015 for frequentist approaches). In addition, it is often straightforward to parallelise the simulations.

Nonetheless, most methodological research has focused on the development of ideas and theories applied to relatively small scale models or data sets, and even some of these simpler examples can take between several hundred thousand model runs to many millions (e.g., Kypraios, Neal and Prangle, 2017). In our opinion, one of the major current challenges in the field is how to perform robust inference when the simulation models are highly computationally intensive; precluding the running of very large numbers of simulations. These systems often go hand-in-hand with high-dimensional input (parameter) and output (data) spaces, and in this paper we illustrate some of these challenges using a complex, large-scale, high-dimensional model of HIV transmission. This model, called Mukwano, is an individual-based stochastic micro-simulation model, that simulates (amongst other things): heterosexual sexual partnerships, sexual activity, HIV transmission and life histories (including births and deaths). Different versions of the model exist, but the version studied here has 22 input parameters, and 18 outputs.

For brevity, we refer the reader to Andrianakis et al. (2015) for full details of the model, but briefly the model simulates heterosexual sexual partnerships (partnership formation, dissolution, and concurrency)

and HIV transmission, alongside demographic events such as births and deaths in a population of individuals. The data come from a long-term (25+ years) longitudinal study of an open cohort of ≈18,000 individuals in rural Uganda. We used informative uniform priors for the 22 inputs, and full details of the model and priors can be found in Andrianakis et al. (2015).

The model has an average run time of ≈5–10 mins per simulation (in the well-supported region—it can be far longer [>3 hours] in some areas of poor support). Based on the discussions in earlier sections, and our own experience, it was decided that an ABC-SMC algorithm, using a single simulation per particle ($n = 1$) would be a sensible choice of routine to try to tackle this problem. Some initial tests using GABC with normally distributed discrepancy terms resulted in extremely high Monte Carlo errors for the GABC likelihood estimate in parts of the space where the model fit was poor. As such we instead used a uniform error term and implemented an intersection approach in which all 18 outputs were matched simultaneously. We set a nonzero minimum bound for the tolerance, relating to roughly twice the observation standard deviation for each output used in Andrianakis et al. (2015).

We implemented the ABC-SMC routine of Toni et al. (2009) [Algorithm 1(b)], using the optimal localised multivariate kernel approach of Filippi et al. (2013). In order for ABC to work well, there must be a large enough number of particles located in areas of high posterior support, and so we generated an initial set of 22,000 particles uniformly from the prior distribution. Here we choose to match to 18 outputs simultaneously, which requires 18 tolerances defined on different scales. We chose initial tolerance values to be the 50th percentile of the simulated metric distances for each of the 18 outputs, and chose tolerances at generation $t + 1$ using a simple bisection method [detailed in Supplement A (McKinley et al., 2017)], where the proportion of generation t particles that would be accepted using the new tolerances was approximately $p_\tau = 0.5$. (We note that this method allows for semi-automatic nonuniform adjustments of the tolerances at each generation of ABC, and can also be applied to outputs that are defined on different scales.)

The results for 11 generations of ABC-SMC are shown in Figure 1, which shows some interesting behaviour. For most outputs there is a steady convergence towards the observed data. However, for one output in

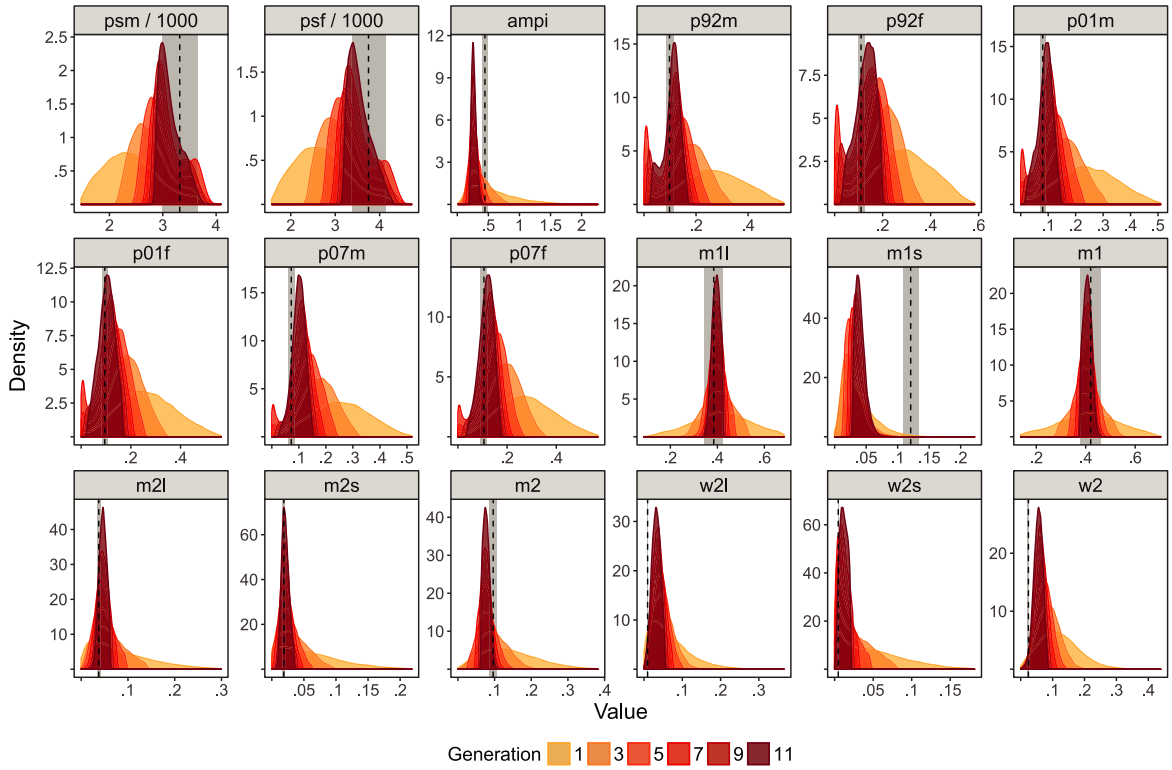

FIG. 1. *Model fits for* 11 *generations of ABC. These show the marginal predictive distributions for the model outputs conditional on the set of ABC particles at different generations of ABC. For brevity we only show generations* 1, 3, 5, 7, 9 *and* 11 *here. The dotted lines denote the data and the target regions are shown in grey.*

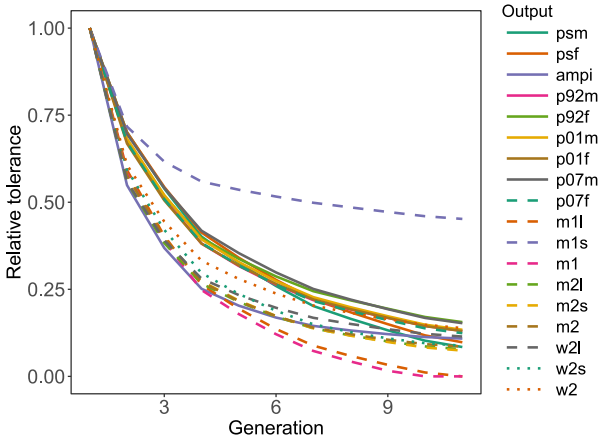

FIG. 2. *Relative tolerance evolution (right panel) for* 11 *generations of ABC-SMC fitted to Mukwano. (Note that a relative tolerance of* 1 *corresponds to the initial tolerance value at the first generation of ABC-SMC, and a relative tolerance of* 0 *corresponds to the target tolerance defined by the observation error, as shown by the grey regions in Figures* 1 *and* 3.)

particular (*m1s*), even after 11 generations of ABC the simulated outputs are far lower than the observed data (see also the *ampi* output). Figure 2 shows the relative tolerances across generations, rescaled such that a value of one corresponds to the initial tolerance, and zero to the target (observation) tolerance. The algorithm should stop when each line crosses zero. The blue line relating to the *m1s* output seems to be asymptoting at a level far higher than we require, and indeed for many of the others the rate-of-change of tolerance values between the generations is also slowing. Although the aim is to generate an acceptance rate of around 0.5 for each generation, in the later generations the actual value is much smaller than this (Table 1). At this point, the final generation took almost 2 days to run on a high-performance cluster.

These anticipated challenges for ABC in higher dimensions present a difficult set of choices for the standard ABC paradigm: do we continue to run the algorithm as before, considering the decreasing convergence rate; do we change criteria in the algorithm, perhaps choosing smaller tolerance thresholds at the cost of decreasing acceptance rates further; do we change metric; or do we stop the algorithm? On the one hand there seems to be convergence towards the data, so one could continue with the ABC. On the other hand the rate of convergence is slowing, and the number of simulations required at each generation is increasing, to the point that we may begin to question the logic of continuing. If the tolerances asymptote to a level that is far away from the data, then the key question is: can the model fit the data adequately, or have we simply not explored the parameter space sufficiently? Sometimes the trade-offs in accuracy required to get ABC algorithms to fit in a computationally feasible manner can be large, and can lead to situations in which the current 'best-fit' from the ABC algorithm is sufficiently poor that we are unable to make useful inferences about the system.

3.1 Emulation

A advance for approximating outputs when a simulation model is computationally expensive is to appeal to the use of an *emulator* (e.g., Sacks et al., 1989). This is a statistical representation of the simulation model that can be used as a surrogate in simulation-intensive routines. Typically, the complex model is run at a series of 'design' points, and the emulator is trained on the simulated outputs at each of these points. Once trained, the emulator can be used to predict the outputs from the complex model (as well as to provide measures of uncertainty in the predictions) very quickly.

Emulators in ABC. There have been several recent applications of using emulators within ABC. Henderson et al. (2009), Jandarov et al. (2014), Wilkinson (2014), Meeds and Welling (2014) and Cameron et al. (2015) each implement MCMC algorithms where the true likelihood is replaced by that derived from an emulator. Important differences between the methods lie in how the emulator is trained. Henderson et al. (2009) use a fixed set of design points (chosen to cover the input space), and Jandarov et al. (2014) use a grid design. These are feasible because the input and output spaces are low-dimensional, but would be challenging in high dimensions, since enough

TABLE 1
Acceptance rates for ABC-SMC

Generation	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**Total**
Acc. rate	0.78	0.64	0.59	0.49	0.44	0.37	0.3	0.27	0.2	0.16	0.15	
Number sims.	28,363	34,407	37,377	45,062	49,816	58,989	72,325	82,236	109,651	139,329	142,764	**800,319**

points would be required to produce a sufficiently accurate emulator. Meeds and Welling (2014) train the emulator as the MCMC progresses, choosing new design points based on local moves around the current point of the chain. Jabot et al. (2014) embed the emulator in both rejection and SMC samplers, using initial design points sampled from the prior. ABC steps are then run using the emulator as a surrogate, and new training points chosen at each generation based on the current set of particles. Cameron et al. (2015) use functional regression to emulate a microsimulation model of malaria infection, which they use to generate an approximate posterior through MCMC.

These approaches look promising, but require an emulator that is sufficiently accurate to represent the complex model across the input space, which in turn requires careful design of training points. In the next section, we discuss an alternative methodology that is specifically designed to rigorously and efficiently explore high-dimensional input spaces to reject areas where the model fits are poor.

3.2 History Matching

History Matching (HM) is a technique developed in the Bayesian computer model literature for finding acceptable inputs to expensive complex models that have high-dimensional input and output spaces (Craig et al., 1997). It has been successfully employed across a range of scientific disciplines, both for deterministic and stochastic models (see Vernon, Goldstein and Bower, 2010, 2014, Andrianakis et al., 2015 and references therein). While there may appear to be superficial similarities between HM and various versions of ABC, the techniques are distinct both in terms of their goal and their implementation. HM is not an inferential procedure, but instead seeks to identify the regions of input space that produce acceptable matches between model and data, where 'acceptable' is defined via an underlying statistical model that incorporates a careful consideration of major uncertainties: observational errors, model discrepancy and others (e.g., stochasticity).

It proceeds by cutting out regions of the input space in iterations or waves, using *implausibility measures*. In each wave t, we design a set of model runs over the current input space Θ_t. The set of outputs is denoted $\mathcal{K} = \{1, \ldots, K\}$. Emulators (such as Gaussian processes) are constructed only over Θ_t to mimic informative outputs of the model (deterministic case) or summaries of outputs (stochastic case), denoted $f_k(\theta)$, providing estimates of the expected values and variances, $\mathrm{E}(f_k(\theta))$ and $\mathrm{Var}(f_k(\theta))$, respectively. At wave t, it is only necessary to choose a set of outputs $k \in \mathcal{K}_t$ than can both be emulated sufficiently accurately, and that are informative: usually this set increases in size at each wave. An implausibility measure can then be constructed for each emulated output $f_k(\theta)$, $k \in \mathcal{K}_t$ (more advanced implausibility measures are available):

$$I_k^2(\theta) = \frac{(\mathrm{E}(f_k(\theta)) - y_k)^2}{\mathrm{Var}(f_k(\theta)) + \mathrm{Var}(\varepsilon_k) + \mathrm{Var}(e_k)}. \tag{13}$$

Here y_k is the observed data, and $\mathrm{Var}(\varepsilon_k)$ and $\mathrm{Var}(e_k)$ are the variances due to model discrepancy and observation error respectively. The structure of $I_k(\theta)$ is derived from an underlying statistical model (Vernon, Goldstein and Bower, 2010), which dictates how to combine the different sources of uncertainty. Because the specified uncertainties are meaningful, unlike the tolerances in standard ABC, the implausibility is also now on a meaningful scale, and we can apply cutoffs on $I_k(\theta)$ directly (motivated by Pukelsheim's 3σ rule Pukelsheim, 1994) to remove implausible parts of the input space if $I_k(\theta) > c$ (where often $c = 3$). Large amounts of the input space Θ_t can often be removed based on a single (or a small combination of) output(s), to define a reduced space Θ_{t+1}. Further waves are performed unless (a) the emulator variances $\mathrm{Var}(f_k(\theta))$ for all outputs of interest are now small in comparison to the other sources of uncertainty $\mathrm{Var}(\varepsilon_k) + \mathrm{Var}(e_k)$, or (b) the entire input space has been deemed implausible.

Why is this a useful approach? HM works well in high dimension for several reasons (Vernon, Goldstein and Bower, 2010). It provides a fast, meaningful decision, based on a subset of outputs, as to whether an input point is implausible that is independent of the rest of the input space, and hence can quickly discard vast regions of input space without modelling the whole set of outputs. Note that these regions will most likely contain extremely low posterior probability, hence, although HM does not seek a Bayesian posterior, it is a very useful precursor if one subsequently wishes to do so. Critically, at each wave the emulator accuracy is expected to improve, and structured emulators involving dimensional reduction can be designed to exploit this. Often an individual output may strongly depend only on a small subset of 'active' inputs (e.g., Vernon, Goldstein and Bower, 2010), and hence the implausibility structure allows us to break a high-dimensional problem into a series of lower-dimensional ones. There may also be several outputs that are difficult to emulate in early waves (perhaps because of their erratic behaviour in uninteresting parts of the input space) but simple to

emulate in later waves in smaller, more realistic input regions. HM thus allows the sequential incorporation of outputs of increasing complexity.

The differences between HM and ABC: at each wave of HM, all the emulators and implausibility cuts from previous waves are also used. Hence, unlike most ABC implementations, HM 'remembers' regions of space that were previously deemed implausible. This is vital in high dimensions to avoid unnecessarily retesting many input locations known to be unacceptable. ABC usually seeks to approximate a Bayesian inference calculation, using ever decreasing tolerances that can cause computational inefficiencies. However, HM is not an inferential procedure, nor is it 'approximate', and uses tolerances that are derived from a well defined statistical model that incorporates realistic assessments of uncertainty that are usually elicited from subject matter experts or by performing simple alternative experiments on the model (Goldstein, Seheult and Vernon, 2013). They are hence interpretable, can be substantial, and are not reduced to arbitrarily small sizes, and can alleviate these computational inefficiencies. The statistical model also facilitates the incorporation of additional uncertainties for example, from the emulator (while exploiting their independence) and the direct use of implausibility, leading to a more efficient parameter search. HM has natural stopping criteria, since either the entire space will be ruled as implausible, implying that the complex model is deficient, or the emulators will achieve sufficient accuracy to determine the acceptable set of inputs. While one can mimic certain parts of a basic HM analysis using ABC (Holden et al., 2016), it is hard to justify this from the ABC paradigm alone, and it would arguably lead to an analysis that is not 'ABC' in nature. (An interesting variation of HM is given in Wilkinson, 2014, in which HM is used to match to the GABC log-likelihood. Once trained the emulator is then used directly in ABC.)

In Andrianakis et al. (2015), a history matching approach was applied to the Mukwano model described previously. The model fits are shown in Figure 3, and required around 355,000 simulation runs (less than half the number of runs required for 11 generations of ABC). Clearly the model *is* capable of producing fits

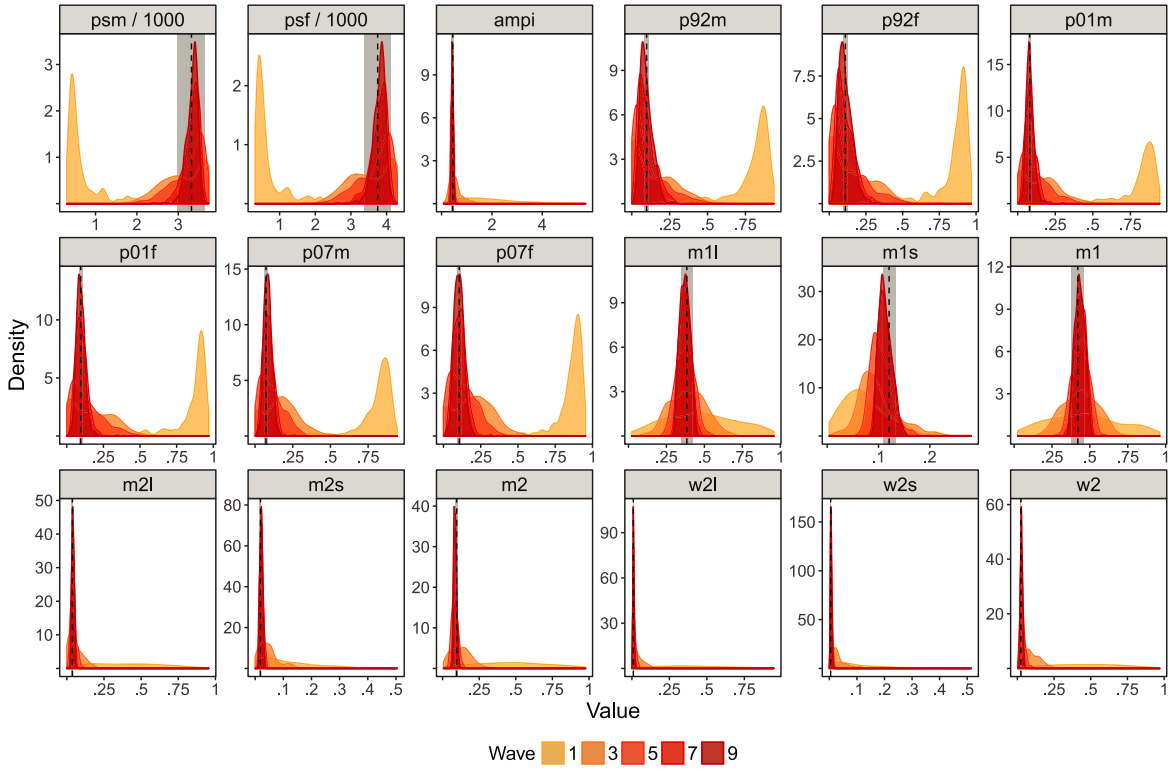

FIG. 3. *Model fits after* 9 *waves of history matching. These show the marginal distributions of the mean outputs* (*from a series of replicate simulations*), *conditional on a set of design points sampled uniformly from the nonimplausible region at each wave. For brevity, we only show waves* 1, 3, 5, 7 *and* 9 *here. The dotted lines denote the data and the target regions are shown in grey. Results from* Andrianakis et al. (2015).

TABLE 2

Acceptance rates for history matching. The top lines correspond to using the tolerances from the final generation (11) *of ABC as shown in Figure 2. The bottom lines correspond to the target tolerance as defined in* Andrianakis et al. (2015). *Since we use repeated simulations per design point for the history matching, these results are shown for 'average' simulations and individual replicate simulations*

		psm	**psf**	**ampi**	**p92m**	**p92f**	**p01m**	**p01f**	**p07m**	**p07f**	
ABC tolerance	Mean	1.00	0.99	1.00	0.93	0.96	0.98	0.98	0.99	0.97	
	Replicate	0.98	0.97	1.00	0.89	0.92	0.97	0.96	0.98	0.94	
Target tolerance	Mean	0.99	0.98	0.77	0.24	0.21	0.33	0.31	0.34	0.28	
	Replicate	0.91	0.85	0.74	0.22	0.20	0.26	0.27	0.30	0.26	
		m1l	**m1s**	**m1**	**m2l**	**m2s**	**m2**	**w2l**	**w2s**	**w2**	**All**
ABC tolerance	Mean	0.71	1.00	0.75	1.00	1.00	1.00	1.00	1.00	1.00	0.48
	Replicate	0.69	1.00	0.74	0.99	0.99	1.00	1.00	1.00	1.00	0.41
Target tolerance	Mean	0.71	0.50	0.75	0.34	0.26	0.43	0.20	0.18	0.24	0.00
	Replicate	0.69	0.50	0.74	0.31	0.25	0.41	0.18	0.15	0.23	0.00

that are close to the observed data, but the nonimplausible region is only a tiny proportion (10^{-11}) of the original space.

As a comparison against the ABC-SMC algorithm, we have also calculated the acceptance rates for the final wave of history matching (Table 2). These results show how many simulations from the wave 9 design points would have been accepted using the generation 11 tolerances from the ABC-SMC run, and also how many would have been accepted according to the target tolerance that we are aiming for. We have produced acceptance rates according to simulated mean outputs (averaged across multiple replicates per design point), in addition to a replicate-specific estimate (making the simplifying assumption that all replicates from all design points are independent). We can see that using the generation 11 tolerances we would have had an acceptance rate of 0.48 (mean) and 0.41 (replicate), compared to a value of 0.15 for the ABC-SMC (Table 1).

The HM procedure produces high acceptance rates for each output considered on its own, but the curse-of-dimensionality is still clear when trying to match all outputs simultaneously. In fact, none of the simulations match all outputs simultaneously at the target tolerance. Nonetheless, we note that the target tolerances for some of the outputs were small, and so we are happy that we have outputs that are relatively close to these targets. In addition, provided the HM has been performed carefully, if there is a region where all outputs can match simultaneously (in terms of realisations) then this region should be contained in the current nonimplausible region.

One extension to the approach described here is to generate multivariate implausibilities. This has been discussed in Vernon, Goldstein and Bower (2010) in the deterministic model case, but the same framework could be used in the stochastic case provided that we can specify a suitable joint multivariate structure between the simulator outputs. However, these have not been developed yet. Without this, it is simpler to use univariate criteria, and to impose cutoffs on the maximum implausibility to identify joint matches (indeed we view it as a strength of history matching that we can carry out such a combined univariate analysis).

Although HM does not produce an approximate posterior in the same sense as ABC, it is possible to view the marginal densities of nonimplausible points (known as depth plots). We have included these as Supplement B (McKinley et al., 2017). One must be careful when directly comparing ABC posteriors and HM depth plots, since the methods are designed to do different things. The aim of HM is to rule out space safely, using whatever aspects of the data are straightforward to exploit in order to do so. ABC on the other hand attempts to identify regions of high (approximate) posterior mass. Nonetheless, we can see that for some variables (e.g., mhag, fchc3, hacr3) the two approaches are targeting quite different parts of the space. We note that in some of these cases the HM has already ruled parts of the space as implausible, and it is not clear whether the ABC is converging towards

these regions, and progress is slow due to the acceptance rates dropping, or whether the requirement for the ABC to match all outputs simultaneously will result in the algorithm converging to slightly different parts of the space. However, the results in Table 2 suggest that the nonimplausible region identified by the HM routine does a better job at finding simulations that match all outputs simultaneously than the ABC does, based on the current waves/generations. The key bottleneck in this particular example is that the ABC convergence is slowing down considerably, and so it may take many more generations to obtain similarly good fits from the ABC than has been achieved thus far with HM.

4. DISCUSSION

ABC methods are exploding in popularity due to their ease-of-implementation. It is often far more straightforward to simulate from an underlying model than to reconstruct (and efficiently) update large numbers of hidden states. In many cases these methods work well, however, matching simulations to data can be challenging, particularly in highly stochastic systems, and this is exacerbated in high dimensions. The computational bottleneck is the speed of the simulations, and ABC methods allow one to trade accuracy and precision of the approximation against computational load. Understanding how much approximation has been introduced and its impact on the inferential properties of the approximate posteriors is often harder to quantify.

Increasing research effort is being employed to come up with more sophisticated sampling and simulation algorithms to help mediate these trade-offs, but these are difficult to scale to highly computational models. The use of an emulator as a surrogate for a complex simulation model can help overcome or mediate some of the challenges that hamper vanilla ABC routines, notably the curse-of-dimensionality (in both the input and output space), the choice of the number of initial particles and the choice of initial tolerances (and subsequent impact on convergence—see, e.g., Vernon, Goldstein and Bower, 2010, Andrianakis et al., 2015); provided that the emulator can be adequately trained. These techniques are harder to implement however, requiring more user input in terms of building, training and interpreting the emulators. Nonetheless, emulation and HM techniques have successfully been used to analyse large models for example, a 96 input, 50 output version of Mukwano that is currently being used to better understand the spread and control of HIV in Uganda (Andrianakis et al., 2017, McCreesh et al., 2017).

Finally, techniques such as history matching do *not* produce an approximate posterior distribution, which can be of key importance in many applications. Hence, we do not argue the use of history matching and emulation as a replacement for ABC (or similar routines), but rather as a *precursor*, enabling us to focus attention on the part of the parameter space in which the model is known to be able to fit the data reasonably well. It may then be possible to use this information to inform the development of ABC or other routines for more systematic inference. For example, HM could be used to ascertain whether the model is capable of fitting the data at all, and if so to inform the generation of a good set of initial particles for seeding the ABC. It may also be possible to use the nonimplausible region, and correlation structure thereof, to inform the perturbation kernel for ABC-SMC and ABC-MCMC routines. In addition, approaches such as that of Fearnhead and Prangle (2012) rely on adequate training runs, which HM can provide. Future research will focus on ascertaining the feasibility of some of these approaches for complex epidemiological models.

ACKNOWLEDEGMENTS

This work was supported by a Medical Research Council (UK) grant on Model Calibration (MR/J005088/1: http://www.mrc.ac.uk/).

SUPPLEMENTARY MATERIAL

Supplement A: Bisection method (DOI: 10.1214/17-STS618SUPPA; .pdf). Details the bisection method used to generate tolerances at each generation of ABC.

Supplement B: Approximate posterior distributions for ABC vs. nonimplausible region for HM (DOI: 10.1214/17-STS618SUPPB; .pdf). Plots of the approximate posterior distributions after 11 generations of ABC, and depth plots after 9 waves of history matching. (Note that HM does not produce posterior samples, rather these correspond to the densities of nonimplausible points.)

REFERENCES

ANDRIANAKIS, I., VERNON, I., MCCREESH, N., MCKINLEY, T. J., OAKLEY, J. E., NSUBUGA, R. N., GOLDSTEIN, M. and

WHITE, R. G. (2015). Bayesian history matching of complex infectious disease models using emulation: A tutorial and a case study on HIV in Uganda. *PLoS Comput. Biol.* **11**. e1003968.

ANDRIANAKIS, I., MCCREESH, N., VERNON, I., MCKINLEY, T. J., OAKLEY, J. E., NSUBUGA, R. N., GOLDSTEIN, M. and WHITE, R. G. (2017). History matching of a high dimensional HIV transmission individual based model. *SIAM/ASA J. Uncertain. Quantificat.* **5** 694–719.

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

BARNES, C. P., FILIPPI, S., STUMPF, M. P. H. and THORNE, T. (2012). Considerate approaches to constructing summary statistics for ABC model selection. *Stat. Comput.* **22** 1181–1197. MR2992293

BEAUMONT, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics* **164** 1139–1160.

BEAUMONT, M. A. (2010). Approximate Bayesian Computation in evolution and ecology. *Annu. Rev. Ecol. Evol. Syst.* **41** 379–406.

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

BEAUMONT, M. A., CORNUET, J.-M., MARIN, J.-M. and ROBERT, C. P. (2009). Adaptive approximate Bayesian computation. *Biometrika* **96** 983–990. MR2767283

BLUM, M. G. B. and FRANÇOIS, O. (2010). Non-linear regression models for approximate Bayesian computation. *Stat. Comput.* **20** 63–73. MR2578077

BORNN, L., PILLAI, N. S., SMITH, A. and WOODARD, D. (2017). The use of a single pseudo-sample in approximate Bayesian computation. *Stat. Comput.* **27** 583–590. MR3613586

BORTOT, P., COLES, S. G. and SISSON, S. A. (2007). Inference for stereological extremes. *J. Amer. Statist. Assoc.* **102** 84–92. MR2345549

BROOKS POLLOCK, E., ROBERTS, G. O. and KEELING, M. J. (2014). A dynamic model of bovine tuberculosis spread and control in Great Britain. *Nature* **511** 228–231.

CAMERON, E., BATTLE, K. E., BHATT, S., WEISS, D. J., BISANZIO, D., MAPPIN, B., DALRYMPLE, U., HAY, S. I., SMITH, D. L., GRIFFIN, J. T., WENGER, E. A., ECKHOFF, P. A., SMITH, T. A., PENNY, M. A. and GETHING, P. W. (2015). Defining the relationship between infection prevalence and clinical incidence of Plasmodium falciparum malaria. *Nat. Commun.* **6** 8170.

CONLAN, A. J. K., MCKINLEY, T. J., KAROLEMEAS, K., POLLOCK, E. B., GOODCHILD, A. V., MITCHELL, A. P., BIRCH, C. P. D., CLIFTON-HADLEY, R. S. and WOOD, J. L. N. (2012). Estimating the hidden burden of bovine tuberculosis in Great Britain. *PLoS Comput. Biol.* **8** e1002730. MR3005930

CRAIG, P. S., GOLDSTEIN, M., SEHEULT, A. H. and SMITH, J. A. (1997). Pressure matching for hydrocarbon reservoirs: A case study in the use of Bayes linear strategies for large computer experiments. In *Case Studies in Bayesian Statistics*. 37–93. Springer.

CSILLÉRY, K., BLUM, M. G. B., GAGGIOTTI, O. E. and FRANÇOIS, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends Ecol. Evol.* **25** 410–418.

DEL MORAL, P., DOUCET, A. and JASRA, A. (2012). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Stat. Comput.* **22** 1009–1020. MR2950081

DIGGLE, P. J. and GRATTON, R. J. (1984). Monte Carlo methods of inference for implicit statistical models. *J. Roy. Statist. Soc. Ser. B* **46** 193–227. MR0781880

DOUCET, A., PITT, M. K., DELIGIANNIDIS, G. and KOHN, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika* **102** 295–313. MR3371005

DROVANDI, C. C. and PETTITT, A. N. (2011). Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics* **67** 225–233. MR2898834

DROVANDI, C. C., PETTITT, A. N. and FADDY, M. J. (2011). Approximate Bayesian computation using indirect inference. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **60** 317–337. MR2767849

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

DROVANDI, C. C., PETTITT, A. N. and MCCUTCHAN, R. A. (2016). Exact and approximate Bayesian inference for low integer-valued time series models with intractable likelihoods. *Bayesian Anal.* **11** 325–352. MR3471993

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

FILIPPI, S., BARNES, C. P., CORNEBISE, J. and STUMPF, M. P. H. (2013). On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. *Stat. Appl. Genet. Mol. Biol.* **12** 87–107. MR3044402

GIBSON, G. J. and RENSHAW, E. (1998). Estimating parameters in stochastic compartmental models using Markov chain methods. *IMA J. Math. Appl. Med. Biol.* **15** 19–40.

GOLDSTEIN, M. and ROUGIER, J. (2009). Reified Bayesian modelling and inference for physical systems. *J. Statist. Plann. Inference* **139** 1221–1239. MR2479863

GOLDSTEIN, M., SEHEULT, A. and VERNON, I. (2013). *Assessing Model Adequacy*, 2nd ed. Wiley, UK.

GOURIÉROUX, C., MONFORT, A. and RENAULT, E. (1993). Indirect inference. *J. Appl. Econometrics* **8** S85–S118.

HENDERSON, D. A., BOYS, R. J., KRISHNAN, K. J., LAWLESS, C. and WILKINSON, D. J. (2009). Bayesian emulation and calibration of a stochastic computer model of mitochondrial DNA deletions in substantia nigra neurons. *J. Amer. Statist. Assoc.* **104** 76–87. MR2663034

HOLDEN, P. B., EDWARDS, N. R., HENSMAN, J. and WILKINSON, R. D. (2016). ABC for climate: Dealing with expensive simulators. Handbook of Approximate Bayesian Computation (ABC). Available at 1511.03475.

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

IONIDES, E. L., NGUYEN, D., ATCHADÉ, Y., STOEV, S. and KING, A. A. (2015). Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. *Proc. Natl. Acad. Sci. USA* **112** 719–724. MR3311541

JABOT, F., LAGARRIGUES, G., COURBAUD, B. and DUMOULIN, N. (2014). A comparison of emulation methods for Approximate Bayesian Computation. Available at http://arxiv.org/abs/1412.7560.

JANDAROV, R., HARAN, M., BJØRNSTAD, O. and GRENFELL, B. (2014). Emulating a gravity model to infer the spatiotemporal dynamics of an infectious disease. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **63** 423–444. MR3238160

JEWELL, C. P., KYPRAIOS, T., CHRISTLEY, R. M. and ROBERTS, G. O. (2009). A novel approach to real-time risk prediction for emerging infectious diseases: A case study in avian influenza H5N1. *Prev. Vet. Med.* **91** 19–28.

JOYCE, P. and MARJORAM, P. (2008). Approximately sufficient statistics and Bayesian computation. *Stat. Appl. Genet. Mol. Biol.* **7**. MR2438407

KYPRAIOS, T., NEAL, P. and PRANGLE, D. (2017). A tutorial introduction to Bayesian inference for stochastic epidemic models using approximate Bayesian computation. *Math. Biosci.* **287** 42–53. MR3634152

LENORMAND, M., JABOT, F. and DEFFUANT, G. (2013). Adaptive approximate Bayesian computation for complex models. *Comput. Statist.* **28** 2777–2796. MR3141363

MARIN, J.-M., PUDLO, P., ROBERT, C. P. and RYDER, R. J. (2012). Approximate Bayesian computational methods. *Stat. Comput.* **22** 1167–1180. MR2992292

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

MCCREESH, N., ANDRIANAKIS, I., NSUBUGA, R. N., STRONG, M., VERNON, I., MCKINLEY, T. J., OAKLEY, J. E., GOLDSTEIN, M., HAYES, R. and WHITE, R. G. (2017). Universal, test, treat, and keep: Improving ART retention is key in cost-effective HIV care and control in Uganda. *BMC Infect. Dis.*. To appear.

MCKINLEY, T., COOK, A. R. and DEARDON, R. (2009). Inference in epidemic models without likelihoods. *Int. J. Biostat.* **5**. MR2533810

MCKINLEY, T. J., ROSS, J. V., DEARDON, R. and COOK, A. R. (2014). Simulation-based Bayesian inference for epidemic models. *Comput. Statist. Data Anal.* **71** 434–447. MR3131981

MCKINLEY, T. J, VERNON, I., ANDRIANAKIS, I., MCCREESH, N., OAKLEY, J. E., NSUBUGA, R. N., GOLDSTEIN, M. and WHITE, R. G. (2017). Supplement to "Approximate Bayesian computation and simulation-based inference for complex stochastic epidemic models." DOI:10.1214/17-STS618SUPPA, DOI:10.1214/17-STS618SUPPB.

MEEDS, E. and WELLING, M. (2014). GPS-ABC: Gaussian process surrogate Approximate Bayesian Computation. Available at http://arxiv.org/abs/1401.2838v1.

NEAL, P. (2012). Efficient likelihood-free Bayesian computation for household epidemics. *Stat. Comput.* **22** 1239–1256. MR2992297

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

O'NEILL, P. D. and ROBERTS, G. O. (1999). Bayesian inference for partially observed stochastic epidemics. *J. R. Stat. Soc., A* **162** 121–129.

O'NEILL, P. D., BALDING, D. J., BECKER, N. G., EEROLA, M. and MOLLISON, D. (2000). Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. C* **49** 517–542. MR1824557

OAKLEY, J. E. and YOUNGMAN, B. D. (2017). Calibration of stochastic computer simulators using likelihood emulation. *Technometrics* **59** 80–92. MR3604191

PITT, M. K., SILVA, R. D. S., GIORDANI, P. and KOHN, R. (2012). On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *J. Econometrics* **171** 134–151. MR2991856

PUKELSHEIM, F. (1994). The three sigma rule. *Amer. Statist.* **48** 88–91. MR1292524

RATMANN, O., JØRGENSEN, O., HINKLEY, T., STUMPF, M., RICHARDSON, S. and WIUF, C. (2007). Using likelihood-free inference to compare evolutionary dynamics of the protein networks of *H. pylori* and *P. falciparum*. *PLoS Comput. Biol.* **3** 2266–2278. MR2369270

RATMANN, O., ANDRIEU, C., WIUF, C. and RICHARDSON, S. (2009). Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proc. Natl. Acad. Sci. USA* **106** 10576–10581.

RATMANN, O., CAMACHO, A., MEIJER, A. and DONKER, G. (2014). Statistical modelling of summary values leads to accurate Approximate Bayesian Computations. Available at arXiv:1305.4283v2.

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172. MR0760681

SACKS, J., WELCH, W. J., MITCHELL, T. J. and WYNN, H. P. (1989). Design and analysis of computer experiments. *Statist. Sci.* **4** 409–435. MR1041765

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

SILK, D., FILIPPI, S. and STUMPF, M. P. H. (2012). Optimizing threshold-schedules for approximate Bayesian computation sequential Monte Carlo samplers: applications to molecular systems. Available at arXiv:1210.3296v1.

SISSON, S. A., FAN, Y. and TANAKA, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **104** 1760–1765. MR2301870

TAVARÉ, S., BALDING, D. J., GRIFFITHS, R. C. and DONNELLY, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics* **145** 505–518.

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

VERNON, I., GOLDSTEIN, M. and BOWER, R. G. (2010). Galaxy formation: A Bayesian uncertainty analysis. *Bayesian Anal.* **5** 619–669. MR2740148

VERNON, I., GOLDSTEIN, M. and BOWER, R. (2014). Galaxy formation: Bayesian history matching for the observable universe. *Statist. Sci.* **29** 81–90. MR3201849

WILKINSON, R. D. (2013). Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Stat. Appl. Genet. Mol. Biol.* **12** 129–141. MR3071024

WILKINSON, R. D. (2014). Accelerating ABC methods using Gaussian processes. In *Proceedings of the* 17*th International Conference on Artificial Intelligence and Statistics* (*AISTATS*) **33** 1015–1023.

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