

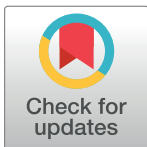
RESEARCH ARTICLE

An approximate diffusion process for environmental stochasticity in infectious disease transmission modelling

Sanmitra Ghosh^{1*}, Paul J. Birrell^{1,2}, Daniela De Angelis^{1,2}

1 MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom, **2** UK Health Security Agency, London, United Kingdom

* sanmitra.ghosh@mrc-bsu.cam.ac.uk



Abstract

Modelling the transmission dynamics of an infectious disease is a complex task. Not only it is difficult to accurately model the inherent non-stationarity and heterogeneity of transmission, but it is nearly impossible to describe, mechanistically, changes in extrinsic environmental factors including public behaviour and seasonal fluctuations. An elegant approach to capturing environmental stochasticity is to model the force of infection as a stochastic process. However, inference in this context requires solving a computationally expensive “missing data” problem, using data-augmentation techniques. We propose to model the time-varying transmission-potential as an approximate diffusion process using a path-wise series expansion of Brownian motion. This approximation replaces the “missing data” imputation step with the inference of the expansion coefficients: a simpler and computationally cheaper task. We illustrate the merit of this approach through three examples: modelling influenza using a canonical SIR model, capturing seasonality using a SIRS model, and the modelling of COVID-19 pandemic using a multi-type SEIR model.

OPEN ACCESS

Citation: Ghosh S, Birrell PJ, De Angelis D (2023) An approximate diffusion process for environmental stochasticity in infectious disease transmission modelling. *PLoS Comput Biol* 19(5): e1011088. <https://doi.org/10.1371/journal.pcbi.1011088>

Editor: Alex Perkins, University of Notre Dame, UNITED STATES

Received: June 6, 2022

Accepted: April 10, 2023

Published: May 18, 2023

Copyright: © 2023 Ghosh et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Code and data (excluding those which were used for modelling COVID-19) supporting the experiments are available on a GitHub repository at <https://github.com/sg5g10/envstoch>. We have also used Zenodo to assign a DOI to the repository: 10.5281/zenodo.7886447. The data used for modelling COVID-19 is owned and handled by the UK Health Security Agency (UKHSA) and is not being made available to the public. The authors accessed the data through data sharing agreement. Anyone interested in accessing the data needs to follow the

Author summary

Over the course of an epidemic, the ability to transmit infection is a key parameter in the epidemic models that seek to quantify and forecast epidemic burden. This parameter will evolve considerably over time due to the emergence of new variants, changing governmental advice and access to testing and possible immunisation. Ignoring these factors can lead to models that produce overly confident estimates and biased predictions. In this paper we propose a computationally efficient method to quantifying this uncertainty in such a way that makes no assumptions about the timing and the magnitude of the impacts of these external stimuli influencing transmission. Our approach relies on the application of an approximate diffusion process to characterise changes in transmission over time. This results in a sizeable computational advantage over previous methods that use a true diffusion process, while retaining a similar quality of estimation. Our methodology would be very valuable to modellers attempting to monitor epidemics in real time.

same procedure by contacting the UKSA Office for Data Release (ODR): <https://www.gov.uk/government/publications/accessing-ukhsa-protected-data>.

Funding: The authors were supported by the Medical Research Council (Unit programme number MC UU 00002/11). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

This is a *PLOS Computational Biology* Methods paper.

1 Introduction

Mathematical modelling of the complex dynamics of infectious diseases remains an essential tool to inform public health policies during epidemic outbreaks. The major focus of such modelling work is describing the intrinsic transmission dynamics and the flow of individuals between compartments that segregate the population as per their disease state. However, an epidemic is also driven by a number of extrinsic factors, including population mobility, social cycles (e.g. holidays), non-pharmaceutical interventions, and climatic variations [1]. In a compartmental model, such factors are often introduced explicitly through the description of the hazard (force) of infection when information about these external drivers is available [2–4]. However, while it is impossible to fully account for all extrinsic factors influencing transmission, yet ignoring this epistemic uncertainty often known as “environmental stochasticity” leads to a structural miss-specification of the model, a “model discrepancy”. Model discrepancy can lead to miss-calibrated models that underestimate uncertainty and produce biased predictions [5]. An elegant approach to account for the un-modelled model discrepancy is to represent the force of infection as a stochastic process. For example, [6, 7] use a diffusion process for this purpose, while [8] use a discrete time stochastic process. Parameter estimation for such stochastic models is, however, challenging. Inference, particularly in a Bayesian context, requires estimation of the joint posterior distribution of both the latent path of the stochastic process and the model parameters. Estimation using a Markov chain Monte Carlo (MCMC) algorithm, involves sampling the realisation of the stochastic process, a high dimensional object, often through *data-augmentation* techniques, which incur a hefty computational cost [9]. As a result efficient calibration of a compartmental model, which embeds a stochastic process, has received significant attention in the literature [10, 11] with the goal of alleviating the computational bottleneck associated with the inference of the stochastic process.

In this paper we propose a new approach to the calibration problem through the use of a path-wise approximation of a diffusion process. Specifically, we apply a truncated Fourier expansion of a Brownian motion to obtain the approximation. Application of this series expansion turns the task of inferring a high dimensional latent diffusion sample path into the task of inferring a smaller dimensional object, the expansion coefficients, which can be carried out without data-augmentation. This method is also applicable in the context of discrete time processes that converge to a diffusion in the continuous time limit. Such processes can be approximated by first carrying out the series expansion of the limiting diffusion and then applying a suitable time discretisation.

We validate the proposed method, firstly, against a data augmentation technique carried out using a particle MCMC sampler proposed in [6], using a dataset from an influenza outbreak in a boarding school. We then further validate this approach using a model where the force of infection is depicted as a sinusoidal signal. Finally, we apply this method to fit a model of COVID-19 spread in England during the first wave.

2 Background: Epidemic models with a time-varying transmission-potential

We consider the canonical SIR (Susceptible-Infected-Removed) model [12] to introduce the stochastic modelling framework, although the methodology can be applied to other more

complex compartmental models. In the SIR model the compartments denote the number of susceptible (S), infected (I), and recovered (R) people in a population subjected to an epidemic at time t . For a population of size N , the SIR model is defined by the following ODE system:

$$\frac{dS_t}{dt} = -\beta S_t \frac{I_t}{N}, \quad \frac{dI_t}{dt} = \beta S_t \frac{I_t}{N} - \gamma I_t, \quad \frac{dR_t}{dt} = \gamma I_t, \quad (1)$$

where $\lambda = \beta \frac{I_t}{N}$ is the force of infection, describing the generation of infections with a transmission-potential β , between susceptible individuals and the fraction, I_t/N , of infectious individuals. The expected period spent in the compartment is given by γ^{-1} . The individual compartment sizes sum to $N = S_t + I_t + R_t$.

To include environmental stochasticity we introduce a time-varying β_t [13–15] to mitigate model discrepancy, leading to a reformulation of the model in Eq (1):

$$\begin{aligned} dx_t &= a(x_t, \xi_a)dt + b(x_t, \xi_b)dW_t \\ \beta_t &= g(x_t) \\ \frac{dS_t}{dt} &= -\beta_t S_t \frac{I_t}{N}, \quad \frac{dI_t}{dt} = \beta_t S_t \frac{I_t}{N} - \gamma I_t, \quad \frac{dR_t}{dt} = \gamma I_t, \end{aligned} \quad (2)$$

where x_t follows a diffusion process described by an Itô stochastic differential equation (SDE) [16] with drift $a(\cdot)$, and diffusion $b(\cdot)$ functions parameterised by the vectors ξ_a and ξ_b respectively; W_t is a standard Brownian motion; and $g(\cdot)$ is a nonlinear transformation that enforces $\beta_t > 0$, such as exponential or inverse-logit transformation. Here we make some mild assumptions about $a(\cdot)$ and $b(\cdot)$ such as, for example, being locally Lipschitz with a linear growth bound [16] to ensure a non-explosive solution.

Inference for the stochastic model in Eq (2) within a Bayesian framework, requires inference of the latent sample path x of the diffusion x_t , which is indirectly observed through the time evolution of the disease states: S_t , I_t , R_t . This is a missing data problem that can be addressed through data-augmentation based MCMC methods [6, 10] in which a high resolution (in time) Euler-Maruyama discretisation of x_t is sampled along with the model parameters. Such MCMC methods incur high computational costs and have reduced efficiency in terms of mixing and speed of convergence. In what follows we will investigate a scalable approximation of x_t that is faster to sample.

3 Methods

Following [17–19], we carry out a Fourier expansion of a Brownian motion W_t and obtain a smooth path-wise series approximation. Using this approximation of a Brownian motion, we can in turn approximate the SDE for x_t with a random ODE. Inference of x_t can then be carried out by inferring coefficients of this ODE, without requiring data-augmentation.

3.1 Fourier expansion of Brownian motion

Within a time interval $[0, T]$, where T is the length of the time horizon within which an epidemic is analysed, the Fourier expansion of a Brownian motion W_t is given by [18]:

$$W_t = \sum_{i=1}^{\infty} \left(\int_0^T \phi_i(s) dW_s \right) \int_0^t \phi_i(u) du. \quad (3)$$

where $\{\phi_i\}_{i=1}^{\infty}$ is a complete orthonormal basis of $L^2[0, T]$ (see Appendix A in S1 Text for

derivation). For example this can be the generalised Fourier cosine basis [20] given by

$$\phi_i(t) = (2/T)^{1/2} \cos\{(2i-1)\pi t/2T\}. \quad (4)$$

We will use the shorthand $Z_i = \int_0^T \phi_i(s) dW_s$. Since the basis functions $\{\phi_i\}$ are deterministic and orthonormal, it follows from standard results of Itô calculus that $Z_i \sim \mathcal{N}(0, 1)$ [18]. By truncating the infinite series in Eq (3) to n -terms we obtain a path-wise approximation of the Brownian motion W_t given by

$$\hat{W}_t = \sum_{i=1}^n Z_i \int_0^t \phi_i(u) du. \quad (5)$$

Following [17], if we order the basis functions by the number of times they change sign on the interval $[0, T]$, then when i is small Z_i governs the low-frequency oscillations of the Brownian motion.

3.2 Approximating a SDE with a random ODE

Taking derivative of $\hat{W}_t$ with respect to time we obtain the following approximation to white noise, the derivative of Brownian motion, given by

$$\frac{d\hat{W}_t}{dt} = \sum_{i=1}^n Z_i \phi_i(t). \quad (6)$$

Now, let us replace the Itô SDE in Eq (2) with the following Stratonovich SDE [16]

$$dx_t = a'(x_t, \xi_a)dt + b(x_t, \xi_b) \circ dW_t, \quad (7)$$

where $(\circ)$ denotes a Stratonovich integral [16] with respect to W_t . The Itô SDE in Eq (2) and the Stratonovich SDE given above are equivalent [16] if

$$a'(x_t, \xi_a) = a(x_t, \xi_a) - \frac{b(x_t, \xi_b)}{2} \frac{\partial b(x_t, \xi_b)}{\partial x_t} b(x_t, \xi_b). \quad (8)$$

By substituting the term dW_t in Eq (7) with the approximation $d\hat{W}_t$ in Eq (6), we obtain the following (random) ODE:

$$\frac{d\hat{x}_t}{dt} = a'(\hat{x}_t, \xi_a) + b(\hat{x}_t, \xi_b) \sum_{i=1}^n Z_i \phi_i(t). \quad (9)$$

The work of [21] shows that as $n \rightarrow \infty$ the solution $\hat{x}_t$ of the above ODE will converge to the solution x_t of the Stratonovich SDE in Eq (7) which, given the choice of $a'(\cdot)$ in Eq (8), is an equivalent representation of the Itô SDE in Eq (2). Thus, the series approximation $\hat{x}_t$ of the solution x_t of an Itô SDE converges to the solution of an equivalent Stratonovich SDE.

Next, we discuss the implications of the above approximation with regards to inference.

3.3 Inference using the series approximation

Using the path-wise series approximation of a diffusion process x_t , presented in the previous sections, we can re-write the canonical SIR model in Eq (2) as a system of coupled ODEs given

by

$$\begin{aligned}\frac{d\hat{x}_t}{dt} &= a'(\hat{x}_t, \xi_a) + b(\hat{x}_t, \xi_b) \sum_{i=1}^n Z_i \phi_i(t) \\ \beta_t &= g(\hat{x}_t) \\ \frac{dS_t}{dt} &= -\beta_t S_t \frac{I_t}{N}, \quad \frac{dI_t}{dt} = \beta_t S_t \frac{I_t}{N} - \gamma I_t, \quad \frac{dR_t}{dt} = \gamma I_t,\end{aligned}\tag{10}$$

where $a'(\cdot)$ is given by Eq (8). Note that the randomness in the above model is now encapsulated in the expansion coefficients $\mathbf{Z} = (Z_1, \dots, Z_n)$. Inference in this model is then relegated to the inference of all the parameters: $\mathbf{Z}$, ξ_a , ξ_b , γ , and the initial values: x_0, S_0, I_0, R_0 . We denote the vector of the parameters governing the dynamics as $\boldsymbol{\theta} = (\xi_a, \xi_b, \gamma)$. We denote the state vector evolving in continuous time by $\mathbf{X}_t = (x_t, S_t, I_t, R_t)$, and by $\mathbf{X}_0 = (x_0, S_0, I_0, R_0)$ the vector of initial values.

In order to explain the inferential framework based on the series approximation, in Eq (10), we assume that the available data $\mathbf{y}_{t_{1:m}} = (y_{t_1}, \dots, y_{t_m})$ are the noisy observations of the state I_t at m time-points. Here we are simply considering prevalence data for the ease of exposition, however the same idea can be extended to more complex observational models where the observed data only provide partial (and often indirect) information of the states $\mathbf{X}_t$ [8].

The inferential goal is to learn the posterior distribution of all the unknown quantities, given the data $\mathbf{y}_{t_{1:m}}$. We place priors $p(\boldsymbol{\theta})$, $p(\mathbf{Z})$, $p(\mathbf{X}_0)$ on the parameters, expansion coefficients and the initial values. Note that, by construction, the $\mathbf{Z} = (Z_1, \dots, Z_n)$ have an independent standard Normal prior, see Section 3.1. We then numerically solve Eq (10) to obtain a likelihood $p(\mathbf{y}_{t_{1:m}} | \mathbf{I}_{t_{1:m}}, \epsilon)$, based on the noise assumption, where $\mathbf{I}_{t_{1:m}}$ is the numerical solution of the state I_t evaluated at the m time-points, and ϵ are the parameters of the chosen data distribution. The posterior distribution, up to a normalisation constant, follows from the Bayes rule:

$$p(\boldsymbol{\theta}, \mathbf{Z}, \mathbf{X}_0, \epsilon | \mathbf{y}_{t_{1:m}}) \propto p(\mathbf{y}_{t_{1:m}} | \mathbf{I}_{t_{1:m}}, \epsilon) p(\boldsymbol{\theta}) p(\mathbf{Z}) p(\mathbf{X}_0) p(\epsilon).\tag{11}$$

Samples from the posterior distribution can be obtained using MCMC. The samples of the latent approximate diffusion path $\hat{\mathbf{x}}$ are simply the numerical solution of the ODE for $\hat{x}_t$ evaluated using samples of $\boldsymbol{\theta}$, $\mathbf{Z}$, $\mathbf{X}_0$ from the posterior distribution.

Note that if we had described x_t using a SDE, as in Eq (2), then we had to infer the sample path of the diffusion along with the parameters. For this we had to discretise the path of x_t using, say, the Euler-Maruyama scheme [22]. Specifically, we had to introduce l points between each pair of successive observation times t_i and t_{i+1} ($i = 0, 1, \dots, m-1$) to allow sufficient accuracy of the Euler-Maruyama scheme. Thus, for a step of $\delta = 1/(l+1)$ we would have ended-up with a discrete skeleton $\mathbf{x} = (x_0, x_\delta, x_{2\delta}, \dots, x_{t_m})$, representing the sample path. This sample path $\mathbf{x}$ along with the parameters $\boldsymbol{\theta}$ can then be estimated using a data-augmentation based MCMC algorithm, where the path $\mathbf{x} \sim p(\mathbf{x} | \mathbf{y}, \boldsymbol{\theta})$ is preferably drawn using a particle filter as proposed in [6]. If we now contrast this inference scheme, where x_t is described as a SDE, to the case where x_t is described by an ODE (after applying the series expansion), it can be seen that as long as $n = \dim(\mathbf{Z})$ is chosen to be substantially smaller than $\dim(\mathbf{x})$, inference using the series expansion will be a simpler problem as we have to deal with the estimation of fewer random variables. Below we show, empirically, that choosing a value of n substantially smaller than $\dim(\mathbf{x})$ still renders an estimate of the posterior distribution that is a reliable approximation to the true posterior.

4 Evaluation

To evaluate the proposed approximation method we fit the model in Eq (10) to the data of an outbreak of influenza at a boarding school [23], on the number of infections for a period of $T = 14$ days among a population of size $N = 763$. This dataset is publicly available in the R package `outbreaks` [24]. This dataset was previously used in [25, 26] to fit a SIR model with a time-varying transmission-potential β_t under the assumption that β_t evolves as a stationary stochastic process, whose dynamics can be modelled using an Ornstein–Uhlenbeck (OU) SDE (since the solution of this SDE evolves as a stationary process). If we describe the dynamics of x_t in Eq (2), through the OU SDE

$$dx_t = (\xi_1 - \xi_2 x_t)dt + \xi_3 dW_t, \quad (12)$$

where $(\xi_1 - \xi_2 x_t)$ is the linear drift function and ξ_3 is a standard deviation parameter, commonly referred to as the *volatility* parameter, representing a constant diffusion function, then the stochastic model in Eq (2) becomes similar to the model in [25]. We can re-write the model in Eq (2) as follows:

$$\begin{aligned} dx_t &= (\xi_1 - \xi_2 x_t)dt + \xi_3 dW_t \\ \beta_t &= \exp(x_t) \\ \frac{dS_t}{dt} &= -\beta_t S_t \frac{I_t}{N}, \quad \frac{dI_t}{dt} = \beta_t S_t \frac{I_t}{N} - \gamma I_t, \quad \frac{dR_t}{dt} = \gamma I_t, \end{aligned} \quad (13)$$

where, following [25] and [26], the logarithm of the transmission-potential evolves as a stationary stochastic process.

Here we specifically want to compare the outcome of inference using the true OU diffusion used above (SDE) with its series approximation (SA), leading to a model such as in Eq (10), where the dynamics of $\hat{x}_t$ is given by the following ODE:

$$\frac{d\hat{x}_t}{dt} = (\xi_1 - \xi_2 \hat{x}_t) + \xi_3 \sum_{i=1}^n Z_i \phi_i(t), \quad (14)$$

where we have chosen the generalised Fourier basis in Eq (4) as the function $\phi_i(t)$.

For the SDE model the latent sample path $\mathbf{x}$, the OU SDE parameters ξ , initial value x_0 and the parameter γ were also estimated together with the initial susceptibility, $s_0 = S(t=0)/N$, assuming the initial recovered fraction $r_0 = 0$ and thus $i_0 = 1 - s_0$. As this is count data we have specified a Poisson likelihood:

$$y_{t_i} | \boldsymbol{\theta}, \mathbf{x}, \mathbf{X}_0 \sim \text{Poisson}(I_{t_i}), \quad i = 1, \dots, m, \quad (15)$$

where in this case $\mathbf{X}_0 = (x_0, s_0)$ and $\boldsymbol{\theta} = (\xi, \gamma)$. For the SA model we used the inferential framework introduced in the previous section and used the Poisson likelihood as above:

$$y_{t_i} | \boldsymbol{\theta}, \mathbf{Z}, \mathbf{X}_0 \sim \text{Poisson}(I_{t_i}). \quad (16)$$

We chose a weakly-informative prior for the parameters governing the dynamics $\xi_1, \dots, \xi_3$, $\gamma \sim \Gamma(2, 2)$. For s_0 a Beta(2, 1), since we expect the true value to be near or greater than 2/3, and for the initial value of the diffusion we used a prior $x_0 \sim \mathcal{N}\left(\xi_1/\xi_2, \left(\frac{\xi_3^2}{2\xi_2}\right)^2\right)$, which is the stationary distribution of the OU diffusion.

For the SDE model, data-augmentation using a particle filter was employed to sample the ‘true’ diffusion’s path, following [6], and produce an unbiased estimate of the likelihood.

Parameters γ , ξ , X_0 were estimated using the Metropolis-Hastings (MH) algorithm, with an adaptive random-walk proposal based on algorithm 4 of [27]. See Appendix B in S1 Text. for further details on this proposal mechanism. The likelihood estimate produced by the particle filter was used in the acceptance step of the MH algorithm. This particle-marginal Metropolis-Hastings (PMMH) MCMC scheme for jointly updating the latent diffusion path along with the parameters has been shown to have superior performance [6] when compared to other data-augmentation approaches. For the PMMH, we used a Bootstrap particle filter [28], where the particles are propagated using Euler-Maruyama discretisation, and set the number of particles to 1000. Following [25], we carried out the Euler-Maruyama iterations with a stepsize $\delta t = 0.1$, leading to $l = 9$ time-points between two observations.

For the SA model we used the Metropolis-Hastings algorithm with the same adaptive random-walk proposal (RWMH) used with the PMMH scheme and the Euler method to numerically solve the ODE adopting the same step-size that is used with the Euler-Maruyama scheme for the SDE.

Note that inference for the SDE model using PMMH will be substantially more computationally heavy compared to the inference for the ODE based SA model, irrespective of the value of n . This is due to the particle filter requiring multiple evaluation of the Euler-Maruyama scheme at each MCMC iteration. Even when parallelised, the particle filter will be bottlenecked by a weight-updating step (see [28] for details) requiring message-passing across processes. The Euler scheme for solving the ODE in Eq (14), in comparison, is evaluated once every iteration of a Metropolis-Hastings algorithm targeting the posterior distribution in Eq (11).

A crucial parameter for the proposed method is the number of basis functions n . If a value of n produces a close match between the marginal densities of the true and approximate diffusion at the end of the analysis period T then the approximation will be valid throughout the course of the epidemic. In this case $T = 14$. In Fig 1 we compare the time T marginal densities $p(\hat{x}_t)|_{t=T}$ obtained by solving the ODE in Eq (14) associated with the SA, and $p(x_t)|_{t=T}$ obtained

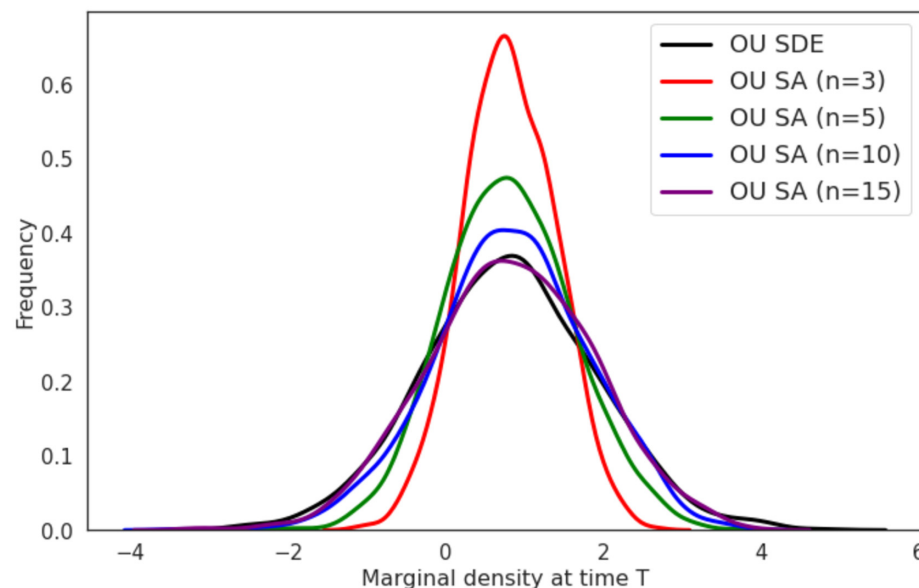


Fig 1. Comparison between the marginal density of the OU SDE at time $T = 14$, with that obtained through the series approximation upon varying the number of basis $n = 3, 5, 10, 15$.

<https://doi.org/10.1371/journal.pcbi.1011088.g001>

from the original OU diffusion, both based on some trial parameters sampled from the prior. The value $n = 15$ produces a close match between the marginal densities. We defer further discussion of the effect of n on estimation to section 4.2.

4.1 Results: Comparison between true and approximate diffusion

We fitted the two models, **SDE** and **SA** respectively, using the associated algorithms as described above to the influenza dataset. We ran two chains of both PMMH, for the **SDE** model, and RWMH, for the **SA** one, for 10^6 iterations where the first 5×10^5 iterations were discarded as burn-in and the remaining samples thinned to obtain 1000 samples from the posterior distribution. The running times were 15907 and 2397 seconds for the PMMH and RWMH with $n = 15$, respectively. We implemented a vectorised particle filter and the Euler solver for the ODE using `Jax` [29]. The adaptive MCMC algorithm was implemented using `Python`.

We notice a good agreement between the parameter estimates obtained using the **SDE** and **SA** counterparts (see Fig 2). Furthermore, in Fig 3 we compare the goodness-of-fit and display the posterior distribution of the latent diffusion paths $p(\mathbf{x}|\mathbf{y}_{t_{1:m}})$ and $p(\hat{\mathbf{x}}|\mathbf{y}_{t_{1:m}})$, corresponding to the **SDE** and **SA**. Additionally, for aid of visualisation, we have also plotted draws from the (posterior) sample paths for both models in Fig 3(c) and 3(d).

We observe a good agreement between the epidemic curves obtained using the **SDE** and the **SA**, but for the posterior distribution of the latent diffusion paths the credible intervals are narrower for the **SA**. The **SA**, due to the truncation of the infinite series expansion, produces smoother paths, slightly underestimating the variability in the evolution of the latent

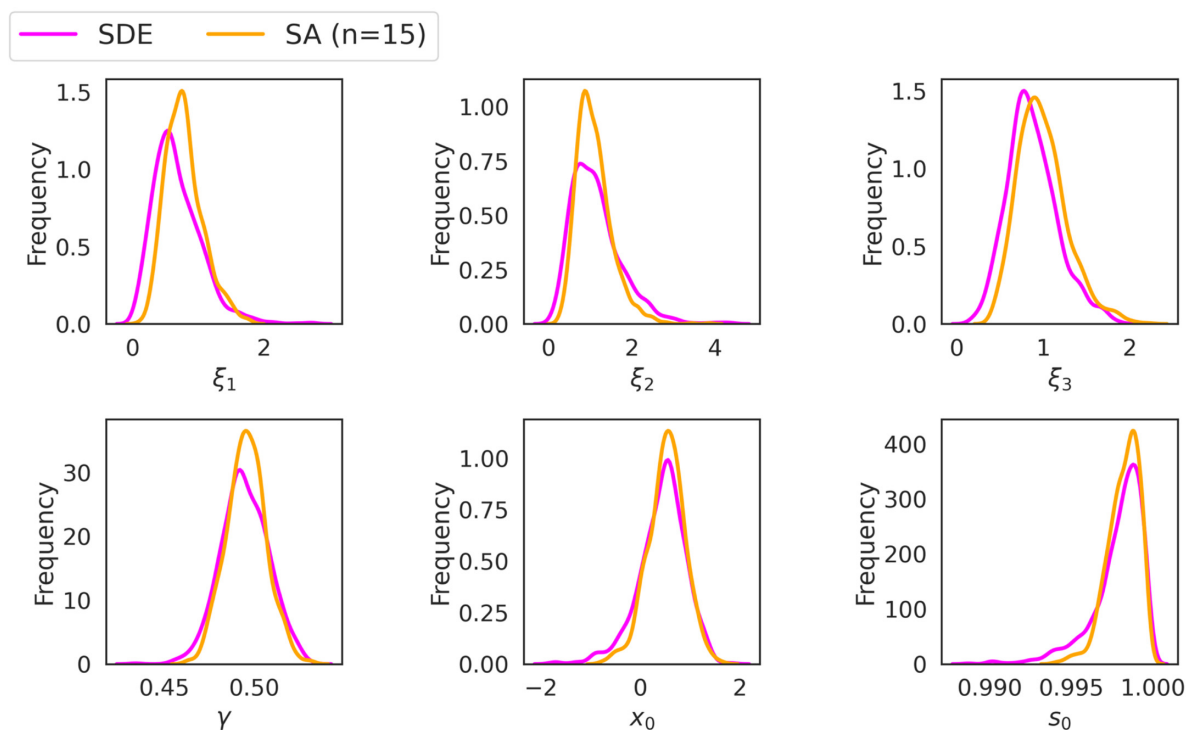


Fig 2. Comparison of the posterior marginal densities of the parameters obtained using the SDE and the SA (with $n = 15$ basis function). These densities are summarised using a kernel density estimate.

<https://doi.org/10.1371/journal.pcbi.1011088.g002>

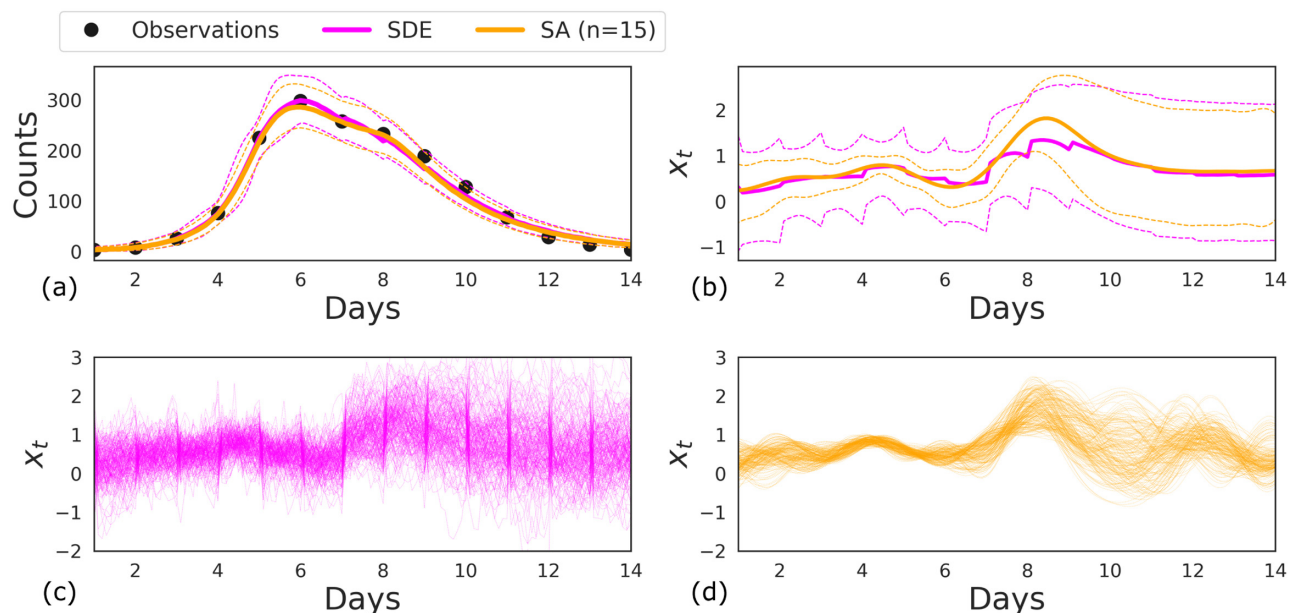


Fig 3. Influenza dataset: Goodness-of-fit (a); posterior distribution of the latent diffusion paths corresponding to the **SDE** and **SA** counterparts (b), with densities summarised by the mean (solid lines) and 95% credible intervals (broken lines); and samples from the posterior distribution of the latent diffusion paths, **SDE** (c) and **SA** (d).

<https://doi.org/10.1371/journal.pcbi.1011088.g003>

diffusion path. On a closer introspection of the posterior means (Fig 3(b)), it is noticeable that the latent diffusion paths drop and increase again in the period between the 4-th and 9-th day, around the peak, indicating an increase in transmissibility. These changes are reflected in the estimates of both **SDE** and **SA**. After the 9-th day, the variability in the latent paths increase for both **SDE** and **SA** and the posterior means match closely. This is expected since after the peak, when the epidemic is receding, a large change in β_t will have negligible effect on the case counts. Note that there is also a considerable amount of uncertainty about the **SA**'s estimate of the maximum value of the transmission-potential around the 9-th day (see the sample paths in Fig 3(d)). That is the **SA** is uncertain whether the maximum value of the transmission-potential occurs on the 8-th or the 9-th day. This increased uncertainty, in turn, results in the credible intervals of **SA** matching the credible intervals of **SDE** after day 9.

These results were confirmed in a simulation study where the simulated datasets mimicked this influenza dataset (see Appendix C in S1 Text).

Additionally, we have fitted the model while using a negative binomial likelihood to allow overdispersion. Details are given in Appendix I in S1 Text.

4.2 The effect of n on the convergence of the posterior

In Fig 1 we noticed that the marginal distribution of the latent diffusion path and its series approximation starts agreeing beyond $n \geq 10$ terms. It is worth investigating whether such a threshold exist for the posterior distributions obtained using the **SDE** and the **SA**. We did this by further comparing the joint posterior distribution $p(\theta, X_0|y)$, from **SDE** and **SA** while varying n . Note that θ and X_0 are quantities which were estimated using both the **SDE** and **SA** counterparts, and thus the joint posterior of these were chosen for comparison. For this comparison we estimated the posterior distribution by fitting the **SA** repeatedly with number

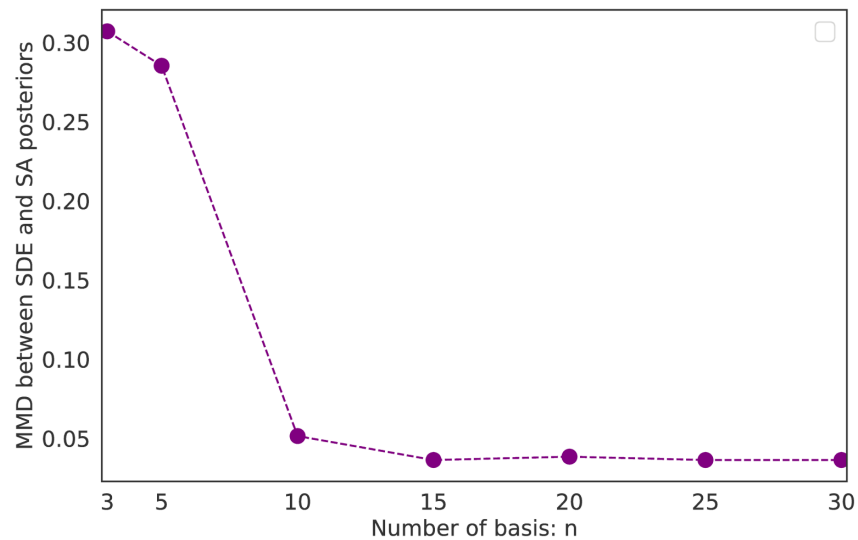


Fig 4. MMD between the joint posterior distributions of the parameters θ and initial values X_0 from SDE and SA (for different n).

<https://doi.org/10.1371/journal.pcbi.1011088.g004>

of basis set to $n = 3, 5, 10, 15, 20, 25, 30$. To compare the posterior distributions, we used the maximum mean discrepancy (MMD) metric [30], a divergence metric that can be calculated using samples from the distributions. See Appendix H in [S1 Text](#) for further details on this metric.

In [Fig 4](#) we plot the MMD between the posteriors from SDE and SA for increasing n . For $n \geq 10$ we found good agreement between the two posteriors, consistent with the results from comparing the marginal densities ([Fig 1](#)). This reinforces our approach of choosing the number of basis by comparing marginals of the latent process, while using the SA. We summarise the runtimes of MCMC with the RWMH proposal for each choice of n in [Table 1](#), noting that the increase in the runtimes as we varied n was negligible, especially when compared to the PMMH with SDE.

5 Approximate diffusion for capturing a complex signal

In the previous section, when fitting the SIR model to an influenza dataset, we considered a stationary stochastic process representing the evolution of the transmission-potential. To further demonstrate the usefulness of the proposed series approximation technique we investigate here the capability of the approximate diffusion process model to recover a complex signal representing the time evolution of the transmission-potential. To this end, we follow closely the experiment furnished in [7], where observations were simulated from a SIRS model whose time-varying transmission potential was given by a deterministic sinusoidal signal. More

Table 1. Runtimes (rounded to nearest integer), in seconds, of MCMC for SA, as a function of the number of basis n , in comparison with the runtime of SDE. These were run on a 3.6 GHz machine with 16 GB memory.

Runtimes in seconds								
SA with	$n = 3$	$n = 5$	$n = 10$	$n = 15$	$n = 20$	$n = 25$	$n = 30$	SDE
	2280	2282	2337	2397	2470	2538	2637	15907

<https://doi.org/10.1371/journal.pcbi.1011088.t001>

specifically, we consider the SIRS model given by [7]

$$\begin{aligned}
 \beta_t &= \beta_0(1 + \beta_1 \sin(2\pi t/365 + 2\pi\beta_2)) \\
 \frac{dS_t}{dt} &= \mu(N - S_t) - \beta_t S_t \frac{I_t}{N} + \alpha R_t \\
 \frac{dI_t}{dt} &= \beta_t S_t \frac{I_t}{N} - (\gamma + \mu)I_t \\
 \frac{dR_t}{dt} &= \gamma I_t - (\alpha + \mu)R_t \\
 \frac{dC_t}{dt} &= \beta_t S_t \frac{I_t}{N},
 \end{aligned} \tag{17}$$

where C_t gives the number of new cases, $1/\alpha$ is the average duration of immunity, γ is the recovery rate and μ is the recruitment or mortality rate. Observation $y_{t_{1:m}}$, from this model, were simulated on a discrete time-grid $t_{1:m} = (t_1, t_2, \dots, t_m)$, with observation interval $\delta t = t_i - t_{i-1}$, ($i = 1, \dots, m$) being 7 days for a total of $m = 156$ weeks constituting an analysis time-span of three years. Again following [7], we assumed a Poisson observation model for the number of new cases. Thus, the generative model is given by

$$y_{t_i} \sim \text{Poisson}(h_{\delta t}(C_t)), \quad i = (1, \dots, m), \tag{18}$$

where $h_{\delta t}(\cdot)$ denotes the cumulative sum of C_t over the observation interval.

The parameters were set to the exact values used in [7] for generating the simulated data. We did the same for the initial values which were chosen in [7] to generate transients realistically reflecting a real application. To fit the data we employed the same SIRS model as in Eq (17), but we assumed the transmission-potential to evolve as an approximate Brownian motion, to capture a non-stationary signal, given by (following Eq (6))

$$\frac{d\beta_t}{dt} = \sigma \sum_{i=1}^n Z_i \phi_i(t), \tag{19}$$

where σ is a volatility parameter and we solve this ODE for β_t along with the compartmental ODEs in Eq (17), using an Euler solver. We chose $n = 20$, through a comparison of the marginal density at $p(\beta_t)|_{t=T}$, where $T = 7 \times 156 = 1092$. We again chose the generalised Fourier basis in Eq (4) as the function $\phi_i(t)$.

The unknown parameters included $\theta = (1/\alpha, 1/\gamma, \sigma)$ (μ was assumed known) and the initial values $X_0 = (S_0, I_0, \beta_0)$. The Poisson observation model in Eq (18) was considered as the likelihood. We retained the same priors as in [7]. The priors and the generative values are given in Appendix J in S1 Text. We ran the adaptive MCMC algorithm, used previously for the influenza dataset, for 5×10^5 iterations, discarded the first half as burn-in, and thinned appropriately to collect 1000 samples representing the posterior distribution. The run-time was 808 seconds, which is orders-of-magnitude faster than what was reported in [7].

In Fig 5 we plot the goodness-of-fit together with the posterior sample path of the transmission-potential. The posterior distribution of the parameters are shown in Appendix J in S1 Text. Note how (Fig 5(b)) the approximate Brownian motion model for β_t can recover well the true underlying signal, which clearly demonstrates that using an approximate diffusion process we can reliably reproduce the time evolution of the transmission-potential.

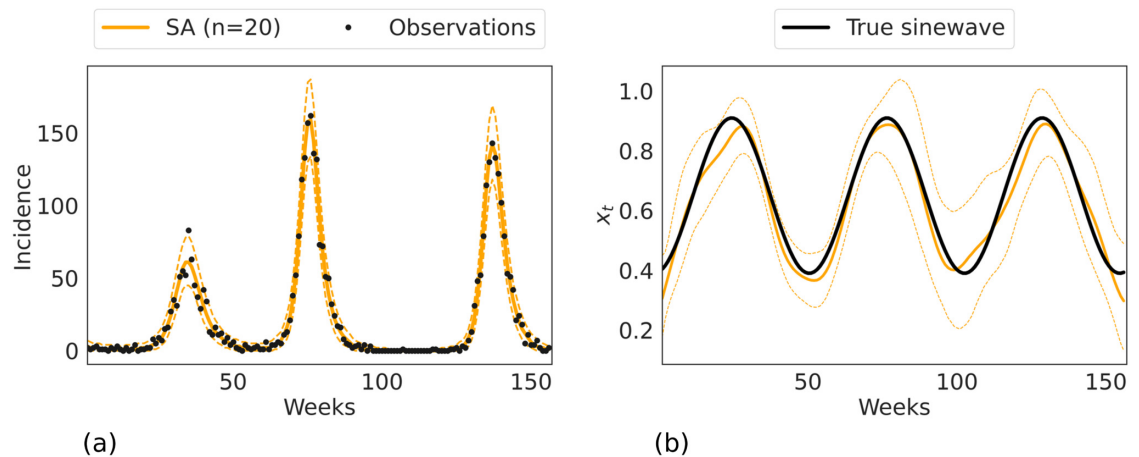


Fig 5. SIRS model with sinusoidal transmission-potential: Goodness-of-fit (a); posterior distribution of the latent diffusion paths corresponding to the SDE and SA approaches (b), with densities summarised by the mean (solid lines) and 95% credible intervals (broken lines).

<https://doi.org/10.1371/journal.pcbi.1011088.g005>

6 Application: Modelling COVID-19 outbreak in England

Our proposed method of modelling the time-varying transmission-potential as an approximate diffusion can also be applied to a discrete time stochastic process that converges to a diffusion in the continuous time limit. For example, an AR(1) process converges to a OU diffusion. Thus, if one is already using an AR(1) process to model the transmission-potential, then a discretised version of the series approximation of OU diffusion, the ODE in Eq (14), can be chosen as its replacement.

To exemplify the application of the series expansion method in replacing a discrete time stochastic process, we have chosen to fit a compartmental model, originally proposed in [8], whose dynamics are described as a set of first order difference equations, to data from the first wave of the COVID-19 outbreak in England, between February and August 2020. This model captures the effect of unknown extrinsic factors on the force of infection through a time-varying transmission-potential modelled as a Gaussian random-walk. We introduce the model of [8] in what follows and introduce an alternative formulation using the series approximation of Brownian motion.

6.1 Transmission model for COVID-19

This is an age and spatially structured transmission model, stratifying the population into $n_A = 7$ age groups and $n_r = 7$ regions. Within each region, the transmission dynamics are governed by a system of first order difference equations:

$$\begin{aligned}
 S_{r,t_k,a} &= S_{r,t_{k-1},a} (1 - \lambda_{r,t_{k-1},a} \delta t) \\
 E_{r,t_k,a}^1 &= E_{r,t_{k-1},a}^1 \left(1 - \frac{2\delta t}{d_L} \right) + S_{r,t_{k-1},a} \lambda_{r,t_{k-1},a} \delta t \\
 E_{r,t_k,a}^2 &= E_{r,t_{k-1},a}^2 \left(1 - \frac{2\delta t}{d_L} \right) + E_{r,t_{k-1},a}^1 \frac{2\delta t}{d_L} \\
 I_{r,t_k,a}^1 &= I_{r,t_{k-1},a}^1 \left(1 - \frac{2\delta t}{d_I} \right) + E_{r,t_{k-1},a}^2 \frac{2\delta t}{d_L} \\
 I_{r,t_k,a}^2 &= I_{r,t_{k-1},a}^2 \left(1 - \frac{2\delta t}{d_I} \right) + I_{r,t_{k-1},a}^1 \frac{2\delta t}{d_I}
 \end{aligned} \tag{20}$$

where: $S_{r,t_k,a}$, $E_{r,t_k,a}^d$, $I_{r,t_k,a}^d$, $d = 1, 2$ represent the time t_k , $k = 1, \dots, K$, partitioning of the population of individuals in a region r , $r = 1, \dots, n_r$, in age-group a , $a = 1, \dots, n_A$, into S (susceptible), E (exposed) and I (infectious) disease states. The average period spent in the exposed and infectious states are given by the parameters d_L and d_I respectively; and $\lambda_{r,t_k,a}$ is the time- and age-varying rate with which susceptible individuals become infected, the force of infection. Time steps of $\delta t = 0.5$ days are chosen to be sufficiently small relative to the latent and infectious periods. Following [31] the initial conditions of the system states S , E^1 , E^2 , I^1 , I^2 at t_0 are given by region-specific parameters ψ_r and $I_{0,r}$, describing the initial exponential growth and the initial number of infectious individuals, respectively. New infections are generated as

$$\Delta_{r,t_k,a}^{\text{infect}} = S_{r,t_k,a} \lambda_{r,t_k,a} \delta t, \quad (21)$$

where $\lambda_{r,t_k,a} \delta t$ is driven over time by a region-specific time-varying transmission potential, $\beta_{t_k,r}$ (see Eq (22)) which moderates the rate at which effective contact take place. This region-specific transmission-potential captures the discrepancy between how actual contact take place between the age groups, and that encoded by a set of time-varying contact matrices. We refer the reader to [8] for further details on the model dynamics and parameterisation.

Over time $\beta_{t_k,r}$ is not allowed to vary unconstrained and a smoothing is imposed by assuming, *a priori* that its evolution follows a Gaussian random-walk process with volatility σ_{β_i} :

$$\begin{aligned} \log(\beta_{t_k,r}) &\sim \mathcal{N}(\log(\beta_{t_{k-1},r}), \sigma_{\beta_i}^2), \quad \text{if } t_k > t_{\text{lock}}, \\ \log(\beta_{t_k,r}) &= 0, \quad \text{if } t_k \leq t_{\text{lock}}, \end{aligned} \quad (22)$$

where t_{lock} indicates the time-point corresponding to the lockdown introduced in England on 23rd March 2020. This random-walk formulation requires the inference of the high-dimensional (due to the choice of δt) sample path of this process, an extremely challenging task using MCMC. We will discuss this inferential difficulty later in Section 6.3.2. To restrict the dimensionality of the process, the model formulation in [8] considered the transmission-potential to be piecewise constant with weekly change-points, and its values at these change-points to evolve following a random-walk. Denote $w_k \equiv w(t_k)$ the week in which time t_k falls. Then the time evolution of the transmission-potential is modelled at a slower weekly time-scale:

$$\begin{aligned} \log(\beta_{w_k,r}) &\sim \mathcal{N}(\log(\beta_{w_{k-1},r}), \sigma_{\beta_w}^2), \quad \text{if } t_k > t_{\text{lock}} + 7/\delta t, \\ \log(\beta_{w_k,r}) &= 0, \quad \text{if } t_k \leq t_{\text{lock}} + 7/\delta t, \\ \beta_{t_k,r} &= \beta_{w_k,r}, \end{aligned} \quad (23)$$

as a Gaussian random-walk, with volatility σ_{β_w} , following the first week of the lockdown. Realisation of the process, for each region, can then be obtained by sampling the vector $\Delta\beta$, of all the weekly increments $\Delta\beta_{w_k,r} = \log(\beta_{w_k,r}) - \log(\beta_{w_{k-1},r})$. It was assumed in [8] that the contact matrices sufficiently described how actual contacts took place between different age groups prior to the lockdown and thus $\beta_{w_k,r} = 1$ over that period. We have retained this assumption in our work.

6.2 Observation model

To fit the model, using a Bayesian framework, surveillance data of age- and region-specific counts of deaths in people with a lab-confirmed COVID-19 diagnosis between 17st February

and 1st August was used. Furthermore, serological data from NHS Blood and Transplant (NHSBT), informing the fraction of the population carrying COVID-19 antibodies, were also used.

Following [8], the observed number of deaths $y_{r,t_k,a}^d$ on day t_k in age group a in region r is the realisation of Negative Binomial distribution:

$$y_{r,t_k,a}^d | d_I, p_a, \psi_r, I_{0,r}, \Delta\beta_r, \eta \sim \text{NegBin}(\mu_{r,t_k,a}, \eta), \quad (24)$$

where $\mathbb{E}y_{r,t_k,a}^d = \mu_{r,t_k,a}$ and $\text{Var}(y_{r,t_k,a}^d) = \mu_{r,t_k,a}(1 + \eta)$. Here η is a dispersion parameter and $\mu_{r,t_k,a} = p_a \sum_{l=0}^k f_{k-l} \Delta_{r,t_l,a}^{\text{infect}}$, with $\Delta_{r,t_l,a}^{\text{infect}}$ obtained by solving the difference equations in Eq (20) and then evaluating Eq (21); f_l giving the probability of the death occurring on the l -th day after infection, assuming the distribution f of the time from infection to death from COVID-19 is known; and p_a representing an age-specific infection-fatality ratio.

The serological data are dependent on two parameters, the sensitivity and the specificity of the serological testing process, k_{sens} and k_{spec} respectively. If, on day t_k , $n_{r,t_k,a}$ blood samples are taken from individuals in region r and age-group a , and the observed number of positive tests is $y_{r,t_k,a}^s$, then we assume following [8],

$$y_{r,t_k,a}^s | d_I, \psi_r, I_{0,r}, \Delta\beta_r, k_{\text{sens}}, k_{\text{spec}} \sim \text{Bin}\left(n_{r,t_k,a}, k_{\text{sens}} \left(1 - \frac{S_{r,t_k,a}}{N_{r,a}}\right) + (1 - k_{\text{spec}}) \frac{S_{r,t_k,a}}{N_{r,a}}\right), \quad (25)$$

where $S_{r,t_k,a}$ is obtained by solving the difference equations in Eq (20). $N_{r,a}$ is the total population in age-group a and region r .

6.3 Inference

The unknown quantities that need to be inferred can be divided into two groups: (i) Global parameters $\theta_g = (\eta, d_I, p_1, \dots, p_{n_A}, k_{\text{sens}}, k_{\text{spec}}, \sigma_{\beta_w})$ shared between regions, and (ii) regional parameters specific to each region: $\theta_r = (\psi_r, I_{0,r}, \Delta\beta_r)$. After placing the same priors as was used in [8] (and listed in Appendix E in S1 Text), the posterior distribution of the unknown quantities is as follows:

$$p(\theta_g, \theta_1, \dots, \theta_{n_r} | y^d, y^s) \propto p(\theta_g) \prod_{k=1}^K \prod_{a=1}^{n_A} \prod_{r=1}^{n_r} p(y_{r,t_k,a}^d | \theta_g, \theta_r) p(y_{r,t_k,a}^s | \theta_g, \theta_r) p(\theta_r), \quad (26)$$

where we denote by $y^d, y^s \in \mathbb{R}^{K \times n_A \times n_r}$ the data for all time-points, ages and regions corresponding to deaths and sero-positive tests, respectively.

6.3.1 Sampling from the posterior. Sampling from the posterior distribution Eq (26) is challenging due to the large number of random-walk increments corresponding to all regions and weeks since lockdown. MCMC with a vanilla RWMH proposal, as applied in [8], due to the linear scaling of convergence time with increasing dimensions mixes poorly and requires a large number of iterations ($\approx 10^7$) of the Markov chain before convergence is reached. To improve convergence we instead used a random-scan Metropolis-within-Gibbs [32] (MwG) algorithm that circumvent the updating of a large parameter vector at each iteration. This MwG algorithm exploits the independence between the regional parameters. Our proposed sampling strategy consists of sampling alternatively, at each MCMC iteration, from the posterior of the global parameters conditioned on all the regional ones: (i) $p(\theta_g | \theta_1, \dots, \theta_{n_r}, y^d, y^s)$, and regional parameters for one randomly chosen region conditioned on the global ones (since the regional parameters are conditionally independent of any other region's

parameters): (ii) $p(\theta_{r^*} | \theta_g, y^d, y^s)$, where $r^* \sim \text{Uniform}(1, n_r)$. Samples from each of these conditional distributions are obtained using an adaptive RWMH move with the same adaptation mechanism used in Section 4.1. The pseudocode for this MwG algorithm is furnished in Appendix F in [S1 Text](#).

6.3.2 An alternative formulation. The number of region-specific random-walk increments $\Delta\beta_{w_k, r}$ that needs to be sampled increases with time. The performance of the MwG algorithm starts deteriorating and exhibiting poor mixing and slow convergence, as this number becomes large. This limits dramatically the usefulness of this model in the context of a real-time application.

For the model in [Eq \(23\)](#), this problem can be tackled by increasing the time between two successive changepoints thus reducing the number of increments to be sampled for a period of analysis. This is however driven by computational convenience, and it would be more meaningful to learn these changes from data. We could model the time evolution of the transmission-potential at a faster time-scale, for example as in [Eq \(22\)](#). However, in this case the number of random-walk increments, to be sampled per region, equals the number of time-points between lockdown and the end of analysis date. Any MCMC sampler, that uses a RWMH proposal, would struggle severely to move efficiently in such a high-dimensional parameter space.

To alleviate these problems we propose to model the transmission-potential as a Brownian motion $W_{t, r}$ with volatility σ_{β_t} evolving in continuous time t and apply the series approximation as follows:

$$\begin{aligned}\beta_{t, r} &= \sigma_{\beta_t} \sum_{i=1}^n Z_{i, r} \int_0^t \phi_i(u) du \\ &= \sigma_{\beta_t} \sqrt{2T} \sum_{i=1}^n Z_{i, r} \frac{\sin\{(i-1/2)\pi t/T\}}{(i-1/2)\pi} \approx W_t,\end{aligned}\quad (27)$$

where the second equality follows from choosing ϕ_i as given in [Eq \(4\)](#) and carrying out the integration. We can then discretise this approximation using the same time-step of δt that is used for the compartmental dynamics to obtain the following path-wise approximation evolving in discrete time t_k :

$$\beta_{t_k, r} = \sigma_{\beta_t} \sqrt{2T} \sum_{i=1}^n Z_{i, r} \frac{\sin\{(i-1/2)\pi t_k/T\}}{(i-1/2)\pi}, \quad (28)$$

where T is the number of days between lockdown and analysis date. Note that in this formulation the problem of sampling a large vector of increments $\Delta\beta_r$ is reduced to that of sampling a n -dimensional vector of the coefficients $\mathbf{Z}_r = (Z_{1, r}, \dots, Z_{n, r})$. From the comparison of the time T marginal distributions of the true and approximate Brownian motion, as for the OU process (see [Fig 1](#)), we found $n = 10$ to produce a good path-wise approximation. Thus, we used $n = 10$ for the subsequent comparative evaluations. The regional parameter vector, $\theta_r = (\psi_r, I_{0, r}, \mathbf{Z}_r)$, now contains the expansion coefficients instead of the random-walk increments $\Delta\beta_r$.

6.4 Results: Comparative evaluations

We ran the MwG algorithm to target the posterior distribution in [Eq \(26\)](#) while using the random-walk based piecewise constant transmission-potential in [Eq \(23\)](#) and the Brownian motion approximation using series expansion with discretisation (SAd) in [Eq \(28\)](#). In both cases we ran 3×10^6 iterations, discarded the first half of the iterations as burn-in and

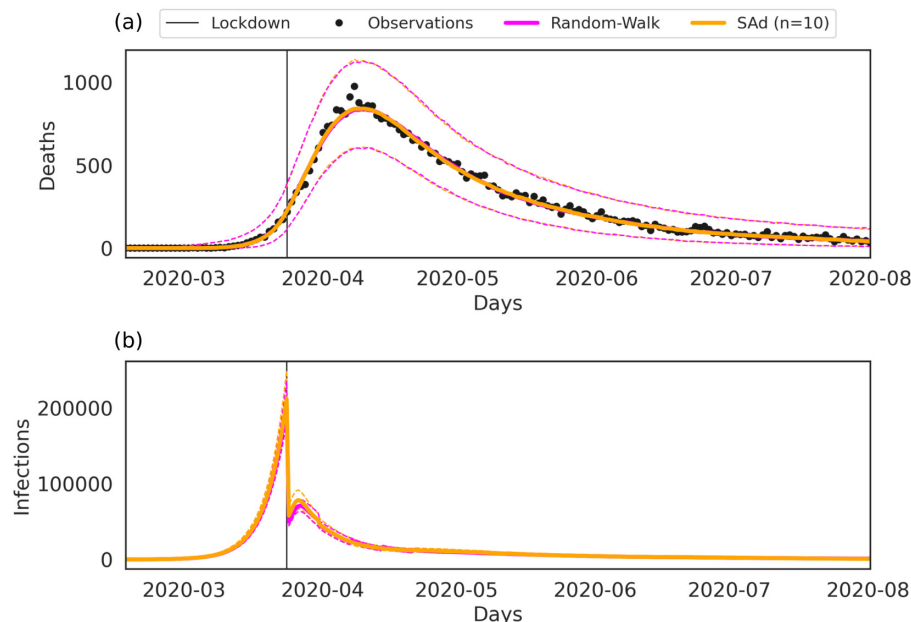


Fig 6. Goodness-of-fit of daily death data (a) and the inferred latent infections (b), produced using the random-walk (magenta lines) and SAd (orange lines). These densities are summarised by the mean (solid lines) and 95% credible intervals (broken lines). The black line indicates the day of lockdown in England 23rd March, 2020.

<https://doi.org/10.1371/journal.pcbi.1011088.g006>

subsequently thinned the remaining samples to obtain 1000 samples. We implemented the epidemic model in C++. The MwG algorithm was implemented using Python.

Fig 6(a) compares, for the two alternative choices of modelling the transmission-potential, the posterior predictive distributions of the death data aggregated across all ages and regions with the observed data (see Appendix G in S1 Text for region-specific plots). Clearly the goodness-of-fit is indistinguishable between the two models. In Fig 6(b) we show summaries of the posterior distributions of the latent infections $p(\Delta^{\text{infect}} | y^d, y^s)$, aggregated across all ages and regions (region-wise infections are shown in Appendix G in S1 Text) again showing close consistency across models, with the exception of a few days immediately following the lockdown where the number of infections estimated by the SAd is slightly higher.

Following [8], we also obtain estimates of the effective region-specific reproduction number $R_{t,k,r}$, their weighted average $R_{t,E}$ representing the reproduction number for all of England, (formulae for these are given in Appendix D in S1 Text). In Fig 7 we show the posterior distributions for $p(R_{t,E} | y^d, y^s)$, using the two alternative models. It is evident that the estimate obtained from the SAd appears to be smoother than what is obtained using the piecewise constant model, more realistically reflecting the actual transmission process that happens in continuous time. In Table 2 we present infection-fatality ratio estimates from the two models, again showing close agreement across models.

Computational gains. The MwG algorithm took around 78 hours to finish for both models of the transmission-potential. However, the SAd allows inference at a faster time-scale producing a smoother estimate of $R_{t,E}$ avoiding artificial model assumptions. Such an inference would be computationally infeasible if using a random-walk model at the more granular time-scale as in Eq (22), given the poor scaling of the RWMH proposal in high dimensions. Thus, using the series approximation we were able to extract more information about the transmission-potential and reproduction-ratio in comparison to the piecewise constant model, while incurring the same computational expense.

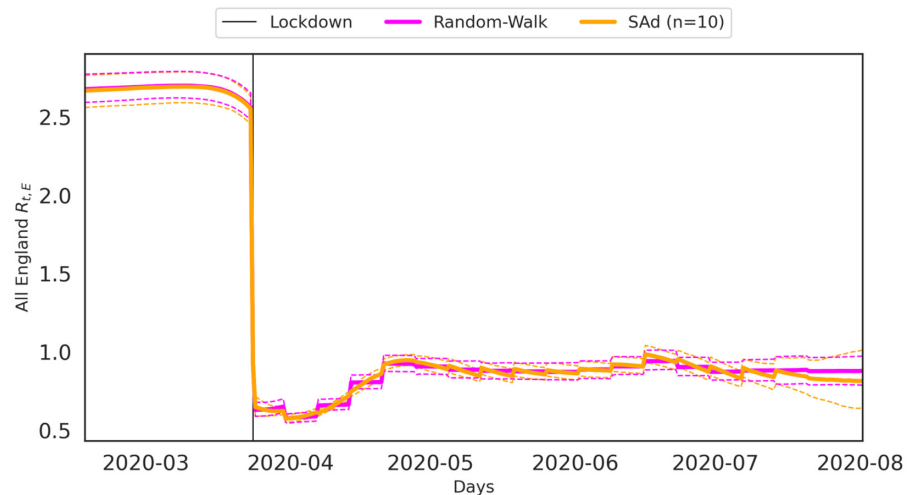


Fig 7. Posterior mean (solid lines) and 95% credible intervals (broken lines) for the all England reproduction number $R_{t,E}$.

<https://doi.org/10.1371/journal.pcbi.1011088.g007>

Had we used the random-walk model in Eq (22), we would have had to further partition each of the regional parameter block in separate chunks to accommodate a large vector of increments $\Delta\beta_{t_k,r} = \log(\beta_{t_k,r}) - \log(\beta_{t_{k-1},r})$. Consequently, multiple Gibbs moves would have been necessary to update all the increments for a randomly chosen region. This, in turn, would have increased the number of likelihood computations, involving the computationally expensive updates of the transmission model, exponentially at each MCMC iteration.

7 Discussion

By modelling the force of infection as the function of a time-varying transmission-potential we can incorporate extrinsic, un-modelled effects in the description of the transmission process within a compartmental model. Describing this transmission-potential, in turn, as a stochastic process, a diffusion in particular, we can inject environmental stochasticity in an otherwise deterministic model. In this paper we proposed a path-wise approximation of a diffusion

Table 2. Posterior mean and 95% credible intervals for the age-specific infection-fatality ratio from the random-walk and SAd models of transmission-potential.

Age group (yrs)	Random-walk	SAd
<5	0.00091% (0.00020%–0.0022%)	0.00093% (0.00071%–0.0019%)
15–14	0.0014% (0.00081%–0.0022%)	0.0014% (0.00062%–0.0022%)
15–24	0.0046% (0.0032%–0.0062%)	0.0044% (0.0029%–0.0060%)
25–44	0.031% (0.028%–0.034%)	0.029% (0.025%–0.034%)
45–64	0.46% (0.44%–0.49%)	0.44% (0.401%–0.49%)
65–74	3.0% (2.9%–3.2%)	2.9% (2.6%–3.3%)
> 74	17% (16%–18%)	17% (15%–18%)

<https://doi.org/10.1371/journal.pcbi.1011088.t002>

process as an alternative to modelling the dynamics of the transmission-potential as a SDE. Through the path-wise approximation we arrive at a random ODE approximating the SDE. As a function of its parameters, the path (solution) of an ODE is completely deterministic. As a result inference of the transmission-potential is simplified, with no need to solve a missing data problem using a computationally expensive data-augmentation procedure.

We demonstrate the efficacy of the proposed path-wise approximation using two epidemic models. In the first one, an influenza model, we replaced an OU SDE with an equivalent path-wise approximation. We noticed similar inference outcomes in terms of parameter estimates and goodness-of-fit using the SDE and its ODE approximation. In the second one, a toy SIRS model, we used an approximation of Brownian motion to reconstruct the non-stationary time evolution of a deterministic sinusoidal transmission potential from simulated data. The path-wise series approximation of Brownian motion was able to faithfully reconstruct the generative values of the sinusoidal transmission potential, as well as the model parameters. Importantly, for both models we observed orders-of-magnitude improvement in computational efficiency of the inference process when using a path-wise approximation of a diffusion process instead of the true diffusion.

We then applied the path-wise approximation to replace a Gaussian random-walk with a discretised path-wise approximation of Brownian motion to model the transmission-potential within a compartmental model of COVID-19 pandemic spread in England. Again we noticed consistent estimates of crucial unknown quantities such as infection-fatality rate, latent infections and a time-varying estimate of the reproduction number. In addition, the path-wise approximation allows the transmission-potential to be modelled at a more granular time-scale providing a smooth estimate of the effective reproduction number. This would be impossible to achieve using the random-walk model due to an exorbitant computational burden.

As an alternative to using our path-wise approximation of Brownian motion to model the transmission-potential, at a faster time-scale, we could have used a different MCMC algorithm, such as the No-U-Turn sampler [33], that is known to perform well for high dimensional problems. This algorithm proposes a move based on the gradient of the target density. Evaluating gradients, however, for the COVID-19 model is challenging as this requires, in addition to extra computations, a complete re-implementation of the model using an automatic differentiation package. However, for modelling studies where such re-implementation is straightforward, we like to point out that by applying a path-wise approximation of a diffusion process we are left with the task of sampling from a posterior distribution with a standard Gaussian prior (over the coefficients). The No-U-Turn sampler generally excels at this task.

In this paper we have used simple diffusion models whose transition densities are known analytically. However, if additional prior information about the force of infection is available, then such information can be incorporated in more complex nonlinear SDEs as models of the time-varying transmission-potential. Moreover, SDEs can be used to model any other time-varying parameters, in addition to the transmission-potential. Our methodology can be seamlessly applied in all such cases to arrive at a path-wise approximation of such complex diffusion processes.

Any time-varying parameter, that is modelled as a diffusion process, can also be modelled using our proposed approximate diffusion process. The usefulness or necessity of considering multiple time-varying parameters, beyond just the transmission potential, depends on the particular application context. In this work our focus primarily was to show the speed-up in inference that can be achieved using an approximate diffusion model of the transmission potential only.

Supporting information

S1 Text. Contains a derivation of the Fourier expansion of Brownian motion (Appendix A), description of the adaptation of the RWMH proposal (Appendix B), simulation study mimicking the influenza dataset (Appendix C), formulae for the reproduction number (Appendix D), priors for the COVID-19 model parameters (Appendix E), pseudocode of the MwG algorithm (Appendix F), region-specific plots for the COVID-19 model (Appendix G), description of the MMD metric (Appendix H), additional experiment using the influenza dataset (Appendix I) and further details of the SIRS model (Appendix J).
(PDF)

Acknowledgments

The authors acknowledge Gareth Roberts, Chris Jewell and Simon Spencer for the helpful discussions at a Bayes4Health meeting where an initial version of this work was presented. The authors also like to thank Angelos Alexopoulos and Colin Starr for discussions and feedback on some aspects of this work.

Author Contributions

Conceptualization: Daniela De Angelis.

Methodology: Sanmitra Ghosh, Paul J. Birrell.

Software: Sanmitra Ghosh.

Supervision: Daniela De Angelis.

Validation: Sanmitra Ghosh, Paul J. Birrell, Daniela De Angelis.

Visualization: Sanmitra Ghosh.

Writing – original draft: Sanmitra Ghosh, Paul J. Birrell, Daniela De Angelis.

Writing – review & editing: Sanmitra Ghosh, Paul J. Birrell, Daniela De Angelis.

References

1. Bretó C, He D, Ionides EL, King AA. Time series analysis via mechanistic models. *The Annals of Applied Statistics*. 2009; p. 319–348.
2. Knock ES, Whittles LK, Lees JA, Perez-Guzman PN, Verity R, FitzJohn RG, et al. Key epidemiological drivers and impact of interventions in the 2020 SARS-CoV-2 epidemic in England. *Science Translational Medicine*. 2021; <https://doi.org/10.1126/scitranslmed.abg4262> PMID: 34158411
3. Keeling MJ, Dyson L, Guyver-Fletcher G, Holmes A, Semple MG, Tildesley MJ, et al. Fitting to the UK COVID-19 outbreak, short-term forecasts and estimating the reproductive number. *medRxiv*. 2020;.
4. Davies NG, Kucharski AJ, Eggo RM, Gimma A, Edmunds WJ, Jombart T, et al. Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: a modelling study. *The Lancet Public Health*. 2020; 5(7):e375–e385. [https://doi.org/10.1016/S2468-2667\(20\)30133-X](https://doi.org/10.1016/S2468-2667(20)30133-X) PMID: 32502389
5. Brynjarsdóttir J, O'Hagan A. Learning about physical parameters: the importance of model discrepancy. *Inverse Problems*. 2014; 30(11):114007. <https://doi.org/10.1088/0266-5611/30/11/114007>
6. Dureau J, Kalogeropoulos K, Baguelin M. Capturing the time-varying drivers of an epidemic using stochastic dynamical systems. *Biostatistics*. 2013; 14(3):541–555. <https://doi.org/10.1093/biostatistics/kxs052> PMID: 23292757
7. Cazelles B, Champagne C, Dureau J. Accounting for non-stationarity in epidemiology by embedding time-varying parameters in stochastic models. *PLoS computational biology*. 2018; 14(8):e1006211. <https://doi.org/10.1371/journal.pcbi.1006211> PMID: 30110322

8. Birrell P, Blake J, Van Leeuwen E, Gent N, De Angelis D. Real-time nowcasting and forecasting of COVID-19 dynamics in England: the first wave. *Philosophical Transactions of the Royal Society B*. 2021; 376(1829):20200279. <https://doi.org/10.1098/rstb.2020.0279> PMID: 34053254
9. De Angelis D, Presanis AM, Birrell PJ, Tomba GS, House T. Four key challenges in infectious disease modelling using data from multiple sources. *Epidemics*. 2015; 10:83–87. <https://doi.org/10.1016/j.epidem.2014.09.004> PMID: 25843390
10. Fuchs C. Inference for diffusion processes: with applications in life sciences. Springer Science & Business Media; 2013.
11. Sottinen T, Särkkä S. Application of Girsanov theorem to particle filtering of discretely observed continuous-time non-linear systems. *Bayesian Analysis*. 2008; 3(3):555–584. <https://doi.org/10.1214/08-BA322>
12. Anderson RM, Anderson B, May RM. Infectious diseases of humans: dynamics and control. Oxford university press; 1992.
13. Ellner SP, Bailey B, Bobashev GV, Gallant A, Grenfell B, Nychka D. Noise and nonlinearity in measles epidemics: combining mechanistic and statistical approaches to population modeling. *The American Naturalist*. 1998; 151(5):425–440. <https://doi.org/10.1086/286130> PMID: 18811317
14. Martinez-Bakker M, King AA, Rohani P. Unraveling the transmission ecology of polio. *PLoS biology*. 2015; 13(6):e1002172. <https://doi.org/10.1371/journal.pbio.1002172> PMID: 26090784
15. Cazelles B, Chau N. Using the Kalman filter and dynamic models to assess the changing HIV/AIDS epidemic. *Mathematical biosciences*. 1997; 140(2):131–154. [https://doi.org/10.1016/S0025-5564\(96\)00155-1](https://doi.org/10.1016/S0025-5564(96)00155-1) PMID: 9046772
16. Oksendal B. Stochastic differential equations: an introduction with applications. Springer Science & Business Media; 2013.
17. Lyons SMJ, Storkey AJ, Särkkä S. The Coloured Noise Expansion and Parameter Estimation of Diffusion Processes. In: Bartlett PL, Pereira FCN, Burges CJC, Bottou L, Weinberger KQ, editors. *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3–6, 2012, Lake Tahoe, Nevada, United States; 2012. p. 1961–1969.*
18. Luo W. Wiener chaos expansion and numerical solutions of stochastic partial differential equations. California Institute of Technology; 2006.
19. Ghosh S, Birrell PJ, De Angelis D. Differentiable Bayesian inference of SDE parameters using a pathwise series expansion of Brownian motion. In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics. vol. 151 of Proceedings of Machine Learning Research. PMLR; 2022. p. 10982–10998.*
20. Lyons SM, Särkkä S, Storkey AJ. Series expansion approximations of Brownian motion for non-linear Kalman filtering of diffusion processes. *IEEE Transactions on Signal Processing*. 2014; 62(6):1514–1524. <https://doi.org/10.1109/TSP.2014.2303430>
21. Wong E, Zakai M. On the Convergence of Ordinary Integrals to Stochastic Integrals. *The Annals of Mathematical Statistics*. 1965; 36(5):1560–1564. <https://doi.org/10.1214/aoms/1177699916>
22. Kloeden PE, Eckhard P. Numerical Solution of Stochastic Differential Equations. Springer Berlin; 1992.
23. Jackson C, Vynnycky E, Hawker J, Olowokure B, Mangtani P. School closures and influenza: systematic review of epidemiological studies. *BMJ open*. 2013; 3(2). <https://doi.org/10.1136/bmjopen-2012-002149> PMID: 23447463
24. Jombart T, Frost S, Nouvellet P, Campbell F, Sudre B. outbreaks: A Collection of Disease Outbreak Data; 2020. Available from: <https://CRAN.R-project.org/package=outbreaks>.
25. Del Moral P, Murray LM. Sequential Monte Carlo with highly informative observations. *SIAM/ASA Journal on Uncertainty Quantification*. 2015; 3(1):969–997. <https://doi.org/10.1137/15M1011214>
26. Ryder T, Golightly A, McGough AS, Prangle D. Black-Box Variational Inference for Stochastic Differential Equations. In: *Proceedings of the 35th International Conference on Machine Learning; 2018. p. 4423–4432.*
27. Andrieu C, Thoms J. A tutorial on adaptive MCMC. *Statistics and computing*. 2008; 18(4):343–373. <https://doi.org/10.1007/s11222-008-9110-y>
28. Gordon N, Salmond D, Ewing C. Bayesian state estimation for tracking and guidance using the bootstrap filter. *Journal of Guidance, Control, and Dynamics*. 1995; 18(6):1434–1443. <https://doi.org/10.2514/3.21565>
29. Bradbury J, Frostig R, Hawkins P, Johnson MJ, Leary C, Maclaurin D, et al.. JAX: composable transformations of Python+NumPy programs; 2018. Available from: <http://github.com/google/jax>.
30. Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A. A kernel two-sample test. *Journal of Machine Learning Research*. 2012; 13(Mar):723–773.

31. Birrell PJ, Ketsetzis G, Gay NJ, Cooper BS, Presanis AM, Harris RJ, et al. Bayesian modeling to unmask and predict influenza A/H1N1pdm dynamics in London. *Proceedings of the National Academy of Sciences*. 2011; 108(45):18238–18243. <https://doi.org/10.1073/pnas.1103002108>
32. Roberts GO, Rosenthal JS. Markov-chain Monte Carlo: some practical implications of theoretical results. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*. 1998; p. 5–20. <https://doi.org/10.2307/3315667>
33. Hoffman MD, Gelman A. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*. 2014; 15(1):1593–1623.