Approximate Bayesian Computation (ABC) for Modelling the COVID-19 Pandemic in the United States of America

Zoë Ganess

MSc in Data Science & Statistics

University of Bath

September 2022

Abstract

Approximate Bayesian Computation (ABC) allows for approximation of the posterior distribution under an intractable or computationally expensive likelihood function. This, along with the fact that ABC can be used even when available data is coarse or complex, makes application of the algorithm suited to infectious disease modelling. Even in other fields, ABC has successfully been applied where more traditional methods such as Markov chain Monte Carlo (MCMC) quickly become infeasible. In this work, the most basic ABC algorithm and a modified Sequential Monte Carlo (SMC) ABC algorithm are implemented for parameter estimation of COVID-19 data in three U.S. states under the stochastic spatial SEIR framework. By using the R package ABSEIR, such simulation studies are conducted and analysed, along with previous theoretical studies, to determine the ability of the algorithms and the SEIR models, to model the pandemic both in an accurate and computationally efficient manner.

This dissertation may be made available for consultation within the University Library and may be photocopied or lent to other libraries for the purposes of consultation.

Signed: Zoë Ganess

Approximate Bayesian Computation (ABC) for Modelling the COVID-19 Pandemic in the United States of America

submitted by
Zoë Ganess

for the degree of MSc in Data Science & Statistics of the
The University of Bath

September 2022

DECLARATION

This dissertation is submitted to the University of Bath in accordance with the requirements of the degree of MSc in Data Science & Statistics in the Department of Computer Science. No portion of the work in this thesis has been submitted in support of an application for any other degree or qualification of this or any other university or institution of learning. Except where specifically acknowledged, it is the work of the author.

Signature of Author……

ZOË GANESS

Contents

CHAPTER 1

INTRODUCTION

This chapter discusses the background and motivation for the utilisation of Approximate Bayesian Computation (ABC) in modelling the COVID-19 pandemic in the United States of America. Here, a general overview of key concepts which encompass the overall project is provided. These include stochastic spatial SEIR (epidemic) models, ABC and the COVID-19 pandemic. As the two former concepts are further discussed in their respective chapters, the reader is aided in developing a solid understanding of components necessary for interpreting results of the conducted simulation study. Lastly, the aims, objectives and motivation for the project are defined.

1.1 Background

Stochastic spatial SEIR models

Statistical inference of infectious disease data is most frequently approached via the method of classifying individuals in a population according to their epidemiological status over time. This classification, formally observed in the formation of compartmental or compartmentalised epidemiological models, can be understood as a simplification of realistic underlying disease progression, such that persons experiencing respective disease states are categorised into stages or 'compartments' [1]. This follows the inherent framework of an epidemic timeline, as individuals in a population will often transverse systematically through several stages of disease development before they no longer contribute to disease spread [2]. With this in mind, one of the most commonly used compartmental models is the SEIR model, where each letter in the abbreviation "SEIR" represents four disease states. The **S**usceptible compartment comprises of individuals who are able to contract the disease based on contact with infectious persons. The **E**xposed compartment consists of individuals who are infected but not yet infectious and therefore, describes the latent period of the respective disease. The

Infectious compartment contains individuals who are infected and are capable of contributing to disease spread. Lastly, the **R**emoved compartment includes individuals who no longer belong to the infectious population due to mortality or recovery with immunity. Due to the possibility of immunity attenuation and therefore, reinfection with some diseases, removed individuals with previously gained immunity may still retransition to the susceptible state, this of which is modelled by the SEIR(S) model. Stochasticity is implemented into these models using specified transition probabilities. Simply put, these represent the probability of an individual moving from one disease state to another in terms of time (temporality) and location (spatiality); the rates at which persons transition between compartments serves as a natural interpretation of these probabilities [2]. As epidemics therefore present themselves as spatiotemporal processes, there is often a trade-off between complexity and practicality in epidemiological modelling. Statistical inference becomes rather challenging in such cases as there are constraints due to computational feasibility as well as limited information available, both in terms of data type and quantity.

Approximate Bayesian Computation (ABC)

Markov Chain Monte Carlo (MCMC) serves as the most traditional method in which simulation studies for complex stochastic models are relatively feasible. However, MCMC quickly becomes impractical in high dimensional problems where evaluation of the posterior distribution is impossible or computationally expensive [3]. As such, numerous alternative inferential approaches have been devised, one of which can be seen via Approximate Bayesian Computation, or ABC. Generally, the conceptual process of ABC is straightforward to understand: simulated pseudo-data generated from the proposed model, conditional on a set of parameters, is compared to observed data and assessed in terms of likeliness of the model producing the same observed data. Direct derivation of the potentially intractable posterior distribution is avoided as samples are only drawn from its likelihood function. Stochastic spatial SEIR models therefore fit nicely under this algorithm due to the way in which its components are parameterised. Under this framework, there is not only direct application to these models, but the prospect of additional posterior approximation techniques are also warranted.

The COVID-19 Pandemic

First declared a global pandemic by the World Health Organization (WHO) on March 11th, 2020, the ongoing spread of the Coronavirus disease 2019 (COVID-19) has brought much attention to the public eye regarding various prospects of infectious disease and the importance of epidemiological studies. In particular, under the amplified need for accurate statistical modelling of the disease, many challenges, such as asymptomatic cases, under-reporting and general lack of testing, have also been established. Estimating predictions such as disease transmission rate and epidemic trajectory therefore presents itself as a much more complex task. With the continuing emergence of different variants, the issue of varying latent and infectious periods also presents itself as a prominent issue [4], as these timeframes are central to achieving accurate predictive results. Further complications arise when considering that these periods may also be age and/or region specific. Globally, interventive methods such as lockdowns, mandated mask-wearing, social distancing requirements, travel restrictions, quarantines/isolation periods and vaccination campaigns, have been shown to reduce a substantial amount of transmission and overall disease spread [5]. With more than 6.5 million recorded deaths worldwide, more than a million of which have occurred in the States, the disease still remains an issue at the forefront of public health agendas.

1.2 Aims, objectives and motivation

The main aim of this project is to determine the suitability of approximate Bayesian computation (ABC) in estimation of SEIR model parameters as it relates to the COVID-19 pandemic in selected American states. This will be assessed in a practical setting in R using the package ABSEIR and relevant data. Predictive results will be compared in terms of estimation performance as well as computational efficiency by conducting a series of simulations. Additionally, theoretical analysis of previous research will aid in exploring the practicality of ABC as related to its relative merits and/or drawbacks. Evaluation and criticisms of the basic SEIR model will also be documented in terms of its ability to accurately model the pandemic. Recommendations regarding any possible supplemental data or alternative methods

which could improve upon basic ABC and SEIR models will also be explored, so as to ameliorate the prediction of infectious-disease dynamics in the future.

Considering the global prevalence of the on-going COVID-19 pandemic, as well as the ubiquity of infectious disease on a whole, the motivation behind this project is driven by the persistent need to improve epidemic modelling, whether it be via the model itself or how it is implemented/fitted. The importance of modelling disease transmission is generally understood as it provides useful insights into the behaviour of epidemic systems. By provision of accurately estimated parameters as well as quantitative predictions using data collected from an epidemic outbreak, subsequent development of potential strategies/public health measures for disease control in terms of curving the outbreak at hand and/or future outbreaks is made possible [2, 6].

To summarise, the project aims can be concisely defined as follows:

- To critically assess the practicality of ABC based on theory as well as simulations based on state-level COVID-19 data using ABSEIR.
- To analyse the suitability of the basic SEIR model in realistically predicting infectious disease dynamics as it relates to the COVID-19 pandemic.
- To propose recommendations which could improve upon basic ABC as well as the basic SEIR model.

CHAPTER 2

STOCHASTIC SPATIAL SEIR MODELS

2.1 Origin

In 1927, Kermack and McKendrick [7] provided the first formal mathematical introduction of the compartmental approach with the initial **S**usceptible, **I**nfectious and **R**emoved (SIR) model. This represented a significant advancement in epidemiological modelling as it served as the basis for the formulation of extended models such as the SEIR model. However, the original approach was still limited due to its inherently deterministic nature. This is contrary to the reality of disease transmission as it presents itself as a stochastic process. As such, another important milestone was recorded in 1999 when the first Bayesian stochastic SEIR compartmental model was presented by O'Neill and Roberts [8]. The initial SIR model is also criticised as it assumes the existence of homogeneous mixing. Under the consideration of natural propagative factors of disease, such as super-spreaders and population characteristics, such as age, the nature of epidemic progression becomes increasingly heterogenous. Additionally, the intuitive belief that most disease contact and transmission is related to the distance between individuals in a population is recognised and as such, spatial heterogeneity should also be considered [9]. Notably, these limitations are further improved upon by the stochastic spatial SEIR model by Brown, et al. [10] as introduced below.

2.2 Methods

As derived by Brown, et al. [10], stochastic spatial SEIR models, compartments and associated transitions are defined over discrete time $\{t_i: t_1, \dots, t_T\}$ as well as discrete space/spatial locations $\{s_j: s_1, \dots, s_n\}$. Observed data is represented by the matrix $\mathbf{Y} = [\mathbf{y_1} \dots \mathbf{y_n}]$, such that $\mathbf{y_j}$ is a $T \times 1$ column vector comprising of data for location s_j. Unknown count parameters to be estimated are represented by **S**, **E**, **I**, and **R**, each of which is a $T \times n$ matrix

corresponding to the susceptible, exposed, infectious and removed compartments, respectively. Similarly, the transition matrices, **S***, **E***, **I***, and **R***, encompass transitions into each respective compartment. Therefore, this temporal relationship can be described by the following set of difference equations:

$$\begin{aligned} \boldsymbol{S}_{i+1} &= \boldsymbol{S}_i - \boldsymbol{E}_i^* + \boldsymbol{S}_i^* \\ \boldsymbol{E}_{i+1} &= \boldsymbol{E}_i - \boldsymbol{I}_i^* + \boldsymbol{E}_i^* \\ \boldsymbol{I}_{i+1} &= \boldsymbol{I}_i - \boldsymbol{R}_i^* + \boldsymbol{I}_i^* \\ \boldsymbol{R}_{i+1} &= \boldsymbol{R}_i - \boldsymbol{S}_i^* + \boldsymbol{R}_i^* \end{aligned} \tag{1}$$

such that $S_i + E_i + I_i + R_i = N \; \forall \; i$, where N represents the vector of fixed population sizes at a particular time [1]. Likewise, the transition matrices can be parametrised by applying a chain binomial structure given by:

$$\begin{aligned} \boldsymbol{S}_{ij}^* &\sim binom\left(\boldsymbol{R}_{ij}, \pi_{ij}^{(RS)}\right) \\ \boldsymbol{E}_{ij}^* &\sim binom\left(\boldsymbol{S}_{ij}, \pi_{ij}^{(SE)}\right) \\ \boldsymbol{I}_{ij}^* &\sim binom\left(\boldsymbol{E}_{ij}, \pi_{i}^{(EI)}\right) \\ \boldsymbol{R}_{ij}^* &\sim binom\left(\boldsymbol{I}_{ij}, \pi_{i}^{(IR)}\right) \end{aligned} \tag{2}$$

such that the transition probabilities are labelled hierarchically according to the two compartments which they link. Using the second equation above as an example, $\boldsymbol{S}_{ij}$ is the number of susceptible individuals in location s_j at time t_i, $\pi_{ij}^{(SE)}$ is the probability that such (susceptible) individuals will transition to the exposed compartment and $\boldsymbol{E}_{ij}^*$ is the number of such individuals transitioning to the exposed compartment at the same time/location [3].

Collectively, the transition probabilities $\pi_{ij}^{(SE)}, \pi_i^{(EI)}$ and $\pi_i^{(IR)}$ encode significant information regarding disease transmission and progression. Formulation of the exposure

probabilities $\{\pi_{ij}^{(SE)}\}$ are highlighted as a means for spatial heterogeneity incorporation. Under a spatial setting, the intensity process of any disease is seen as a combination of population mixing, pathogen infectivity and spatial heterogeneity. Parameterisation of the exposure probabilities governs this intensity such that the probability of infection for an individual living in location s_j at time t_i is as specified below:

$$\pi_{ij}^{(SE)} = 1 - exp\left\{-\delta_{ij}e^{\theta_{ij}} - \sum_{\{l \neq j\}} (f(d_{jl})\delta_{il}e^{\theta_{il}})\right\} \quad (3)$$

where:

- δ_{ij} represents the proportion of infectious individuals in location s_j at time t_i .
- $f(d_{jl})$ represents some known function proportional to the contact between spatial locations; d_{jl} specifies a distance metric between spatial locations s_j and s_l.
- $\theta_{ij} = \log(\lambda_{ij}p)$ is the exposure intensity parameter for location s_j and time t_i .
    - p is the probability in which an exposed individual becomes infected with a specified disease.
    - λ_{ij} parameterises the Poisson distribution under the assumption that the number of 'contacts' K_{ij} between a person of interest and other individuals in location s_j at time t_i is $K_{ij} \sim \text{Poisson}(\lambda_{ij})$ [1, 10].

The above spatial formulation was originally derived using the work of Lekone and Finkenstädt [11] in the non-spatial case as more generally defined below:

$$\pi_i^{(SE)} = 1 - exp\left\{-e^{\theta_i}\frac{I_i}{N}\right\} \quad (4)$$

$\pi_i^{(EI)}$ and $\pi_i^{(IR)}$ represent the latent and infectious periods, respectively. The latent period is therefore characterised by the length of time from exposure to symptom onset (i.e.

time spent in the **E**xposed compartment) whereas the infectious period describes the length of time from which an individual becomes infectious until recovery or death (i.e. time spent in the **I**nfectious compartment). In the simplest case, these probabilities may be assumed to be distributed under the exponential distribution as follows:

$$\pi_i^{(EI)} = 1 - \exp\left(-\gamma_{(EI)}\right)$$
$$\pi_i^{(IR)} = 1 - \exp\left(-\gamma_{(IR)}\right) \quad (5)$$

where $\gamma_{(EI)}$ and $\gamma_{(IR)}$ are the rates at which individuals transition from one respective compartment to another. The mean latent and infectious periods are therefore given by $1/\gamma_{(EI)}$ and $1/\gamma_{(IR)}$, respectively. Additionally, by placing gamma priors on these rate terms, such that $\gamma_{(EI)} \sim gamma(\alpha_{(EI)}, \beta_{(EI)})$ and $\gamma_{(IR)} \sim gamma(\alpha_{(IR)}, \beta_{(IR)})$, these probabilities are provided both a more flexible and proper range [1]. However, it is relevant to emphasise that such exponentially distributed periods imply that there is a constant probability of transition due to the memorylessness property of the distribution. This is problematic as in reality, compartment membership times are dependent on the length of time already spent in a particular disease state [1, 9]. Despite this, these constant transition probabilities usually produce acceptable fits as well as provide computational benefits [3].

The advent of the path-specific (PS SEIR) structure by Porter and Oleson [12] allows for the incorporation of non-exponential compartment membership times. As such, a much more reasonable disease process is represented, as latent and infectious periods are now defined by continuous random variables equivalent to the amount of time spent in respective compartments. The PS SEIR transition probabilities can be formulated as follows:

$$\pi_w^{(EI)} = P(W_1 \leq w + 1 \mid W_1 > w)$$
$$\pi_w^{(IR)} = P(W_2 \leq w + 1 \mid W_2 > w) \quad (6)$$

where the random variables W_1 and W_2 define the latent and infectious distributions and w corresponds to the discrete time spent in the exposed or infectious compartments [9]. Common distributions utilised for aforementioned random variables are gamma or Weibull.

2.3 Literature Review

Strengths and limitations

As explained by Whiteley and Rimella [13], the main advantage of such compartmental models is the allowance of joint modelling of disease dynamics and multimodal data. As such, due to this flexibility associated with the baseline model, there have been several extended models which include additional compartments in an attempt to better capture the scope of epidemiological processes. The main problem associated with stochastic spatial SEIR models, as well as other compartmental models, is that inference of their parameters is non-trivial in a statistical sense. As discussed by Boys and Giles [2], due to the nature of partially observed epidemic outbreaks, it becomes increasingly difficult to obtain necessary/accurate data. An example of this can be seen via the feasibility of accurately recording times at which individuals contract the pathogen of interest. Furthermore, as stated by several authors [3, 6, 14, 15], under higher dimensional problems, parameter estimation becomes computationally expensive/intractable due to increasing complexity and population size. Whiteley and Rimella [13] provide further explanation for this, stating that this intractability associated with the likelihood function, arises due to "summation over a prohibitively large number of configurations of latent variables representing counts of subpopulations in disease states which cannot be observed directly." The traditional model is also limited structurally when considering the convoluted and intricate nature of infectious diseases and their interactions with the human population, such that all pathogens, strains and disease processes cannot be fully accommodated. To overcome these issues of data paucity and parameter estimation, statistical methodologies incorporating "likelihood-free" inference algorithms, such as Approximate Bayesian Computation (ABC), have become increasingly developed and utilised for simulation-based approximation [13, 16, 17].

Extended SEIR models for COVID-19 data

The flexibility of compartmental models has led to the genesis of a variety of extended SEIR models, in response to the need for accurate predictions under complex infectious disease processes. As such, several of such models have been developed to model the dynamics of the COVID-19 pandemic as accurately as possible. One of the earlier studies conducted in 2020 by Lin, et al. [18] utilised a conceptual SEIR model with extra compartments corresponding to the cumulative number of cases as well as public perception of risk. Particularly, the inclusion of the cumulative number of cases increases the probability of making more realistic estimates due to the existence of unreported cases, as suggested by Wu and McGoogan [19]. A more recent study by Chen, et al. [20] also addressed the issue of unreported cases by dividing the singular infectious compartment into two, one for reported symptomatic cases (I) and unreported symptomatic cases (U), to give a SEIUR model. Along with other characteristics associated with the virus, such as asymptomatic and symptomatic manifestations, models such as the one presented by Grzybowski, et al. [21], include compartments corresponding to quarantined infected (Q) and confined susceptible (C) individuals (SEIRCQ). The relative inclusion of these extra compartments has its benefits, but as explained by Reis, et al. [22], accurate model calibration may be hindered due to these additions which result in an increase in the number of unknown parameters to be estimated.

CHAPTER 3

APPROXIMATE BAYESIAN COMPUTATION (ABC)

3.1 Origin

In 1984, Rubin [23] provided the first description of approximate Bayesian computational (ABC) methods, stating that "Bayesian statistics and Monte Carlo methods are ideally suited to the task of passing many models over one dataset," [23, p. 1159]. It was not until 1997 when this concept was proposed under the specification of a rejection algorithm, this of which can be attributed to Tavare, et al. [24]. Rather than 'approximate', this initial rejection algorithm was termed to be the 'exact' Bayesian computation algorithm. This is precisely because this algorithm was conditioned by the acceptance of the simulated data being exactly equal to the observed data. However, as raised by Kypraios, et al. [25], this exactitude is only practical when the observed data is discrete. A more elegant approach under continuous distributions or discrete ones in which the probability of the aforementioned exactitude is unacceptably low was suggested by Pritchard, et al. [26] in a 1999 paper on population genetics. In particular, this serves as the most basic algorithm used today and as such, will be discussed further.

3.2 Methods

Contrary to a frequentist approach, the assumption of model parameters as random variables in Bayesian analysis implies the existence of probability distributions of these parameters. One such distribution, which is central to Bayesian inference is the posterior distribution, this of which is to be approximated for respective model parameters. Consider the traditional equation of the posterior distribution derived using Bayes' theorem:

$$f_{\boldsymbol{\theta}}(\theta|\boldsymbol{Y}) = \frac{f_{\boldsymbol{Y}}(\boldsymbol{y}|\theta)\pi_{\boldsymbol{\theta}}(\theta)}{\int_{\theta} f_{\boldsymbol{Y}}(\boldsymbol{y}|\theta)\pi_{\boldsymbol{\theta}}(\theta)d\theta} \propto f_{\boldsymbol{Y}}(\boldsymbol{y}|\theta)\pi_{\boldsymbol{\theta}}(\theta) \quad (7)$$

where θ is a $p \times 1$ vector representing the unknown parameters, with p dimensional parameter space $\boldsymbol{\Theta}$ and prior distribution $\pi_{\boldsymbol{\theta}}(\theta)$. The prior distribution is based off of (prior) information regarding the parameters to be estimated, such as the mean latent and infectious periods in epidemiological modelling. Once data is observed ($N \times 1$ vector represented by $\boldsymbol{y}$), the prior distribution is updated and renormalised using the likelihood of the observed data $f_{\boldsymbol{Y}}(\boldsymbol{y}|\theta)$ [3, 25].

Origination of Approximate Bayesian Computation (ABC) was driven by the increasing need to tackle laborious inferential analysis in population genetics and epidemics. The considerable complexity associated with models such as SEIR models, for reasons discussed previously, renders the associated likelihood function(s) $f_{\boldsymbol{Y}}(\boldsymbol{y}|\theta)$ (and consequently, the corresponding posterior distribution $f_{\boldsymbol{\theta}}(\theta|\boldsymbol{Y})$) computationally expensive and therefore, intractable. In particular, it is rarely possible to analytically calculate the likelihood in a temporal setting, as it involves integration over all possible infection times. Thus, being a "likelihood-free" algorithm, ABC is highly favoured as an effective and intuitively accessible method for approximate analysis in a Bayesian setting [25, 27]. Formally, this concept can be summarised by the ABC rejection algorithm below:

Algorithm 1 ABC Rejection Algorithm

Require: Define a tolerance $\boldsymbol{\epsilon > 0}$ and let ' ← ' denote assignment

1: **for** $\boldsymbol{i} \leftarrow \boldsymbol{1}$ to $\boldsymbol{n}$ **do**
2: $\quad \boldsymbol{d} \leftarrow \infty$
3: $\quad$ **while** $\boldsymbol{d > \epsilon}$ **do**
4: $\quad\quad$ draw $\boldsymbol{\theta_i} \sim \boldsymbol{\pi(\Theta)}$
5: $\quad\quad$ draw $\boldsymbol{x_i} \sim \boldsymbol{f_Y(y|\theta)_i}$
6: $\quad\quad \boldsymbol{d} \leftarrow \boldsymbol{\rho(y, x_i)}$

Using the prior distribution $\boldsymbol{\pi(\Theta)}$, samples $\boldsymbol{\theta_i}$ (candidate parameter vector(s)) are generated repeatedly and subsequently utilised for generation of a simulated dataset $\boldsymbol{x_i}$ from the model/ likelihood function $\boldsymbol{f_Y(y|\theta)_i}$. This is not to be confused with evaluation of the potentially problematic likelihood as samples are only being generated from it in this case [23, 28]. The algorithm 'accepts' the suggested values for the parameters according to the distance function $\boldsymbol{\rho(y, x_i)}$ which represents some distance measure between the model output and the observed data. More specifically, retention of a parameter is warranted once there is generation of replicate datasets such that the simulated data is sufficiently 'close' to the observed data $\boldsymbol{y}$, on the basis of some specified distance and tolerance $\boldsymbol{\epsilon}$ [3, 17, 29]. In other words, if $\boldsymbol{\rho(y, x_i) \leq \epsilon}$, accept $\boldsymbol{\theta_i}$. Direct comparison of simulated data versus observed data is therefore feasible due to the presence of this distance measure. To demonstrate this, consider data in terms of daily infection cases of a disease where a SEIR model would be applicable. In terms of estimating and selecting the respective parameters, it would be sensible to compare the sum of the squared differences between the number of infected persons, corresponding to the actual and predicted data [15].

While the basic ABC algorithm improves upon estimation under intractable likelihood functions, it has been observed that sampling performance tends to decrease when priors are diffuse to the posterior distribution. This shortcoming is further exacerbated when dealing with high dimensional problems [28, 30]. As such, various extensions to the basic algorithm have been devised in order to increase sampling efficiency [3]. One example of this is the Sequential Monte Carlo (SMC) ABC algorithm originally proposed by Toni, et al. [29]. Rather than directly sampling the posterior, a sequence of proxy distributions constructed by gradually decreasing tolerance(s) $\boldsymbol{\epsilon}$ are utilised instead, such that there is convergence to the posterior. For a sequence of K distributions, a finite number of parameter sets (or particles), are randomly sampled from the prior distribution. Each set of parameters under subsequent distributions are then generated using a series of sequential sampling steps, such that it is weighted from the previous, corresponding distribution [6, 15].

3.3 Literature Review

MCMC versus ABC

Unlike ABC, posterior inference via Markov chain Monte Carlo (MCMC) methods directly utilise the posterior distribution of interest for Markov chain simulation; these methods are therefore likelihood-dependent [2]. In the case of epidemiological data, Neal [16] notes that applications of MCMC have been attempted either in terms of temporal epidemics or final size epidemic data. However, in such cases, inferential deductions under MCMC become relatively convoluted due to data insufficiency, hence rendering the model likelihood intractable. Methodologically, this intractability does not directly affect the efficiency of MCMC. Neal [16] further emphasises that tractability can be deduced via data augmentation such that additional data is included as model parameters. Such inclusion results in a trade-off between a tractable likelihood and the relative efficiency of the MCMC algorithm. This notion is supported by Marin, et al. [31] who state that induction of increased dimension due to the data augmentation significantly decreases the convergence properties of the MCMC algorithm such that they are no longer able to be considered in a practical setting. Following the reasons specified in the previous sections, it is known why ABC has become increasingly popular under such high dimensional settings. Specifically, it provides an almost automated resolution to models with intractable likelihoods via straightforward simulations from the equivalent distribution.

Strengths and limitations

ABC is conventionally viewed in a positive light as it allows for inference under conditions where traditional algorithms such as Markov chain Monte Carlo (MCMC) methods become less practical or simply fail. However, little is known about the quality of the approximation provided by ABC beyond results shown in simulation studies [32]. Even within simulation studies, the convergence results (i.e. obtaining an exact match between simulated

and actual data) either require the tolerance to approach zero or the sample size to approach infinity. This renders the probability of obtaining such a result to be miniscule due to the impracticality of such a request. Another fundamental challenge of ABC is seen through the use of summary statistics which aids in dimensionality reduction under large datasets. Not only does this impose the threat of information loss, but one must now consider the possibility of biases and how they may have manifested in any inferential analysis [31, 33]. Due to the inherent nature of the curse of dimensionality, this cannot be avoided in higher dimensional problems, but rather, the relative balance between low dimensionality and informativeness of a model must be preserved as much as possible. In a practical setting, this is may be achieved by subset selection methods, these of which attempt to select an informative subset from a candidate of summary statistics [27].

A summary of the potential advantages as it relates to the use of ABC for fitting epidemic models is discussed in a paper by McKinley, et al. [6]. It claims that due to the simulative nature of the models, imputation of missing data is naturally facilitated. Additionally, inclusion of prior information as it relates to model parameters is afforded due to the Bayesian setting. In the Bayesian MCMC framework it is straightforward to impute missing data at each iteration of the Markov chain which allows us to numerically integrate over the probability distributions of the unobserved process. Additionally, Bayesian inference allows for the incorporation of prior knowledge about the disease process, which is often available in the form of epidemiological case studies, virology studies, and other modelling efforts. The importance of this point is seen via the fact that highly correlated parameters exist at a substantial degree in epidemic processes. In a frequentist framework, this high correlation induces the risk of multicollinearity in the model, however, if prior information is known about these parameters, say through past experimental studies, then this inclusion may reduce the chance of this problem occurring. Lastly, the issue of data paucity is bypassed via the use of some summary or sufficient statistic such that key epidemiological features are most often always utilised to drive model fit. Studies by McKinley, et al. [33] and Wilkinson [32] both expand on the point of summary statistics, attributing this inclusion to the algorithm's ability to facilitate dimensionality reduction. Götte [27] further strengthens the understanding of this concept by emphasising that the number of summary statistics should not be exceedingly large, as this either would lower acceptance rates to prohibitive levels or require the tolerance to be increased, leading to a distortion in the relative approximation.

As with any statistical method, the efficiency of basic ABC should be considered. As explained by McKinley, et al. [33], implementation of basic ABC in an efficient manner is usually deemed difficult. Brown, et al. [3], as well as several other researchers, go on to provide the same possible reason for this, stating that sampling performance is significantly degraded due to generation of a low acceptance probability, once the prior and posterior distributions are diffused or quite different from each other. In other words, the basic algorithm maintains its usefulness once acceptable prior information is available, such that prior distribution proposals fall within regions corresponding to high posterior density [34]. A well-known solution to this problem is the previously introduced SMC-ABC algorithm developed by Toni, et al. [29].

CHAPTER 4

ABSEIR

The package ABSEIR created by Brown [34], provides a convenient means of implementing approximate Bayesian computation for SEIR models in R. Using the previously described model parameters, Brown, et al. [3] propose two important modifications in order to incorporate key model components into the spatial framework and as such, the software. Firstly, using the path-specific transition probabilities in equation 6, spatiality is added by replacing the exposure and transition matrices, $\boldsymbol{E}_{ij}^{*}$ and $\boldsymbol{I}_{ij}^{*}$, from equation 2 with the following:

$$\boldsymbol{E}_{ij}^{*} \sim \sum_{l=1}^{m_1} binom\left(\boldsymbol{E}_{ijl}, P(W_1 \leq l + h_i \mid W_1 > l)\right)$$

$$\boldsymbol{I}_{ij}^{*} \sim \sum_{l=1}^{m_2} binom\left(\boldsymbol{I}_{ijl}, P(W_2 \leq l + h_i \mid W_1 > l)\right) \quad (8)$$

$\boldsymbol{E}$ and $\boldsymbol{I}$ are first defined as a $T \times n \times m_1$ and $T \times n \times m_2$ arrays, where m_1 and m_2 represent the maximum times which an individual may remain in a specified latent and infectious states, respectively; h_i is an included temporal offset. $\boldsymbol{E}_{ij}^{*}$ and $\boldsymbol{I}_{ij}^{*}$ therefore represent the (i, j) element of the respective transition matrices, obtained by summation of the exposure and infectious arrays over the corresponding maximum time periods [3].

Secondly, using the exposure probability $\pi_{ij}^{(SE)}$ in equation 3, Brown [34] further specifies each distance measure using an $n \times n$ matrix such that the set $\{\mathbf{D}_z : z = 1, \dots, Z\}$ is defined with corresponding spatial autocorrelation parameters $\{\rho_z\}$; these spatial autocorrelation parameters are constrained such that $\sum_{z=1}^{Z} \rho_z \leq 1$ and $\{0 \leq \rho_z < 1 : z = 1, \dots, Z\}$. The reformulated exposure probability is therefore:

$$\pi_{ij}^{(SE)} = 1 - exp\left(\left\{-\eta_i - \sum_{z=1}^{Z} \rho_z(\mathbf{D}_z \eta_i)\right\}_j^{h_i}\right) \quad (9)$$

Central to this term is the intensity matrix $\boldsymbol{\eta}$. In order to compute this matrix, Brown, et al. [3] first begin by assuming that each location has a time-varying epidemic intensity. This term is structured by associating each location with a $T \times p$ design matrix $\mathbf{X}_j$ where p represents any associated exposure parameters such as intercepts, intervention dates and demographical information. By concatenating each of these $\mathbf{X}_j$ matrices row-wise into a singular $\mathbf{X}^{SE}$ matrix, computation of the $T \times n$ intensity matrix $\boldsymbol{\eta}$ becomes more efficient as it is derived in one step from the Tn by 1 column vector $\mathbf{X}^{SE}\boldsymbol{\beta}^{SE}$; here $\boldsymbol{\beta}^{SE}$ is a shared parameter vector. Under this specification it is therefore impossible to estimate distinctive exposure parameters for each spatial location and time point. Instead, this linear predictor prior structure can be viewed a dimension reduction strategy such that the overall intensity process is still intuitive and flexible [1].

Brown, et al. [3] also alter the SMC-ABC algorithm by introducing three primary modifications. Firstly, a batch size (`batch_size`) $N \geq n$ (requested samples) is used to parallelly conduct simulations and distance evaluations, such that rarity of acceptances under decreasing $\boldsymbol{\epsilon}$ is alleviated. Secondly, by allowing an initial batch size (`init_batch_size`) larger than that of the subsequent sequential step, the algorithm begins at a $\boldsymbol{\epsilon}$, such that runtimes are decreased. Lastly, rather than manual specification of sequenced $\boldsymbol{\epsilon}$ values, the software utilises a distinct '$\boldsymbol{\epsilon}$ schedule' such that $\epsilon_{t+1} = c\epsilon_t$ where $0 < c \leq 1$. Thus, under these modifications, there a two implied modes of convergence: a user-specified number of sampling epochs (`epochs`) or maximum batches (`max_batches`) may be provided. In the latter, for a particular $\boldsymbol{\epsilon}$ value, a maximum N batches are executed before returning to the current sample of n parameter values. A termination acceptance rate (usually selected on the basis of available computational resources) is therefore included, such that termination of sequential $\boldsymbol{\epsilon}$ values are adapted to the current sampling difficulty [3].

For the process of model selection, Brown, et al. [3] highlight the problems associated with traditional methods such as Bayes factor comparisons, under the SMC-ABC algorithm.

Where needed, only analogous results of the algorithm should be utilised. To ensure that this is always the case when comparing models, the software functions under the assumption that models were either "run to the same terminating minimum acceptance rate (i.e. arbitrarily large T, identical n, N), or were forced to run until the same $\boldsymbol{\epsilon}$ threshold was reached [3]." As such, the effects of diffuse priors are attenuated and the potential of running poor models to the same terminating value as reasonable ones is avoided. Under this specification, it is then possible to use the acceptance rates ratio at the next iteration for approximation of the Bayes factor comparing two candidate models (`compareModels`); this is calculated under the assumption that each distribution sequence has converged [3].

Numerical and graphical summaries of model results are produced on the basis of a specified set of model components. `DataModel` describes the relationship between the observed data and the epidemic quantity of interest [3, 34]. For example, in the context of COVID-19, for modelling new infectious counts, this function would be configured by using relevant data regarding new cases (recorded daily, weekly, etc.) and selection of the $\boldsymbol{I}_{ij}^{*}$ compartment by specifying *`compartment="I_star"`*. `ExposureModel` relates to $\pi_{ij}^{(SE)}$, that is, the probability that susceptible individuals will transition to the exposed compartment, and describes factors relating to changes in epidemic intensity [3, 34]. As seen, this is directly correlated with the exposure covariate $\mathbf{X}^{SE}$, where parameters such as intervention measures and vaccinate rates may be specified. Such inclusion not only results in a more realistic model in terms of epidemic intensity, but allows one to determine the possible effect which the covariates had on the epidemic. Covariates, such as temporal basis splines of varying degrees of freedom, offer further versatility in exposure modelling. Through this method of splines, reproduction of flexible curves are made possible as it assumes the structure of a piecewise polynomial structure. The type of polynomial (termed degree) and the number/placement of knots is what then defines the type of spline [35]; the degrees of freedom combine these two numbers. For example, a cubic spline (degree = 3) with 4 knots corresponds to 7 degrees of freedom. This flexibility does come with a trade-off as with an increasing number of knots, overfitting is possible whereas decreasing this number may result in a rigid and restrictive function [35]. Prior parameters for the shared parameter vector $\boldsymbol{\beta}^{SE}$ is also specified under this function. `TransitionPriors` allows for specification of transition processes between the **E**, **I**, and **R** compartments ($\pi_i^{(EI)}$ and $\pi_i^{(IR)}$), thus governing the duration of the latent and

infectious periods of the disease of interest. Two different configurations are currently offered, namely, the exponential compartment membership model, and a path specific generalisation [3, 34]. Such formulation requires relevant prior probabilities and effective sample sizes (ESS), the latter of which is related to the confidence/precision associated with the prior information. In the instance where reinfection is possible, such as with COVID-19, ABSEIR allows users to fit a SEIR(S) model via `ReinfectionModel` by specifying *"SEIRS"*; the *"SEIR"* input returns the SEIR model where there is no reinfection. In terms of the United States, different states may be treated as homogeneous groups, or locations, with heterogeneous mixing between groups. Therefore, for such models which make use of state-level data and hence, describe more than one spatial location, `DistanceModel` is used to introduce spatial dependence through a set of distance matrices $\{\mathbf{D}_z\}$; prior distributions for the spatial autocorrelation parameters $\{\rho_z\}$ are also encoded using this function [3, 34]. `InitialValues` initialises the epidemic model by incorporating the starting population values corresponding to each compartment, namely, $\boldsymbol{S}_0$, $\boldsymbol{E}_0$, $\boldsymbol{I}_0$, and $\boldsymbol{R}_0$. Lastly, `SamplingControl` determines which ABC algorithm (either basic ABC or SMC-ABC) is to be used, and how it is configured/tuned using specifications such as `batch_size`, `init_batch_size, max_batches and epochs`. Once the aforementioned components are successfully created, they are combined and fitted, after which samples are drawn from the posterior distribution using the `SpatialSEIRModel` function [3, 34].

Consideration of the previously discussed model components presents itself as a task rather than a problem. Specifically, substantial deliberation should be used when configuring the components so as to model the disease progression as accurately as possible. Availability of numerous variables makes this task both feasible in a sense but possibly also exhaustive considering the volume of data. Available data regarding COVID-19 parameters is particularly complex due to the sheer number of variables as well as the density of datapoints corresponding to each country/location on a daily basis; as such the pandemic certainly can be viewed as a higher dimensional problem. Furthermore, as with most real-world analysis, missing data, especially with respect to the early developmental stages of the pandemic, presents itself as an issue. This is compounded under the assumption of unreported and asymptomatic cases, such that the actual number of cases may be underestimated. As previously discussed, ABC improves estimation under these circumstances, however, from evaluation of the `SamplingControl` function, it is also clear that ABSEIR is able to utilise other ABC

algorithms than the basic one, hence raising the question of the relative merits of basic ABC. In particular, in cases where a non-informative prior is used SMC-ABC show improved efficiency when compared to basic ABC [31].

CHAPTER 5

SIMULATION STUDY

In order to investigate the highlighted project aims, the previously described methods were each evaluated via a comprehensive simulation study. It was decided to divide the study so as to firstly focus on the non-spatial aspect and then the spatial aspect of the problem. The non-spatial analysis is used to evaluate whether the algorithm and model function well in the simplest case. As such, deductions from this analysis may be beneficial in analysis of the more complex, spatial case. The non-spatial case is centred on the state of Florida whereas the spatial case incorporates two additional states which neighbour Florida, namely Alabama and Georgia, to give a total of three separate spatial locations. Rather than putting the main focus on building a perfectly working model, there is more emphasis on determining which aspects of the methods perform well and diagnostically analysing components which hinder desirable performance.

5.1 Description of data

Being focused on applicability to COVID-19, state-level data regarding aspects such as daily cases and deaths were compiled via the publicly available dataset provided by the New York Times [36] as sourced from state/local governments and health departments. Likewise, vaccination information was compiled via Our World in Data [37] as sourced from the United States Centers for Disease Control and Prevention (CDC). Available population estimates were used as defined by the United States Census Bureau [38]. In particular, it should be noted that these population estimates are for the year 2019 rather than 2020; considering the timeline of the pandemic and the nature of SEIR models, it seemed more sensible to use 2019 data as 2020 estimates (recorded from April 1st to July 1st) may not be an accurate representation of the population size at the beginning of the pandemic (as a result of deaths due to the virus) during early 2020. Due to the reputability of these data sources, there is assumed maintenance of data reliability.

The code used to produce the results and graphs as seen in the subsequent sections can be sourced at the repository https://github.com/2oe-g/dissertation, along with supplemental results and graphs.

5.2 Non-Spatial Analysis

Florida was chosen for the non-spatial analysis due to distinctive features which characterise development in the state's pandemic timeline. In particular, the state currently maintains the seventh highest number of cases per 100,000 residents in the United States. Additionally, the overall lack of substantially restrictive preventative methods, such as state-mandated mask wearing and vaccination mandates, offer an interesting prospective for this analysis; this is further compacted by executively signed orders which prohibit local governments from implementing COVID-19 mitigations.

Simulation Set-up

For the non-spatial case, the best working model was used to evaluate the goals of the simulation which are summarised follows:

1. Determine the effect of a small number of timepoints versus a larger number of timepoints on computational efficiency.

2. Determine the effect of the number of parameters to be estimated on computational efficiency.

3. Evaluate the estimation and computational performance of models with exponentially distributed transition probabilities versus Weibull distributed transition probabilities.

4. Compare the performance of the basic ABC algorithm versus the SMC-ABC algorithm.

In order to achieve the first goal, a smaller epidemic timeline of 77 timepoints starting from the first recorded case on March 1st, 2020 until May 16th, 2020 was compared to a larger timeline of 184 timepoints from the first recorded case until August 31st, 2020. The second goal was evaluated by incorporating additional parameters such as intervention dates and temporal basis splines of varying degrees of freedom. In particular, two intervention dates were implemented into the models, the first of which represents the initial interventive response of state-wide school closure on March 16th, 2020 and the subsequent state-wide stay-at-home order which came into effect on April 2nd, 2020. For the third goal, transition probabilities of the latent and infectious periods for the best working model were modelled separately using the exponential distribution and the Weibull distribution, after which aspects such as computational runtimes, posterior predictive distributions and posterior estimates were compared. Finally, for the fourth goal, considering the discussed conclusions from previous studies, it was first deemed more efficient to use the SMC-ABC algorithm to deduce the best working model. Not only this, but some functionality of ABSEIR is only available when using this algorithm. One important instance of this is usage of the software's built-in function to compare models based on the approximate Bayes factor for model comparison. Once the best working model was established, it was then run again using the basic algorithm for comparison.

Informative priors were used to describe the latent period in all models as this would be done in practice. Using estimates from previous studies [4, 39], the mean length of the latent period was centred at 6 days. Under the exponential transition process, this prior information is encoded using the exposed-to-infectious rate term, $\gamma^{(EI)} \sim \Gamma(1,6)$ with a high precision effective sample size (ESS) of 1000. This also results in a median latent period of approximately 4.16 days and 95% of exposed individuals transitioning to the infectious compartment within approximately 17.97 days. Under the Weibull distribution, gamma

hyperpriors were used to assign an ESS of 1000 to the latent shape and scale parameters such that $\alpha^{lat} \sim \Gamma(shape = 686.3788,\ rate = 313.6212)$ and $\beta^{lat} \sim \Gamma(shape = 867.046,\ rate = 132.954)$. α^{lat} maps the probability of the latent period shape being between approximately 2.08 and 2.29 days and so, the minimum of these values is retained as it provides the broadest possible range for the latent period; in particular, with increasing shape values, the Weibull distribution becomes narrower. Likewise, β^{lat} maps the probability of the scale value between approximately 6.24 and 6.81 days, but here, the maximum of these values is retained as with decreasing scale values, the Weibull distribution becomes narrower. This ensures that $Z_1 \sim Weibull(\alpha^{lat} = 2.08, \beta^{lat} = 6.81)$ represents an approximate mean latent period of 6 days with a median of 5.7 days and maximum membership time of 19.8 days. By approximately 11.5 days, 95% of exposed individuals would have transitioned to the infectious compartment. The specified exponential and Weibull distributions for the latent periods are depicted in Figure 1 and Figure 2, respectively.

Due to increased variability of the infectious period durations (usually ranges from 10 to 24 days) , less informative priors were chosen. Using estimates from a previous study which found the mean time from symptom onset to two negative RT-PCR tests (an indicator of recovery) to be 13.4 days and from symptom onset to hospital discharge or death to be 18.1 days [40], the mean length of the infectious period was centred at 16 days. Decreased confidence in this estimate is encoded with an ESS of 100; under the exponential distribution, $\gamma^{(IR)} \sim \Gamma(1,16)$ which results in a median latent period of approximately 11.09 days and 95% of exposed individuals transitioning to the infectious compartment within approximately 47.93 days (as seen in Figure 1). Under the Weibull distribution, $\alpha^{inf} \sim \Gamma(shape = 81.5249,\ rate = 18.4751)$ (maps probability of infectious period shape between 3.79 and 5.05 days) and $\beta^{inf} \sim \Gamma(shape = 94.06012,\ rate = 5.939876)$ (maps probability of infectious period shape between 13.78 and 17.96 days) hyperpriors are chosen such that $Z_2 \sim Weibull(\alpha^{inf} = 3.79,\ \beta^{inf} = 17.96)$ represents an approximate mean infectious period of 16 days with a median of 16.3 days and maximum membership time of 32.2 days; as seen in Figure 2, by approximately 23.97 days, 95% of infected individuals would have transitioned to the removed compartment.

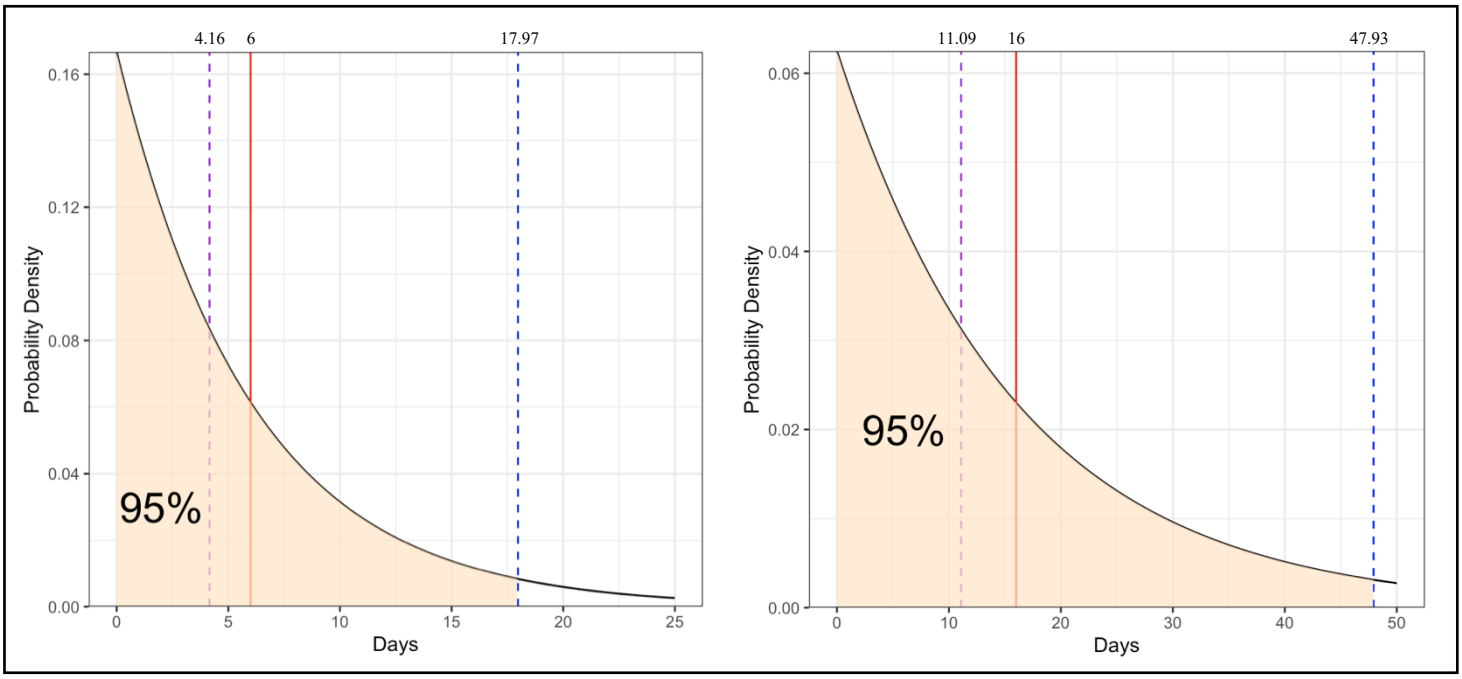

Figure 1: Latent (left) and infectious (right) period under exponential distribution showing median (purple dashed line), mean (red line) and 95% upper bound (blue dashed line)

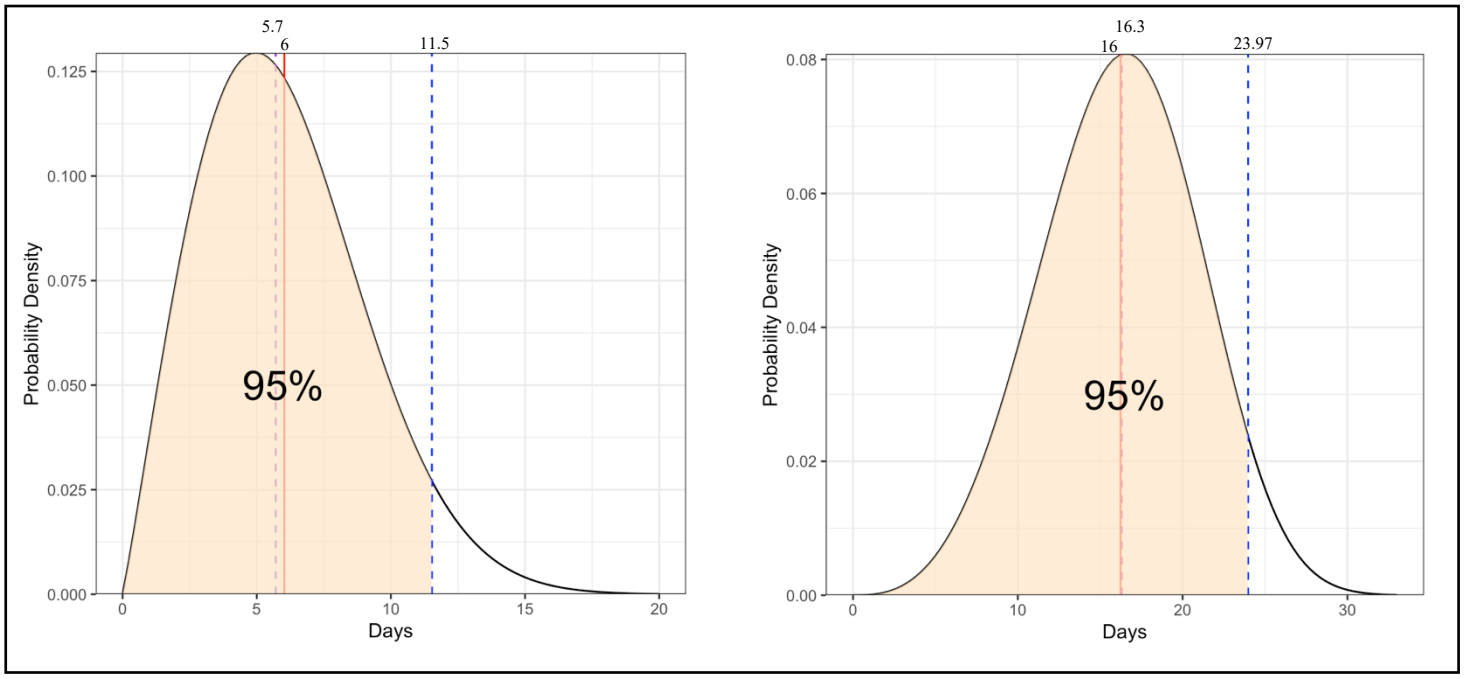

Figure 2: Latent (left) and infectious (right) period under Weibull distribution showing median (purple dashed line), mean (red line) and 95% upper bound (blue dashed line)

As much less is known about the effect of the exposure process parameters, $\beta^{(SE)}$, on epidemic intensity, each parameter was assigned independent $\mathrm{N}(0,2)$ (i.e. prior precision of 0.5) prior distributions. Usage of such priors on the exposure parameters is seen to be supported as with a relatively small variance, more realistic and flexible epidemics are simulated, such that the entire range of probable epidemics is evaluated without assigning considerable prior probability on the extremes of epidemic behaviour [3, 9].

Using the aforementioned prior distributions, 100 parameter samples were generated for each proposed model and used to simulate 5000 epidemics with an initial susceptible population $S_0 = 21{,}477{,}737$ and $I_0 = 2$ infectious individuals. These simulations were conducted under an initial batch size of 1000000, a batch size of 2000, a max batch size of 250, 1000000 epochs and shrinkage of 0.99; shrinkage defines the multiplicative constant by which the maximum distance between simulated and observed epidemics is shrunk between each iteration [34].

Simulation Results

In order to establish the behaviour of SEIR models under ABC, the simplest case was first analysed for the shorter timeline. In particular, emphasis is placed on specifications of the exposure process parameters as these govern predictions of epidemic intensity, this of which ultimately affects the rate at which epidemic growth occurs.

Model 1 is an underspecified model in which exposure was estimated using a single intercept ($\beta_1^{(SE)}$); this term ultimately captures unconstrained epidemic growth as influenced by both pathogen infectivity and population mixing [1]. Models 2 and 3 both parameterised transmission probability using a baseline intensity ($\beta_1^{(SE)}$) as well as intervention terms associated with a piecewise linear covariate, which is equal to zero up until the time interventions began, at which point it becomes linear in time. More specifically, model 2 only includes the first intervention ($\beta_2^{(SE)}$) whereas model 3 includes both ($\beta_2^{(SE)}$, $\beta_3^{(SE)}$). Model 3

(now termed Model 4 for identifiability and distinction from the model under the shorter timeline) was ultimately used for initial diagnostic analysis under the longer timeline and demonstration of how temporal basis splines (Models 5-8) may improve estimation.

Results of simulations for the first timeline are presented in Figure 3. From first glance, it is clear that the underspecified model performs poorly in estimating new cases; this is because the model assumes (and predicts) a constant mean intensity process ($\boldsymbol{\eta}$) of approximately $\beta_1^{(SE)} = -1.29$. At the beginning of the pandemic, intuitively there are a small number of infected individuals such that the probability of exposure (recall that this probability is linked to the intensity process) is relatively small. However, as more persons come into contact with infected individuals over time, under a constant $\boldsymbol{\eta}$, that probability grows such that the overall number of new cases grows exponentially, as more suspectable persons become exposed and hence, infected. To prove this, recall the more general form of the transition probability for the non-spatial case, $\pi_i^{(SE)}$, from equation 4:

$$\pi_i^{(SE)} = 1 - \exp\left(-e^{\boldsymbol{\eta}_i}\frac{I_i}{N}\right)$$

Now, consider two differing timepoints, say t = 5 and t = 70, for easier comparability. From the simulated data, it is also possible to infer that the mean number of infectious persons at these times are estimated to be $I_5 = 3$ and $I_{70} = 1770$. Then, it is clear that:

$$\pi_5^{(SE)} = 1 - \exp\left(-e^{-1.29}\frac{3}{21477737}\right) \approx 3.844969e^{-8}$$

and:

$$\pi_{70}^{(SE)} = 1 - \exp\left(-e^{-1.29}\frac{1770}{21477737}\right) \approx 2.268506e^{-5}$$

such that $\pi_{70}^{(SE)} > \pi_5^{(SE)}$

■

As such, public health efforts are aimed at avoiding such a constant, or even increasing, intensity process as it produces an uncontrolled epidemic outbreak.

In comparison , models 2 and 3 perform much better as they account for variation in the epidemic intensity via inclusion of intervention terms. Both models predict a decrease in epidemic intensity after the first intervention (day 16), with model 2 estimating a posterior mean intensity of $\beta_2^{(SE)} = -0.897$ and model 3 estimating $\beta_2^{(SE)} = -1.490$; this difference is apparent as the decreasing slope in the intensity prediction plot is much steeper under model 3. This then possibly implies that this interventive method of state-wide school closure had a positive impact, such that there was decrease in disease spread. Upon inspecting the posterior predictive distributions, it is clear that both models flatten the curve and thus, avoid substantial exponential growth. This follows the same notion that with decreasing intensity, the probability of exposure also decreases. However, it is also apparent that model 3 is a much better fit of the data compared to model 2; more specifically, model 2 tends to underestimate the number of new cases as it lacks the adequate amount of epidemic intensity due to the constant rate at which it decreases after the first intervention.

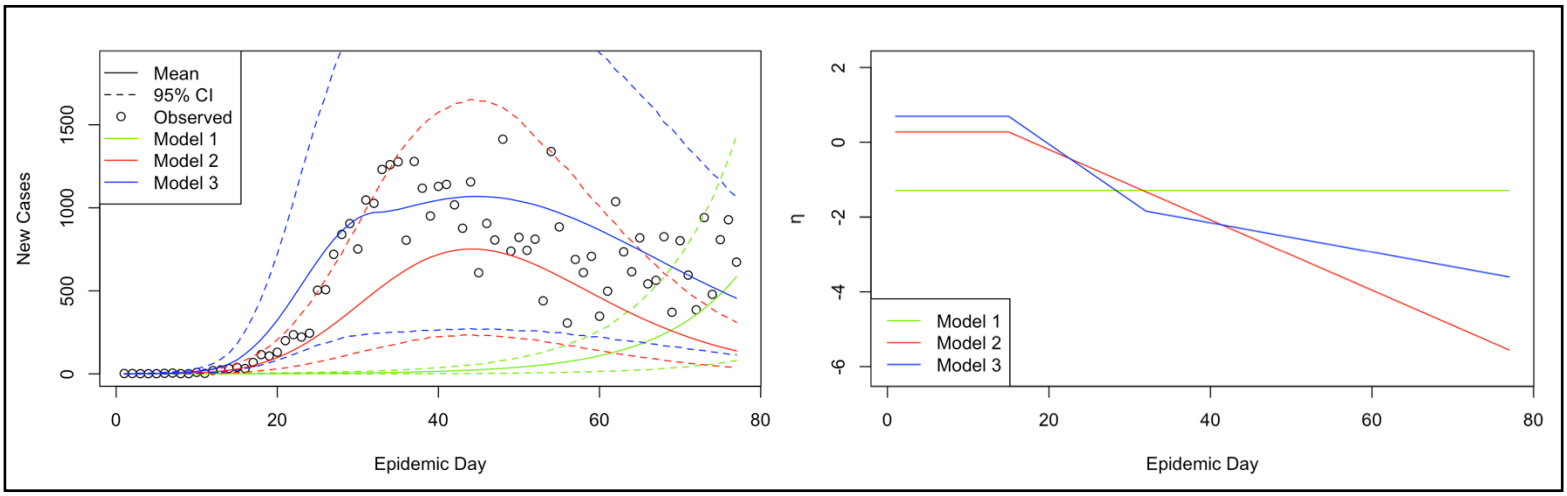

Figure 3: Posterior predictive distribution and intensity prediction for first timeline

The difference in estimation performance can therefore be attributed to the inclusion of the second intervention (day 33) term in model 3, which in turn, predicts that the state mandated stay-at-home order saw a posterior mean intensity of $\beta_3^{(SE)} = 1.098$, and thus, possibly

resulted in increased disease spread. Intuitively, this seems odd considering intervention methods are aimed at decreasing spread. However, it should be emphasised that the true effect of a single intervention may not be detectable due to other factors occurring in the population which may also affect epidemic intensity. While the stay-at-home order was imposed on April 1st, 2020, there was also the reopening of a number of beaches soon after on April 17th [41]. Hence, it is possible that this resulted in increased contact and spread rather than the stay-at-home order. Ultimately, the importance of including accurate and relevant intervention information is seen to be essential for reasonable estimations, as ignoring such details can bias intensity parameters [1]. In particular, the intensity intercept estimate ($\beta_1^{(SE)}$) for model 1 (-1.29) implies a less infectious epidemic process than that of models 2 (0.055) and 3 (0.692), which account for changes in intensity as a result of interventions.

Using model 3, further simulations were run using the basic ABC algorithm, one run using the exponential distribution for the transition periods and the other using the Weibull distribution; model 3 was also run again under the SMC-ABC algorithm using the Weibull distribution. From Figure 4, the overall difference in the posterior predictive distributions do not appear to be significant, with both algorithms and transition distributions performing well in terms of estimation. Notably, the Weibull distribution estimations present a different behaviour in terms of the overall shape of the curve, as seen in the small decrease of predicted cases just before day 40. Additionally, when examining the intensity prediction plot, there is a notable difference under the SMC Weibull estimation, such that the predicted intensity after the second intervention is seen to be higher when compared to the other predictions. This increased intensity is reflected in the posterior estimates such that the SMC Weibull curve predicts a higher amount of new infectious cases at the end of the timeline when compared to the other curves.

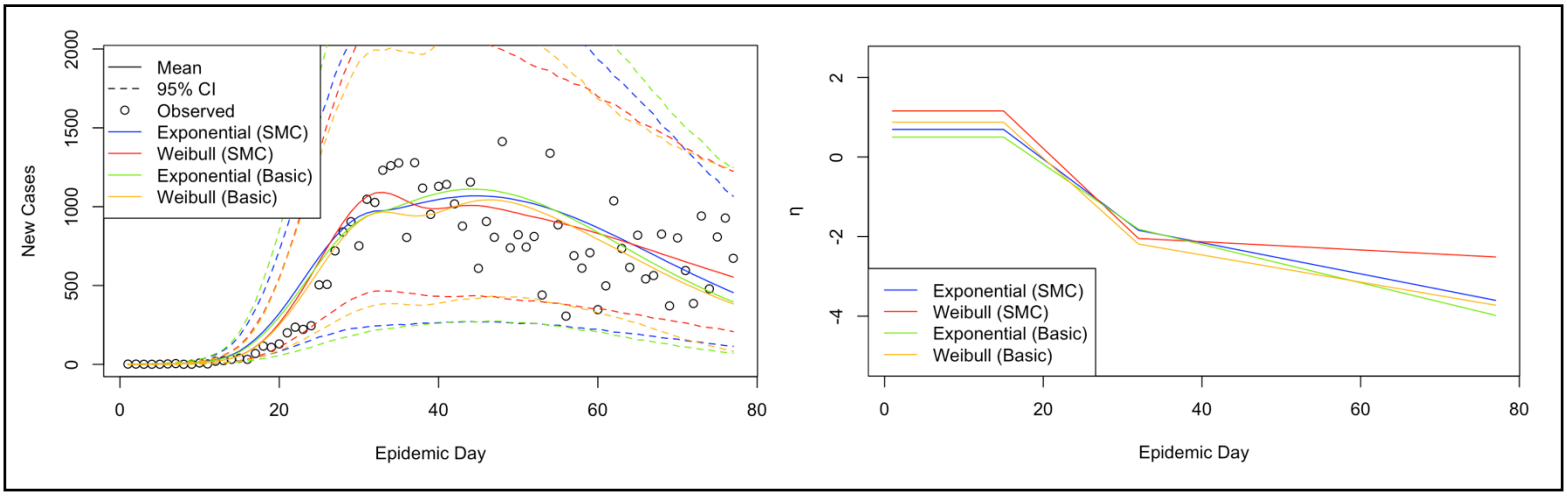

Figure 4: Posterior predictive distribution and intensity prediction for first timeline under Model 3

When considering the reason as to why the curves of the Weibull distribution estimations appear differently to those of the exponential distribution estimations, it is important to recall the properties of each distribution. In particular, the exponential distribution assumes that there is a constant probability of transitioning from one compartment to the next, regardless of membership time. This is not a realistic representation of most infectious disease processes as it is expected that the longer an individual has been infectious, the more likely they are to be removed (by recovery or death). This is improved upon via the Weibull distribution which assumes that transition probabilities correspond to the length of time an individual has spent in a respective compartment [9]. Additionally, under the specified transition priors, though centred at the same mean, the exponential distribution assumes that 95% of infected individuals transition to the removed compartment within 47.93 days whereas the Weibull distribution is within 23.97 days. As seen in Figure 5, the exponential distribution allows for substantially longer membership times for the infected compartment. The relative proportion of infected individuals under the exponential distribution is substantially larger than that of the Weibull distribution. This is because infected individuals may transition to the removed compartment at a much quicker rate under the Weibull distribution compared to the exponential one. It is therefore easy to understand how this would affect estimates of the transition probabilities; with an increased proportion of infected individuals in the population, there is a decrease in the transition probability such that the number of new infections decreases.

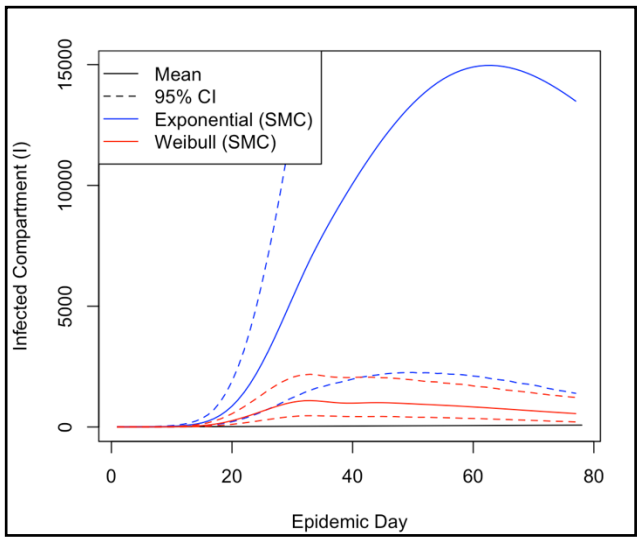

Figure 5: Posterior predictive distribution of infectious compartment under Model 3

Using the approximate Bayes factors presented in Table 1, it is apparent that model 3 under the Weibull distribution results in factors uniformly greater than 1 and strongly so, such that this distribution performs the best while incorporating two interventions for this timeline. For further analysis, the posterior parameter estimates for model 3 under the SMC-ABC algorithm are presented in Table 2, for both the exponential and Weibull distribution estimations.

Table 1: Approximate Bayes factors for Model 2, Model 3 (Exp) and Model 3 (Weibull) (row versus column index)

	Model 2	Model 3 (Exp)	Model 3 (Weibull)
Model 2	NaN	0.00	0.00
Model 3 (Exp)	Inf	1.00	0.05
Model 3 (Weibull)	Inf	19.47	1.00

Table 2: SMC-ABC posterior parameter summary for Model 3

	Mean		**SD**		**95% LB**		**95% UB**	
	Exponential	Weibull	Exponential	Weibull	Exponential	Weibull	Exponential	Weibull
$\beta_1^{(SE)}$	0.692	1.157	0.218	0.197	0.309	0.830	1.132	1.574
$\beta_2^{(SE)}$	-1.490	-1.885	0.141	0.141	-1.741	-2.166	-1.207	-1.618
$\beta_3^{(SE)}$	1.098	1.781	0.134	0.148	0.850	1.555	1.349	2.094
$\gamma^{(\mathrm{EI})}$	0.151	-	0.039	-	0.073	-	0.224	-
$\gamma^{(\mathrm{IR})}$	0.085	-	0.056	-	0.009	-	0.212	-
α^{lat}	-	2.619	-	0.747	-	1.304	-	4.285
β^{lat}	-	7.296	-	1.131	-	5.352	-	9.286
α^{inf}	-	8.738	-	4.780	-	2.084	-	20.911
β^{inf}	-	8.636	-	2.407	-	4.354	-	13.117

From Table 2, the Weibull distribution parameterisation results in a more infectious epidemic process ($\beta_1^{(SE)}$) compared to that of the exponential distribution. Both parameterisations predict a negative posterior mean intensity associated with the first intervention ($\beta_2^{(SE)}$), with the Weibull estimate inferring a greater decrease in epidemic intensity; likewise, the second intervention ($\beta_3^{(SE)}$) is estimated to have had a greater increase (though not necessarily attributed to this intervention, as previously explained) in epidemic intensity when compared to the exponential estimate. The exponential distribution parameterisation estimates a posterior mean latent period of $\frac{1}{\gamma^{(\mathrm{EI})}} = \frac{1}{0.151} \approx 6.6$ days with a 95% credible interval of (4.5, 13.7) and a posterior mean infectious period of $\frac{1}{\gamma^{(\mathrm{IR})}} = \frac{1}{0.085} \approx 11.8$ days with a 95% credible interval of (4.7, 11.1). The Weibull distribution parameterisation estimates a posterior mean latent period of $\beta^{\mathrm{lat}}\,\Gamma\left(1+\frac{1}{\alpha^{\mathrm{lat}}}\right) = 7.296\Gamma\left(1+\frac{1}{2.619}\right) \approx 6.5$ days with a 95% credible interval of (4.9, 8.5) and a posterior mean infectious period of $\beta^{\mathrm{inf}}\,\Gamma\left(1+\frac{1}{\alpha^{\mathrm{inf}}}\right) = 8.738\Gamma\left(1+\frac{1}{8.636}\right) \approx 8.2$ days with a 95% credible interval of (3.9, 12.8). Both parameterisations therefore estimate a very similar mean latent period which is similar to the proposed prior mean. On the other hand, the estimated mean infectious period is substantially different between both parameterisations; in this case, considering the aforementioned approximate Bayes factor estimates, it is therefore more likely that the true mean infectious period (for this timeline) is around 8 days as opposed to the proposed prior mean of 16 days.

When considering the longer timeline, the same specifications of model 3 are retained and used under the new name of model 4. From Figure 6, it is clear that the model does a relatively acceptable job of identifying the peak corresponding to the second wave of cases but does poorly at detecting the first wave. This is because with the longer timeline, there is much more variation in epidemic intensity taking place, such that the current model is unable to accurately predict such changes. In particular, when observing the intensity prediction plot, there is a predicted increase in intensity corresponding to implementation of the first intervention and a constant decrease after the second intervention. The model tries to compensate for this under-specification by attempting to predict an intensity process which captures the drastic difference in new cases between the first and second waves, however, the rate at which predicted intensity rapidly declines results in a much less intense epidemic process than the observed cases. Therefore, with such a rigid specification for the exposure process parameters, it becomes increasingly difficult to obtain reliable estimates when more complex epidemic factors are taking place in the population.

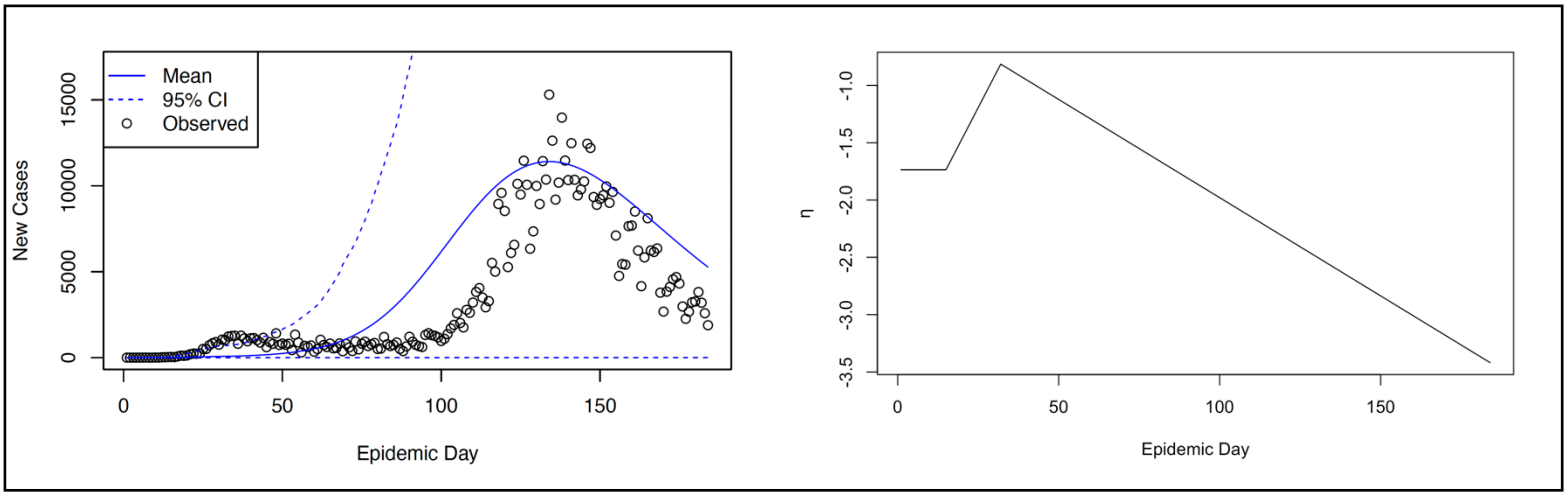

Figure 6: Posterior predictive distribution and intensity prediction for second timeline under Model 4

As proposed by several papers [1, 3, 9, 10], temporal basis expansion techniques, such as basis splines and trigonometric functions, aid in improving estimation by allowing for construction of a more flexible model, even in the absence of structural information about exhibited, complex epidemic behaviour. Following this notion, additional exposure process parameters were included in the model by using varying degrees of freedom; starting at 3 degrees of freedom under model 5, with an increasing basis up until 6 degrees of freedom under model 8. Using the approximate Bayes factors presented in Table 3, model 6, this of which corresponds to 4 degrees of freedom, is the only model which produced factors greater than 1

and so, is retained as the best model; as done in the first timeline, more simulations were run using this model under the different algorithms and transition distributions.

Table 3: Approximate Bayes factors for Models 4-8 (row versus column index)

	Model 4	Model 5	Model 6	Model 7	Model 8
Model 4	1.00	0.80	0.77	0.91	1.15
Model 5	1.25	1.00	0.97	1.14	1.43
Model 6	1.29	1.03	1.00	1.18	1.48
Model 7	1.10	0.88	0.85	1.00	1.26
Model 8	0.87	0.70	0.67	0.79	1.00

From Figure 7, it is clear that inclusion of this splines basis does in fact improve estimation such that both algorithms and transition distributions are able to better predict new cases during the first wave (as the observed cases are contained within the upper bound of the respective 95% CI). Notably, the SMC algorithm is seen to predict a greater mean number of new cases corresponding to the second peak, when compared to the basic algorithm. This is because the predicted intensity under the basic algorithm decreases at a much slower rate after the second intervention date. In comparison, the SMC algorithm predicts a relative increase in epidemic intensity after this date, after which there is exponential decay. As such, the SMC algorithm is able to predict a much more flexible intensity process such that a more complex epidemic situation is represented.

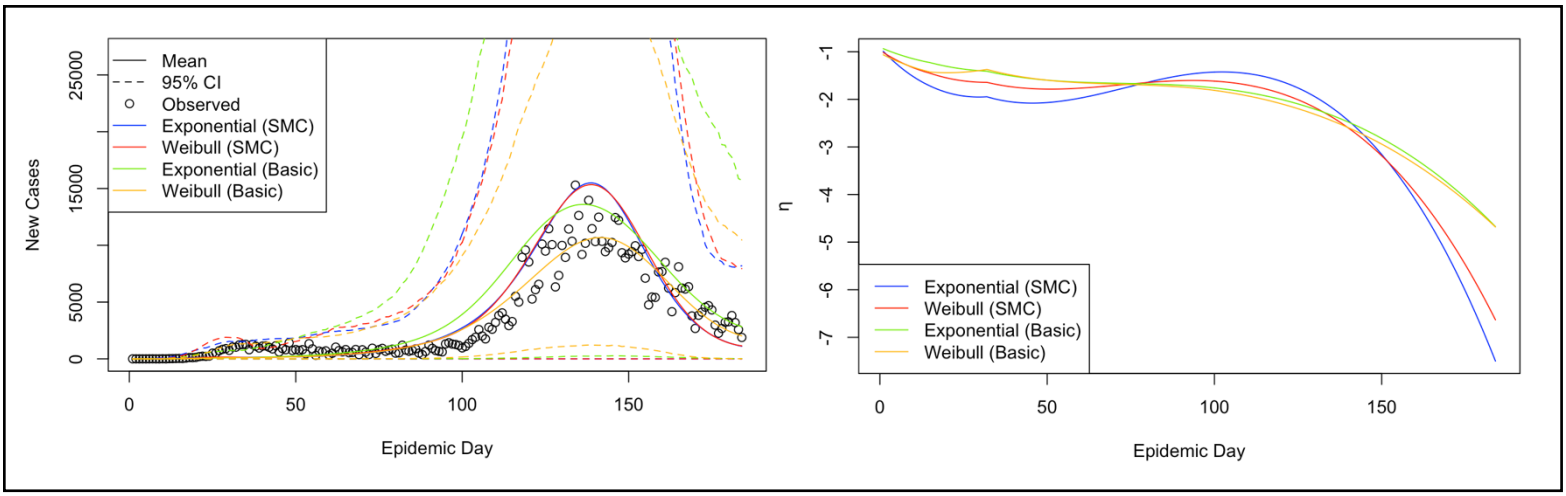

Figure 7: Posterior predictive distribution and intensity prediction for second timeline under Model 6

However, as a rule of caution, it should be noted that interpretations of intensity estimates for specific interventions should generally be avoided. Due to the extrapolative nature of splines, exact estimates of the exposure process parameters may not be as reliable [1], however, their inclusion still allows for overall deductions to be made on epidemic behaviour. In particular, rather than attributing any changes in epidemic intensity to a specific/defined intervention, it is safer to generalise inferences from the posterior predictive distribution. For example, after the second intervention, observed cases appear to remain rather constant/show a slight decrease, up until around day 90; it is then possible to infer that some interventions (not specific/limited to ones included in the model) taking place in the population could have accounted for this. One case of this can be seen where there is increased public awareness of the disease such that persons take more personal preventative methods, such as mask wearing and social distancing, even though such mandates may not be in place. On the other hand, it is intuitive to think that the drastic increase in cases around day 100 would have resulted due to possible removal of interventions, such as expirations of any mandated orders; this is in fact plausible as the state-mandated stay-at-home order expired on May 4th, 2020 (day 65), after which 'Full Phase 1 Reopening' was implemented on May 18th, 2020. This allowed for some facilities, such as professional sports venues, to operate at full capacity whereas others, such as indoor restaurants and gyms, were allowed 50% capacity [41]. Notably, as this analysis only considers overall cases at the state level and therefore, only considers state-wide intervention methods, it is important to remember that interventive methods at the county level were also put in place. For example, several counties such as Broward county and Fort Lauderdale, issued county-level mask mandates from around mid-April, which required individuals to wear masks in public spaces [42]. As such, with the inability to feasibly include all these interventive methods, there is under-specification of the intensity process such that epidemic infectiousness is underestimated; the true epidemic intensity process is seen to be a cumulative result of state-level, county-level and personal-level interventive methods. With this in mind, interpretations of posterior parameters (excluding exposure process estimates) for model 6 with Weibull parameterised transition processes (this of which gives greater approximate Bayes factors as seen in Table 4) under the basic and SMC algorithm are presented in Table 5 for further analysis.

Table 4: Approximate Bayes factors for Model 6; Exponential and Weibull (row versus column index)

	Exponential	Weibull
Exponential	1.00	0.98
Weibull	1.02	1.00

Table 5: SMC-ABC posterior parameter summary and estimation performance for Model 6 under Weibull parameterisation

	Mean		**SD**		**95% LB**		**95% UB**	
	Basic	SMC	Basic	SMC	Basic	SMC	Basic	SMC
α^{lat}	2.198	2.176	0.076	0.136	2.064	1.937	2.368	2.416
β^{lat}	6.518	6.607	0.197	0.384	6.183	5.857	6.868	7.244
α^{inf}	4.358	4.384	0.450	0.841	3.506	2.946	5.251	6.392
β^{inf}	15.870	15.940	1.500	1.902	13.014	12.019	18.922	19.425

From Table 5, the basic ABC algorithm estimates a posterior mean latent period of $\beta^{lat}\,\Gamma\left(1+\frac{1}{\alpha^{lat}}\right) = 6.518\Gamma\left(1+\frac{1}{2.198}\right) \approx 5.8$ days with a 95% credible interval of (5.5, 6.1) and a posterior mean infectious period of $\beta^{inf}\,\Gamma\left(1+\frac{1}{\alpha^{inf}}\right) = 15.870\Gamma\left(1+\frac{1}{4.358}\right) \approx 14.5$ days with a 95% credible interval of (11.7, 17.4). On the other hand, the SMC-ABC algorithm estimates a posterior mean latent period of $\beta^{lat}\,\Gamma\left(1+\frac{1}{\alpha^{lat}}\right) = 6.607\Gamma\left(1+\frac{1}{2.176}\right) \approx 5.9$ days with a 95% credible interval of (5.2, 6.4) and a posterior mean infectious period of $\beta^{inf}\,\Gamma\left(1+\frac{1}{\alpha^{inf}}\right) = 15.940\Gamma\left(1+\frac{1}{4.384}\right) \approx 14.5$ days with a 95% credible interval of (10.7, 18.1). Both algorithms therefore estimate similar mean latent periods which are close to the proposed prior mean of approximately 6 days. Interestingly, both algorithms similarly predict the same mean infectious period, with the proposed prior mean of approximately 16 days being contained within both of their predicted 95% credible intervals. Inherently, there therefore appears to be not much of a significant difference between estimation performance in terms of both algorithms in this case.

5.3 Spatial Analysis

Simulation Set-up

Using inferences gained from the non-spatial analysis, further considerations are now investigated in this spatial analysis. In particular, since spatiality is introduced, different methods of representing this component are explored between three neighbouring states: Florida, Alabama and Georgia. The same timeframes as the non-spatial case are used and the same models, along relatively the same intuitive approach, with the same specified priors are maintained under the assumption that they are sufficient for this analysis; though this notion may be flawed, conceptually it aids in making the choice of models substantially easier as the models were fine tuned for Florida (which is also included in this analysis). There are two main changes to the models: firstly, rather than one overall intercept which represents the baseline intensity ($\beta_1^{(SE)}$) , each of the three states are assumed to have a separate and constant intensity value up until the start of interventions; $\beta_1^{(SE)}$ therefore corresponds to Alabama, $\beta_2^{(SE)}$ to Florida and $\beta_3^{(SE)}$ to Georgia. Additionally, there is inclusion of a distance (spatial) component which is modelled under two overall assumptions, the first of which assumes that contact intensity/rates arising from the neighbouring states is the same (CAR or Conditionally Autoregressive model), and the other which assumes that there are different contact intensities/rates (distance model and gravity model).

Under the CAR model, the presiding assumption is that epidemic spread between states (spatial locations) depends on a single overall spatial correlation term which is represented by inclusion of a simple distance matrix such that all spatial locations are assigned a 'neighbourhood indicator'; simply put, neighbouring states (i.e. those that border each other) are encoded with 1 whereas non-neighbouring states are encoded with 0 as seen below:

$$\mathbf{D}_z = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

such that the first row/column represents Alabama, the second represents Florida and the third represents Georgia. Contrarily, in order to determine any potential difference in cross-border spread, each set of states was also assigned a separate spatial autocorrelation parameter; simply termed 'distance model', this is encoded by specifying three separate distance matrices such that the relationship between pairs of bordering states is represented by using the same 'neighbourhood indicator'. For example, consider the first relationship between Alabama and Florida as encoded below:

$$\mathbf{D}_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

A more traditional approach is seen via the gravity model which allows for continuous spatial dependence via inclusion of more specific distance metrics, such as city centroids, between discrete point locations; in this analysis, the weighted distance is used and derived via the *usaww* matrix of the *splm* package. Using the respective population sizes for each state, it is then assumed that the contact process between each pair of states is proportional to their populations divided by the squared 'distance'.

Spatial autocorrelation terms, ρ_z, were each assigned Beta(1,10) priors such that the contribution arising from contact between states reflects less than approximately 23% of contact intensity occurring within states; such a constraint is relatively conservative yet realistic considering that there are generally no restrictions regarding state-to-state movement.

Simulation Results

Using the two-intervention model from the non-spatial case, each state was assigned corresponding exposure process parameters via intervention dates; in particular, both Alabama and Georgia implemented state-wide school closure on March 18th, 2020 whereas a state-wide stay-at-home order was mandated on April 4th, 2020 in Alabama and on April 2nd, 2020 in Georgia. Additionally, the aforementioned distance matrices were incorporated to give models

1a, 1b and 1c, representing the distance, CAR and gravity models, respectively. Using the approximate Bayes factors presented in Table 6, it is apparent that there is a strong preference of the distance model over the other two models, thus implying that the contact rates between states is possibly a dynamic process rather than a constant one; model 1a was therefore retained for further analysis.

Table 6: Approximate Bayes factors for Model 1a, Model 1b and Model 1c
(row versus column index)

	Model 1a	Model 1b	Model 1c
Model 1a	1.00	16.46	9.35
Model 1b	0.06	1.00	1.02
Model 1c	0.72	0.98	1.00

From Figure 8, it is evident that the basic algorithm performs very poorly; this is initially unexpected considering the estimation performance observed in the non-spatial case. However, as explained by several authors, sampling efficiency/performance tends to decrease in situations where diffuse prior distributions are present, with such problems being especially true for high dimensional cases [3, 28]. Under the premise of a spatial analysis, the dimensionality of the problem does indeed increase as the algorithm has to sample and predict estimates for each location as opposed to just one location in the non-spatial case.

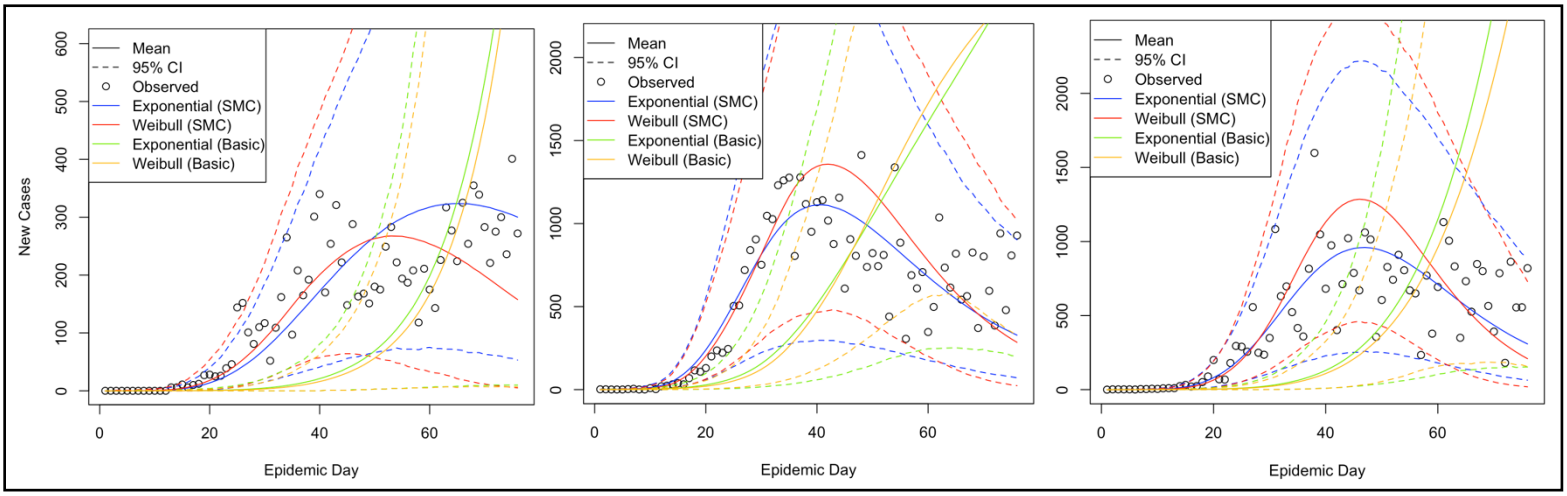

Figure 8: Posterior predictive distribution for first timeline under Model 1a;
Alabama (left), Florida (middle) and Georgia (right)

There is also a higher possibility of the priors being diffuse with respect to the posterior, such that proposals from the prior distribution are not contained within high posterior density regions with relative frequency. With more than one location, it becomes more likely that the specified prior means of the latent and infectious periods do not accurately represent the actual period lengths as observed in all of these three particular states. Notably, since the basic algorithm performed rather well in the non-spatial analysis of Florida, there is possible implication that the provided priors were sufficient. However, with different locations, this may not be the case. As showed by Cheng, et al. [4], the mean latent period of COVID-19 varies with age, as observed both in a global meta-analysis of conducted studies, and their own observational study. In particular, it was noted that middle-aged persons (41-60 years) had the shortest incubation period. Drawing from this, it is then possible that one state had a much greater demographic composition of one age group compared to others, such that the mean periods are variable from state to state.

As seen in Figure 9, the SMC-ABC algorithm models predict the highest intensity process in Florida whereas Alabama is predicted to have the lowest. This is expected once considering the drastic difference in the scale of the observed cases in Alabama compared to Florida. The Weibull parameterisation tends to underestimate the number of cases approaching the end of the timeline due to the rapid decrease in predicted intensity, especially in Alabama. The exponential parameterisation performs much better such that the change in intensity is able to adequately maintain sufficient epidemic infectiousness. This is also supported by the approximate Bayes factors presented in Table 7 below; a posterior parameter summary is provided in Table 8 for further analysis.

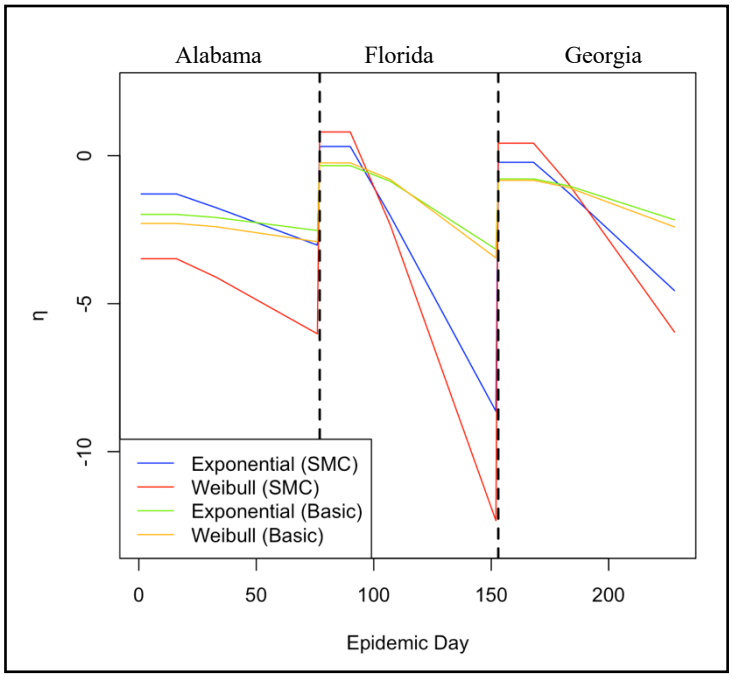

Figure 9: Intensity prediction for first timeline under Model 1a

Table 7: Approximate Bayes factors for Model 1a; Exponential and Weibull (row versus column index)

	Exponential	Weibull
Exponential	1.00	3.12
Weibull	0.32	1.00

Table 8: SMC-ABC posterior parameter summary for Model 1a (Exponential)

	Mean	**SD**	**95% LB**	**95% UB**
$\beta_1^{(SE)}$	-1.292	0.300	-1.971	-0.743
$\beta_2^{(SE)}$	0.310	0.169	-0.001	0.667
$\beta_3^{(SE)}$	-0.223	0.165	-0.539	0.135
$\beta_4^{(SE)}$	-13.740	2.247	-17.908	-9.854
$\beta_5^{(SE)}$	-0.951	4.690	-9.492	8.177
ρ_1	0.158	0.098	0.022	0.388
ρ_2	0.164	0.090	0.019	0.321
ρ_3	0.340	0.067	0.209	0.463
$\gamma^{(EI)}$	0.199	0.048	0.125	0.289
$\gamma^{(IR)}$	0.062	0.032	0.010	0.120

Although there is some overlap in the posterior credible regions for the three state-specific intercepts, as seen in Table 8, the SMC-ABC algorithm under exponential parameterisation estimates Florida to have the highest mean epidemic potential of 0.310 followed by Georgia with -0.223 and lastly Alabama with -1.292. Due to the form of the intensity matrix, it becomes infeasible to estimate a distinct parameter for each spatial location and time point and so, the intensity process ultimately takes on the form of a linear predictor prior structure [1]; as such, exposure process estimates corresponding to the first and second interventions are seen to be bigger in magnitude when compared to the non-spatial analysis. Therefore, in a generalised manner, it can be assumed that the first intervention resulted in a greater decrease in epidemic infectiousness when compared to the second intervention, across all three states. The three included spatial parameters are seen to be distinctly non-zero, implying cross-border spread; in particular, the estimated contact rate between Florida and Georgia (ρ_3) is seen to be significantly larger than that of Alabama and Florida (ρ_1), and Alabama and Georgia (ρ_2). As substantially lower cross-border spread is predicted for both borders with Alabama, it is possible to infer the existence of heterogeneity. In the context of COVID-19, Großmann, et al. [43] highlights super-spreader events and different individual levels of infectiousness as examples of such heterogeneous factors. Lastly, the posterior mean of the latent period is estimated to be approximately 5 days with a 95% credible interval of (3.5, 8.0) whereas the posterior mean of the infectious period is estimated to be approximately 16.1 days with a 95% credible interval of (8.3, 100). The former is acceptable considering the fact that the proposed prior latent period mean of 6 days lies within the 95% credible posterior interval; the latter is of concern considering the upper bound estimate of 100 days. When theorising why such a broad infectious period would have been estimated, it is certainly possible that the model accounts for the difference in the shape of the distributions as seen in the posterior predictive curves. In particular, Alabama is observed to have a relatively different shape compared to Florida and Georgia such that there is yet to have the development of a distinctive ‘first wave’. As such, new cases in Alabama are predicted to be decreasing at a slower rate towards the end of the timeline as compared to the other two states; this is also inferred by the intensity prediction plots such that the gradient of decrease is much less in Alabama. As a result, there is much more uncertainty about the infectious period length. This notion is explored further using the second timeline.

Compared to a temporal spline basis of 4 degrees of freedom, which was deemed to fit the second timeline the best in the non-spatial analysis, it was found that the use of 3 degrees of freedom provided a better fit to all three spatial locations. Following the aforementioned theory, it is now clear that Alabama does indeed only exhibit one distinct peak/wave whereas the other two states characteristically exhibit two, as seen in Figure 10.

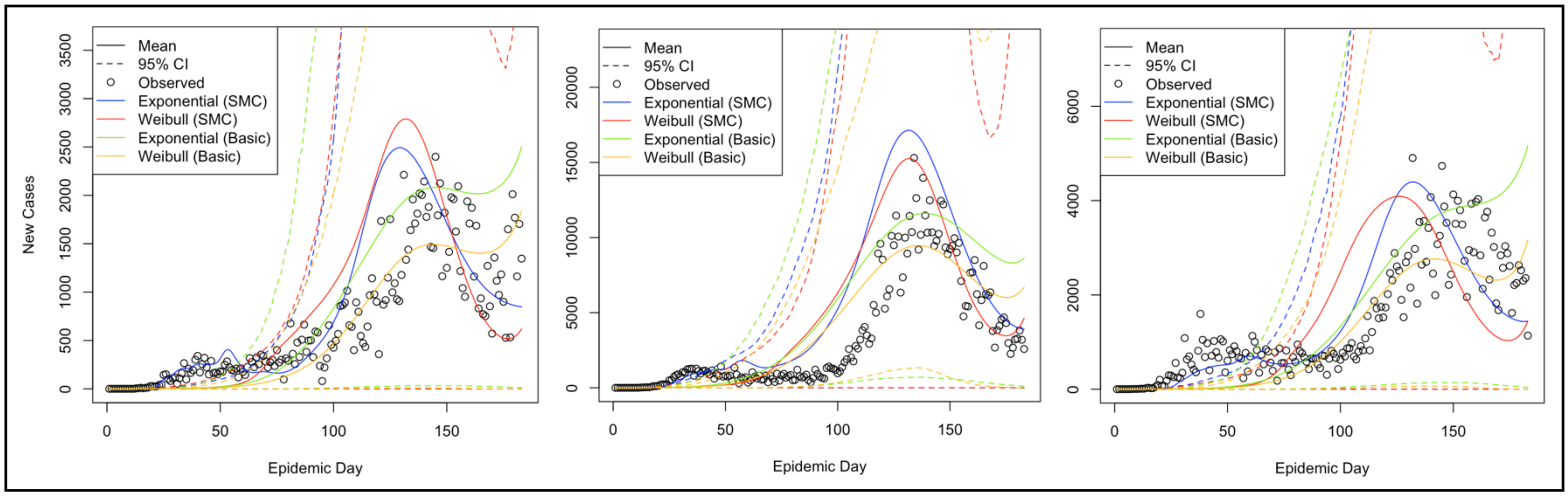

Figure 10: Posterior predictive distribution for first timeline under Model 2a;
Alabama (left), Florida (middle) and Georgia (right)

It is apparent that the model fails to detect the first wave in Florida and Georgia, despite the inclusion of splines. This follows the same notion proposed in the non-spatial case whereby model 4 attempts to predict an intensity process which captures the drastic difference in new cases between the first and second waves, however, the rate at which predicted intensity rapidly declines results in a much less intense epidemic process than the observed cases. Now, under consideration of the different spatial locations, it is possible that the model attempts to compensate for the substantial difference in the scale of new cases, such that estimation performance is hindered; in particular, the highest observed value for Alabama is seen to be around 2400 new cases whereas Florida sees a value of around 15,000. When examining the intensity prediction in Figure 11, a much more negative baseline epidemic intensity parameter is observed for Florida and Georgia, when compared to what was predicted under model 1a. As explained by Brown [1], as intensity parameters become increasingly negative, there are fewer adequate epidemic simulations as a result of lower epidemic activity, ultimately deeming

intervention terms to be "not meaningfully estimable in cases where the epidemic dies out before the hypothetical intervention occurs."

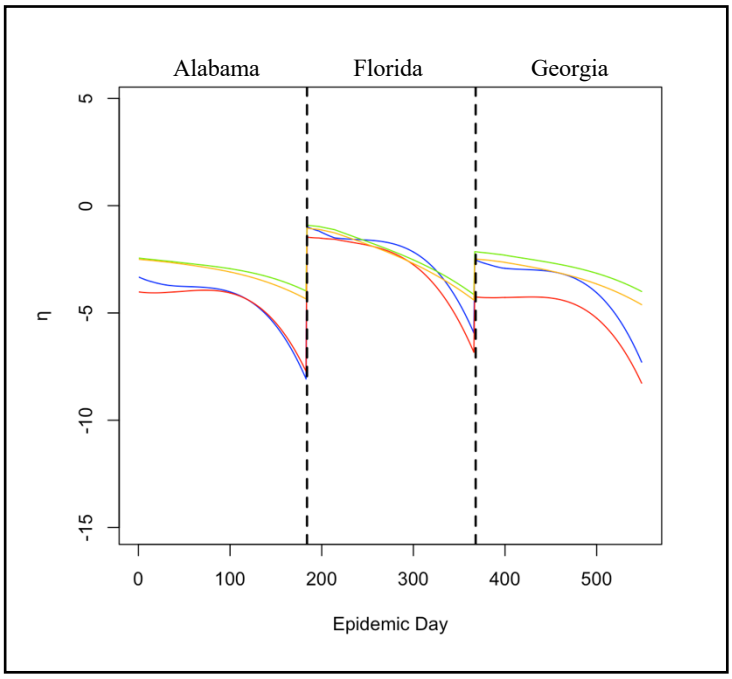

Figure 11: Intensity prediction for first timeline under Model 2a

From the approximate Bayes factors presented in Table 9 below, the model 2a under the Weibull parameterisation is preferred over the exponential parameterisation; for completeness, its posterior parameter summary (exposure process parameters excluded due to previously explained reasons) is provided in Table 10.

Table 9: Approximate Bayes factors for Model 2a; Exponential and Weibull (row versus column index)

	Exponential	Weibull
Exponential	1.00	0.62
Weibull	1.61	1.00

Table 10: SMC-ABC posterior parameter summary for Model 2a (Weibull)

	Mean	**SD**	**95% LB**	**95% UB**
ρ_1	0.334	0.163	0.023	0.614
ρ_2	0.150	0.103	0.005	0.365
ρ_3	0.263	0.142	0.033	0.547
α^{lat}	2.139	0.252	1.627	2.595
β^{lat}	6.829	0.751	5.396	8.167
α^{inf}	4.959	1.826	1.298	8.169
β^{inf}	16.933	4.818	9.094	25.779

From Table 10, the posterior mean latent period is estimated to be approximately 6.1 days with a 95% credible interval of (4.8, 7.3) whereas the posterior mean infectious period is estimated to be approximately 15.5 days with a 95% credible interval of (8.4, 24.3). Again, the three included spatial parameters are seen to be distinctly non-zero, implying cross-border spread. However, compared to the shorter timeline, the estimated contact rate between Alabama and Florida (ρ_1), is seen to be significantly larger than that of Alabama and Georgia (ρ_2) and Florida and Georgia (ρ_3). In particular, there is now a shift such that the estimated contact intensity between Florida and the two other states are the highest whereas Alabama and Georgia (ρ_2) have the lowest. Once more, this further emphasises a difference in the scale at which Florida's epidemic timeline develops compared to Alabama and Georgia.

5.4 Computational Efficiency

In order to evaluate how specific aspects such as the number of parameters, the number of timepoints, parameterisation of the transition periods as well as algorithmic differences affect computational efficiency, the relative runtimes for relevant models are compared. In particular, rather than specifying exact runtime values, more relaxed and generalised estimates are provided. This is done due to the unreliability of timing R packages. Computational runtimes are not necessarily specific to the tasks being run in R but rather, they are affected by other background tasks and the amount of available computational memory as well. As fine-tuning parameters, such as the number of samples, batches and epochs to be run iteratively, remain constant throughout the entire simulation study, it is then easier to infer any differences in runtimes to the aforementioned model considerations.

Table 11: Runtimes for non-spatial models (77 timepoints)

Model	Algorithm	Parameters	Runtime (seconds)
1	SMC-ABC	5	54.145
2	SMC-ABC	6	76.667
3	SMC-ABC	7	198.104
3	Basic ABC	7	52.107

For simplicity, the non-spatial case is firstly considered as seen in Table 11. For the first timeline with 77 timepoints, model 3 required a runtime around three times as long as that of model 1. In comparison, model 2 only required approximately 1.4 times as much runtime than that of model 1. Using the same number of parameters, model 3 under the basic ABC algorithm sees a significant decrease in computational cost as it only ran for approximately a quarter of the time spent under the SMC-ABC algorithm. In this case, considering the relatively similar predicted posterior distributions and posterior estimates, the basic ABC algorithm can be said to be more computationally efficient.

Table 12: Runtimes for non-spatial models (184 timepoints)

Model	Algorithm	Parameters	Runtime (seconds)
4	SMC-ABC	5	195.776
5	SMC-ABC	8	178.343
6	SMC-ABC	9	210.122
6	Basic ABC	9	92.888
7	SMC-ABC	10	198.717
8	SMC-ABC	11	135.069

The second timeline, as presented in Table 12, increases the number of timepoints to 184. Despite this, the computational burden of increased dimensional matrices does not appear to have any inherent effect on runtimes such that model 3 and model 4 (recall that they are the same model just under different timelines) required approximately the same amount of time for assumed convergence. With the exploratory inclusion of basis splines of varying degrees of freedom (models 5-8), it is apparent that increasing the number of exposure parameters also do not result in much differing runtimes (the greatest difference is observed to be around 30 seconds); in particular, this is true up until the inclusion of a sixth temporal component under the 6 degrees of freedom spline basis (model 8) which resulted in the shortest runtime. Recalling how the algorithm (and software) works, this relatively short runtime can probably be attributed to overspecification of model (exposure) parameters such that the number of accepted samples from the prior distribution quickly exceeds the maximum number of 250 batches; in other words, the prior distribution fails to converge to the posterior. Using the 'best' model, it is again observed that the basic ABC algorithm is computationally more efficient compared to the SMC-ABC algorithm even with an increased number of parameters under model 6; the SMC-ABC algorithm takes approximately 2.3 times longer.

Table 13: Runtimes for non-spatial models (transition period parameterisations)

Model	Timepoints	Algorithm	Parameters	Runtime (seconds)
3 (Exp)	77	SMC-ABC	5	198.104
3 (Weibull)	77	SMC-ABC	7	683.723
3 (Exp)	77	Basic ABC	5	52.107
3 (Weibull)	77	Basic ABC	7	182.847
6 (Exp)	184	SMC-ABC	9	210.122
6 (Weibull)	184	SMC ABC	11	524.453
6 (Exp)	184	Basic ABC	9	92.888
6 (Weibull)	184	Basic ABC	11	242.690

For considerations of transition period parameterisations as seen in Table 13, model 3 and model 6 are chosen. For the smaller timeline, the Weibull parameterisation ran significantly (approximately 3.5 times) longer compared to the exponential parameterisation, both using the SMC-ABC algorithm; similarly for the longer timeline the Weibull parameterisation required a runtime 2.5 times greater than that of the exponential parameterisation. The basic algorithm also produced reduced runtimes under the Weibull parameterisation, however they are still observed to be significantly greater than exponential parameterisation under the same algorithm. Interestingly, the corresponding ratios produced using the SMC-ABC algorithm (3.5 times and 2.5 times longer runtimes), are also observed under the basic algorithm. Such a difference in runtimes can be attributed to the fact that parameterisation using the Weibull distribution, is computationally expensive as there is an increased number of parameters, corresponding to the shape and scale of the latent and infectious periods (4 total parameters), to be sampled and thus estimated; in comparison, the exponential distribution only has two parameters. Ultimately, there is a trade-off between computational efficiency and estimation performance depending on the epidemic at hand. For example, in the non-spatial case, it was observed that the Weibull distribution was strongly preferred over the exponential distribution for model 3 (such that there were moderately different posterior estimates), as inferred by the approximate Bayes factors. This is in comparison to the Weibull parameterised model 6 which is still estimated to be preferred over the exponential parameterisation by the approximate Bayes factors; however, the difference in factors is much less drastic here, this of which can also be inferred from the posterior predictive distributions which are essentially the same. Thus, the idea of spending longer runtimes to get similar posterior estimates to that of a model/parameterisation which takes a shorter amount of time is conceptually tedious.

Table 14: Runtimes for spatial models

Model	Timepoints	Algorithm	Parameters	Runtime (seconds)
1a (Exp)	77	SMC-ABC	10	1082.683
1a (Weibull)	77	SMC-ABC	12	5291.655
1a (Exp)	77	Basic ABC	10	11.010
1a (Weibull)	77	Basic ABC	12	433.531
1b (Exp)	77	SMC-ABC	8	1174.808
1c (Exp)	77	SMC ABC	8	1025.711
2a (Exp)	184	SMC-ABC	13	677.711
2a (Weibull)	184	SMC-ABC	15	2386.547
2a (Exp)	184	Basic ABC	13	23.787
2a (Weibull)	184	Basic ABC	15	82.880

Finally, for the spatial case, the overall runtimes for all models were significantly longer than that of the non-spatial case. As such, just from a generalised overview as seen in Table 14, it is clear that substantial increases in dimensions of the corresponding matrices does in fact result in a decrease in computational efficiency. For example, in the non-spatial case, consider the two-intervention model for the smaller timeline; specified exposure parameters are defined using a [77 x 3] exposure matrix, where 77 represents the number of timepoints and 3 corresponds to the number of exposure parameters. Now consider the same model for the spatial analysis; the dimensions of the exposure matrix increase drastically to [228 x 5] such that 228 corresponds to the intensity time series for each location, with an increase in the number of columns to 5 as two additional parameters are added for estimation of the additional two locations. Not only this, but there is additional evaluation of distance matrices, this of which may also contribute to this drastic difference. Under the previously described model, there is in fact an approximately 5.5 times longer runtime requirement for the spatial case when compared to the non-spatial case. Much of the implied inferences remain the same for the spatial case such that Weibull parameterisation ran for extensively long periods of time, especially under the SMC-ABC algorithm (at least 20 minutes to an hour and a half at most).

CHAPTER 6

CONCLUSION

In this project, two approximate Bayesian computation (ABC) algorithms, the basic ABC algorithm and the SMC-ABC algorithm, have been implemented in the estimation of stochastic spatial SEIR parameters for modelling COVID-19 in three U.S. states. Along with this simulation study, an in-depth review of the respective methods as well as associated origins, merits and limitations were presented. In this chapter, the project findings are summarised. Following this, the original aims of the project are recalled and the project's degree of execution with regards to such is also evaluated. Finally, areas of further work are suggested.

6.1 Summarised Findings

The non-spatial simulation study served as a means of establishing key behaviour associated with both the ABC algorithms and SEIR models. The inclusion of accurate and relevant intervention terms as parameters of the exposure probability was seen to be central to modelling data of new case counts. With an increased number of such terms, the intensity process became much more flexible, hence improving estimation of the epidemic timeline. However, the feasibility of this increased amount of inclusion decreases with more complex epidemics. This can be seen in two contexts as it pertains to accessibility of data and computational efficiency. In the first case, the example in which there was a proposed increased public awareness of the disease is clearly applicable here. In particular, it is difficult to accurately quantitate such personal preventative methods, such as mask wearing and social distancing, on a spatiotemporal level, especially if mandates are not officially be in place. In the second case, it was shown that with an increasing amount of exposure parameters, the dimensionality of the problem also increases, especially in the spatial case. Though the observed runtimes are not exceedingly extreme, it is unclear how the algorithm will perform in estimating more than 15 parameters, the maximum amount used in the study. It was also noted that other population factors/events may cause the true estimated effect of intensity parameters

to become unidentifiable. Additional time-varying covariates, such as basis splines, were shown to alleviate decreased modelling capabilities, even in the absence of such detailed intervention data. However, while overall fit of the model to the data was improved, interpretations based on exact intensity parameter estimations were avoided due to the extrapolative nature of splines. Instead, estimates were treated as a guide and were used to make more general inferences based on the relative complexity of the underlying population dynamics. Under the preferred Weibull parameterisation, there was not much difference between the basic ABC algorithm and the SMC-ABC algorithm in terms of estimation performance. However, when considering the difference in runtimes, the basic algorithm was deemed more computationally efficient in this case.

Contrary to the non-spatial case, the basic algorithm performed poorly in the spatial analysis, possibly due to the use of diffuse priors in such a high dimensional problem. Population demographics and age-varying latent periods may account for this discrepancy. In the spatial analysis, the preferred distance model inferred the presence of cross-border contact between the three states. Florida was predicted to have the highest mean epidemic potential whereas Alabama was predicted to have the lowest. Despite the inclusion of splines, the model was not able to accurately capture both peaks in the longer timeline, potentially due to the substantial difference in the scale of Florida's cases versus that of the other two states. Evidence for this is further emphasised by the contact intensity estimates. In particular, the estimated contact intensity between Florida and the two other states were seen to be the highest whereas Alabama and Georgia had the lowest. Runtimes were increased to a noticeable extent due to the rate at which dimensionality increased owing to both: the two additional locations and their corresponding parameters.

For both the non-spatial and spatial case, it was noted that the Weibull parameterisation had considerably longer runtimes than the exponential parameterisation. It was deliberated whether the trade-off between computation time and estimation performance was warranted. Using the approximate Bayes factors as a guide, it was deemed sensible to spend more time computing more reasonable estimates as seen the first non-spatial model (model 3). On the other hand, when estimates between the two distributions were almost the same, as seen in the second non-spatial model (model 6), the additional runtimes were seen as exhaustive.

6.2 Overall aims and their degree of execution

Revisiting the initial aims of the project, the first goal was to critically assess the practicality of ABC based on theory, as well as simulations based on state-level COVID-19 data using ABSEIR. Theory-based discussion was presented in the literature review section of chapter 3, much of which tied into the simulation-based evaluations. As discussed in the literature review, ABC allows for inference under high dimensional settings where traditional algorithms such as MCMC methods become less practical or simply fail. Through the simulation study of the COVID-19 pandemic in the U.S., a core high dimensional problem, the most basic ABC algorithm was then assessed in terms of producing reasonable estimates and computational efficiency. Much of the theory stated that the basic algorithm tends to falter under diffuse priors; this was specifically confirmed in the spatial analysis. Following the next aim of making recommendations for methodological improvements, the SMC-ABC algorithm was evaluated as it had been shown to increase efficiency when the basic algorithm became less practical. This was also confirmed in the spatial analysis. As such, basic ABC is only seen to be practical once informative priors are provided whereas SMC-ABC offers improved results in a diffused setting.

The second goal was to analyse the suitability of the basic SEIR model in realistically predicting infectious disease dynamics as it relates to the COVID-19 pandemic. Only the basic model was completely implemented in the simulation studies; in other words, no extra compartments were added to the model. Obtained estimates were reasonable in most cases and provided for characterisation of overall epidemic behaviour. However, it was noted that the increased complexity of the epidemic process warranted the inclusion of additional covariates which govern the epidemic intensity. As discussed in the literature review, this may be achieved by the creation of extended models; for example, an extra compartment may be included such that vaccinated individuals are assigned a different probability of contracting the virus compared to unvaccinated persons. Therefore, as this concept was not implemented directly using ABSEIR, the aim of making recommendations for methodological improvements of the model may be considered as partially completed.

6.3 Further Work

As the inferences of this study are only based on the three specified states, a promising aspect for further studies using these methods can be seen through expansion of the number spatial locations involved. This is because results cannot be generalised to all cases of ABC for stochastic spatial SEIR models. The aforementioned uncertainty associated with the algorithm's estimation of more than the maximum number of parameters used in this study also offers another interesting prospective. This uncertainty may also be extended to the number of spatial locations as well. As such, one could investigate whether there exists some range or threshold value for the number of these parameters to be included, after which estimation becomes impossible. Additionally, as highlighted in the previous section, extended models may show improvements to modelling. Thus, another intuitive study could include much more compartments regarding additional COVID-19 data (or just any epidemic data in general). The simplest case is seen via inclusion of vaccination rates for respective locations. More complex cases can be devised following specifications discussed in chapter 2. For example, consider one of the earliest models formulated by Lin, et al. [18]. Public perception of risk is included as an extra compartment, however more complicated models may be supplemented by government policy information, Google mobility data or even social media activity; as such a more detailed epidemic process is developed around the human reaction to outbreak severity over time [9]. Lastly, with the constant and evolutionary nature of infectious disease processes, further studies could implement these methods for analysis of newly emerging diseases. In such early stages of epidemic development, a sound understanding of epidemic behaviour (such as associated latent and infectious periods) is not quite established as yet. As such, evaluation of the different algorithms under truly non-informative priors could allow for even more evidence regarding estimation performance and computational efficiency.

References

[1] G. D. Brown, *Application of heterogeneous computing techniques to compartmental spatiotemporal epidemic models*. The University of Iowa, 2015.

[2] R. J. Boys and P. R. Giles, "Bayesian inference for stochastic epidemic models with time-inhomogeneous removal rates," *Journal of mathematical biology,* vol. 55, no. 2, pp. 223-247, 2007, doi: 10.1007/s00285-007-0081-y.

[3] G. D. Brown, A. T. Porter, J. J. Oleson, and J. A. Hinman, "Approximate Bayesian computation for spatial SEIR(S) epidemic models," *Spatial and spatio-temporal epidemiology,* vol. 24, pp. 27-37, 2018, doi: 10.1016/j.sste.2017.11.001.

[4] C. Cheng *et al.*, "The incubation period of COVID-19: a global meta-analysis of 53 studies and a Chinese observation study of 11 545 patients," *Infectious Diseases of Poverty,* vol. 10, no. 1, p. 119, 2021/09/17 2021, doi: 10.1186/s40249-021-00901-9.

[5] S. Flaxman *et al.*, "Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe," *Nature,* vol. 584, no. 7820, pp. 257-261, 2020/08/01 2020, doi: 10.1038/s41586-020-2405-7.

[6] T. McKinley, A. R. Cook, and R. Deardon, "Inference in epidemic models without likelihoods," *The International Journal of Biostatistics,* vol. 5, no. 1, 2009.

[7] W. O. Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epidemics," *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character,* vol. 115, no. 772, pp. 700-721, 1927.

[8] P. D. O'Neill and G. O. Roberts, "Bayesian inference for partially observed stochastic epidemics," *Journal of the Royal Statistical Society: Series A (Statistics in Society),* vol. 162, no. 1, pp. 121-129, 1999.

[9] C. Ward, "Bayesian Methods for Spatio-Temporal Epidemic Models to Accurately Capture Complex Dynamics of Disease Spread," The University of Iowa, 2021.

[10] G. D. Brown, J. J. Oleson, and A. T. Porter, "An empirically adjusted approach to reproductive number estimation for stochastic compartmental models: A case study of two Ebola outbreaks: Empirically Adjusted Reproductive Number," *Biometrics,* vol. 72, no. 2, pp. 335-343, 2016, doi: 10.1111/biom.12432.

[11] P. E. Lekone and B. F. Finkenstädt, "Statistical Inference in a Stochastic Epidemic SEIR Model with Control Intervention: Ebola as a Case Study," *Biometrics,* vol. 62, no. 4, pp. 1170-1177, 2006, doi: 10.1111/j.1541-0420.2006.00609.x.

[12] A. T. Porter and J. J. Oleson, "A path-specific SEIR model for use with general latent and infectious time distributions," (in eng), *Biometrics,* vol. 69, no. 1, pp. 101-8, Mar 2013, doi: 10.1111/j.1541-0420.2012.01809.x.

[13] N. Whiteley and L. Rimella, "Inference in Stochastic Epidemic Models via Multinomial Approximations," 2020.

[14] E. Ramirez-Torres *et al.*, "Mathematical modeling and forecasting of COVID-19: experience in Santiago de Cuba province," 2021.

[15] A. Minter and R. Retkute, "Approximate Bayesian Computation for infectious disease modelling," *Epidemics,* vol. 29, pp. 100368-100368, 2019, doi: 10.1016/j.epidem.2019.100368.

[16] P. Neal, "Efficient likelihood-free Bayesian Computation for household epidemics," *Statistics and computing,* vol. 22, no. 6, pp. 1239-1256, 2010, doi: 10.1007/s11222-010-9216-x.

[17] H. F. Lopes, "Handbook of Markov Chain Monte Carlo by BROOKS, S., GELMAN, A., JONES, G. L. and MENG, X," *Biometrics,* vol. 69, no. 3, pp. 801-801, 2013, doi: 10.1111/biom.12099.

[18] Q. Lin *et al.*, "A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action," *International journal of infectious diseases,* vol. 93, pp. 211-216, 2020, doi: 10.1016/j.ijid.2020.02.058.

[19] Z. Wu and J. M. McGoogan, "Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention," *JAMA : the journal of the American Medical Association,* vol. 323, no. 13, pp. 1239-1242, 2020, doi: 10.1001/jama.2020.2648.

[20] Z. Chen, L. Feng, H. A. Lay, K. Furati, and A. Khaliq, "SEIR model with unreported infected population and dynamic parameters for the spread of COVID-19," *Mathematics and computers in simulation,* vol. 198, pp. 31-46, 2022, doi: 10.1016/j.matcom.2022.02.025.

[21] J. M. V. Grzybowski, R. V. da Silva, and M. Rafikov, "Expanded SEIRCQ Model Applied to COVID-19 Epidemic Control Strategy Design and Medical Infrastructure Planning," *Mathematical problems in engineering,* vol. 2020, pp. 1-15, 2020, doi: 10.1155/2020/8198563.

[22] R. F. Reis *et al.*, "Characterization of the COVID-19 pandemic and the impact of uncertainties, mitigation strategies, and underreporting of cases in South Korea, Italy, and Brazil," *Chaos, solitons and fractals,* vol. 136, pp. 109888-109888, 2020, doi: 10.1016/j.chaos.2020.109888.

[23] D. B. Rubin, "Bayesianly justifiable and relevant frequency calculations for the applied statistician," *The Annals of statistics,* vol. 12, no. 4, pp. 1151-1172, 1984, doi: 10.1214/aos/1176346785.

[24] S. Tavare, D. J. Balding, R. C. Griffiths, and P. Donnelly, "Inferring Coalescence Times From DNA Sequence Data," *Genetics (Austin),* vol. 145, no. 2, pp. 505-518, 1997, doi: 10.1093/genetics/145.2.505.

[25] T. Kypraios, P. Neal, and D. Prangle, "A tutorial introduction to Bayesian inference for stochastic epidemic models using Approximate Bayesian Computation," *Mathematical biosciences,* vol. 287, pp. 42-53, 2017, doi: 10.1016/j.mbs.2016.07.001.

[26] J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman, "Population growth of human Y chromosomes: a study of Y chromosome microsatellites," *Molecular biology and evolution,* vol. 16, no. 12, pp. 1791-1798, 1999, doi: 10.1093/oxfordjournals.molbev.a026091.

[27] H. Götte, "Handbook of Approximate Bayesian Computation. Edited by Scott A.Sisson, YananFan, Mark A.Beaumont (2019). London, UK: Chapman & Hall/CRC Press. 662 pages, ISBN: 978-1-4398-8150-7," *Biometrical journal,* vol. 61, no. 6, pp. 1601-1602, 2019, doi: 10.1002/bimj.201900141.

[28] M. A. Beaumont, J.-M. Cornuet, J.-M. Marin, and C. P. Robert, "Adaptive approximate Bayesian computation," *Biometrika,* vol. 96, no. 4, pp. 983-990, 2009, doi: 10.1093/biomet/asp052.

[29] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf, "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical

systems," *Journal of the Royal Society interface,* vol. 6, no. 31, pp. 187-202, 2009, doi: 10.1098/rsif.2008.0172.
[30] M. G. B. Blum, "Approximate Bayesian Computation: A Nonparametric Perspective," *Journal of the American Statistical Association,* vol. 105, no. 491, pp. 1178-1187, 2010, doi: 10.1198/jasa.2010.tm09448.
[31] J.-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder, "Approximate Bayesian computational methods," *Statistics and computing,* vol. 22, no. 6, pp. 1167-1180, 2011, doi: 10.1007/s11222-011-9288-2.
[32] R. D. Wilkinson, "Approximate Bayesian computation (ABC) gives exact results under the assumption of model error," 2008, doi: 10.1515/sagmb-2013-0010.
[33] T. J. McKinley *et al.*, "Approximate Bayesian Computation and Simulation-Based Inference for Complex Stochastic Epidemic Models," *Statistical science,* vol. 33, no. 1, pp. 4-18, 2018, doi: 10.1214/17-STS618.
[34] G. D. Brown. "ABSEIR." https://github.com/grantbrown/ABSEIR (accessed 14 Feb, 2022).
[35] A. Perperoglou, W. Sauerbrei, M. Abrahamowicz, and M. Schmid, "A review of spline function procedures in R," *BMC Medical Research Methodology,* vol. 19, no. 1, 2019, doi: 10.1186/s12874-019-0666-3.
[36] N. Y. Times. "COVID-19 Cases and Deaths Rolling Averages and Anomalous Days." https://github.com/nytimes/covid-19-data/tree/master/rolling-averages (accessed 15 May, 2022).
[37] O. W. i. Data. "Data on COVID-19 (coronavirus) vaccinations by Our World in Data." https://github.com/owid/covid-19-data/tree/master/public/data/vaccinations#united-states-vaccination-data (accessed 20 May, 2022).
[38] U. S. C. Bureau. "State Population Totals: 2010-2019." https://www.census.gov/data/datasets/time-series/demo/popest/2010s-state-total.html (accessed 15 May, 2022).
[39] Y. Weng and G. Y. Yi, "Estimation of the COVID-19 mean incubation time: Systematic review, meta-analysis, and sensitivity analysis," (in eng), *J Med Virol,* vol. 94, no. 9, pp. 4156-4169, Sep 2022, doi: 10.1002/jmv.27841.
[40] A. W. Byrne *et al.*, "Inferred duration of infectious period of SARS-CoV-2: rapid scoping review and analysis of available evidence for asymptomatic and symptomatic COVID-19 cases," *BMJ Open,* vol. 10, no. 8, p. e039856, 2020, doi: 10.1136/bmjopen-2020-039856.
[41] Wikipedia. "COVID-19 pandemic in Florida." https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Florida (accessed 16 Aug, 2022).
[42] J. Ogles. "Beyond the veil: What mask requirements are in place in Florida?" https://floridapolitics.com/archives/342364-beyond-the-veil-what-face-mask-requirements-are-in-place-in-florida/ (accessed 15 Aug, 2022).
[43] G. Großmann, M. Backenköhler, and V. Wolf, "Heterogeneity matters: Contact structure and individual variation shape epidemic dynamics," *PLOS ONE,* vol. 16, no. 7, p. e0250050, 2021, doi: 10.1371/journal.pone.0250050.