Math. Model. Nat. Phenom. 17 (2022) 8
https://doi.org/10.1051/mmnp/2022014

Mathematical Modelling of Natural Phenomena
www.mmnp-journal.org

EFFICIENT UNCERTAINTY QUANTIFICATION AND VARIANCE-BASED SENSITIVITY ANALYSIS IN EPIDEMIC MODELLING USING POLYNOMIAL CHAOS

Bjørn C.S. Jensen[1,*], Allan P. Engsig-Karup[2] and Kim Knudsen[2]

Abstract. The use of epidemic modelling in connection with spread of diseases plays an important role in understanding dynamics and providing forecasts for informed analysis and decision-making. In this regard, it is crucial to quantify the effects of uncertainty in the modelling and in model-based predictions to trustfully communicate results and limitations. We propose to do efficient uncertainty quantification in compartmental epidemic models using the generalized Polynomial Chaos (gPC) framework. This framework uses a suitable polynomial basis that can be tailored to the underlying distribution for the parameter uncertainty to do forward propagation through efficient sampling via a mathematical model to quantify the effect on the output. By evaluating the model in a small number of selected points, gPC provides illuminating statistics and sensitivity analysis at a low computational cost. Through two particular case studies based on Danish data for the spread of Covid-19, we demonstrate the applicability of the technique. The test cases consider epidemic peak time estimation and the dynamics between superspreading and partial lockdown measures. The computational results show the efficiency and feasibility of the uncertainty quantification techniques based on gPC, and highlight the relevance of computational uncertainty quantification in epidemic modelling.

Mathematics Subject Classification. 62J10, 65C60, 92D30.

Received August 13, 2021. Accepted April 10, 2022.

1. Introduction

In this manuscript we demonstrate the application of techniques from uncertainty quantification (UQ) to epidemic modelling to provide insight and identify the parameters that have the strongest influence on the results of the predictive modelling. Such knowledge is crucial for mitigation strategies, restriction policies, etc. targeting controlling or reducing the impact of the spread of diseases for securing public health. This will in part also improve the ability to deal with uncertainty in predictive modelling.

Uncertainty quantification as an independent field grew out of problems in various other fields such as probability theory, dynamical systems and numerical simulations. Sampling based techniques, such as Markov

Keywords and phrases: Uncertainty quantifictaion, global statistics, sobol indices, epidemic modelling, Covid-19.

[1] Department of Mathematics and Statistics, University of Helsinki, 00560 Helsinki, Finland.
[2] Department of Applied Mathematics and Computer Science, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark.

* Corresponding author: bjorn.jensen@helsinki.fi

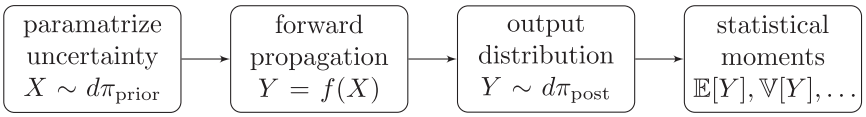


FIGURE 1. The workflow used for handling uncertainty quantification through forward propagation.

Chain Monte Carlo (MCMC) methods and bootstrapping [7, 9], are commonly used in epidemic modelling as seen in the studies [13, 19, 21], and by the expert group[1] providing the Covid-19 related modelling for the Danish government. We propose an alternative approach in the generalized Polynomial Chaos [3, 10, 23, 24] as an efficient general non-iterative framework to do UQ analysis using forward modelling where the uncertainties are parameterized. The outcome are estimates of the expected value and variance of possible solutions under the uncertainty.

Mathematical models are widely applied for modelling the spread of infectious diseases [12], and model predictions are used to inform political decisions for societal counter measures. Models of various complexity, flexibility, restrictions and assumptions exist; see [5] and [2] for a deeper introductions to different kinds of models. If appropriately combined with data the models provide insight into the behaviour of a disease. It may yield estimates for its duration, the peak infection and other various aspects.

In this paper we use extended versions of the SIR model; a compartmental epidemic model named after the division of a population into three compartments: (S)usceptible individuals, (I)nfection individuals, and (R)ecovered individuals; see *e.g.* [5]. A SIR model is simple in nature and generally assumes a homogeneously mixed population across compartments. The model is yet flexible and it easily allows for extensions *e.g.* by separation into age groups or local, geographic regions. One should note that SIR models exhibit exponential growth in the modelling of infected people during the early stage of an epidemic, and hence they are sensitive to model parameters and best suited for short term predictions.

Model parameters are often estimated from various kinds of real-world data, and for that reason they are intrinsically affected by uncertainty. Mathematically, the uncertainty is parameterized by a probability distribution, and with this distribution in place, the parameter uncertainty can be propagated to the model output providing a distribution of possible outputs in place of a single prediction. Computational UQ computes various statistics from the output distribution, *e.g.* mean, variance, confidence intervals, etc. See Figure 1 for a conceptual illustration of UQ.

It can be quite challenging to explore the potentially complicated output distribution. Sampling methods such as Markov Chain Monte Carlo (MCMC) are flexible and commonly employed tools for the purpose. However, MCMC methods often require substantial sampling due to slow convergence. When the model evaluation is expensive, MCMC may not be feasible [15].

Generalized Polynomial Chaos (gPC) poses an efficient alternative non-sampling based method, which can provide very good estimates using significantly fewer model evaluations provided the dimension of the parameter space is sufficiently low. The drawback is that gPC suffers from the curse of dimensionality, *i.e.* when the parameter count (*i.e.* the dimension) grows, the computational requirements may grow much more. The method utilizes orthogonal polynomials and Gaussian quadrature to optimize the number of model evaluations necessary to compute statistics by means of an orthonormal expansion. gPC has also been used on Spanish Covid-19 data in [18].

When various statistics of the output distribution are in place it is important what input parameters gave most influence to the output uncertainty. In other words, are some parameters significantly more contributing to the uncertainty in the model output? Variance-based sensitivity analysis gives the answer in terms of the so-called Sobol indices [1, 20]. Sobol indices have for example been applied in analysis of the Bristish COVIDSIM model in [6].

[1] https://covid19.ssi.dk/analyser-og-prognoser/modelberegninger (accessed April 9th, 2021; in Danish).

The main novelty of our work is in Section 4, where we demonstrate the utility of gPC analysis and Sobol indices in epidemic models and apply them to Danish data from the early phases of Coronavirus SARS-CoV-2 (Covid-19). The versatility of the tools is illustrated in two different cases based on SIR models. Case 1 concerns the estimation of the timing and size of the peak of an epidemic. Case 2 provides a way for UQ modelling of superspreaders in a compartmental model inspired by [19].

This manuscript is structured as follows: Sections 2 and 3 provides the theoretical background. In Section 2, we give an introduction to gPC and illustrate how various basic statistics are directly computable from the expansion coefficients. Sobol indices are introduced in Section 3. We provide a short derivation of their formulation, and we relate Sobol indices to Polynomial Chaos by providing formula for their computation in terms of the gPC expansion coefficients.

The computations included in this manuscript were done in MATLAB and the framework is available as a small toolbox on the DTU GITLAB server[2]. The methods used for computing the various quadratures have been ported to MATLAB from NUMPY[17].

2. Polynomial Chaos expansion

We will consider a model described by the input-output map $f\colon \Omega \subseteq \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} = \mathbb{R}^n$ is the parameter space and Ω is a subset and $\mathcal{Y} = \mathbb{R}^m$ is the output space. The aim is to quantify the effects uncertainty in input $X \in \mathcal{X}$ to that of the output of a model $f(X) = Y \in \mathcal{Y}$.

Polynomial Chaos decomposes $f \in L^2(\Omega, \mathcal{Y}, \mathrm{d}\mu)$ in a basis of orthonormal polynomials $\{\phi_\alpha\}$ of increasing order given by the multi-index $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$. We will use the notational convention that $\alpha = 0$ when $\alpha_i = 0$ for all $1 \leq i \leq n$. Note that $\phi_0(x) = 1$; the zeroth order polynomial is always constant. The decomposition is standard

$$f(X) = \sum_{0 \leq \alpha} \langle f, \phi_\alpha \rangle_\mu \phi_\alpha(X), \quad \text{where} \quad \langle f, \phi_\alpha \rangle_\mu = \int_\Omega f(x)\phi_\alpha(x)\,\mathrm{d}\mu(x). \tag{2.1}$$

Here $\langle f, \phi_\alpha \rangle_\mu$ are the coefficients of f computed via the $L^2(\Omega, \mathrm{d}\mu)$ inner product. In practice the series is truncated and $f(X)$ is approximated by a finite sum. When f is smooth the decay of coefficients is exponential[4] and hence the error introduced by the approximation is small.

For a number of common probability measures $\mathrm{d}\mu$ the orthonormal polynomials are well-known and easy to generate. They also yield Gaussian quadratures with respect to these probability measures, which makes the computation of the involved integrals fast [4].

Consider as an example the standard normal distribution $\mathcal{N}(0, 1^2)$, which up to a scaling constant has probability measure $\exp(-x^2/2)$. The (probabilists) Hermite polynomials form an orthogonal sequence with respect to this measure. The corresponding Gaussian-Hermite quadrature is formed by picking a degree n_{quad} as

$$\int_{\mathbb{R}} f(x) e^{-\frac{x^2}{2}}\,\mathrm{d}x \approx \sum_{i=1}^{n_{\text{quad}}} w_i f(\xi_i)$$

with ξ_i denoting the roots of the Hermite polynomial of the corresponding degree and weights w_i; this quadrature is exact whenever f is a polynomial with $\deg(f) \leq 2n_{\text{quad}} - 1$.

2.1. Statistical properties

While stochastic phenomena come with expressive distributions, they are often well characterized by basic statistical properties like the mean, variance and covariance. Given a model f with parameters characterized by the random variable X, the resulting output $Y = f(X)$ is a new random variable. In this section we will

[2] https://gitlab.gbar.dtu.dk/bcsj/covid-19-ctrl-public-code.

discuss how basic statistical properties of Y become directly computable from the coefficients in our polynomial expansion for f.

Assume that $\mathrm{d}\mu$ is a probability measure on the parameter set Ω, and that $\{\phi_\alpha\}$ is an orthonormal basis as above. Consider the random variable $Y = f(X)$, where $X \sim \mathrm{d}\mu$, *i.e.* it follows the distribution defined by $\mathrm{d}\mu$. Let us denote by $c_\alpha = \langle f, \phi_\alpha \rangle_\mu$, then it is easy to see that we immediately obtain the mean value in terms of the first coefficient

$$\mathbb{E}[Y] = \int_\Omega f(x)\,\mathrm{d}\mu(x) = \int_\Omega f(x)\phi_0(x)\,\mathrm{d}\mu(x) = c_0. \tag{2.2}$$

In a similarly way the variance may be derived as

$$\mathbb{V}[Y] = \sum_{0 \leq \alpha} c_\alpha^2 \mathbb{V}\left[\phi_\alpha(X)\right] = \sum_{0 < \alpha} c_\alpha^2, \tag{2.3}$$

using $\mathbb{E}[\phi_\alpha] = \delta_\alpha$ and $\mathbb{E}[\phi_\alpha \phi_\beta] = \delta_{\alpha-\beta}$ by orthonormality.

Consider now the random variables $Y_1 = f_1(X)$ and $Y_2 = f_2(X)$ with coefficients $\{c_{1,\alpha}\}$ and $\{c_{2,\alpha}\}$, then a computation analogous to that for the variance yields the covariance as

$$\mathrm{Cov}(Y_1, Y_2) = \mathbb{E}[Y_1 Y_2] - \mathbb{E}[Y_1]\mathbb{E}[Y_2] = \sum_{0<\alpha}^{\infty} c_{1,\alpha} c_{2,\alpha}. \tag{2.4}$$

In higher generality, if $\mathbf{Y}$ is a random vector with $Y_i = f_i(X)$, $1 \leq i \leq k$ and we have coefficients $c_{i,\alpha} = \langle f_i, \phi_\alpha \rangle_\mu$. Forming the infinite matrix

$$Q = \begin{bmatrix} c_{1,\alpha(1)} & c_{1,\alpha(2)} & \cdots & c_{1,\alpha(j)} & \cdots \\ c_{2,\alpha(1)} & c_{2,\alpha(2)} & \cdots & c_{2,\alpha(j)} & \cdots \\ \vdots & \vdots & & \vdots & \\ c_{k,\alpha(1)} & c_{k,\alpha(2)} & \cdots & c_{k,\alpha(j)} & \cdots \end{bmatrix} \in \mathbb{R}^{k \times \mathbb{N}},$$

where $\alpha(\cdot)\colon \mathbb{N} \to \mathbb{N}^n$ is some traversal of the multi-index space with $\alpha(0) = (0, 0, \ldots, 0)$, the covariance matrix $C = \mathrm{Cov}(\mathbf{Y}, \mathbf{Y})$ is of the form $C = QQ^T \in \mathbb{R}^{k \times k}$.

3. Variance-based sensitivity analysis

The uncertainty of the output $Y = f(X)$ is quantified in terms of its statistical properties like mean and variance, but it is natural to trace the uncertainty in Y back to the individual parameters in X. The quantification of individual input parameters' influence on the output variance is called variance-based sensitivity analysis. If such information is available, we would know the relative importance of the uncertainty in a single parameter on the overall output variance, and this knowledge would indicate in which parameters of X, if possible, we should aim at reducing the uncertainty.

The Sobol indices form a quantification of the variance contribution on the output Y from each individual parameter and each combination of the parameters X. Like the global statistical properties, the Sobol indices are computable from the polynomial expansion coefficients for f. We give a brief example here, then present the formulation of the Sobol indices, and follow up with the derivation in terms of the coefficients.

Consider the map $f\colon \Omega \subset \mathcal{X} \to \mathcal{Y}$, $\mathcal{X} = \mathbb{R}^n$, $\mathcal{Y} = \mathbb{R}$ (note that the following derivation extends naturally to $\mathcal{Y} = \mathbb{R}^n$). Let $\mathrm{d}\mu = \prod_{i=1}^n \mathrm{d}\mu_i$ be a probability measure on $\Omega \subseteq \mathcal{X}$ and $X = (X_1, \ldots, X_n)$, $X_i \sim \mathrm{d}\mu_i$. Our example is a simple case: Let $X_1 \sim \mathcal{N}(0, a^2)$ and $X_2 \sim \mathcal{N}(0, b^2)$ be normally distributed. Say $Y = X_1 + X_2$, then the Sobol indices would be S_1, S_2 and S_{12} corresponding to each non-empty combination of X_1 and X_2. Their

values would be $S_1 = \frac{a^2}{a^2+b^2}$, $S_2 = \frac{b^2}{a^2+b^2}$ and $S_{12} = 0$. In other words, say $a > b$, then it is better to decrease the uncertainty in X_1 rather than X_2. In contrast, if $Y = X_1X_2$ then the Sobol indices are $S_1 = S_2 = 0$ and $S_{12} = 1$ so decreasing the uncertainty in either parameter is equally beneficial.

3.1. Sobol indices

To compute the Sobol indices we rely on a decomposition into marginalizations of f. We give a derivation of the Sobol indices based on the exposition in [20] to illustrate their computation.

Let $U = \{1, \dots, n\}$.

$$f(X) = \sum_{u \subseteq U} f_u(X_u), \tag{3.1}$$

where we use the notation $X_u = (X_i)_{i \in u}$ and $f_\emptyset(X_\emptyset) \coloneqq f_0 = \mathbb{E}[f(X)]$. The remaining functions f_u, $u \neq \emptyset$, are then recursively defined by

$$f_u(X_u) = \mathbb{E}_{U \setminus u}[f(X)] - \sum_{u' \subsetneq u} f_{u'}(X_{u'}), \tag{3.2}$$

where $\mathbb{E}_{U \setminus u}[f(X)]$ is the marginalization over the parameters X_i for $i \in U \setminus u$,

$$\mathbb{E}_{U \setminus u}[f(X)] \coloneqq \int_{\mathbb{R}^k} f(X) \prod_{i \in U \setminus u} \mathrm{d}\mu_i(X_i), \quad k = |U \setminus u|. \tag{3.3}$$

Here $\mathrm{d}\mu = 0$ outside the domain of f, *i.e.* outside Ω.

Note that the sum is telescopic, each component corresponding to a set u subtracting subset components again. Hence we may compute each $f_u(X_u), u \subseteq U$ starting from the smallest subsets progressing hierarchically upward.

We consider now the variance of $f(X)$ and apply the expansion (3.1) to obtain

$$\mathbb{V}[f(X)] = \sum_{u \subseteq U, u \neq \emptyset} \mathbb{V}[f_u(X_u)]. \tag{3.4}$$

By dividing by the left hand side in (3.4) we get

$$1 = \sum_{u \subseteq U, u \neq \emptyset} \frac{\mathbb{V}[f_u(X_u)]}{\mathbb{V}[f(X)]} = \sum_{u \subseteq U, u \neq \emptyset} S_u, \tag{3.5}$$

which divides the variance into partitional contributions by parameter combinations. The partitions $S_u \coloneqq \mathbb{V}[f_u(X_u)]/\mathbb{V}[f(X)]$, $u \subseteq U = \{1, \dots, n\}$ are the Sobol indices.

3.2. Relation to polynomial chaos

The Sobol indices are efficiently computable from the gPC coefficients. The marginalizations of the distribution arise as restrictions to certain subsets of the coefficients.

Let c_α be the gPC coefficients of f. To compute the Sobol indices we wish to compute the terms $\mathbb{V}[f_u(X_u)]$. Taking the variance on both sides in (3.2) we get

$$\mathbb{V}[f_u(X_u)] = \mathbb{V}\left[\mathbb{E}_{U \setminus u}[f(X)]\right] - \sum_{u' \subsetneq u} \mathbb{V}[f_{u'}(X_{u'})].$$

Clearly, if we compute bottom up hierarchically using the partial ordering $u \leq v$ if $u \subseteq v$, we simply need to compute the marginalizations $\mathbb{V}[\mathbb{E}_{U\setminus u}[f(X)]]$ and then subtract formerly computed values.

Due to the maginalizations of f we will need to consider the marginal structure of our basis functions $\{\phi_\alpha\}$ too. For a multi-index $\alpha \in \mathbb{N}^n$ we shall use the notation

$$\phi_\alpha(x) = \psi_{1,\alpha_1}(x_1) \cdots \psi_{n,\alpha_n}(x_n),$$

where $\{\psi_{i,j}\}_j$ is the orthonormal polynomial basis for parameter X_i. With this we may derive

$$\begin{aligned}
\mathbb{E}_{U\setminus u}[f(X)] &= \int_{\mathbb{R}^k} f(X) \prod_{i\in U\setminus u} \mathrm{d}\mu_i(X_i) \\
&= \int_{\mathbb{R}^k} \sum_{0\leq\alpha} c_\alpha \phi_\alpha(X) \prod_{i\in U\setminus u} \mathrm{d}\mu_i(X_i) \\
&= \sum_{0\leq\alpha} c_\alpha \left(\prod_{i\in u} \psi_{i,\alpha_i}(X_i)\right) \left(\prod_{i\in U\setminus u} \int_{\mathbb{R}} \psi_{i,\alpha_i}(X_i)\, \mathrm{d}\mu_i(X_i)\right) \\
&= \sum_{0\leq\alpha} c_\alpha \left(\prod_{i\in u} \psi_{i,\alpha_i}(X_i)\right) \left(\prod_{i\in U\setminus u} \mathbb{E}[\psi_{i,\alpha_i}(X_i)]\right)
\end{aligned}$$

(note that this product of mean values is 0 unless $\alpha_i = 0$ for all $i \in U\setminus u$; we write simply $\alpha_{U\setminus u} = 0$)

$$= \sum_{0\leq\alpha,\alpha_{U\setminus u}=0} c_\alpha \prod_{i\in u} \psi_{i,\alpha_i}(X_i)$$

(as $\psi_{i,0}(x) = 1$ this product extends to all of α again now that $\alpha_{U\setminus u} = 0$ is fixed)

$$= \sum_{0\leq\alpha,\alpha_{U\setminus u}=0} c_\alpha^2 \phi_\alpha(X).$$

Taking the variance of the above and using the fact that $\mathbb{V}[\phi_\alpha(X)] = 1$ for $\alpha \neq 0$ and zero otherwise we get

$$\mathbb{V}\left[\mathbb{E}_{U\setminus u}[f(X)]\right] = \sum_{0\leq\alpha,\alpha_{U\setminus u}=0} c_\alpha^2 \mathbb{V}[\phi_\alpha(X)] = \sum_{0<\alpha,\alpha_{U\setminus u}=0} c_\alpha^2 \tag{3.6}$$

Visually, if we consider just two parameters, we see in the coefficient grid below how the different coefficients distribute themselves among the

$$\begin{array}{c|cccc|c} \cancel{c^2_{0,0}} & c^2_{0,1} & c^2_{0,2} & c^2_{0,3} & \cdots & \sum\square = \mathbb{V}[f_2(x_2)] \\ \hline c^2_{1,0} & c^2_{1,1} & c^2_{1,2} & c^2_{1,3} & \cdots & \\ c^2_{2,0} & c^2_{2,1} & c^2_{2,2} & c^2_{2,3} & \cdots & \\ c^2_{3,0} & c^2_{3,1} & c^2_{3,2} & c^2_{3,3} & \cdots & \\ \vdots & \vdots & \vdots & \vdots & \ddots & \\ \sum\square = \mathbb{V}[f_1(x_1)] & & & & & \sum\square = \mathbb{V}[f_{12}(x_{12})] \end{array}$$

Here "$\sum\square$" is simply intended as a placeholder symbol for the *sum of each of the elements in the sectioned off region of the grid*. Note that $c_{0,0}$ is crossed out, as it is the mean, which does not contribute to the variance.

4. Uncertainty quantification in modelling spread of diseases using Polynomimal Chaos

In this section we present two cases to demonstrate the flexibility of the above techniques by applying them to compartmental models. The first case considers a SEIR model and compute distributions for the size and timing of the peak of the modelled epidemic.

For the second case a more extensive compartmental model is considered. Inspired by the agent based modelling of superspreaders discussed in [19] we construct a multi-compartment model and formulate a modelling approach for superspreaders leading to comparable results despite the differences in modelling assumptions. With this model we perform a UQ analysis on the coefficients modelling the government imposed restrictions.

In both cases the population size N is taken as 5.8×10^6 matching the size of the Danish population.

4.1. Case 1: Epidemic peak

For this simple case we consider a SEIR model, *i.e.* a model with the compartments (S)usceptible, (E)xposed, (I)nfectious and (R)ecovered/removed. The model is visualized in the diagram in Figure 2.

As an ODE system the model is of the form

$$\frac{\partial S}{\partial t} = -\beta \frac{I(t)S(t)}{N}, \tag{4.1a}$$

$$\frac{\partial E}{\partial t} = \beta \frac{I(t)S(t)}{N} - \sigma E(t), \tag{4.1b}$$

$$\frac{\partial I}{\partial t} = \sigma E(t) - \gamma I(t), \tag{4.1c}$$

$$\frac{\partial R}{\partial t} = \gamma I(t), \tag{4.1d}$$

where β, σ and γ are transition coefficients and N the total size of the population. σ is the rate at which people progress from being exposed (incubating) to becoming infectious, and γ is the rate at which one recovers (or dies) from the disease. Their reciprocals are the average time an individual spends in the exposed and infectious compartments respectively. β denotes the average rate of infection happening in the population. This quantity is depending on infectiousness of the virus and the social patterns of the population; *e.g.* higher hygiene standard

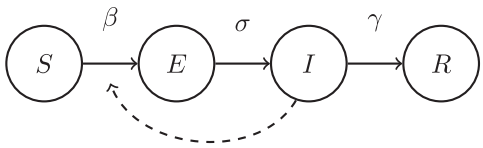
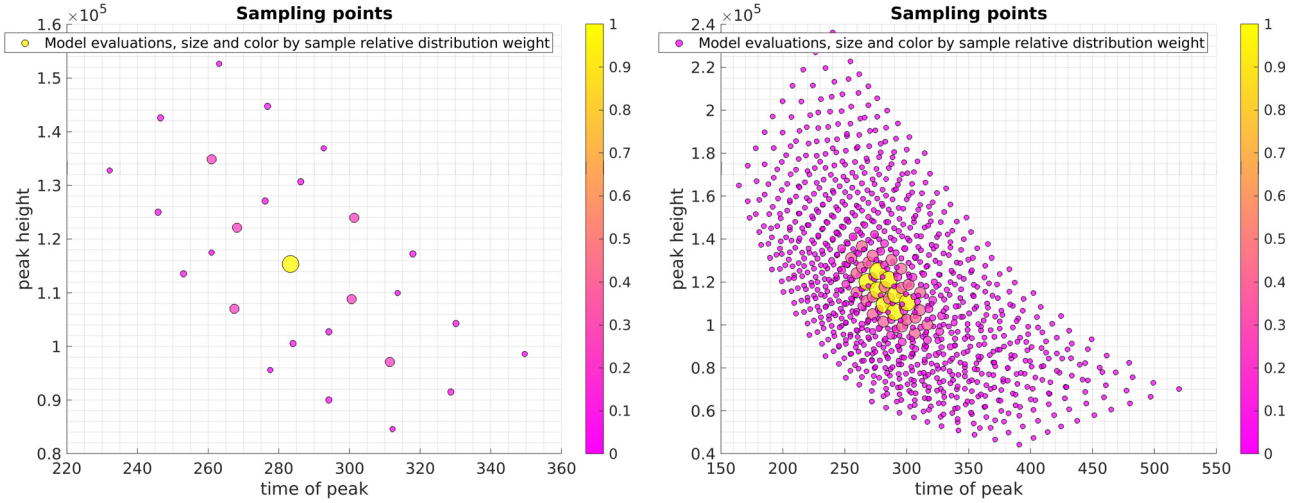


FIGURE 2. Illustration of compartments and transmission rates for a SEIR model.



(A) (B)

FIGURE 3. Comparison of the model evaluation points for the epidemic peak model. There are 27 points in (A) and 1000 points in (B). The color and size of a point is scaled relative to the weight w_i the point is given by the quadrature rule; a minimum size was fixed for visibility.

in the population would lead to a lower β. Note that as a consequence of the model, $S + E + I + R = N$ and hence constant at all times.

In an epidemic the number of infected individuals will rise rapidly as each infected individual will infect several others. However, as the population becomes saturated with infected and recovered individuals (henceforth assumed immune to reinfection) the chance for a meeting between an infected and a susceptible will decrease. This effect is often referred to as *herd immunity*; see [8] for further reading on the concept. Hence, the epidemic is expected to peak at some time t_{peak} where the number of infectious individuals are at its highest.

We consider in this example each parameter β, σ and γ uncertain. The uncertainties are given as uncertainty in the reproduction number $R_0 = \frac{\beta}{\gamma}$, in the duration in the exposed compartment $\tau_{\text{inc}} = \sigma^{-1}$ and the duration in the infectious compartment $\tau_{\text{inf}} = \gamma^{-1}$. As these are positive quantities we assume each log-normally distributed. We thus consider the map

$$\mathcal{F}\colon (R_0, \tau_{\text{inc}}, \tau_{\text{inf}}) \mapsto (t_{\text{peak}}, I_{\text{peak}}), \tag{4.2}$$

where $I_{\text{peak}} \coloneqq I(t_{\text{peak}})$. As the log-normal distribution is simply a transformation of the normal distribution, it is a simple task to transform the quadrature nodes accordingly.

We can thus apply the theory from the previous sections to propagate the uncertainty in the arguments of $\mathcal{F}$ to the output using only few evaluations. As the output quantities are known to be positive as well, we shall assume log-normal distributions for these as well and fit them by computed means and variances.

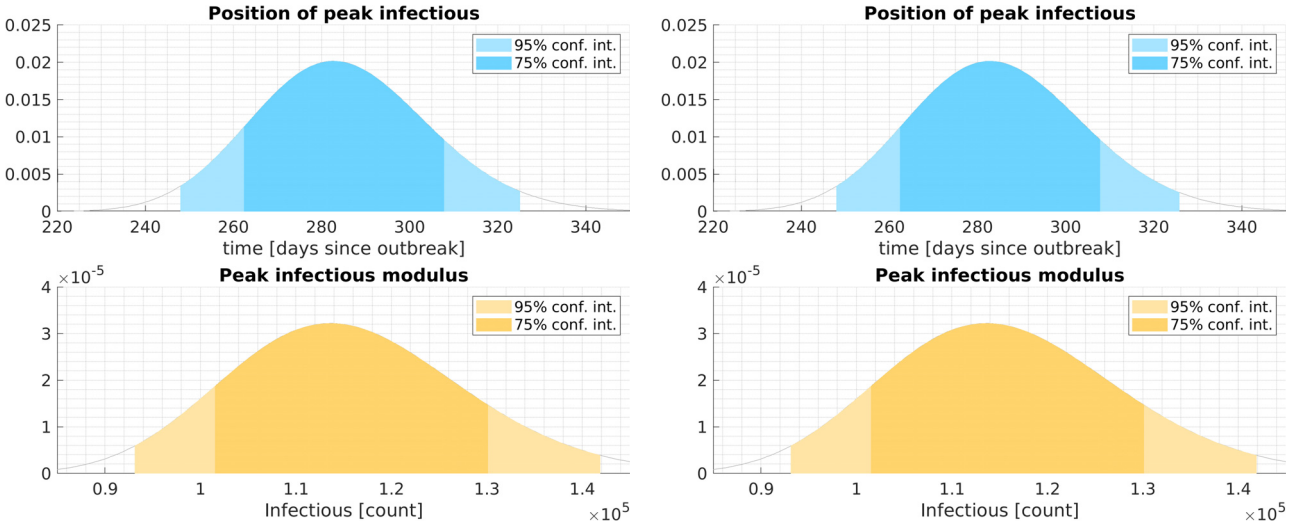


FIGURE 4. Comparison of the distributions for the epidemic peak model output under an assumption that they are lognormal distributed. For the 27 model evaluations in (A) and for the 1000 model evaluations in (B).

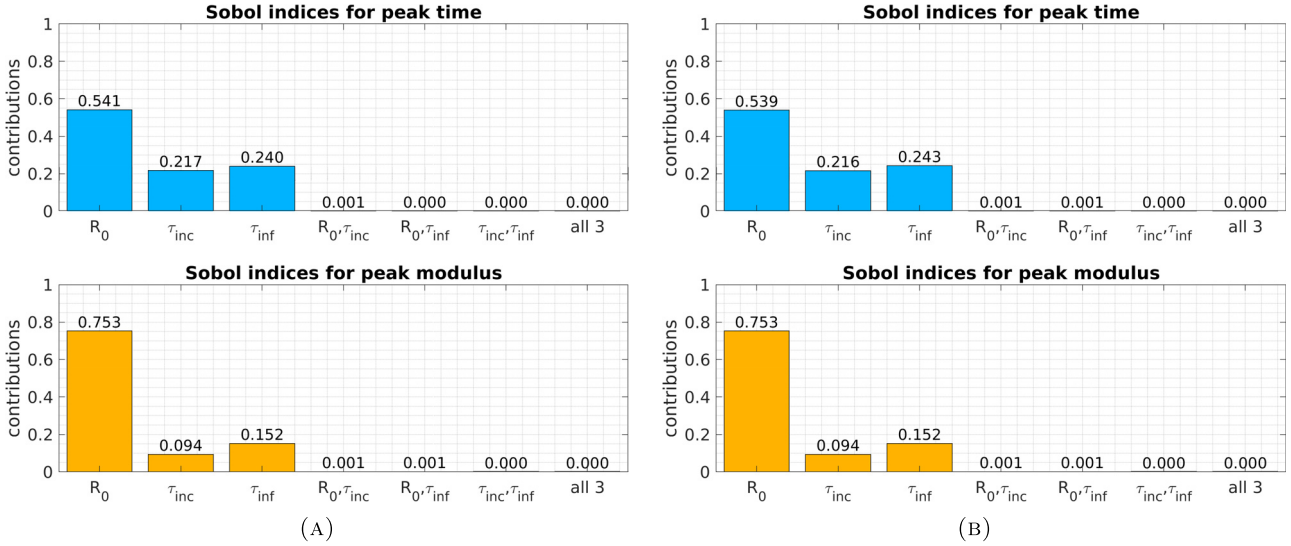


FIGURE 5. Comparison of the Sobol indices for the epidemic peak model. R_0 is clearly the most influential parameter. For the 27 model evaluations in (A) and for the 1000 model evaluations in (B).

The result of the case can be seen in Figures 3, 4 and 5 where we have used the following hyper-parameters; chosen based on early reported numbers for Covid-19 [11, 14, 22] with some level of contact restriction assumed. However, we stress that for the purpose of demonstrating the technique the specific values are of less importance.

		mean	variance
$R_0 \sim$	LogNormal	1.4	0.025^2
$\tau_{\text{inc}} \sim$	LogNormal	4.2	0.25^2
$\tau_{\text{inf}} \sim$	LogNormal	3.3	0.25^2

We compute the mean and standard deviation for the model with a 3rd order of quadrature, 27 model evaluation, and a 10th order, 1000 model evaluations. For comparison we run an MCMC method, a Metropolis-Hastings sampler, for 10,000 iterations to explore the joint distribution. With an acceptance rate of about 26.6% this lead to approximately 2660 model evaluations. We run the MCMC method 8 times to give an idea about the variation in the results it produces. The results are show in the table below where the $\mathcal{X} \pm \mathcal{Y}$ indicates the value's average $\mathcal{X}$ and standard deviation $\mathcal{Y}$ across the 8 MCMC runs.

		t_{peak}		I_{peak}	
method	#evaluations	mean	std	mean ($\times 10^3$)	std ($\times 10^3$)
gPC	27	284.78	19.89	115.63	12.56
gPC	10^3	284.84	19.90	115.64	12.56
MCMC	(8×) ∼2660	285.01 ± 1.37	20.28 ± 0.82	115.73 ± 0.67	12.64 ± 0.27

The results are further explored in Figures 3–6. In the (A) and (B) subfigures of Figures 3–5 we show the 3rd order quadrature and 10th order quadrature results respectively. In Figure 6 we show the explored samples in image space by one of the MCMC runs and the corresponding marginalized distributions. We also show how the mean and standard deviations stabilize across the 10,000 iterations for all the 8 runs.

Figure 3 illustrates all samples in the image space. Their colors and sizes have been scaled by their corresponding quadrature weights w_i, thought a minimum size was fixed for visibility. Figure 4 shows the lognormal distribution for the peak time and the magnitude of the peak in infectious individuals. In Figure 5 the corresponding Sobol indices are illustrated for the selected values the variance of R_0 is the primary concern if we wanted to narrow down the peak time further.

We observe how the 27 model evaluations provides similar information as to the 1000 model evaluations, showing that this problem is handled well already at this low number of evaluations. Furthermore, we observe how we need many more iterations from the MCMC method to reach similar levels of precision.

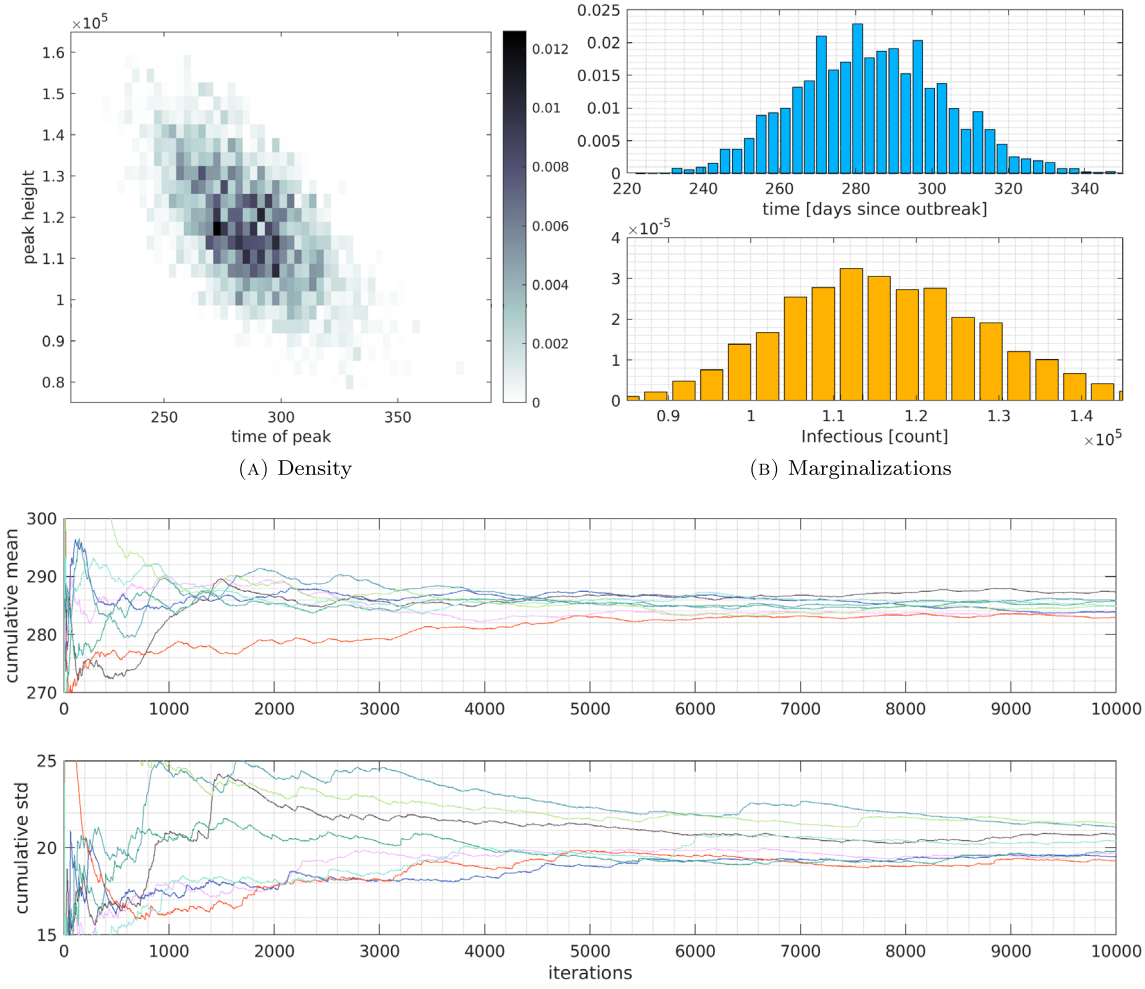
4.2. Case 2: Superspreaders

Superspreaders are infected individuals who during an epidemic are responsible for the infection of a significantly larger amount of individuals than the observed average. Historical observations of diseases have shown that incidents with superspreaders play an important role to the development of epidemics [16].

Various aspects play into causing an individual to become a superspreader, which may both be physiological and sociodynamic in nature. An individual exhaling an increased amount of pathogens relative to the norm could lead to a significantly larger number of infections during regular social interactions compared to a "normal" infectious individual. But it could also simply be the participation in a large scale social event for instance a party, concert or a festival, where the physical distancing may be very low and number of contacts proportionally higher, which results in mass infection.

Inspired by [19] we attempt in this case to replicate some of their results in a computationally fast way using a compartmental model. We employ the structure from their agent based model to construct the compartmental model depicted in the diagram in Figure 7. In the diagram we have the following compartments: First, as in the former model susceptible and exposed. Then there are asymptomatic infectious I_1 and symptomatic infectious I_2. We note that this is a legacy structure from [19], where it is used mostly for book keeping. Neither there nor here is behavior assumed to differ between the compartments. We have a (W)ait compartment, which signifies a short time, where the individual is either so sick that they have isolated themselves as to not infect anyone before admission to the hospital, or they are in non-infecting recovery. There is a branch with (H)ospitalized and (C)ritical care before all ending in the recovered/removed compartment.

The parameters choices are taken as in [19], but we restate them for completeness in Table 1. For z_1 and



(A) Density

(B) Marginalizations

(C) Cumulative means and standard deviations

FIGURE 6. The density (A) and the marginalized densities (B) explored by one run of 10,000 samples of the MCMC method. In (C) we show the cumulative mean and standard deviation for t_{peak}, meaning the mean and standard deviation of all the samples up to the particular iteration.

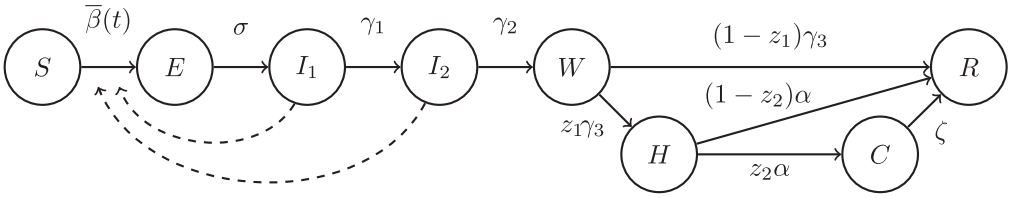


FIGURE 7. Illustration of the expanded compartmental model taking into account superspreaders; based on [19].

TABLE 1. Superspreader model parameters [19]. Units are in [days].

σ^{-1}	γ_1^{-1}	γ_2^{-1}	γ_3^{-1}	α^{-1}	ζ^{-1}
1.2	1.2	3	2	5	12

TABLE 2. Population distribution and hospitalization probability data [19]. Legend: D: Distribution of the population; H: Probability of hospitalization; C: Probability of moving to critical care.

	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80+
D ($\{d_i\}$) [%]	10.9	11.9	13.3	11.7	13.6	13.6	11.7	8.9	4.4*
H ($\{h_i\}$) [%]	0.001**	0.013	0.37	1.1	1.4	2.7	3.9	5.5	5.5
C ($\{\kappa_i\}$) [%]	5	5	5	5	6.3	12.2	27.4	43.2	70.9

*) 0.1% was added here since the numbers from the source table didn't actually add to 100%.
**) This number was 0 in the table, it is known that some kids end up hospitalized, so we changed it to a small but strictly positive value.

z_2 we compute them from the hospitalization rates listed in the *Supplementary material Table 1* in [19] which comes from Norwegian data. We present the data in Table 2 where d_i, h_i and κ_i are the data rows. From these quantities z_1 and z_2 are computed as

$$z_1 = \sum_{i=1}^{9} d_i h_i, \quad \text{and} \quad z_2 = \sum_{i=1}^{9} \frac{d_i h_i}{z_1} \kappa_i.$$

The expression for the last parameter $\overline{\beta}(t)$ is given in (4.6) and (4.5). The modelling approach is covered in Section 4.2.1.

4.2.1. Modelling varying infectivity

We model superspreaders by assuming a distribution of infectivity amongst individuals in the population. Consider the normalized population $[0,1]$ and assign to each $a \in [0,1]$ an infectivity $\beta(a)$. That is, if U is some ordered set of possible infectivities and ψ is a probability measure on U with cumulative probability function Ψ, then $\beta(\cdot) = (1-\Psi)^{-1}(\cdot)$. We assume that the population is ordered by decreasing infectivity; *i.e.* $a < a'$ implies that $\beta(a) \geq \beta(a')$. Assuming a well-mixed distribution such that infection is equally likely to hit any individual $a \in [0,1]$ we may readily calculate the contribution to infection caused by the fraction p most infectious individuals, C_p.

In a SIR model the number of people getting infected at a time t is commonly, as seen in (4.1a), of the form

$$\overline{\beta} \frac{I(t)S(t)}{N},$$

where $I(t)$ is the number of infected individuals, $S(t)$ the number of susceptible individuals, and N is the population size. Here $\overline{\beta}$ is the average infectivity; $\overline{\beta} = \int_0^1 \beta(a)\,\mathrm{d}a$.

The p most infectious individuals would then be contributing the fraction

$$C_p = \overline{\beta}^{-1} \int_0^p \beta(a)\,\mathrm{d}a. \tag{4.3}$$

This quantity informs the choice of probability measure ψ if one works under the scheme that superspreaders form some fraction p of the population and is responsible for infecting the fraction C_p of the population. Fixing these two quantities limits the admissible measures ψ.

This way of modelling a variation in infectivity also admits fairly easy extensions to control scenarios where some rules may change behavioral dynamics over time. We may for instance consider a time-dependent infectivity

$$\overline{\beta}_{\text{restricted}}(t) = \int_0^1 \phi(\beta(a), a, t)\,\mathrm{d}a,$$

where $\phi(b, a, t)$ is some restriction function describing a change in the behavior over time. In practice, however, $\phi(b, a, t) \equiv \phi(b, t)$ will typically be independent of a as we cannot feasibly identify an individual as more infectious than another until after the fact. And so we have to make rules that are uniform for everyone. A simple example could be a strict limitation in how many individuals anyone meet, which could be crudely modelled as

$$\phi(b, a, t) = \min(b, c(t)), \tag{4.4}$$

where $c\colon \mathbb{R}_+ \to \mathbb{R}_+$ is some time-dependent upper bound.

Of course, if $\phi(b, a, t) \equiv \phi(a, t)$ is independent of b, we may impose any kind of other ordering on the population, *e.g.* by age, and apply hard restrictions based on that one. But then we lose all information from the infectivity β, which might be undesirable.

We take for our model $\beta(a)$ as a simple piecewise constant function

$$\beta(a) = \begin{cases} sA & \text{if } a < p, \\ s & \text{if } a >= p, \end{cases} \qquad a \in [0, 1]. \tag{4.5}$$

Here p is the assumed concentration of superspreaders; *e.g.* if we assume 10% are superspreaders $p = 0.1$. sA is the infection rate for superspreaders and s the infection rate for the remaining population. It is an easy calculation that C_p from (4.3) is independent of s, so choosing an assumed pair (p, C_p) determines A and choosing s then determines the mean rate $\overline{\beta}$.

We shall model a social restriction as a hard cap on the amount of individuals any single person gets to interact with. We model this with a restriction function as in (4.4); *i.e.*

$$\overline{\beta}(t) = \int_0^1 \min(\beta(a), c(t))\,\mathrm{d}a. \tag{4.6}$$

with c being a piecewise constant function which changes values at approximately 1) the time of the Danish lockdown, 2) the timing of the Danish reopening's phase 1 (about a month later), 3) the timing of the Danish reopening's phase 2 (another about 40 days later).

4.2.2. Fitting the model using real-world data

From an assumption about the prevalence of superspreaders we first fit β's scaling parameter s and an initial condition I_0 for the epidemic from hospital admission by day[3] (only for the pre-lockdown part of the data set) and an assumption of an unmitigated growth rate at about 23% per day [19]. For the initial condition we fit a number I_0 and we assume that $S(0) = N - I_0$, $E(0) = \frac{I_0}{2}$, $I_1(0) = \frac{I_0}{3}$ and $I_2(0) = \frac{I_0}{6}$. The remaining

[3]Danish data available from SSI (www.ssi.dk), the Danish Ministry of Health. The data was public and accessed on June 14th, 2020; it does not remain available anymore. The used data file is available as a CSV-file with the codes in the GitLab repository.

compartments start at 0. The problem is formulated as

$$\arg\min_{s,I_0} \frac{w_0}{2}|\mathcal{G}(s,I_0)-0.23|^2+\alpha\frac{w_1}{2}\|\mathcal{H}_{t<t_1}(s,I_0)-H_{\text{ssi},t<t_1}\|^2, \tag{4.7}$$

where the weights $w_0^{-1}=0.23^2\ \left[\frac{\text{individuals}^2}{\text{day}^2}\right]$ and $w_1^{-1}=\|H_{\text{ssi}}\|^2\ \left[\text{individuals}^2\right]$ balance the widely different scales of the two terms, and H_{ssi} is the data set of newly admitted hospitalized by day. $\mathcal{G}$ computes the average initial daily growth rate and $\mathcal{H}$ computes the newly admitted hospitalizations from the model. By the subscript $t<t_1$ we mean only the part of the data corresponding to this constraint; $t_1=16$ before which $\mathcal{H}$ is independent of our restriction function. We chose $\alpha=0.01$.

Using the now determined quantities (s,I_0) we fit the three restriction levels in $c(t)$ from the whole data set of hospital admission by day. Fixing $(t_1,t_2,t_3)=(16,46,86)$ we have

$$c(t)=\begin{cases}1 & \text{if } t\le t_1,\\ c_1 & \text{if } t_1<t\le t_2,\\ c_2 & \text{if } t_2<t\le t_3,\\ c_3 & \text{if } t_3<t.\end{cases} \tag{4.8}$$

Then the parameter fitting problem becomes

$$\arg\min_{c_1,c_2,c_3}\frac{1}{2}\|\mathcal{H}(c_1,c_2,c_3)-H_{\text{ssi}}\|^2-w_2(\min(0,c_2-c_1)+\min(0,c_3-c_2)), \tag{4.9}$$

where w_2 is some arbitrary large number so the last term forms a soft constraint enforcing $c_1<c_2<c_3$.

Assuming $(p,C_p)=\left(\frac{1}{10},\frac{4}{5}\right)$, *i.e.* that only 10% contribute 80% of all infections, the above fitting schemes resulted in

$$s=0.602\ \left[\frac{1}{\text{day}}\right],\quad I_0=473.572\ [\text{individuals}],\quad \text{and}\quad (c_1,c_2,c_3)=(0.130,0.187,0.188).$$

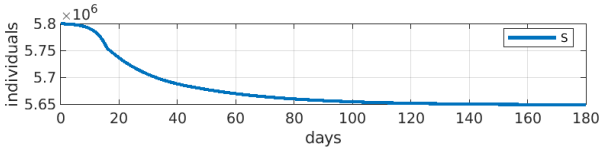
The fractional number of people is due to the continuous nature of the model. These results depended slightly on the choice of initial condition but the differences were on the order of 10^{-3}. The simulation, when done using these data, may be viewed in Figure 8.

We note that the difference between restriction levels c_2 and c_3 is almost insignificant. There are likely various reasons for this. In the agent based model of [19] a social structure is incorporated, which allow them to close particular sectors of society. In comparison our compartmental model has no such structure and thus cannot model that a only particular part of society is closed. A possible explanation might be that the phase 2 reopening didn't really affect the overall amount of contacts for people. This could also be due to a lack of data as small variations in c_3 did not change the value of the optimization functional significantly.

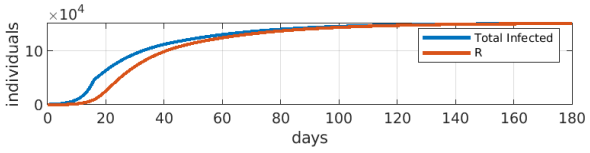
4.2.3. Adding uncertainty

Assuming a level of uncertainty in the fitted restriction levels we may compute confidence intervals for the model. In Figure 9 we assume normally distributed priors for the restriction levels with means c_i, $i=1,2,3$, and relatively scaled standard deviations of $0.1c_i$, $i=1,2,3$. Assuming the posteriors may be approximated reasonably by a truncated normal distributions, 95% confidence intervals are visualized. We see that with uncertainty of this level on the parameters the development is expected to keep declining.

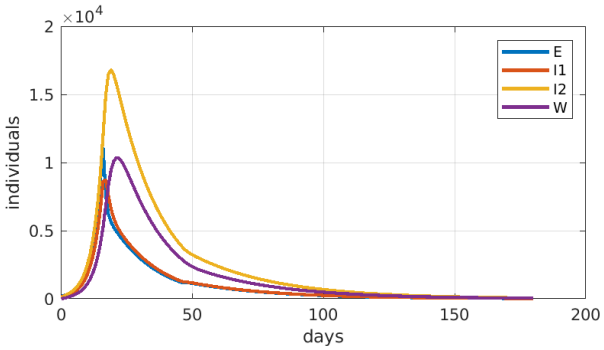
The evolution of the Sobol indices over time is drawn up in Figure 10 illustrating the variance contributions from the parameters, which show as expected how c_1 is the most important initially but is gradually taken over



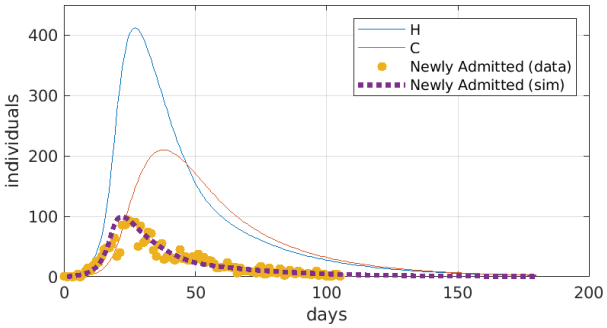
(A) Susceptible



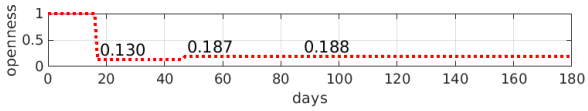
(B) Recovered and total infected



(C) Exposed and infected compartments



(D) Hospitalized, intensive care and newly admitted



(E) Openness level of society

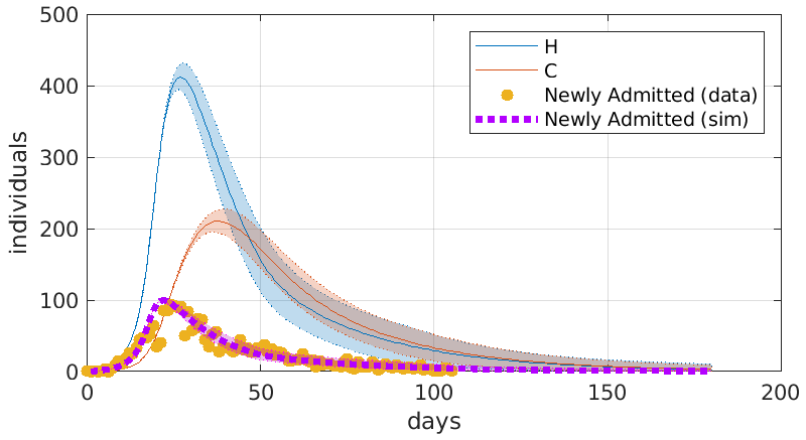
FIGURE 8. Superspreader case simulation using fitted data. The time-evolution of different compartments have been visualized grouped by their relative y-scale to make sure patterns are discernible. The y-scale counts individuals. The red dashed curve in both lower charts (E) visualizes the openness level of the society. The yellow dots in (D) is charted data on newly hospital admissions by day from the Danish authorities and the purple dashed line is the model fit.

by c_2 and then c_3 in the later stages. Notably c_1 remains fairly important even during the time span where c_2 controls the level of interaction, and likewise c_2 into the time span where c_3 is active.

5. Conclusion

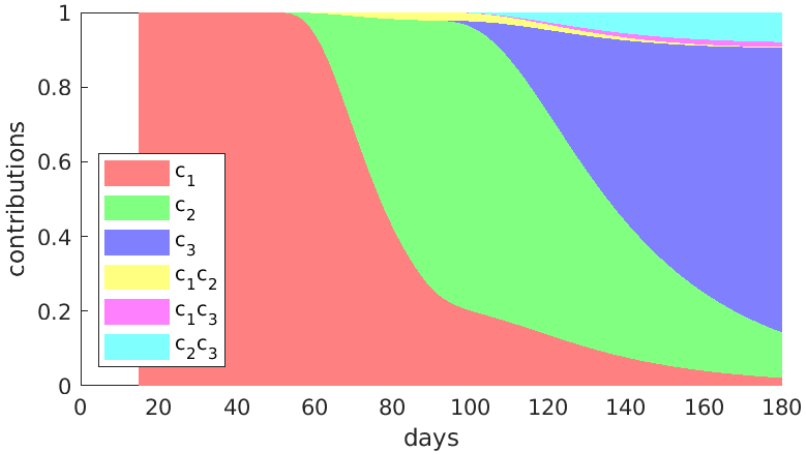
In this paper, we have summarized modern methods for efficient computational uncertainty quantification with the key elements of generalized Polynomial Chaos and the related Polynomial Chaos Expansions. Moreover, we have introduced the related variance-based sensitivity analysis in terms of Sobol indices with applications to the modelling spread of diseases. The techniques have been applied to epidemic models of SIR type based on official Danish Covid-19 data. Our results have demonstrated that the chosen tools are well suited in this setting and reproduce conclusions similar to those of the more computationally expensive model used *e.g.* in [19].

Taking uncertainty into account in predictions from epidemic models is an important means to access possible scenarios and influence of uncertainty on the predictions. Doing this in a transparent and efficient way should be a priority. We recommend that future studies using similar models consider the use of the presented techniques as a tool for quantifying the uncertainty and sensitivity in their computations.



(A) Confidence intervals

FIGURE 9. Shows a superspreader case simulation using fitted data taking into account uncertainty. Hospitalized (H), critical care (C) and newly hospitalized are shown with confidence intervals. The yellow dots are newly hospital admissions by day from the Danish authorities.



(A) Evolution of Sobol indices

FIGURE 10. Shows the Sobol indices as an evolution over time in the superspreader case simulation using fitted data with added uncertainty. At each time the color distribution above determine the part of the variance contributed by each parameter; the corresponding parameters are listed in the legend-box. The c_ic_j parts are the joint variance contributions of c_i and c_j, as is visible the joint parts are mostly unrelated here.

Acknowledgements. This work was supported by the project *Estimation, Simulation, and Control for Optimal Containment of COVID 19* from the Novo Nordisk Fonden; project no. NNF20SA0063089 (Application no. 67). BCSJ was supported by the Academy of Finland (grant no. 320022).

REFERENCES

[1] A. Alexanderian, P.A. Gremaud and R.C. Smith, Variance-based sensitivity analysis for time-dependent processes. *Reliab. Eng. Syst. Saf.* **196** (2020) 106722.

[2] S.N. Arifin, G.R. Madey and F.H. Collins, Spatial agent-based simulation modeling in public health: design, implementation, and applications for malaria epidemiology. John Wiley & Sons (2016).

[3] D. Bigoni, *Uncertainty Quantification with Applications to Engineering Problems*. Ph.D. thesis, Technical University of Denmark (2015).
[4] C. Canuto, M.Y. Hussaini, A. Quarteroni and T.A. Zang, Spectral methods: fundamentals in single domains. Springer Science & Business Media (2007).
[5] V. Capasso and V. Capasso, vol. 88 of Mathematical structures of epidemic systems. Springer (1993).
[6] W. Edeling, H. Arabnejad, R. Sinclair, D. Suleimenova, K. Gopalakrishnan, B. Bosak, D. Groen, I. Mahmood, D. Crommelin and P.V. Coveney, The impact of uncertainty on predictions of the CovidSim epidemiological code. *Nat. Comput. Sci.* **1** (2021) 128–135.
[7] B. Efron, The bootstrap and markov-chain monte carlo. *J. Biopharmaceut. Stat.* **21** (2011) 1052–1062.
[8] P. Fine, K. Eames and D.L. Heymann, "Herd Immunity": A Rough Guide. *Clin. Infect. Dis.* **52** (2011) 911–916.
[9] J. Friedman, T. Hastie, R. Tibshirani *et al.*, The elements of statistical learning. *Springer series in statistics*, New York (2001), Vol. 1.
[10] R. Ghanem and P.D. Spanos, Polynomial chaos in stochastic finite elements. *J. Appl. Mech.* **57** (1990) 197–202.
[11] K. Græsbøll, L.E. Christiansen, U.H. Thygesen and C. Kirkeby, Delaying the peak of the COVID-19 epidemic with travel restrictions. *Epidemiol. Methods* **10** (2021) 20200042.
[12] H.W. Hethcote, The Mathematics of Infectious Diseases. *SIAM Rev.* **42** (2000) 599–653.
[13] T. House, A. Ford, S. Lan, S. Bilson, E. Buckingham-Jeffery and M. Girolami, Bayesian uncertainty quantification for transmissibility of influenza, norovirus and Ebola using information geometry. *J. Royal Soc. Interface* **13** (2016) 20160279.
[14] Q. Lin, S. Zhao, D. Gao, Y. Lou, S. Yang, S.S. Musa, M.H. Wang, Y. Cai, W. Wang, L. Yang *et al.*, A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action. *Int. J. Infectious Diseases* **93** (2020) 211–216.
[15] J.S. Liu and J.S. Liu, Vol. 10 of Monte Carlo strategies in scientific computing. Springer (2001).
[16] J.O. Lloyd-Smith, S.J. Schreiber, P.E. Kopp and W.M. Getz, Superspreading and the effect of individual variation on disease emergence. *Nature* **438** (2005) 355–359.
[17] T.E. Oliphant, vol. 1 of A guide to NumPy. Trelgol Publishing USA (2006).
[18] A. Olivares and E. Staffetti, Uncertainty quantification of a mathematical model of COVID-19 transmission dynamics with mass vaccination strategy. *Chaos Solitons Fractals* **146** (2021) 110895.
[19] K. Sneppen and L. Simonsen, Impact of superspreaders on dissemination and mitigation of COVID-19. *Proc. of the National Academy of Sciences* **118** (2021) e2016623118.
[20] B. Sudret, Global sensitivity analysis using polynomial chaos expansions. *Reliab. Eng. Syst. Saf.* **93** (2008) 964–979.
[21] L. Taghizadeh, A. Karimi and C. Heitzinger, Uncertainty quantification in epidemiological models for the COVID-19 pandemic. *Comput. Biol. Med.* **125** (2020) 104011.
[22] E. Unlu, H. Léger, O. Motornyi, A. Rukubayihunga, T. Ishacian and M. Chouiten, Epidemic analysis of COVID-19 outbreak and counter-measures in France. medRxiv (2020). https://doi.org/10.1101/2020.04.27.20079962
[23] D. Xiu, Numerical methods for stochastic computations: a spectral method approach. Princeton University Press (2010).
[24] D. Xiu and G.E. Karniadakis, The Wiener–Askey polynomial chaos for stochastic differential equations. *SIAM J. Sci. Comput.* **24** (2002) 619–644.