Uncertainty quantification in mechanistic epidemic models via cross-entropy approximate Bayesian computation

Americo Cunha Jr, David A.W. Barton, Thiago Ritto

Nonlinear Dynamics manuscript No.
(will be inserted by the editor)

Uncertainty quantification in mechanistic epidemic models via cross-entropy approximate Bayesian computation

Americo Cunha Jr · David A. W. Barton · Thiago G. Ritto

Received: date / Accepted: date

Abstract This paper proposes a data-driven approximate Bayesian computation framework for parameter estimation and uncertainty quantification of epidemic models, which incorporates two novelties: (i) the identification of the initial conditions by using plausible dynamic states that are compatible with observational data; (ii) learning of an informative prior distribution for the model parameters via the cross-entropy method. The new methodology's effectiveness is illustrated with the aid of actual data from the COVID-19 epidemic in Rio de Janeiro city in Brazil, employing an ordinary differential equation-based model with a generalized SEIR mechanistic structure that includes time-dependent transmission rate, asymptomatics, and hospitalizations. A minimization problem with two cost terms (number of hospitalizations and deaths) is formulated, and twelve parameters are identified. The calibrated model provides a consistent description of the available data, able to extrapolate forecasts over a few weeks, making the proposed methodology very appealing for real-time epidemic modeling.

A. Cunha Jr (Corresponding author)
Rio de Janeiro State University – UERJ, Institute of Mathematics and Statistics, Rio de Janeiro, Brazil
ORCID: 0000-0002-8342-0363
E-mail: americo.cunha@uerj.br

D. A. W. Barton
University of Bristol, Faculty of Engineering, Bristol, UK
ORCID: 0000-0002-0595-4239
E-mail: david.barton@bristol.ac.uk

T. G. Ritto
Federal University of Rio de Janeiro – UFRJ, Department of Mechanical Engineering, Rio de Janeiro, Brazil
ORCID: 0000-0003-0649-6919
E-mail: tritto@mecanica.ufrj.br

Keywords COVID-19 modeling · machine learning · uncertainty quantification · cross-entropy method · ABC inference

1 Introduction

Since the COVID-2019 outbreak became a public information in January 2020 [79], many researchers have contributed with a variety of epidemic models to deal with this epidemic spread. They seek a better understanding of the disease's propagation mechanism and make short-term forecasts to guide public and private agents in related decision-making. In this context, mechanistic compartmental models with classical structures such as the SIR (susceptible, infected and removed), SEIR (susceptible, exposed, infected and removed) [2, 21, 75], or their variants with additional compartments [6, 7, 36, 37, 42] has been widely explored in literature.

These models are exciting tools to aid an epidemiologist since they can explain the past and explore future scenarios for an epidemic outbreak from qualitative and quantitative points of view. Thus they generate insight and can support decision-making. Their balance between simplicity (fast to run simulations) and complexity (good to represent the phenomenology of the problem) may be an advantage for situations where analysis needs to be done in near-real-time (like in an evolving outbreak).

In this direction, Pacheco et al. [46] analyzed an SEIR-type model and investigated different scenarios for Brazil, highlighting the importance of social isolation to avoid a collapse of the hospital infrastructure (in the early COVID-19 outbreak). In the meantime, Vyasarayani and Chatterjee [74] studied an SEIR model with an additional compartment for quarantine, con-

sidering time delays for latency and an asymptomatic phase, while Yu et al. [82] and Cai et al. [4] proposed fractional versions of the SEIR model that capture memory effects of epidemic dynamics.

At this point, it is essential to emphasize that to forecast real-world epidemic outbreaks, especially in real-time, the employed compartmental mechanistic models must be calibrated and validated with actual and reliable data. Often, this identification is a dynamic process, with the models being updated whenever new data becomes available. In addition, uncertainties play a significant role in epidemiological models [60,66]. Values of the model parameters, the model structure, and the epidemic data are uncertain. Thus, beyond identifying values for parameters, it is crucial to perform an uncertainty quantification (UQ) study, which can take into account the variability of the parameters. Probability theory might be used in this endeavor [8,25,62], and the Bayesian learning strategy [27,59,68] is convenient because prior knowledge is updated consistently with data and UQ occurs automatically. Studies dealing with these aspects can be seen in the recent literature of computational epidemiology [20,26,31,32,34–36,83].

For instance, He et al. [20] analyzed an SEIR model with hospitalization and quarantine, using the particle swarm optimization algorithm (a population-based stochastic optimization algorithm) to identify the model parameters from data and considering the stochastic nature of the infection by introducing a Gaussian white noise. Using Poisson and binominal processes to incorporate uncertainty in case observations within an SEIR model, Kucharski et al. [31] describe the dynamics of newly symptomatic cases, reported onsets of new infections, reported confirmation of cases, and the infection prevalence on evacuation flights. A different stochastic system is used by Lobato et al. [35], where a set of stochastic differential equations is employed to describe the random evolution of time-dependent parameters of a compartmental model.

On the other hand, Jha et al. [26] employed a set of partial differential equations governing the spatial-temporal evolution of susceptible, exposed, infectious, recovered, and deceased individuals, considering a strategy for model calibration, validation, and UQ based on Bayesian learning. Within this UQ framework, they considered additive Gaussian noise to construct the likelihood function and assumed log-Normal priors. This UQ approach for computational epidemiology is very general and powerful, being considered the standard methodology to build data-driven mechanistic epidemiological models for use in real-time [32,53]. Variations of this general methodology are also employed by Libotte et al. [34], Lyra et al. [36], Zhang et al. [83] — among many authors — and the general setting is very well described in the excellent book by E. Kuhl [32].

Data-driven epidemic modeling via Bayesian learning has some natural advantages, which stand out: (i) combines the identification of the model parameters (model calibration) and quantifies the effects of underlying uncertainties (uncertainty quantification) into a single framework; (ii) allows new data (information) to be incorporated into the data-driven model in a very straightforward way, via Bayes theorem. However, some weaknesses of this framework often cannot be ignored, such as: (i) the use of sampling techniques like Markov Chain Monte Carlo, which often translates into a computationally intensive process; (ii) the great sensitivity of the inference results to the choice of the prior distribution, which encapsulates a priori knowledge about the model parameters; (iii) the typical difficulty of inferring the initial conditions of the dynamic model when information about these parameters is scarce.

The computational cost can be alleviated in several ways, for example, by exploring parallelization strategies, using surrogate models, or employing approaches that avoid evaluating the likelihood function (typically the most expensive step in the Bayesian inference process), etc. A technique that tackles this problem by replacing the evaluation of the likelihood function with the calculation of a computationally cheaper error metric is known as Approximate Bayesian Computation (ABC) [33,38,40,43]. This is a likelihood-free learning strategy where the prior probability distribution of the parameters is updated, with the aid of available data, only comparing the discrepancy between predictions and observations. Nevertheless, this strategy maintains a prior sensitivity, making informative priors essential, and offers no advantage when inferring initial conditions with reduced information.

To construct an informative prior distribution for the dynamic model parameters, the present work exploits a non-convex optimization technique known as the cross-entropy (CE) method [9,16,29]. This iterative optimization technique starts with an initial probability distribution for the parameters and sequentially updates it, seeking to minimize a cost function that measures the discrepancy between model predictions and data observations, achieving the global optimum asymptotically. This metaheuristic is an exciting methodology to identify parameters in dynamical systems, primarily because of its simplicity and theoretical guarantees of convergence, with excellent results reported in recent literature [11–13,71,76].

The identification of a reasonable initial condition from a small set of information can be made from the knowledge of dynamic states (obtained from the sys-

tem of differential equations) that are compatible with observations of variables accessed via surveillance data. Starting the evolution of the dynamics from a state like this ensures that all temporal variables present, at the initial instant, plausible values, which increases the consistency of the simulation of the epidemic outbreak. The combination of this approach for initial conditions with a Bayesian inference process for the other parameters of the epidemiological model can be advantageous in several real problems of computational epidemiology, thus being the object of interest of this paper.

This paper proposes a novel methodology for calibrating and uncertainty quantifying a mechanistic epidemic model that combines the CE and ABC techniques with a clever strategy for inferring realistic initial conditions. First, the vector of initial conditions is estimated by a combination of dynamic states compatible with observations of the epidemic outbreak. The second step employs the CE method to construct (solving a non-convex optimization problem) an informative prior distribution that represents the parametric uncertainties. Then, it uses ABC to refine (update) the optimal parameter distributions and propagate the underlying uncertainties through the model. It can infer realistic initial conditions with a theoretical guarantee of building an informative a priori distribution. In addition to calibrating/updating the dynamic model, it also considers the effects of the parametric uncertainties underlying the problem. To the best of the authors' knowledge, this formulation of the Bayesian learning process for epidemiological models combining CE and ABC has not yet been explored, contributing towards improving the methodology's inference capacity and, thus, to developing a robust framework UQ on mechanistic epidemic models. The proposed methodology's effectiveness is illustrated with an SEIR-type epidemic model with seven compartments (susceptible, exposed, infectious, asymptomatic, hospitalized, recovered, and deceased) [47], and actual data from the city of Rio de Janeiro.

The paper is organized as follows. Section 2 depicts the mechanistic epidemic model. The proposed methodology that combines CE with ABC is presented in section 3. The results are shown in section 4. The manuscript body is closed with the concluding remarks in section 5.

2 SEIR(+AHD) epidemic model

2.1 Modeling of the contagion process

The compartmental model employed in this work to describe a COVID-19 outbreak in Rio de Janeiro city is schematically illustrated in Figure 1, where the population is segmented into seven disjoint compartments: susceptible (S); exposed (E); infectious (I); asymptomatic (A); hospitalized (H); recovered (R); deceased (D). This model is dubbed here as the SEIR(+AHD) model.

In this dynamic contagion model, the infection spreads via direct contact between a susceptible and an infected (infectious, asymptomatic, or hospitalized) individual. For simplicity, it is assumed that infectious and asymptomatic individuals are equally likely to transmit the disease to a susceptible person, while this risk is reduced in hospitalized individuals. The latency period between a person becoming infected, starting to have symptoms, and transmitting the disease, is taken into account by the presence of an exposed compartment, which counts those individuals who, despite carrying the pathogen, still do not show symptoms nor can infect other people. Among the infected, some individuals are asymptomatic; only a fraction display symptoms after incubation; they are dubbed infectious. Asymptomatic individuals can recover or die (a rare event). On the other hand, infectious individuals, in addition to recovery and death, may result in hospitalization. Hospitalized people reduce their probability of dying from the disease, but they can still have this outcome or recover. The recovered compartment is just an accumulator receiving individuals from various groups but does not directly interfere with the dynamics. This model was proposed by Pavack et al. [47], who were inspired by the age-structured model presented in [36], and its variant which considers ICU admissions presented in [45].

The population in each of the compartments at time t is measured by the following state variables: susceptible $S(t)$; exposed $E(t)$; infectious $I(t)$; asymptomatic $A(t)$; hospitalized $H(t)$; recovered $R(t)$; and deceased $D(t)$. Variable $N = N(t)$ represents the alive population size at time t. This contagion model has the following parameters: initial alive population N_0 (number of individuals); transmission rate β (days^{-1}); latent rate α (days^{-1}); fraction of symptomatic f_E (non-dimensional); recovery rate γ (days^{-1}); hospitalization rate ρ (days^{-1}); death rate δ (days^{-1}); asymptomatic mortality-factor κ_A (non-dimensional); hospitalization mortality-factor κ_H (non-dimensional); and hospitalization infectivity-factor ϵ_H (non-dimensional).

The deterministic non-autonomous dynamical system associated to this compartmental model (see [47]

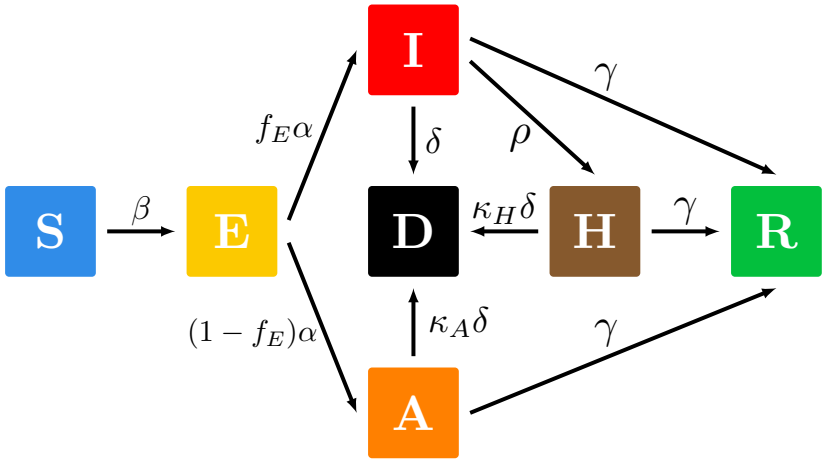

Fig. 1 Schematic representation of the SEIR(+AHD) compartmental model considering latency period, asymptomatic individuals, hospitalizations, and deaths. This model is used here to describe the COVID-19 dynamics.

for details) is written as

$$\begin{aligned}
\dot{S} &= -\beta(t)\, S\,(I + A + \epsilon_H\, H)/N\,,\\
\dot{E} &= \ \ \beta(t)\, S\,(I + A + \epsilon_H\, H)/N - \alpha\, E\,,\\
\dot{I} &= f_E\,\alpha\, E - (\gamma + \rho + \delta)\, I\,,\\
\dot{R} &= \gamma\,(I + A + H)\,,\\
\dot{A} &= (1 - f_E)\,\alpha\, E - (\kappa_A\,\delta + \gamma)\, A\,,\\
\dot{H} &= \rho\, I - (\gamma + \kappa_H\,\delta)\, H\,,\\
\dot{D} &= \delta\,(I + \kappa_A\, A + \kappa_H\, H)\,,\\
\dot{N} &= -\dot{D}\,,
\end{aligned} \tag{1}$$

where the corresponding initial conditions are given by $\boldsymbol{u}(0) = (S_0, E_0, I_0, A_0, H_0, R_0, D_0, N_0)$. Obviously, if convenient, the size of this system can be reduced by one unit if the last equation is replaced by the algebraic constraint $N = N_0 + D_0 - D$, which represents the total population evolution over time.

2.2 Time dependence of the transmission rate

As the disease spreads, the parameter β might change, and this temporal dependence can be taken into account through the following expression (taken from [72]):

$$\beta(t) = \beta_0 + \frac{(\beta_\infty - \beta_0)}{2}\left(1 + \tanh\left(\eta\,\frac{(t - t_\beta)}{2}\right)\right)\,, \tag{2}$$

where β_0 is the initial value of β, β_∞ the final value, the adaptation time η defines how fast β reaches β_∞, and t_β is the transition time (when $t = t_\beta$ then $\beta = (\beta_0 + \beta_\infty)/2$). This model allows β to smoothly vary between two distinct levels of disease transmission (from lower to higher, or vice versa), a situation typically encountered in the COVID-19 contagion dynamics [18,72].

2.3 Associated dynamic system

The dynamic state of the epidemic system (1) at time t can be represented, in a compact way, by the time-dependent vector

$$\boldsymbol{u}(t) = (S, E, I, A, H, R, D, N)\,, \tag{3}$$

while the model parameters may be lumped into the parameter vector

$$\boldsymbol{x} = (\beta_0, \alpha, f_E, \gamma, \rho, \delta, \kappa_A, \kappa_H, \epsilon_H, \beta_\infty, \eta, t_\beta)\,, \tag{4}$$

so that the dynamic model can be written as

$$\dot{\boldsymbol{u}}(t) = F(t,\, \boldsymbol{u}(t),\, \boldsymbol{x})\,, \tag{5}$$

where the map $(t,\, \boldsymbol{u}(t),\, \boldsymbol{x}) \mapsto F(t,\, \boldsymbol{u}(t),\, \boldsymbol{x}) \in \mathbb{R}^8$ represents the nonlinear evolution law defined by the right hand side of the dynamical system in (1).

2.4 Applicability and limitations

The system of differential equations defined in (1) gives rise to a predictive computational model to describe the dynamics of COVID-19 contagion in a context where tracking the number of hospitalizations is essential.

Such a model can help study possible epidemiological scenarios and may lead to qualitative and quantitative insights into the epidemic dynamics. Such information can help guide decision-makers in managing their local health system. For instance, the model can check

whether there is a risk of overloading the hospitals in a particular city. Also, they could estimate when they will suffer the most significant demand. From a more qualitative perspective, the model can assess the impact on hospitalizations of different strategies to mitigate (or even suppress) the epidemic.

But like any computational model, it is subject to limitations, and its use outside the proper context can lead to entirely erroneous predictions [69]. Since it is a deterministic compartmental model, obtained as a mean-field approximation in the thermodynamic limit, it is only applicable in regions where the population density can be modeled as a continuous function, a situation typically valid in urban centers of large cities. As it is a model derived from the SEIR family, it assumes a population with a homogeneous contact structure, which does not correspond to the reality of practically anywhere. Therefore, one must care about the potential effects of population heterogeneity.

Other unmodeled effects that may be significant are related to social behavior change due to the course of the epidemic (e.g. risk perception, change in the pattern of social iterations, mask use, etc.) [78], reinfections [51], etc. These can be included in the model, but this is not the goal of the present paper.

3 Uncertainty quantification framework

3.1 Quantities of interest

Among all the quantitative information that can be estimated with the epidemic model in (1), this paper is particularly interested in two time-dependent quantities, the number of hospitalizations, and the total number of deaths, i.e., the quantities of interest (QoIs) here are the time series $H(t)$ and $D(t)$.

None of these time series, individually or together, correspond to the response of the dynamic model itself. The model response is given by the parametric curve $t \mapsto \boldsymbol{u}(t)$, so that the above time series correspond to a derived quantity $t \mapsto (H(t), D(t))$, extracted from $\boldsymbol{u}(t)$ through a projection.

On the theoretical plane, $t \mapsto (H(t), D(t))$ is defined over a continuous-time domain and, consequently, is an infinite-dimensional object. However, for computational purposes, it is necessary to discretize both time-series so that, in practice, the dynamic model returns finite-dimensional representations of them. Once the computational representation of each time series materializes itself in the form of an n-dimensional numerical sequence, one might think that the model's discrete response is given by the quantities of interest vector

$$\boldsymbol{y} = [H(t_1), \cdots, H(t_n), D(t_1), \cdots, D(t_n)]\,, \tag{6}$$

where $t_1, \cdots, t_n$ are the time-instants underlying the temporal discretization. If other observables become the quantities of interest, the vector $\boldsymbol{y}$ can be modified straightforwardly. Just as if observables defined in different temporal grids are needed.

3.2 Abstraction of the epidemic model

In an abstract perspective, the computational model can be represented by an equation of form

$$\boldsymbol{y} = \mathcal{M}(\boldsymbol{x})\,, \tag{7}$$

which indicates that the vector $\boldsymbol{y}$ is obtained from the vector of parameters $\boldsymbol{x}$ through a mapping $\mathcal{M}$, which represents the discretized version of the dynamic model that is coded in the computer. Therefore, whenever convenient, the notation $\boldsymbol{y}(\boldsymbol{x})$ is adopted.

Furthermore, if is necessary to distinguish the components of $\boldsymbol{y}$ that are related to $H(t)$ from those associated with $D(t)$, the following partition is adopted

$$\boldsymbol{y}(\boldsymbol{x}) = [\boldsymbol{y}^H(\boldsymbol{x})\ \boldsymbol{y}^D(\boldsymbol{x})]\,. \tag{8}$$

In a case where there are K quantities of interest, the response vector is written

$$\boldsymbol{y}(\boldsymbol{x}) = [\boldsymbol{y}^1(\boldsymbol{x})\ \cdots\ \boldsymbol{y}^K(\boldsymbol{x})]\,. \tag{9}$$

This abstract representation helps to simplify the formulation of the uncertainty quantification framework presented in sequence.

3.3 Data from the epidemic surveillance system

Epidemiological surveillance data, represented in this paper by the vector quantity $\boldsymbol{y}_{data}$, can be used to monitor and understand (in real-time or a posterior) the course of an epidemic through direct observation, or in conjunction with computational models that can be calibrated and validated against them.

In this work, the data used are related to the records of hospitalizations and deaths due to COVID-19 in the city of Rio de Janeiro city, from Jan 01, 2020, until Dec 31, 2020. The choice of data from this period, rather than more recent observations, is motivated by the fact that this was one of the critical moments of the COVID-19 pandemic in the city of Rio de Janeiro, with high pressure in both health and funeral systems. These

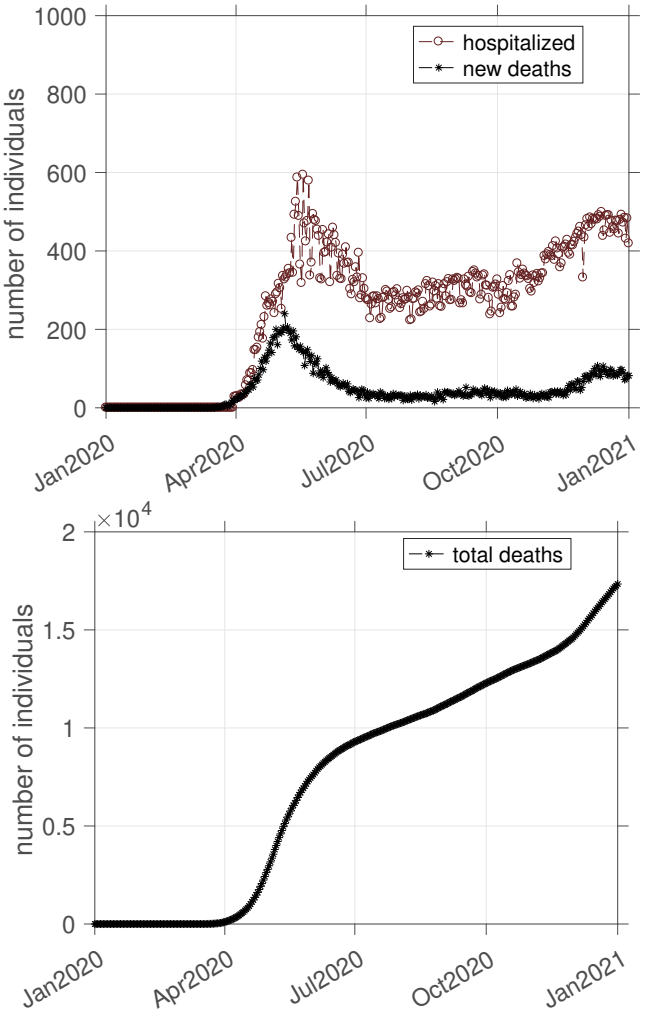

Fig. 2 Surveillance data of COVID-19 outbreaks in Rio de Janeiro city between Jan 01, 2020, until Dec 31, 2020 [49]. The number of hospitalized individuals and new deaths appears at the top and the total number of deaths at the bottom.

data, shown in Figure 2, are cataloged and made available by local health authorities [49], being anonymous for ethical reasons and patient privacy.

For the sake of compatibility with the structure of $\boldsymbol{y}(\boldsymbol{x})$, the data vector $\boldsymbol{y}_{data}$ which lumps hospitalizations and total deaths (or simply deaths) time series is partitioned as follows

$$\boldsymbol{y}_{data} = [\boldsymbol{y}^H_{data} \;\; \boldsymbol{y}^D_{data}] \,. \tag{10}$$

A combination between these data and the epidemic model predictions is done in the uncertainty quantification framework presented below.

3.4 Quantification of the discrepancy between the mathematical model and available data

The comparison between data and predictions can be done by means of the following discrepancy (error estimation) function

$$\mathcal{J}(\boldsymbol{x}) = \omega \, \frac{||\boldsymbol{y}^H_{data} - \boldsymbol{y}^H(\boldsymbol{x})||^2}{||\boldsymbol{y}^H_{data}||^2} \;\; + \;\; (1-\omega) \, \frac{||\boldsymbol{y}^D_{data} - \boldsymbol{y}^D(\boldsymbol{x})||^2}{||\boldsymbol{y}^D_{data}||^2} \,, \tag{11}$$

where $\omega \in [0,1]$ is a weight parameter which controls how the hospitalization/deaths data contributes to this discrepancy function. For $\omega = 0$, only death data are taken into account. Conversely, for $\omega = 1$ only hospitalization data matter. Between these extremes the error metric considers a balance between the two data sets. If $\omega = 0.5$ they have the same weight. It is worth mentioning that we could identify ω together with the other model parameters, including it in vector $\boldsymbol{x}$, possibility not explored in this paper.

In the case where K quantities of interest are available in the form of data, partitioned as

$$\boldsymbol{y}_{data} = [\boldsymbol{y}^1_{data} \;\; \cdots \;\; \boldsymbol{y}^K_{data}] \,, \tag{12}$$

so that the model response reads as in Eq.(9), and the discrepancy function is written as

$$\mathcal{J}(\boldsymbol{x}) = \sum_{k=1}^{K} \omega_k \, \frac{||\boldsymbol{y}^k_{data} - \boldsymbol{y}^k(\boldsymbol{x})||^2}{||\boldsymbol{y}^k_{data}||^2} \,, \tag{13}$$

with the weights defining a convex combination, i.e.,

$$\omega_1 + \cdots + \omega_K = 1 \,. \tag{14}$$

3.5 Baseline calibration of model parameters via the Cross-Entropy (CE) method for optimization

The process of calibrating the computational model against the available data requires that the discrepancy function defined by Eq.(11) (or by Eq.(13)) be minimized by an optimal choice of parameters, a task that is mathematically formulated as the following optimization problem

$$\boldsymbol{x}^\star = \arg \min_{\boldsymbol{x} \, \in \, \mathcal{X}} \mathcal{J}(\boldsymbol{x}) \,, \tag{15}$$

where the set of admissible parameters is defined by

$$\mathcal{X} = \{ \; \boldsymbol{x} \mid \boldsymbol{x}_{min} \preceq \boldsymbol{x} \preceq \boldsymbol{x}_{max} \; \} \,, \tag{16}$$

where $\boldsymbol{x}_{max}$ and $\boldsymbol{x}_{min}$ represent, respectively, upper and lower bound vectors for the model parameters, and the generalized inequality $\preceq$ is understood to hold for each component of the vectors.

In general, the optimization problem defined by (15) is nonconvex, so the use of gradient-based techniques may not be effective in capturing the parameter configuration that best fits the model to the data. Due to the nonconvexity, the solution obtained may be a local optimum (perhaps quite distinct from the global optimum). To avoid this type of situation, the present work tackles this optimization problem with the aid of the cross-entropy method [16,54,56], a simplistic gradient-free iterative procedure for global optimization that has guarantees of convergence in certain typical situations [55,56]. This method has been successfully used in recent literature for the identification of parameters in nonlinear computational models [11–13,71,76].

The fundamental idea of the cross-entropy method is to transform the optimization problem, defined by (15), into a rare event estimation problem. In this way, a sequence of approximations to the global optimum is constructed with the aid of an importance sampling process, where the probability of sampling the rare event (global optimum) grows over time [55,56].

The CE algorithm consists of two phases:

- Sampling – where the objective function domain is sampled according to a certain distribution to explore the feasible region;
- Learning – where the distribution parameters are updated, with the aid of a set of elite samples. The goal is to shrink the distribution simultaneously as it is translated towards the global optimum.

To mathematically analyze the behavior of this algorithm, let $\Gamma^\star = J(\boldsymbol{x}^\star)$ be the global minimum sought, and $\boldsymbol{X}$ a randomized version of $\boldsymbol{x}$, with probability distribution characterized by the probability density function (PDF) $p(\cdot, \boldsymbol{v})$, i.e., $\boldsymbol{X} \sim p(\cdot, \boldsymbol{v})$, for a parameter vector $\boldsymbol{v} = (\boldsymbol{\mu}, \boldsymbol{\sigma})$, with mean vector $\boldsymbol{\mu}$ and standard deviation vector $\boldsymbol{\sigma}$.

Since $\Gamma^\star$ is the global optimal value of $\mathcal{J}$, there are few points $\boldsymbol{x}$ in the domain $\mathcal{X}$ that produce a value $\Gamma = \mathcal{J}(\boldsymbol{x})$ very close to $\Gamma^\star$. In this way,

$$\mathcal{P}\left\{\mathcal{J}(\boldsymbol{x}) \leq \Gamma\right\} \approx 0 \;\text{ for } \Gamma \approx \Gamma^\star\,, \tag{17}$$

which states that $\mathcal{J}(\boldsymbol{x}) \leq \Gamma$ is a rare event for $\Gamma \approx \Gamma^\star$.

The solution to this rare event estimation problem involves sampling the domain $\mathcal{X}$ with N_{ce} samples – drawn according the distribution $p(\cdot, \boldsymbol{v})$, evaluate the objective function at the samples $\boldsymbol{X}_k$, and then construct of a sequence of estimators $(\hat{\Gamma}_\ell, \hat{\boldsymbol{v}}_\ell)$ such that

$$\hat{\Gamma}_\ell \xrightarrow{a.s.} \Gamma^\star \;\text{ and }\; p\left(\boldsymbol{x}, \hat{\boldsymbol{v}}_\ell\right) \xrightarrow{a.s.} \delta\left(\boldsymbol{x} - \boldsymbol{x}^\star\right), \tag{18}$$

where the parameter vector $\boldsymbol{v} = (\boldsymbol{\mu}, \boldsymbol{\sigma})$ is updated by solving the following nonlinear program

$$\hat{\boldsymbol{v}}_\ell = \arg\max_{\boldsymbol{v}} \sum_{k \,\in\, \mathcal{E}_\ell} \mathbf{1}\left\{\mathcal{J}(\boldsymbol{X}_k) \leq \hat{\Gamma}_\ell\right\} \ln p\left(\boldsymbol{X}_k;\, \boldsymbol{v}\right), \tag{19}$$

being $\mathbf{1}\{\cdot\}$ the indicator function, and $\mathcal{E}_\ell$ an elite sample set, defined by a fixed percentage of the samples $\boldsymbol{X}_k$ that produced the values closest to the global optimum. Among the values associated with the elite set, the largest one defines the estimator $\hat{\Gamma}_\ell$.

The above sequence is optimal in the sense that the importance sampling process tries to minimize the Kullback-Leibler divergence (also known as the cross-entropy function) between the sampling distribution $p\left(\cdot;\, \boldsymbol{v}\right)$ and a Dirac delta function centered on the global optimum [55,56].

For sake of numerical implementation, $p\left(\cdot,\, \boldsymbol{v}\right)$ is assumed as a truncated Gaussian distribution with bounds defined in a conservative way (assuming broad intervals for the random variables supports). Distributions from the exponential family, like truncated Gaussian, allow the nonlinear program from Eq.(19) to be solved analytically [55,56], so that the low-order statistics in the parameters vector $\boldsymbol{v} = (\boldsymbol{\mu}, \boldsymbol{\sigma})$ are updated by simply calculating the sample mean and sample standard deviation from the elite sample set $\mathcal{E}_\ell$.

From a theoretical point of view, the process described above is guaranteed to converge to the global optimum [55,56]. However, in practical terms, the distribution may degenerate before it gets close enough to this optimum point [29,56]. To avoid this situation, the following smoothing scheme is employed

$$\hat{\boldsymbol{\mu}}_\ell := a\,\hat{\boldsymbol{\mu}}_\ell + (1-a)\,\hat{\boldsymbol{\mu}}_{\ell-1}\,, \tag{20}$$

$$\hat{\boldsymbol{\sigma}}_\ell := b_\ell\,\hat{\boldsymbol{\sigma}}_t + (1-b_\ell)\,\hat{\boldsymbol{\sigma}}_{\ell-1}\,, \tag{21}$$

$$b_\ell = b - b\left(1 - \frac{1}{\ell}\right)^q, \tag{22}$$

for a set of smooth parameters such that $0 < a \leq 1$, $0.8 \leq b \leq 0.99$ and $5 \leq q \leq 10$ [29,56], with the estimations at iterations ℓ and $\ell - 1$ obtained by solving the Eq.(19) analytically.

The convergence of the sampling process is controlled by the test

$$||\boldsymbol{\sigma}_\ell - \boldsymbol{\sigma}_{\ell-1}||_w \leq 1\,, \tag{23}$$

where the weighted root-mean-square norm of the difference vector $\boldsymbol{x} - \boldsymbol{y} \in \mathbb{R}^N$ is defined as

$$||\boldsymbol{x} - \boldsymbol{y}||_w = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(w_j\left(x_j - y_j\right)\right)^2}\,, \tag{24}$$

for the error weights

$$w_j = \frac{1}{\texttt{atol}_j + 0.5\,|x_j + y_j|\,\texttt{rtol}}\,, \tag{25}$$

with $\texttt{atol}_j$ and $\texttt{rtol}$ denoting absolute and relative tolerances, respectively. Due to the normalization provided by the weights of Eq.(25), a weighted norm of the order of 1 in (23) can be considered small. This type of convergence test, frequently used in the best differential equation solvers [22,58], provides robust error control.

An overview of the cross-entropy method can be seen in Figure 3. More details about the CE implementation can be seen in Algorithm 1, section 3.7, and in the references [9,55,56].

3.6 Model update and uncertainty quantification through Approximate Bayesian Computation (ABC)

The update of the model calibration process involves a Bayesian inference scheme [27,53,67], where the data set $\boldsymbol{y}_{data}$ and a prior distribution for the parameters $\pi(\boldsymbol{X})$ are combined with the aid of a likelihood function $\pi(\boldsymbol{y}_{data} \mid \boldsymbol{X})$ to estimate a posterior parameter distribution $\pi(\boldsymbol{X} \mid \boldsymbol{y}_{data})$ through Bayes' theorem

$$\pi(\boldsymbol{X} \mid \boldsymbol{y}_{data}) \propto \pi(\boldsymbol{y}_{data} \mid \boldsymbol{X})\ \pi(\boldsymbol{X})\,, \tag{26}$$

which combines prior information and available data in an optimal way [67,68].

For inference purposes in this setting, the approximate Bayesian computation (ABC) scheme proposed by Toni et al. [70] is employed. A likelihood function form is not assumed, so the usual hypothesis of additive independent Gaussian noise is unnecessary. Alternatively, the model prediction and the epidemic data are directly compared with the aid of a discrepancy function $\mathcal{J}(\boldsymbol{x})$ — such as those defined by Eq.(11) or Eq.(13) — to measure the representation quality of the drawn model. Monte Carlo simulation [10,29], employing an acceptance-rejection sampling strategy, is used in the inference process, in a way that a sample $\boldsymbol{X}_k$ drawn from the prior distribution $\pi(\boldsymbol{X})$ is accepted only if $|\mathcal{J}(\boldsymbol{X}_k)| < \texttt{tol}$, where $\texttt{tol}$ is a (problem-dependent) tolerance prescribed by the user. Once the discrepancy function of Eq.(11) is defined as a kind of relative error, it is not necessary to employ two tolerances to control the convergence of the ABC process, as done in the case of CE. But for other definitions of $\mathcal{J}(\boldsymbol{x})$ this kind of convergence criterion may be helpful.

The good practice of this technique dictates that all known information about the model parameters should be encapsulated into a prior distribution $\pi(\boldsymbol{X})$, to obtain an informative inference process.

Typically, the iterative process of the CE method provides a lot of information about the parameters, so it is beneficial to take advantage of this knowledge to build the prior. Therefore, the methodology proposed in this paper adopts as prior distribution, for the ABC inference step, the truncated Gaussian distribution with support bounds $\boldsymbol{x}_{min}$, and $\boldsymbol{x}_{max}$, central tendency $\boldsymbol{\mu}$, and dispersion information $\boldsymbol{\sigma}$ that comes from the last iteration of CE algorithm, i.e.,

$$\pi(\boldsymbol{X}) \sim \mathcal{TN}(\boldsymbol{\mu}, diag(\boldsymbol{\sigma}), \boldsymbol{x}_{min}, \boldsymbol{x}_{max})\,. \tag{27}$$

It is important to note that although this prior distribution is defined in the same (broad) region where the initial truncated Gaussian of the CE method was defined, it is much more informative. Despite the support limits being kept invariant, the central tendency encapsulated in the mean $\boldsymbol{\mu}$, and the dispersion defined by the standard deviation $\boldsymbol{\sigma}$ are updated by the CE iteration several times, obtaining a substantial gain of information in this process.

Conceptually, the posterior distribution for the computational model parameters $\pi(\boldsymbol{X} \mid \boldsymbol{y}_{data})$ is obtained from the samples accepted in the acceptance/rejection process, through some technique of statistical inference. Armed with this probabilistic distribution, in theory any statistical information about the model parameters can be obtained, as well as the intrinsic uncertainty of the parameters can be propagated to the response of the dynamic system. In practice, this distribution is not always inferred, and it is very common that only partial statistical characterizations (e.g. low-order moments) are calculated, or that the accepted samples are used directly in the uncertainty propagation process that follows the definition of the distribution of the model input.

It should be noted at this point that the calibration process described above concerns only the coefficients of the system of differential equations that define the epidemiological model. The initial conditions of the dynamics were not considered. They are inferred by a heuristic process, guided by intuition about the behavior of nonlinear systems, which is described in section 4.2.

An overview of the ABC can be seen in Figure 4. Further details are available in Algorithm 1 from section 3.7, and in the references [33,38,40,43].

3.7 The novel metaheuristic CE-ABC framework for model calibration and uncertainty quantification

The combination of CE and ABC gives rise to a novel algorithm for epidemic model calibration and UQ. A

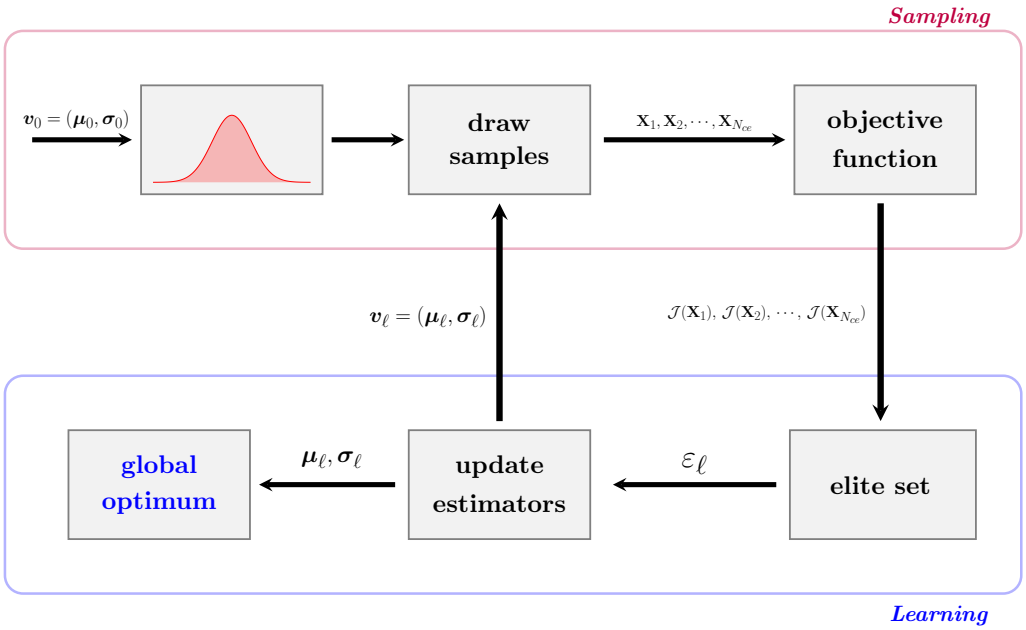

Fig. 3 Schematic representation of the two-phases CE algorithm: (i) sampling – where the domain is sampled according to a given distribution to explore the feasible region, and (ii) learning – where the distribution parameters are updated with the aid of an elite set, to improve the optimum estimation.

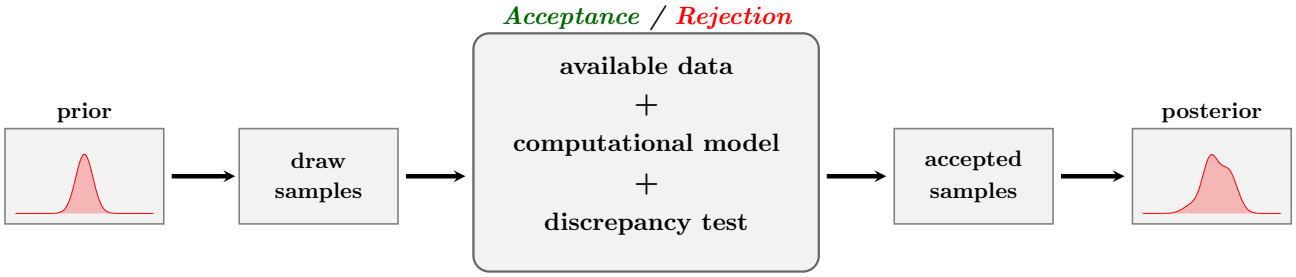

Fig. 4 Schematic representation of the ABC algorithm. An a priori distribution is used to generate samples that are selected, if a discrepancy function is small, in an acceptance/rejection process, to generate a posterior distribution.

schematic representation of this new UQ framework, called here CE-ABC, can be seen in Figure 5, where the available data (from epidemic surveillance system) and the computational model — defined by Eq.(1) — are combined to evaluate the discrepancy function $\mathcal{J}(\boldsymbol{X})$ — defined by Eq.(11).

First, a truncated Gaussian distribution, defined with aid of conservative (board) bounds and informative values for central tendency, is used by CE method to sample the domain and obtain a first informative estimation for the model parameters values. After the convergence of this iterative process, the updated truncated Gaussian is used to define a prior distribution to be used in the ABC algorithm. Then the ABC combine this informative prior distribution with data, using a discrepancy function, to obtain a posterior distribution of the model parameters. The accepted samples from the Monte Carlo sampling, which define the posterior, are also used to draw credible envelopes. Other statistical information (e.g. low-order moments, MAP, etc) may be obtained in the same way. The computational recipe for the CE-ABC procedure is shown in Algorithm 1.

3.8 Remarks about the CE-ABC algorithm

The novel CE-ABC framework presented here is based on two general statistical methods which have already been applied to several complex problems [40, 56, 65, 70]. The resulting algorithm inherits two interesting properties from these methods: (i) from CE, the guarantee of convergence to the global optimum in typical situations; and (ii) from ABC, the likelihood shape independence and relative computational efficiency, when compared to approaches that require a direct assessment of the likelihood function. Such a mixture of good properties generates a robust framework for stochastic simulations

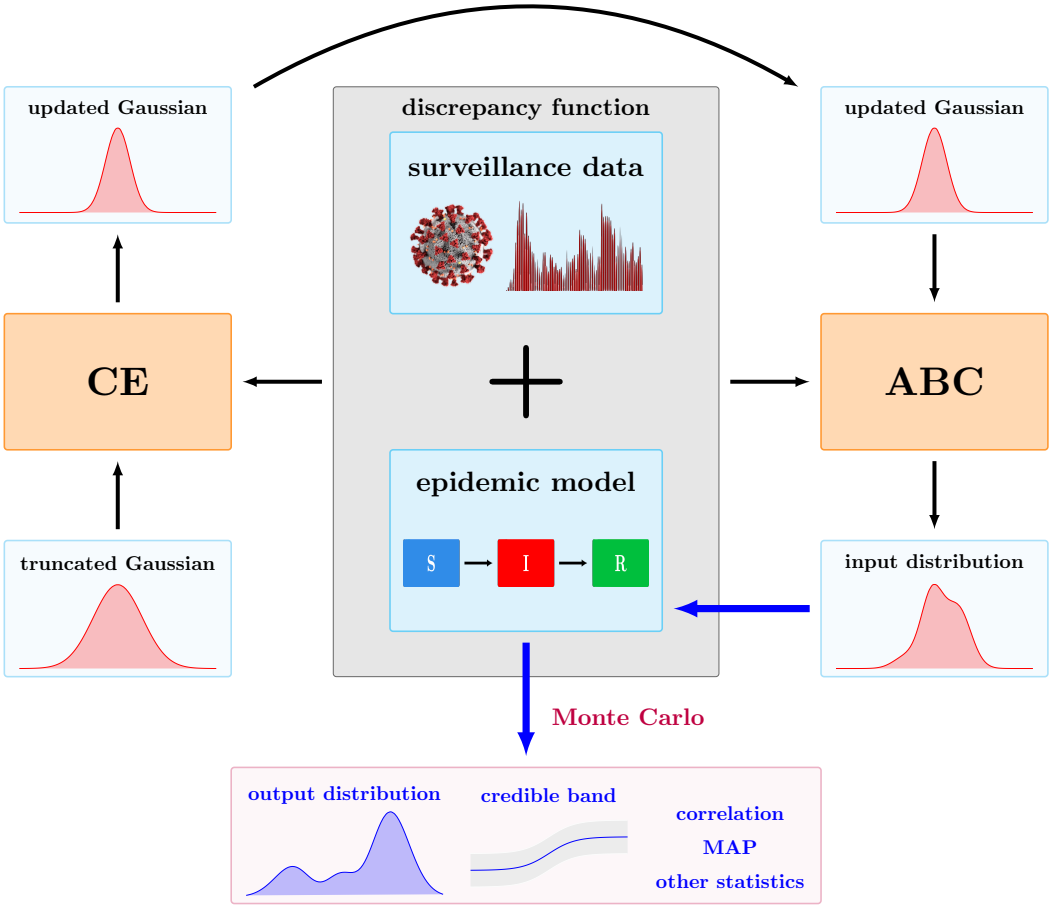

Fig. 5 Schematic representation of the CE-ABC framework for estimating parameters and uncertainty quantification in mechanistic epidemic models. First, starting from a truncated Gaussian distribution, an estimate for the model parameters is obtained with the cross-entropy (CE) method. Then, the estimation of the parameters is refined through an inference process employing approximate Bayesian computation (ABC), which also propagates the uncertainties through an acceptance-rejection Monte Carlo simulation to obtain, in the end, a statistical characterization of the model output uncertainty.

involving epidemic models, which are typically difficult to calibrate and have a limited predictability horizon, requiring quantification of uncertainties for any minimally reliable forecast.

Although the good theoretical properties of the CE-ABC framework are observed in the numerical studies developed by the authors with the epidemic model employed in this paper, its use in conjunction with other types of computational models (e.g. computational mechanics) requires a more comprehensive theoretical analysis. Such a formal analysis for a broad class of models is beyond the scope of this work and the present journal, but it would be a fascinating work on applied mathematics, which the authors leave as a suggestion for future work.

Due to the generality of CE and ABC methods, but in a context where a rigorous mathematical analysis to ensure algorithm functionality for a broad class of computational models is missing, the authors consider that the proposed CE-ABC framework is a metaheuristic[1] for model calibration and UQ.

Despite the CE-ABC algorithm's excellent convergence properties, it is impossible to make an accurate inference if "bad" values (physically/biologically inconsistent or very discrepant with reality) are assigned to the model's nominal parameters, initial conditions, and bounds. Defining the bounds and nominal values for the parameters and initial conditions is an important task that must be done carefully. It is necessary to have biological intuition (in parallel to the importance of physical intuition when in the context of computational physics). The analyst's experience with the problem of interest is essential; it is a kind of expert knowledge that must be embedded into the priors distributions. Besides, exploratory tests with the computational model

[1] A technique for efficiently solving a computational problem (approximately) that is generally suboptimal in some sense for practical use.

Algorithm 1 CE-ABC is a metaheuristic that combines CE and ABC for parameters estimation and uncertainty quantification in mechanistic epidemic models. It receives as input the computational model $\mathcal{M}$, the discrepancy function $\mathcal{J}$, the sampling distributions bounds $\boldsymbol{x}_{min}$ and $\boldsymbol{x}_{max}$, the number of CE samples N_{ce}, the elite sample set size $N_{\mathcal{E}_\ell}$, the number of ABC samples N_{abc}, an absolute and a relative tolerance for CE `atol` and `rtol`, a tolerance for ABC `tol`, and an upper bound for CE iterations `maxiter`. The algorithm returns the best parameter estimate obtained by both CE and ABC, and the samples accepted during the ABC iteration process.

procedure CE-ABC($\mathcal{M}$, $\mathcal{J}$, $\boldsymbol{x}_{min}$, $\boldsymbol{x}_{max}$, N_{ce}, $N_{\mathcal{E}_\ell}$, N_{abc}, `atol`, `rtol`, `tol`, `maxiter`)
Require: $\boldsymbol{x}_{min} \preceq \boldsymbol{x}_{max}$
Require: $N_{\text{ce}} > 0$
Require: $N_{\text{abc}} > 0$
Require: $N_{\mathcal{E}_\ell} > 0$ and $N_{\mathcal{E}_\ell} < N_{\text{ce}}$
Require: `atol` ≥ 0
Require: `rtol` > 0
Require: `tol` > 0
Require: `maxiter` > 0
//- - - - - CE step - - - - -//
$\ell := 0$
$\boldsymbol{\mu} := (\boldsymbol{x}_{max} + \boldsymbol{x}_{min})/2$
$\boldsymbol{\sigma} := (\boldsymbol{x}_{max} - \boldsymbol{x}_{min})/\sqrt{12}$
Draw $\boldsymbol{X} \sim \mathcal{TN}(\boldsymbol{\mu}, diag(\boldsymbol{\sigma}), \boldsymbol{x}_{min}, \boldsymbol{x}_{max})$ // total of N_{ce} samples
while $||\boldsymbol{\sigma}_\ell - \boldsymbol{\sigma}_{\ell-1}||_w > 1$ and $\ell <$ `maxiter` **do**
  $\ell := \ell + 1$
  Evaluate $\boldsymbol{Y}_k = \mathcal{M}(\boldsymbol{X}_k)$ for $k = 1 : N_{ce}$
  Evaluate $\mathcal{J}(\boldsymbol{X}_k)$ for $k = 1 : N_{ce}$
  Define elite sample set $\mathcal{E}_\ell$
  Update $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ using $N_{\mathcal{E}_\ell}$ samples from $\mathcal{E}_\ell$
end while
//- - - - - ABC step - - - - -//
$\mathcal{J}_{min} = \infty$
$\boldsymbol{X}_{best} =$ `NaN`
$\boldsymbol{Y}_{best} =$ `NaN`
Define prior $\pi(\boldsymbol{X}) = \mathcal{TN}(\boldsymbol{\mu}, diag(\boldsymbol{\sigma}), \boldsymbol{x}_{min}, \boldsymbol{x}_{max})$
Draw $\boldsymbol{X} \sim \pi(\boldsymbol{X})$ // total of N_{abc} samples
for $k = 1 : N_{\text{abc}}$ **do**
  Evaluate $\boldsymbol{Y}_k = \mathcal{M}(\boldsymbol{X}_k)$
  Evaluate $\mathcal{J}(\boldsymbol{X}_k)$
  if $\mathcal{J}(\boldsymbol{X}_k) <$ `tol` **then**
    Accept $\boldsymbol{X}_k$
    Save $\boldsymbol{X}_k$ and $\boldsymbol{Y}_k$
    if $\mathcal{J}(\boldsymbol{X}_k) < \mathcal{J}_{min}$ **then**
      $\boldsymbol{X}_{best} := \boldsymbol{X}_k$
      $\boldsymbol{Y}_{best} := \boldsymbol{Y}_k$
      $\mathcal{J}_{min} := \mathcal{J}(\boldsymbol{X}_k)$
    end if
  else
    Reject $\boldsymbol{X}_k$
  end if
end for
Return $(\boldsymbol{X}, \boldsymbol{Y})^{ce}_{opt}$, $(\boldsymbol{X}, \boldsymbol{Y})^{abc}_{best}$, and $(\boldsymbol{X}, \boldsymbol{Y})^{abc}_{saved}$
end procedure

and information from previous works may be precious to discover a suitable interval of values.

The results obtained with the CE-ABC framework also strongly depend on the tolerances `atol`, `rtol`, and `tol`, chosen by the user. There are no canonical values for these parameters that are valid for all types of inference; good values are problem-dependent. In this way, the analyst's experience and intuition are crucial in defining these values and a little numerical experimentation with the computational model. In the numerical experiments repeated in the session 4 these tolerances are defined as being `atol` $= 0.001$, `rtol` $= 0.05$, and `tol` $= 0.1$.

Once again, it is worth emphasizing the observations made in section 2.4 about the limitations and applicability of the model. No matter how robust the calibration and uncertainty propagation algorithm is, how good is the choice of model parameters and bounds. If the model does not describe the reality in a minimally reliable way, terribly wrong (or in the limit nonsense)

Table 1 Plausible nominal values and bounds for the parameters of the SEIR(+AHD) epidemic model.

	Unit	Nominal	Min	Max	Refs
β or β_0	1/day	1/7	1/14	1/2	[39,80]
α	1/day	1/5	1/10	1/2	[5,39]
f_E	—	0.8	0.7	0.9	[3,39,44]
γ	1/day	1/14	1/21	1/7	[39,84]
ρ	1/day	1/700	1/2100	1/100	[39,77,84]
δ	1/day	1/14000	1/21000	1/100	[63]
κ_A	—	0.0010	0.0005	0.0050	—
κ_H	—	0.05	0.01	0.10	[19,39,80]
ϵ_H	—	0.2	0.1	0.5	—
β_∞	1/day	1/7	1/14	1/2	[72]
η	1/day	5	0	10	[72]
τ_β	day	60	0	120	[49]

predictions will emerge from the simulations. Choosing a suitable model is a primary exercise and of great importance in this type of analysis.

4 Results and discussion

This section presents several numerical experiments conducted with the SEIR(+AHD) epidemic model and the proposed CE-ABC algorithm. The plausible nominal values used in the integration of the dynamics of a virgin population for COVID-19 infections are presented in Table 1, which also shows numerical bounds (upper and lower) that are used to delineate the feasible domain limits in the CE method. The plausible values from Table 1 correspond to a COVID-19 outbreak in a virgin population to the disease, such as those observed worldwide in 2020. They were determined by information from the literature [3,5,19,39,44,49,63,72,77,80,84] or numerical experimentation.

Numerical experiments with these parameters are not focused on being very reliable reproductions of the COVID-19 outbreaks in 2020. They only aim to have the main characteristics of the epidemic dynamics so that they offer a good test for the methodology proposed in this paper.

The objective is to show that the CE-ABC framework is a powerful tool for data-driven epidemic modeling and can be used, together with a suitable epidemic model, in near real-time to predict the course of an epidemic outbreak of an emerging disease (such as COVID-19) in a time horizon compatible with the limits of predictability of the underlying dynamics.

4.1 Dynamic evolution of a fully susceptible population subjected to an initial infection

The first analysis presented here concerns the situation of a population virgin to COVID-19 infections, where the disease is introduced into the community by a single individual externally exposed to the viral agent that causes the disease.

This is a hypothetical case (possibly unrealistic), as it does not consider any measures to mitigate or suppress the outbreak during its occurrence. However, its study may be essential to delineate a possible baseline behavior related to a potential epidemic of COVID-19, providing projections of a worst-case scenario and some intuition about the free evolution of the disease.

In this scenario, a constant value for $\beta = \beta_0$ is considered, as well as an initial population $N_0 = 5.5 \times 10^6$ (compatible with the city of Rio de Janeiro, Brazil), a single exposed individual $E_0 = 1$, and all other initial conditions are set to zero, except the suscetibles, which is set as the difference between N_0 and all other variables. The time-step for simulation is equal to 1 day. The model parameters values are defined in the third column of Table 1.

The dynamic evolution of the SEIR(+AHD) model, for a temporal interval of 2 years, can be seen in Figure 6, which shows the corresponding time series in linear (top) and logarithmic (bottom) scales. In the bottom part, it is possible o better observe the trends of time series in which the maximum values are small compared to the initial size of the population.

Despite community transmission starting in the first moments of the dynamics, due to the transmission structure of this type of model, an outbreak only takes on notable proportions after 200 days of disease circulation in the population. In other words, it may take more than six months after the start of transmission of the disease within this population for the outbreak to be noted by the major public.

However, after the outbreak became noticeable, a wave of contagion by COVID-19 quickly emerged, characterized by a rapid increase in exposed compartments, concomitant with a decrease at the same rate in the susceptible population. Most recover directly, while a small portion dies without medical care. The other part of those infected are hospitalized, most of them recover, and a small amount dies.

The peak of infections occurs around 350 days after the insertion of the first exposed case in the population, almost a year after the disease arrives in the community. There were about 700 thousand people with active disease (exposed, infectious and asymptomatic) in the community during the peak, almost 13% of the initial

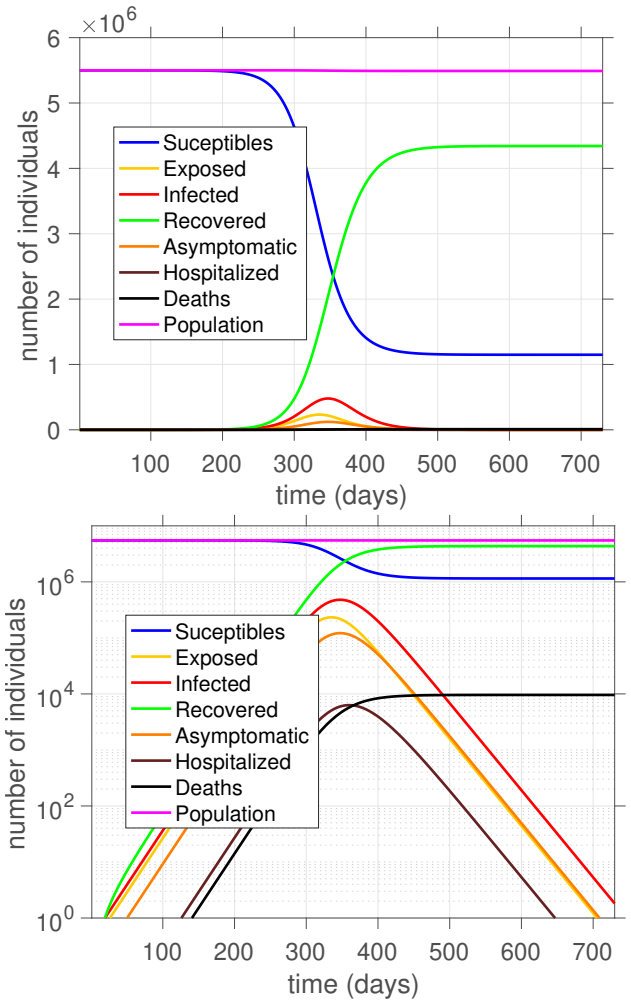

Fig. 6 Dynamic response of the SEIR(+AHD) epidemic model in a scenario of a totally susceptible population, with a single individual exposed. Time series in linear scale are shown at the top and in logarithmic scale at the bottom, which allows displaying better the curves in which maximum values are small compared to the initial population value.

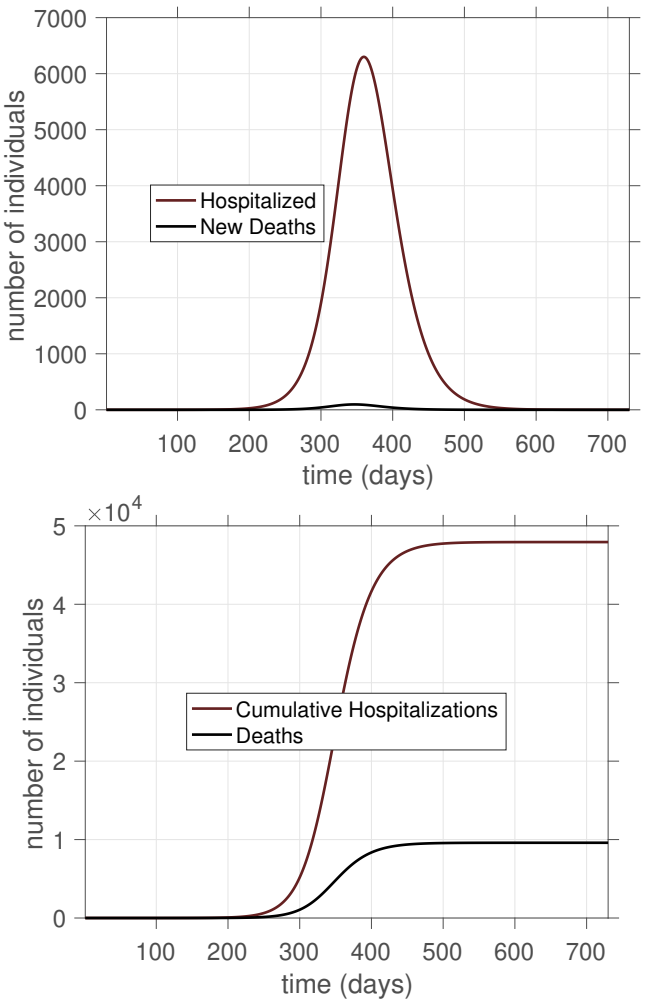

Fig. 7 Dynamic response of the two QoIs for the SEIR(+AHD) epidemic model in a scenario of a totally susceptible population, with a single individual exposed. Time series for the number of hospitalized (upper brown curve) and new deaths (lower black curve) are shown at the top, while the corresponding cumulative numbers can be seen at the bottom.

population. The susceptible corresponded to close to 20% of the people after two years, while the recovered account for almost 80% of the people.

The reader can better appreciate the evolution of the number of hospitalizations and new deaths per day in Figure 7 (top), as well as their respective cumulative values throughout the epidemic outbreak (bottom). The peaks of hospitalizations and deaths occur a few days after the peak of infections, involving more than 6000 people under medical care and around 100 deaths. At the end of the two-year window, almost 48,000 people were hospitalized at some point, and another 10,000 people died from complications inherent to the disease.

Throughout the outbreak, the variation in population size is slight compared to its initial size (10 thousand is a small number compared to 5.5 million, around 0.2%) but highly significant in demographic terms. This hypothetical outbreak is responsible for losing approximately 10 thousand lives in about 300 days. This value would correspond to something around 15% of all deaths that occurred in the city of Rio de Janeiro in 2019[2]. But note that, in this case, such an unusual amount of deaths is due to a single disease.

4.2 Determination of a dynamically consistent initial state for the epidemic model calibration

Typically, in the process of calibrating a dynamic model with the aid of data, the bottleneck is identifying the

[2] Demographic data for the city of Rio de Janeiro can are available at `https://transparencia.registrocivil.org.br`.

initial conditions since, often, the initial state of the system of interest is partially (or even totally) unknown. This is the case when dealing with compartmentalized epidemic systems in the form of an SEIR model or its variants. Observations on the infected compartment are usually available (subject to delay and underreporting), but data from recovered rarely (and even when they are, they are often unreliable). However, for the practical impossibility of measuring them, the susceptible and exposed practically are never known directly. Other possible compartments can also be challenging to measure in practice [17,32].

In this scenario, the determination of initial conditions (or part of its components) is usually done via direct inference from the data [14,15,30,41] or by indirect means, with plausible assumptions or educated estimates about the actual values [6,46]. But while the latter approach is highly subject to epistemic errors, the former may suffer from identifiability issues. Thus, novel methodologies to identify initial conditions of epidemiological systems are welcome.

By the existence and uniqueness theorem for ODEs [1,23,48,64,73], the dynamical system defined by (1) and a given initial condition has only one dynamic state for each instant of time. Once the value of one of the components is fixed (e.g. hospitalizations) for a particular moment, only one combination of values in the other compartments produces a dynamic state compatible with the fixed value. For this reason, it is practically impossible to infer a consistent initial condition from assumptions or ansatz to values (especially in an actual setting where surveillance data are imperfect representations of the dynamics of interest, and there are compartments for which data are not available).

To avoid the above difficulties, a three-step procedure to determine a suitable set of initial conditions that is compatible with the observed data is proposed:

1. Given a reference value for hospitalizations H_{ref}, the dynamics of a population virgin to the disease (such as presented in section 4.1) is used to determine the time instant for which $H(t)$ is closest to H_{ref}. The corresponding dynamic state is recorded;
2. Analogously, given a reference value for deaths D_{ref}, the dynamic state corresponding to the shortest distance between $D(t)$ and D_{ref} is obtained and recorded;
3. Finally, a dynamic state corresponding to a weighted average between the two states determined above is calculated and assumed to be a dynamically consistent initial condition.

When using the dynamics of a population totally susceptible to the disease to identify a dynamic state close to the reference values for H (or for D), the procedure guarantees that this state is "dynamically consistent", as it is a solution to the initial value problem associated with the epidemic model. Although, in general, such a state does not exactly satisfy the reference value, by the continuous dependence of the solutions on the initial conditions, one can guarantee that such a dynamic state is "sufficiently close" to the state associated with the exact value of the reference. By making a convex combination of initial conditions obtained this way, we still have a dynamic state close to all the reference values. In this way, the procedure described above can generate an initial condition that is "dynamically consistent" with the available data. The procedure is naturally generalized if there are reference values for other compartments.

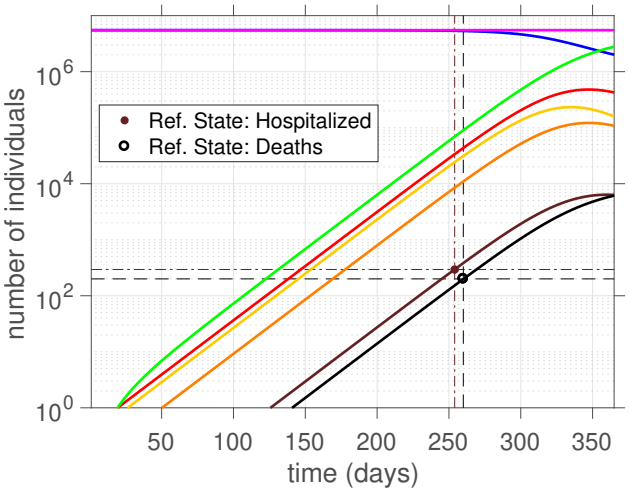

Fig. 8 Dynamic response of the SEIR(+AHD) epidemic model in a scenario of a totally susceptible population, with a single individual exposed. Two states (for different times) are highlighted, corresponding to prescribed values of H and D. The initial condition of subsequent simulations is defined by a convex combination of these two dynamic states. The colors of the curves follow the same scheme as in Figure 6.

To illustrate of the methodology, the reader can observe Figure 8, which shows two distinct dynamic states obtained from the data on hospitalizations and deaths together with the time series on a semi-logarithmic scale, corresponding to the response of a virgin population to the disease with a single exposed individual. The convex combination of these two dynamic states, considering the same weights used in Eq.(11), is used as an initial condition in the numerical experiments presented in the following sections.

4.3 Calibration and validation of the SEIR(+AHD) epidemic model and its descriptive capacity

In this section the proposed CE-ABC framework is used to calibrate and quantify the parametric uncertainties inherent to the SEIR(+AHD) dynamic model. To this end, the following hyperparameters are adopted in the CE-ABC algorithm: discrepancy function weight $\omega = 0.75$ for hospitalizations (and thus, $1 - \omega = 0.25$ for deaths); CE samples $N_{ce} = 100$; CE elite sample set size $N_{\mathcal{E}_\ell} = 10\%$ of N_{ce}; CE absolute tolerance `atol` $= 0.001$; CE relative tolerance `rtol` $= 0.05$; CE mean value smoothing parameter $a = 0.7$; CE variance dynamic smoothing parameters $b = 0.8$ and $q = 5$; CE maximum number of iterations `maxiter` $= 150$; ABC samples $N_{abc} = 2000$; ABC tolerance `tol` $= 0.1$. The bounds for the model parameters are adopted according to the values shown in Table 1.

The data considered here for the training of the dynamic model include the records of hospitalizations and total deaths in the city of Rio de Janeiro between May 1 and 31, 2020. The statistical validation process of the calibrated model uses the data corresponding to the following month, between June 1 and 30, 2020. Data for April 2020 are ignored because they are unreliable since the city's epidemiological surveillance system was still adapting to the new reality at the beginning of the pandemic.

The results regarding the calibration, quantification of uncertainties, and validation of the SEIR(+AHD) model with the aid of the CE-ABC algorithm can be seen in Table 2 and Figures 9 and 10.

Table 2 shows the values calculated by the CE-ABC algorithm for the parameters of the SEIR(+AHD) model, showing the estimates obtained by the CE optimizer in the third column; the respective standard deviation values in the fourth column; the best sample of the ABC simulation in the fifth column; and the standard deviation values of the posterior distributions obtained by ABC in the sixth column. The two sets of parameters identified present very close values, and the ABC result is a kind of refinement of the estimate obtained by the CE.

Regarding the posterior joint distribution of the model parameters, the reader finds this information in Figure 9, which presents the histograms and scatter plots for each of the model parameters (estimated with the samples accepted by the ABC simulation). Scatter plots give information about the correlation between the parameters. In this figure, the order of the parameters is the same as shown in Table 2.

In Figure 10 the reader can see the time series of hospitalizations (left) and total deaths (right) for the city of Rio de Janeiro in a time window that covers the months of May and June 2020. Time trajectories, accepted by the ABC simulations (87% of the 2000 total)[3], are displayed as thin solid lines in light gray; the trajectory that corresponds to the optimal set of parameters obtained by the CE optimizer is displayed as a dash-dotted line; the best sample trajectory obtained by the ABC simulation (the one with the smallest error) is indicated as a dashed line; while the median calculated with the samples accepted in the ABC simulation is indicated as a thick solid line. In addition, a 95% credibility envelope is displayed in the form of a filled region above the ABC samples. Training data are

Table 2 Parameters identified for SEIR(+AHD) epidemic model via CE-ABC framework, and the respective standard deviation values.

	Unit	CE Optimal	CE std dev	ABC Best	ABC std dev
β_0	1/day	0.12	0.02	0.13	0.02
α	1/day	0.20	0.07	0.27	0.06
f_E	—	0.81	0.03	0.84	0.03
γ	1/day	0.13	0.01	0.12	0.01
ρ	1/day	0.0006	0.0001	0.0005	0.0001
δ	1/day	0.0021	0.0004	0.0015	0.0004
κ_A	—	0.0026	0.0008	0.0027	0.0008
κ_H	—	0.0563	0.0130	0.0575	0.0128
ϵ_H	—	0.25	0.07	0.33	0.07
β_∞	1/day	0.31	0.06	0.43	0.06
η	1/day	5.8	1.9	6.2	1.8
τ_β	day	146	7	153	7

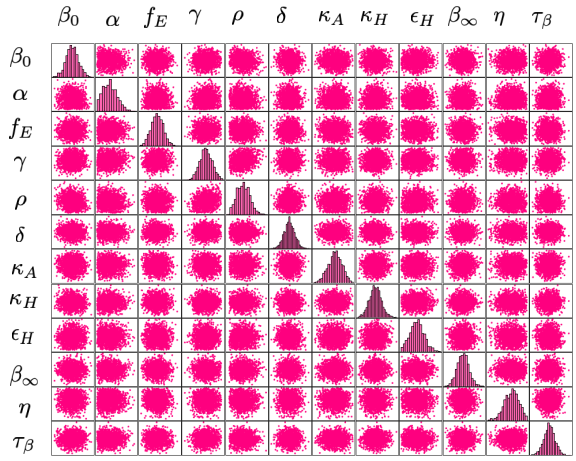

Fig. 9 Histograms and scatter plots of the SEIR(+AHD) epidemic model, estimated with the samples accepted by the ABC simulations. The order of the parameters is the same as shown in Table 2.

[3] This high acceptance rate, which may seem very high at first glance, is due to the informative prior obtained by CE.

displayed as magenta circles, while validation data are shown as cyan asterisks.

The comparison between the CE-ABC time series and the training data shows that, in this scenario, the dynamic model can reproduce well the epidemic outbreak experienced by the city of Rio de Janeiro in May 2020. For both time series, the curve corresponding to the optimal set of parameters identified by the CE optimizer, the best scenario simulated by ABC, and the median calculated with the cases accepted in the ABC simulation provide good descriptions of the epidemic data trend. Furthermore, the training data fit is robust once it includes the 95% credibility interval obtained by ABC accepted samples, covering most of data fluctuations.

In terms of validation, by comparing the predictions (extrapolations) made by the dynamic model and future data (not used in the calibration), one can note that the dynamic model captures the trend of the outbreak. It takes over the data due to a 95% credibility band around the calculated evolution curves. Strictly speaking, the forecasts are reasonably accurate for the first seven days of extrapolation, starting to shift from the simulated curves from this point onwards. In what follows, hospitalizations are slightly underestimated by the ABC median by approximately 25%, while the median overestimates total deaths up to a limit of around 10%.

In light of the minimal horizon of predictability that epidemic systems present[4], these predictions can be considered very good, as they provide accurate values in the short term (one week) and bring some reasonable information in the medium term (one month). Although 10-25% uncertainty in forecasting the number of hospital beds/expected deaths is not highly accurate for an immediate sizing of hospitals or funeral units, it is still informative in indicating to decision makers the correct order of magnitude for these outcomes. For instance, knowing a month in advance, in the course of an epidemic, that a few hundred (not thousands, or vice-versa) of hospital beds/burials will be required per day can be crucial information to prevent a hospital or cemetery from collapsing. Added to this is that the model can be recalibrated weekly (or daily), updating the short/medium forecasts whenever new data become available.

[4] In an epidemic where people are aware of what is happening, there is a feedback between the rate of infection and people's social behavior. Being aware of the severity of the outbreak beforehand can help reduce its intensity or vice versa. The great difficulty in modeling such feedback is one of the factors (perhaps the main one) that limits the predictability horizon of epidemic models.

It is also worth mentioning that, in addition to making predictions about the QoIs for which epidemic data are available, a well-calibrated mechanistic model can provide information on latent quantities (for which data are not available), such as the number of susceptible, exposed, asymptomatic, etc. In this sense, to illustrate this possibility, the reader is invited to observe Figure 11, which presents the evolution of the time series associated with eight dynamics state coordinates of the epidemic model in a time window that includes the months of May and June 2020, considering the best estimate of the ABC simulation for the model parameters.

From a qualitative point of view, this simulation allows the analyst to infer that, in this two-month interval, there is a slight but notable decrease in the number of susceptible people in the general population due to the increase in COVID-19 infections, followed by an increase in total recoveries. The number of symptomatic infected is always more significant than the number of exposed, greater than the asymptomatic infected. Quantitatively, it can be seen that these last three groups have sizes of the same order of magnitude, which is hundreds of times greater than the number of hospitalized patients. At the end of this two-month interval, the total number of accumulated deaths reaches a value comparable to the active infected.

Of course, the accuracy of such information largely depends on the extent to which the structure of the epidemic model provides a reliable representation of epidemic dynamics. If it is a good representation, the simulations should offer great insight; if it is a feeble representation, the simulations do not tell anything useful at the limit. In intermediate cases, where the model is more or less accurate, useful information can be obtained, but not all the information from the simulations is reliable. Separating what is helpful from what is not requires deep knowledge of some basic principles of epidemiology.

4.4 Influence of CE-ABC hyperparameters on the description of the epidemic dynamical system

The CE-ABC framework combines two advanced stochastic simulation techniques, thus inheriting all the control parameters (a.k.a. hyperparameters) underlying the two methods. Consequently, the model calibration process and the propagation of parametric uncertainties depend, nonlinearly, on these hyperparameters. Thus, a study of how such quantities affect the modeling is desirable and recommendable. The present section of the manuscript seeks to shed light on this.

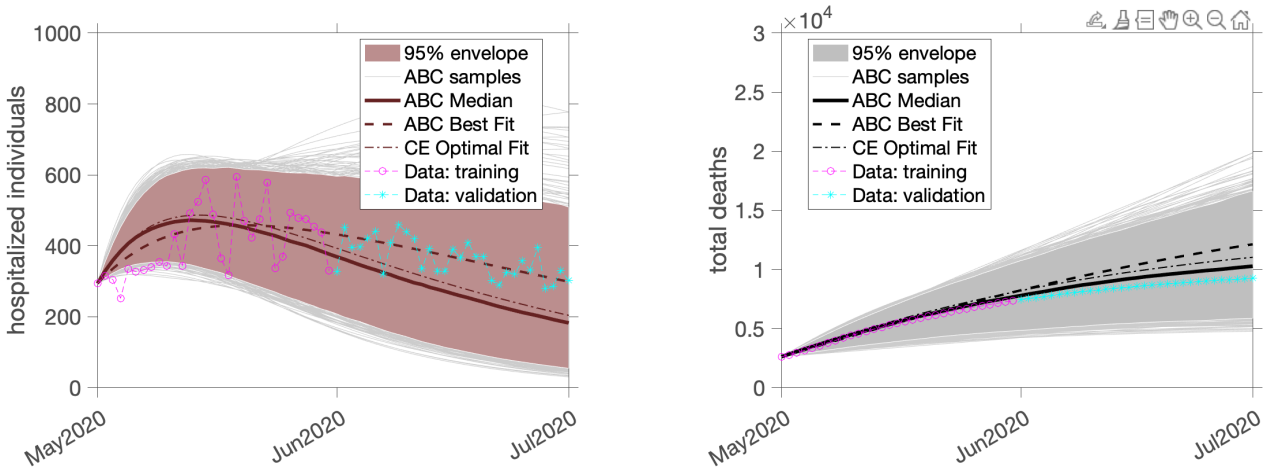

Fig. 10 Time series generated by the CE-ABC algorithm for the number of hospitalized individuals (left) and total deaths (right) obtained with the SEIR(+AHD) model, which is calibrated with Rio de Janeiro epidemic data from May 2020 and validated for a temporal window covering the month of June 2020. Here the discrepancy function weight is $\omega = 0.75$, and ABC acceptance rate is 87%.

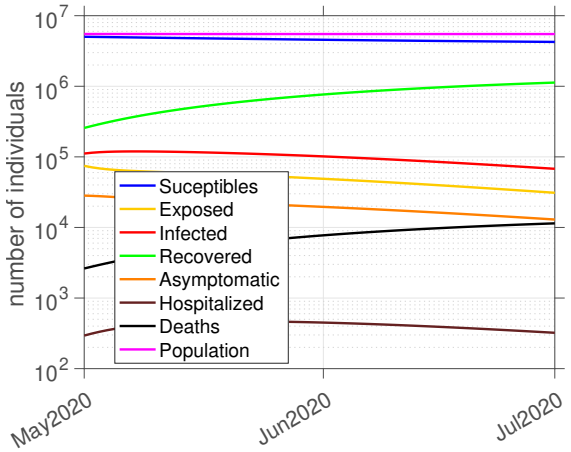

Fig. 11 Dynamic response of the SEIR(+AHD) epidemic model, in a time window that includes the months of May and June 2020, considering the best estimate of the ABC simulation for the model parameters. Here the discrepancy function weight is $\omega = 0.75$.

Initially, we investigated the effect of the discrepancy function weight parameter ω on the results. For that, Figure 12 presents the two QoIs calculated by the SEIR(+AHD) epidemic model, considering five different values for the weight: $\omega \in \{0, 0.25, 0.5, 0.75, 1\}$.

By visual inspection, it is possible to see that the best fits are obtained when $\omega = 0.25$, $\omega = 0.5$ or $\omega = 0.75$. The first and the last case favor one QoI, but without totally disregarding the effect of the other, while the intermediate case balances both. In this setting, deciding which QoI should be given more weight is a matter of convenience. If the information on hospitalizations is more important, higher ω values should be considered. The opposite is true if the primary interest is to follow the evolution of deaths.

The $\omega = 0$ scenario considers a limiting case, where the discrepancy function defined by Eq.(11) considers only the total number of deaths. In contrast, considering only hospitalizations, the other extreme situation is counted when $\omega = 1$. These two limit cases have a terrible fit in the disregarded QoI, and should only be considered in situations in which only one of the QoI is of reliable.

Once the value of ω is fixed, the results are also affected by the choices of $\boldsymbol{x}_{min}$, $\boldsymbol{x}_{max}$, N_{ce}, N_{ε_ℓ}, N_{abc}, `atol`, `rtol`, `tol`, and `maxiter`.

The effect of varying the number ABC simulation samples N_{abc} can be seen in Figure 13, where results are presented for $N_{abc} = 100$ (top) and $N_{abc} = 1000$ (bottom). Note that the number of accepted samples is directly proportional to the total number of simulated samples. This variation directly impacts the estimation of the posterior distribution, with a consequence in the obtained median and credibility band. Of course, these estimates are also affected by the tolerance `tol`, which improves or worsens the results as it is decreased or increased. As it is an obvious effect, numerical experiments in this sense are not shown. A tolerance of the order of 10%, i.e., `tol` $= 0.1$, proved to be effective for the simulations in this paper.

The parameters `atol`, `rtol`, and `maxiter` have an influence on how much faster the CE optimizer will stop, with consequent gain/loss of accuracy, followed by an increase/decrease in the computational cost. The reader is invited to do numerical experiments with these parameters to see their effect in practice. In the prelim-

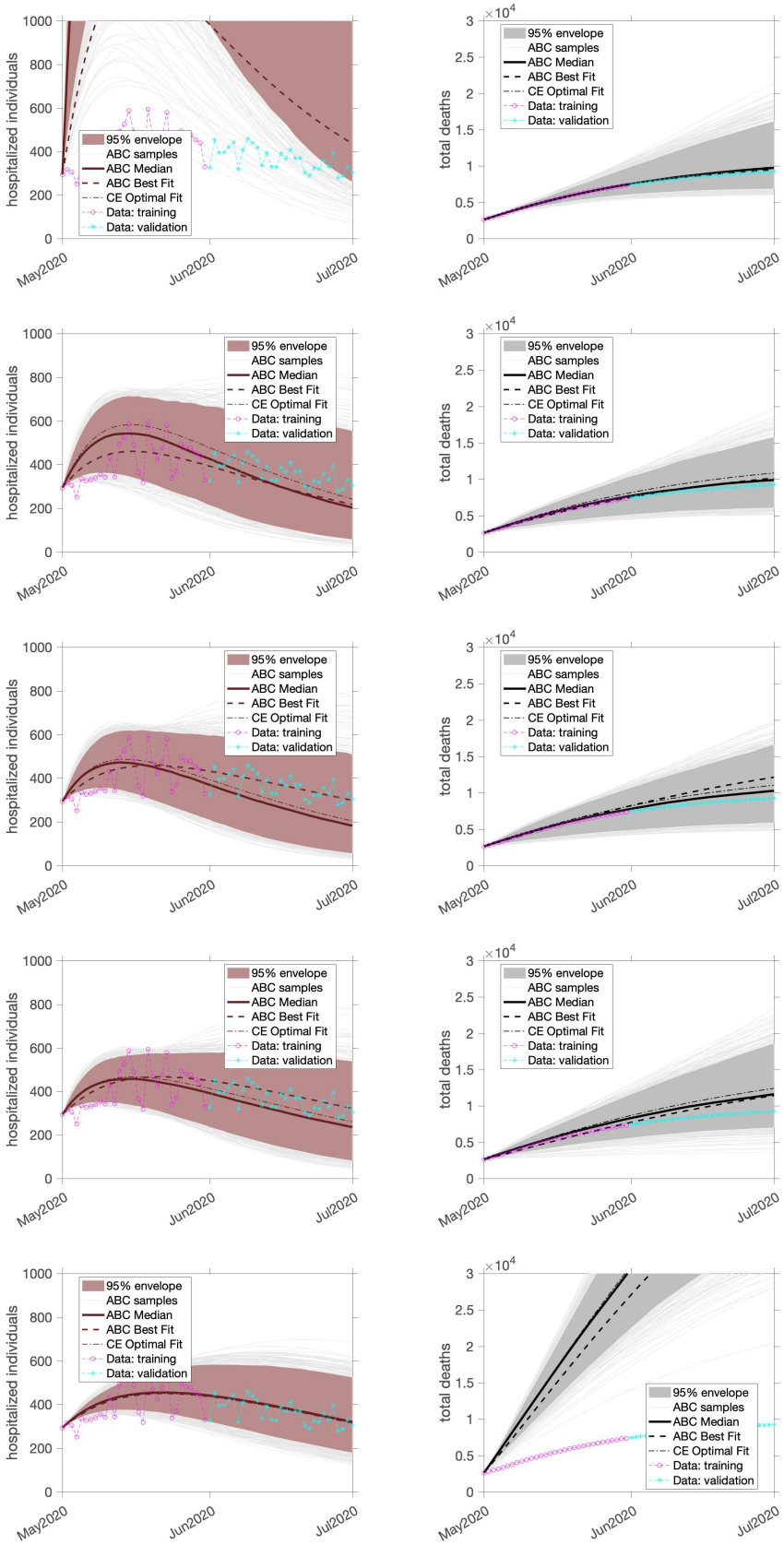

Fig. 12 Quantities of interest calculated by the SEIR(+AHD) epidemic model with different values of the weight parameter: $\omega = 0$ (first line); $\omega = 0.25$ (second line); $\omega = 0.5$ (third line); $\omega = 0.75$ (fourth line); $\omega = 1$ (fifth line).

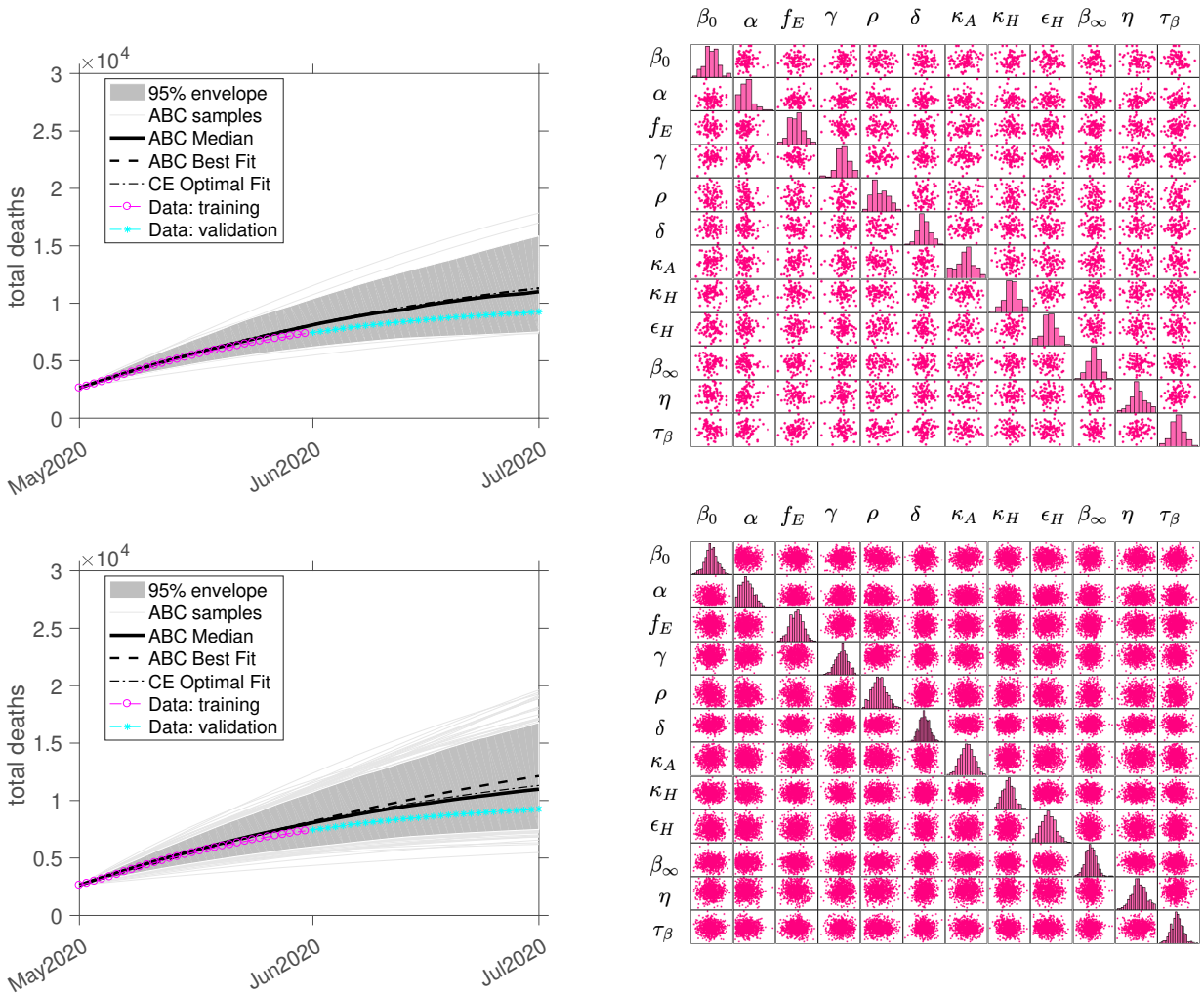

Fig. 13 Effect of varying the number ABC simulation samples on the QoIs calculations and posterior inference. At the top $N_{abc} = 100$, while at the bottom $N_{abc} = 1000$. Here the discrepancy function weight is $\omega = 0.75$ and $N_{ce} = 100$.

inary numerical studies conducted by the authors, we observed that there are no great gains in obtaining an estimate of the optimal parameters with great precision. Relative tolerance values of the order of 5%, i.e., `rtol` $= 0.05$, provide a good compromise between accuracy and computational cost.

However, it is interesting to observe the effect of N_{ce} variation in practice, as shown in Figure 14, which considers different values for CE samples, $N_{ce} = 50$ (top) and $N_{ce} = 200$ (bottom). The number of CE samples influences the selection of the optimal set of parameters and the inference process performed by ABC, since the a priori distribution used by ABC is constructed with the help of statistics calculated by the CE. It is interesting to note that, unlike ABC, where a greater number of samples typically leads to a better inference result, this is not necessarily the case with CE optimization. In the example shown in Figure 14, the model calibrated with only 50 samples has better adherence to the data than the counterpart that uses $N_{ce} = 200$. A variation in the size of the elite set $N_{\mathcal{E}_\ell}$ can positively or negatively impact accuracy, depending on the case. Prior numerical experiments help identifying which case of the problem of interest.

As it is a stochastic algorithm, obviously the results depend on the value of the statistical seed used. Figure 15 shows two simulations with the same hyperparameters, but with slightly different inference results.

Finally, but not least, it is worth mentioning that the results are strongly influenced by choice of bounds $\boldsymbol{x}_{min}$ and $\boldsymbol{x}_{max}$. Indeed, in the authors' experience, these are the parameters that have the most significant impact on the quality of results (together with the epidemic surveillance data). A bad choice for the parameter bounds can lead to unreliable models for the actual behavior of the outbreak. Good choices for these parameters demand detailed knowledge about the biological

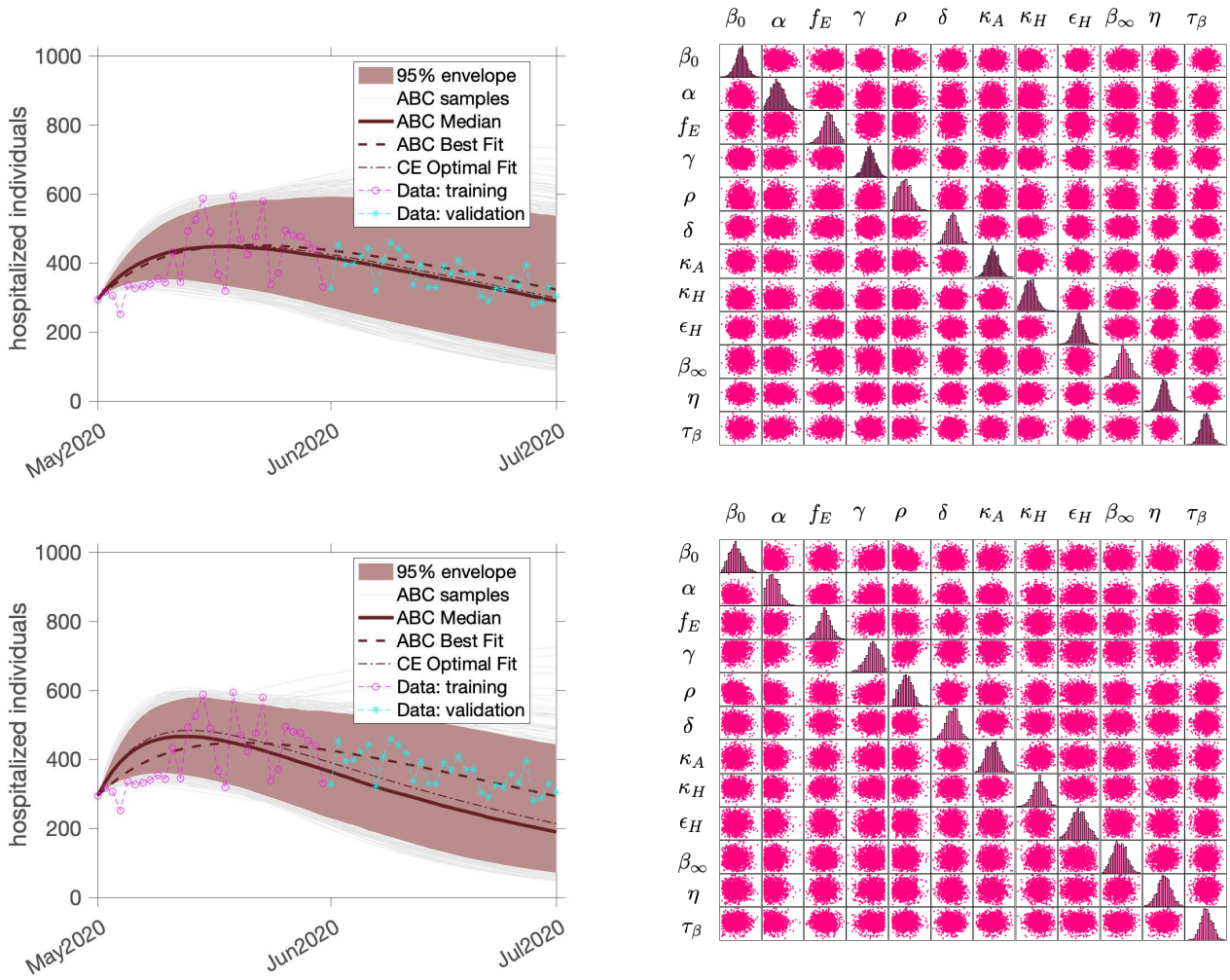

Fig. 14 Effect of varying the number CE samples on the QoIs calculations and posterior inference. At the top $N_{ce} = 50$, while at the bottom $N_{ce} = 200$. Here the discrepancy function weight is $\omega = 0.75$ and $N_{abc} = 2000$.

aspects of the problem. Numerical experimentation can also be of great help in finding plausible values.

4.5 Predictability limit for the epidemic dynamics using the CE-ABC and the SEIR(+AHD) model

This section presents a study to delineate the predictability limit of the SEIR(+AHD) model as a tool to predict the dynamics of COVID-19 in the city of Rio de Janeiro in the year 2020.

To this end, Figures 16 and 17 show the evolution of the two QoIs, for various training data sets, extrapolating forecasts over a 30-day horizon. Training data are incremented every seven days, including information from the last seven days, starting with the period between May 1 and 7, 2020, and ending with May 1 and July 9, 2020.

For the first three calibrations (calibrations between the first and third week), the model presents a modest descriptive capacity of the data, with the trend of the short and mid-term forecasts being quite discrepant to that observed in the following weeks. However, the respective credibility intervals encompass the observations.

As the weeks go by, with more (and better quality) data feeding into the model (calibrations between the fourth and seventh week), the description of the training data improves substantially, as does the predictive ability. In such cases, within a week, the model's median predicts the numbers of hospitalizations and deaths with reasonable accuracy for epidemic estimates. However, it starts to lose accuracy from the second week of extrapolation gradually, although it continues to follow, more or less, the trend of the data for 30 days (and covers them via the credibility band).

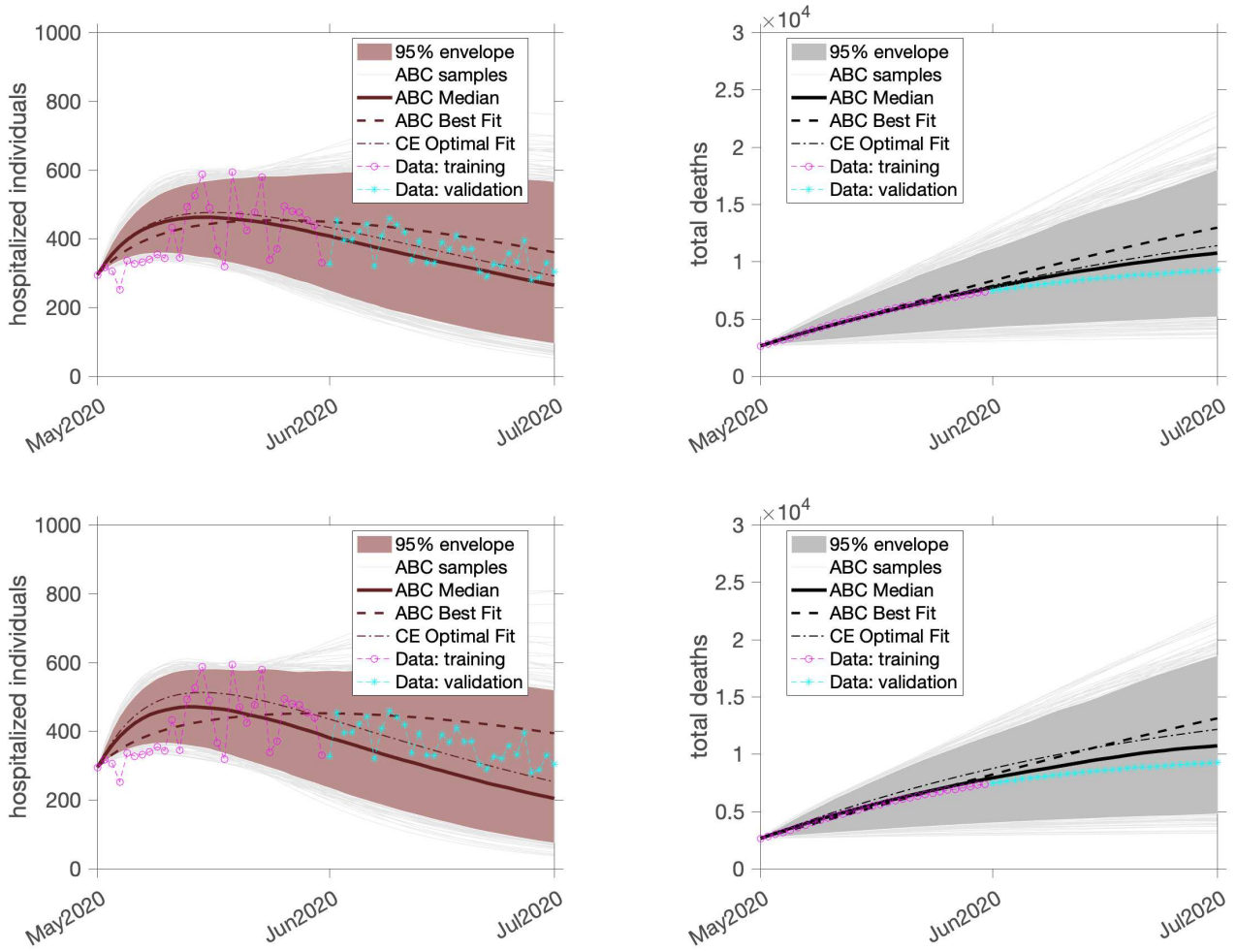

Fig. 15 Effect of the statistical seed on the QoIs calculations. Here the discrepancy function weight is $\omega = 0.75$, $N_{abc} = 2000$ and $N_{ce} = 200$.

The descriptive capacity remains impeccable for the last three weeks of model calibration (weeks eight to ten), with good short-term (one week) prediction. However, it is possible to notice a significant divergence between predictions and observations after seven days of extrapolation. Such a loss of predictability is not directly related to the quantity or quality of the calibration data but rather to a structural change in the dynamic behavior of the outbreak. In July 2020, the second wave of contagion began in the city of Rio de Janeiro [18], drastically changing the trend of evolution of QoIs. As the infection rate $\beta(t)$ was modeled to only contemplate a single change in the infection plateau, the present model cannot accurately describe the new wave of infections. One possibility to make the model regain its predictive capacity would be the inclusion of a new infection plateau in Eq.(2), as done in [72]. In general, this strategy can be adopted to address not just one but several waves of infection. Due to space limitations, the authors did not include results in this sense in the manuscript, but it would be interesting to test this strategy in future work.

The above results show that the model's descriptive capacity and predictability limit are strongly influenced by the amount and quality of data used in the calibration process.

In general, the more data, the greater the predictability horizon of the model. It presents an excellent or good capacity to extrapolate within a horizon of one or two weeks, with some capacity to predict the trend up to thirty days (depending on the outbreak's dynamic behavior). The quality of the outbreak's data also matters, as it is clear from the first three numerical experiments. In these cases, hospitalization data do not show the typical fluctuation of this time series (probably due to underreporting, since at the beginning of May 2020, the surveillance system in Rio de Janeiro was still adapting to the pandemic), which affects the model's fitting.

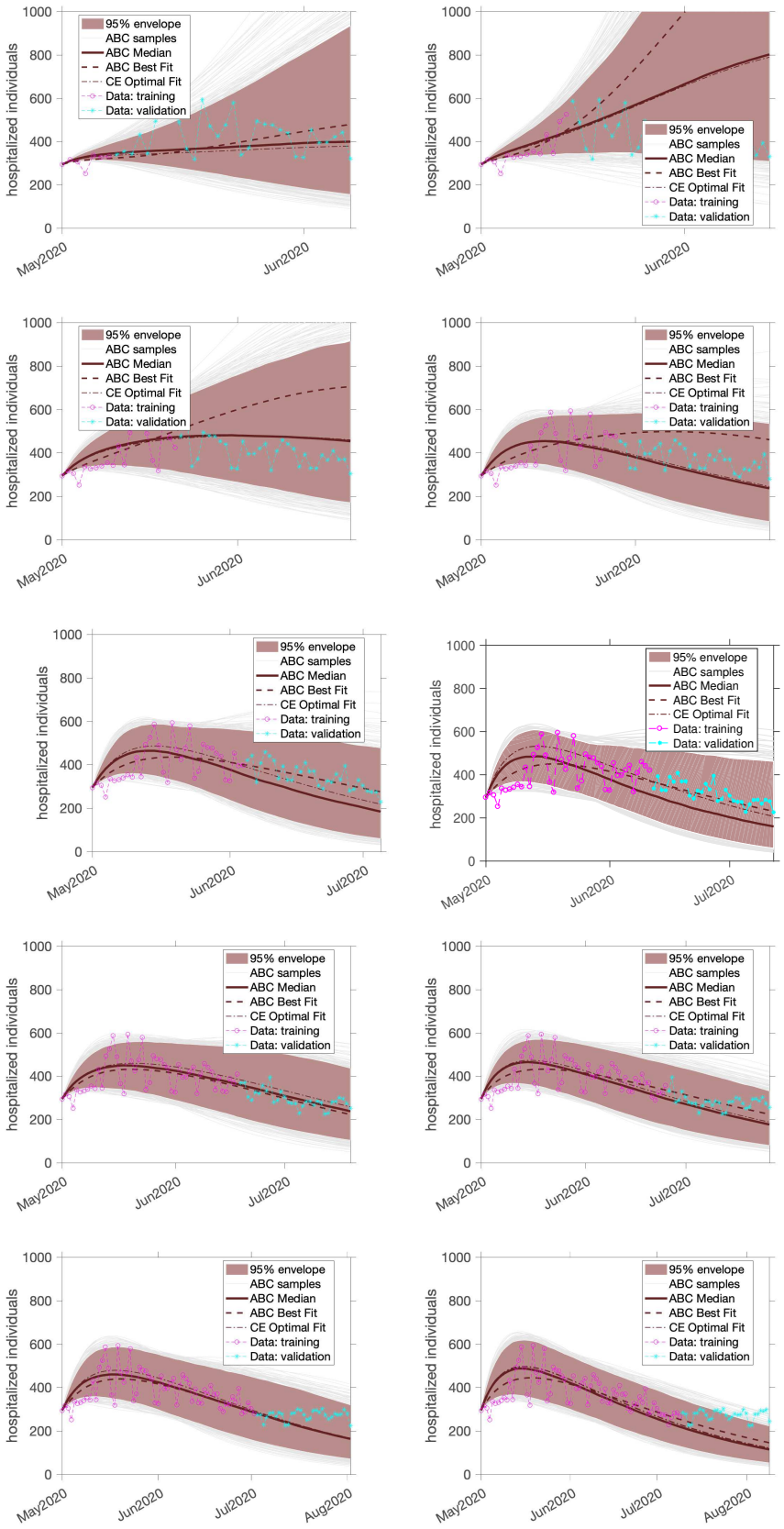

Fig. 16 Evolution of the hospitalizations time series, calibrated with different datasets of Rio de Janeiro epidemic, extrapolating forecasts over a 30-day horizon. Training data are incremented every seven days, including information from the last seven days, starting with the period between May 1 and 7, 2020, and ending with May 1 and July 9, 2020.

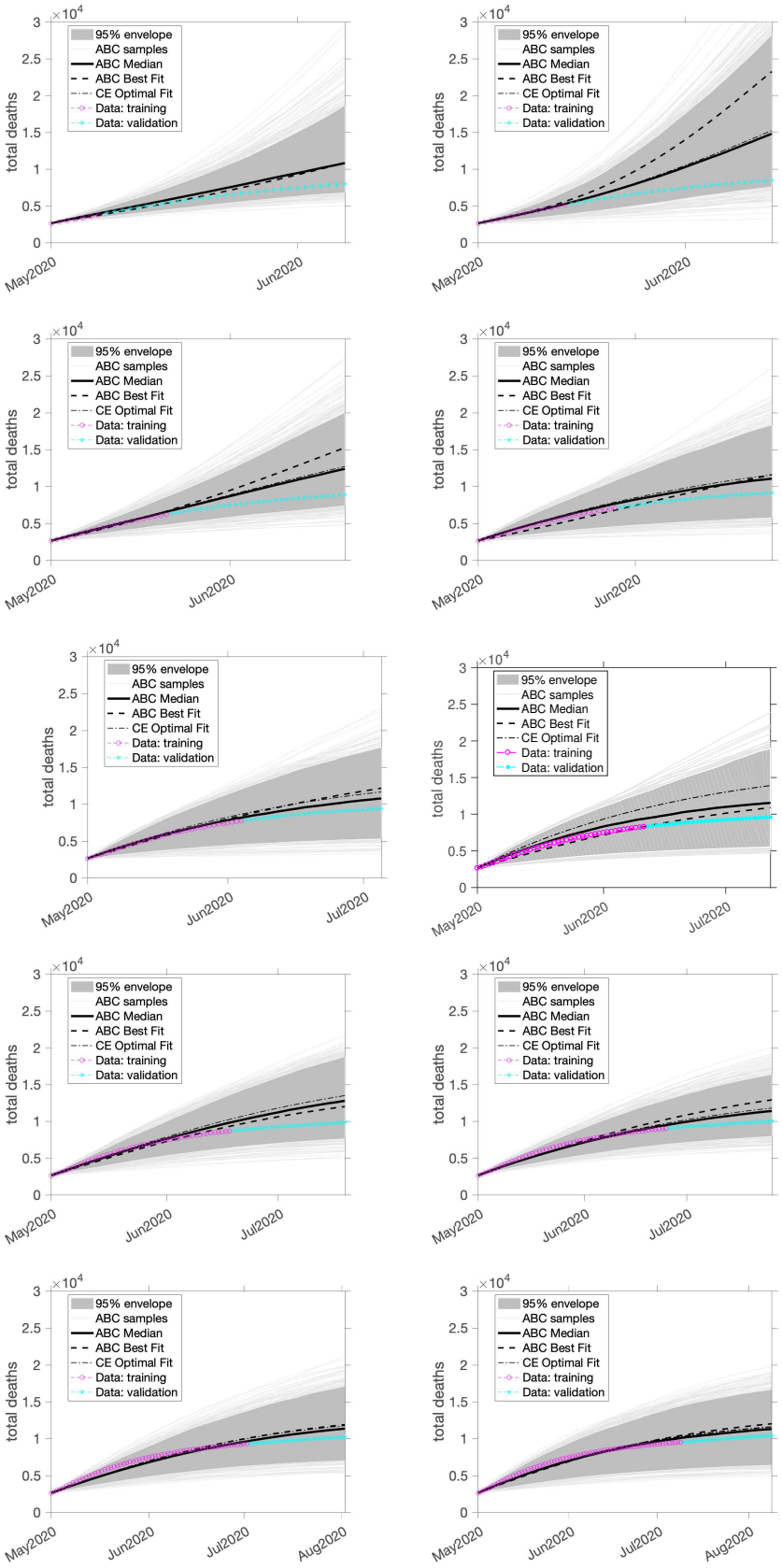

Fig. 17 Evolution of the total deaths time series, calibrated with different datasets of Rio de Janeiro epidemic, extrapolating forecasts over a 30-day horizon. Training data are incremented every seven days, including information from the last seven days, starting with the period between May 1 and 7, 2020, and ending with May 1 and July 9, 2020.

Insufficient or poor quality data can compromise the model's fit, generating unrealistic predictions. However, structural changes in the dynamic behavior of the epidemic (e.g., the emergence of a new wave of contagion) have an even more pronounced effect in compromising the predictive capacity of the model. In the periods preceding such changes, even quality data cannot guarantee that the model will extrapolate the data well in the mid (or even in the short) term.

4.6 Verification of efficiency for the CE-ABC framework for data-driven epidemic modeling

To conclude the discussion of the results, this section presents a computational experiment to address the computational-statistical efficiency of the CE-ABC framework in comparison with a standard ABC approach (without using CE method to refine the prior).

For this purpose, four test cases are compared in Figure 18, which shows histograms and scatter plots for the epidemic model parameters obtained with different strategies of statistical learning:

1. ABC with a uniform prior;
2. ABC with lognormal prior;
3. ABC with truncated Gaussian prior;
4. CE-ABC with truncated Gaussian prior.

In all these tests, the values adopted for the limits of the distributions (when of limited support) are shown in Table 1, where the mean values also come from. The respective standard deviations are defined as those corresponding to a uniform distribution on the finite support of this table. The other parameters are similar to the case discussed in section 4.3, except the number of samples used in Bayesian inference, which is $N_{abc} = 100k$ for the three case that use standard ABC, and just $N_{abc} = 2k$ for the novel CE-ABC framework.

For the three cases where only ABC is used, the acceptance rate is at most a modest 1%, while a substantial value of 87% is obtained in the case that uses CE-ABC. These results show the efficiency of the new CE-ABC approach proposed here concerning the standard ABC method. It is also clear that the CE-ABC framework provided a gain in information about the distribution of parameter values much higher than the traditional method, with average values reasonably far from the boundaries. Both characteristics result from using the CE method to refine the prior, which becomes much more informative than the distribution used to sample the domain at the beginning of the optimization process that identifies the model parameters' nominal values.

5 Concluding remarks

5.1 Contributions

This paper proposes a new framework for model calibration and uncertainty quantification that combines the cross-entropy (CE) method for optimization with approximate Bayesian computation (ABC) for statistical learning. In this approach CE is used to obtain an initial and informative estimation of the model parameters. Then, central tendency and dispersion information obtained from CE are used to construct a informative prior distribution for an inference process that uses ABC to refine the model calibration and propagate the underlying uncertainties via acceptance-rejection Monte Carlo sampling. This framework also employs a heuristic strategy for identification of the initial conditions by using plausible dynamic states that are compatible with observational data.

This combination of well-established algorithms gives rise to a framework for uncertainty quantification with several good features. CE and ABC are intuitive and straightforward algorithms. Their combination gives rise to a simplistic and robust framework, with few control parameters of clear interpretation, where no gradient computation is required. In the update step with ABC, the initial knowledge about the model parameters obtained by CE optimization is incorporated into the prior distribution and updated with the available data to produce an informative posterior distribution. Also, there is no need to assume an additive Gaussian error. The uncertainty propagation is performed when the parameters are identified, generating considerable computational savings. The methodology's major limitation is its sampling nature, so many simulations might be needed to achieve convergence. This characteristic is not a problem for applications using epidemic models based on differential equations, where each deterministic simulation has a low computational cost. But for other domains (e.g. computational mechanics), where models may take hours/days to run a single instance, the CE-ABC framework may not be competitive.

The proposed methodology was tested on an epidemic model with an SEIR-type structure that also considers asymptomatic individuals, hospitalizations, deaths, and time-dependent transmission rate. Actual data from COVID-19 outbreaks in Rio de Janeiro city were employed in the model calibration process. The results were consistent, and the methodology seems promising. They show that it is possible to perform good calibrations of the epidemic model with the CE-ABC formalism in scenarios that require a descriptive model (to explain past outbreaks) and those where the objective

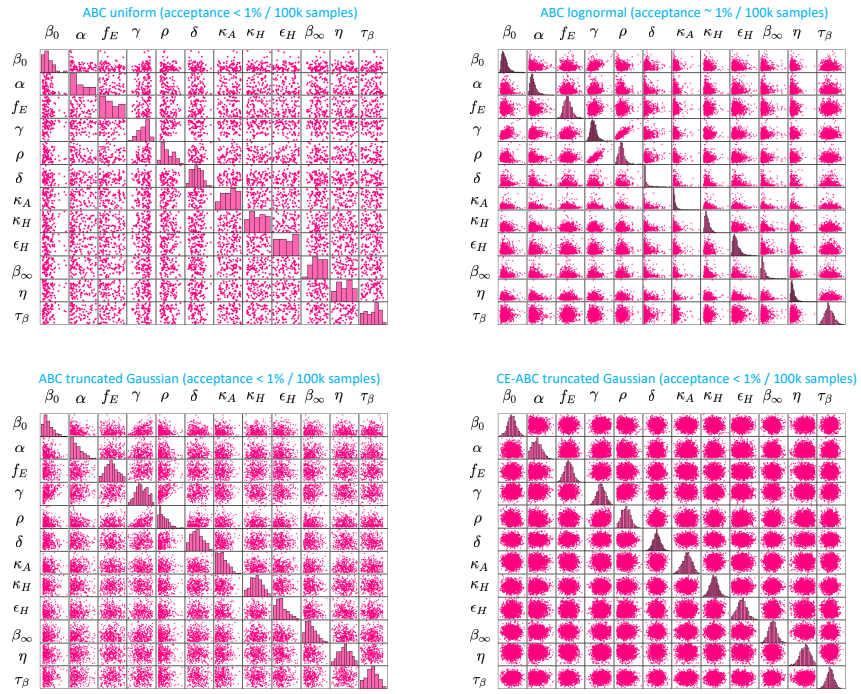

Fig. 18 Histograms and scatter plots for the model parameters obtained with different statistical learning strategies. ABC with a uniform prior and 100k samples (top left); ABC with lognormal prior and 100k samples (top right); ABC with truncated Gaussian prior and 100k samples (bottom left); CE-ABC with truncated Gaussian prior and 2k samples (bottom right).

is to obtain a predictive model (to infer future behavior of epidemics). In scenarios where the epidemic model structure is a good abstraction of the contagion dynamics, a horizon of good quantitative predictability of up to two weeks can be achieved using CE-ABC for model calibration and uncertainty quantification, with a good capacity for qualitative description of the data trend for up to one month.

5.2 Future directions

There are some possibilities to continue this research. One can apply the proposed methodology to other dynamical systems, including other COVID-19 models with different data, or the possibility of contemplating reinfection. Another branch that can be explored is related to model (epistemic) uncertainties. For instance, it can be very appealing to combine our CE-ABC framework with methodologies that compensate for deficiencies in the structure of the mathematical model. For instance, the random matrix-based nonparametric probabilistic approach by C. Soize [61]; the universal differential equations (UDE) for scientific machine learning by Rackauckas et al. [50]; or one of the physics-informed neural networks approached for epidemic modeling available on the literature [28,52,57,81]. It would also be exciting and natural to insert the CE-ABC algorithm proposed here as a calibration/UQ module in the integrated framework for data-driven epidemic models developed by Zhang et al. [83].

5.3 Disclaimer

A model is always wrong, but some of them are useful. This idea has a more pronounced meaning in computational epidemiology than in physics, as the first principles of epidemic dynamics are unknown. Although a mechanistic model such as the SEIR(+AHD) used here is a (typically very rough) approximation of epidemic dynamics, it allows exploring qualitative scenarios (short, medium, and long term) that can provide great insight into the evolution of the outbreak. Thus, being an extremely valuable tool for epidemiologists [24]. Undoubtedly, such an approach is much more rational and conservative than being guided by the (well-intentioned or not) opinion of curious people with no training in the area or the general public (layman by definition).

However, a final observation is necessary before one uses an epidemic model to guide decision-making during a real-time outbreak. It concerns the interpretation of results. As epidemiology is a highly interdisciplinary

area, it is practically impossible for a single professional to hold all the necessary skills to assess the results of an epidemic simulation unequivocally and, above all, understand the consequences of intervention measures that can be taken. Scientific, ethical, and humanistic aspects are equally important in this context and must be discussed by an interdisciplinary panel of professionals. In this sense, the authors of this paper strongly recommend that simulations of this nature, especially made with our model and framework, be evaluated and used with great caution when making decisions, preferably being scrutinized by a team of experts.

Dedication

The authors dedicate this work to the memory of all the victims of the COVID-19 worldwide tragedy. In particular, the first author dedicates the paper to his friends Natasha Zadorosny and Victor Costa Silva, who left in their prime.

Acknowledgements

The authors thank Prof. Luiz Max Fagundes de Carvalho (EMAp/FGV) and the anonymous reviewers for their critical reading of this text and for the excellent comments that helped improve the manuscript's final version.

Funding

The first author received financial support from the Carlos Chagas Filho Research Foundation of Rio de Janeiro State (FAPERJ) under the following grants: 210.167/2019, 211.037/2019, and 201.294/2021. The second author would like to acknowledge the Engineering and Physical Sciences Research Council (EPSRC) via grant number EP/R006768/1. The last author would like to acknowledge the financial support from the Brazilian agencies *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES) under the grants Finance Code 001, PROEX 803/2018, and CAPES-PRINT 88887.569759/2020-00, and FAPERJ under the grant 201.183/2022.

Code availability

To facilitate the reproduction of this paper's results the code used in the simulations is available on GitHub: https://github.com/americocunhajr/CE-ABC

Declarations

Conflict of Interest

The authors declare they have no conflict of interest.

References

1. Arnold, V.I.: Ordinary Differential Equations, 2nd edn. Springer (1992)
2. Brauer, F.: Mathematical epidemiology: Past, present, and future. Infectious Disease Modelling **2**(2), 113–127 (2017). DOI 10.1016/j.idm.2017.02.001
3. Byambasuren, O., Cardona, M., Bell, K., Clark, J., McLaws, M.L., Glasziou, P.: Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: Systematic review and meta-analysis. Official Journal of the Association of Medical Microbiology and Infectious Disease Canada **5**(4), 223–234 (2020). DOI 10.3138/jammi-2020-0030
4. Cai, M., Karniadakis, G.E., Li, C.: Fractional SEIR model and data-driven predictions of COVID-19 dynamics of Omicron variant. Chaos **32**, 071101 (2022). DOI 10.1063/5.0099450
5. Cheng, C., 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 **10**(1), 119 (2021). DOI 10.1186/s40249-021-00901-9
6. Costa, G.S., Cota, W., Ferreira, S.C.: Outbreak diversity in epidemic waves propagating through distinct geographical scales. Physical Review Research **2**, 043306 (2020). DOI 10.1103/PhysRevResearch.2.043306
7. Cotta, R.M., Naveira-Cotta, C.P., Magal, P.: Mathematical parameters of the COVID-19 epidemic in Brazil and evaluation of the impact of different public health measures. Biology **9**(8) (2020). DOI 10.3390/biology9080220
8. Cunha Jr, A.: Modeling and Quantification of Physical Systems Uncertainties in a Probabilistic Framework. In: S. Ekwaro-Osire, A.C. Goncalves, F.M. Alemayehu (eds.) Probabilistic Prognostics and Health Management of Energy Systems, pp. 127–156. Springer International Publishing (2017). DOI 10.1007/978-3-319-55852-3_8
9. Cunha Jr, A.: Enhancing the performance of a bi-stable energy harvesting device via the cross-entropy method. Nonlinear Dynamics **103**, 137–155 (2021). DOI 10.1007/s11071-020-06109-0
10. Cunha Jr, A., Nasser, R., Sampaio, R., Lopes, H., Breitman, K.: Uncertainty quantification through Monte Carlo method in a cloud computing setting. Computer Physics Communications **185**, 1355–1363 (2014). DOI 10.1016/j.cpc.2014.01.006
11. Dantas, E.: A cross-entropy strategy for parameters identification problems. Monograph, Universidade do Estado do Rio de Janeiro (2019). Https://dx.doi.org/10.13140/RG.2.2.18045.51688
12. Dantas, E., Cunha Jr, A., Silva, T.A.N.: A numerical procedure based on cross-entropy method for stiffness identification. In: 5th International Conference on Structural Engineering Dynamics (ICEDyn 2019). Viana do Castelo, Portugal (2019)
13. Dantas, E., Cunha Jr, A., Soeiro, F.J.C.P., Cayres, B.C., Weber, H.I.: An inverse problem via cross-entropy method for calibration of a drill string torsional dynamic

model. In: 25th ABCM International Congress of Mechanical Engineering (COBEM 2019). Uberlândia, Brazil (2019)
14. Dantas, E., Tosin, M., Cunha Jr, A.: Calibration of a SEIR-SEI epidemic model to describe Zika virus outbreak in Brazil. Applied Mathematics and Computation **338**, 249–259 (2018). DOI 10.1016/j.amc.2018.06.024
15. Dantas, E., Tosin, M., Cunha Jr, A.: An uncertainty quantification framework for a Zika virus epidemic model. Journal of Computational Interdisciplinary Sciences **10**, 91 (2019)
16. De Boer, P.T., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A tutorial on the cross-entropy method. Annals of Operations Research **134**, 19–67 (2005). DOI 10.1007/s10479-005-5724-z
17. Gamerman, D., Prates, M.O., Paiva, T., Mayrink (Editors), V.D.: Building a Platform for Data-Driven Pandemic Prediction From Data Modelling to Visualisation - The CovidLP Project. Chapman and Hall/CRC (2022)
18. Gianfelice, P.R.L., Oyarzabal, R.S., Cunha, A., Grzybowski, J.M.V., Batista, F.C., Macau, E.E.N.: The starting dates of COVID-19 multiple waves. Chaos: An Interdisciplinary Journal of Nonlinear Science **32**(3), 031101 (2022). DOI 10.1063/5.0079904
19. Grasselli, G., Zangrillo, A., Zanella, A., Antonelli, M., Cabrini, L., Castelli, A., Cereda, D., Coluccello, A., Foti, G., Fumagalli, R., Iotti, G., Latronico, N., Lorini, L., Merler, S., Natalini, G., Piatti, A., Ranieri, M.V., Scandroglio, A.M., Storti, E., Cecconi, M., Pesenti, A., Network, C..L.I.: Baseline Characteristics and Outcomes of 1591 Patients Infected With SARS-CoV-2 Admitted to ICUs of the Lombardy Region, Italy. Journal of the American Medical Association **323**(16), 1574–1581 (2020). DOI 10.1001/jama.2020.5394
20. He, S., Peng, Y., Sun, K.: SEIR modeling of the COVID-19 and its dynamics. Nonlinear Dynamics **101**(3), 1667–1680 (2021). DOI 10.1007/s11071-020-05743-y
21. Hethcote, H.W.: The mathematics of infectious diseases. SIAM Review **42**(4), 599–653 (2000). DOI 10.1137/S0036144500371907
22. Hindmarsh, A.C., Brown, P., Grant, K., Lee, S., Serban, R., Shumaker, D.E., Woodward, C.S.: SUNDIALS: Suite of Nonlinear and Differential/Algebraic Equation Solvers. ACM Transactions on Mathematical Software **31**(3), 363–396 (2005). DOI 10.1145/1089014.1089020
23. Hirsch, M.W., Smale, S., Devaney, R.L.: Differential Equations, Dynamical Systems, and an Introduction to Chaos, 3rd edn. Academic Press (2012)
24. Holmdahl, I., Buckee, C.: Wrong but Useful — What Covid-19 Epidemiologic Models Can and Cannot Tell Us. New England Journal of Medicine **383**(4), 303–305 (2020). DOI 10.1056/NEJMp2016822
25. Jaynes, E.T.: Probability Theory: the logic of science. Cambridge University Press (2003)
26. Jha, P., Cao, L., Oden, J.: Bayesian-based predictions of COVID-19 evolution in Texas using multispecies mixture-theoretic continuum models. Computational Mechanics **66**(5), 1055–1068 (2020). DOI 10.1007/s00466-020-01889-z
27. Kaipio, J., Somersalo, E.: Statistical and Computational Inverse Problems. Springer (2004)
28. Kharazmi, E., Cai, M., Zheng, X., Zhang, Z., Lin, G., Karniadakis, G.E.: Identifiability and predictability of integer- and fractional-order epidemiological models using physics-informed neural networks. Nature Computational Science **1**(11), 744–753 (2021). DOI 10.1038/s43588-021-00158-0
29. Kroese, D.P., Taimre, T., Botev, Z.I.: Handbook of Monte Carlo Methods. Wiley (2011)
30. Kucharski, A.J., Funk, S., Eggo, R.M., Mallet, H.P., Edmunds, W.J., Nilles, E.J.: Transmission dynamics of Zika virus in island populations: a modelling analysis of the 2013–14 French Polynesia outbreak. PLoS Neglected Tropical Disiases **10**(5) (2016). DOI 10.1371/journal.pntd.0004726
31. Kucharski, A.J., Russell, T.W., Diamond, C., Liu, Y., J., E., Funk, S., Eggo, R.M.: Early dynamics of transmission and control of COVID-19: a mathematical modelling study. The Lancet Infectious Diseases **20**(5), 553–558 (2020). DOI 10.1016/S1473-3099(20)30144-4
32. Kuhl, E.: Computational Epidemiology: Data-Driven Modeling of COVID-19. Springer (2021)
33. Kypraios, T., Neal, P., Prangle, D.: A tutorial introduction to bayesian inference for stochastic epidemic models using approximate bayesian computation. Mathematical Biosciences **287**, 42–53 (2017). DOI 10.1016/j.mbs.2016.07.001
34. Libotte, G.B., dos Anjos, L., Almeida, R.C.C., Malta, S.M.C., Silva, R.S.: Framework for enhancing the estimation of model parameters for data with a high level of uncertainty. Nonlinear Dynamics **107**, 1919–1936 (2022). DOI 10.1007/s11071-021-07069-9
35. Lobato, F.S., Libotte, G.B., Platt, G.M.: Mathematical modelling of the second wave of COVID-19 infections using deterministic and stochastic SIDR models. Nonlinear Dynamics **106**, 1359–1373 (2021). DOI 10.1007/s11071-021-06680-0
36. Lyra, W., do Nascimento, J.D., Belkhiria, J., de Almeida L., Chrispim, P.P.M., de Andrade, I.: COVID-19 pandemics modeling with modified determinist SEIR, social distancing, and age stratification. the effect of vertical confinement and release in Brazil. PLoS ONE **15**(9), e0237627 (2020). DOI 10.1371/journal.pone.0237627
37. Martcheva, M.: An Introduction to Mathematical Epidemiology. Springer, New York (2015)
38. McKinley, T.J., Vernon, I., Andrianakis, I., McCreesh, N., Oakley, J.E., Nsubuga, R.N., Goldstein, M., White, R.G.: Approximate Bayesian Computation and Simulation-Based Inference for Complex Stochastic Epidemic Models. Statistical Science **33**, 4 – 18 (2018). DOI 10.1214/17-STS618
39. MIDAS Network: MIDAS 2019 Novel Coronavirus Repository. `https://github.com/midas-network/COVID-19` (2020)
40. Minter, A., Retkute, R.: Approximate bayesian computation for infectious disease modelling. Epidemics **29**, 100368 (2019). DOI 10.1016/j.epidem.2019.100368
41. Morrison, R.E., Cunha Jr, A.: Embedded model discrepancy: A case study of Zika modeling. Chaos: An Interdisciplinary Journal of Nonlinear Science **30**, 051103 (2020). DOI 10.1063/5.0005204
42. Müller, J., Kuttler, C.: Methods and Models in Mathematical Biology: Deterministic and Stochastic Approaches. Springer, New York (2015)
43. Neal, P.J.: Approximate Bayesian Computation Methods for Epidemic Models. In: L. Held, N. Hens, P. O'Neill, J. Wallinga (eds.) Handbook of Infectious Disease Data Analysis, p. . Chapman and Hall/CRC (2019). DOI 10.1201/9781315222912
44. Nogrady, B.: What the data say about asymptomatic COVID infections. Nature **587**, 534–535 (2020). DOI 10.1038/d41586-020-03141-3

45. Oliveira, J.F., Jorge, D.C.P., Veiga, R.V., Rodrigues, M.S., Torquato, M.F., da Silva, N.B., Fiaccone, R.L., Cardim, L.L., Pereira, F.A.C., de Castro, C.P., Paiva, A.S.S., Amad, A.A.S., Lima, E.A.B.F., Souza, D.S., Pinho, S.T.R., Ramos, P.I.P., Andrade, R.F.S.: Mathematical modeling of COVID-19 in 14.8 million individuals in Bahia, Brazil. Nature Communications **12**, 333 (2021). DOI 10.1038/s41467-020-19798-3
46. Pacheco, P.M.C.L., Savi, M.A., Savi, P.V.: COVID-19 dynamics considering the influence of hospital infrastructure: an investigation into Brazilian scenarios. Nonlinear Dynamics **106** (2021). DOI 10.1007/s11071-021-06323-4
47. Pavlack, B., Grave, M., Dantas, E., Basilio, J., de la Roca, L., Norenberg, J.a., Tosin, M., Chaves, L., Matos, D., Issa, M., Luo, R., Guyt, A., Soares, L., Burgos, R., Lovisolo, L., Cunha, A.: EPIDEMIC - Epidemiology Educational Code. Journal of Open Source Education **5**, 149 (2022). DOI 10.21105/jose.00149
48. Perko, L.: Differential Equations and Dynamical Systems, 3rd edn. Springer (2006)
49. Prefitura do Rio de Janeiro: Painel Rio COVID-19. `http://coronavirus.rio/painel` (2022)
50. Rackauckas, C., Ma, Y., Martensen, J., Warner, C., K, Z., Supekar, R., Skinner, D., Ramadhan, A., Edelman, A.: Universal differential equations for scientific machine learning. arxiv p. 2001.04385 (2020). DOI 10.48550/arXiv.2001.04385
51. Rahman, S., Rahman, M.M., Miah, M., Begum, M.N., Sarmin, M., Mahfuz, M., Hossain, M.E., Rahman, M.Z., Chisti, M.J., Ahmed, T., Arifeen, S.E., Rahman, M.: COVID-19 reinfections among naturally infected and vaccinated individuals. Scientific Reports **12**, 1438 (2022). DOI 10.1038/s41598-022-05325-5
52. Raissi, M., Ramezani, N., Seshaiyer, P.: On parameter estimation approaches for predicting disease transmission through optimization, deep learning and statistical inference methods. Letters in Biomathematics **6**(2), 1–26 (2019). DOI 10.1080/23737867.2019.1676172
53. Roda, W.C.: Bayesian inference for dynamical systems. Infectious Disease Modelling **5**, 221–232 (2020). DOI 10.1016/j.idm.2019.12.007
54. Rubinstein, R.Y.: The cross-entropy method for combinatorial and continuous optimization. Methodology and Computing in Applied Probability **2**, 127–190 (1999). DOI https://doi.org/10.1023/A:1010091220143
55. Rubinstein, R.Y.: Simulation and the Monte Carlo Method, 3rd edn. Wieley (2016)
56. Rubinstein, R.Y., Kroese, D.P.: The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning. Information Science and Statistics. Springer-Verlag (2004)
57. Shaier, S., Raissi, M., Seshaiyer, P.: Data-driven approaches for predicting spread of infectious diseases through DINNs: Disease Informed Neural Networks (2021). DOI 10.48550/arxiv.2110.05445
58. Shampine, L.F., Reichelt, M.W.: The MATLAB ODE Suite **18**(1), 1–22 (1997). DOI 10.1137/S1064827594276424
59. Sivia, D.S.: Data analysis – a Bayesian tutorial. Oxford Science (2006)
60. Smith, R.C.: Uncertainty Quantification: Theory, Implementation and Applications. SIAM (2014)
61. Soize, C.: A nonparametric model of random uncertainties for reduced matrix models in structural dynamics. Probabilistic Engineering Mechanics **15**(3), 277–294 (2010). DOI 10.1016/S0266-8920(99)00028-4
62. Soize, C.: Uncertainty Quantification: An Accelerated Course with Advanced Applications in Computational Engineering. Springer (2017)
63. Statistica: Coronavirus (COVID-19) death rate in countries with confirmed deaths and over 1,000 reported cases as of november 2, 2021, by country. `https://www.statista.com/statistics/1105914/coronavirus-death-rates-worldwide/` (2021)
64. Strogatz, S.H.: Nonlinear Dynamics and Chaos: With Applications To Physics, Biology, Chemistry, And Engineering, 2nd edn. Westview Press (2014)
65. Sunnåker, M., Busetto, A.G., Numminen, E., Corander, J., Foll, M., Dessimoz, C.: Approximate Bayesian Computation. PLOS Computational Biology **9**(1), 1–10 (2013). DOI 10.1371/journal.pcbi.1002803
66. Taghizadeh, L., Karimi, A., Heitzinger, C.: Uncertainty quantification in epidemiological models for the COVID-19 pandemic. Computers in Biology and Medicine **125**, 104011 (2020). DOI 10.1016/j.compbiomed.2020.104011
67. Tarantola, A.: Inverse Problem Theory and Methods for Model Parameter Estimation. SIAM (2005)
68. Tenorio, L.: An Introduction to Data Analysis and Uncertainty Quantification for Inverse Problems. SIAM (2017)
69. Tolles, J., Luong, T.: Modeling Epidemics With Compartmental Models. Journal of the American Medical Association **323**(24), 2515–2516 (2020). DOI 10.1001/jama.2020.8420
70. Toni, T., Welch, D., Strelkowa, N., Ipsen, A., Stumpf, M.P.H.: Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of the Royal Society Interface **6**(31), 187–202 (2009). DOI 10.1098/rsif.2008.0172
71. Tosin, M.: Modeling and uncertainty quantification in the nonlinear dynamics of epidemiological phenomena: Application to Zika virus and COVID-19 outbreaks. Master's thesis, Universidade do Estado do Rio de Janeiro, Rio de Janeiro (2021)
72. Vasconcelos, G.L., Brum, A.A., Almeida, F.A.G., Macêdo, A.M.S., Duarte-Filho, G.C., Ospina, R.: Standard and anomalous waves of COVID-19: A multiple-wave growth model for epidemics. Brazilian Journal of Physics **51**, 1867–1883 (2021). DOI 10.1007/s13538-021-00996-3
73. Verhulst, F.: Nonlinear Differential Equations and Dynamical Systems, 2nd edn. Springer (2012)
74. Vyasarayani, C.P., Chatterjee, A.: New approximations, and policy implications, from a delayed dynamic model of a fast pandemic. Physica D: Nonlinear Phenomena **414**, 132701 (2021). DOI 10.1016/j.physd.2020.132701
75. Vynnycky, E., White, R.: An Introduction to Infectious Disease Modelling. Oxford University Press (2010)
76. Wang, B.: Parameter estimation for ODEs using a cross-entropy approach. Master's thesis, University of Toronto, Toronto (2012)
77. Wang, D., et al.: Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. Journal of the American Medical Association **323**(11), 1061–1069 (2021). DOI 10.1001/jama.2020.1585
78. Weitz, J.S., Park, S.W., Eksin, C., Dushoff, J.: Awareness-driven behavior changes can shift the shape of epidemics away from peaks and toward plateaus, shoulders, and oscillations. Proceedings of the National Academy of Sciences **117**(51), 32764–32771 (2020). DOI 10.1073/pnas.2009911117
79. WHO: Coronavirus disease 2019 (COVID-19). Situation report 24. Geneva: World Health Organization (2020)

80. Wu, P., Hao, X., Lau, E.H.Y., Wong, J.Y., Leung, K.S.M., Wu, J.T., Cowling, B.J., Leung, G.M.: Real-time tentative assessment of the epidemiological characteristics of novel coronavirus infections in wuhan, china, as at 22 january 2020. Eurosurveillance **25**(3), (2020). DOI 10.2807/1560-7917.es.2020.25.3.2000044
81. Yazdani, A., Lu, L., Raissi, M., Karniadakis, G.E.: Systems biology informed deep learning for inferring parameters and hidden dynamics. PLOS Computational Biology **16**(11), 1–19 (2020). DOI 10.1371/journal.pcbi.1007575
82. Yu, X., Lu, L., Shen, J., Li, J., Xiao, W., Chen, Y.: A fractional-order SEIHDR model for COVID-19 with inter-city networked coupling effects. Nonlinear Dynamics **101**(3), 1717–1730 (2021). DOI 10.1007/s11071-020-05848-4
83. Zhang, S., Ponce, J., Zhang, Z., Lin, G., Karniadakis, G.: An integrated framework for building trustworthy data-driven epidemiological models: Application to the COVID-19 outbreak in New York City. PLOS Computational Biology **17**(9), 1–29 (2021). DOI 10.1371/journal.pcbi.1009334
84. Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X., Guan, L., Wei, Y., Li, H., Wu, X., Xu, J., Tu, S., Zhang, Y., Chen, H., Cao, B.: Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. The Lancet **395**, 1054–1062 (2020). DOI 10.1016/S0140-6736(20)30566-3