# A stochastic Bayesian bootstrapping model for COVID-19 data

Julia Calatayud [a], Marc Jornet [b], Jorge Mateu [c]

[a] Departament de Matemàtiques, Universitat Jaume I, 12071 Castellón, Spain.
email: calatayj@uji.es
ORCID: 0000-0002-9639-1530

[b] Departament de Matemàtiques, Universitat de València, 46100 Burjassot, Spain.
email: marc.jornet@uv.es
ORCID: 0000-0003-0748-3730

[c] Departament de Matemàtiques, Universitat Jaume I, 12071 Castellón, Spain.
email: mateu@uji.es
ORCID: 0000-0002-2868-7604

**Abstract.** We provide a stochastic modeling framework for the incidence of COVID-19 in Castilla-Leon (Spain) for the period March 1, 2020 to February 12, 2021, which encompasses four waves. Each wave is appropriately described by a generalized logistic growth curve. Accordingly, the four waves are modeled through a sum of four generalized logistic growth curves. Pointwise values of the twenty input parameters are fitted by a least-squares optimization procedure. Taking into account the significant variability in the daily reported cases, the input parameters and the errors are regarded as random variables on an abstract probability space. Their probability distributions are inferred from a Bayesian bootstrap procedure. This framework is shown to offer a more accurate estimation of the COVID-19 reported cases than the deterministic formulation.

*Keywords:* Bayesian bootstrap, COVID-19 reported infections and waves, Deterministic and stochastic modeling, Least-squares fitting, Multiple generalized logistic growth curves, Random parameters and errors

## 1. Introduction

COVID-19 is an infectious disease caused by coronavirus SARS-CoV-2. It was detected for the first time in Wuhan, China, in December 2019, and quickly spread around the globe becoming an ongoing pandemic. The virus is rapidly transmitted between persons through small droplets and aerosols. The most common symptoms of the disease are fever, dry cough and fatigue. Data to date suggest that 80% of infections are mild or asymptomatic, 15% are severe, and 5% are critical. Lethality strongly depends on age and comorbidities. As of July 2021, more than 190 million people have been infected and more than 4 million people have died. To contain the spread of the virus and alleviate the pressure on the health systems, governments put several restrictions such as city or country lockdowns, quarantines, social distancing and hygiene measures, curfews, mandatory masks, etc. [19, 44, 45, 47].

The use of mathematical models is an effective tool to describe and predict the evolution of epidemics and to propose targeted measures [6, 33]. Due to the fast transmissibility of SARS-CoV-2 and the containment measures frequently implemented by governments, the modeling of its spread is a difficult problem. For example, as already noticed by other researchers [1, 28, 29], the usual autonomous SIR (susceptible-infected-recovered) model cannot capture the quick variations of COVID–19 reported infections.

When aggregated time-series data are present, the logistic differential equation model may be useful to fit the measurements on COVID-19 infections [43]. The curve is characterized by an increasing growth in the beginning period, combined with a decreasing growth at a later stage.

Generalizations of the logistic growth curve, to allow for more general sigmoid shapes, have been employed to model COVID-19 [3,22,31,46]. Those approaches seem to be adequate when fitting a single wave of the COVID-19 epidemic. For multiple phases of growth, as occurs with COVID-19 infection waves, a single $S$-shaped trajectory is not applicable. In this line, ideas from [27] on the bi-logistic model have been used to fit cumulative cases of COVID-19 along two different outbreaks or waves [9,36,49]. These types of models, usually called phenomenological or statistical, are often useful to reproduce and forecast the course of an epidemic [7,32], when the insight is limited, treatments and interventions are rapidly changing, data vary abruptly, and mechanistic models (compartmental models with laws of transmission) present difficulties.

The incidence of an epidemic and its modeling have intrinsic uncertainties that are irreducible (random uncertainty). Thus, to better mimic reality, models should incorporate stochastic components [38]. For example, reference [22] also included stochastic effects through a Bayesian formalization. Bayesian inference has been suggested for other models to accommodate COVID-19 data, such as the Gompertz curve [4]. However, the use of the Bayesian bootstrap [34] to infer input uncertainties does not seem to have been investigated so far.

In this paper, we investigate the use of four combined generalized logistic differential equations to model the four waves of the COVID-19 epidemic in Castilla-Leon (a Spanish autonomous region with 2.5 million inhabitants) at once. The temporal period runs from March 1, 2020 to February 12, 2021. The calibration of model parameters is conducted by a least-squares minimization procedure. Due to data measurement errors, some form of stochasticity is introduced into the model. The input parameters and the errors are considered as random variables, whose probability distributions are inferred from a Bayesian bootstrap technique.

The plan of the paper is the following. Section 2 provides the methodology, including a brief description of the data, and the deterministic and stochastic approaches. Section 3 presents the results (numerical calculations, fittings and plots). A discussion comes in Section 4. The paper ends with some final conclusions in Section 5.

## 2. Methods

2.1. **Data.** We have data on the number of new daily COVID-19 reported infections in the Spanish autonomous region of Castilla-Leon. This region is the largest community in Spain by area, it is located in the northwest of Spain, and it has a population of around 2.5 million. The data correspond to the temporal period of almost one year, from March 1, 2020 to February 12, 2021, which encompasses four waves of the epidemic. The cases have been retrieved from the open data portal of Castilla-Leon [1]. This dataset only captures a small fraction of the true burden, due to asymptomatic cases, lack of resources and omission of suspected but not confirmed cases. In this paper, we treat Castilla-Leon as a whole where people interact homogeneously.

In Figure 1, the number of daily new reported infections, for 349 consecutive days, is depicted in the left panel, while the accumulated number of daily reported infections is shown in the right panel. We note that there are four clear waves of the epidemic. The first wave corresponds to the first entrance of the virus in Spain, which ended up in summer 2020 due to the severe lockdown imposed by authorities. The second wave started after summer 2020 due to relaxation of measures and overlapped with a larger third wave along autumn. Finally, the fourth wave began after the end-of-year vacations and ended in February 2021 due to some restrictions and the vaccination program. A significant variability in the daily data, with abruptly increasing and decreasing magnitudes, is observed between nearby days, highlighting some sort of uncertainty entailing some stochastic nature in the data. This may be due to

[1] `https://datosabiertos.jcyl.es/web/es/datos-abiertos-castilla-leon.html`

highly variable factors, such as the quantity of tests available and performed, symptomatology, etc. The implementations and computations are performed with Mathematica®, version 12.0, and are included as supplementary material, where the data are available (variable *vtotal*).

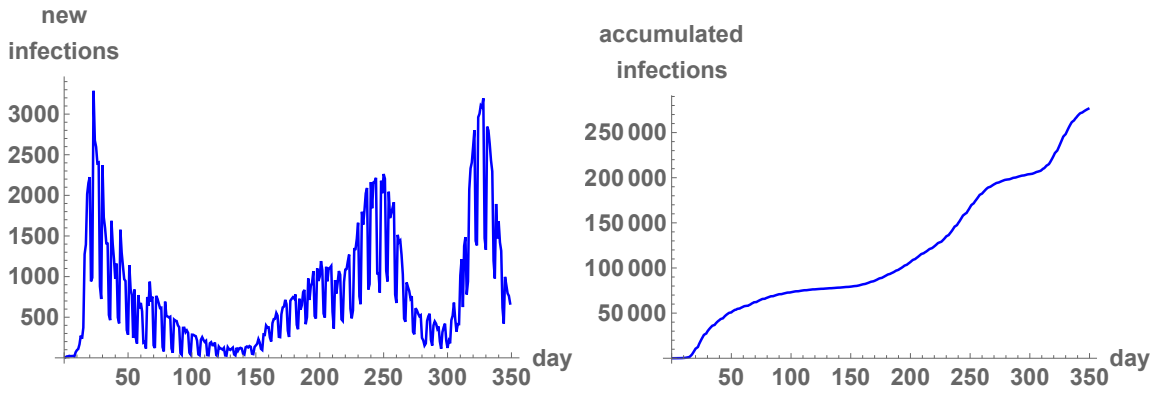


FIGURE 1. *Left panel*: number of new daily reported infections. *Right panel*: accumulated daily number of reported infections.

2.2. **Multiple generalized logistic growth curves.** In this subsection, the reported infections of the COVID-19 disease presented in Figure 1 are modeled. Each wave of the epidemic – the accumulated version from the right panel of Figure 1 – is described by means of a generalized logistic growth curve. The four waves are then modeled by juxtaposing four generalized logistic growth curves. The input parameters of the complete model are calibrated by certain optimization procedures. This framework provides a smooth curve that fits the data from both panels of Figure 1. Note that in this subsection, stochastic effects are not taken into account yet.

2.2.1. *A generalized logistic differential equation.* The Malthusian model, proposed by T.R. Malthus in 1798 in an essay [24], describes an exponential growth in a population through the ordinary differential equation

$$y'(t) = ay(t),$$

where $a$ is a positive parameter defined as the intrinsic growth rate. Given an initial condition $y(t_0) = y_0$, the solution of the above equation is given by

$$y(t) = y_0 \mathrm{e}^{a(t-t_0)}.$$

A modern formulation of the Malthusian growth model can be read at any introductory text [30]. In the field of population ecology, it is considered as the first law of population dynamics [41].

In order to capture the decrease in the growth rate with time, P.F. Verhulst proposed in 1838 [18, 42] the logistic model, given by

$$y'(t) = ay(t)\left(1 - \frac{y(t)}{K}\right),$$

where $K > 0$ is the carrying capacity (the limit of $y(t)$ as $t$ tends to infinity). This is a Bernoulli-type ordinary differential equation. Given an initial condition $y(t_0) = y_0$, the solution is known in closed form

$$y(t) = \frac{K}{1 + \left(-1 + \frac{K}{y_0}\right)\mathrm{e}^{-a(t-t_0)}}.$$

This function was employed to forecast an Ebola epidemic [8]. The saturation effect implicitly captures public health interventions, without complex mechanistic assumptions about the transmission process.

To allow for more flexible $S$-shaped curves to model growth phenomena over time, the following modification of the logistic differential equation has been suggested in the literature

$$y'(t) = ay(t)\left(1 - \left(\frac{y(t)}{K}\right)^b\right). \tag{2.1}$$

Originally, (2.1) was used for the analysis of tumor growth [5, 26, 35, 39], though applications in epidemiology are also found. Examples of diseases include SARS [14, 16], dengue fever [17], influenza H1N1 [15], Zika [7], Ebola [32], and COVID-19 [3, 22, 31, 46]. Here $b > 0$ is a power that controls the asymmetry of the curve and how fast the limiting number $K$ is approached. It endows the model with higher flexibility. When $b = 1$, the classical logistic differential equation is obtained, and when $b$ tends to 0, the Gompertz equation is recovered. This generalization belongs to the class of Bernoulli differential equations too. If $y(t_0) = y_0$ is the initial condition, the solution takes the form

$$y(t) = \frac{K}{\left[1 + \left(-1 + \left(\frac{K}{y_0}\right)^b\right)\mathrm{e}^{-ab(t-t_0)}\right]^{\frac{1}{b}}}. \tag{2.2}$$

This is called a generalized logistic growth curve (or sometimes Richards' curve). It may be appropriate to model the aggregated cases of a single wave of the COVID-19 epidemic. For new cases (not accumulated), consecutive differences $y(t) - y(t-1)$ are considered.

2.2.2. *Combination of growth curves.* For multiple phases of growth, a single sigmoid curve is not appropriate to describe such data. Thus, we propose a combination of generalized logistic growth curves of the form (2.2). This is an extension of the work initiated by P.S. Meyer for the bi-logistic model [27], with subsequent applications in sociology [13], agriculture [37] or epidemiology [9, 20, 36, 49], for instance.

Mathematically, a combination of generalized logistic growth curves takes the following form

$$y(t) = \sum_i \frac{K_i}{\left[1 + \left(-1 + \left(\frac{K_i}{y_{0,i}}\right)^{b_i}\right)\mathrm{e}^{-a_i b_i (t-t_{0,i})}\right]^{\frac{1}{b_i}}}. \tag{2.3}$$

This sum of trajectories, supplied with four terms $i = 1, 2, 3, 4$, allows modeling the accumulated cases of the four concatenated COVID-19 waves. In practice, one should try fitting several concatenated models and compare their goodness-of-fit. In our case, we tried three terms, but the fit was not good. For new cases (not accumulated), consecutive differences $y(t) - y(t-1)$ are considered. It is important to note that the four generalized logistic growth curves are not independent (the four waves are not treated independently). Finally, notice that $t_{0,i}$ captures the beginning of the $i$-th wave. Each $K_i$ measures the highest infection level of the $i$-th wave.

2.2.3. *A deterministic fit.* Let $d_l$ be the number of new reported cases at time $l \in \{1, \ldots, 349\}$. Let $I_l$ be the number of accumulated reported cases at time $l \in \{1, \ldots, 349\}$, scaled by the total population in Castilla-Leon ($N_0 = 2.408 \times 10^6$ inhabitants). The following simple relations hold

$$d_l = (I_l - I_{l-1}) \times N_0,$$

$$I_l \times N_0 = \sum_{j=1}^{l} d_j.$$

A combination of generalized logistic growth curves as in (2.3) is used to model $\{I_l\}_{l=1}^{349}$. The parameters of (2.3) are estimated at once by a deterministic least-squares procedure [40],

$$\min_{\text{parameters}} \sum_{l=1}^{349} \left(I_l - y(l)\right)^2. \tag{2.4}$$

In most of the cases, the model curve will not go through all the data, so this minimum value will not be zero. The minimum gives a measure of how good the fit is.

2.3. **Parametric randomization of the model.** We now proceed with modeling the daily random variability of the data through a probabilistic setting. Following a Bayesian formalism, the parameters and the errors will be regarded as random variables on an abstract probabilistic space. Their probability distributions will be then inferred by means of Bayesian bootstrapping. In this context, the output of the model is a stochastic process, which will render a stochastic fit of the data, rather than an averaged estimation.

2.3.1. *Randomization.* Part of the variability of the data is not captured by the deterministic model. We assume the errors are random variables naturally defined in a probabilistic framework. We also consider the lack of knowledge on the input parameters, which prescribe the constitutive laws of the system, represented within a probabilistic framework [38, chapter 1], [21, chapter 1]. Thus, we consider in this new setup the parameters and the errors of the model as random variables.

The field of uncertainty quantification studies the impact of random uncertainties on models [38]. This quantification is necessary to evaluate the discrepancies between the model predictions and the current system behavior. Inverse uncertainty quantification deals with inference of the probability distributions of the parameters from the data. These probability distributions are not, in general, independent. Forward uncertainty quantification extracts the main statistical content of the model output, once the probability distributions of the parameters are fixed [21,48]. The various stages of the probabilistic modeling process are schematically illustrated in Figure 2. Inverse uncertainty quantification is not an easy task. Here we rely on the Bayesian bootstrap technique. Forward uncertainty quantification will be conducted via Monte Carlo simulation.

2.3.2. *Bayesian bootstrap.* Given a mathematical model, the model error coming from a deterministic least-squares optimization technique may be regarded as a random variable $X$. The error varies with time, and thus the errors at different instants of time may be seen as copies of $X$. In this context, the errors of the model are identically distributed and independent random variables $X_1, \ldots, X_m$, where $m$ is the length of the data. Actually, these random variables are unknown; only different realizations $x_1, \ldots, x_m$ are available from the data and the deterministic fit. The bootstrap methodology assumes that the observed residuals $x_1, \ldots, x_m$ are all possible distinct values of $X$, based on the principle that all observed variables are discrete.

The Bayesian bootstrap, developed by D.B. Rubin [34], infers the distribution of $X$ by resampling $x_1, \ldots, x_m$ with Dirichlet weights of coefficients $1, \ldots, 1$. This procedure corresponds to the following hierarchical Bayesian statistical model

$$(X_1, \ldots, X_m)|(p_1, \ldots, p_m) \sim \underset{m \text{ times}}{\otimes} \text{Cat}(m, (p_1, \ldots, p_m)),$$

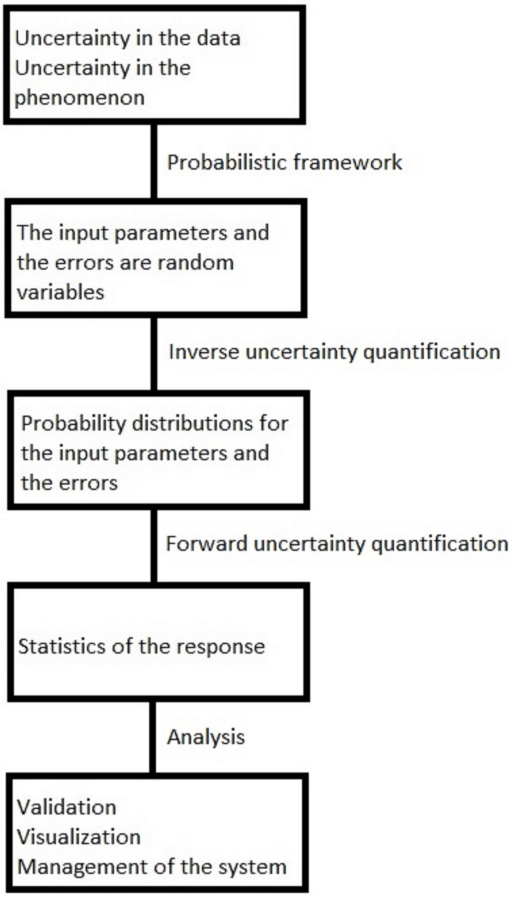


FIGURE 2. Schematic illustration of the various stages of the probabilistic modeling process.

with

$$(p_1, \ldots, p_m) \sim \mathrm{Dir}(m, (0, \ldots, 0)).$$

Bayes' theorem yields the posterior distribution

$$(p_1, \ldots, p_m)|(x_1, \ldots, x_m) \sim \mathrm{Dir}(m, (1, \ldots, 1)),$$

because, for $p_1 + \ldots + p_m = 1$,

$$\begin{aligned}\pi(p_1, \ldots, p_m|x_1, \ldots, x_m) &\propto \pi(x_1, \ldots, x_m|p_1, \ldots, p_m) \times \pi(p_1, \ldots, p_m) \\ &\propto \prod_{k=1}^{m} p_k \times \prod_{k=1}^{m} p_k^{-1} = 1 \sim \mathrm{Dir}(m, (1, \ldots, 1)).\end{aligned}$$

That is, the posterior proportions are uniformly distributed on the simplex. Here Cat is the Categorical distribution on $\{x_1, \ldots, x_m\}$ and Dir stands for the Dirichlet distribution, which is a conjugate prior. Parameters $(0, \ldots, 0)$ and $(1, \ldots, 1)$ come from the frequencies in $x_1, \ldots, x_m$ before and after observing them, respectively. Note that $\mathrm{Dir}(m, (0, \ldots, 0))$ is an improper prior.

The $\mathrm{Dir}(m, (1, \ldots, 1))$ distribution can be sampled as follows. From independent realizations $u_1, \ldots, u_{m-1} \sim \mathrm{Unif}(0, 1)$, assume that these values are ordered as $u_1 \leq \ldots \leq u_{m-1}$. Set

$u_0 = 0$ and $u_m = 1$. Define $g_k = u_k - u_{k-1}$, $k = 1, \dots, m$. Then $g_1, \dots, g_m$ are $m$ independent realizations of $\mathrm{Dir}(m, (1, \dots, 1))$.

For each resampling of $(x_1, \dots, x_m)$ with Dirichlet weights of coefficients $(1, \dots, 1)$, the input parameters of the model are determined by a least-squares fitting. Formally, it is assumed that

$$\text{parameters} = \Lambda(X_1, \dots, X_m) \text{ almost surely}, \quad \Lambda = \text{least-squares fitting operator}.$$

This gives rise to samples of the input parameters, so their (posterior) probability distributions may be inferred as

$$\begin{aligned}&\text{parameters}|(x_1, \dots, x_m) \propto \\ &\qquad \text{parameters}|(X_1, \dots, X_m) \times (X_1, \dots, X_m)|(p_1, \dots, p_m) \times (p_1, \dots, p_m)|(x_1, \dots, x_m).\end{aligned}$$

This allows solving the problem of inverse uncertainty quantification.

2.3.3. *Monte Carlo simulation.* Monte Carlo simulation is a popular method for forward uncertainty quantification. It is simple to implement and robust. It uses a collection $\{\gamma_1, \dots, \gamma_M\}$ of independent random realizations of the model response, usually obtained from deterministic numerical techniques. The statistics of the model response are derived from the statistics of that sample. For example, the mean of a function $h$ of the model output is estimated by $\frac{1}{M}\sum_{k=1}^{M} h(\gamma_k)$, by the law of large numbers. Monte Carlo simulation essentially amounts to conducting $M$ deterministic resolutions, where $M$ is generally large. The robustness of the method is due to its independence of the random dimensionality, the variable $t$, or regularity issues. Thus, in contrast to spectral methods, its use is advantageous when there is a large number of input random parameters or the variable $t$ may be large. Further, if the inverse parameter estimation method generates realizations of the parameters (such as Bayesian methods), then Monte Carlo simulation seems the logic option for forward uncertainty quantification. The convergence of the Monte Carlo estimate behaves as $M^{-1/2}$ due to the central limit theorem. It is assessed based on the estimated statistics. The reader is referred to [21, 38, 48] for similar discussions on Monte Carlo sampling.

2.3.4. *A stochastic fit.* The aim here is to randomize the deterministic model. The input parameters and the model errors are assumed to be random variables. Note that to apply the Bayesian bootstrap methodology, one needs errors that are identically distributed and independent. However, the fit for the accumulated infections does not yield independent residuals. Indeed, two consecutive residuals are correlated, because of the increasing character of the curve. Also, in addition, the fit for the daily new infections does not yield identically distributed errors. Certainly, a resampling of residuals may give rise to negative data points, which does not make sense. To fix these issues and achieve, as far as possible, an independent and identically distributed sample, the (natural) logarithms of the daily new infections will be considered. While the parameters of model (2.3) for cumulative infections are calibrated by least-squares fitting, the resampling is performed for the residuals obtained for the logarithms of the daily new infections. For each resampling of these residuals, we come back to cumulative infections and refit model (2.3). In this way, realizations of the model parameters are obtained for each refit.

The following steps summarize the procedure for estimating the probability distributions of the parameters:

**Step 1:** Determine the residuals for the logarithms of the daily new infections: $x_l = \log(d_l) - \log((y(l) - y(l-1)) \times N_0)$. The parameters of $y$ were previously determined by the deterministic least-squares fitting (2.4).

**Step 2:** Start a FOR loop to generate $M$ bootstrap samples.

**Step 3:** Resample the residuals with Dirichlet weights of coefficients $(1,\dots,1)$. Keep the resampling as a vector $x^*_{\text{iteration}}$ with components $x^*_{\text{iteration}}(l)$.
**Step 4:** Define $z_l = \log(d_l) + x^*_{\text{iteration}}(l)$. These are the new generated data for the logarithms of the daily new infections.
**Step 5:** Let $\tilde{I}_l = (\sum_{j=1}^{l} \mathrm{e}^{z_j})/N_0$. These are the new generated data for the accumulated infections.
**Step 6:** Fit the deterministic model (2.3) to $\tilde{I}_l$ by least-squares fitting. Keep the parameters as a vector $\lambda^*_{\text{iteration}}$.
**Step 7:** End the FOR loop.
**Output:** An ensemble of realizations $(x^*_1,\dots,x^*_M)$ for the residuals (errors) of the logarithms of the daily new infections. An ensemble of realizations $(\lambda^*_1,\dots,\lambda^*_M)$ for the input parameters.

Once $(x^*_1,\dots,x^*_M)$ and $(\lambda^*_1,\dots,\lambda^*_M)$ are available, the following steps outline the procedure to calculate the realizations of the model output:

**Step 1:** Start a FOR loop over $k = 1,\dots,M$.
**Step 2:** With parameters $\lambda^*_k$, evaluate $\alpha_k(l) = \log((y(l) - y(l-1)) \times N_0)$. This gives a vector $\alpha_k$, a discrete sample path of the model for the logarithms of the daily new infections.
**Step 3:** Incorporate the error: $\beta_k = \alpha_k + x^*_k$. This is a a discrete sample path of the model for the logarithms of the daily new infections, taking into account the random errors.
**Step 4:** Let $\gamma_k(l) = (\sum_{j=1}^{l} \mathrm{e}^{\beta_k(j)})/N_0$. Here $\gamma_k$ is a discrete sample path for the accumulated infections.
**Step 5:** End the FOR loop.
**Output:** The $M$ discrete sample paths $(\gamma_1,\dots,\gamma_M)$ of the model output for the accumulated infections. From them, statistics such as the mean, the variance, quantiles, etc. may be determined by means of Monte Carlo simulation.

## 3. Results

3.1. **Deterministic fit.** We have used the built-in function *FindFit* in Mathematica® to obtain the optimal parameters in (2.4), which are (up to four significant digits) as follows:

$$y_{0,1} = 2.9 \times 10^{-6},\ b_1 = 0.1675,\ a_1 = 0.7396,\ K_1 = 0.01982,\ t_{0,1} = 0,$$

$$y_{0,2} = 0.008758,\ b_2 = 0.8369,\ a_2 = 0.1142,\ K_2 = 0.008834,\ t_{0,2} = 113.7,$$

$$y_{0,3} = 0.01955,\ b_3 = 4.595,\ a_3 = 0.02152,\ K_3 = 0.05543,\ t_{0,3} = 212.1,$$

$$y_{0,4} = 0.001298,\ b_4 = 0.8733,\ a_4 = 0.1505,\ K_4 = 0.03213,\ t_{0,4} = 303.6.$$

The fitted model is depicted in Figure 3. In the left panel, the aggregated reported infections in percentages, $y(l) \times 100$, are shown, while the right panel shows the estimated new daily infections, $(y(l) - y(l-1)) \times N_0$. The combination of generalized logistic growth curves allows for a good fit of the COVID-19 incidence, at least from an averaged (smoothed) point of view. In the following subsection, stochasticity will be taken into account to deal with the daily random variability of the measurements. In Figure 4, the predictability of the model is quantified. We use two train sets, up to $t = 200$ and up to $t = 300$, and predict the epidemic size a few days later. The forecast is reasonably good for two weeks. It seems to be better at $t = 200$ than at $t = 300$, possibly due to the fact that $t = 200$ corresponds to the peak of the second wave, while $t = 300$ corresponds to the very early growth phase of the fourth wave. Predictions seem to be better when a larger dataset of the wave is available.

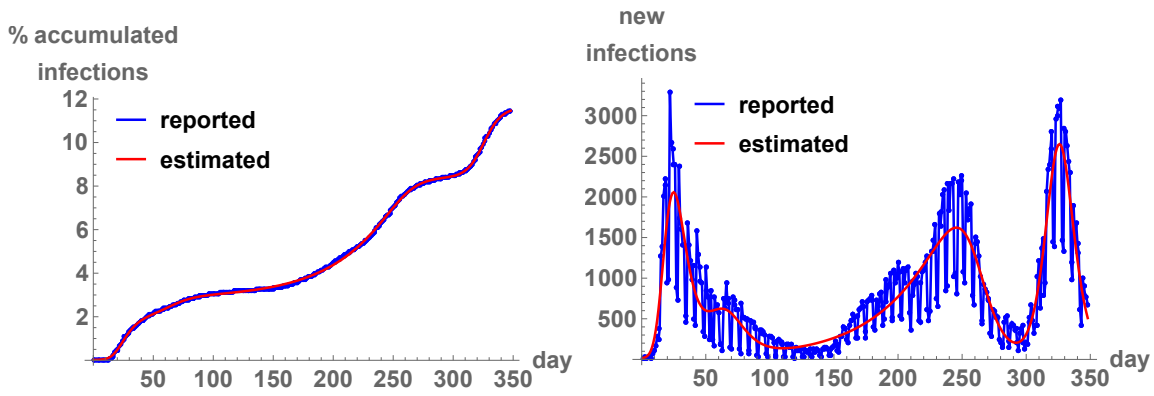


FIGURE 3. *Left panel*: deterministic fit for the daily accumulated reported infections in percentages. *Right panel*: deterministic fit for the new daily reported infections.

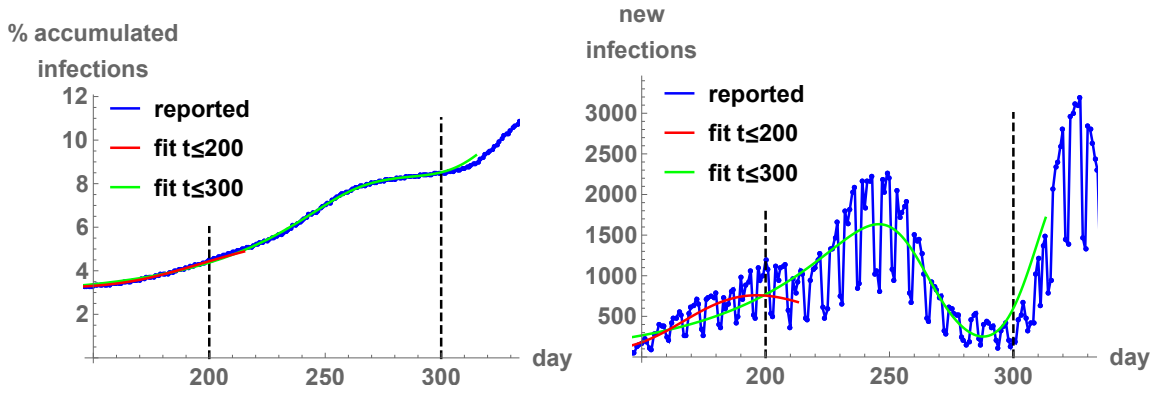


FIGURE 4. *Left panel*: deterministic prediction for the daily accumulated reported infections in percentages. *Right panel*: deterministic prediction for the new daily reported infections. The vertical dashed line indicates the end of the calibration period.

3.2. **Stochastic fit.** Table 1 reports the estimated marginal posterior statistics (mean, median, standard deviation, and quantiles 0.025 and 0.975) of the corresponding parameters. Figure 5 plots histograms for some marginal posterior distributions. Realizations $(\lambda_1^*, \ldots, \lambda_M^*)$ are employed, for $M = 1000$ bootstrap samples (larger bootstrap samples do not render significant differences at the scale of the figures). In Figure 6, the fit of the randomized model is shown, both for the accumulated reported infections (left panel) and the new reported infections per day (right panel). We draw mean values and regions of probability 0.95 for $(\gamma_1, \ldots, \gamma_M)$. Optical inspection shows that the estimated mean values for the model response are very similar to the deterministic fit from Figure 3. Thus, the deterministic fit is extended by incorporating probabilistic features. In the stochastic approach, the probability regions must contain the variability of the data, but in a correct way, in the sense that the realizations generated should resemble the pattern of the data. A stochastic method whose probability regions are unnecessarily wide is not good, despite containing all recorded measurements. To better appreciate the similarity between the real data and the stochastic model, compared to the deterministic

counterpart, some realizations ($\gamma_1$, $\gamma_2$ and $\gamma_3$) are plotted in Figure 7. In Figure 8, the predictability of the model is illustrated. We use two sets of incidence data, up to $t = 200$ and up to $t = 300$ as in the deterministic subsection.

| | mean | median | standard deviation | quantile 0.025 | quantile 0.975 |
|---|---|---|---|---|---|
| $b_1$ | 0.350 | 0.283 | 0.224 | 0.103 | 0.892 |
| $a_1$ | 0.580 | 0.513 | 0.191 | 0.361 | 0.962 |
| $K_1$ | 0.018 | 0.018 | 0.004 | 0.010 | 0.024 |
| $y_{0,2}$ | 0.011 | 0.010 | 0.003 | 0.005 | 0.018 |
| $b_2$ | 0.564 | 0.674 | 0.316 | 0.074 | 0.930 |
| $a_2$ | 0.248 | 0.137 | 0.203 | 0.104 | 0.769 |
| $K_2$ | 0.011 | 0.010 | 0.005 | 0.006 | 0.019 |
| $t_{0,2}$ | 115.1 | 114.8 | 2.839 | 111.7 | 118.4 |
| $y_{0,3}$ | 0.019 | 0.019 | 0.002 | 0.016 | 0.021 |
| $b_3$ | 4.183 | 4.490 | 0.857 | 2.146 | 4.985 |
| $a_3$ | 0.025 | 0.022 | 0.043 | 0.020 | 0.027 |
| $K_3$ | 0.055 | 0.055 | 0.004 | 0.049 | 0.060 |
| $t_{0,3}$ | 212.6 | 212.4 | 1.412 | 211.5 | 213.7 |
| $y_{0,4}$ | 0.001 | 0.001 | 0.0003 | 0.0004 | 0.002 |
| $b_4$ | 0.712 | 0.814 | 0.269 | 0.131 | 0.984 |
| $a_4$ | 0.220 | 0.168 | 0.138 | 0.122 | 0.690 |
| $K_4$ | 0.032 | 0.032 | 0.003 | 0.026 | 0.037 |
| $t_{0,4}$ | 304.4 | 304.2 | 1.138 | 302.5 | 306.8 |

TABLE 1. Estimated marginal posterior statistics (mean, median, standard deviation, and quantiles 0.025 and 0.975) of the input random parameters.

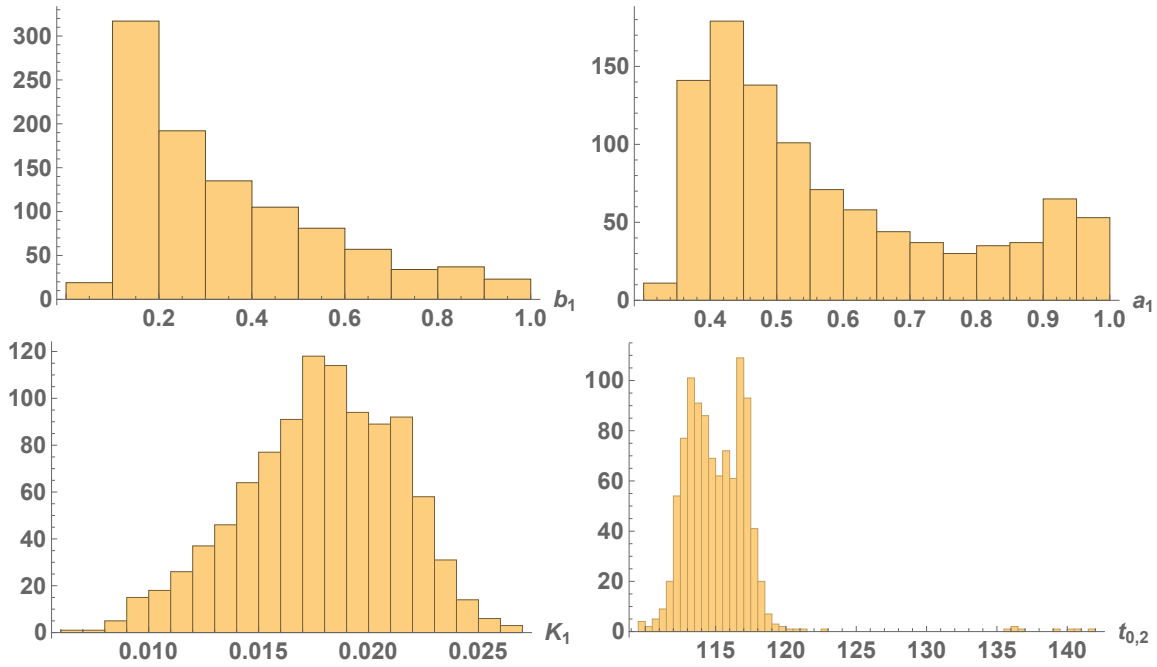


FIGURE 5. Histograms for some marginal posterior distributions.

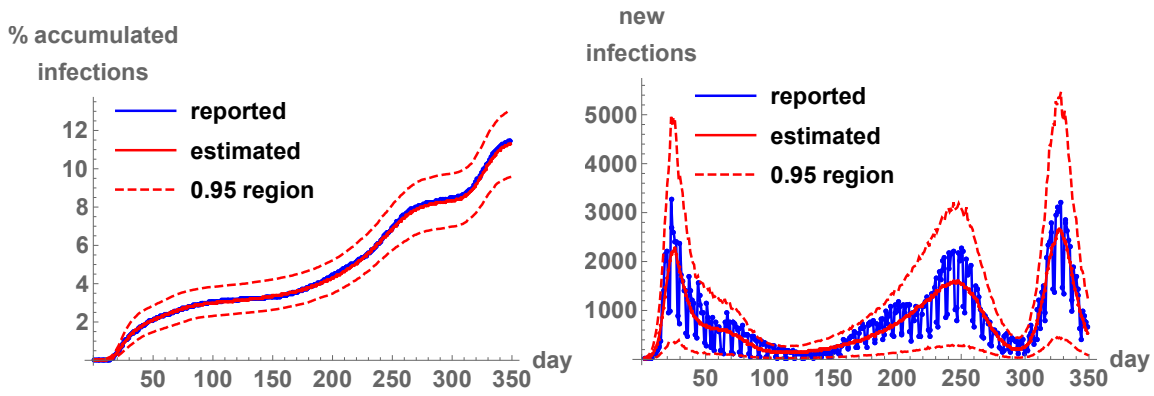


FIGURE 6. *Left panel*: stochastic fit for the accumulated daily reported infections in percentages. *Right panel*: stochastic fit for the new daily reported infections.

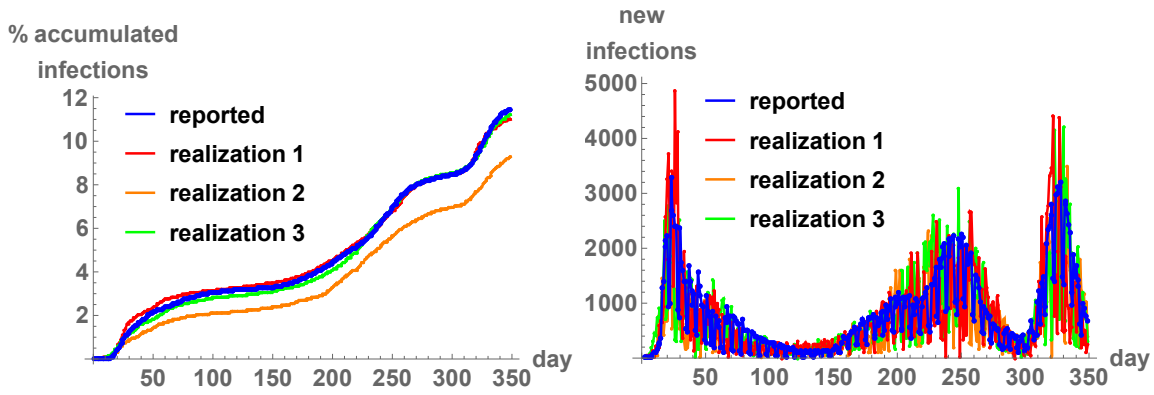


FIGURE 7. *Left panel*: some realizations of the stochastic model for the accumulated daily reported infections in percentages. *Right panel*: some realizations of the stochastic model for the new daily reported infections.

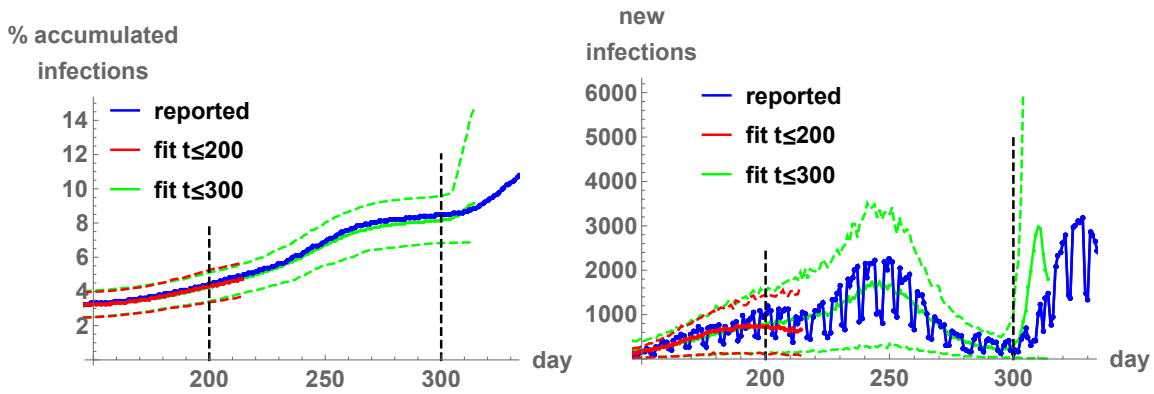


FIGURE 8. *Left panel*: stochastic prediction for the daily accumulated reported infections in percentages. *Right panel*: stochastic prediction for the new daily reported infections. The vertical dashed line indicates the end of the calibration period.

## 4. Discussion

The COVID-19 spread in Castilla-Leon is modeled along the four waves in a single equation, both from the deterministic and the stochastic points of view. Generalized logistic differential equations are concatenated four times to deal with the four waves at once. This procedure extends the use of the generalized logistic differential equation to more than one wave, compared to [3, 22, 31, 46], and to more than two waves, compared to [9, 36, 49] from the work [27] on the bi-logistic model. The deterministic model gives a good fit for COVID-19 reported cases, especially for accumulated cases. Daily new reported infections present significant variability, perhaps due to highly variable factors, such as amount of tests available and performed, symptomatology, etc. Thus, stochasticity is incorporated, to extend the deterministic model and to obtain realizations that resemble the irregular dynamics of the reported new cases closer [38]. The Bayesian bootstrap [34], together with a trick to deal with residuals, is employed to set probability distributions for the input parameters and the errors. The potential of the Bayesian bootstrap for mathematical models with uncertainties does not seem to have been investigated. Our approach resembles the previous use of the frequentist bootstrap [12] for mathematical models with deterministic parameters and random errors [10]. Apart from fitting the available data, predictions have been performed by using two train sets, for a quite large forecast period of 15 days. Phenomenological models like the one proposed in this paper may be useful to generate reasonable forecasts in near time of the incidence of starting and advanced epidemic outbreaks, without entering in accounting for possible underlying mechanisms of the studied phenomenon (such as temporal or spatial dependencies), which would clearly be the base for further analysis. This idea was also discussed in [7, 32] for a single generalized logistic model.

We justify the use of the Bayesian bootstrap by commenting the unfeasibility of other techniques:

- *Maximum entropy principle.* This principle has been used in the literature to infer consistent probability distributions for input parameters [11]. The probability density function of the parameter is taken by maximizing the Shannon entropy functional, often restricted to a certain support, to a mean value equal to the deterministic estimate, and rarely to a variance if available. In the case studied in the present paper, reliable supports of the parameters are unknown a priori.
- *General Bayesian model.* Not restricted to the Bayesian bootstrap, Markov Chain Monte Carlo algorithms may solve the problem for any set of prior distributions [23, 38]. However, the large amount of input random parameters in our case study prevented us from using this option.
- *Itô-type stochastic differential equations.* One could naturally ask about the incorporation of a white noise (formal derivative of Brownian motion) into the generalized logistic differential equation (2.1) [2, 25]. However, accumulated infections give rise to increasing sample paths, which would contradict the everywhere non-differentiability of Itô processes.

Some limitations of the present work, which define potential avenues for future research, are the following: (a) Ideally, the probabilistic interval for the model output should be narrower. It would be of interest to investigate alternative deterministic models (the averaged, smooth curve) or Bayesian approaches. (b) We needed to consider the logarithms of the daily new infections to have an adequate resampling of residuals for which increasing sample paths (accumulated infections) or negative values (daily new infections) were not a problem. It would be interesting to directly define a model error for the accumulated infections that takes into account the correlation between successive days. This correlation is due to the increasing character of the

curve. (c) Spatial, behavioral, or environmental effects have been neglected. These conditions may provide a more faithful fit of the COVID-19 data, albeit at the expense of higher complexity. Nonetheless, when uncertainty is present on the phenomenon itself and the data, sometimes it may be preferable to simply consider the model as uncertain, rather than augmenting its complexity.

## 5. Conclusion

We have shown in the paper that the sum of four generalized logistic growth curves allows for a proper fit of the accumulated reported infections along the four waves of the COVID-19 epidemic in Castilla-Leon (Spain). Daily new reported infections are described by consecutive differences. The input parameters are pointwise calibrated by least-squares fitting. However, this calibration lacks of probabilistic interpretations.

Taking into account the significant variability in the daily reported data, with noisy features, stochasticity is incorporated into the model by treating the input parameters and the model errors as random variables. This conception of uncertainty matches with the Bayesian formalism of Statistics. The Bayesian bootstrap is an adequate approach for inverse uncertainty quantification and infers the probability distributions of the parameters. The model response is stochastic and includes realizations that permit a more reliable fit of the daily new COVID-19 reported cases, compared to the smooth deterministic counterpart.

## Funding

This paper has been partially funded by projects PID2019-107392RB-I00 from Spanish Ministry of Science, AICO/2019/198 from Generalitat Valenciana, and PID2020-115270GB-I00 from Spanish Agencia Estatal de Investigación (MCIN/AEI/10.13039/501100011033).

## Conflict of Interest Statement

The authors declare that there is no conflict of interests regarding the publication of this article.

## Data Availability Statement

The implementations and computations in Mathematica® are included as supplementary material, where the data are available (variable *vtotal*). The cases have been retrieved from the open data portal of Castilla-Leon.

## References

[1] Acedo L, Moraño JA, Santonja FJ, Villanueva RJ (2016) A deterministic model for highly contagious diseases: The case of varicella. Physica A 450:278–286.

[2] Allen E (2007) Modeling With Itô Stochastic Differential Equations. Springer Science & Business Media, Dordrecht, Netherlands.

[3] Aviv-Sharon E, Aharoni A (2020) Generalized logistic growth modeling of the COVID-19 pandemic in Asia. Infectious Disease Modelling 5:502-509.

[4] Berihuete A, Sánchez-Sánchez M, Suárez-Llorens A (2021) A Bayesian model of COVID-19 cases based on the Gompertz curve. Mathematics 9(3):228.

[5] Birch CP (1999) A new generalized logistic sigmoid growth equation compared with the Richards growth equation. Annals of Botany 83(6):713–723.

[6] Chitnis N, Schpira A, Smith D, Hay SI, Smith T, Steketee R (2010) Mathematical modelling to support malaria control and elimination. Roll Back Malar Prog Impact Ser (World Health Organization, Progress & impact series) 5:1–48.

[7] Chowell G, Hincapie-Palacio D, Ospina JF, Pell B, Tariq A, Dahal S, Moghadas SM, Smirnova A, Simonsen L, Viboud C (2016) Using phenomenological models to characterize transmissibility and forecast patterns and final burden of Zika epidemics. PLoS Currents 8.
[8] Chowell G, Simonsen L, Viboud C, Kuang Y (2014) Is west Africa approaching a catastrophic phase or is the Ebola epidemic slowing down? Different models yield different answers for Liberia. PLoS Curr 2014(6).
[9] da Silva EV, da Silva Melo J, Leite MA (2020) Modelo bi-logístico aplicado aos primeiros 1015 casos de COVID-19 em indígenas do Estado do Amapá e norte do Pará. Science and Knowledge in Focus 3(2):77–88.
[10] Dogan G (2007) Bootstrapping for confidence interval estimation and hypothesis for parameters of system dynamics models. System Dynamics Review 23:415–436.
[11] Dorini FA, Sampaio R (2012) Some results on the random wear coefficient of the Archard model. Journal of Applied Mechanics 79(5):051008–051014.
[12] Efron B (1979) Bootstrap methods: Another look at the jackknife. The Annals of Statistics 7(1):1–26.
[13] Fenner T, Levene M, Loizou G (2013) A bi-logistic growth model for conference registration with an early bird deadline. Open Physics 11(7):904–909.
[14] Hsieh YH (2009) Richards model: a simple procedure for real-time prediction of outbreak severity. In: Ma Z, Zhou Y, Wu J (ed) Modeling and Dynamics of Infectious Diseases, World Scientific, pp 216–236.
[15] Hsieh YH (2010) Pandemic influenza A (H1N1) during winter influenza season in the southern hemisphere. Influenza and Other Respiratory Viruses 4(4):187–197.
[16] Hsieh YH, Lee JY, Chang HL (2004) SARS epidemiology modeling. Emerging infectious diseases 10(6):1165.
[17] Hsieh YH, Ma S (2009) Intervention measures, turning point, and reproduction number for dengue, Singapore, 2005. The American journal of tropical medicine and hygiene 80(1):66–71.
[18] Kingsland S (1982) The Refractory Model: The Logistic Curve and the History of Population Ecology. The Quarterly Review of Biology 57:29–52.
[19] Kraemer MU, Yang CH, Gutierrez B, Wu CH, Klein B, Pigott DM, Brownstein JS (2020) The effect of human mobility and control measures on the COVID–19 epidemic in China. Science 368(6490):493–497.
[20] Lavrova AI, Postnikov EB, Manicheva OA, Vishnevsky BI (2017) Bi-logistic model for disease dynamics caused by *Mycobacterium tuberculosis* in Russia. Royal Society Open Science 4(9):171033.
[21] Le Maître OP, Knio OM (2010) Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics. Springer Science & Business Media, Netherlands.
[22] Lee SY, Lei B, Mallick B (2020) Estimation of COVID-19 spread curves integrating global data and borrowing information. PloS One 15(7):e0236860.
[23] Lesaffre E, Lawson AB (2012) Bayesian Biostatistics. Wiley, Statistics in Practice, New York.
[24] Malthus TR (1999) An Essay on the Principal of Population. Oxford World's Classics Paperbacks Oxford University Press, Oxford.
[25] Mao X (2007) Stochastic Differential Equations and Applications. Elsevier.
[26] Marusic M, Bajzer Z, Vuk-Pavlovic S, Freyer JP (1994) Tumor growth in vivo and as multicellular spheroids compared by mathematical models. Bulletin of Mathematical Biology 56:617–631.
[27] Meyer PS (1994) Bi-logistic growth. Technological Forecasting and Social Change 47(1):89–102.
[28] Moein S, Nickaeen N, Roointan A, Borhani N, Heidary Z, Javanmard SH, Ghaisari J, Gheisari Y (2021) Inefficiency of SIR models in forecasting COVID–19 epidemic: a case study of Isfahan. Scientific Reports 11(1):1–9.
[29] Muñoz-Fernández GA, Seoane JM, Seoane-Sepúlveda JB (2021) A SIR–type model describing the successive waves of COVID–19. Chaos Soliton Fract 144:110682.
[30] Murray JD (2002) Mathematical Biology I. An Introduction. Springer-Verlag, New York.
[31] Pelinovsky E, Kurkin A, Kurkina O, Kokoulina M, Epifanova A (2020) Logistic equation and COVID-19. Chaos Soliton Fract 140:110241.
[32] Pell B, Kuang Y, Viboud C, Chowell G (2018) Using phenomenological models for forecasting the 2015 Ebola challenge. Epidemics 22:62–70.
[33] Remuzzi A, Remuzzi G (2020) COVID-19 and Italy: What next?. Lancet 395(10231):1225–1228.
[34] Rubin DB (1981) The Bayesian bootstrap. The Annals of Statistics 9(1):130–134.
[35] Sachs RK, Hlatky LR, Hahnfeldt P (2001) Simple ODE models of tumor growth and anti-angiogenic or radiation treatment. Mathematical and Computer Modelling 33(12–13):1297–1305.
[36] Salpasaranis K, Stylianakis V (2020) Forecasting models of the coronavirus (COVID-19) cumulative confirmed cases using a hybrid genetic programming method. European Journal of Engineering and Technology Research 5(12):52–60.

[37] Shehu V (2015) Simple Logistic and Bi-Logistic Growth used as forecasting models of greenhouse areas in Albanian agriculture. Journal of Multidisciplinary Engineering Science and Technology 2(9):2648–2653.
[38] Smith RC (2013) Uncertainty Quantification: Theory, Implementation, and Applications. SIAM.
[39] Spratt JS, Meyer JS, Spratt JA (1996) Rates of growth of human neoplasms: Part II. Journal of Surgical Oncology 61(1):68–83.
[40] Stanescu D, Chen-Charpentier BM, Jensen BJ, Colberg PJS (2009) Random coefficient differential models of growth of anaerobic photosynthetic bacteria. Electron T Numer Ana 34:44–58.
[41] Turchin P (2001) Does population ecology have general laws?. Oikos 94(1):17–26.
[42] Verhulst PF (1838) Notice sur la loi que la population suit dans son accroissement. Corr Math et Phys 10:113–121.
[43] Wang P, Zheng X, Li J, Zhu B (2020) Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. Chaos Soliton Fract 139:110058.
[44] World Health Organization (WHO) (2021) Coronavirus disease (COVID–19) pandemic. Available at `https://www.who.int/emergencies/diseases/novel-coronavirus-2019` (accessed 22nd July 2021).
[45] Wu YC, Chen CS, Chan YJ (2020) The outbreak of COVID-19: An overview. J Chin Med Assoc 83(3):217.
[46] Wu K, Darcet D, Wang Q, Sornette D (2020) Generalized logistic growth modeling of the COVID-19 outbreak: comparing the dynamics in the 29 provinces in China and in the rest of the world. Nonlinear dynamics 101(3):1561–1581.
[47] Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YL (2020) A new coronavirus associated with human respiratory disease in China. Nature 579(265):265–269.
[48] Xiu D (2010) Numerical Methods for Stochastic Computations: A Spectral Method Approach. Cambridge Texts in Applied Mathematics, Princeton University Press, New York.
[49] Zhang L, Tao Y, Zhuang G, Fairley CK (2020) Characteristics analysis and implications on the COVID-19 reopening of Victoria, Australia. The Innovation 1(3):100049.