Physica D 413 (2020) 132674

Contents lists available at ScienceDirect

Physica D

journal homepage: www.elsevier.com/locate/physd

Inversion of a SIR-based model: A critical analysis about the application to COVID-19 epidemic

Alessandro Comunian [a], Romina Gaburro [b], Mauro Giudici [a,*]

[a] Università degli Studi di Milano, Dipartimento di Scienze della Terra "A. Desio", via Cicognara 7, 20129 Milano, Italy
[b] University of Limerick, Department of Mathematics and Statistics, Health Research Institute (HRI), Limerick, Ireland

ARTICLE INFO

Article history:
Received 29 May 2020
Received in revised form 22 July 2020
Accepted 31 July 2020
Available online 12 August 2020
Communicated by V.M. Perez-Garcia

Keywords:
Inverse problems
SIR models
COVID-19

ABSTRACT

Calibration of a SIR (Susceptibles–Infected–Recovered) model with official international data for the COVID-19 pandemics provides a good example of the difficulties inherent in the solution of inverse problems. Inverse modeling is set up in a framework of discrete inverse problems, which explicitly considers the role and the relevance of data. Together with a physical vision of the model, the present work addresses numerically the issue of parameters calibration in SIR models, it discusses the uncertainties in the data provided by international authorities, how they influence the reliability of calibrated model parameters and, ultimately, of model predictions.

1. Introduction

Epidemic modeling is usually performed with compartmental models, often called SIR (Susceptibles–Infected–Recovered) models, which are claimed to go back to the work by Ronald Ross and Hilda P. Hudson more than one century ago [1,2] and, ten years later, to the work of Anderson Gray McKendrick and William Ogilvy Kermack [3,4]. This class of models shares several characteristics with models of population dynamics and with conceptual lumped models in hydrology. These models simulate the temporal evolution of some compartments of the population, which is normally subdivided among *Susceptibles* (i.e., those individuals who have not yet been affected by the virus and who could be subject to infection), *Infected* (i.e., those individuals who have been infected by the virus) and *Recovered* (i.e., those individuals who have recovered, after having been infected). For this reason, these models are usually referred to as SIR models. They are based on phenomenological laws to describe the transfer of individuals from one class to another.

These models have found wide application both in life sciences, mostly in epidemiology, and in the field of economic, political and social sciences, e.g., in the context of addressing the costs of policies designed to block epidemics and the diffusion of viruses and in the realm of optimal control to assess the political measures which guarantee the best equilibrium between reduction of the epidemic spread and harmful secondary socio-economical impacts [5–7]. Several extremely interesting papers have been devoted to the mathematical properties of SIR models, often by applying the theory of dynamical systems; see [8–17] among many others. However, previous works on SIR models have, to the best of the authors' knowledge, seldom addressed the calibration of SIR models with real data, i.e., the issue of a proper fitting of epidemiological data with model outcomes. Some examples refer to applications to dengue transmission [18], H5N1 avian influenza [19], HIV epidemic [20] and Severe Acute Respiratory Syndrome (SARS) [21].

In this paper, we address the problem of calibrating the epidemiological parameters of a SIR model describing the evolution in time of the current COVID-19 pandemic. This is achieved by solving numerically the underlying inverse problem via the minimization of an objective function that measures the discrepancy between the simulated solutions to the discretized SIR model and the official data on COVID-19.

Model calibration is a common problem in geophysical and environmental modeling. The present paper follows the general framework introduced in [22] to handle discrete inverse problems for model calibration and analyzes the role of data following the discussion in [23]. The continuous SIR model considered here is discretized via a forward-time finite-differences scheme which is implemented in a specifically designed code, developed using the Python programming language, to provide at each discrete time $n \in \mathbb{Z}$ a vector state of the discretized SIR system, which is, in turn, matched against real data in order to calibrate the parameters of the system via a minimization problem (the inverse problem). The results presented in this paper consider the application of the model to a given nation, Italy in this instance, avoiding any subdivision in provinces, regions or states. The great

* Corresponding author.
E-mail addresses: alessandro.comunian@unimi.it (A. Comunian), romina.gaburro@ul.ie (R. Gaburro), mauro.giudici@unimi.it (M. Giudici).

https://doi.org/10.1016/j.physd.2020.132674
0167-2789/

amount of data collected during the COVID-19 pandemic due to the diffusion of the SARS-CoV-2 virus (also called "coronavirus") provides an exceptional basis to test calibration of SIR models via the solution of an inverse problem. It is well known that inverse problems are ill-posed, due to the lack of uniqueness and stability of these problems. Non-uniqueness will be addressed in this paper by the application of different algorithms for the minimization of the misfit between reference observed quantities and model predictions. The other relevant topic for ill-posedness is the lack of stability, i.e., the lack of continuous dependence of the parameters to be identified on the data, so that small errors in the data can lead to large discrepancies in the parameters one is trying to identify via the inverse problem. We do not provide a full review here on these topics, but we mention [24] for a general-purpose description, [25,26] for a deep discussion on the instability issue in the context of the so-called inverse conductivity problem and [27] for recent results about optical tomography.

The main objective of this paper is to fix some concepts about SIR models and their calibration and to discuss the relevance of data for reliability of model outcomes in the context of inverse problems and their application to the specific study of the spread of COVID-19 [23].

The paper is designed to advance the current knowledge about the functioning, potentialities and limitations of epidemic models. It also highlights certain similarities among geophysical, environmental and epidemic modeling, therefore providing further insights in epidemic model calibration. On the other hand, this work does not aim to provide forecasts of the pandemic evolution at this stage. It is in the authors' opinion that the quality of the data that are currently available does not allow to perform reliable forecast and model outcomes should be used with high prudence. It will be material of future work to further develop and refine the SIR model presented in this work and to address the issue of providing forecasts of the epidemics, when the data will be better understood.

For instance, the correct number of infected people "remains unknown because asymptomatic cases or patients with very mild symptoms might not be tested and will not be identified", as recognized, e.g., by [28]. In an interview published on March 23rd, 2020, by the Italian newspaper "La Repubblica", Angelo Borrelli, Head of Dipartimento della Protezione civile (national civil protection department) stated that a ratio of one certified case out of every 10 total cases is credible. Furthermore, different criteria have been adopted by different countries and institutions to define the various categories of infected, recovered and deceased people by or with COVID-19. This fact has been widely recognized as a cause of uncertainty in the collected data. Finally, censorship on COVID-19 pandemics is reported by journalists and organizations in some of the countries affected by the pandemic.

The paper is organized as follows. Section 2 contains the description of the SIR model in both the continuous and the discrete case (Section 2.1) together with a precise formulation of the inverse problem addressed in this paper in the discrete setting (Section 2.2). In particular, inverse modeling, i.e., model calibration, is set up and discussed computationally within the framework proposed by [22]. The results obtained by applying our SIR model to the COVID-19 pandemic are shown in Section 3. Section 4 is devoted to a discussion about the main assumptions on which the SIR model is based and some possible future developments. Section 5 summarizes the most relevant results of this work.

2. Methods and materials

2.1. The continuous and the discrete models

We start by defining the objects involved in the continuous SIR model considered in this paper.

Definition 2.1. We denote by $S(t)$, $I(t)$, $R(t)$ and $D(t)$ the number of *susceptible*, *infected*, *recovered* and *deceased individuals* of the population under study at time t, respectively, for t varying in some interval $\mathcal{I} \subset \mathbb{R}$. Here D includes only those individuals who died while being infected, whereas the total population, at time t, is given by $P(t) = S(t) + I(t) + R(t)$.

Definition 2.2. We denote by β and δ the *birth* and *death rates*, respectively, under normal conditions, i.e., without considering deaths caused by the epidemic. We also denote by γ, ρ and ϕ the *infection*, *recovery* and *fatality rates*, respectively. The dimension of these coefficients is [time^{-1}].

Notice that ϕ accounts for the deaths related to the pandemic, i.e., it represents the increase in the death rate due to the pandemic. The normal death rate is considered through δ.

Note that β and δ in Definition 2.2 are rarely considered in epidemic modeling, as the time variation of P due to the normal evolution of the population is either negligible or smoother than its variation due to the presence of an epidemic. This is due to the fact that typical values of β and δ are smaller than the ones of γ, ρ and ϕ by one or more orders of magnitude, as shown in Section 3.2. We keep birth and death rates in the model, in order to facilitate a thorough discussion of the assumptions behind this model, which is given in Section 4. We make the following assumptions.

Assumption 2.1. The coefficients β, δ, γ, ρ and ϕ are assumed to be constant.

Assumption 2.2. The number of contacts of each infected person per unit time does not vary among the infected population and it is assumed to be constant in time. Moreover the fraction of such contacts who are susceptible to the infection is given by S/P, whereas $(I + R)/P$ is the fraction of those persons who cannot be infected, as it is also assumed that recovered people are immunized.

The following equations, based on the seminal papers [1–4], are used to describe the time evolution of S, I, D and R:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \beta S - \gamma \frac{IS}{P} - \delta S, \tag{1}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta I + \gamma \frac{IS}{P} - \rho I - \phi I - \delta I, \tag{2}$$

$$\frac{\mathrm{d}D}{\mathrm{d}t} = \phi I, \tag{3}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \beta R + \rho I - \delta R \tag{4}$$

together with the initial conditions $S(t_{\mathrm{ini}}) = P_{\mathrm{ini}} - 1$, $I(t_{\mathrm{ini}}) = 1$, $R(t_{\mathrm{ini}}) = 0$ and $D(t_{\mathrm{ini}}) = 0$, where $t_{\mathrm{ini}} \in \mathcal{I} \subset \mathbb{R}$ is the time at which the first individual is infected and P_{ini} is the population at t_{ini}. Notice that from Eqs. (1) to (4) one can easily deduce

$$\frac{\mathrm{d}P}{\mathrm{d}t} = \beta P - \delta P - \phi I \tag{5}$$

and if we couple (5) with (2)

$$\begin{cases} \dfrac{\mathrm{d}P}{\mathrm{d}t} = (\beta-\delta)P - \phi I, & \text{in } \mathcal{I}, \\ \dfrac{\mathrm{d}I}{\mathrm{d}t} = -\alpha I + \gamma\dfrac{IS}{P}, & \text{in } \mathcal{I}, \\ P(t_{\text{ini}}) = P_{\text{ini}}, \\ I(t_{\text{ini}}) = 1, \end{cases} \tag{6}$$

where

$$\alpha = \phi + \rho - \beta + \delta. \tag{7}$$

We can approximate (6) to the simple system of autonomous linear ordinary differential equations

$$\begin{cases} \dfrac{\mathrm{d}P}{\mathrm{d}t} = (\beta-\delta)P, & \text{in } (t_{\text{ini}}, t_{\text{ini}}+h), \\ \dfrac{\mathrm{d}I}{\mathrm{d}t} = (\gamma-\alpha)I, & \text{in } (t_{\text{ini}}, t_{\text{ini}}+h), \\ P(t_{\text{ini}}) = P_{\text{ini}}, \\ I(t_{\text{ini}}) = 1, \end{cases} \tag{8}$$

for some small $h > 0$. This rough approximation is justified by thinking that, for h small enough, $I(t) \ll P_{\text{ini}} \simeq S(t)$ and therefore $IS/P \simeq I$ in (6).

The system

$$\begin{cases} \dfrac{\mathrm{d}P}{\mathrm{d}t} = (\beta-\delta)P, & \text{in } (t_{\text{ini}}, t_{\text{ini}}+h), \\ P(t_{\text{ini}}) = P_{\text{ini}}, \end{cases} \tag{9}$$

describes the population evolution taking into account demographic aspects only, i.e., in absence of the perturbation caused by epidemics and by assuming that the birth and death rates are constant, whereas

$$\begin{cases} \dfrac{\mathrm{d}I}{\mathrm{d}t} = (\gamma-\alpha)I, & \text{in } (t_{\text{ini}}, t_{\text{ini}}+h), \\ I(t_{\text{ini}}) = 1 \end{cases} \tag{10}$$

describes the time evolution of the number of infected cases during a short time after the beginning of the infection at time $t = t_{\text{ini}}$. The solution to (10), $I(t) \simeq \exp[(\gamma-\alpha)\cdot(t-t_{\text{ini}})]$ and, for h small enough, its linear approximation near t_{ini}, $I(t) \simeq 1 + (\gamma-\alpha)\cdot(t-t_{\text{ini}})$, give a first rough explanation about why, during the first phases of the epidemics, i.e., for $t \simeq t_{\text{ini}}$, the number of infected individuals, $I(t)$, seems to grow linearly. This fact motivates the difficulties in the design of an efficient early warning system. In fact, once $I(t)$ increases to a significant level to be detected, the exponential growth had already kicked in and the containment measures can be effective only if quite drastic.

The discrete model is a simple forward-time finite-differences discretization of Eqs. (1) to (4). For $n \in \mathbb{Z}$, we denote the discrete time steps, at a uniform spacing Δt, by $t_n = n\Delta t$. The following definition is useful for the discrete model.

Definition 2.3. We denote by S_n, I_n, R_n and D_n the number of *susceptible*, *infected*, *recovered* and *deceased individuals* of the population under study at time t_n, respectively, for $n = n_{\text{ini}}, \dots, n_{\text{ini}} + N^{(\text{mod})} - 1$, where n_{ini} is such that $t_{\text{ini}} = n_{\text{ini}}\Delta t$ and $N^{(\text{mod})}$ is the number of modeled time steps. The total population at time t_n is given by $P_n = S_n + I_n + R_n$.

Then the resulting algebraic iterative equations are of the form

$$\begin{cases} S_{n+1} &= \left[1 + \left(\beta - \gamma\dfrac{I_n}{P_n} - \delta\right)\Delta t\right] S_n, \\ I_{n+1} &= \left[1 + \left(\beta + \gamma\dfrac{S_n}{P_n} - \rho - \phi - \delta\right)\Delta t\right] I_n, \\ D_{n+1} &= D_n + \phi I_n \Delta t, \\ R_{n+1} &= [1 + (\beta-\delta)\Delta t]\, R_n + \rho I_n \Delta t, \end{cases} \tag{11}$$

for $n = n_{\text{ini}}, \dots, n_{\text{ini}} + N^{(\text{mod})} - 1$, with initial conditions

$$S_{n_{\text{ini}}} = P_{\text{ini}} - 1, \qquad I_{n_{\text{ini}}} = 1, \qquad D_{n_{\text{ini}}} = R_{n_{\text{ini}}} = 0 \tag{12}$$

and the discrete counterpart of (5) is

$$P_{n+1} = [1 + (\beta-\delta)\Delta t]\, P_n - \phi I_n \Delta t. \tag{13}$$

Here the time spacing $\Delta t = 1$ day, in agreement with the sampling of the available data set on COVID-19 pandemic (see Section 2.3). Eqs. (11) are implemented in a specifically designed code, developed using the Python programming language.

The choice $n \in \mathbb{Z}$ allows to simplify the notation adopted in the formulation of the inverse problem in Section 2.2. It is important to notice that $n = 0$, i.e., $t_0 = 0$, represents the first day for which epidemic data are available and in general it does not coincide with $n = n_{\text{ini}}$, which corresponds to t_{ini}, the day when the first person was infected in a given nation, according to our model. We will call $t_0 = 0$ ($n = 0$) and t_{ini} ($n = n_{\text{ini}}$) the *monitoring initial time* and the *modeling initial time*, respectively. Accordingly, we will also call $P_{\text{ini}} = P(t_{\text{ini}})$ the *model initial population*.

2.2. The inverse problem: model calibration

As stated in the introduction, the inverse problem addressed here is defined in the discrete setting by making use of the conceptual framework and the notation of [22]. The numerical task in treating the inverse problem consists in solving iteratively (11) and matching such solutions with the data collected within a certain time-frame $[t_{\min}, t_{\max})$. Such (discrete) time-varying vector-solutions s_n are collected in an array $\mathbf{s}$, called the *state of the system*

$$\begin{aligned} \mathbf{s} = \Big\{ & s_n = (s_n^{(1)}, s_n^{(2)}, s_n^{(3)}, s_n^{(4)}) \in \mathbb{R}^4 \mid \\ & s_n^{(1)} = S_n,\ s_n^{(2)} = I_n,\ s_n^{(3)} = R_n,\ s_n^{(4)} = D_n, \\ & n = n_{\text{ini}}, \dots, n_{\text{ini}} + N^{(\text{mod})} - 1 \Big\}, \end{aligned} \tag{14}$$

where $N^{(\text{mod})}$ and $n = n_{\text{ini}}$ have been introduced in Definition 2.3. $\mathbf{s}$ is the model outcome used to forecast the number of infected, recovered and dead individuals. To this end, we also introduce the *model forecast*, an array $\mathbf{y}$ defined by

$$\begin{aligned} \mathbf{y} = \Big\{ & y_n = (y_n^{(1)}, y_n^{(2)}, y_n^{(3)}) \in \mathbb{R}^3 \mid y_n^{(1)} = I_n,\ y_n^{(2)} = R_n,\ y_n^{(3)} = D_n, \\ & n = n_{\min}, \dots, n_{\max} - 1 \Big\}, \end{aligned} \tag{15}$$

for some $n_{\min}$, $n_{\max}$, with $n_{\text{ini}} \le n_{\min} < n_{\max} \le n_{\text{ini}} + N^{(\text{mod})}$. The available data are collected in an array $\mathbf{d}$. In the specific case considered here, the subset of *data* denoted by $\mathbf{d}' \subset \mathbf{d}$ includes the cumulative number of the confirmed infected cases, together with the number of the recovered and dead persons, released by

health official organizations

$$\mathbf{d}' = \left\{ d'_n = (d'^{(1)}_n, d'^{(2)}_n, d'^{(3)}_n) \in \mathbb{R}^3 \mid d'^{(1)}_n = C^{(\mathrm{ref})}_n,\ d'^{(2)}_n = R^{(\mathrm{ref})}_n,\ d'^{(3)}_n = D^{(\mathrm{ref})}_n,\ n = 0, \ldots, N^{(\mathrm{ref})} - 1 \right\}, \tag{16}$$

where $N^{(\mathrm{ref})}$ is the number of data time steps, i.e., the number of time steps for which data are available. Notice that $C^{(\mathrm{ref})}_n$ is the cumulative number of confirmed infected cases, so that the number of infected cases at a given time n is given by

$$I^{(\mathrm{ref})}_n = C^{(\mathrm{ref})}_n - R^{(\mathrm{ref})}_n - D^{(\mathrm{ref})}_n. \tag{17}$$

$\mathbf{d}$ can include also other data, e.g., demographic data used to infer the values of some model parameters (β and δ). $n = 0$ represents the so-called *monitoring initial time* introduced in 2.1, which corresponds to the first day for which epidemic data $\mathbf{d}'$ are available; recall that, in general, it does not coincide with the day $n = n_{\mathrm{ini}}$ when the first person was infected in a given country.

Model calibration requires that the model forecast be close to a *calibration target*, an array $\mathbf{t}$ that collects the values which should be attained by the model forecast, if the model were physically "correct" and the model parameters were "optimal". In this specific case $\mathbf{t}$ is defined by

$$\mathbf{t} = \left\{ t_n = (t^{(1)}_n, t^{(2)}_n, t^{(3)}_n) \in \mathbb{R}^3 \mid t^{(1)}_n = I^{(\mathrm{ref})}_n,\ t^{(2)}_n = R^{(\mathrm{ref})}_n,\ t^{(3)}_n = D^{(\mathrm{ref})}_n,\ n = n_{\min} \ldots, n_{\max} - 1 \right\}, \tag{18}$$

where $I^{(\mathrm{ref})}_n$ is given by (17) and $n_{\min}$, $n_{\max}$ are such that $0 \le n_{\min} < n_{\max} \le N^{(\mathrm{ref})}$. The *model parameters* are placed in an array $\mathbf{p}$:

$$\mathbf{p} = (\beta, \delta, \Delta t, \rho, \phi, \gamma, n_{\mathrm{ini}}, P_{\mathrm{ini}}) \in \mathcal{P} \subset {\mathbb{R}_+}^6 \times \mathbb{Z} \times (\mathbb{N} \setminus \{0\}), \tag{19}$$

where $\mathbb{R}_+ = (0, +\infty)$ and we recall that $\Delta t = 1$ day and P_{ini} is the *model initial population* introduced in Section 2.1.

If we summarize the algebraic equations in the discrete model (11) together with the initial conditions (12) with

$$\mathbf{f}(\mathbf{p}, \mathbf{s}) = 0, \tag{20}$$

the *forward problem* can be stated as: *given* $\mathbf{p}$, *find the unique state* $\mathbf{s} = \mathbf{g}(\mathbf{p})$ that solves (20). In other words, given the parameters $\mathbf{p}$, the solution to the forward problem will give the state of the system, $\mathbf{s}$. In order to introduce the corresponding *inverse problem*, it is convenient to write $\mathbf{p}$ as

$$\mathbf{p} = \left(\mathbf{p}^{(\mathrm{fix})}, \mathbf{p}^{(\mathrm{cal})}\right), \tag{21}$$

where

$$\mathbf{p}^{(\mathrm{fix})} = (\beta, \delta, \Delta t), \quad \mathbf{p}^{(\mathrm{cal})} = (\rho, \phi, \gamma, n_{\mathrm{ini}}, P_{\mathrm{ini}}) \in \mathcal{P}^{(\mathrm{cal})}. \tag{22}$$

$\mathbf{p}^{(\mathrm{fix})}$ and $\mathbf{p}^{(\mathrm{cal})}$ include the model parameters whose values are fixed before the simulation and the model parameters whose values are obtained from the solution of the underlying inverse problem, which is yet to be stated, respectively. $\mathcal{P}^{(\mathrm{cal})}$ is the set of the admissible values for $\mathbf{p}^{(\mathrm{cal})}$, possibly defined by fixing lower and upper bounds for each parameter.

Remark 2.1. Some remarks on $\mathbf{p}$, $\mathbf{y}$ and $\mathbf{t}$ are in order.

1. The array of fixed parameters is a function of $\mathbf{d}$: $\mathbf{p}^{(\mathrm{fix})} = \mathbf{p}^{(\mathrm{fix})}(\mathbf{d})$;
2. The model forecast $\mathbf{y}$ is a function of $\mathbf{s}$, $\mathbf{p}$ and $\mathbf{d}$: $\mathbf{y} = \mathbf{y}(\mathbf{d}, \mathbf{s}, \mathbf{p})$;
3. $\mathbf{t}$ may depend on $\mathbf{d}$ and $\mathbf{p}^{(\mathrm{fix})}$, but must be independent of $\mathbf{p}^{(\mathrm{cal})}$: $\mathbf{t} = \mathbf{t}\left(\mathbf{d}, \mathbf{p}^{(\mathrm{fix})}\right)$.

The misfit between model predictions and the target values is computed by means of the following objective function:

$$\mathsf{O}_{\mathbf{y},\mathbf{t}}\left(\mathbf{p}^{(\mathrm{cal})}\right) = \sum_{i=1}^{3} \mathsf{O}^{(i)}_{\mathbf{y},\mathbf{t}}\left(\mathbf{p}^{(\mathrm{cal})}\right) \tag{23}$$

where $\mathsf{O}^{(i)}_{\mathbf{y},\mathbf{t}}\left(\mathbf{p}^{(\mathrm{cal})}\right)$ is defined by

$$\mathsf{O}^{(i)}_{\mathbf{y},\mathbf{t}}\left(\mathbf{p}^{(\mathrm{cal})}\right) = \left\{ \frac{1}{n_{\max} - n_{\min}} \sum_{n=n_{\min}}^{n_{\max}-1} \left[\frac{y^{(i)}_n - t^{(i)}_n}{\max\left\{\xi, t^{(i)}_n\right\}} \right]^2 \right\}^{1/2}, \tag{24}$$

for $i = 1, 2, 3$, where $\xi \ge 1$ is a threshold-and-weight parameter and $n_{\min}$, $n_{\max}$ are such that

$$\max\{0, n_{\mathrm{ini}}\} \le n_{\min} < n_{\max} \le \min\left\{N^{(\mathrm{mod})} + n_{\mathrm{ini}}, N^{(\mathrm{ref})}\right\}. \tag{25}$$

In other words, $\mathsf{O}_{\mathbf{y},\mathbf{t}}$ is the sum of three functions, each of which considers one of the three reference quantities, separately. The model calibration is then performed by solving the following *inverse problem*:
Given $\mathbf{p}^{(\mathrm{fix})}$ *and* $\mathbf{d}$, *given the solution* $\mathbf{s} = \mathbf{g}(\mathbf{p})$ *to* (20), *determine* $\mathbf{y}(\mathbf{d}, \mathbf{g}(\mathbf{p}), \mathbf{p})$, $\mathbf{t}$ *and find* $\mathbf{p}^{(\mathrm{cal})\star}$, *such that*

$$\mathbf{p}^{(\mathrm{cal})\star} = \arg \min_{\mathbf{p}^{(\mathrm{cal})} \in \mathcal{P}^{(\mathrm{cal})}} \mathsf{O}_{\mathbf{y},\mathbf{t}}\left(\mathbf{p}^{(\mathrm{cal})}\right),$$

i.e.

$$\mathsf{O}_{\mathbf{y},\mathbf{t}}\left(\mathbf{p}^{(\mathrm{cal})\star}\right) \le \mathsf{O}_{\mathbf{y},\mathbf{t}}\left(\mathbf{p}^{(\mathrm{cal})}\right), \qquad \forall \mathbf{p}^{(\mathrm{cal})} : \left(\mathbf{p}^{(\mathrm{fix})}, \mathbf{p}^{(\mathrm{cal})}\right) \in \mathcal{P}. \tag{26}$$

In other words, the objective of model calibration is to find the parameter values which best fit the reference data in a given time interval, $n_{\min} \le n < n_{\max}$.

The parameter ξ plays a double role. First of all, it is a threshold which keeps positive the denominator of the quantity appearing in (24), even in the particular case when $t^{(i)}_n = 0$. Furthermore, it controls some characteristics of the objective function. For $\xi = 1$ and if $t^{(i)}_n \ge 1$, $\mathsf{O}^{(i)}_{\mathbf{y},\mathbf{t}}$ is nothing but the root-mean-squared relative difference between target, $t^{(i)}_n$, and modeled values, $y^{(i)}_n$, of the ith component of t_n and y_n. When ξ assumes large values, it acts as a weight which dampens the errors corresponding to low and moderate values of $t^{(i)}_n$, namely $t^{(i)}_n \ll \xi$. As a consequence, for large values of ξ, relative errors corresponding to large values of $t^{(i)}_n$ will be dominant; from a practical point of view, this means that early-time behavior, when $I^{(\mathrm{ref})}_n$, $R^{(\mathrm{ref})}_n$ and $D^{(\mathrm{ref})}_n$ are small, is less relevant to the model fitting. In particular, if $\xi > \max\left\{t^{(i)}_n,\ n_{\min} \le n < n_{\max}\right\}$, then $\mathsf{O}^{(i)}_{\mathbf{y},\mathbf{t}}$ reduces to the standard root-mean-squared error. Notice that the latter condition implies that the value of ξ could be very large. A sensible upper limit for ξ, $\xi_{\max}$, could be for instance given by the upper bound of P_{ini}, which is identified with the total population of the country for which the simulation is performed, as shown in Section 3.2.

It is worth stressing that the explicit use of an interval $n_{\min} \le n < n_{\max}$ for the definition of $\mathbf{t}$, $\mathbf{y}$ and the objective function, although somehow cumbersome, is useful to assess changes in the physical parameters with time. Some examples of it will be shown in Section 3.2.

2.3. Data and computer implementation for COVID-19

The application of the model introduced in Section 2.1 and of the model calibration introduced in Section 2.2 can be attempted thanks to publicly available data on COVID-19 pandemic. The application will be performed at national level, i.e., the considered population will be the whole population of some countries. For

Algorithm 1 Pseudo-code for SIR model inversion by global optimization.

```
function MAIN
    read parameter file
    read data downloaded from JHU repository at github.com
    read demographic data
    set t
    set fixed = (p^(fix), t)
    set bounds = range for p^(cal)
    for run = 0 to Number_of_runs parallel do
        call single_run (bounds, fixed)
    end for
    read and assembles results from single runs
    save and draw results
end function
function SINGLE_RUN(bounds, fixed)
    call differential_evolution (Objective_function, bounds, fixed)
    save results of a single run
end function
function OBJECTIVE_FUNCTION(p, t)
    s = SIRmodel (p)
    objfun= O_{y(s),t}(p)
    return objfun
end function
function SIRMODEL(p)
    initialize s_{n_ini}
    for n = n_ini + 1 to n_ini + N^(mod) − 1 do
        compute s_n from s_{n−1} by (10)
    end for
    return s
end function
```

Fig. 1. Pseudo-code for SIR model inversion by global optimization.

each country, the array **d** is populated with data coming from two basic sources.

Data on COVID-19 pandemic are available from the GitHub repository managed by the Johns Hopkins University [29]. This is a collection of publicly available data from multiple sources, which are processed and delivered by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). Notice that the data are provided to the public strictly for educational and academic research purposes. The data are updated daily and the files used in this paper have been downloaded from the GitHub platform on May 2, 2020. The array **t** has been filled in by using those files.

Tailored codes have been developed under Python 3.7.6 to download data from the Github repository, perform the forward model introduced in Section 2.1 and calibrate the model by solving the corresponding inverse problem defined in Section 2.2. The inversion is based on the functions of the `optimize` module from SciPy v1.4.1 and profit from multi-core execution through the standard `multiprocessing` package. The pseudo-code for inversion is given in Fig. 1. The optimization algorithms that have been tested are based on constrained minimization, so that some bounds on $\mathbf{p}^{(\mathrm{cal})}$ should be prescribed. Best results have been obtained by global optimization with the function `differential_evolution` [30]. Since this function implements a stochastic algorithm, the pseudo-code of Fig. 1 shows that several runs of the algorithms are executed in an easily parallelized loop.

Fig. 2 shows the trend of confirmed cases, recovered and deceased people for a number of countries that have been considered the most relevant for the analysis of COVID-19 pandemic not only by the scientific community, but also by mass media. These plots show different trends of the three curves describing the evolution in time of the confirmed, recovered and dead cases among the various countries considered in this study.

Aside from China, for which the starting phase is not reported, since the virus diffusion started earlier than the first date for which data are available in the data set, the number of confirmed cases (plots A in Fig. 2) shows a first slow increase, followed by an exponential increase and possibly a slowdown after few weeks. It is highly questionable whether this behavior is related to the number of tests performed to confirm virus infection.

The most regular trends are clearly the ones describing the number of deceased people (plots C in Fig. 2), after about one week since the first reported case in each country considered in this study. Doubts about comprehensiveness of official data on deaths caused by coronavirus have been raised by several sources of information and by some commentators. Nevertheless, from a visual analysis of the data shown in Fig. 2, it seems safe to state that the number of deaths represents the time series with the smoothest variation and possibly the less affected by uncertainties in the data. We interpret this as an effect of the variability, in time and among countries, of the procedures adopted to assess the infected. For instance, the number of confirmed infected depends on the number of conducted tests and on the elapsed time from the collection of swabs to the completion of laboratory analysis and to the release of information to the public. In other words, estimates of infected and recovered are affected by errors different form the estimates of deceased persons. Moreover, Eqs. (3) and (4) show that D and R depend on some integrals of I with respect to time, so that the time variations of the numbers of deceased and recovered persons are expected to be more regular than the number of infected.

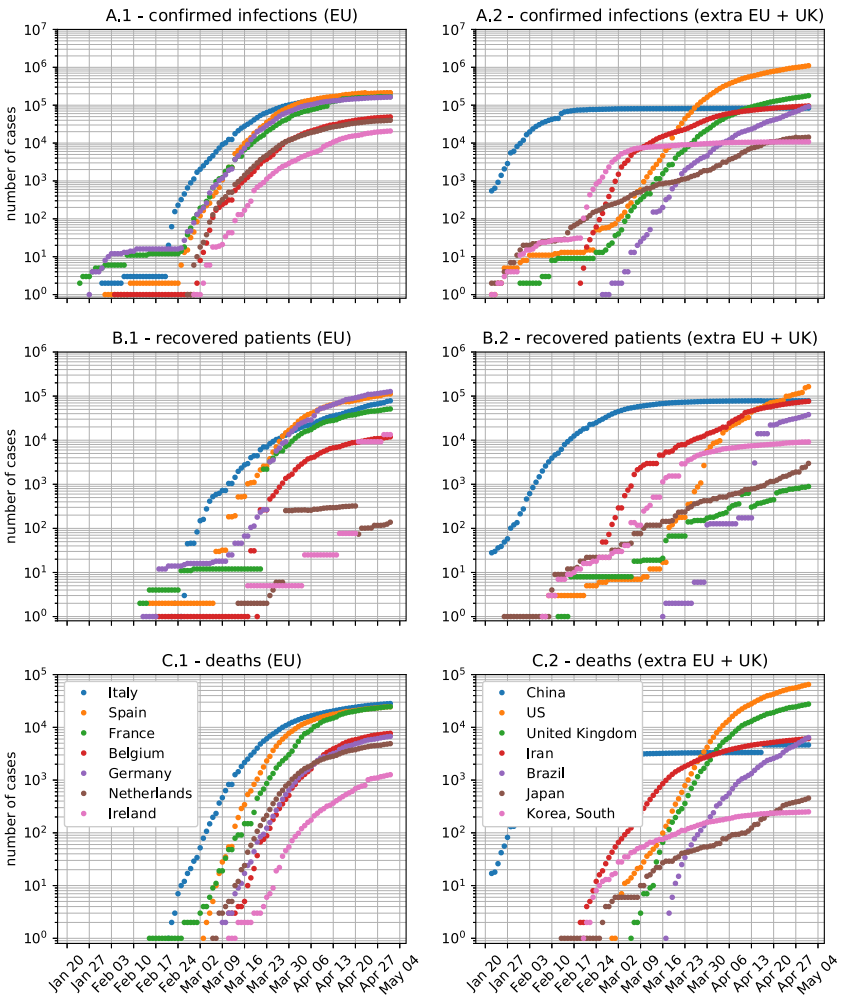

Fig. 2. Data about COVID-19 pandemic in selected EU (left column) and extra EU + United Kingdom (right column) countries: A – confirmed infections; B – recovered patients; C – deaths.

The second data source is the most updated version of the UN Demographic Yearbook [31]. Demographic data have been extracted from this volume. The values of population, birth and death rates of each country, for which the model has been tested, are included in **d**. They are used to fix the values of β and δ and to provide a first estimate of P_{ini}.

Notice that the daily sampling rate of epidemiologic data induces to choose $\Delta t = 1\,\text{day}$. Moreover, the coefficients β and δ are expressed on a daily basis, i.e., they are converted to the same measurement units as γ, ϕ and ρ, namely day^{-1} (see Definition 2.2).

3. Results

3.1. Model results

First of all, the behavior of the model is shown with test case 1, which includes four model runs for which all the model parameters, but ρ, are kept fixed. Exploratory tests were run with diverse parameters sets. However, the selection of the parameter sets used to obtain the results presented here was inspired by a preliminary calibration test performed on the data available for Italy. In this sense, these parameters, listed in Table 1, could be considered quite realistic. The results of the model for a one-year-long simulation period are shown in Fig. 3.

The general behavior shows an exponential increase in the number of infected persons (notice that the vertical axis is in logarithmic scale) followed by an exponential decrease but with a longer characteristic time. The number of deaths obviously decreases if ρ increases and in particular, we have four different situations for the four runs: (a) for the smallest value of ρ, the curve of susceptible persons dramatically decreases from some days before the peak of infections and reaches very small values after few weeks; (b) for a slightly higher value of ρ, the high number of deceased people causes a clear reduction of the population at the end of the simulation period and the whole population is recovered; (c) for the third tested value of ρ, $\rho = 0.056\,\text{day}^{-1}$, the number of susceptible and dead people reaches a stationary condition after about 8 months from the start of the epidemic and they share approximately the same value; (d) for the highest tested value of ρ, the number of susceptible people decreases with time, but remains high. Notice that, for the latter run, the reduction of the total population is limited, less than 10%, and after one year almost all the living population is recovered. The overall behavior of run (d) is coherent with the fact that a high value of ρ has the effect that a large fraction of infected people recovers in short time.

It is important to stress that this test case has the goal of showing how the model can predict different behavior and these results should not be considered as a forecast of the actual behavior of any real pandemic.

SIR models are often applied using the ratio of the number of individuals in each category with respect to the total population as state variables, namely S/P, I/P, R/P. Test case 1 showed that

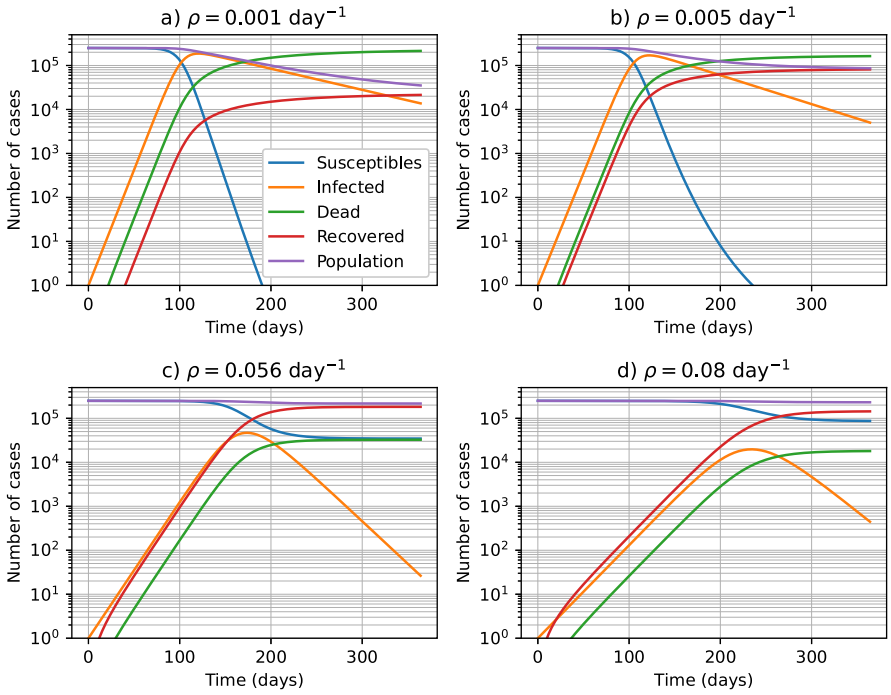

Fig. 3. Model results for test case 1.

Table 1
Parameter values for test case 1.

Parameter	run (a)	run (b)	run (c)	run (d)
β	$2.5 \cdot 10^{-5}$ day^{-1}	idem	idem	idem
δ	$3 \cdot 10^{-5}$ day^{-1}	idem	idem	idem
γ	0.14 day^{-1}	idem	idem	idem
ρ	0.001 day^{-1}	0.005 day^{-1}	0.056 day^{-1}	0.08 day^{-1}
ϕ	0.01 day^{-1}	idem	idem	idem
P_{ini}	250 000	idem	idem	idem

for three sets of model parameters, which differ only for the value of ρ, the total population has only a limited variation, so that approximating P to a constant value could appear reasonable. Nevertheless, the term used to compute the infection rate is directly proportional to both I and S and inversely proportional to P so that it introduces a non-linearity in the model. Therefore test case 2 is designed to assess the effect of P_{ini} on model results. To this goal, P_{ini} values span four orders of magnitude, from 10^6 to 10^9, whereas the other parameters are fixed at the values of run (a) of test case 1. The results are shown in Fig. 4 as functions of the normalized quantities versus time. The values of each function at the end of the simulation period are very similar. The main differences are in the evolving phase, for which the response of a small population appears to be more rapid than that of a large population. Roughly speaking, the curves corresponding to high populations show a delay with respect to the curve for the smallest population of about 15 days per an increase in P_{ini} by an order of magnitude. This remark, if confirmed by runs with more reliable parameter sets, could have fundamental consequences in the design of early warning systems. In fact, the time at which a given threshold of cases over the total population is exceeded increases with the population size.

3.2. Model calibration

Model calibration for the COVID-19 pandemic by solution of the inverse problem is a very challenging problem. This is not surprising at all. In fact, Fig. 3 shows some typical trends of the model time series, which are smoother than those observed from the reference data and drawn in Fig. 2. In other words, the choice of optimal fitting parameters for a simple SIR model, in order to properly simulate the observed trend, is very difficult. This is due both to the sources of uncertainties on the data and to the assumptions behind SIR models.

In particular, this paper is focused on the results obtained with data from Italy, but the same qualitative remarks hold also for the application to data from other countries.

The basic properties of the performed tests are listed in Table 2. The comparison between reference and fitted time series for test A, which is to be considered as the ideal one, because all the data are used and the standard settings are applied, is shown in Fig. 5. The discrepancy between reference and modeled values in log scale is greater for the initial phase of the epidemic; the model does not reproduce the sharp reduction of the rate of increase of deaths which appears in the reference time series around mid March.

Notice that for tests B, C and D three subsets of data are used, corresponding to three non overlapping time intervals, each of which is 33-days-long. In particular, the first day for which data are available is January 22, 2020 and the data series used in this paper ends on May 2, 2020. Therefore, the data set for test B ends on February 24, 2020, the data set for test C covers the interval from February 25 to March 29, 2020, and the data set for test D starts on March 30, 2020. The goal of these three tests is to examine possible differences in the optimal values of the parameters and in the behavior of the inversion procedure, for successive temporal phases of the epidemic. For test E, $\xi = 1$, so that each of the functions $\mathsf{O}^{(i)}_{\mathbf{y},\mathbf{t}}$ given by (24) is nothing but the root-mean-squared relative difference between reference and modeled values of I, R and D for $i = 1, 2, 3$, respectively. Therefore, test E has been designed in order to assess the dissimilarities in the inversion results due to the application of different objective functions. This is achieved by comparing tests A and E, which in ultimate essence are founded on absolute versus relative errors between model predictions and calibration targets, respectively. Test F is based on a subset of the data, in particular, for this

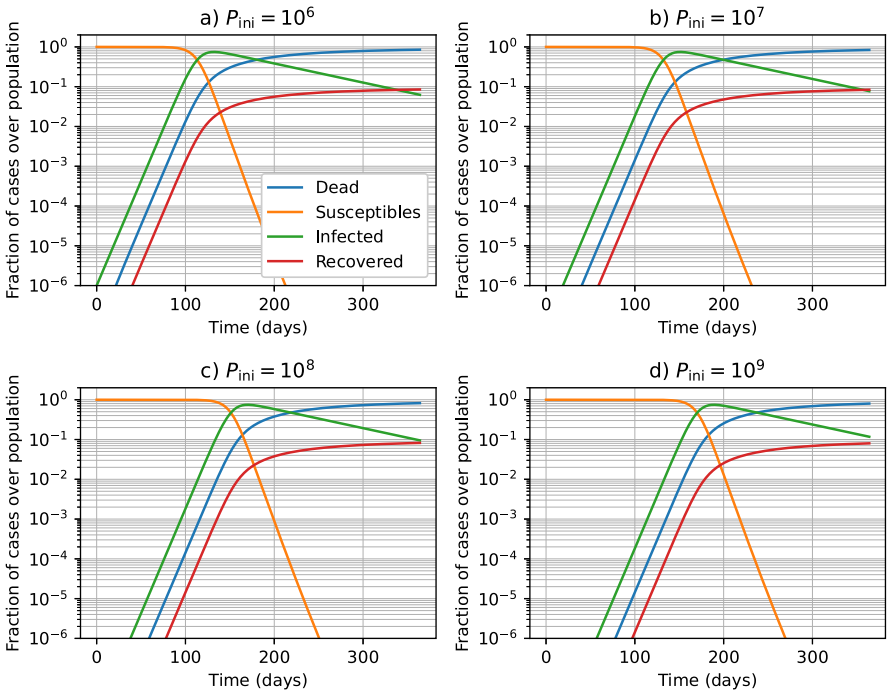

Fig. 4. Model results for test case 2.

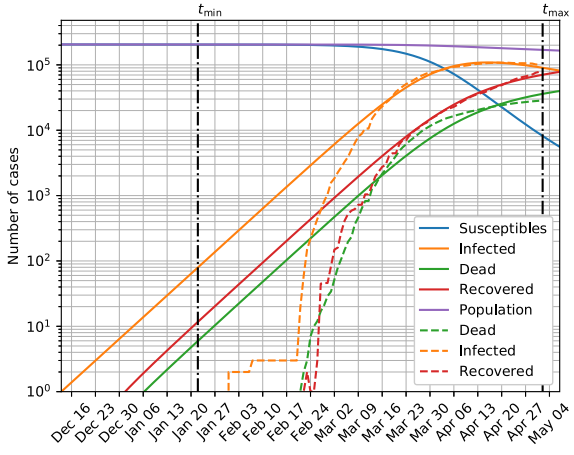

Fig. 5. Comparison of reference (dashed lines) and modeled values (continuous lines) for Italy with the parameters obtained by solution of the inverse problem for test A (see Table 2). The vertical dotted black lines delimit the time-frame of the data set used for model calibration, i.e., they correspond to t_{min} and t_{max}.

Table 2
Inversion tests with data referred to Italy. The standard approach uses the settings described in Section 2.2 with $\xi = 10^6$ and the data described in Section 2.3. Test G is based on the hypotheses that (i) the numbers of infected and recovered persons are ten times those reported by official fonts and (ii) the number of deaths is twice the official one.

Test	n_{min}	n_{max}	Notes
A	0	101	Standard
B	0	33	Standard
C	34	67	Standard
D	68	101	Standard
E	0	101	$\xi = 1$
F	0	101	$O^{(3)}_{\mathbf{y},\mathbf{t}}(\mathbf{p}^{(cal)})$
G	0	101	Modified data

Table 3
Lower and upper bounds for the parameters to be calibrated by inversion of data referred to Italy.

	γ	ρ	ϕ	t_{ini}	P_{ini}
Minimum	10^{-4} day^{-1}	10^{-5} day^{-1}	10^{-6} day^{-1}	-60	$2 \cdot 10^5$
Maximum	1 day^{-1}	0.1 day^{-1}	0.1 day^{-1}	20	10^8

test the number of dead patients only is fitted; the rationale behind this test is that $D^{(ref)}$ should be less uncertain than the other data of $\mathbf{d}'$. Finally, test G is an attempt to consider the hints raised by several authorities and scientists, suggesting that official numbers could be heavily underestimated. In this test, it is assumed that the number of infected and recovered persons be 10 times greater than those reported in official documents; analogously the number of deaths is assumed to be twice the official value. Notice that this does not mean that these estimates are more accurate than the official ones; test G is designed to be a first attempt of sensitivity analysis, by considering a data set very different from the reference one, which is used for test A.

Minimization of the objective function $O_{\mathbf{y},\mathbf{t}}$ was performed with different functions of the SciPy's module `optimize` which implements several methods to find a minimum, also by taking into account possible bounds on $\mathbf{p}^{(cal)}$, which limit the admissible values of each parameter and are used to build the $\mathcal{P}^{(cal)}$ set. The reader is referred to the on-line SciPy's documentation for details; here, it suffices to recall that tested methods include Nelder–Mead simplex algorithm [32] and conjugate-gradient based methods [33]. The bounds have been assigned on the basis of preliminary gross estimates from available data and they are listed in Table 3. These are used to construct $\mathcal{P}^{(cal)}$ for this work. Notice that the upper bound for P_{ini} is 10^8, hence here $\xi_{max} = 10^8$, i.e., $\xi \in [1, 10^8]$.

Several runs have been conducted with a routine for local minimization and the best results were obtained with the L-BFGS-B method, which is a variation of the Broyden–Fletcher–Goldfarb–Shannon (BFGS) algorithm [33] to reduce memory requirements and to handle simple constraints. The results of these runs are not presented here, for two basic motivations. That method is part of

Table 4
Results of model calibration by inversion of data referred to Italy for γ, ρ and ϕ. For the details about the performed tests see Table 2.

Test	γ (in day^{-1})	ρ (in day^{-1})	ϕ (in day^{-1})
A	0.1381 ± 0.0002	$(1.761 \pm 0.002) \times 10^{-2}$	$(8.24 \pm 0.02) \times 10^{-3}$
B	0.26 ± 0.05	$(3.27 \pm 0.6) \times 10^{-3}$	$(6 \pm 1) \times 10^{-3}$
C	0.185 ± 0.004	$(1.88 \pm 0.01) \times 10^{-2}$	$(1.52 \pm 0.01) \times 10^{-2}$
D	0.12229 ± 0.00001	$(1.694 \pm 0.001) \times 10^{-2}$	$(7.929 \pm 0.005) \times 10^{-3}$
E	0.17 ± 0.04	$(1.3 \pm 0.2) \times 10^{-2}$	$(1.2 \pm 0.3) \times 10^{-2}$
F	0.1384 ± 0.0002	$(1.556 \pm 0.002) \times 10^{-2}$	$(1.450 \pm 0.004) \times 10^{-3}$
G	0.28 ± 0.01	$(2.3 \pm 0.2) \times 10^{-2}$	$(1.2 \pm 0.9) \times 10^{-2}$

Table 5
Results of model calibration by inversion of data referred to Italy for t_{ini} and P_{ini}. For the details about the performed tests see Table 2.

Test	t_{ini}	P_{ini}
A	(-42.8 ± 0.2) day	$(2.111 \pm 0.003) \times 10^5$
B	(7.6 ± 3) day	$(5.44 \pm 0.86) \times 10^7$
C	(-18.5 ± 2) day	$(3.2 \pm 1.2) \times 10^5$
D	-59 day	$(2.1940 \pm 0.0084) \times 10^5$
E	(9.3 ± 2.8) day	$(3.5 \pm 1.1) \times 10^7$
F	(-56.4 ± 0.2) day	$(2.042 \pm 0.002) \times 10^6$
G	(-10.1 ± 5.8) day	$(1.0 \pm 0.6) \times 10^7$

a wide family of algorithms which move towards the minimum by means of gradient-based searches. However, it is not possible to compute analytically derivatives of $\mathsf{O}_{\mathbf{y},\mathbf{t}}$ with respect to t_{ini} and P_{ini}, which are integer, and not real, variables. Therefore, that family of methods cannot be applied in a rigorous way. Although the results of the performed runs, possibly fixing the value of t_{ini}, are not shown and discussed in detail here, it is nevertheless useful to mention them, because they confirm the existence of multiple local minima for $\mathsf{O}_{\mathbf{y},\mathbf{t}}$. As a consequence, these preliminary tests called for the application of a global minimization algorithm.

Global minimization by application of `differential_evolution` [30], even with the default settings, yielded good results, which are listed in Tables 4 and 5. The mean value of each parameter and the corresponding standard deviation have been estimated after 10 runs of this stochastic algorithm, for which the random initializing seed introduces variations among the returned results. When looking at Table 5, it is important to recall again that t_{ini} and P_{ini} are integer numbers, but in the table the averages and the relative standard errors are computed after 10 runs and this explains the float numbers notation.

Besides the optimal values of $\mathbf{p}^{(\text{cal})}$ listed in Tables 4 and 5, it is important and useful to consider also some properties of the inversion procedure for each test; they are listed in Table 6.

Table 4 shows that, apart from few tests, the optimal values of γ, ρ and ϕ are relatively similar, sharing the same order of magnitude and the relationship $\gamma > \rho > \phi$ among different tests. These inequalities are violated by the results of test B and possibly of test E, for which the values of ρ and ϕ are very close, if the standard error is considered; test B refers to the very initial days of the epidemic, whereas test E refers to the use of relative errors in the computation of the objective function. Notice that using relative errors gives some more weight to the small values of the elements of $\mathbf{t}$, which are those recorded at the beginning of the epidemic. Therefore, these results are quite consistent. Notice also that tests B and E are the only tests for which $t_{\text{ini}} > 0$. In these tests, like in test G, the calibrated parameters display the highest coefficient of variation (the ratio between the standard deviation and the average); in other words, these are the tests for which the optimal values show more uncertainty.

A first rough qualitative analysis of the pandemic peak in the continuous model, together with the values of γ, ϕ and ρ listed in Table 4 would also suggest that at the pandemic peak a large fraction of the population would have already been infected, and possibly recovered. In fact, in the continuous model, when I reaches its maximum value we have

$$\frac{\mathrm{d}I}{\mathrm{d}t} = 0 \quad \Rightarrow \quad \beta + \gamma \frac{S}{P} - \delta - \phi - \rho = 0, \tag{27}$$

and, after simple algebraic manipulations,

$$I + R = \frac{\gamma - \alpha}{\gamma} P, \tag{28}$$

where α is defined in (7). The calibration results listed in Table 4 show that α is about one order of magnitude smaller than γ. In particular for the calibration tests performed in this study (Table 4) $(\gamma - \alpha) \cdot \gamma^{-1}$ assumes a relatively high value, close to 0.8.

Two facts should be mentioned about the results of test A shown in Table 5: first, $t_{\text{ini}} < 0$, i.e., it seems that the infection started before the official appearance of the first confirmed case; second, P_{ini} is close to the lower bound chosen in Table 3, so that the model predicts that the population which has been involved in the infection could be relatively small. These qualitative remarks are confirmed by most of the other tests. Notice, in particular, that even if one considers tests B and E, which give the highest average value of P_{ini} among different runs, the runs which yield the least values of $\mathsf{O}_{\mathbf{y},\mathbf{t}}$ give a value of P_{ini} close to $2 \cdot 10^5$, as for test A. Recall that tests B and E share the property of emphasizing the role of early stage data.

Table 6 shows that tests A, D and G are those for which the results of different runs are more consistent with each other. This is important, because it shows that the identification of $\mathbf{p}^{(\text{cal})^\star}$ with the proposed approach appears to be robust for these tests. On the other hand, for the remaining tests, it is important to carefully check the outcomes of each single run. In fact, the initial seed could introduce some bias which cannot be overcome by the `differential_evolution` routine with its default settings and the final result could yield a local minimum, instead of the global one. This is illustrated by the comparison in Fig. 6, which shows the results of test F for the optimal parameters and those averaged among the 10 runs and listed in Tables 4 and 5. This test was designed to fit the data on the deceased people, as shown in Fig. 6(a); the fit seems extremely good, in fact, the two green curves overlap almost perfectly for a large time interval. On the other hand, from Fig. 6(b) it is evident that some of the inversion runs yielded parameters which do not permit to properly and satisfactorily reproduce the data.

From Table 6, it is also apparent that the objective function is computed a great number of times for each single run of the tests. This number strongly varies among the tests. Recall that each computation of $\mathsf{O}_{\mathbf{y},\mathbf{t}}$ requires a run of the model, so that the computational costs could become important. The tests discussed in this paper run on a PC with an `Intel core i7 9th Gen` processor; the execution time of a single run varied from 34 s for test B to 722 s for test F.

4. Discussion on the SIR model

Some basic assumptions, on which the mathematical model developed in this work is founded, deserve to be recalled and discussed.

The model relies on the assumption that the population under consideration is homogeneous. In other words, no distinction is made in terms of gender, age, economic wealth, health and wellness, working conditions, life style, home state, genetic background and so on. In particular, the γ parameter is assumed to be independent of factors like working/living conditions that could be responsible for social distancing and the duration of contacts

Table 6
Properties of inversion of data referred to Italy; the values are based on 10 runs of the minimization algorithm for each test. For the details about the performed tests see Table 2.

Test	Minimum of $O_{\mathbf{y},\mathbf{t}}$	Number of iterations of the algorithm	Maximum number of evaluations of $O_{\mathbf{y},\mathbf{t}}$ for a single run
A	$(7.462 \pm 0.001) \times 10^{-3}$	125 ± 4	11,316
B	$(2.102 \pm 0.03) \times 10^{-5}$	31 ± 2	3,312
C	$(2.60 \pm 1.8) \times 10^{-3}$	190 ± 14	19,338
D	$(8.3164 \pm 0.0008) \times 10^{-3}$	140 ± 4	11,802
E	2.35 ± 0.09	41 ± 12	8,859
F	$(4.9 \pm 3.9) \times 10^{-4}$	320 ± 27	31,566
G	$(5.2032 \pm 0.0013) \times 10^{-2}$	83 ± 3	7,788

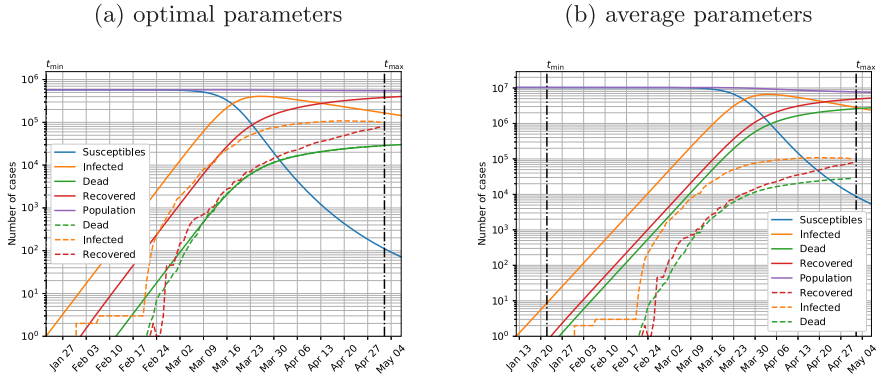

Fig. 6. Comparison of reference (dashed lines) and modeled values (continuous lines) for Italy with the parameters obtained by solution of the inverse problem for test F (see Table 2): (a) optimal parameters corresponding to the global minimum; (b) parameters averaged among the 10 inversion runs (Tables 4 and 5). The vertical dotted black lines delimit the time-frame of the data set used for model calibration, i.e., they correspond to $t_{\min}$ and $t_{\max}$.

of infected – and therefore infectious – individuals within the working/living environments.

The recovery and fatality parameters, ρ and ϕ, respectively, are assumed to be constant too. This is not based on the homogeneity assumption mentioned above only. In fact, this implies that recovery and fatality are modeled as instantaneous processes, i.e., independent of the time passed since each infection occurs; moreover, no relation is considered between death or healing of infected people and the strength of the symptoms of these individuals or to the facility where they are being treated (home, non-intensive care hospital units, Intensive Care Units — ICUs). The latter condition could be modeled by subdividing the class of infected people among sub-classes, e.g., asymptomatic, with light symptoms, admitted to hospital non-intensive care units, admitted to ICUs [34,35].

One could handle the approximation of constant parameters ρ and ϕ, by replacing them with functions $(\tilde{\phi}, \tilde{\rho})$ of elapsed time since infection. Such functions should enter in a deconvolution product involving the number of persons who have been infected at a given time and are still infected, i.e., are not yet recovered or passed away. With this approach, ϕI and ρI in (2) to (4) could be replaced by

$$\int_0^{\tau_{\max}} \tilde{\phi}(\tau)\tilde{I}(t-\tau)\,\mathrm{d}\tau \quad \text{and} \quad \int_0^{\tau_{\max}} \tilde{\rho}(\tau)\tilde{I}(t-\tau)\,\mathrm{d}\tau, \quad \text{where}$$
$$\tilde{I}(t-\tau) = \gamma \frac{I(t-\tau)S(t-\tau)}{P(t-\tau)} \exp\left\{-\int_0^{\tau} \left[\tilde{\rho}(\tau') + \tilde{\phi}(\tau')\right] \mathrm{d}\tau' - \delta\tau\right\}, \tag{29}$$

where δ is the death rate introduced in Definition 2.2.

An attempt to account for time-varying parameters, depending on varying conditions, can be found in [36].

The assumption of homogeneity could be relaxed by considering distributed models, similar to those applied to the study of transport phenomena, e.g., for diffusion of contaminants in the environment. Those models can account for "diffusive" spread and for "advective" transport. However, the required parametrization is often much finer than the one for lumped models, so that the number of parameters to be calibrated strongly increases, and therefore in absence of good quality data it could be difficult to perform a reliable calibration and validation of the model for a practical application.

It is also assumed that the population under study is a closed system, thus disregarding variations induced by short-time, tourist or business travels, by intermediate-time mobility of students and workers, and by long-time effects of migrant fluxes.

The model is also independent of climatic and environmental conditions, i.e., the processes considered by the model are assumed to be independent of the variability of weather conditions and environmental quality at any temporal and space scale. In particular, this means that neither sharp and rapid variations nor annual or seasonal cycles are considered as possible factors affecting the modeled processes.

Epidemic models rarely consider birth and death rates, because the corresponding terms in the underlying equations are usually negligible. In this work, however, these terms have been kept, as they play a significant role in our discussion. In particular, following the assumption of population homogeneity mentioned above, it is assumed that infected pregnant women give birth to infected babies and that this occurs at the same rate as for susceptible women.

With regard to infection rate, which is described by the term $\gamma IS/P$ in (1) and (2), some remarks are in order. This term is computed by assuming that each infected individual has a given, constant number of contacts with other persons per unit time. Our model assumes that the number of persons who cannot be infected is $I + R$, so that the fraction of contacted persons who cannot be infected is given by $(I + R)/P$; on the other hand, the fraction of contacted individuals who can be infected is given by S/P. This is equivalent to assuming that recovered people become

immune to the virus. Notice that timing, magnitude and longevity of immunity against SARS-CoV-2 are still open questions for the scientific community (see, e.g., [37–39]). Moreover, recovered people are assumed to be not infectious, which is the case if the response of their immune system is fast enough so that, once they come in contact with the virus again, the virus is destroyed by the immune system before it can be spread to susceptible persons.

Other promising classes of models are stochastic models [40], either under a Monte Carlo framework or by using assimilation techniques, e.g., the Ensemble Kalman Filter (EnKF, see, e.g., [41]). In principle, Monte Carlo models might be adapted in a relatively easy way to account for several phenomena and also to consider the role of aspects like gender, age, health and wellness on the probability of infection, recovery and decease. On the other hand, EnKF could provide a firm theoretical framework to improve model predictions by means of uncertain data. Other models in the Bayesian framework [42] could be very helpful to handle discrepancies between model predictions and reference values. Unfortunately, in this case the systematic and random errors could be so high as to make it very difficult to handle them even in a stochastic framework.

5. Conclusions

The problem of calibrating the epidemiological parameters of a SIR model describing the evolution in time of the current COVID-19 pandemic is addressed in this work. The calibration is performed by solving numerically the underlying inverse problem via the minimization of an objective function measuring the discrepancy between the simulated solutions to the discretized SIR model and the official data on COVID-19. The iterative optimization process, depending upon the choice of a threshold-and-weight parameter ξ, allows also for the calibration of the initial time t_{ini} accounting for the day when the first infected case occurred, and of the initial population P_{ini} involved at the start of the epidemic.

Several tests were performed in order to study the impact of the data, the time frames over which the data were collected and the performance of different objective functions, depending on the choice of ξ, on the calibration of the parameters (see Table 2). Test A can be considered as the reference one, as it is performed by making use of the official data over the full time frame of 101 days considered here and with a weight $\xi = 10^6$. The results obtained in this test are quite consistent with the ones obtained in tests D and F, performed with full official data over the last 33 days of the full time frame considered here and with official data regarding deaths only but over the full time frame of 101 days, respectively. The only slight inconsistency is shown in the optimal value of the initial population P_{ini} for test F, which turned out to be of an order of magnitude larger than the values for P_{ini} obtained in tests A and D.

The choice of the weight $\xi \in [1, 10^8]$, leading to different objective functions, is responsible for the calibration of the various parameters during various time stages of the epidemic. For $\xi = 1$ (test E) the objective function is akin to the root-mean-squared relative error, and therefore the corresponding results are very similar to test B, where the data from the early time of the monitoring interval are considered, i.e., when the numbers of infected, recovered and deceased people are still quite small. Assigning a higher value of ξ (e.g., $\xi = 10^6$ like in tests from A to D and in test G) yields an objective function akin to the root-mean-squared absolute misfit between target and predicted signals. The differences between these two extreme types of objective functions are evidenced by the different results they produce, but also by the difficulties involved to obtain the global minimum via their minimization. More specifically, for tests B and E, each run of the optimization algorithm requires a relatively small number of iterations for the convergence to a minimum, therefore a relatively small number of evaluations of the objective function is needed. On the other hand, the values of the local minima obtained in these cases are often quite far from the optimal calibrated parameters. A more in-depth analysis of this will be part of future work.

The results of tests B, C and D (see Tables 4 and 5), where the calibration is performed by data collected in successive time intervals $[0, 33]$, $[34, 67]$ and $[68, 101]$ respectively, show that the optimal values of γ and t_{ini} decrease from test B to test D. In other words, the model fits the data in the early stage of the epidemic $[0, 33]$ by relying on the fact that the epidemic started around a week after the official time zero, i.e., when the official first infected case was reported and that at that time the infection was strong, i.e., the infection coefficient γ was large. On the other hand, the calibration of γ and t_{ini} during the time intervals when the epidemic has widely developed, i.e., from day 34 onwards, shows an infection strength that is relatively low compared to the one in the initial stage and that the epidemic outbreak started earlier than the time when the official first case was identified. Table 5 shows in fact a negative t_{ini} for both tests C and D. This fact could be explained by the difficulty in recognizing the appearance of the first infected cases in the epidemic. In other words, the number of infected people could be underestimated in the early stage of epidemic evolution and this might strongly affect the solution of the inverse problem.

Overall, our results show some of the classical, well known difficulties of non-linear least-squares inversion, in particular the dependence of the solution on the starting values, related to the existence of multiple local minima, and the flatness of the objective function around the local minima. To overcome such difficulties we applied the "differential evolution" algorithm [30] and the results obtained were very good. Other relevant algorithms for global optimization that could be tested as part of future work are genetic algorithms [35,43], particle swarm optimization [44], simulated annealing [45].

One of the limitations to the current SIR model is given by the assumptions of homogeneity and steadiness, i.e., the assumption that the epidemiological parameters to be calibrated, γ, ρ and ϕ, are constant (Assumption 2.1). It is the authors' intention to further develop and refine the model considered here as part of future work. This could include for example a division of the class of infected individuals in subclasses that might take care of the gravity of the infection together with the facility where the infected individual is being treated (home, hospital, ICUs, etc.).

Another limitation is given by the uncertainty in the available data. Test G was performed with data ten times greater than the official ones with the intent of starting a sensitivity analysis. The results obtained in test G show that the calibration of the parameters via these data, in particular the values for γ and ρ are not very consistent with tests A, D and F mentioned above. These results emphasize the ill-posed nature of the underlying inverse problem, by providing evidence about the great care that has to be given to the quality of pandemic data, when used to calibrate or validate epidemic models. In fact, poor quality data might yield unrealistic parameter values and, therefore, unreliable model predictions. This fact, together with the limitations in the models, should always be carefully considered especially when these models are used as engines of decision support systems.

Even though it would be improvident to draw quantitative conclusions because of the limitations mentioned above, the preliminary results of our model calibration qualitatively confirm that the infection started earlier than the official appearance of the first episodes of infection. This is a result of paramount

importance also from a practical point of view, as it is a lesson to be considered in the design of early warning systems for future epidemics and for epidemiological risk analysis.

The results of model inversion also suggest that the calibrated model could be reliable for a portion of the entire population. Somehow, the model itself, through its calibration, seems to suggest the width of the population for which its approximations could be valid. In particular, this is shown by the results obtained for a high value of ξ ($\xi = 10^6$) via the calibration of the parameter P_{ini} included in $\mathbf{p}^{(\text{cal})}$. Including P_{ini} among the parameters to be calibrated provides in principle a, possibly very rough, estimate of the width of the initial population involved at the start of the epidemic. This observation seems to go in tandem with the well-known fact that in the countries most affected by COVID-19, the epidemic spread of the virus had mostly concentrated in specific areas: the province of Hubei, and above all the city of Wuhan, in China; the Lombardy region, and above all the provinces of Bergamo, Brescia, Lodi and Milan, in Italy; the city of New York in the first phase of epidemic spread in the United States; Île-de-France in France; Madrid and Catalunya in Spain; London in the UK.

CRediT authorship contribution statement

Alessandro Comunian: Improved the implementation in the computer codes; Performed calculations; Discussed the results; Contributed to the revision of the final manuscript. **Romina Gaburro:** Revised the analytical and formal mathematical aspects of the work; Discussed the results; Contributed to the revision of the final manuscript. **Mauro Giudici:** Designed the work, Wrote the first draft of the manuscript, Developed the first version of the computer codes, Performed calculations; Discussed the results; Contributed to the revision of the final manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work presents results of a purely curiosity-driven research, which has received support only through the standard working facilities of the authors' institutions. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

The data on COVID-19 epidemic have been downloaded from the following URL: https://github.com/CSSEGISandData/COVID-19.

Finally, the authors wish to thank the anonymous referees and the editor, Prof. Víctor M Pérez-García, for their very valuable comments.

References

[1] R. Ross, An application of the theory of probabilities to the study of a priori pathometry.–Part I, Proc. R. Soc. A 92 (1916) 204–230, http://dx.doi.org/10.1098/rspa.1916.0007.

[2] R. Ross, H.P. Hudson, An application of the theory of probabilities to the study of a priori pathometry.–Part II, Proc. R. Soc. A 93 (1917) 212–225, http://dx.doi.org/10.1098/rspa.1917.0014.

[3] W.O. Kermack, A.G. McKendrick, A contribution to the mathematical theory of epidemics, Proc. R. Soc. A 115 (1927) 700–721, http://dx.doi.org/10.1098/rspa.1927.0118.

[4] W.O. Kermack, A.G. McKendrick, Contributions to the mathematical theory of epidemics. II. —The problem of endemicity, Proc. R. Soc. A 138 (1932) 55–83, http://dx.doi.org/10.1098/rspa.1932.0171.

[5] T.C. Reluga, Game theory of social distancing in response to an epidemic, PLoS Comput. Biol. 6 (2010) e1000793, http://dx.doi.org/10.1371/journal.pcbi.1000793.

[6] E.P. Fenichel, Economic considerations for social distancing and behavioral based policies during an epidemic, J. Health Econ. 32 (2013) 440–451, http://dx.doi.org/10.1016/j.jhealeco.2013.01.002.

[7] M.S. Eichenbaum, S. Rebelo, M. Trabandt, The Macroeconomics of Epidemics, in: Working Paper Series, (No. 26882) National Bureau of Economic Research, 2020, http://dx.doi.org/10.3386/w26882.

[8] V. Capasso, G. Serio, A generalization of the Kermack-McKendrick deterministic epidemic model, Math. Biosci. 42 (1978) 43–61, http://dx.doi.org/10.1016/0025-5564(78)90006-8.

[9] V. Capasso, Mathematical structures of epidemic systems, Lect. Notes Biomath., 97, Springer, 1983, http://dx.doi.org/10.1007/978-3-540-70514-7.

[10] E. Beretta, Y. Takeuchi, Global stability of an SIR epidemic model with time delays, J. Math. Biol. 33 (1995) 250–260, http://dx.doi.org/10.1007/BF00169563.

[11] H.W. Hethcote, The mathematics of infectious diseases, SIAM Rev. 42 (2000) 599–653, http://dx.doi.org/10.1137/S0036144500371907.

[12] M. Kamo, A. Sasaki, The effect of cross-immunity and seasonal forcing in a multi-strain epidemic model, Physica D 165 (2002) 228–241, http://dx.doi.org/10.1016/S0167-2789(02)00389-5.

[13] J. Greenman, M. Kamo, M. Boots, External forcing of ecological and epidemiological systems: A resonance approach, Physica D 190 (2004) 136–151, http://dx.doi.org/10.1016/j.physd.2003.08.008.

[14] A. Korobeinikov, Lyapunov functions and global stability for SIR and SIRS epidemiological models with non-linear transmission, Bull. Math. Biol. 68 (2006) 615–626, http://dx.doi.org/10.1007/s11538-005-9037-9.

[15] K. Wang, W. Wang, H. Pang, X. Liu, Complex dynamic behavior in a viral model with delayed immune response, Physica D 226 (2007) 197–208, http://dx.doi.org/10.1016/j.physd.2006.12.001.

[16] R. Peng, F. Yi, Asymptotic profile of the positive steady state for an SIS epidemic reaction–diffusion model: Effects of epidemic risk and population movement, Physica D 259 (2013) 8–25, http://dx.doi.org/10.1016/j.physd.2013.05.006.

[17] C.C. McCluskey, Complete global stability for an SIR epidemic model with delay — distributed or discrete, Nonlinear Anal. RWA 11 (2010) 55–59, http://dx.doi.org/10.1016/j.nonrwa.2008.10.014.

[18] A. Pandey, A. Mubayi, J. Medlock, Comparing vector–host and SIR models for dengue transmission, Math. Biosci. 246 (2013) 252–259, http://dx.doi.org/10.1016/j.mbs.2013.10.007.

[19] L.M.A. Bettencourt, R.M. Ribeiro, Real time Bayesian estimation of the epidemic potential of emerging infectious diseases, PLOS ONE 3 (2008) 1–9, http://dx.doi.org/10.1371/journal.pone.0002185.

[20] H. Joshi, S. Lenhart, K. Albright, K. Gipson, Modeling the effect of information campaigns on the HIV epidemic in Uganda, Math. Biosci. Eng. 5 (2008) 757 – 770, http://dx.doi.org/10.3934/mbe.2008.5.757.

[21] T.W. Ng, G. Turinici, A. Danchin, A double epidemic model for the SARS propagation, BMC Infect. Dis. 3 (19) (2003) http://dx.doi.org/10.1186/1471-2334-3-19.

[22] M. Giudici, F. Baratelli, L. Cattaneo, A. Comunian, G.D. Filippis, C. Durante, F. Giacobbo, S. Inzoli, M. Mele, C. Vassena, A conceptual framework for discrete inverse problems in geophysics, 2019, arXiv:1901.07937.

[23] M. Giudici, Development, calibration and validation of physical models, in: K.C. Clarke, B.O. Parks, M.P. Crane (Eds.), Geographic Information Systems and Environmental Modeling, Prentice-Hall, Upper Saddle River (NJ), 2001, pp. 100–121.

[24] A.N. Tikhonov, V.Y. Arsenin, Solutions of Ill-Posed Problems, V.H. Winston & sons, 1977.

[25] G. Alessandrini, Stable determination of conductivity by boundary measurements, Appl. Anal. 27 (1988) 153–172, http://dx.doi.org/10.1080/00036818808839730.

[26] G. Alessandrini, Open issues of stability for the inverse conductivity problem, J. Inverse and Ill-posed Probl. 15 (2007) 451–460, http://dx.doi.org/10.1515/jiip.2007.025.

[27] O. Doeva, R. Gaburro, W.R.B. Lionheart, C.J. Nolan, Lipschitz stability at the boundary for time-harmonic diffuse optical tomography, Appl. Anal. (2020) http://dx.doi.org/10.1080/00036811.2020.1758314.

[28] D. Baud, X. Qi, K. Nielsen-Saines, D. Musso, L. Pomar, G. Favre, Real estimates of mortality following COVID-19 infection, Lancet Infect. Dis. (2020) http://dx.doi.org/10.1016/S1473-3099(20)30195-X.

[29] E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time, Lancet Infect. Dis. (2020) http://dx.doi.org/10.1016/S1473-3099(20)30120-1.

[30] R. Storn, K. Price, Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces, J. Global Optim. 11 (1997) 341–359, http://dx.doi.org/10.1023/A:1008202821328.

[31] Department of Economic and Social Affairs, 2018 Demographic Yearbook Annuaire démographique, sixty ninth ed., United Nations, 2019.

[32] F. Gao, L. Han, Implementing the Nelder-Mead simplex algorithm with adaptive parameters, Comput. Optim. Appl. 51 (2012) 259–277, http://dx.doi.org/10.1007/s10589-010-9329-3.
[33] R. Fletcher, Practical Methods of Optimization, second ed., John Wiley & Sons, 1987, http://dx.doi.org/10.1002/9781118723203.
[34] G. Giordano, F. Blanchini, R. Bruno, P. Colaneri, A. Di Filippo, A. Di Matteo, M. Colaneri, Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy, Nat. Med. (2020) http://dx.doi.org/10.1038/s41591-020-0883-7.
[35] M.T. Rouabah, A. Tounsi, N.-E. Belaloui, Early dynamics of COVID-19 in Algeria: A model-based study, 2020, arXiv:2005.13516.
[36] L. Ferrari, G. Gerardi, G. Manzi, A. Micheletti, F. Nicolussi, E. Biganzoli, S. Salini, Modelling provincial COVID-19 epidemic data in italy using an adjusted time-dependent SIRD model, 2020, arXiv:2005.12170.
[37] Y. Shi, Y. Wang, C. Shao, J. Huang, J. Gan, X. Huang, E. Bucci, M. Piacentini, G. Ippolito, G. Melino, COVID-19 infection: The perspectives on immune responses, Cell Death Differ. (2020) http://dx.doi.org/10.1038/s41418-020-0530-3.
[38] P. Kellam, W. Barclay, The dynamics of humoral immune responses following SARS-CoV-2 infection and the potential for reinfection, J. Gen. Virol. (2020) http://dx.doi.org/10.1099/jgv.0.001439.
[39] M. Baay, B. Lina, A. Fontanet, A. Marchant, M. Saville, P. Sabot, P. Duclos, J. Vandeputte, P. Neels, SARS-Cov-2: Virology, epidemiology, immunology and vaccine development, Biologicals (2020) http://dx.doi.org/10.1016/j.biologicals.2020.06.005.
[40] V. Isham, Stochastic models for epidemics with special reference to AIDS, Ann. Appl. Probab. 3 (1993) 1–27, http://dx.doi.org/10.2307/2959726.
[41] G. Evensen, The Ensemble Kalman Filter: Theoretical formulation and practical implementation, Ocean Dyn. 53 (2003) 343–367, http://dx.doi.org/10.1007/s10236-003-0036-9.
[42] D. Calvetti, A. Hoover, J. Rose, E. Somersalo, Bayesian Dynamical estimation of the parameters of an SE(A)IR COVID-19 spread model, 2020, arXiv:2005.04365.
[43] L. Davis, Handbook of Genetic Algorithms, Van Nostrand Reinhold, New York, 1991.
[44] J. Kennedy, R.C. Eberhart, Particle swarm optimization, in: Proceedings of ICNN'95 - International Conference on Neural Networks, vol. 4, 1995, pp. 1942–1948.
[45] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Optimization by simulated annealing, Science 220 (1983) 671–680, http://dx.doi.org/10.1126/science.220.4598.671.