SUIHTER: A new mathematical model for COVID-19. Application to the analysis of the second epidemic outbreak in Italy

Nicola Parolini[1], Luca Dede'[1], Paola F. Antonietti[1], Giovanni Ardenghi[1], Andrea Manzoni[1], Edie Miglio[1], Andrea Pugliese[2], Marco Verani[1], and Alfio Quarteroni[1,3]

[1]MOX, Department of Mathematics, Politecnico di Milano, Italy
[2]Department of Mathematics, University of Trento, Italy
[3]Institute of Mathematics, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
(Professor Emeritus)

Abstract

The COVID-19 epidemic is the last of a long list of pandemics that have affected humankind in the last century. In this paper, we propose a novel mathematical epidemiological model named SUIHTER from the names of the seven compartments that it comprises: susceptible uninfected individuals (S), undetected (both asymptomatic and symptomatic) infected (U), isolated infected (I), hospitalized (H), threatened (T), extinct (E), and recovered (R). A suitable parameter calibration that is based on the combined use of least squares method and Markov Chain Monte Carlo (MCMC) method is proposed with the aim of reproducing the past history of the epidemic in Italy, surfaced in late February and still ongoing to date, and of validating SUIHTER in terms of its predicting capabilities. A distinctive feature of the new model is that it allows a one-to-one calibration strategy between the model compartments and the data that are daily made available from the Italian Civil Protection. The new model is then applied to the analysis of the Italian epidemic with emphasis on the second outbreak emerged in Fall 2020. In particular, we show that the epidemiological model SUIHTER can be suitably used in a predictive manner to perform scenario analysis at national level.

1 Introduction

The Coronavirus pandemic of coronavirus disease 2019 (COVID-19) is a tremendous threat to global health. Since the outbreak in early December 2019 in China, more than 1 834 573 global deaths have been registered, while the estimated total number of confirmed cases is 84 511 153 up to January 2nd, 2021 [2]. The real number of people infected is unknown, but probably much higher.

In this scenario, predicting the trend of the epidemic is of paramount importance to mitigate the pressure on the health systems and activate control strategies (e.g. quarantines, lock-downs, and suspension of travel) aiming at containing the disease and delaying the spread.

As these predictions have vital consequences on the different actions taken from governments to limit and control the COVID-19 pandemic, the recent period has seen considerable flowering of epidemiological mathematical models; see, e.g, [6, 14, 21, 22, 30, 39]. However, estimates and scenarios emerging from modeling highly depend on different factors, ranging from epidemiological assumptions to, perhaps most importantly, the completeness and quality of the data based on

which models are calibrated. Since the beginning of the COVID-19 emergency, the quality of data on infections, deaths, tests, and other factors have been spoiled by under-detection or inconsistent detection of cases, reporting delays, and poor documentation. This inconvenient has affected, and still is to date hampering, the intrinsic predictive capability of mathematical models.

Despite the lack or incompleteness of the available data, which makes modeling the current COVID-19 outbreak challenging, mathematical models are still vital to establish predictions within reasonable ranges, and can be adapted to incorporate the effects of public health authority interventions in order to estimate in advance their effectiveness and their impact on the COVID-19 spread. Building upon the celebrated SIR (susceptible (S), infectious (I), and recovered (R)) model proposed in 1927 by Kermack and McKendrick [25], several generalizations have been formulated over the years by enriching the number of compartments, e.g. Susceptible – Exposed – Infectious – Recovered (SEIR), Susceptible - Infectious - Susceptible (SIS), Susceptible - Exposed - Infected - Recovered - Deceased (SEIRD), Susceptible – Exposed – Infectious – Asymptomatic – Recovered (SEIAR), Susceptible - Infectious - Susceptible - Recovered (SIRS), Susceptible - Exposed - Infectious - Quarantined - Recovered (SEIQR), Maternally - derived immunity - Susceptible – Exposed – Infectious – Recovered (MSEIR), ... ; we refer to, e.g., [7, 24, 32] for an overview. Overall, these models have been abundantly applied to locally analyze COVID-19 outbreak dynamics in various countries (see, e.g., [26, 28, 30, 31]).

However, the peculiar epidemiological traits of the COVID-19 ask for models better able to accurately portray the mutable dynamic characteristics of the ongoing epidemic, with particular emphasis on two critical aspects: (i) the crucial role played by the undetected (both asymptomatic and symptomatic) individuals; (ii) the number of individuals that require Intensive Care Unit (ICU) admission. This latter aspect is of paramount importance in designing realistic scenarios that incorporate the pressure of the epidemic on the national health systems.

In this paper we introduce a new mathematical model, named SUIHTER, based on the initials of the seven compartments that it comprises: susceptible uninfected individuals (S), undetected (both asymptomatic and symptomatic) infected (U), isolated infected (I), hospitalized (H), threatened (T), extinct (E), recovered (R). It is a system of coupled ordinary differential equations (ODEs) that are driven by a set of parameters that are indeed piecewise constant time dependent functions. A first set of parameters denote the transmission rates due to contacts between susceptible and undetected, quarantined or hospitalized subjects. A second set of parameters mimics the rates at which I (isolated) and H (hospitalized) individuals develop clinically relevant or life-threatening symptoms. A further parameter indicates the probability rate of detection of previously undetected infected individuals. Another set of parameters indicates the rate of recovery for the four classes of infected subjects. Finally, the last parameters denote the mortality rates for the different compartments.

This SUIHTER model has been conceived to face some of the limitations that can be found in existing epidemiological models applied to the COVID-19 pandemic. On the one hand, some studies adopt simple SIR-like models [26, 30, 31], which have the advantage of having a limited number of parameters to be calibrated, but pay the price of being unable to track the dynamics of different categories of infected individuals. On the other hand, other multi-compartmental models (see e.g. [6, 22]) have been proposed to account for the detailed knowledge of the clinical characterization for different classes of infected individuals according to the actual level of disease severity. However, it is not always possible (and, even when possible, it is not easy) to associate the multiple infected compartments to the available data. The SUIHTER model has been designed with the objective of

2

creating the most compact model able to predict the different categories of infectious individuals which are considered relevant by the policy makers.

A key challenge in modelling the dynamics of COVID-19 epidemic is represented by the large number of the undetected cases. Indeed, the contribution to the spread of the epidemics due to the (often asymptomatic) undetected cases is too relevant to be neglected. Several authors [22, 45, 29] have attempted, using different strategies, to quantify the number of undetected infections and their effect on epidemic spread. In the present work, we propose a strategy about the initialization of those compartments that are not covered by the data (such as *Susceptible*, *Undetected* and *Recovered*).

The model adopts a two-step calibration process based on a preliminary estimation of the model parameters that uses a Least Squares minimization, followed by a Bayesian calibration performed through a Markov Chain Monte Carlo algorithm.

The model has been adopted to simulate the second COVID-19 epidemic outbreak in Italy arisen in Fall 2020 (and still ongoing). In particular, we have investigated the capability of the model in forecasting the occurrence of a peak for the most relevant compartments with an adequate advance notice. Results of the calibration, simulation by `SUIHTER` and predictions for Italy and the six largest Italian regions are also reported.

The outline of the paper is as follows: in Section 2 we introduce the `SUIHTER` mathematical model; Section 3 is devoted to the description of the calibration procedure, Section 4 contains the numerical results along with their discussion. In Section 5, we draw our conclusions and we discuss some model's limitations.

2   Mathematical model

The spread of COVID-19 had made it clear that it is of paramount importance to include in epidemiological models a compartment describing the dynamics of infected individuals that are still undetected. This is, e.g., the case of [22]. However, some compartments presented in [22] (undetected asymptomatic infected and undetected symptomatic infected) are virtually impossible to be validated since these classes of individuals cannot be traced in public databases (cf. [1]). For this reason, building upon [22] we propose a new model more suited to taking full advantage of publicly available data. In particular, our model is described by the following system of ordinary differential equations

$$
\begin{aligned}
\dot{S}(t) &= -S(t)\frac{\beta_U U(t) + \beta_I I(t) + \beta_H H(t)}{N}, \\
\dot{U}(t) &= S(t)\frac{\beta_U U(t) + \beta_I I(t) + \beta_H H(t)}{N} - (\delta + \rho_U)U(t), \\
\dot{I}(t) &= \delta U(t) - (\rho_I + \omega_I + \gamma_I)I(t) + \theta_H H(t), \\
\dot{H}(t) &= \omega_I I(t) - (\rho_H + \omega_H + \theta_H + \gamma_H)H(t) + \theta_T T(t), \\
\dot{T}(t) &= \omega_H H(t) - (\theta_T + \gamma_T)T(t), \\
\dot{E}(t) &= \gamma_I I(t) + \gamma_H H(t) + \gamma_T T(t), \\
\dot{R}(t) &= \rho_U U(t) + \rho_I I(t) + \rho_H H(t)
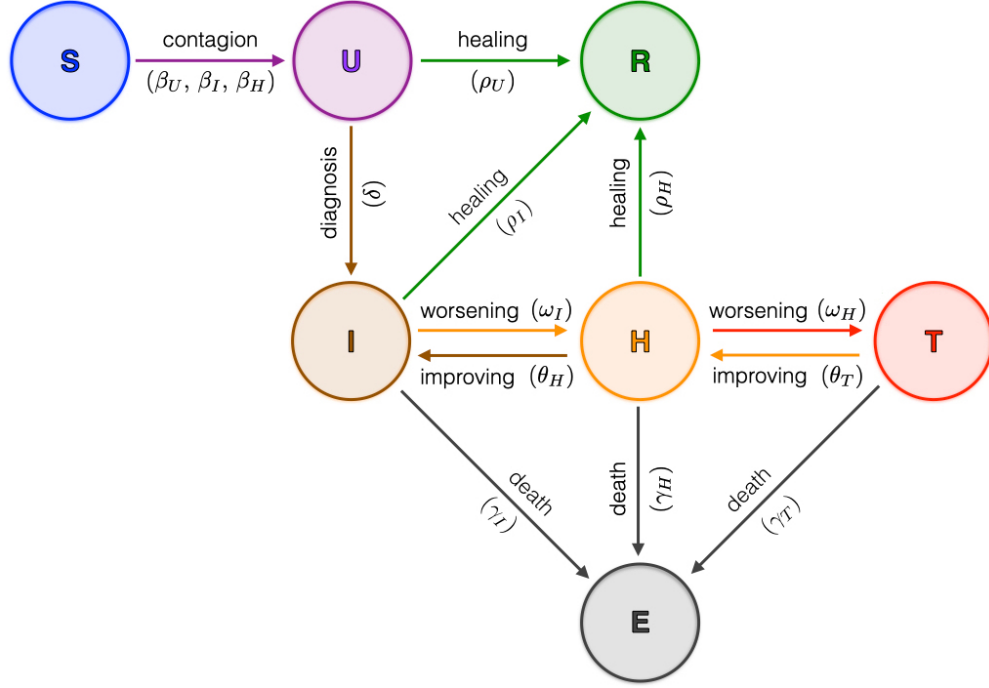\end{aligned}
\tag{1}
$$

3

Figure 1: Interactions among compartments in `SUIHTER` model

where the compartments of the model are defined as follows (see Figure 1):

- S: number of *susceptible* (uninfected) individuals;

- U: number of *undetected* (both asymptomatic and symptomatic) infected individuals;

- I: number of infected individuals *isolated* at home;

- H: number of infected *hospitalized* individuals;

- T: number of infected *threatened* individuals hosted in ICUs;

- E: number of *extinct* individuals;

- R: number of *recovered* individuals,

and $N = S + U + I + H + T + E + R$ denotes the total population (assumed constant).

The model is characterized by the following 14 parameters, some of which are possibly chosen as time dependent piecewise polynomial functions:

- β_U, β_I, β_H denote the transmission rates due to contacts between a susceptible subject and an undetected infected, a quarantined, or a hospitalized subject, respectively;

- ω_I denotes the rate at which I-individuals develop clinically relevant symptoms, while ω_H denotes the rate at which H-individuals develop life-threatening symptoms;

- θ_H and θ_T denote the rates at which H and T-individuals improve their health conditions and return to the less critical I and H compartments, respectively;

- δ denotes the probability rate of detection, relative to undetected infected individuals;

- ρ_U, ρ_I and ρ_H denote the rate of recovery for three classes (U, I and H, respectively) of infected subjects;

- γ_I, γ_H and γ_T denote the mortality rates for the individuals isolated at home, hospitalized and hosted in ICUs, respectively.

Since data available for the *Recovered* cases do not include those individuals who recovered before being detected, we also propose a novel indicator, that will be denoted as *Recovered from detected*, and that we define as

$$R_D(t) = \int_{t_I}^{t} \left(\rho_I I(\tau) + \rho_H H(\tau) \right) \, d\tau.$$

This indicator can be obtained in postprocessing from computed compartments and collects those individuals who recovered after being detected.

In mathematical epidemiology a fundamental quantity is the *basic reproduction number* (denoted by $\mathcal{R}_0$), which is used to measure the transmission potential of a disease. It represents the average number of secondary infections produced by a typical case of an infection in a population where everyone is susceptible (see [16, 7, 32]). For our model, by using a similar argument to the one adopted in the proof of Proposition 1 in [22], we find

$$\mathcal{R}_0 = \frac{\beta_U}{r_1} + \frac{\delta}{r_1} \left(\frac{\beta_I(r_3 r_4 - \theta_T \omega_H) + \beta_H \omega_I r_4}{r_2 r_3 r_4 - r_4 \theta_H \omega_I - r_2 \theta_T \omega_H} \right), \tag{2}$$

where $r_1 = \delta + \rho_U, r_2 = \rho_I + \omega_I + \gamma_I, r_3 = \rho_H + \omega_H + \theta_H + \gamma_H$, and $r_4 = \theta_T + \gamma_T$. For the sake of comparison (cf. Eq. (32) in [22]), we observe that in the present context the characteristic polynomial $q(s)$ of the Jacobian matrix associated to the linearization of (1) around the equilibrium configuration $(\bar{S}, 0, 0, 0, 0, \bar{E}, \bar{R})$ with $\bar{S} + \bar{E} + \bar{R} = N$ is

$$q(s) = s^3 p(s) \qquad \text{with} \quad p(s) = D(s) - \bar{S} N(s)$$

where

$$D(s) = (s + r_1)(s + r_2)(s + r_3)(s + r_4) - (s + r_1)\theta_H \omega_I - (s + r_1)(s + r_2)\theta_T \omega_H$$

and

$$N(s) = (s + r_4)\{\beta_U[(s + r_2)(s + r_3) - \omega_I \theta_H] + \beta_I \delta(s + r_3) + \beta_H \delta \omega_I\} - \beta_U \omega_H \theta_T(s + r_2) - \beta_I \delta \theta_T \omega_H.$$

From the mathematical point of view, the reproduction number $\mathcal{R}_0$ plays the role of a threshold value at the outset of the epidemic. If $\mathcal{R}_0 > 1$, the disease spreads in the population; if $\mathcal{R}_0 < 1$, the number of infected gradually declines to zero. Note that all factor in Eq. (2) are, as expected, actually positive. Furthermore, the expression (2) considerably simplifies upon assuming, as done in the remainder of this paper, that $\beta_I = \beta_H = \theta_H = \gamma_H = 0$. Our SUIHTER model, as other compartmental models, corresponds to a particular case of an integral model with arbitrary distribution of infectious time, for which $\mathcal{R}_0$ is well-known [16].

2.1 Model initialization

A critical issue is the way those compartments for which data are unavailable (*Susceptible*, *Undetected* and *Recovered*) are initialized. In particular, when the analysis focuses on a late phase of the epidemics, as in the present investigation of the second epidemic outbreak in the Fall 2020, it may difficult to estimate those "initial" values as a result of the simulation from day 0.

For these reasons, we have devised a strategy to estimate the number of *Recovered* and *Undetected* individuals based on the value of the *Infection Fatality Ratio* (IFR) defined as the ratio between the number of deaths and the number of resolved cases (dead or recovered) at a specific time (ideally at the end of the epidemic)

$$\text{IFR} = \frac{E}{R+E}. \tag{3}$$

We assume that IFR will be roughly constant in time, at least over the first wave. By using the age-dependent estimates given in [8] the IFR can be estimated around $1.2\,\%$ for Italy. The number of *Recovered* individuals on a given day can be then computed from (3) based on IFR and the number of *Expired* individuals. Moreover, the number of *Undetected* individuals at a given time can be obtained by exploiting the detecting ratio at any given time as follows. We introduce a time dependent *Case Fatality Ratio* (CFR) defined as:

$$\text{CFR}(t) = \frac{\Delta E(t)}{\Delta R_D(t) + \Delta E(t)}, \tag{4}$$

where $\Delta E(t) = E(t + \Delta t/2) - E(t - \Delta t/2)$ and $\Delta R_D(t) = R_D(t + \Delta t/2) - R_D(t - \Delta t/2)$, denote the deaths and (detected) recovered cases observed in a time-window of size $\Delta t = 28$ days around a given time t. The number of *Undetected* individuals at a given time is estimated by assuming that the detection ratio, that is the percentage of detected cases with respect to the total number of positive cases, be computed as

$$\frac{I(t) + H(t) + T(t)}{U(t) + I(t) + H(t) + T(t)} \approx \frac{\Delta I(t) + \Delta H(t) + \Delta T(t)}{\Delta U(t) + \Delta I(t) + \Delta H(t) + \Delta T(t)} = \frac{\text{IFR}}{\text{CFR}(t+d)}. \tag{5}$$

Here we have assumed that the variation of the number of total positive individuals in the time window $[t - \Delta t, t + \Delta t]$ can be approximated by the variation of the resolved case shifted by a confirmation-to-death delay $d = 13$ (see [44]), namely

$$\Delta U(T) + \Delta I(t) + \Delta H(t) + \Delta T(t) \approx \Delta R(t+d) + \Delta E(t+d) = \frac{\Delta E(t+d)}{\text{IFR}}.$$

Similarly, we have assumed that the variation of detected positive individuals in the time window $[t - \Delta t, t + \Delta t]$ can be approximated with the variation of the resolved detected cases shifted by the same time delay:

$$\Delta I(t) + \Delta H(t) + \Delta T(t) \approx \Delta R_D(t+d) + \Delta E(t+d) = \frac{\Delta E(t+d)}{\text{CFR}(t+d)}.$$

By using the available data for I, H, T, and E, other than the value of S deduced under the assumption of constant total population and the estimate of $\text{CFR}(t)$ given in Eq. (4), we estimate the initial conditions for R and U as

$$R(t) = \left(\frac{1}{\text{IFR}} - 1 \right) E(t),$$

$$U(t) = \left(\frac{\mathrm{CFR}(t+d)}{\mathrm{IFR}} - 1 \right) (I(t) + H(t) + T(t)),$$

from Eqs. (3) and (5), respectively.

3 Parameter calibration

Model calibration through data fitting is essential to reproduce the past history of the epidemic and to perform short-term forecasts by inferring the epidemiological characteristics of COVID-19.

Here we use reported isolated, hospitalized, threatened and extinct cases data to estimate the parameters of the proposed SUIHTER model. In particular, we perform the calibration in two steps. Firstly, we find a set of parameter values using an (ordinary) least squares (LS) estimator. Then, we perform a Bayesian calibration using a Markov Chain Monte Carlo (MCMC) algorithm, starting from a prior distribution of the parameters centered about the LS estimate. Calibration of epidemiological models has been already performed in a Bayesian framework, following the pioneering paper by O'Neill and Roberts [37], for several infectious diseases [9, 17, 27]. In the case of COVID-19 epidemic, Bayesian inference has been performed using simpler SIR [40, 47], meta-community SEIR-like [6, 19, 21, 28, 33] and SEIAR [39] models, in this latter case aiming at estimating nine parameters – including a dynamic, time-dependent contact rate $\beta(t)$ – during the first outbreak of the COVID-19 epidemic. In addition to model calibration, our analysis also provides a numerical assessment of the predictive capability of the model, in forecasting with an adequate advance notice the occurrence of a peak for the most relevant compartments.

System (1) can be recast in the following general form describing a system of ODEs for a state vector $\mathbf{Y}$ with n_e components (or compartments): find $\mathbf{Y}(t) : [t_I, t_F] \to \mathbb{R}^{n_e}$ with $\mathbf{Y}(t) = [Y_1(t), \ldots, Y_{n_e}(t)]^T$ such that

$$\mathbf{Y}'(t) = \mathbf{F}(t, \mathbf{Y}(t); \mathbf{p}(t)) \quad t \in (t_I, t_F] \tag{6}$$

$$\mathbf{Y}(t_I) = \mathbf{Y}_0. \tag{7}$$

where $\mathbf{F}$ is the right-hand-side of system (1) and $\mathbf{Y}_0 \in \mathbb{R}^{n_e}$ denotes the initial condition at time t_I evaluated as discussed in Section 2.1. The evolution of the system depends on n_{par} time-dependent parameters, collected into the function $\mathbf{p}(t) : (t_I, t_F] \to \mathbb{R}^{n_p}$. Note that, at the discrete level, the nonlinear nature of system (1) is treated by considering an explicit (fourth-order Runge-Kutta) time discretization with time step $\Delta t = 1$ day

Let us partition the interval $I = [t_I, t_F]$ into n_{ph} phases, corresponding to different epidemic stages due to, e.g., partial restrictions (such as lock-down measures) or different containment rules introduced by the Government or by the local Authorities, so that on each phase the environmental conditions may be considered fixed. Thus, on each phase, the values of the n_{par} model parameters are assumed constant (but unknown), so that we can introduce the following set of admissible parameters

$$\mathcal{P}_{ad} = \{\mathbf{p}(t) : \mathbf{p}(t)|_{I_k} = \mathbf{p}_k \in [\mathbf{p}_{L,k}, \mathbf{p}_{U,k}], \ k = 1, \ldots, n_{ph}\} \tag{8}$$

where $\mathbf{p}_{L,k}, \mathbf{p}_{U,k}$ are given constant vectors. For the sake of notation, let us denote by $\mathbf{p} \in \mathbb{R}^{n_p}$ the vectors of unknown parameters to be estimated, with $n_p = n_{par} n_{ph}$, and let $\mathbf{Y} = \mathbf{Y}(t, \mathbf{p})$ highlight the dependence of the states on the parameters. Consequently, $\mathcal{P}_{ad}$ is the n_p-dimensional hypercube delimited by the constraints (8). Additional constraints on the parameters are assumed, by imposing that some of them are constant over all phases.

Let Δt be a positive time step, for which we consider n_{me} measurements of $n_{com} = 5 < n_e$ compartments at equally spaced times $t_j = j\Delta t$, $j = 1, \ldots, n_{me}$ over the interval $I = [t_I, t_F]$, with $t_1 = t_I + \Delta t$, $t_{n_{me}} = t_F$; in total, we have $n_{com} \times n_{me} = 5 \times n_{me}$ reported data, say $\hat{\mathcal{D}}(t) = \{\hat{\mathbf{Y}}_{I,H,T,E,R_D}(t_j)\}_{j=1}^{n_{me}} \in \mathbb{R}^{5 \times n_{me}}$, that is,

$$\hat{\mathcal{D}}(t) = \{(\hat{I}(t_1), \hat{H}(t_1), \hat{T}(t_1), \hat{E}(t_1), \hat{R}_D(t_1))^T, \ldots, (\hat{I}(t_{n_{me}}), \hat{H}(t_{n_{me}})), \hat{T}(t_{n_{me}}), \hat{E}(t_{n_{me}}), \hat{R}_D(t_{n_{me}}))^T\}.$$

The first stage of the calibration process is then performed by seeking a LS estimate of the parameters vector, given by the solution of the following minimization problem,

$$\hat{\mathbf{p}} = \arg \min_{\mathbf{p} \in \mathcal{P}_{ad}} \tag{9}$$

where

$$\mathcal{J}(\mathbf{p}) := \sum_{j=1}^{n_{me}} \sum_{k=\{I,H,T,E,R_D\}} \alpha_k(t_j) \|\mathbf{Y}_k(t_j, \mathbf{p}) - \hat{\mathbf{Y}}_k(t_j)\|_2^2, \tag{10}$$

where $\mathbf{Y}_k(t_j, \mathbf{p})$, $k = \{I, H, T, E, R_D\}$ denote the components of the solution $\mathbf{Y}$ of Problem (6)-(7), corresponding to the compartments I, H, T, E, and R_D, respectively, evaluated at time instant $t_j, j = 1, \ldots, n_{me}$, for the set of parameters $\mathbf{p}$. For a balanced distribution of the error across the different compartments, whose amplitudes vary along time, the dynamical weight coefficients are defined as $\alpha_k(t_j) = 1/\hat{\mathbf{Y}}_k(t_j)$.

We considered the official epidemiological data supplied daily by the Italian Civil Protection, hereafter called "raw data" and freely available at `https://github.com/pcm-dpc/COVID-19`, [1]. The accuracy of these data is highly questioned, in particular concerning the estimate of the total number of reported positive cases (strongly dependent on the daily screening effort) [21]. This lack of accuracy is indeed one of the most critical aspects of simple SIR-like models in which the calibration of the Infected compartment is often performed using data on the reported positive cases. In our model, the $n_{com} = 5$ time series selected for the calibration (*Isolated, Hospitalized, Threatened, Extincts* and *Recovered from detected*) are the data that have been daily supplied by the Italian authorities since the beginning of the pandemic. One of the key features of the proposed `SUIHTER` model is indeed the one-to-one correspondence of the compartments with the categories for which reliable data, as the ones provided on a daily-basis by the Italian Civil Protection, are available [1].

When n_{ph} phases are considered, equation (9) leads to the optimization of $n_p = 14n_{ph}$ parameters in total. Namely, for each phase of the epidemic, we have the 14 parameters given by $[\beta_U, \beta_I, \beta_H, \omega_I, \omega_H, \delta, \rho_U, \rho_I, \rho_H, \theta_H, \theta_T, \gamma_I, \gamma_H, \gamma_T]$.

Unfortunately, so many parameters make the calibration process problematic. In what follows, we calibrate our model under the following simplifying assumptions:

- β_I and β_H are set to zero, by assuming that the infection only occurs through a contact between a Susceptible individual and an Undetected infected individual (see, for instance, [12]);

- θ_H is set to zero as we assume that a Hospitalized individual can return back to home only once he has recovered, since this parameters may be difficult to estimate in the absence of specific data on the H to I flux.;

- γ_H are set to zero, by assuming that when a Hospitalized individual is in life-threatening conditions, he/she is moved to ICU;

- δ, ρ_U, ρ_I, ρ_H, γ_I, $\theta_T \in \mathbb{R}$ are constant on $[t_I, t_F]$.

With these restrictions, the total number of parameters to be calibrated is reduced to $4n_{ph} + 6$.

The first stage of the calibration process has been performed by solving the minimization problem (9) numerically. We have used a parallel version of the limited memory Broyden-Fletcher-Goldfarb-Shanno algorithm with box constraints (L-BFGS-B), see [49] for details.

The second stage of the calibration process aims at quantifying uncertainties and has been carried out employing a Bayesian framework, since the latter provides probability densities of the input parameters that can be propagated through the model.

Bayesian inference allows us to construct a probability distribution function (PDF) for the unknown parameters merging prior information and available data, these latter entering in the expression of the likelihood function. At this stage, in order to account for the uncertainty on the initial conditions, we extend the set of parameters to be estimated to $\bar{\mathbf{p}} = (\mathbf{p}, \mathbf{q})$ where $\mathbf{q} = (U(t_I), R(t_I))$ collects the initial conditions for the *Undetected* and *Recovered* compartments whose values are not available from the data and can only be estimated. The posterior PDF can then be obtained through the Bayes theorem on conditional probabilities. For the case at hand, we quantify the likelihood of the parameter vector $\bar{\mathbf{p}}$ corresponding to the model outcome $\mathbf{Y}_k(t_j, \bar{\mathbf{p}})$, $k = \{I, H, T, E, R_D\}$ in correlation to the reported cases $\hat{\mathcal{D}}(t)$ as

$$\pi(\hat{\mathcal{D}}(t) \mid \bar{\mathbf{p}}) = \prod_{k=\{I,H,T,E,R_D\}} \frac{1}{(2\pi\sigma^2)^{n_{me}/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^{n_{me}} (\hat{\mathbf{Y}}_k(t_j) - \mathbf{Y}_k(t_j, \bar{\mathbf{p}}))^2\right),$$

where the (unknown) variance σ^2 is assumed to be constant for each compartment.

Note that a usual assumption when estimating parameters of epidemiological models is that reported cases, given their discrete nature, follow a Poisson or negative binomial distribution [17, 21, 12, 33]. However, given the relatively large number of cases, the normal distribution represents a good and flexible approximation, allowing for faster computations.

Using Bayes' theorem, we obtain the posterior distribution of the parameters $\bar{\mathbf{p}}$ accounting for the prior knowledge on the parameters and the reported cases, as

$$\pi(\bar{\mathbf{p}} \mid \hat{\mathcal{D}}(t)) = \frac{\pi(\hat{\mathcal{D}}(t) \mid \bar{\mathbf{p}})\pi(\bar{\mathbf{p}})}{\pi(\hat{\mathcal{D}}(t))} = \frac{\pi(\hat{\mathcal{D}}(t) \mid \bar{\mathbf{p}})\pi(\bar{\mathbf{p}})}{\int_{\mathcal{P}} \pi(\hat{\mathcal{D}}(t) \mid \bar{\mathbf{p}})\pi(\bar{\mathbf{p}})d\bar{\mathbf{p}}},$$

where $\pi(\bar{\mathbf{p}})$ denotes the (uniform) prior distribution for the parameters. Here, we assume that the prior PDF for the model parameters $\mathbf{p}$ is centered at the LS estimate $\hat{\mathbf{p}}$ obtained during the former calibration stage, on a range $[0.9\hat{\mathbf{p}}, 1.1\hat{\mathbf{p}}]$, while priors of the initial values $\mathbf{q}$ is centered around an estimate $\hat{\mathbf{q}}$ obtained based on the IFR, *infection fatality ratio* (see Eq. (3)), on a range $[0.7\hat{\mathbf{q}}, 1.3\hat{\mathbf{q}}]$. The larger relative amplitude of the latter prior interval reflects the higher uncertainty on the initial value for the *Recovered* and *Undetected* compartments.

An alternative, more common and rigorous procedure, would require to specify informative priors for the parameters, starting from key epidemiological features, as done, e.g., in [21, 12]. However, given the large numbers of parameters to be estimated – some of which do not find explicit counterparts in epidemiological literature – we have assumed uniform priors, centered about the

LS estimates, as a practical shortcut to overcome the difficulty in specifying the prior distribution. In terms of predictive capability of the model, numerical results provided in Section 4 allows us to assess the proposed approach.

Since we cannot obtain the posterior distribution over the model parameters $\mathbf{p}$ analytically, we adopt approximate-inference techniques based on Monte Carlo (MC) methods, which aim at generating a sequence of random samples from a Markov chain whose distribution approaches the posterior distribution asymptotically, whence the name of Markov chain Monte Carlo (MCMC) [43]. In particular, we have used the delayed rejection adaptive Metropolis (DRAM) algorithm [23] implemented in `pymcmcstat`, see [34] for the details. The first $500\,000$ samples of the chain serve to tune the sampler and are later discarded (burn-in period). We use the next $500\,000$ samples to approximate the posterior distribution for the parameters $\bar{\mathbf{p}}$.

From the generated chains, we draw N_{MC} samples of the parameters $\bar{\mathbf{p}}_1, \ldots, \bar{\mathbf{p}}_{N_{MC}}$ that we use to perform forward propagation of uncertainty through the model, and to compute predictive envelopes of the `SUIHTER` model compartments (or predictive distributions).

We report the MC samples of the trajectories on the time interval $(t_I, t_{for}]$, including a forecast window $(t_F, t_{for}]$ that extends beyond the time window $(t_I, t_F]$ where data have been reported, to assess the predictive capability of the model.

4   Results and discussion

In this section we present three batteries of numerical results assessing the forecasting capabilities of the `SUIHTER` model. Our analysis focuses on the second wave of the epidemic that started at the end of the Summer of 2020 and, at the time of this writing, is still affecting Italy. In Section 4.1, we present the simulation of the second wave obtained with the `SUIHTER` model using for its calibration all the data between August 20, 2020 and December 31, 2020. By limiting the time range of the data used for the calibration, we also investigate the model capability in forecasting the peaks of the different compartments (see Section 4.2).

Our results at the national level for the second outbreak have been obtained by initializing the *Isolated*, *Hospitalized*, *Threatened* and *Extinct* compartments with the data provided by the Dipartimento della Protezione Civile [1] at August 20, 2020, namely $I = 15\,063$, $H = 883$, $T = 68$ and $E = 35\,418$. The initial value for the *Undetected* and *Recovered* compartments are estimated using the strategy based on IFR and time-dependent CFR introduced in Section 2.1, resulting in the values $U = 12\,274$ and $R = 2\,916\,082$, respectively. Finally, the initial condition for the *Susceptible* compartments is given by $S = (N - I - U - H - T - E - R) = 57\,504\,185$. Note that this would imply that, by the end of the first wave, around 4.8% of the Italian population had been infected. A serosurvey organized by ISTAT and ISS had estimated that 2.5% of the Italian population had been infected [35, 46]; the survey however had a low compliance, so that its results may be biased. A corresponding survey in Spain [41] with a much higher compliance rate estimated seropositivity to 4.6% or 5%, depending on the methodology used for the seroprevalence analysis. Using an ensemble model calibrated over several countries [36] estimate the proportion infected in Italy at September 1 around 4.5%. Using instead a dynamical model calibrated over detailed data, [33] estimate the proportion infected in Italy at September 30 of 4.78%. Thus, the value of *Recovered* cases obtained for August 20 looks rather realistic.

4.1 Simulation of the second epidemic wave

The `SUIHTER` model has been used to simulate the second epidemic outbreak, starting from August 20, 2020 until December 31, 2020. The different phases in which the parameters can take different values have been identified according to the occurrence of some critical events:

- September 24, 2020: all schools at the national level reopened after the summer (and spring lockdown) closure (schools calendars vary by grades and by region level in Italy);

- October 8, 2020: new rules imposing the mandatory use of masks in all locations (either indoor or outdoor) accessible to public;

- October 26, 2020: confinement rules including distance learning for most secondary schools, limitations on the activity of shops, bars and restaurants, strong limitation of sport and leisure activities[1];

- November 6, 2020: stricter confinement rules including distance learning from 9th grade, further restrictions on commercial activities, limitations on the circulation outside the own municipality (for some Italian regions, classified as *red* regions)[2];

- November 15, 2020: additional confinement rules as more regions turned to *red* color [3];

- November 19, 2020: additional confinement rules as more regions turned to *red* color[4];

- November 29, 2020: relaxation of confinement rules in some regions turned to *orange* color[5];

- December 6, 2020: relaxation of confinement rules in some regions turned to *yellow* color[6];

- December 18: stricter confinement rules are introduced for Christmas holidays[7].

By considering a time lag of 4 days (to account for the incubation period) [20], the corresponding phases on which the model parameters are defined and possibly changing) are:

- Phase 1: August 20, 2020 - September 28, 2020;

- Phase 2: September 29, 2020 - October 11, 2020;

- Phase 3: October 12, 2020 - October 29, 2020;

- Phase 4: October 30, 2020 - November 9, 2020;

- Phase 5: November 10, 2020 - November 18, 2020;

- Phase 6: November 19, 2020 - November 23, 2020;

---

[1]DPCM October 24, 2020, `http://www.governo.it/sites/new.governo.it/files/DPCM_20201024.pdf`
[2]DPCM November 4, 2020, `https://www.gazzettaufficiale.it/eli/gu/2020/11/04/275/so/41/sg/pdf`
[3]`http://www.salute.gov.it/imgs/C_17_notizie_5171_0_file.pdf`
[4]`http://www.regione.abruzzo.it/system/files/atti-presidenziali/ordinanze/2020/ordinanza-n-102.pdf`
[5]`http://www.salute.gov.it/imgs/C_17_notizie_5197_0_file.pdf`
[6]`https://www.gazzettaufficiale.it/eli/id/2020/12/05/20A06781/sg`
[7]Decree Law December 18, 2020, n. 172, `https://www.gazzettaufficiale.it/eli/id/2020/12/18/20G00196/sg`

- Phase 7: November 24, 2020 - December 3, 2020;

- Phase 8: December 4, 2020 - December 10, 2020;

- Phase 9: December 11, 2020 - December 22, 2020;

- Phase 10: December 23, 2020 - December 31, 2020.

As mentioned in Section 3, the compartments employed for calibration are only those with more reliable data, namely *Isolated* (I), *Hospitalized* (H), *Threatened* (T), *Extinct* (E) and *Recovered from detected* individuals.

We performed the model calibration by employing the MCMC parameter estimation procedure described in Section 3, over the 10 phases using the data over the full time range from August 20, 2020 to December 31, 2020. The simulations were run for the subsequent 15 days beyond the date associated to the last set of data used for the calibration forecasting the evolution of the epidemic until January 15, 2021. For the new additional phase the values of the parameters are obtained by linearly extrapolating the two (constant) values of the corresponding parameter of the last two phases, located at the final day of each phase, namely phases 9 and 10.

In Figure 2, we report the expected values for the time evolution of the 7 compartments of the SUIHTER model as well as the time evolution of additional compartment of the *Daily new positive*, which corresponds to $\delta U(t)$, and the corresponding 95% prediction intervals obtained by propagating input uncertainties through the model.

We notice that the calibrated compartments (*Isolated*, *Hospitalized*, *Threatened*, *Extinct* and *Recovered from detected*) accurately fit the corresponding data time series in the calibration phase. For all the compartment the 15-day forecast also indicates the capability of the model in predicting the evolution of the epidemic at the national level.

Moreover, the time history of the *Daily new positives* is also in reasonable agreement with the data, proving that the model is able to capture the main dynamics of the system also for quantities that are not directly driven by the data calibration. Our calibration indicates that from August 20, 2020 to January 15, 2021, $4\,056\,118 \pm 284\,533$ individuals have been infected, of which $48.0\% \pm 2.0\%$ have been detected. In addition, we estimate a posteriori, i.e. by using the outputs of the simulation, that the IFR during the same period is $1.3\% \pm 0.1\%$. This latter figure is compatible with the IFR estimated at 1.2% for Italy by using estimates by age reported in [8] while somewhat higher than the estimate shown in [36]. We also observe that our calculated estimates are likely to be underestimated as the second outbreak is still ongoing at the present time and compartments of isolated and extinct individuals become populated at different time scales.

The median values and the 95% credibility intervals computed by the MCMC calibration are reported in Table 1 for the parameters that are constant over the simulation and for the initial values of *Undetected* and *Recovered*, while in Table 2 we report the parameters that are free to change in each phase. The posterior distributions for all the constant and time-dependent parameters are reported in Figures 1 and 2 of the *Supplementary Material*. The traceplots for all the parameters (also available in the *Supplementary Material* in Figures 3 and 4) indicate that the MCMC method reached a good convergence after the first $500\,000$ samples.

The former parameters and time dependent functions represent rates that can be used to interpret the dynamics of the second Italian outbreak. For example, large values of β_U indicate sustained transmission rates at the corresponding phases. Values of healing rates ρ_U, ρ_I and ρ_H are proportional to the probability of healing for individuals in the compartments U, I and H, but
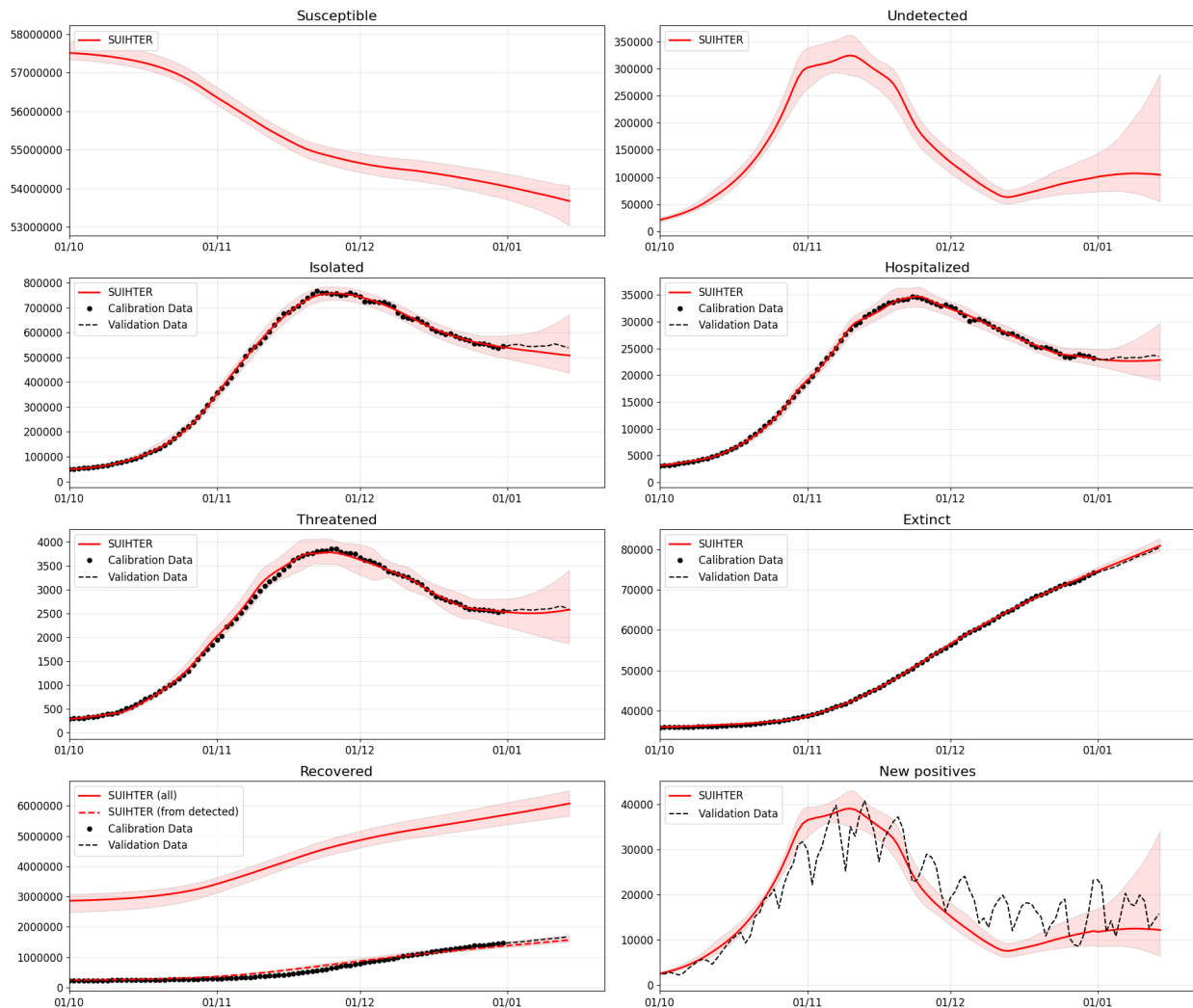
Figure 2: Expected values (solid lines) and 95% prediction intervals (shaded areas) for the 7 compartments of the SUIHTER model plus the additional *Daily new positives* compartment. The data are indicated with black dots (in the calibration phase) and with a dashed line in the validation phase

are inversely proportional to the corresponding average time of healing; the rate ρ_I also incorporates the healing on isolated individuals who are however asymptomatic. To better understand the role of the parameters, note that if they were constant, $\frac{\rho_I}{\rho_I+\omega_I+\gamma_I}$ would represent the probability for an isolated individual to recover without being hospitalized, and similarly $\frac{\rho_H}{\rho_H+\omega_H+\gamma_H}$ represents the probability for a hospitalized individual to recover without being transferred to ICUs. In the same way, $\frac{\gamma_T}{\gamma_T+\theta_T}$ represents the probability of dying for an individual in ICUs, and $\frac{\delta}{\delta+\rho_U}$ represents the probability that an infected individual is detected.

Finally, Table 2 also reports the value of the basic reproduction number $\mathcal{R}_0$ calculated as in Eq.(2) for the SUIHTER model. The calculation uses the model parameters reported in Tables 1 and 2 (columns $1-4$). Note that the amplitude of the credibility intervals are strongly influenced

13

	Median	95% CI
δ	0.12041	[0.10739, 0.12841]
γ_I	3.78e-5	[3.43e-5, 4.15e-5]
ρ_U	0.12320	[0.11303, 0.13593]
ρ_I	0.02408	[0.02197, 0.02658]
ρ_H	0.06677	[0.06171, 0.07212]
θ_T	0.05026	[0.04517, 0.05456]
$U(t_I)$	12 429	1 762
$R(t_I)$	2 727 163	150 286

Table 1: Median values and 95% credibility intervals of constant parameters and U and R initial values

	β_U		ω_I		ω_H		γ_T	
Phase	Median	95% CI	Median	95% CI	Median	95% CI	Median	95% CI
1	0.2640	[0.2475, 0.2825]	0.0059	[0.00537, 0.00648]	0.0132	[0.0121, 0.0146]	0.0760	[0.0691, 0.0837]
2	0.3658	[0.3329, 0.3936]	0.00771	[0.00701, 0.00847]	0.0192	[0.0173, 0.0210]	0.1252	[0.1133, 0.1372]
3	0.3449	[0.3223, 0.3685]	0.00933	[0.00849, 0.01018]	0.0223	[0.0202, 0.0243]	0.0886	[0.0793, 0.0958]
4	0.2756	[0.2485, 0.2972]	0.00691	[0.00629, 0.00755]	0.0264	[0.0238, 0.0286]	0.1561	[0.1400, 0.1689]
5	0.2421	[0.2202, 0.2658]	0.00496	[0.00445, 0.00537]	0.0259	[0.0233, 0.0281]	0.1673	[0.1517, 0.1830]
6	0.1779	[0.1615, 0.1952]	0.00422	[0.00383, 0.00464]	0.0269	[0.0243, 0.0293]	0.1909	[0.1741, 0.2103]
7	0.2093	[0.1906, 0.2307]	0.00340	[0.00309, 0.00373]	0.0263	[0.0238, 0.0286]	0.1900	[0.1726, 0.2079]
8	0.1924	[0.1743, 0.2109]	0.00313	[0.00283, 0.00342]	0.0251	[0.0226, 0.0272]	0.1872	[0.1708, 0.2055]
9	0.3052	[0.2780, 0.3354]	0.00309	[0.00281, 0.00339]	0.0244	[0.0223, 0.0269]	0.1924	[0.1729, 0.2086]
10	0.2949	[0.2686, 0.3251]	0.00351	[0.00319, 0.00385]	0.0249	[0.0226, 0.0272]	0.1867	[0.1700, 0.2053]

Table 2: Median values and 95% credibility intervals of the parameters that changes over the phases and the corresponding $\mathcal{R}_0$

by the choice of the prior in the interval centered about the values of the model parameters obtained by the least squares procedure $\pm 10\%$. They should mainly be judged in relative terms.

We observe that the value of $\mathcal{R}_0$ obtained by the calibration reflects the full reopening of educational activities and work restart after holidays, as well as the public health measures and restrictions later introduced by authorities to contain the second epidemic outbreak. In particular, the rise of $\mathcal{R}_0$ in Phases 2 and 3 follows the full schools reopening and restart of working activities from mid September, and probably accounts for seasonality effects too. Restrictions on mobility, schools, businesses and partial lock-downs were introduced in late October at regional and national levels, as reflected by the decrease of $\mathcal{R}_0$ from Phase 5 to 6, when $\mathcal{R}_0$ became smaller than one. Partial reopening and easing restrictions were gradually introduced in some regions and at the national level from late November, as the new increment of $\mathcal{R}_0$ from Phase 9 indicates.

4.1.1 Simulating the second outbreak for Italian regions

The results obtained simulating the epidemic at the national scale can indeed hide specific local outbreaks. The SUIHTER model can also simulate the evolution of the epidemic for everyone of the 20 Italian Regions for which the same data time series as those used for the national calibration are available. Unfortunately, this is not true for the finer geographical level (the 107 provinces) since only the number of total cases from the beginning of the epidemic is provided.
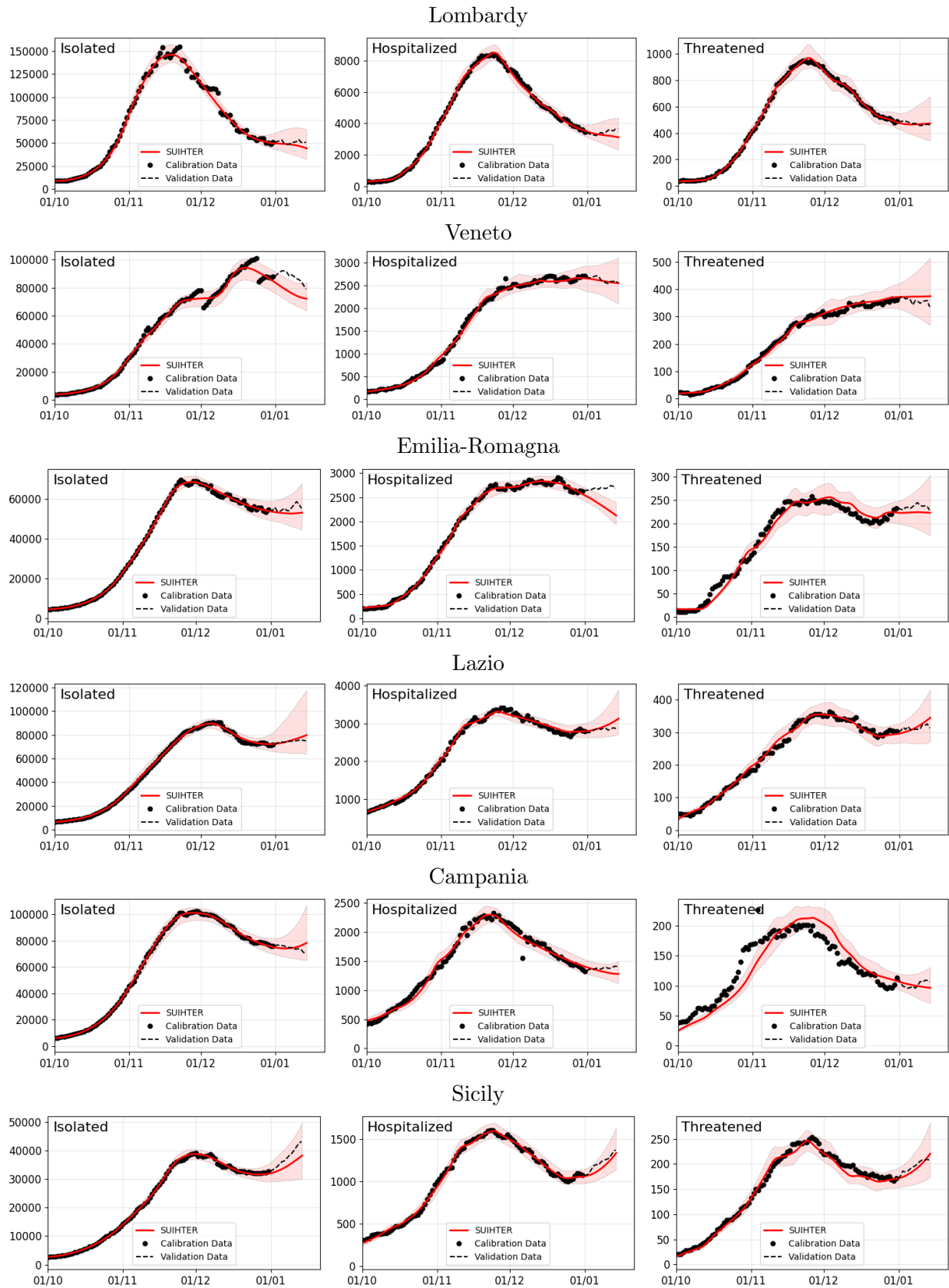
Figure 3: Expected values (solid lines) and 95% prediction intervals (shaded areas) for the *Isolated*, *Hospitalized* and *Threatened* compartments, from left to right, in the six larger Italian regions.

Following the same initialization and calibration strategies adopted at the national level, we have carried out the simulation of the second epidemic outbreak in the six larger Italian regions, namely Lombardy, Veneto, Emilia-Romagna, Lazio, Campania and Sicily. In Figure 3, the expected value for the time evolution of the three infectious compartments used for the calibration and the corresponding 95% prediction intervals are reported for the former three regions. The calibration has been carried out using the same setting as for the national level, i.e. calibrating the model with the data available until December 31, 2020 and then simulating until January 15, 2021. The results obtained by numerical simulations stand in good agreement with the real data, with few exceptions, namely the *Isolated* compartment in Veneto (where the time series is clearly affected by some reporting problems) and the *Hospitalized* compartment in Emilia-Romagna.

4.2 Predicting the peaks

Predicting the peak of an epidemic outbreak is a tremendous challenge for an epidemiological model. Yet, the predictive capability of epidemiological models is of paramount importance to inform policymakers about the dynamics of the disease and foresee timing and level of peaks of infected, hospitalized and ICU treated individuals, as well as the potential effects of policy responses.

With the goal of investigating to which extent our `SUIHTER` model is able to predict the occurrence of the epidemic wave peak, we repeated the calibration using the data over limited time ranges.

In particular, we have considered three different cases: in *Case 0* we used all the data time histories available until December 3, while in *Cases 1, 2* and *3*, the data employed for the calibration were limited to November 23, November 18, and November 9, respectively. For each case, the simulations were run for the subsequent 30 days beyond the date associated to the last set of data used for the calibration and the linear extrapolation carried out as indicated before.

In Figure 4, we report the expected value for the time evolution of the four compartments used for the calibration, and the 95% prediction intervals obtained by propagating input uncertainties through the model. The accuracy of the forecast, as expected, improves when a richer set of data are employed in the calibration. Our simulations show the occurrence of a peak for each of the three compartments, not only for *Case 0* in which the time lapse of the data used for the calibration covers the peaks, but also for *Case 1, 2,* and *3*, when the data time-series employed for the calibration are still rising. However, we should remark that if the model is calibrated with a shorter time series, namely available data stop more than 30 days before the peak, the occurrence of the peak cannot be correctly predicted.

As already noticed, because of the overall complexity of the problem and the limited data available for its calibration, by no means we intend here to certify in rigorous terms the actual values of the future compartments. However, in spite of the widths of the predictive intervals (which depend, at some extent, on the widths of the chosen prior distributions), we nonetheless observe that the expected values (solid lines in Figure 4) carry meaningful prediction capabilities.

To further assess the accuracy of the prediction, it is interesting to compare this peak forecasting with respect to the actual data, i.e. the day and value that have been reported for the different compartments at the end of November 2020. Moreover, we propose a comparison with the predictions obtained using the two different strategies based on data fitting. The first is based on a simple polynomial fit of degree 2 on the last recorded 10 days, while the second is obtained by using a curve registration (see [42] for an overview on this subject) by exploiting the similarities between the first and second waves. The registration procedure is performed by first computing
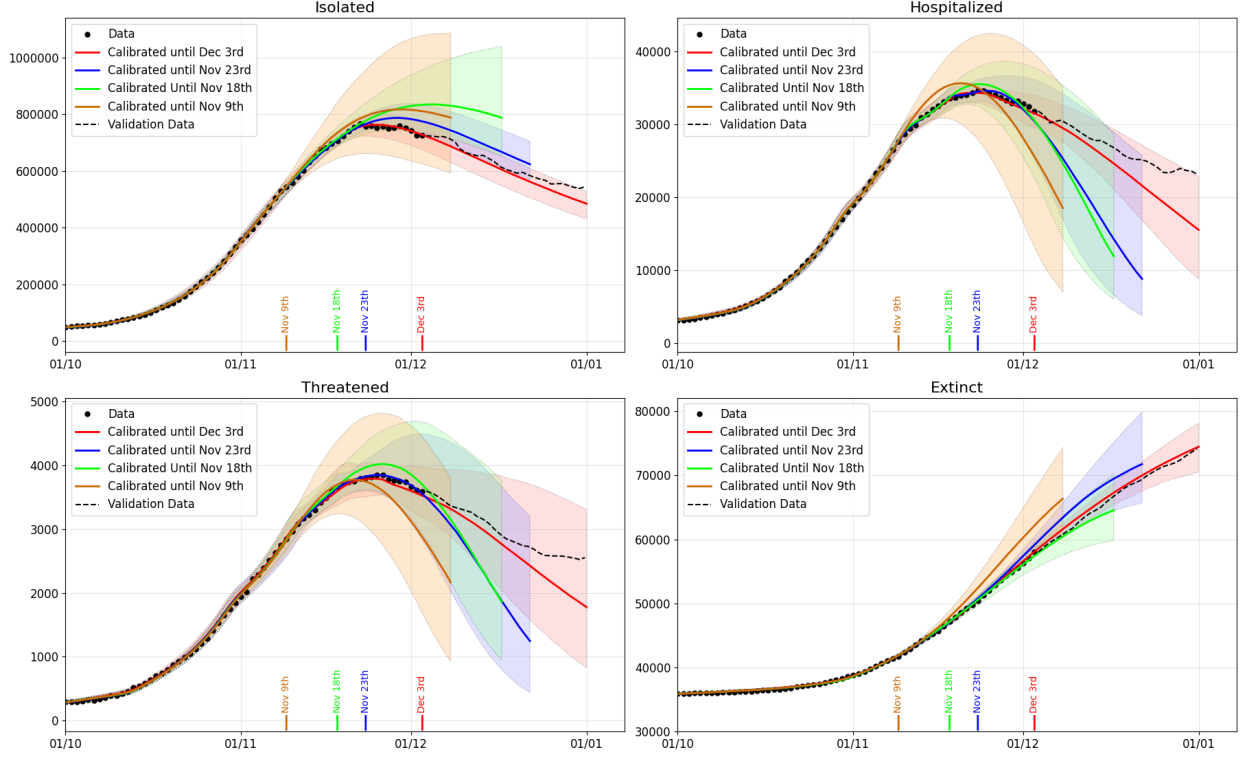
Figure 4: Peak forecast obtained by the `SUIHTER` model with different data ranges for the *Isolated*, *Hospitalized*, *Threatened* and *Extinct* compartments

the Exponentially Modified Gaussian (EMG) function that best fits the first wave. We denote this function as $w(t), t_0 \le t \le t_1$, with t_0 the first day of the recorded data (February 24) and t_1 equal to August 1. Then a second minimization problem is solved to compute the time shift and scaling factors to apply to the computed EMG function to best fit the rising portion of the second wave in the time range $[t_k, t_n]$, with t_k coinciding with October 15 and t_n with the last recorder date. Namely, we look for the optimal time shift $\bar{h}$, and the scaling factors $\bar{s}_1$ and $\bar{s}_2$ such that

$$(\bar{h}, \bar{s}_1, \bar{s}_2) = \underset{h,s_1,s_2}{\arg\min} \sum_{i=t_k}^{t_n} \left(s_1 \, w(s_2 \, t_i + h) - d_i \right)^2,$$

where d_i is the value of the considered data series at day t_i.

For each data series, the fitted EMG function and the optimal values for shift and scaling factors are computed and, in this way the shape of the first wave can be used to complete the second wave for the different compartments.

A comparison between the peak forecast obtained with the `SUIHTER` model, the quadratic extrapolation (based on the last 10 days), and the registration approach is displayed in Figure 5, for the *Isolated*, *Hospitalized* and *Threatened* compartments. The curves show how the prediction in terms of day of peak occurrence and peak value changes as far as an increasing number of data are used (the last data day is reported on the horizontal axis). To minimize the effect of data daily noise, the reference value (in dashed line) is obtained smoothing the data with a Savitzky-Golay polynomial smoothing filters of degree 3 [38].
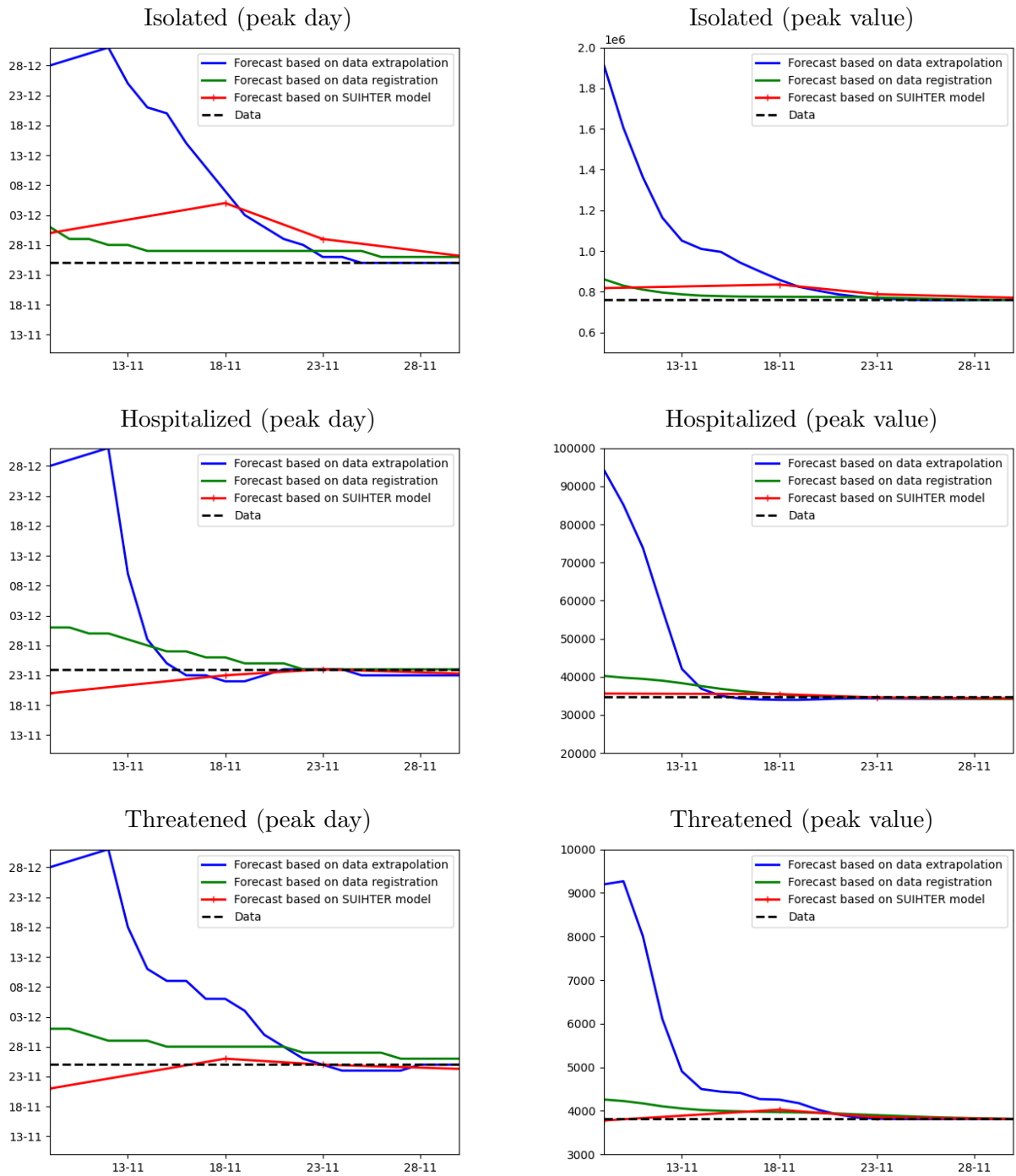
17

Figure 5: Peak day (left) and peak value (right) vs. last used data by day for the three compartments *Isolated* (top), *Hospitalized* (middle) and *Threatened* (bottom), estimated with data extrapolation, data registration and `SUIHTER` model

By comparing the peak predictions with the day and value of the measured (smoothed) data peak (reported with a dashed line in Figure 5), we should first remark that the SUIHTER prediction largely outperforms those obtained with polynomial extrapolation. Moreover, even when compared with predictions based on the registration with the first epidemic wave, the SUIHTER model is more accurate for most of the considered quantities. When making this comparison, it is worthy noticing that, while prediction based on the registration strongly depends on the evolution of the different compartment during the first epidemic wave, the predictions based on SUIHTER do not require any a-priori knowledge of previous epidemic waves.

5   Conclusions and model limitations

In this paper, we have introduced a new mathematical model, named SUIHTER, to describe the ongoing pandemic of coronavirus disease 2019 (COVID-19). This epidemiological model is constructed on seven compartments – susceptible uninfected individuals (S), undetected (both asymptomatic and symptomatic) infected (U), isolated (I), hospitalized (H), threatened (T), extinct (E) and recovered (R) – and we exploit it to study and analyse the second Italian outbreak emerged in Fall 2020 and still ongoing. In particular, our model is suited for calibration against data made available daily by the Italian Civil Protection [1]. On the basis of these data at the national level, our calibration populates the compartments I, H, T, E and R_D, which we purposely use to determine transmission rates, rates of recovery, infection fatality ratio, etc. In particular, SUIHTER allows to estimate the infected, but undetected population, a compartment (U) that is crucial for studying and understanding the epidemic, especially considering that large shares of infected individuals went uncounted during the first and even the second outbreaks in Italy. Moreover, thanks to our approach transmission rates, and thus the basic reproduction number $\mathcal{R}_0$, can be estimated. Finally, our calibration is made robust by exploiting Bayesian estimation using the Markov Chain Monte Carlo method.

The SUIHTER model calibrated at the Italian national level is validated against data related to the last part of the second outbreak. Comparisons are made against basic statistical models, namely quadratic regression and registration of the first epidemic wave. The comparison demonstrates the better accuracy of SUIHTER for predictive purposes. This is made possible by using extrapolated transmission rates that are calibrated at earlier times through regression models, a feature that allows capturing peaks of the second Italian outbreak correctly, and enables using SUIHTER in a predictive fashion by leveraging data available at the current date. This novel approach attempts to circumvent a common issue of the use of epidemiological compartmental models for forecasting [39], that is accurately capturing transmission rates. However, as our approach is based on interpolating values of these transitions rates, the accuracy of their extrapolation and, consequently, their exploitation for prediction within SUIHTER can only be limited to restricted time windows, especially when government interventions and citizen behaviours are changing). Note that, although the calibration procedure did not make any assumptions about the temporal changes in parameters, the estimates accurately reflect the policy changes: estimates of $\mathcal{R}_0$ decrease as control measures are tightened and increase when they are relaxed. The results of the simulation of the second wave carried out at the national and regional levels showed the capability of the model in predicting accurately the time-evolution in a time frame of 15 days past the data used in the calibration. Longer term predictions the model should account for the possible changes in restriction rules that may occur in the future to supply analyses of different scenarios (as recently done in [38] based on

the `SUIHTER` model).

A further limitation of our approach is that we are currently calibrating the Italian epidemic outbreaks at the national level, that is as a whole, without summing up the different contributions at the level of the 20 Italian regions for which data are available [1]; we indeed performed the calibration of the six larger Italian regions. Populating compartments at the national level by summing up results obtained by tailored calibrations of each Italian region would allow a better capturing of the spatio-temporal heterogeneity of the Italian outbreaks, which reflects different mobility patterns and density of population. In this respect, several different approaches have been proposed in literature, see, e.g. [11] and the references therein, ranging from the use of network based models [5, 13], to systems of ordinary differential equations on network [3, 4], as well as non-local partial differential equations [48]. Among the contributions appeared during the COVID-19 pandemic, we also recall the recent papers [6, 21, 28], where a meta-community SEIR-like model has been proposed and employed to reproduce the contagion in Italy. Still, calibrating our `SUIHTER` at the regional level, and for all the regions, would require a more sophisticated design due to the intrinsic ill-posedness of the inverse problem, especially when taking mobility patterns into account. Nevertheless, we plan to better address spatio-temporal heterogeneity of the Italian outbreaks in the future by generalizing our `SUIHTER` model to incorporate suitable spatial-multicity mobility terms at the regional level. Even if a more spatially detailed compartment model is surely desirable, to act, for example, at the province level (Italy is comprised of 107 provinces), at the time being no detailed data for its calibration have been made available.

Albeit the `SUIHTER` model is namely very sophisticated and involves 14 time-dependent parameters and functions to be determined based on available data, we limited our calibration to a subset of the possible control variables, by forcibly setting to zero some parameters that we deemed to be less relevant for the transmission of the epidemic and by assuming as constant over time some other ones. We also neglected incubation time, and we implicitly assumed that all distributions in the states are exponential, which is far from correct [18]. Still, we believe that this qualifies as an acceptable compromise among the complexity of the `SUIHTER` model and its calibration procedure, the associated computational costs, and the accuracy of the results. Some of the calibrated parameters assume values that are able to compensate for those parameters prescribed a-priori, even if their interpretation may not result straightforward in explaining the outbreak. In this respect, we plan to assess the robustness of our approach by allowing the calibration of additional parameters. Further, our multi-compartment `SUIHTER` model does not consider stratification of age groups within the compartments. This is namely an important aspect as some compartments like H, T and E are mostly populated by the elderly, while the transmission mechanisms widely differ by age and context of infection (workplace, school, family, etc.) [15, 33]. We also plan to improve `SUIHTER` by considering age stratification within its compartments.

Among the limitations of our work is the identifiability of model parameters as `SUITHER` calls indeed for calibrating a relatively large set of such parameters. Although we acknowledge its importance, the current study does not present a direct verification of the identification conditions on model parameters nor, in general, a identifiability analysis. Some, albeit indirect, verification of the identifiability properties of the model comes from the very large sets of time series on which calibration has been successfully performed. These are however aspects that we may better address in future studies.

Finally, in consideration of the ongoing emergency situation amidst the second Italian outbreak, we believe that our `SUIHTER` model is well suited to be used in a predictive manner to support and

motivate public health measures. To the best of our knowledge, apart from [10] wherein a SEIRD model is used at the regional level, `SUIHTER` stands as one of the first models to analyze the second Italian COVID-19 outbreak and that can readily serve the purpose of predicting the short-term epidemic trends and perform longer term scenario analyses.

Acknowledgements

Funding statement

References

[1] Presidenza del Consiglio dei Ministri, Dipartimento della Protezione Civile, Italia, last accessed January 16, 2021. https://github.com/pcm-dpc/COVID-19.

[2] Johns Hopkins University. COVID-19 Dashboard by the Center for Systems Science and Engineering, last accessed January 3, 2021. https://coronavirus.jhu.edu.

[3] Linda J.S. Allen, B.M. Bolker, Yuan Lou, and A.L. Nevai. Asymptotic profiles of the steady states for an *SIS* epidemic patch model. *SIAM Journal on Applied Mathematics*, 67(5):1283–1309, 2007.

[4] Julien Arino and P. van den Driessche. A multi-city epidemic model. *Mathematical Population Studies*, 10(3):175–193, 2003.

[5] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J. Ramasco, and Alessandro Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489, 2009.

[6] Enrico Bertuzzo, Lorenzo Mari, Damiano Pasetto, Stefano Miccoli, Renato Casagrandi, Marino Gatto, and Andrea Rinaldo. The geography of COVID-19 spread in italy and implications for the relaxation of confinement measures. *Nature Communications*, 11:4264, 2020.

[7] Fred Brauer, Carlos Castillo-Chavez, and Zhilan Feng. *Mathematical Models in Epidemiology*. Springer, 2019.

[8] Nicholas F. Brazeau, Robert Verity, Sara Jenks, Han Fu, Charles Whittaker, Peter Winskill, Ilaria Dorigatti, Patrick Walker, Steven Riley, Ricardo P. Schnekenberg, Henrique Hoeltgebaum, Thomas A. Mellan, Swapnil Mishra, H. Juliette T. Unwin, Oliver J. Watson, Zulma M. Cucunubá, Marc Baguelin, Lilith Whittles, Samir Bhatt, Azra C. Ghani, Neil M. Ferguson, and Lucy C. Okell. COVID-19 infection fatality ratio: estimates from seroprevalence. Technical Report 34, Imperial College London, 2020. https://doi.org/10.25561/83545.

[9] Simon Cauchemez, Achuyt Bhattarai, Tiffany L. Marchbanks, Ryan P. Fagan, Stephen Ostroff, Neil M. Ferguson, David Swerdlow, and the Pennsylvania H1N1 working group. Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. *Proceedings of the National Academy of Sciences*, 108(7):2825–2830, 2011.

[10] Giulia Cereda, Cecilia Viscardi, Luca Gherardini, Fabrizia Mealli, and Michela Baccini. A SIRD model calibrated on deaths to investigate the second wave of the SARS-CoV-2 epidemic in Italy. *Epidemiologia & Prevenzione – Rivista dell'Associazione Italiana di Epidemiologia*, (2052), 2020.

[11] Dongmei Chen, Bernard Moulin, and Jianhong Wu. *Analyzing and modeling spatial and temporal dynamics of infectious diseases.* John Wiley & Sons, 2014.

[12] Weihsueh A Chiu, Rebecca Fischer, and Martial L Ndeffo-Mbah. State-level needs for social distancing and contact tracing to contain COVID-19 in the United States. *Nature Human Behaviour*, 4(10):1080–1090, 2020.

[13] Vittoria Colizza, Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences*, 103(7):2015–2020, 2006.

[14] Fabio Della Rossa, Davide Salzano, Anna Di Meglio, Francesco De Lellis, Marco Coraggio, Carmela Calabrese, Agostino Guarino, Ricardo Cardona-Rivera, Pietro De Lellis, Davide Liuzza, Francesco Lo Iudice, Giovanni Russo, and Mario di Bernardo. A network model of italy shows that intermittent regional strategies can alleviate the COVID-19 epidemic. *Nature Communications*, 11(5106), 2020.

[15] Laura Di Domenico, Giulia Pullano, Chiara Sabbatini, Pierre-Yves Boelle, and Vittoria Colizza. Impact of lockdown on COVID-19 epidemic in Île-de-France and possible exit strategies. *BMC Medicine*, 18:240, 07 2020.

[16] Odo Diekmann, Hans Heesterbeek, and Tom Britton. *Mathematical Tools for Understanding Infectious Disease Dynamics:.* Princeton University Press, 2013.

[17] Ilaria Dorigatti, Simon Cauchemez, Andrea Pugliese, and Neil Morris Ferguson. A new approach to characterising infectious disease transmission dynamics from sentinel surveillance: Application to the Italian 2009–2010 A/H1N1 influenza pandemic. *Epidemics*, 4(1):9–21, 2012.

[18] Luca Ferretti, Chris Wymant, Michelle Kendall, Lele Zhao, Anel Nurtay, Lucie Abeler-Dörner, Michael Parker, David Bonsall, and Christophe Fraser. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science*, 368(6491), 2020.

[19] Seth Flaxman, Swapnil Mishra, Axel Gandy, H. Juliette T. Unwin, Thomas A. Mellan, Helen Coupland, Charles Whittaker, Harrison Zhu, Tresnia Berah, Jeffrey W. Eaton, Mélodie Monod, Pablo N. Perez-Guzman, Nora Schmit, Lucia Cilloni, Kylie E.C. Ainslie, Marc Baguelin, Adhiratha Boonyasiri, Olivia Boyd, Lorenzo Cattarino, Laura V. Cooper, Zulma Cucunubà, Gina Cuomo-Dannenburg, Amy Dighe, Bimandra Djaafara, Ilaria Dorigatti, Sabine L. van Elsland, Richard G. FitzJohn, Katy A.M. Gaythorpe, Lily Geidelberg, Nicholas C. Grassly, William D. Green, Timothy Hallett, Arran Hamlet, Wes Hinsley, Ben Jeffrey, Edward Knock, Daniel J.

Laydon, Gemma Nedjati-Gilani, Pierre Nouvellet, Kris V. Parag, Igor Siveroni, Hayley A. Thompson, Robert Verity, Erik Volz, Caroline E. Walters, Haowei Wang, Yuanrong Wang, Oliver J. Watson, Peter Winskill, Xiaoyue Xi, Patrick G.T. Walker, Azra C. Ghani, Christl A. Donnelly, Steven Riley, Michaela A.C. Vollmer, Neil M. Ferguson, Lucy C. Okell, Samir Bhatt, and Imperial College COVID-19 Response Team. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584:257–261, 2020.

[20] Tapiwa Ganyani, Cecile Kremer, Dongxuan Chen, Andrea Torneri, Christel Faes, Jacco Wallinga, and Niel Hens. Estimating the generation interval for COVID-19 based on symptom onset data. *Euro Surveill.*, 25(17), 2020.

[21] Marino Gatto, Enrico Bertuzzo, Lorenzo Mari, Stefano Miccoli, Luca Carraro, Renato Casagrandi, and Andrea Rinaldo. Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures. *Proceedings of the National Academy of Sciences*, 117(19):10484–10491, 2020.

[22] Giulia Giordano, Franco Blanchini, Raffaele Bruno, Patrizio Colaneri, Alessandro Di Filippo, Angela Di Matteo, and Marta Colaneri. Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nature Medicine*, pages 1–6, 2020.

[23] Heikki Haario, Marko Laine, Antonietta Mira, and Eero Saksman. DRAM: Efficient adaptive MCMC. *Statistics and Computing*, 16:339–354, 2006.

[24] Herbert W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653, 2000.

[25] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of London. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[26] Adam J. Kucharski, Timothy W. Russell, Charlie Diamond, Yang Liu, John Edmunds, Sebastian Funk, Rosalind M. Eggo, Fiona Sun, Mark Jit, James D. Munday, Nicholas Davies, Amy Gimma, Kevin van Zandvoort, Hamish Gibbs, Joel Hellewell, Christopher I. Jarvis, Sam Clifford, Billy J. Quilty, Nikos I. Bosse, Sam Abbott, Petra Klepac, and Stefan Flasche. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *The Lancet Infectious Diseases*, 20(5):553 – 558, 2020.

[27] Phenyo E. Lekone and Bärbel F. Finkenstädt. Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics*, 62(4):1170–1177, 2006.

[28] Ruiyun Li, Sen Pei, Bin Chen, Yimeng Song, Tao Zhang, Wan Yang, and Jeffrey Shaman. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science*, 368(6490):489–493, 2020.

[29] Zhihua Liu, Pierre Magal, and Glenn Webb. Predicting the number of reported and unreported cases for the COVID-19 epidemics in China, South Korea, Italy, France, Germany and United Kingdom. *Journal of Theoretical Biology*, 509:110501, 2021.

[30] Elena Loli Piccolomini and Fabiana Zama. Monitoring italian COVID-19 spread by a forced SEIRD model. *PLOS ONE*, 15(8):1–17, 08 2020.

[31] Benjamin F. Maier and Dirk Brockmann. Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China. *Science*, 368(6492):742–746, 2020.

[32] Maia Martcheva. *An Introduction to Mathematical Epidemiology*, volume 61. Springer, 2015.

[33] Valentina Marziano, Giorgio Guzzetta, Bruna Maria Rondinone, Fabio Boccuni, Flavia Riccardo, Antonino Bella, Piero Poletti, Filippo Trentini, Patrizio Pezzotti, Silvio Brusaferro, Giovanni Rezza, Sergio Iavicoli, Marco Ajelli, and Stefano Merler. Retrospective analysis of the Italian exit strategy from COVID-19 lockdown. *Proceedings of the National Academy of Sciences*, 118(4), 2021.

[34] Paul R. Miles and Ralph C. Smith. Parameter estimation using the python package `pymcmcstat`. In *Proceedings of the 18th Python in Science Conference (SCIPY 2019)*, 2019.

[35] Ministero della Salute, Italy. Covid-19, the results of the seroprevalence survey illustrated, last accessed January 8, 2021. `http://www.salute.gov.it/portale/news/p3_2_1_1_1.jsp?lingua=italiano&menu=notizie&p=dalministero&id=5012`.

[36] Megan O'Driscoll, Gabriel Santos, Lin Wang, Derek Cummings, Andrew Azman, Juliette Paireau, Arnaud Fontanet, Simon Cauchemez, and Henrik Salje. Age-specific mortality and immunity patterns of SARS-CoV-2 infection in 45 countries. *Nature*, 590(7844):140–145, 2021.

[37] Philip D. O'Neill and Gareth O. Roberts. Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(1):121–129, 1999.

[38] Nicola Parolini, Giovanni Ardenghi, Luca Dede', and Alfio Quarteroni. A mathematical dashboard for the analysis of italian COVID-19 epidemic data. *International Journal for Numerical Methods in Biomedical Engineering*, e3513, 2021.

[39] Mathias Peirlinck, Kevin Linka, Francisco Sahli Costabal, Jay Bhattacharya, Eran Bendavid, John P.A. Ioannidis, and Ellen Kuhl. Visualizing the invisible: The effect of asymptomatic transmission on the outbreak dynamics of COVID-19. *Computer Methods in Applied Mechanics and Engineering*, 372:113410, 2020.

[40] Chiara Piazzola, Lorenzo Tamellini, and Raúl Tempone. A note on tools for prediction under uncertainty and identifiability of SIR-like dynamical systems for epidemiology. *Mathematical Biosciences*, page 108514, 2020.

[41] Marina Pollán, Beatriz Pérez-Gómez, Roberto Pastor-Barriuso, Jesús Oteo, Miguel A. Hernán, Mayte Pérez-Olmeda, Jose L. Sanmartín, Aurora Fernández-García, Israel Cruz, Nerea Fernández de Larrea, Marta Molina, Francisco Rodríguez-Cabrera, Mariano Martín, Paloma Merino-Amador, Jose León Paniagua, Juan F. Muñoz Montalvo, Faustino Blanco, Raquel Yotti, and ENE-COVID Study Group. Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study. *Lancet*, 396(10250):535–544, 2020.

[42] James Ramsay. Curve registration. *The Oxford Handbook of Functional Data Analysis*, 2010.

[43] Christian Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Science & Business Media, 2013.

[44] Timothy W. Russell, Nick Golding, Joel Hellewell, Sam Abbott, Laurence Wright, Carl A. B. Pearson, Kevin van Zandvoort, Christopher I. Jarvis, Hamish Gibbs, Yang Liu, Rosalind M. Eggo, W. John Edmunds, Adam J. Kucharski, and CMMID COVID-19 working group. Reconstructing the early global dynamics of under-ascertained COVID-19 cases and infections. *BMC Medicine*, 18, 2020.

[45] Lucia Russo, Cleo Anastassopoulou, Athanasios Tsakris, Gennaro Nicola Bifulco, Emilio Fortunato Campana, Gerardo Toraldo, and Constantinos Siettos. Tracing day-zero and forecasting the COVID-19 outbreak in Lombardy, Italy: A compartmental modelling and numerical optimization approach. *PLOS ONE*, 15(10):1–22, 10 2020.

[46] Linda L. Sabbadini. First results of the SARS-Cov-2 seroprevalence survey, last accessed January 8, 2021. http://www.salute.gov.it/imgs/C_17_notizie_4998_0_file.pdf.

[47] Leila Taghizadeh, Ahmad Karimi, and Clemens Heitzinger. Uncertainty quantification in epidemiological models for the COVID-19 pandemic. *Computers in Biology and Medicine*, 125:104011, 2020.

[48] Fei-Ying Yang, Wan-Tong Li, and Shigui Ruan. Dynamics of a nonlocal dispersal SIS epidemic model with Neumann boundary conditions. *Journal of Differential Equations*, 267(3):2011–2051, 2019.

[49] Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560, December 1997.

SUIHTER: A new mathematical model for COVID-19. Application to the analysis of the second epidemic outbreak in Italy
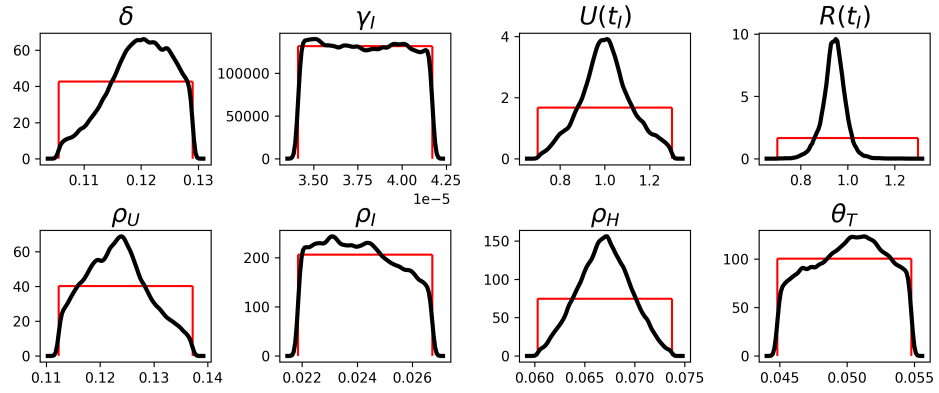

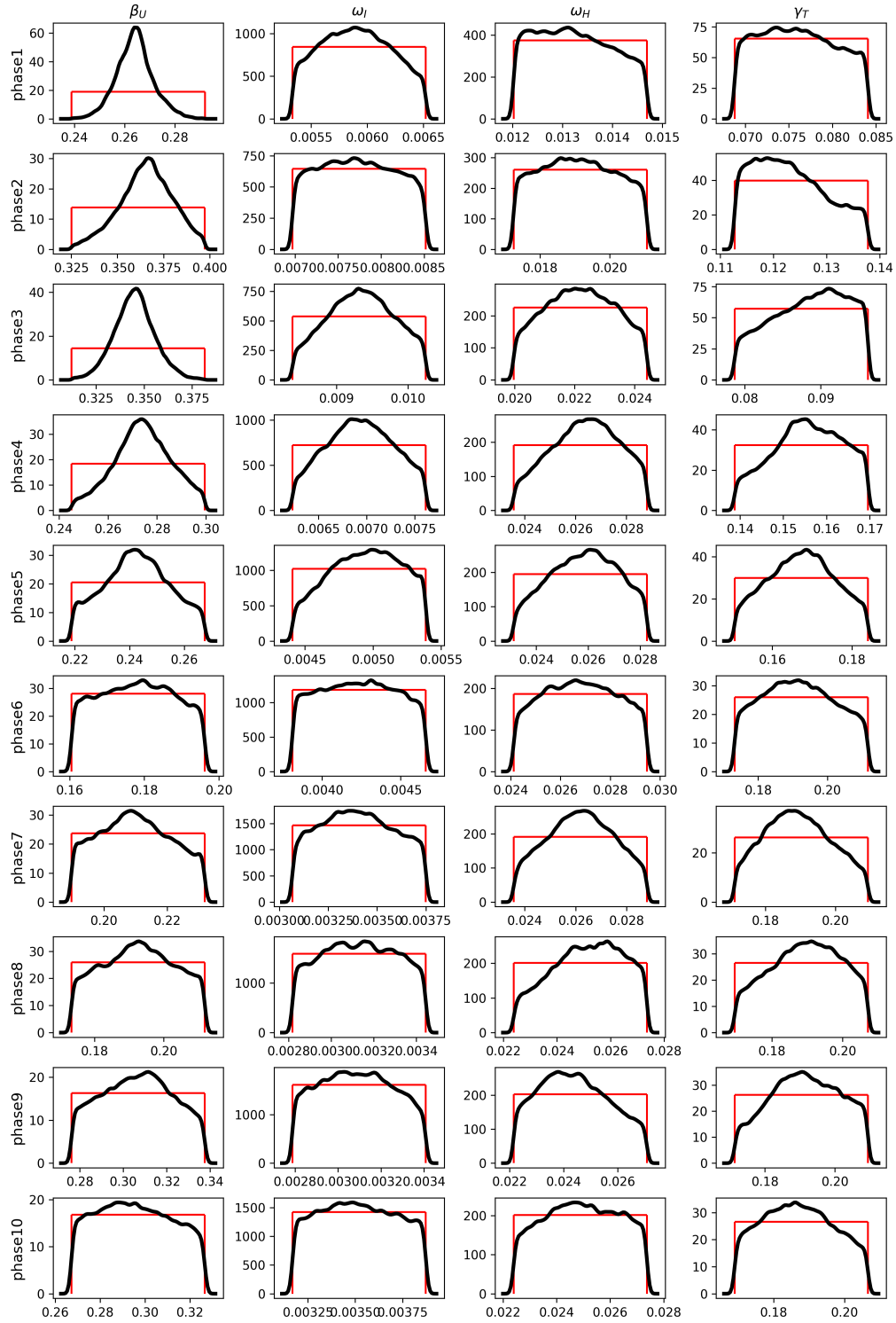Figure 1: Posterior (black) and uniform prior (red) distributions for all the fixed model parameters.

Figure 2: Posterior (black) and uniform prior (red) distributions for all the varying model parameters.
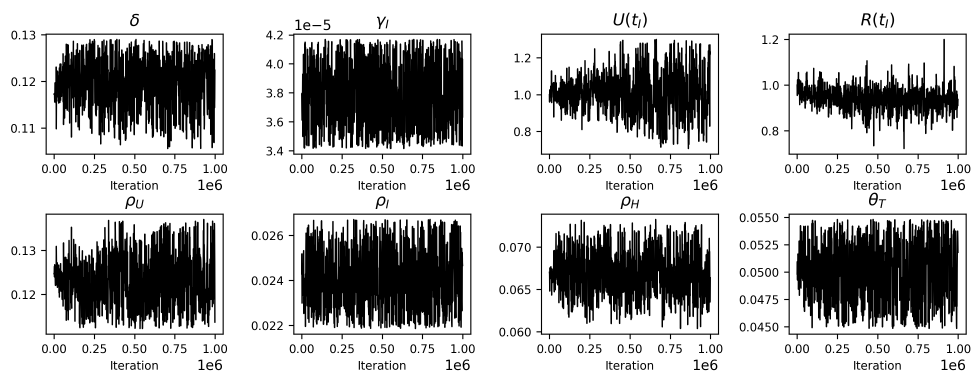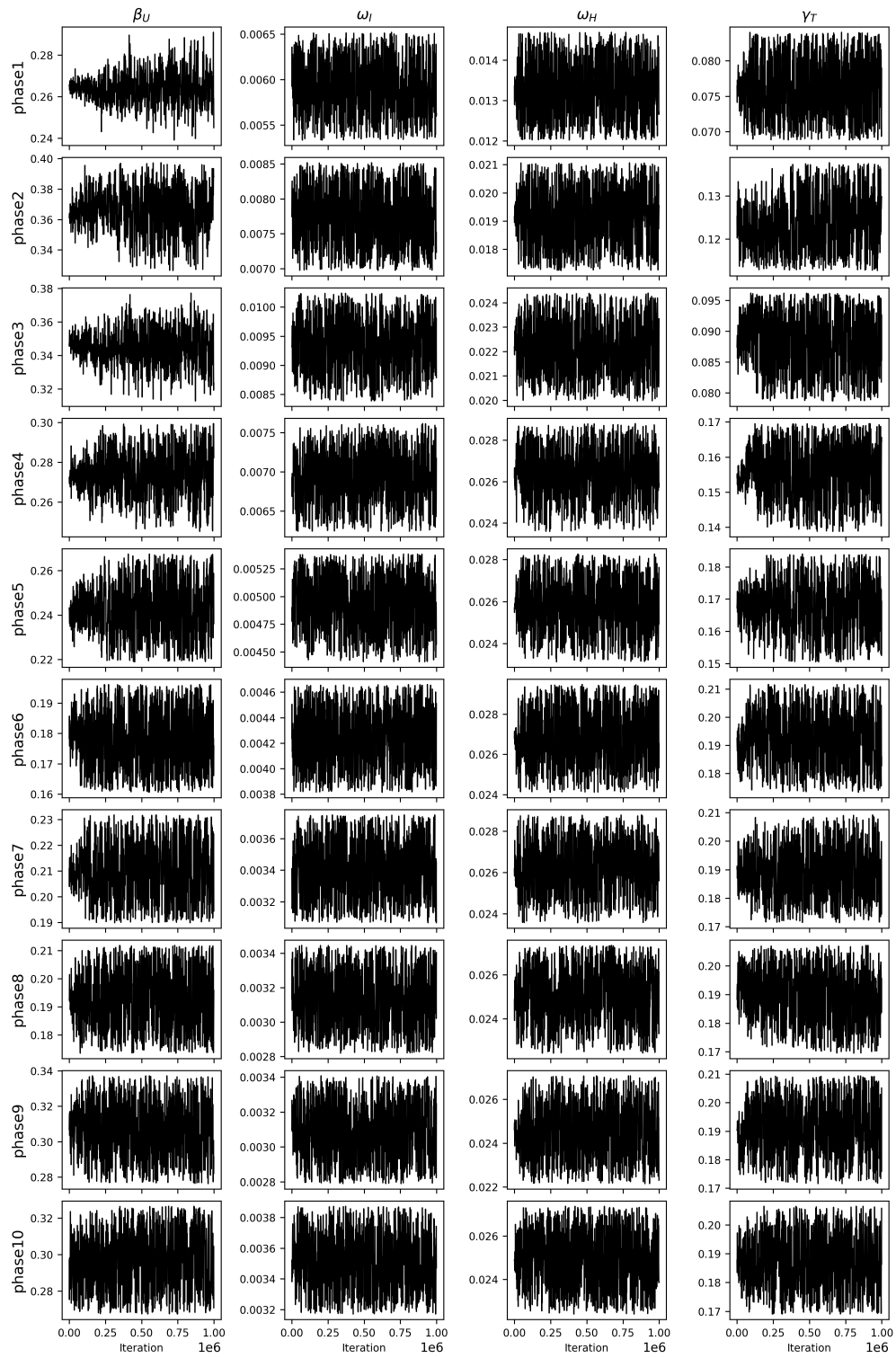
Figure 3: Traceplot for the fixed model parameters.

Figure 4: Traceplot for all the varying model parameters.