DISENTANGLING THE ROLE OF VIRUS INFECTIOUSNESS AND AWARENESS-BASED HUMAN BEHAVIOR DURING THE EARLY PHASE OF THE COVID-19 PANDEMIC IN THE EUROPEAN UNION

MARCOS A. CAPISTRÁN*, JUAN-ANTONIO INFANTE, ÁNGEL M. RAMOS, AND JOSÉ M. REY

Abstract. In this work, we manage to disentangle the role of virus infectiousness and awareness-based human behavior in the COVID-19 pandemic. Using Bayesian inference, we quantify the uncertainty of a state-space model whose propagator is based on an unusual SEIR-type model since it incorporates the effective population fraction as a parameter. Within the Markov Chain Monte Carlo (MCMC) algorithm, Unscented Kalman Filter (UKF) may be used to evaluate the likelihood approximately. UKF is a suitable strategy in many cases, but it is not well-suited to deal with non-negativity restrictions on the state variables. To overcome this difficulty, we modify the UKF, conveniently truncating Gaussian distributions, which allows us to deal with such restrictions. We use official infection notification records to analyze the first 22 weeks of infection spread in each of the 27 countries of the European Union (EU). It is known that such records are the primary source of information to assess the early evolution of the pandemic and, at the same time, usually suffer underreporting and backlogs. Our model explicitly accounts for uncertainty in the dynamic model parameters, the dynamic model adequacy, and the infection observation process. We argue that this modeling paradigm allows us to disentangle the role of the contact rate, the effective population fraction, and the infection observation probability across time and space with an imperfect first principles model. Our findings agree with phylogenetic evidence showing little variability in the contact rate, or virus infectiousness, across EU countries during the early phase of the pandemic, highlighting the advantage of incorporating the effective population fraction into pandemic modeling for heterogeneity in both human behavior and reporting. Finally, to evaluate the consistency of our data assimilation method, we performed a forecast that adequately fits the actual data.

Significance. Data-driven and model-based epidemiological studies aimed at learning the number of people infected early during a pandemic should explicitly consider the behavior-induced effective population effect. Indeed, the non-isolated, or effective, fraction of the population during the early phase of the pandemic is time-varying, and first-principles modeling with quantified uncertainty is imperative for an adequate analysis across time and space. We argue that, although good inference results may be obtained using the classical SEIR type model, the model posed in this work has allowed us to disentangle the role of virus infectiousness and awareness-based human behavior during the early phase of the COVID-19 pandemic in the European Union from official infection notification records.

1. Introduction

In this paper, we carry out a model-driven retrospective analysis of the daily records of new COVID-19 infection notifications in each of the 27 countries of the European Union during

Date: May 29, 2023.

2010 *Mathematics Subject Classification.* Primary: 65N21; Secondary: 62F15.

Key words and phrases. Data assimilation. Forecasting. Epidemics.

∗ To whom correspondence should be addressed.

the initial 22 weeks of the pandemic. This paper aims to contribute toward disentangling the role of virus infectiousness and human behavior during the early phase of the COVID-19 pandemic.

The awareness-based human behavior we are referring to is the change in general living habits in terms of isolation and hygiene measures due to risk perception and governmental control measures. This behavior reduces contact between people and therefore transmission between them. We use a state-space formulation that explicitly accounts for uncertainty in the dynamic model parameters, the dynamic model adequacy, and the infection observation process. We argue that this modeling paradigm allows us to analyze early pandemic dynamics across time and space using the same imperfect first-principles model for all EU countries. We argue that our findings are in agreement with phylogenetic evidence, see Figure 1 of Hodcroft *et al.* [1]: small genetic variability of the SARS-CoV-2 virus across EU countries during the early phase of the pandemic should correspond to little variability in the contact rate or virus infectiousness. This argument highlights the importance of explicitly accounting for human behavior in the dynamical model and underreporting to explain the incidence data. The inferred effective population fraction and observation probability vary considerably across countries. At the same time, the uncertainty of the model adequacy is dispersed across several orders of magnitude, which is expected since a different health system collected each country's data. We studied the early phase in order to choose a critical and important phase of the pandemic that would allow us to make a fair comparison between different countries. Other phases of the pandemic could be studied using the same methodology. During the early phase of the pandemic, there is no genotypic divergence of the virus and very different measures were taken by different countries. We argue that part of the evolution of infections is due to the disease itself and part to human behavior. In fact, since the parameters β (contact rate), ω (effective population fraction) and q (infection observation probability) are globally structurally identifiable, Figures 4 and 5 support our hypothesis: the contact rate β is roughly the same across countries during the early phase of the pandemic, while there is variation in the effective population fraction ω and the infection observation probability q. We believe that our findings are general and can be applied to other outbreak/epidemic/pandemic processes. However, the amount of data and information available about the COVID-19 pandemic in the European Union provides a suitable experimental design for our work. For our analysis, we have used official records from the European Centre for Disease Prevention and Control [2].

Related work. During the last two years, a large number of studies have developed SIR-type models to account for the different processes relevant to the COVID-19 evolution, including undetected cases, new variants, and vaccination (see, for example, Ivorra *et al.* [3] and Ramos *et al.* [4, 5]). It is well known that daily records of new infections and death notifications are the primary source of information for assessing the evolution of the early phase of the pandemic. For example, Villani *et al.* [6] compare death rates across Europe during the first wave of COVID-19. However, at the same time, these records suffer from underreporting and backlogs [7]. The capacity of countries' health systems to respond to and detect new cases during a pandemic emergency [8], as well as people's perceptions of disease risk [9], are determinants of data quality. These issues demand formal inference methods [10, 11, 12] as well as proper evaluation methods to assist pandemic surveillance and forecast [13, 14]. Engbert *et al.* [15], Daza *et al.* [16]. Evensen *et al.* [17] and others have used data assimilation methods to analyze daily records of new COVID-19 infection reports. Engbert *et al.* consider a dynamical model defined by a continuous-time Markov jump process to obtain a consistent

predictive model at the regional level in two German cities. However, a Markov jump process only accounts for intrinsic noise in the epidemic process. In addition, intrinsic and extrinsic factors are known to influence epidemic processes [18]. Engbert *et al.* point out that more realistic pandemic modeling approaches must account for spatial heterogeneity. Furthermore, it is known that short-term awareness of fatalities affects the trend of the pandemic [19], and may induce bias in the forecasts [20]. In this regard, models using an effective population fraction for epidemic data assimilation and forecasting have been used successfully in the past [16, 21, 22, 23]. Likewise, modeling underreporting and identifiability of underreporting parameters has been studied [24, 25]. Of note, there is evidence that during the early phase of the COVID-19 pandemic in Europe, there was little variability in the circulating virus across countries, see Hodcroft *et al.* [1]. Therefore, it is a matter of taste whether to model the non-stationary nature of the pandemic in either the contact rate, the effective population fraction, or the infection reporting probability. Finally, with respect to the prediction of new cases and deaths during the pandemic, the leading teams were composed of forecasting hubs. In these hubs, a group of models provided standardized forecasts that were integrated into the hub using statistical methods such as Bayesian model averaging. See Sherratt *et al.* [26] for the European forecasting hub, and Ray *et al.* [27] for the American forecasting hub. The models in both the European and the American forecasting hubs make predictions from one to four weeks out. The hub predictions are evaluated using appropriate scores. This illustrates the difficulty of making a reliable forecast of an epidemic in the short term.

In this paper, the dynamics of the pandemic is defined using a classical deterministic SEIR model, with an effective population fraction. We do not explicitly model the intrinsic, extrinsic, or external uncertainty in the epidemic model. Instead, we use Galioto and Gorodetsky [28] modeling strategy to account for the different sources of uncertainty through dynamical model parameters, model adequacy parameters, and observations model parameters. In this modeling strategy, the likelihood is approximately evaluated using the Unscented Kalman Filter (UKF). According to Dan Simon [29], pp. 480, "*The Unscented Kalman Filter provides a balance between the low computational effort of Kalman Filter and high performance of particle filters*". Indeed, UKF is well suited for the problem addressed here, provided the dimension of the state vector is small. Following the ideas of Mitchell and Arnold [30], who analyze the effect of different observation functions on epidemic data assimilation, we use an incidence observation model with underreporting (see Section 2). Of note, our observation model is equivalent to the one employed by Engbert *et al.* We set the beginning of community transmission in each analyzed country to the day when the first case was recorded. To assess the robustness of our data assimilation model, we use the posterior predictive distribution of the dynamic model to make short-term forecasts, shutting down the model adequacy and observation uncertainty terms.

Contributions and limitations.

- We show that the contact rate is comparable across most countries of the European Union during the early phase of the pandemic.
- We relate the posterior distribution for the effective population ω to the efficacy of the social contention measures (lockdown, use of masks, self-isolation,...) across different countries of the European Union.
- We use proper forecasting scores to evaluate the quality of the data assimilation and prediction.

- We propose a modification of the Unscented Kalman Filter (UKF) to address the non-negativity constraint on state variables.
- The methodology developed in this work is not thought for forecasting purposes, but to show the disentanglement between the role of virus infectiousness and awareness-based human behavior during the early phase of the COVID-19 pandemic. In any case, the model and methods used here also allow us to carry out forecasts but, similar to other forecasting methods, the forecast reliability decreases if we increase the forecasting period. Furthermore, the present work does not include the late evolution of the pandemic since that analysis corresponds to a new research question.
- The forecasting method presented here is for short periods and the results show its robustness when considering the whole set of countries studied. We think this methodology can be used as the basis to develop other forecasting techniques in particular cases, number of deaths, hospital demand, etc.

2. Theoretical framework

This section proposes a pandemic model accounting for the effective population effect and summarizes a series of implementation and validation tools. In this regard, the two leading model validation tools are the structural identifiability of the mapping from parameters to observables in the noise-free limit, and the data assimilation and forecasting with quantified uncertainty. We have found that the proposed dynamic model is globally identifiable if the initial conditions are prescribed. Thus, we have conducted a thorough Bayesian analysis of the parameter estimation problem. As stated above, the main goal of this paper is to provide a modeling scenario to analyze, at a national level, daily records of new infection notifications during the early phase of the pandemic with an emphasis on inferring the contact rate β, the effective population fraction ω, and the infection observation probability q.

In order to achieve this objective, our proposal develops the following program:

(1) Given a research question, design the data collection: compare records of new infection notifications in 27 countries of the European Union.
(2) Define an inverse problem: quantities that we want to know (contact rate β, effective population fraction ω, and infection observation probability q), quantities that we can measure (daily reports of new infection notifications), a forward mapping that relates them (SEIR model and an observation operator).
(3) Carry out structural identifiability analysis to establish whether the parameters of interest are identifiable given the data, the forward mapping, and the observation operator.
(4) Pose a Bayesian model: model the prior distribution of the parameters, model the likelihood.
(5) Use UKF and MCMC to make an histogram of the posterior distribution of the parameters given the data. (It is not necessary to recover the posterior distribution of the states in this application, which is computationally expensive. Actually, we can recover the distribution of the states from the posterior distribution of the parameters if necessary. Instead, we marginalize with respect to the states. In particular, we use UKF in order to marginalize the likelihood).
(6) Analyze the posterior distribution to establish consistency and robustness of the inference process.
(7) Interpret the results and communicate the findings in the narrative of the problem at hand.

Data. We use data from the first 154 days of the pandemic in each country of the European Union (see Figure 1). The reasoning behind the experimental design is to capture the diversity of situations over the European Union countries across time and space with one single model. Indeed, we look at the development of the pandemic asynchronously. The data assimilation period starts on the day each country started pandemic vigilance, which is different for each country. Despite the varying quality of the data, we have chosen to work directly with those records officially published by the different countries [2] (see also the "Code and data availability" section at the end of Section 3). Negative values were adjudicated with linear interpolation of the adjacent points. In Figure 1, the data for each country are plotted, scaled by their maximum value, so that the fluctuations in the data over the period considered can be appreciated.

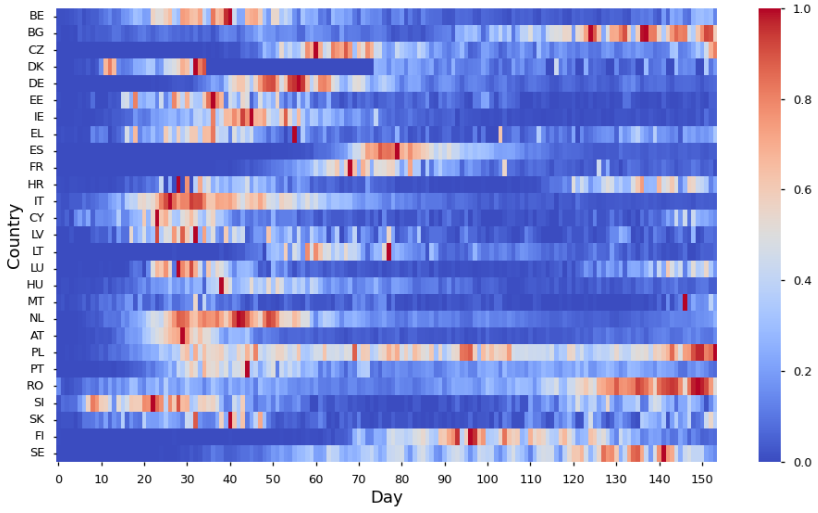


FIGURE 1. **Scaled COVID-19 data of the European Union.** Aligned first 154 days of daily infection notification records in the 27 countries of the European Union [2]. Data for each country starts on the day they started pandemic vigilance and is scaled with respect to its maximum value. Consecutive red and blue squares highlight rapid changes. Each country is represented by two letters, following the European Interinstitutional Style Guide (see [31]).

Note that when very close cold and hot colors are present, it is due to a great oscillation in the data.

Dynamical model. Let us denote the total and isolated populations by N and A, respectively. Then, we define the effective population fraction $\omega = \frac{N-A}{N}$ as the fraction of the population that is not isolated. Aiming to quantify the effective population fraction during the early phase of the pandemic, we shall consider the scaled SEIR model with population $\omega N = S + E + I + R$. The variables S, E, I, and R denote the compartments of susceptible, latently infected, infectious, and removed individuals. We do not consider other compartments, such as people in quarantine, hospitalized, deceased, etcetera. Specifically, the SEIR model with effective population fraction used here is

$$\begin{aligned} \frac{dE}{dt} &= \beta \frac{I}{\omega N}(\omega N - E - I - R) - \sigma E \\ \frac{dI}{dt} &= \sigma E - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned} \tag{1}$$

with initial conditions $(E(0), I(0), R(0))^T = (E_0, I_0, R_0)^T$. Here, N is the official population size of a given country according to [32]. The constants $1/\sigma$ and $1/\gamma$ denote, respectively, the latently infected and infectious residence times. Of note, both residence times are known now that the virology, transmission, and pathogenesis of the SARS-CoV-2 has been described [33,

34, 35], see Table 1. We shall consider, the contact rate β, the effective population fraction ω, and the infection observation probability q (see **Observation model** below) as parameters to be inferred. We shall assume that the initial conditions are approximately known, stating $R(0) = 0$ and

$$E_0, I_0 \sim \text{Gamma}(1, 10). \tag{2}$$

With this parameterization we postulate that there are roughly 10 latently infected individuals and 10 infectious individuals in the population at the time when data started to be recorded (see [16]) with variance 100.

Observation model. We shall assume that the expected value of the daily number of new COVID-19 infection notifications y_k is equal to the product of the aggregated number of individuals entering the infectious compartment between two consecutive observation times, and an observation probability q, namely

$$\mathbb{E}[y_k] = q \int_{t_{k-1}}^{t_k} \sigma E(s) ds, \tag{3}$$

where $[t_{k-1}, t_k]$ denotes the one day long observation time interval. Following Engbert *et al.* [15], the expected value of the reported cumulative number of infected individuals at time t_k is $q(I(t_k) + R(t_k))$. From the last two equations of the dynamical model (1), we notice that $I(t) + R(t) = I(0) + R(0) + \int_0^t \sigma E(s) ds$. Our observation model (3) follows immediately. We recall that q is a parameter to be inferred.

Identifiability. Equations (1) and (3) define a nonlinear parameter-to-observable map in the noise-free limit, or forward mapping

$$y_k = y_k(\theta). \tag{4}$$

We say that the forward mapping (4) is globally structurally identifiable (see [36]) when

$$y_k(\theta) = y_k(\theta') \text{ for all } k \in \mathbb{N} \Leftrightarrow \theta = \theta'. \tag{5}$$

There are many approaches to structural identifiability analysis of non-linear forward mappings defined by a nonlinear system of ordinary differential equations and an observation operator similar to equation (4). Of particular interest to our problem are the definitions and methods presented in [36, 37, 38]. We use the differential algebra suite DAISY by Bellu *et al.* [37] to show that for fixed values of N, σ and γ, the parameters β, ω and q are globally identifiable if the initial conditions E_0, I_0 and R_0 are given. For this reason, although the values of E_0, I_0 are unknown, we have fixed them as realizations of a suitable Gamma distribution.

Data assimilation. For the sake of explicitly accounting for the adequacy of the dynamic model in the data assimilation process, it is convenient to reformulate the dynamical model (1) in terms of a propagator operator. Let $x(t_k) = (E(t_k), I(t_k), R(t_k))^T$ denote the vector of state variables at time $t = t_k$. Using equation (1) with initial condition $x_0 = (E_0, I_0, 0)^T$ we define the propagator operator

$$x_k = \Psi(x_{k-1}, \theta_\Psi), k \in \mathbb{N} \tag{6}$$

recursively by the solution of the initial value problem for equation (1), with initial condition $x(t_{k-1}) = x_{k-1}$, evaluated at time $t = t_k$. Here, we denote $\theta_\Psi = (\beta, \omega)$. Likewise, we denote the observation model defined by equation (3) as

$$h(x_k, \theta_h) = q \int_{t_{k-1}}^{t_k} \sigma E(s) ds,$$

where $\theta_h = q$. Following the notation of Galioto and Gorodetsky [28], data assimilation is carried out with a state space model of the form

$$\begin{aligned} x_k &= \Psi(x_{k-1}, \theta_\Psi) + \xi, \quad \xi \sim \mathcal{N}(0, \Sigma(\theta_\Sigma)), \\ y_k &= h(x_k, \theta_h) + \eta, \quad \eta \sim \mathcal{N}(0, \Gamma(\theta_\Gamma)). \end{aligned} \tag{7}$$

Here, the model uncertainty parameters $\theta_\Sigma = (\Sigma_{11}, \Sigma_{22}, \Sigma_{33})$ allow us to quantify the dynamical model adequacy, and $\Sigma(\theta_\Sigma)$ is the diagonal matrix with diagonal θ_Σ. Likewise, $\theta_\Gamma = \Gamma_{11}$ allows us to quantify the error in the observation process. Note that the 1×1 matrix $\Gamma(\theta_\Gamma)$ is the number Γ_{11} itself.

The numerical approximation of Ψ using the Runge-Kutta method *RK45* (see [39]), and the evaluation of the observation operator h using the composite trapezoidal rule, allows us to obtain a stable numerical approximation of the continuous parameter estimation problem, see [40, 41]. Other, more general error models for ξ and η in equation (7) are described elsewhere. However, the error model provided here is amenable to addressing the different sources of uncertainty in the present work, as shown in the remainder of the paper.

Parameter	Dimension	Value	Reference
β	days^{-1}	Inferred	
ω	adimensional	inferred	
q	adimensional	inferred	
Σ_{11}	population	inferred	
Σ_{22}	population	inferred	
Σ_{33}	population	inferred	
Γ_{11}	population	inferred	
$E(0)$	population	Gamma(1,10)	[16]
$I(0)$	population	Gamma(1,10)	[16]
$1/\sigma$	days	5	[33, 34]
$1/\gamma$	days	14	[33, 34, 35]
N	population	per country	[32]

TABLE 1. **Model parameters.**

Bayesian model. The state-space model (7) defines two conditional probabilities

$$\begin{aligned} \pi(x_k | x_{k-1}, \theta_\Psi, \theta_\Sigma) &= \frac{\exp\left(-\frac{1}{2}||x_k - \Psi(x_{k-1}, \theta_\Psi)||^2_{\Sigma(\theta_\Sigma)}\right)}{(2\pi)^{3/2}|\Sigma(\theta_\Sigma)|^{1/2}}, \\ \pi(y_k | x_k, \theta_h, \theta_\Gamma) &= \frac{\exp\left(-\frac{1}{2}||y_k - h(x_k, \theta_h)||^2_{\Gamma(\theta_\Gamma)}\right)}{(2\pi)^{1/2}|\Gamma(\theta_\Gamma)|^{1/2}}, \end{aligned} \tag{8}$$

for $k = 1, \ldots, n$. Denoting by $\theta = (\theta_\Psi, \theta_h, \theta_\Sigma, \theta_\Gamma)$ and $\mathscr{Y}_n = (y_1, \ldots, y_n)^T$, the joint likelihood is given by

$$\pi_{Y|X,\Theta}(\mathscr{Y}_n|x,\theta)\pi_{X|\Theta}(x|\theta) \tag{9}$$

and, from equations (8), it can be expressed as

$$\pi_{Y|X,\Theta}(\mathscr{Y}_n|x,\theta)\pi_{X|\Theta}(x|\theta) = \prod_{k=1}^{n} \pi(x_k|x_{k-1},\theta_\Psi,\theta_\Sigma)\pi(y_k|x_k,\theta_h,\theta_\Gamma). \tag{10}$$

If we pose a model $\pi_\Theta(\theta)$ for θ of the prior distribution, we obtain a posterior distribution

$$\pi_{\Theta|X,Y}(\theta|x,\mathscr{Y}_n) \propto \pi_\Theta(\theta)\pi_{Y|X,\Theta}(\mathscr{Y}_n|x,\theta)\pi_{X|\Theta}(x|\theta). \tag{11}$$

Since the propagator Ψ is nonlinear, it becomes necessary to explore the posterior distribution (11) numerically, e.g., through Markov Chain Monte Carlo (MCMC). However, it is apparent that the posterior distribution of the states X is not needed since these can be computed from the posterior distribution of the parameters θ and the propagator Ψ. Thus we require to explore the marginal posterior distribution

$$\pi_{\Theta|Y}(\theta|\mathscr{Y}_n) = \int \pi_{\Theta|X,Y}(\theta|x,\mathscr{Y}_n)dX \propto \pi_\Theta(\theta)\pi_{Y|\Theta}(\mathscr{Y}_n|\theta), \tag{12}$$

where $\pi_{Y|\Theta}(\mathscr{Y}_n|\theta)$ is the marginal likelihood

$$\pi_{Y|\Theta}(\mathscr{Y}_n|\theta) = \int \pi_{Y|X,\Theta}(\mathscr{Y}_n|x,\theta)\pi_{X|\Theta}(x|\theta)dX. \tag{13}$$

There are no straightforward exact methods to compute the marginal posterior distribution (12). However, Theorem 1 of Galioto and Gorodetsky [28] provides a recursive algorithm to approximate sequentially the marginal likelihood $\pi_{Y|\Theta}(\mathscr{Y}_n|\theta)$, implemented in Algorithm 3 of the same reference, using the Unscented Kalman Filter (UKF), thus allowing to use MCMC to sample the marginal posterior distribution

$$\pi_{\Theta|Y}(\theta|\mathscr{Y}_n) \propto \pi_\Theta(\theta)\pi_{Y|\Theta}(\mathscr{Y}_n|\theta). \tag{14}$$

For the sake of exposition clarity, we write here the following

Theorem 2.1. *(Marginal Likelihood, Theorem 1 of* [28]*, Theorem 12.1* [42]*) Let* $\mathscr{Y}_n = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ *denote a set of observations up to time* t_n*. Let the initial state have distribution* $\pi(x_0|\theta)$*. Denoting* $\pi(x_0|\theta, \mathscr{Y}_0) = \pi(x_0|\theta)$ *then the marginal likelihood* $\pi_{Y|\Theta}(\mathscr{Y}_n|\theta)$ *is defined recursively, for* $k = 0, \ldots, n-1$*, in three steps*

Prediction:

$$\pi(x_{k+1}|\theta,\mathscr{Y}_k) = \int \frac{\exp\left(-\frac{1}{2}||x_{k+1} - \Psi(x_k,\theta_\Psi)||^2_{\Sigma(\theta_\Sigma)}\right)}{(2\pi)^{3/2}|\Sigma(\theta_\Sigma)|^{1/2}} \times \pi(x_k|\theta,\mathscr{Y}_k)dx_k. \tag{15}$$

Update:

$$\pi(x_{k+1}|\theta,\mathscr{Y}_{k+1}) = \pi(x_{k+1}|\theta,\mathscr{Y}_k) \times \frac{\exp\left(-\frac{1}{2}||y_{k+1} - h(x_{k+1},\theta_h)||^2_{\Gamma(\theta_\Gamma)})\right)}{(2\pi)^{1/2}|\Gamma(\theta_\Gamma)|^{1/2}} \tag{16}$$

Marginalization:

$$\pi(\mathscr{Y}_{k+1}|\theta) = \int \pi(x_{k+1}|\theta,\mathscr{Y}_{k+1}) \frac{\exp\left(-\frac{1}{2}||y_{k+1} - h(x_{k+1},\theta_h)||^2_{\Gamma(\theta_\Gamma)})\right)}{(2\pi)^{1/2}|\Gamma(\theta_\Gamma)|^{1/2}} dx_{k+1}. \tag{17}$$

In the present case, Algorithm 3 of Galioto and Gorodetsky needs to be adapted to limit the state variables to take only nonnegative values. To enforce this condition we do not modify the distribution of the state variables. Instead, we have given zero probability to proposals that give rise to (Gaussian) evidence distributions that either, its mean has any negative component, or the covariance is not positive definite (of note, in the present case, the evidence distribution is univariate). This is described in Algorithm 1 below, where $\mathscr{Y}_0 = \emptyset$, $\mathscr{Y}_k = \{y_i : i \leq k\}, k = 1, \ldots, n$ and $\left[\sqrt{P_{k-1}}\right]_i$ stands for the ith column of Cholesky factor of matrix P_{k-1}. Inspired by Engbert *et al.* [15], we set the distribution of the initial conditions in Algorithm 1 as a Gaussian distribution $\mathcal{N}(m_0, P_0)$, where $m_0 = (E_0, I_0, 0)^T$, with E_0, I_0 as a draw of Gamma$(1, 10)$ and $P_0 = 10I_3$, where $I_3 \in \mathbb{R}^{3\times 3}$ is the identity matrix.

In this paper we use the affine invariant probability transition kernel t-walk from Christen and Fox [43] to implement the UKF-MCMC algorithm to sample the marginal posterior distribution (14) given data $\mathscr{Y}_n$ (daily records of new infections).

Prior elicitation. We elicit prior distributions for the inferred parameters as follows: We argue that early in the pandemic, there should be little variability in the contact rate β across countries. Based on Figure 2 of Park *el al.* [44] and Brauer *el al.* [45], we pose a prior on β such that $3.0 = \mathbb{E}(\mathscr{R}_0) = \frac{\beta}{\gamma}$. Similar to epidemic data assimilation models [23, 46], we account for the non-homogeneous nature of the epidemic as a dynamical system letting $\omega = \omega_k$ and $q = q_k$ follow Gaussian random walks in the logarithmic scale. The resulting models are as follows

$$\begin{aligned} \log(\beta) &\sim \mathscr{N}(\log(3\gamma), \sigma_\beta^2), \\ \omega_0 &\sim \text{Beta}(1.1, 1.1), \\ \text{logit}(\omega_k) &\sim \mathscr{N}(\text{logit}(\omega_{k-1}), \sigma_\omega^2), \\ q_0 &\sim \text{Beta}(1.1, 1.1), \\ \text{logit}(q_k) &\sim \mathscr{N}(\text{logit}(q_{k-1}), \sigma_q^2), \end{aligned} \tag{18}$$

where $\text{logit}(z) = \log(\frac{z}{1-z})$.

For the dynamic model adequacy parameters θ_Σ we elicit as prior a distribution Gamma$(1, N)$ in order to scale it to each country's population size N. Finally, the prior distribution for $\theta_\Gamma = \Gamma_{11}$ is a distribution Gamma$(1, N)$. The prior distribution of θ is the product of the prior distribution of each parameter.

Forecasting. To assess the consistency our data assimilation method, we carry out 15 days forecasting with a proper forecast evaluating score. Forecasting is defined using the posterior distribution $\pi_{\Theta|Y}(\theta|\mathscr{Y}_n)$. For each sample $\theta = (\theta_\Psi, \theta_h, \theta_\Sigma, \theta_\Gamma)$ of this posterior distribution of the parameters given data, we consider the state vector $x_k = x_k(\theta)$ at the time t_k and the corresponding observation y_k. Then, taking $k_0 = 154$ and x_{k_0} as the initial value, we propagate the state along the propagator and observation operators, ignoring model adequacy errors

$$\begin{aligned} x_{k_0+j} &= \Psi(x_{k_0+j-1}, \theta_\Psi) \\ y_{k_0+j} &= h(x_{k_0+j}, \theta_h) \end{aligned} \tag{19}$$

for $j = 1, \ldots, 15$. In order to evaluate the forecast we consider the simplest proper score mentioned in Bracher *et al.* [14]. For the central $(1-\alpha)100\,\%$ prediction interval, the following

interval score is considered:

$$\text{IS}_\alpha(F, y) = (u - l) + \frac{2}{\alpha} \times (l - y) \times \mathbf{1}(y < l) + \frac{2}{\alpha} \times (y - u) \times \mathbf{1}(y > u). \tag{20}$$

Here, F represents the predictive distribution, y the observed data, l and u denote, respectively, the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of F and $\mathbf{1}$ means the indicator function.

Additionally, this score is complemented with the values of the percentages of coverage of the central $(1 - \alpha)100\,\%$ prediction interval.

3. Results

This section presents the main results obtained after applying our methodology. We show data assimilation results in the first 22 weeks of reported cases in each EU country, as well as the disentangling of SARS-CoV-2 virus infectivity (linked to the contact rate) from human behavior (modeled by the effective population fraction). Additional figures are shown in the Supplementary material. We also discuss the appropriateness of the imperfect model for our interests and present a short-term forecast (15 days) that is appropriately evaluated.

Data assimilation. Figure 2 shows the high-quality data assimilation achieved with our modeling. For each country and day of the data assimilation period, the difference between the reported datum and the median of the predictive posterior distribution is plotted, scaled by the maximum of the incidence data for each country.

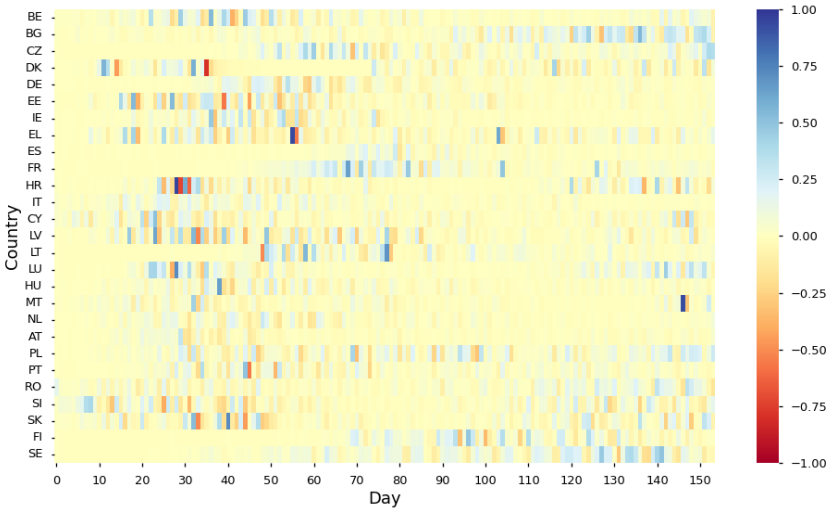


FIGURE 2. **Data assimilation for the 27 countries of the European Union.** The score depicted in this figure is given by the residue of the median of the model and the data, scaled with respect to the maximum of the data for each country.

It can be seen that this difference is near zero in practically all countries along the 154 days. The results for Spain (ES), Italy (IT), Netherlands (NL), Austria (AT) and Romania (RO) are excellent. In the case of Malta (MT), there is always a very good assimilation of data except on day 146. The most significant discrepancies occur when the quality of the

reported data is low, with significant oscillations in consecutive days, or null data, or inclusive negative, as can be seen by comparing with Figure 1 (for example, Denmark (DK) in the day 32, Greece (EL) around day 55, Croatia (HR) near day 30 or Slovakia (SK) in day 32).

On the other hand, Figure 3 shows the results for three representative countries, selected in function of their population size: a country with a large population (Spain), one with a medium population (Belgium), and a country with a small number of inhabitants (Luxembourg).

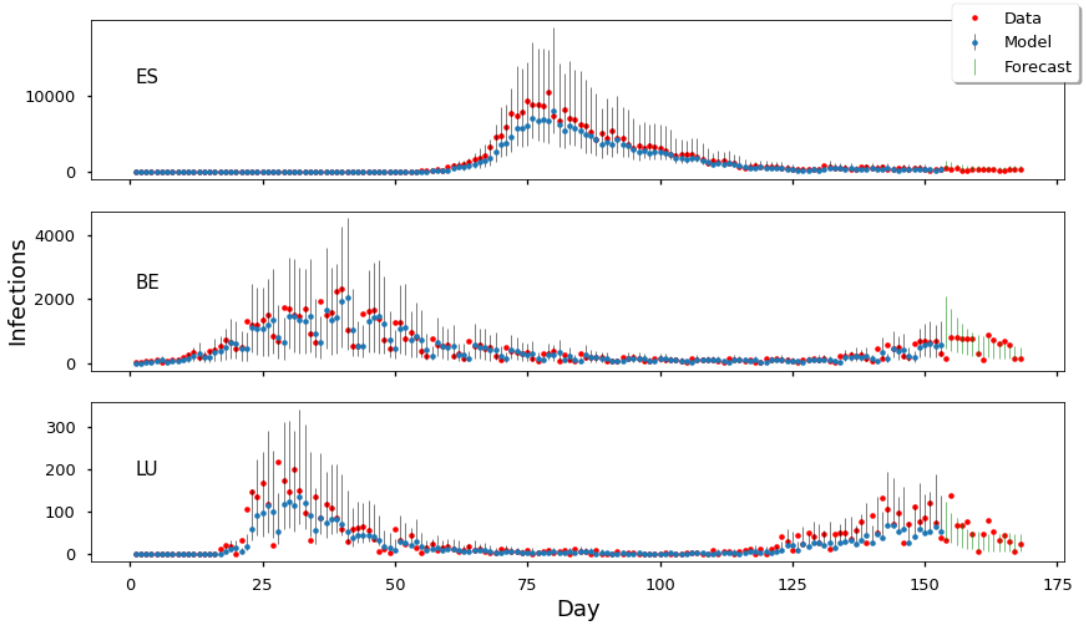


FIGURE 3. **Data assimilation and forecast: Spain, Luxembourg and Belgium.** The first 154 items correspond to the data assimilation. The last 15 items represent the obtained forecast. Red dots represent data, blue dots represent the medians of the posterior distribution and vertical bars (black for the assimilation, green for the forecast) show the quantile range $[0.05, 0.95]$.

The assimilated data correspond to the first 154 days of the graph. The excellent quality of the approximation obtained can be observed for Spain, where the reported data have a smooth behavior. We highlight the closeness of the medians (blue dots) of the distribution to the real data (red dots). On the other hand, although Belgium and Luxembourg have one and two zones, respectively, with widely dispersed data, the medians are very close to the data and only a few data fall outside the quantile range $[0.05, 0.95]$.

Contact rate β. Figure 4 shows the posterior distribution of parameter β for each country using a standard box plot.

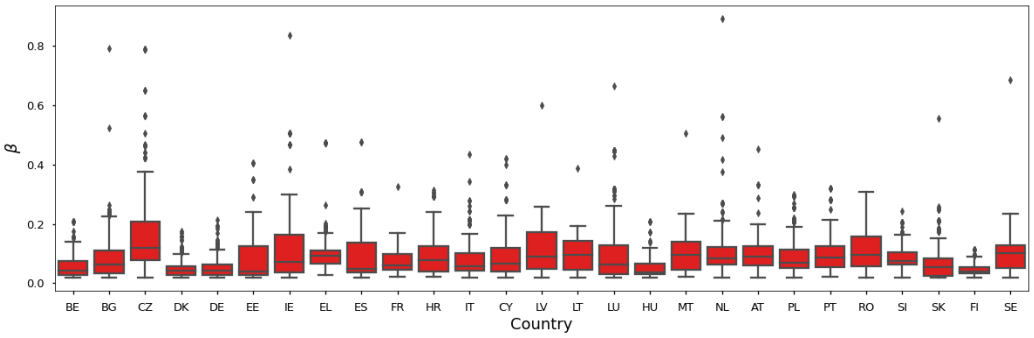


FIGURE 4. **Posterior distributions of the contact rates**, represented by using standard box plots (box limits: Q1 and Q3; maximum length of each whisker is chosen as one and a half times the corresponding interquartile range).

The values of β remain within a similar range for all countries, being the interquartile range lower than 0.2 and the median values near 0.1. This uniformity of the values of the contact rate shows how this parameter is characteristic of the disease. Therefore, it is parameter ω the one that models the effect of human behavior and social containment measures.

Effective population and observation probabilities. Figure 5 shows box plots of the posterior distributions of the effective population fractions and infection observation probabilities.

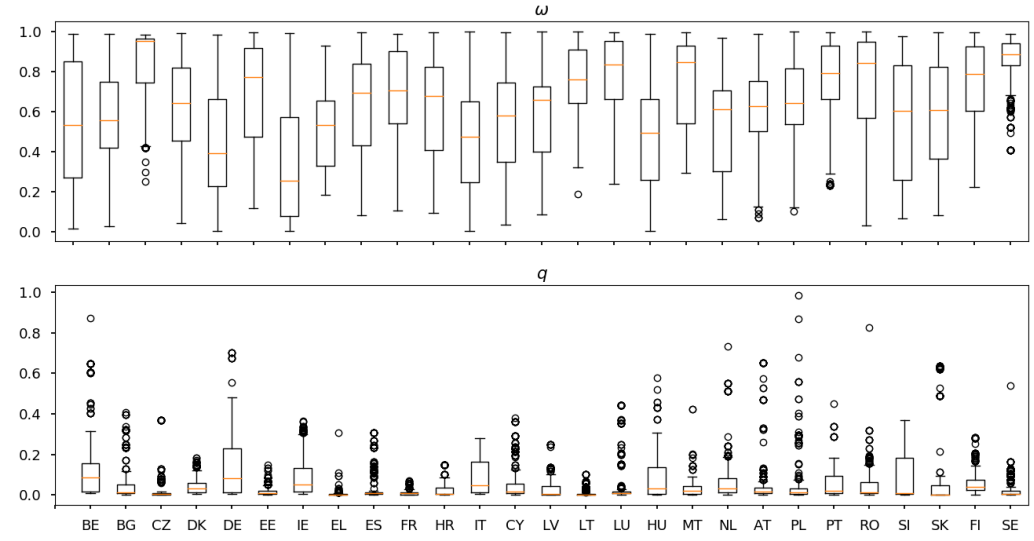


FIGURE 5. **Effective population fractions ω (top) and infection observation probabilities q (bottom)**, represented by using standard box plots. We remark that unlike the contact rate, the effective population fraction and the infection observation probability are not uniform across the analyzed countries.

Both parameters take values in the interval $[0, 1]$. It can be seen that there is a great variety of distributions, depending on the chosen country. The diversity of the distributions of ω (Figure 5), top) captures the different behavior of the population and the different containment measures adopted in the different countries. In the same way, it can be seen (Figure 5, bottom) how the heterogeneous distributions of the observation probability q account for the diverse detection capacity of the infected population by each health system. The low probabilities observed are compatible with a possible deficit in the detection systems' functioning in the pandemic's early phase.

Model adequacy. As stated before, our approximation uses the same imperfect first principles model (a simple SEIR type model) for all countries. The fact that we consider dynamical model uncertainty explicitly allows us to obtain robust results despite the imperfect model. Figure 6 shows, in decimal logarithmic scale, box plots of the posterior distribution of parameters θ_Σ of the correlation matrix.

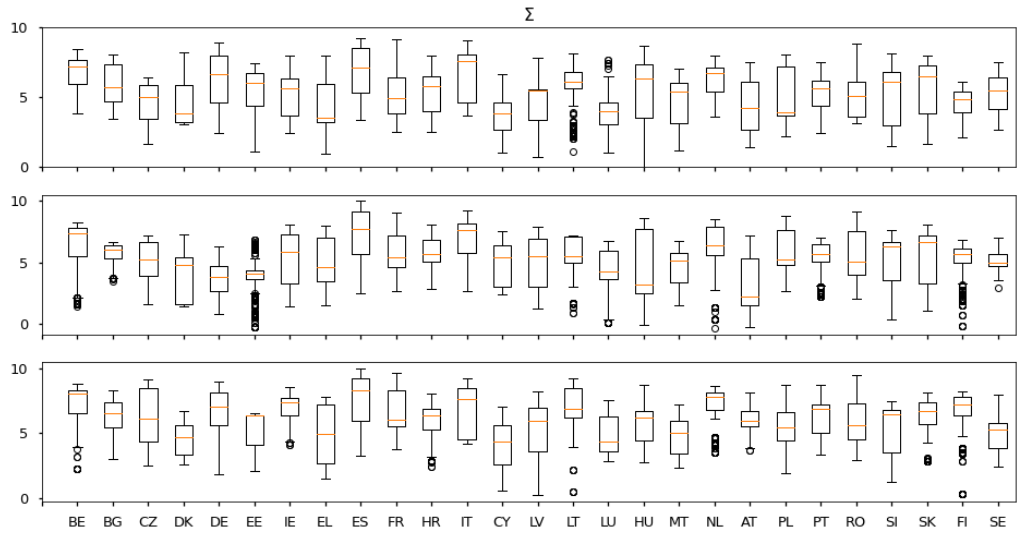
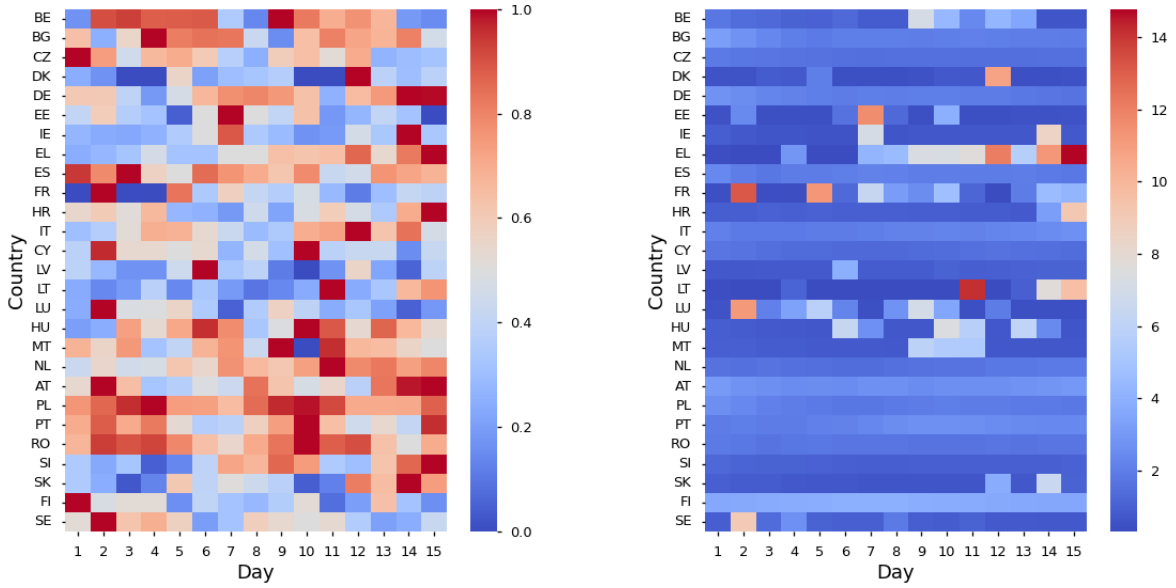


FIGURE 6. **Model adequacy.** The model noise parameters, represented here by standard box plots on a logarithmic scale, show that the model adequacy varies across several orders of magnitude over the analyzed countries. This fact highlights how modeling explicitly the different sources of uncertainty allows us to focus on the analysis of the contact rate, effective population fraction, and observation probability using an imperfect dynamical model.

Significant variability of behaviors between countries and the parameters corresponding to the same country can be appreciated, as well as values of great magnitude. Despite this, the modeling achieves good data assimilation and, as shown below, adequate forecasting.

Forecast evaluation. Finally, we present the forecast results for the 15 days following the assimilation period. Figure 7a -analogous to Figure 1- represents the daily infection notification records in the 27 countries of the European Union for these 15 days, scaled with respect to its maximum value. Figure 7b shows the proper score (20) for the forecast. In order to

relativize with respect to the size of the different populations in each country, this score is scaled with respect to the maximum value of the records of each country in the considered 15 days. We have chosen a central prediction interval of 90%, corresponding to $\alpha = 0.1$ in (20). We note that the lower is the value of the score, the better is the forecast.



(a) Forecast data

(b) Forecast evaluation

FIGURE 7. **Forecast for the 15 days following the assimilation period**. Scaled daily infection notification records (a) and scaled score (20) for the forecast (b).

The quality of the forecast is high in most countries. The score is adequate and uniform for all days in countries such as Bulgaria (BG), Czechia (CZ), Spain (ES), Italy (IT), Cyprus (CY), Netherlands (NL), Romania (RO) and Slovenia (SI). In countries such as Ireland (IE), Croatia (HR), Latvia (LV) and Slovakia (SK) the forecast is satisfactory except for one or two days. However, high score values are obtained on the twelfth day in Denmark (DK), the seventh day in Estonia (EE), the last day in Greece (EL), the second and fifth days in France (FR), the eleventh day in Lithuania (LT) and the second day in Luxembourg (LU). The reason for this poor performance may lie in the fact that the data for each of these days are far from those corresponding to nearby days, as seen in Figure 7a.

Country	BE	BG	CZ	DK	DE	EE	IE	EL	ES
Mean (15 days)	1888.2	451	314.4	503.5	1435.5	37.7	317.9	1240.6	946.2
Coverage (%)	60	100	100	60	100	60	86.7	33.3	100
Country	FR	HR	IT	CY	LV	LT	LU	HU	MT
Mean (15 days)	5901.4	211.4	1173.7	37.7	19.7	51.7	390.2	111.6	125.3
Coverage (%)	20	86.7	100	100	86.7	66.7	40	60	73.3
Country	NL	AT	PL	PT	RO	SI	SK	FI	SE
Mean (15 days)	1273	505.1	1655.3	666.5	2566.8	43	163.2	93.5	876.1
Coverage (%)	100	100	100	100	100	93.3	73.3	100	60

TABLE 2. **Forecast score mean and coverage percentage.** For each country, this table shows the mean of the 15 values of the forecast evaluation and the percentage of coverage of the 15 prediction intervals.

Table 2 shows, for each country, the mean of the forecast evaluations for each of the 15 days. Also, the percentages of coverage of central 90 % prediction intervals are shown. The range of mean values is very wide, since they are linked to the size of the population of each country. This is the reason why in the Figure 7b the values have been scaled, in each country, with respect to the largest data. It is worth noting that the mean values obtained for France (FR), Greece (EL) and Luxembourg (LU) are large when compared to those of countries with similar population sizes. On the other hand, we note that 16 countries achieve more than 85 % of coverage. Only 3 countries (again France, Greece and Luxembourg) obtain scores lower than 50 %.

Finally, regarding the three countries selected for the study in Figure 3, namely Spain, Belgium and Luxembourg, it can be observed an excellent forecast result for Spain and a poorer result for Belgium and Luxembourg, which is related with the oscillation of the data used in the assimilation process, as well as during the forecasting period. Actually, in Table 2 the values of the percentage coverage of this countries are 100 % (ES), 60 % (BE) and 40 % (LU). However, in all three countries, the forecast captures the trend of the actual data, as can be seen in Figure 3.

Code and data availability. The code necessary to reproduce the results of this manuscript is available at the GitHub repository ([47]). The European Union COVID-19 databases are publicly available at the European Centre for Disease Prevention and Control website [2]. COVID-19 data should be downloaded using the program *gather_data.py*. As mentioned above, demographic information about the European Union was obtained from the Eurostat website [32].

4. Conclusions

In this paper, we contribute toward disentangling the roles of virus infectiousness and human behavior in early COVID-19 pandemic dynamics in the European Union. We have posed a state-space model that accounts explicitly for the uncertainty in the data, dynamic model, and parameters. For our inferences, we use the fact that the contact rate, the effective population fraction, and the infection observation probability are structurally identifiable under mild conditions for the proposed SEIR type model. This identifiability result allows us

to assimilate and forecast official records of infection notifications from the 27 countries of the European Union. We argue that our modeling approach allows us to make a fair comparison of the early evolution of the pandemic under different scenarios. These countries had different levels of data reliability, regarding the initial phase of the pandemic. Taking that into account, some countries had better predictions than others, but the overall forecast results are quite good. Indeed, we have found little variability in the contact rate across countries and time. At the same time, there is considerable variability in the effective population fraction and the infection observation probability. Our short-term forecast reflects the quality of the data assimilation performed. We argue that our modeling approach is general and could be used further for hypothesis testing using first principles dynamical models.

Acknowledgements

MAC acknowledges the generous support of Universidad Complutense de Madrid through a visiting professor appointment that made his contributions to this work possible. This work was carried out thanks to the financial support from the Spanish *Ministry of Science and Innovation* under Project PID2019-106337GB-I00.

References

[1] Emma B Hodcroft, Moira Zuber, Sarah Nadeau, Timothy G Vaughan, Katharine HD Crawford, Christian L Althaus, Martina L Reichmuth, John E Bowen, Alexandra C Walls, Davide Corti, et al. Spread of a sars-cov-2 variant through europe in the summer of 2020. *Nature*, 595(7869):707–712, 2021.

[2] Covid-19 european union. Available online: `https://www.ecdc.europa.eu/en/covid-19/data`. (Accessed on 20/07/2022).

[3] B. Ivorra, M.R. Ferrández, M. Vela-Pérez, and A.M. Ramos. Mathematical modeling of the spread of the coronavirus disease 2019 (covid-19) taking into account the undetected infections. the case of china. *Communications in Nonlinear Science and Numerical Simulation*, 88:105303, 2020, doi:https://doi.org/10.1016/j.cnsns.2020.105303.

[4] A.M. Ramos, M.R. Ferrández, M. Vela-Pérez, A.B. Kubik, and B. Ivorra. A simple but complex enough θ-sir type model to be used with covid-19 real data. application to the case of italy. *Physica D: Nonlinear Phenomena*, 421:132839, 2021, doi:https://doi.org/10.1016/j.physd.2020.132839.

[5] A.M. Ramos, M. Vela-Pérez, M.R. Ferrández, A.B. Kubik, and B. Ivorra. Modeling the impact of sars-cov-2 variants and vaccines on the spread of covid-19. *Communications in Nonlinear Science and Numerical Simulation*, 102:105937, 2021, doi:https://doi.org/10.1016/j.cnsns.2021.105937.

[6] Leonardo Villani, Martin McKee, Fidelia Cascini, Walter Ricciardi, and Stefania Boccia. Comparison of deaths rates for covid-19 across europe during the first wave of the covid-19 pandemic. *Frontiers in public health*, 8:620416, 2020.

[7] David García-García, María Isabel Vigo, Eva S Fonfría, Zaida Herrador, Miriam Navarro, and Cesar Bordehore. Retrospective methodology to estimate daily infections from deaths (remedid) in covid-19: the spain case study. *Scientific reports*, 11(1):1–15, 2021.

[8] Zachary Desson, Lisa Lambertz, Jan Willem Peters, Michelle Falkenbach, and Lukas Kauer. Europe's covid-19 outliers: German, austrian and swiss policy responses during the early stages of the 2020 pandemic. *Health policy and technology*, 9(4):405–418, 2020.

[9] María-José Mendoza-Jiménez, Tessa-Virginia Hannemann, and Josefine Atzendorf. Behavioral risk factors and adherence to preventive measures: evidence from the early stages of the covid-19 pandemic. *Frontiers in Public Health*, 9:674597, 2021.

[10] Thomas McAndrew and Nicholas G Reich. An expert judgment model to predict early stages of the covid-19 pandemic in the united states. *PLoS Computational Biology*, 18(9):e1010485, 2022.

[11] Andrea L Bertozzi, Elisa Franco, George Mohler, Martin B Short, and Daniel Sledge. The challenges of modeling and forecasting the spread of covid-19. *Proceedings of the National Academy of Sciences*, 117(29):16732–16738, 2020.

[12] Cleo Anastassopoulou, Lucia Russo, Athanasios Tsakris, and Constantinos Siettos. Data-based analysis, modelling and forecasting of the covid-19 outbreak. *PloS one*, 15(3):e0230405, 2020.
[13] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.
[14] Johannes Bracher, Evan L Ray, Tilmann Gneiting, and Nicholas G Reich. Evaluating epidemic forecasts in an interval format. *PLoS computational biology*, 17(2):e1008618, 2021.
[15] Ralf Engbert, Maximilian M Rabe, Reinhold Kliegl, and Sebastian Reich. Sequential data assimilation of the stochastic seir epidemic model for regional covid-19 dynamics. *Bulletin of mathematical biology*, 83(1):1–16, 2021.
[16] María L Daza-Torres, Marcos A Capistrán, Antonio Capella, and J Andrés Christen. Bayesian sequential data assimilation for covid-19 forecasting. *Epidemics*, page 100564, 2022.
[17] Geir Evensen, Javier Amezcua, Marc Bocquet, Alberto Carrassi, Alban Farchi, Alison Fowler, Pieter L Houtekamer, Christopher K Jones, Rafael J de Moraes, Manuel Pulido, et al. An international initiative of predicting the sars-cov-2 pandemic using ensemble data assimilation. *Foundations of Data Science*, 3(3), 2021.
[18] Katia Koelle and Mercedes Pascual. Disentangling extrinsic from intrinsic factors in disease dynamics: a nonlinear time series approach with an application to cholera. *The American Naturalist*, 163(6):901–913, 2004.
[19] Joshua S Weitz, Sang Woo Park, Ceyhun Eksin, and Jonathan Dushoff. Awareness-driven behavior changes can shift the shape of epidemics away from peaks and toward plateaus, shoulders, and oscillations. *Proceedings of the National Academy of Sciences*, 117(51):32764–32771, 2020.
[20] Ceyhun Eksin, Keith Paarporn, and Joshua S Weitz. Systematic biases in disease forecasting–the role of behavior change. *Epidemics*, 27:96–105, 2019.
[21] Marcos A Capistrán, J Andrés Christen, and Jorge X Velasco-Hernández. Towards uncertainty quantification and inference in the stochastic sir epidemic model. *Mathematical biosciences*, 240(2):250–259, 2012.
[22] Marisa C Eisenberg, Joseph NS Eisenberg, Jeremy P D'Silva, Eden V Wells, Sarah Cherng, Yu-Han Kao, and Rafael Meza. Forecasting and uncertainty in modeling the 2014-2015 ebola epidemic in west africa. *arXiv preprint arXiv:1501.05555*, 2015.
[23] Jason Asher. Forecasting ebola with a regression transmission model. *Epidemics*, 22:50–55, 2018.
[24] Kokouvi M Gamado, George Streftaris, and Stan Zachary. Modelling under-reporting in epidemics. *Journal of mathematical biology*, 69(3):737–765, 2014.
[25] Pierre Magal and Glenn Webb. The parameter identification problem for sir epidemic models: identifying unreported cases. *Journal of mathematical biology*, 77(6):1629–1648, 2018.
[26] Katharine Sherratt, Hugo Gruson, Rok Grah, Helen Johnson, Rene Niehus, Bastian Prasse, Frank Sandman, Jannik Deuschel, Daniel Wolffram, Sam Abbott, et al. Predictive performance of multi-model ensemble forecasts of covid-19 across european nations. *medRxiv*, pages 2022–06, 2022.
[27] Evan L Ray, Nutcha Wattanachit, Jarad Niemi, Abdul Hannan Kanji, Katie House, Estee Y Cramer, Johannes Bracher, Andrew Zheng, Teresa K Yamana, Xinyue Xiong, et al. Ensemble forecasts of coronavirus disease 2019 (covid-19) in the us. *MedRXiv*, pages 2020–08, 2020.
[28] Nicholas Galioto and Alex Arkady Gorodetsky. Bayesian system id: optimal management of parameter, model, and measurement uncertainty. *Nonlinear Dynamics*, 102(1):241–267, 2020.
[29] Dan Simon. *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*. John Wiley & Sons, 2006.
[30] Leah Mitchell and Andrea Arnold. Analyzing the effects of observation function selection in ensemble kalman filtering for epidemic models. *Mathematical biosciences*, 339:108655, 2021.
[31] European Interinstitutional Style Guide. Consulted on April of 2022, `https://publications.europa.eu/code/en/en-370100.htm`.
[32] Eurostat. Life expectancy. Data retrieved on April 15$^{\text{th}}$ of 2022 from Eurostat, `https://ec.europa.eu/eurostat/web/main/home`.
[33] Muge Cevik, Krutika Kuppalli, Jason Kindrachuk, and Malik Peiris. Virology, transmission, and pathogenesis of sars-cov-2. *BMJ*, 371, 2020, doi:10.1136/bmj.m3862.
[34] Hiroshi Nishiura, Natalie M Linton, and Andrei R Akhmetzhanov. Serial interval of novel coronavirus (covid-19) infections. *International journal of infectious diseases*, 93:284–286, 2020.

[35] Ron Sender, Yinon Bar-On, Sang Woo Park, Elad Noor, Jonathan Dushoff, and Ron Milo. The unmitigated profile of covid-19 infectiousness. *Elife*, 11:e79134, 2022.
[36] Hongyu Miao, Xiaohua Xia, Alan S Perelson, and Hulin Wu. On identifiability of nonlinear ode models and applications in viral dynamics. *SIAM review*, 53(1):3–39, 2011.
[37] Giuseppina Bellu, Maria Pia Saccomani, Stefania Audoly, and Leontina D'Angiò. Daisy: A new software tool to test global identifiability of biological and physiological systems. *Computer methods and programs in biomedicine*, 88(1):52–61, 2007.
[38] Necibe Tuncer and Trang T Le. Structural and practical identifiability analysis of outbreak models. *Mathematical biosciences*, 299:1–18, 2018.
[39] scipy.org. Consulted on April of 2022, `https://docs.scipy.org/doc/scipy/reference/generated/scipy.integrate.solve_ivp.html`.
[40] Marcos A Capistrán, J Andrés Christen, and Sophie Donnet. Bayesian analysis of odes: solver optimal accuracy and bayes factors. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):829–849, 2016.
[41] J Andrés Christen and José Luis Pérez-Garmendia. Weak and tv consistency in bayesian uncertainty quantification using disintegration. *Boletín de la Sociedad Matemática Mexicana*, 27(1):1–23, 2021.
[42] Simo Särkkä. *Bayesian Filtering and Smoothing*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2013.
[43] J. A. Christen, C. Fox, et al. A general purpose sampling algorithm for continuous distributions (the t-walk). *Bayesian Analysis*, 5(2):263–281, 2010.
[44] Sang Woo Park, Benjamin M Bolker, David Champredon, David JD Earn, Michael Li, Joshua S Weitz, Bryan T Grenfell, and Jonathan Dushoff. Reconciling early-outbreak estimates of the basic reproductive number and its uncertainty: framework and applications to the novel coronavirus (sars-cov-2) outbreak. *Journal of the Royal Society Interface*, 17(168):20200144, 2020.
[45] Fred Brauer, Carlos Castillo-Chavez, and Zhilan Feng. *Mathematical models in epidemiology*, volume 32. Springer, 2019.
[46] Dave Osthus, Kyle S Hickmann, Petruţa C Caragea, Dave Higdon, and Sara Y Del Valle. Forecasting seasonal influenza with a state-space sir model. *The annals of applied statistics*, 11(1):202, 2017.
[47] Marcos A. Capistrán, Juan-Antonio Infante, Ángel M. Ramos, and José M. Rey. Disentangling the role of virus infectiousness and human behavior method. `https://github.com/MarcosACapistran/covid_eu`, 2022.

(Marcos A. Capistrán) Centro de Investigación en Matemáticas (CIMAT), Jalisco S/N, Valenciana, Guanajuato, 36023, México

Email address, Marcos A. Capistrán: `marcos@cimat.mx`

(Juan-Antonio Infante, Ángel M. Ramos and José María Rey) Instituto de Matemática Interdisciplinar and Departamento de Análisis Matemático y Matemática Aplicada, Facultad de CC. Matemáticas, Universidad Complutense de Madrid, Plaza de Ciencias 3, 28040, Madrid, Spain

Email address, Juan-Antonio Infante: `infante@mat.ucm.es`

Email address, Ángel M. Ramos: `angel@mat.ucm.es`

Email address, José M. Rey: `jrey@mat.ucm.es`

Algorithm 1: Unscented Kalman filtering algorithm for approximating the posterior distribution $\pi_{\Theta|Y}(\theta|\mathscr{Y}_n)$ with truncated Gaussian distributions.

Input:
- System parameters $\theta = (\theta_\Psi, \theta_h, \theta_\Sigma, \theta_\Gamma)$.
- Prior distribution $\pi_\Theta(\theta)$ (see **Prior elicitation**).
- Distribution on initial conditions $m_0 = (E_0, I_0, 0)^T$ (2), $P_0 = 10I_3$.
- SEIR model parametrization $\Psi(\theta_\Psi)$ (1).
- Observation model parametrization $h(\theta_h)$ (3).
- Covariance matrices $\Sigma(\theta_\Sigma)$ and $\Gamma(\theta_\Gamma)$ (7)
- UKF parameters α, κ, β, λ, d as in [28].
- Weights W_i^m, W_i^c, for $i = 0, ..., 2d$ as in [28].

Output:
- Approximate evaluation of the marginal posterior distribution $\pi_{\Theta|Y}(\theta|\mathscr{Y}_n)$ (14).

Step 1. Compute the prior distribution $\pi(\theta|\mathscr{Y}_0) = \pi_\Theta(\theta)$

for $k = 1$ to n **do**

Step 2. Predict $\pi(X_k|\theta, \mathscr{Y}_{k-1}) \approx \mathcal{N}(m_k^-, P_k^-)$

- Form the sigma points

$$\begin{aligned}\mathscr{X}_{k-1}^0 &= m_{k-1}\\ \mathscr{X}_{k-1}^i &= m_{k-1} + \sqrt{d+\lambda}\left[\sqrt{P_{k-1}}\right]_i\\ \mathscr{X}_{k-1}^{i+d} &= m_{k-1} - \sqrt{d+\lambda}\left[\sqrt{P_{k-1}}\right]_i, \quad \text{for } i = 1, ..., d\end{aligned}$$

- Propagate the sigma points through the dynamical model (6)

$$\hat{\mathscr{X}}_k^i(\theta) = \Psi(\mathscr{X}_{k-1}^i, \theta_\Psi), \quad \text{for } i = 0, ..., 2d$$

- Compute the mean and covariance

$$\begin{aligned}m_k^-(\theta) &= \sum_{i=0}^{2d} W_i^m \hat{\mathscr{X}}_k^i\\ P_k^-(\theta) &= \sum_{i=0}^{2d} W_i^c(\hat{\mathscr{X}}_k^i - m_k^-)(\hat{\mathscr{X}}_k^i - m_k^-)^T + \Sigma(\theta_\Sigma)\end{aligned}$$

Step 3. Compute the evidence $\pi(y_k|\theta, \mathscr{Y}_{k-1}) \approx \mathscr{N}(\mu_k, S_k)$

- Propagate sigma points through observation operator

$$\hat{\mathscr{Y}}_k^i(\theta) = h(\hat{\mathscr{X}}_k^i, \theta_h) \quad \text{for } i = 0, ..., 2d$$

- Compute the mean and covariance

$$\begin{aligned}\mu_k(\theta) &= \sum_{i=0}^{2d} W_i^m \hat{\mathscr{Y}}_k^i\\ S_k(\theta) &= \sum_{i=0}^{2d} W_i^c(\hat{\mathscr{Y}}_k^i - \mu_k)(\hat{\mathscr{Y}}_k^i - \mu_k)^T + \Gamma(\theta_\Gamma)\end{aligned}$$

- If $\mu_k(\theta)$ has any negative component or $S_k(\theta)$ is not positive definite then set $\pi(\theta|\mathscr{Y}_n) = 0$ and break the loop.

Step 4. Update $\pi(X_k|\theta, \mathscr{Y}_k) \approx \mathscr{N}(m_k, P_k)$

$$\begin{aligned}C_k(\theta) &= \sum_{i=0}^{2d} W_i^c(\hat{\mathscr{X}}_k^i - m_k^-)(\hat{\mathscr{Y}}_k^i - \mu_k)^T\\ m_k(\theta) &= m_k^- + (C_k S_k^{-1})(y_k - \mu_k)\\ P_k(\theta) &= P_k^- - (C_k S_k^{-1}) S_k (C_k S_k^{-1})^T\end{aligned}$$

Step 5. Update $\pi(\theta|\mathscr{Y}_k) \propto \pi(y_k|\theta, \mathscr{Y}_{k-1})\pi(\theta|\mathscr{Y}_{k-1})$

end for