Characterizing the COVID-19 dynamics with a new epidemic model: Susceptible-exposed-asymptomatic-symptomatic-active-removed

Grace Y. YI[1,2], Pingbo HU[1], and Wenqing HE[1*]

[1]*Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada N6A 5B7*
[2]*Department of Computer Science, University of Western Ontario, London, Ontario, Canada N6A 5B7*

Key words and phrases: Basic reproduction number; COVID-19; epidemic model; IF-EAKF algorithm; ordinary differential equations; transmission.

MSC 2020: Primary 92B15; secondary 62P10.

Abstract: The coronavirus disease 2019 (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has spread stealthily and presented a tremendous threat to the public. It is important to investigate the transmission dynamics of COVID-19 to help understand the impact of the disease on public health and the economy. In this article, we develop a new epidemic model that utilizes a set of ordinary differential equations with unknown parameters to delineate the transmission process of COVID-19. The model accounts for asymptomatic infections as well as the lag between symptom onset and the confirmation date of infection. To reflect the transmission potential of an infected case, we derive the *basic reproduction number* from the proposed model. Using the daily reported number of confirmed cases, we describe an estimation procedure for the model parameters, which involves adapting the *iterated filter-ensemble adjustment Kalman filter* (IF-EAKF) algorithm. To illustrate the use of the proposed model, we examine the COVID-19 data from Quebec for the period from 2 April 2020 to 10 May 2020 and carry out sensitivity studies under a variety of assumptions. Simulation studies are used to evaluate the performance of the proposed model under a variety of settings. *The Canadian Journal of Statistics* 50: 395–416; 2022 © 2022 Statistical Society of Canada

Résumé: La maladie à coronavirus 2019 (COVID-19), causée par le coronavirus 2 du syndrome respiratoire aigu sévère (SARS-CoV-2), s'est rapidement propagée et représente une grande menace pour le public. Pour mieux comprendre l'impact de cette maladie sur la santé publique et l'économie, il est important d'étudier la dynamique de sa transmission. A cette fin, les auteurs de cet article proposent un nouveau modèle épidémiologique basé sur un ensemble d'équations différentielles ordinaires avec des paramètres inconnus et qui tient compte des infections asymptomatiques ainsi que du décalage entre l'apparition des symptômes et la date de confirmation de l'infection. Ils en déduisent le *taux de reproduction de base* qui traduit le potentiel de transmission d'un cas infecté. En utilisant le nombre rapporté de cas confirmés, les auteurs décrivent une procédure d'estimation des paramètres du modèle qui repose sur une adaptation de l'algorithme *filtre itéré - filtre de Kalman énsemble àjustement* (IF-EAKF). Une mise en application du modèle proposé est illustrée à travers l'examen des données COVID-19 du Québec pour la période du 2 avril 2020 au 10 mai 2020. Une analyse de sensibilité du modèle construit est explorée sous diverses

Additional Supporting Information may be found in the online version of this article at the publisher's website.
* *Corresponding author: whe@stats.uwo.ca*

© 2022 Statistical Society of Canada

hypothèses. Enfin, les auteurs ont fait appel à des études de simulation pour évaluer la performance du modèle proposé et ce sous différents scénarios. *La revue canadienne de statistique* 50: 395–416; 2022 © 2022 Société statistique du Canada

1. INTRODUCTION

The coronavirus disease 2019 (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has spread stealthily and represented a tremendous threat to worldwide public health. To help understand the virus transmissions, it is imperative to investigate the transmission dynamics quantitatively. In the literature, a variety of epidemic models have been developed to study various types of infectious diseases, including the *susceptible-infectious-recovered* (SIR) model, the *susceptible-infectious-susceptible* (SIS) model, the *susceptible-exposed-infectious-recovered* (SEIR) model, the Reed–Frost model, and their variants. A review of epidemic models can be found in Duan et al. (2015).

Applications of those epidemic models have been extensive. To name a few, Osthus et al. (2017) used the SIR model to forecast seasonal influenza. Shah & Gupta (2013) applied the SEIR model to examine the transmission processes of vector-borne diseases. Ng & Orav (1990) proposed a generalized Reed–Frost model to predict human immunodeficiency virus (HIV) incidence in San Francisco's homosexual population, and Ng, Turinici & Danchin (2003) developed a modified SEIR model, called the *susceptible-exposed-infectious-recovered-protection* (SEIRP) model, to study the outbreak of the *severe acute respiratory syndrome* (SARS) in China.

Recently, a number of new models have been explored to study the dynamics of COVID-19. For example, Tang et al. (2020) proposed a generalized SEIR model to incorporate presymptomatic COVID-19 cases and study the implications of the intervention measures such as contact tracing, quarantine, and isolation in China. Tuite, Fisman & Greer (2020) generalized the SEIR model by incorporating the information on interventions and severities of clinical symptoms to examine the potential impact of case-based and noncase-based nonpharmaceutical interventions for the population of Ontario, Canada. Mandal et al. (2021) proposed a deterministic model by stratifying the population into three age groups and incorporating asymptomatic cases to explore various strategies for lifting lockdowns in India. Also, the IHME COVID-19 Forecasting Team (2021) used the SEIR model to characterize the trajectories of COVID-19 infections and the effects of nonpharmaceutical interventions in the United States.

In this article, we propose a new epidemic model, called the *susceptible-exposed-asymptomatic-symptomatic-active-removed* (SEASAR) model, to delineate the COVID-19 transmission dynamics. We describe the target population by stratifying it into six subpopulations, consisting of individuals who are, respectively, susceptible, exposed, asymptomatic, symptomatic, active, and removed. This model generalizes the SIR and SEIR models by accounting for asymptomatic infections and the lag between symptom onset and the diagnosis time. Consistent with the SIR and SEIR models, we make two routine assumptions: (1) the population is homogeneous and (2) there are no inbound and outbound travels. Similar to the SIR and SEIR models, the SEASAR model is a deterministic model that utilizes ordinary differential equations to describe the transmission dynamics of COVID-19. We derive the basic reproduction number from the proposed model to provide a scalar measure of the pandemic.

To implement the proposed model, we develop an estimation procedure for the model parameters by adapting the iterated filter-ensemble adjustment Kalman filter (IF-EAKF) algorithm (e.g., Ionides, Bretó & King, 2006), where sampling from Bayesian posterior distributions is employed. We illustrate the use of the proposed model by analyzing the COVID-19 data from Quebec for the period from 2 April 2020 to 10 May 2020. We conduct sensitivity analyses to assess how the estimation of the model parameters and the predicted results may change as the model assumptions are altered. We compare the analysis results by applying the proposed method in contrast to the SIR and SEIR models as well as a neural network model. Simulation studies provide a basis for assessing the model performance under different settings.

The remainder of this article is organized as follows. We introduce the deterministic SEASAR model and elaborate on its rationale in Section 2. In Section 3, we present the stochastic model for the observed data and establish its connection with the SEASAR model, together with the initialization setup. Section 4 describes the estimation procedure by adapting the IF-EAKF algorithm. In Section 5, we utilize the proposed SEASAR model to analyze the Quebec COVID-19 data for the period from 2 April 2020 to 10 May 2020, and Section 6 reports our simulation studies. The article concludes with a discussion of some outstanding issues.

2. MODEL FRAMEWORK

Via a meta-analysis, He, Yi & Zhu (2020) estimated that about 46% of individuals with COVID-19 are asymptomatic. Because of the incubation period, there is a time lag between symptom onset and being confirmed as an infected individual. To incorporate these features of COVID-19, we develop the SEASAR model to be described as follows.

2.1. Illustration of the Proposed Model

To illustrate the ideas, we first consider a *static* framework by focusing on a given time point. We divide the target population into six subpopulations with specific features, denoted by S, E, I_a, I_s, A, and R, respectively. Specifically, S represents the subpopulation of *susceptible* cases (i.e., those at risk of becoming infected with the novel coronavirus), E is the subpopulation of *exposed* cases (i.e., those who are infected but do not have the infectious ability yet and are still in the latent period) (e.g., Bai et al., 2020), I_a stands for the subpopulation of *asymptomatic* infections (i.e., those cases who are infectious but exhibit no symptoms), I_s represents the subpopulation of *symptomatic* infections (i.e., those cases who exhibit symptoms and are infectious, but who are not yet confirmed), A is the subpopulation of *active* cases (i.e., those confirmed cases who have either not recovered or died), and R denotes the subpopulation of *removed* cases (i.e., those confirmed cases who have recovered or died from COVID-19).

Next, we introduce parameters to facilitate dynamic changes among the subpopulations. Let Z denote the *average latent period*, defined as the average time (in days, say) from being infected to having the infectious ability. Various studies have been conducted to estimate the value of Z (e.g., Bai et al., 2020; Guan et al., 2020), so here we take Z as being available. Let θ denote the *average symptomatic transmission rate*, defined as the average number of individuals infected by a symptomatic case per unit time. Let the *average asymptomatic transmission rate* be denoted by $\mu\theta$, defined as the average number of individuals infected by an asymptomatic case per unit time. As asymptomatic infections are regarded as less infectious than symptomatic cases (e.g., Li et al., 2020), μ is a constant between 0 and 1. Let α denote the average fraction of *symptomatic* infections relative to all infections, let β denote the average rate for *asymptomatic* infections to develop symptoms per unit time, and let γ denote the average recovery rate of *asymptomatic* infections per unit time. Let F stand for the average time (in days, say) from symptom onset to the time of being a confirmed case, let B denote the average time (in days, say) from being a confirmed case to having recovered, and let J represent the average time (in days, say) from being a confirmed case to death due to COVID-19.

Figure 1 is a flowchart showing the relationship among the six subpopulations. A black solid arrow between two subpopulations indicates that the members of one subpopulation can transition into the other subpopulation; a red dashed arrow between two subpopulations indicates that members in one subpopulation can be infected by members in the other subpopulation. The parameters on black solid lines determine the number of people who move from one subpopulation to another subpopulation per unit time, and the corresponding parameters on red

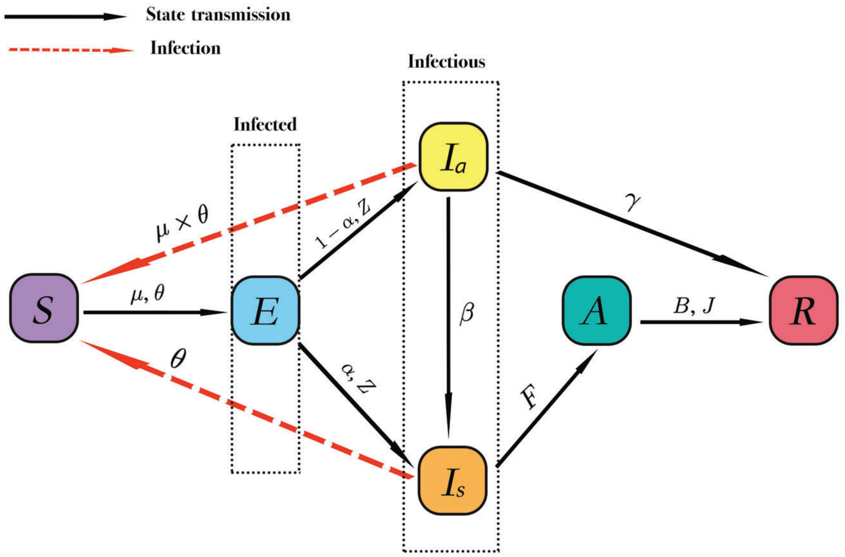

FIGURE 1: Illustration of the SEASAR model. The population is divided into six compartments: S (susceptible), E (exposed), I_a (asymptomatic), I_s (symptomatic), A (active), and R (removed).

dashed lines determine the number of people infected by asymptomatic or symptomatic cases per unit time. We assume that we have a homogeneous population and that any confirmed COVID-19 case must be quarantined immediately and cannot infect other cases thereafter. Thus, there is no transition from A to other compartments except R. Further, we assume that asymptomatic individuals are not tested for COVID-19, whereas all symptomatic cases are assumed to be confirmed at a certain time. The former assumption is fairly reasonable, especially in the early stage of the pandemic when test kits are scarce. However, the latter assumption is less realistic, because in reality, some symptomatic cases may never be confirmed because of false negative results or not being tested. These two assumptions basically consider the setting where an initially asymptomatic individual can move into compartment A only if this person shows symptoms before recovery or death, and those asymptomatic cases who never show symptoms cannot directly enter compartment A.

2.2. Deterministic Dynamic Model

Figure 1 shows a static chart for the transitions among the six subpopulations at a given time point. However, for any time period, the transitions are not static but dynamic. To characterize this *dynamic* feature, we modify the six subpopulations discussed in Section 2.1 by showing their dependence on time t, and let $S^*(t)$, $E^*(t)$, $I_a^*(t)$, $I_s^*(t)$, $A^*(t)$, and $R^*(t)$ denote the respective sizes of the corresponding six subpopulations (or state compartments) at time t. We assume that the size of each state compartment at t follows a certain distribution with the mean, denoted by the same symbol with the asterisk removed (e.g., $S^*(t) \sim \text{Poisson}(S(t))$ for $t > 0$). While those sizes are treated as random, here we focus on delineating the dynamic changes in their means to reflect the underlying links.

To be specific, let $\phi(t) = (S(t), E(t), I_a(t), I_s(t), A(t), R(t))^{\mathrm{T}}$ denote the vector of the six average subpopulation sizes at time t. We represent the dynamic changes in $\phi(t)$ via the ordinary differential equations (ODEs):

$$\frac{dS(t)}{dt} = -\frac{\theta S(t) I_s(t)}{N} - \frac{\mu \theta S(t) I_a(t)}{N}; \tag{1}$$

$$\frac{dE(t)}{dt} = \frac{\theta S(t) I_s(t)}{N} + \frac{\mu \theta S(t) I_a(t)}{N} - \frac{E(t)}{Z}; \tag{2}$$

$$\frac{dI_a(t)}{dt} = (1-\alpha)\frac{E(t)}{Z} - \beta I_a(t) - \gamma I_a(t); \tag{3}$$

$$\frac{dI_s(t)}{dt} = \alpha \frac{E(t)}{Z} - \frac{I_s(t)}{F} + \beta I_a(t); \tag{4}$$

$$\frac{dA(t)}{dt} = \frac{I_s(t)}{F} - \frac{A(t)}{B} - \frac{A(t)}{J}; \tag{5}$$

$$\frac{dR(t)}{dt} = \gamma I_a(t) + \frac{A(t)}{B} + \frac{A(t)}{J}; \tag{6}$$

where, under the assumption of no immigration or emigration of individuals, N represents the time-invariant total size $N \triangleq S(t) + E(t) + I_a(t) + I_s(t) + A(t) + R(t)$ at any time point t. Any of the Equations (1)–(6) is determined by the other five equations because of the total size constraint. Such a constraint, however, is not applied to the sum of $S^*(t)$, $E^*(t)$, $I_a^*(t)$, $I_s^*(t)$, $A^*(t)$, and $R^*(t)$.

Our reasoning for Equations (1)–(6) may be found in Appendix B of the accompanying Supplementary Material. For ease in referring to Equations (1)–(6), let $\eta = (\theta, \mu, \alpha, \beta, \gamma, F, B, J)^{\mathrm{T}}$ denote the vector of parameters of primary interest; then

$$\frac{d\phi(t)}{dt} = g(\phi(t), \eta),$$

where $g(\cdot, \cdot)$ represents the vector function determined by the right-hand side of Equations (1)–(6), and we call these six equations the SEASAR model. Figure S.1 in the Supplementary Material displays a flowchart of the transmission dynamics for the six subpopulations together with the associated values.

2.3. The Basic Reproduction Number

Knowing the value of η allows us to describe the six subpopulation sizes using Equations (1)–(6). Further, it enables us to describe the severity of the pandemic using a simple measure, the *basic reproduction number*, denoted R_0, which is defined as the expected number of cases infected by one case in a population consisting of individuals susceptible to infection.

A large value of R_0 indicates a severe pandemic. Usually, comparing R_0 to 1 describes the spread of the disease. "$R_0 > 1$" suggests that the infection is spreading in the population, and "$R_0 < 1$" indicates a dying-down situation. In Appendix A of the Supplementary Material, we show that the value of R_0 derived from the SEASAR model equals

$$R_0 = \frac{\theta(F\alpha\gamma + \beta F - \mu\alpha + \mu)}{\beta + \gamma}.$$

3. DATA AND THE STOCHASTIC MODEL

3.1. The Observed Data and the Stochastic Model

For $t > 0$, let $Y(t)$ denote the number of confirmed cases to be reported *on* day t, which we regard as a random variable. We assume that $Y(t)$ follows a normal distribution:

$$Y(t) \sim N\left(\mu_c(t), \sigma_t^2\right) \tag{7}$$

with mean $\mu_c(t)$ and variance σ_t^2. As in Section 2.1, we assume that only symptomatic individuals may be confirmed as cases; thus, by Figure 1, a case that is being confirmed at time t corresponds to a transition from state I_s to state A at t. By the definition of F, $\frac{1}{F}$ can be regarded as the average proportion of symptomatic cases that are confirmed on any given day, and, hence, $I_s(t)/F$ represents the mean number of confirmed cases on day t. Consequently, $\mu_c(t)$ in (7) is given by $I_s(t)/F$, which links $Y(t)$ with the SEASAR model.

Let τ_0 denote the initial time point from which we start examining the data. Since the number of confirmed COVID-19 cases is reported on a daily basis, we take the fundamental unit of time to be a day, and let $\mathcal{T} = \{\tau_0, \tau_0 + 1, \ldots, \tau_0 + T\}$ denote the examination days with T being a specified positive integer.

In the study period $\mathcal{T}$, let $\mathcal{D}_c = \{y_t : t \in \mathcal{T}\}$ represent the observed values for the process $\{Y(t) : t \in \mathcal{T}\}$. Being called the *observational error variance* (OEV) (e.g., Li et al., 2020), σ_t^2 in (7) is often characterized heuristically based on previously observed data $\{y_s \in \mathcal{D}_c : s < t\}$ via an assumed function form. For example, for $t \in \mathcal{T}$, σ_t^2 may be fixed as

$$\sigma_t^2 = \max\left(30, \frac{y_{t-1}}{20}\right), \tag{8}$$

bearing in mind that other specifications to characterize the value of σ_t^2 are also possible.

By time τ_0, let R_c, C_0, and D_0 denote the reported cumulative number of recoveries, of confirmed cases, and of deaths from COVID-19, respectively. For $t \in \mathcal{T}$, let $Q(t)$ denote the number of individuals who report that their COVID-19 symptoms appeared on day t but are not yet confirmed to have COVID-19 on day t; $Q(t)$ is related to $I_s(t)$ in the SEASAR model via $I_s(t) = \sum_{s\le t} Q(s) - \sum_{s\le t} y_s$ for $t \in \mathcal{T}$. Let $\mathcal{D}_b = \{R_c, C_0, D_0\}$ record the available data when the study begins and write $\mathcal{D}_s = \{Q(t) : t \in \mathcal{T}\}$.

In contrast, the unobserved variables $S^*(t)$, $E^*(t)$, $I_a^*(t)$, $I_s^*(t)$, $A^*(t)$, and $R^*(t)$, $\mathcal{D}_b$, $\mathcal{D}_s$, and $\mathcal{D}_c$ represent the observed data that are used to describe an estimation procedure for the model parameters in Equations (1)–(6), where $\mathcal{D}_b$ and $\mathcal{D}_s$ are used to initialize the mean sizes of the six subpopulations, together with the assumptions as outlined in Section 3.2; and $\mathcal{D}_c$ is used to estimate the model parameters η, as described in Section 4.

3.2. Initial Mean Sizes of the Subpopulations

At the beginning point τ_0 of the study, the initial mean sizes of the six subpopulations, $\phi(\tau_0) = (S(\tau_0), E(\tau_0), I_a(\tau_0), I_s(\tau_0), A(\tau_0), R(\tau_0))^{\mathrm{T}}$, are given. Table S.1 in the Supplementary Material summarizes the relationship of the initial mean sizes of the six subpopulations to be used in Section 4.2, with r_1, r_2, and r_3 being defined as in the following.

In contrast to R_c, which is defined in Section 3.1, let R_a denote the cumulative number of recovered asymptomatic cases by time τ_0; however, R_a is unavailable. To facilitate the relationship between the observed and unobserved values, let $r_1 = R_a/R_c$ denote the ratio of the unobserved value R_a to the observed value of R_c, and let $r_2 = I_a(\tau_0)/I_s(\tau_0)$ represent the ratio of the unobserved size $I_a(\tau_0)$ to the observed size $I_s(\tau_0)$. Motivated by Hao et al. (2020), let $r_3 = E(\tau_0)/\left\{\sum_{t=\tau_0}^{\lfloor \tau_0+Z \rfloor} Q(t)\right\}$ denote the ratio of the unobserved $E(\tau_0)$ to the observed data in $\mathcal{D}_s$ over the time window $[\tau_0, \lfloor \tau_0 + Z \rfloor]$, where the function $\lfloor x \rfloor$ represents the largest integer that is less than or equal to x. Notationally, one may write $E(\tau_0) = r_3 \times \sum_{t=\tau_0}^{\lfloor \tau_0+Z \rfloor} Q(t)$, though it is understood that $E(\tau_0)$ and the data in $\mathcal{D}_s$ do not have an intrinsic connection.

The introduction of the ratios r_1, r_2, and r_3 does not enable us to determine the values of unobservable R_a, $I_a(\tau_0)$, and $E(\tau_0)$, but these ratios do offer us an informative way to describe the pandemic situation in relative scales of the observed values by time τ_0. For instance, at the early stage of the pandemic, testing kits are limited, so the number of recoveries from *confirmed* cases is likely to be much smaller than the corresponding number of recoveries by *asymptomatic* individuals, and such a scenario can be informatively described by a large value of r_1. If r_2 is greater than 1, then there are more *asymptomatic* infections than *symptomatic* infections. Despite the lack of information about r_1, r_2, and r_3, one may conduct sensitivity analyses by changing their values to describe different scenarios with different degrees of severity of the pandemic, as we report in Section 5.

4. ESTIMATION PROCEDURE

Here we adapt the IF-EAKF algorithm in combination with the fourth-order Runge–Kutta (RK4) method (e.g., Süli & Mayers, 2003, p. 328) to estimate the SEASAR model parameter η. The basic idea is to embed the original SEASAR model with its *time-invariant* parameter η into an expanded yet "artificial" model with a *time-varying* parameter, say $\eta(t)$, so that the problem of estimating the original parameter η is converted to estimating $\eta(t)$ sequentially, where the enhanced model assumes the same form as we assumed in Equations (1)–(6) with the parameter η replaced by $\eta(t) = (\theta(t), \mu(t), \alpha(t), \beta(t), \gamma(t), F(t), B(t), J(t))^{\mathrm{T}}$. The introduction of the "artificial" time-varying parameter $\eta(t)$ allows us to implement the IF-EAKF algorithm for the estimation of η by sequentially examining the observed data over $\mathcal{T}$ for each $t \in \mathcal{T}$.

Considering this enlarged model offers us the flexibility of using the observed data $\mathcal{D}_c$ *sequentially* to update estimates of the model parameters using the Bayesian principle. For $t > 0$, the time-dependent parameter $\eta(t)$ in the enlarged model is treated as random, and its prior and posterior distributions are updated for the sequential time points in $\mathcal{T}$, as described in Section 4.2. To implement the IF-EAKF algorithm, we specify the prior information concerning the parameters at the study entry point τ_0 in Section 4.1 and then present the implementation procedure in Section 4.2.

4.1. Initialization of the Model Parameters

Here we assume that the prior information concerning the parameters $\eta(\tau_0)$ at the initial time point τ_0 is not informative except for constraining the parameters to certain ranges to reflect our prior knowledge about them.

To be specific, the transmission rate θ of symptomatic infections is such that $0 \le \theta \le 7$ to cover the reported values in the literature, including Hao et al. (2020) and Li et al. (2020). The multiplicative factor μ is assumed to satisfy $0.1 \le \mu \le 1$, the fraction α of symptomatic infections relative to all infections is restricted so that $0.1 \le \alpha \le 1$, the average rate β of asymptomatic infections who develop symptoms per unit time is constrained to be $0.0002 \le \beta \le 0.8$, and the average recovery rate γ of asymptomatic infections per unit time is such that $0.1 \le \gamma \le 1$. The average time F from symptom onset to being confirmed is considered to be between 1 and 10 days (Kramer, 2020), the average time B from being confirmed to recovery is between 7 and 42 days (WHO, 2020), and the average time J from being confirmed as a symptomatic case until death occurs ranges from 14 to 56 days (WHO, 2020).

4.2. An Estimation Algorithm

In this subsection, we describe the detail of adapting the IF-EAKF algorithm to iteratively update the estimates of the parameters of the SEASAR model. Let n be a prespecified integer, which is reasonably large, and let L denote the iteration number. The estimation procedure consists of iterations for $l = 1, \dots, L$ as follows:

First, we describe how to obtain an initial estimate of $\eta(t)$ for $t \in \mathcal{T}$ at iteration $l = 1$, elaborated in the following four parts.

Part 1: ***At iteration*** $l = 1$ ***and at time*** $t = \tau_0$, ***determine prior and posterior values of*** $\eta(t)$ ***and*** $\mu_c(t)$:

- **Stage 1**: Determine prior values for $\eta(\tau_0)$ and $\mu_c(\tau_0)$.

- **Step 1**: Let π_η denote a prior distribution for parameter $\eta(\tau_0)$, which we assume is the uniform distribution over the ranges specified in Section 4.1, with the assumption that the parameter components in $\eta(\tau_0)$ are independent of each other.
- **Step 2**: Generate n values from π_η, denoted $\left\{\eta^i_{\text{pri},\tau_0} : i = 1, \ldots, n\right\}$, where for each i, $\eta^i_{\text{pri},\tau_0} = \left(\theta^i_{\text{pri},\tau_0}, \mu^i_{\text{pri},\tau_0}, \alpha^i_{\text{pri},\tau_0}, \beta^i_{\text{pri},\tau_0}, \gamma^i_{\text{pri},\tau_0}, F^i_{\text{pri},\tau_0}, B^i_{\text{pri},\tau_0}, J^i_{\text{pri},\tau_0}\right)^{\text{T}}$.
- **Step 3**: Using the initial size $I_s(\tau_0)$ of symptomatic infections, we generate n prior values for $\mu_c(\tau_0)$, denoted $\left\{\mu^i_{c,\text{pri}}(\tau_0) : i = 1, \ldots, n\right\}$, by setting $\mu^i_{c,\text{pri}}(\tau_0) = \frac{I_s(\tau_0)}{F^i_{\text{pri},\tau_0}}$ for $i = 1, \ldots, n$. Then calculate the sample mean and variance:

$$\bar{o}_{\text{pri},\tau_0} = \frac{\sum_{i=1}^n \mu^i_{c,\text{pri}}(\tau_0)}{n} \quad \text{and} \quad \sigma^2_{\text{pri},\tau_0} = \frac{\sum_{i=1}^n \left\{\mu^i_{c,\text{pri}}(\tau_0) - \bar{o}_{\text{pri},\tau_0}\right\}^2}{n-1},$$

together with the pairwise sample covariances:

$$\sigma^{cov}_{X,\mu_c(\tau_0),\text{pri}} = \frac{1}{n-1}\sum_{i=1}^n \left\{\mu^i_{c,\text{pri}}(\tau_0) - \frac{\sum_{i=1}^n \mu^i_{c,\text{pri}}(\tau_0)}{n}\right\}\left\{X^i_{\text{pri},\tau_0} - \frac{\sum_{i=1}^n X^i_{\text{pri},\tau_0}}{n}\right\},$$

where X is a symbol for θ, μ, α, β, γ, F, B, and J.

- **Stage 2**: Generate posterior values for $\eta(\tau_0)$ and $\mu_c(\tau_0)$.

We employ the following steps to generate n posterior values for $\mu_c(\tau_0)$ and $\eta(\tau_0)$ from their posterior distribution. The derivations may be found in Appendix E of the Supplementary Material.

- **Step 1**: Generate n posterior values for $\mu_c(\tau_0)$, denoted $\left\{\mu^i_{c,\text{post}}(\tau_0) : i = 1, \ldots, n\right\}$. For each i, $\mu^i_{c,\text{post}}(\tau_0)$ is determined by

$$\mu^i_{c,\text{post}}(\tau_0) = \frac{\sigma^2_{\tau_0}}{\sigma^2_{\tau_0} + \sigma^2_{\text{pri},\tau_0}}\bar{o}_{\text{pri},\tau_0} + \frac{\sigma^2_{\text{pri},\tau_0}}{\sigma^2_{\tau_0} + \sigma^2_{\text{pri},\tau_0}} y_{\tau_0} + \sqrt{\frac{\sigma^2_{\tau_0}}{\sigma^2_{\tau_0} + \sigma^2_{\text{pri},\tau_0}}}\left\{\mu^i_{c,\text{pri}}(\tau_0) - \bar{o}_{\text{pri},\tau_0}\right\}, \tag{9}$$

where $\sigma^2_{\tau_0}$ is given by Equation (8) with $t = \tau_0$ and y_{t-1} taken as y_{τ_0}, the number of confirmed cases reported on day τ_0.

- **Step 2**: Generate n posterior values for $\eta(\tau_0)$, denoted $\left\{\eta^i_{\text{post},\tau_0} : i = 1, \ldots, n\right\}$, with $\eta^i_{\text{post},\tau_0} = \left(\theta^i_{\text{post},\tau_0}, \mu^i_{\text{post},\tau_0}, \alpha^i_{\text{post},\tau_0}, \beta^i_{\text{post},\tau_0}, \gamma^i_{\text{post},\tau_0}, F^i_{\text{post},\tau_0}, B^i_{\text{post},\tau_0}, J^i_{\text{post},\tau_0}\right)^{\text{T}}$, where each component of $\eta^i_{\text{post},\tau_0}$ equals

$$X^i_{\text{post},\tau_0} = X^i_{\text{pri},\tau_0} + \left(\frac{\sigma^{cov}_{X,\mu_c(\tau_0),\text{pri}}}{\sigma^2_{\text{pri},\tau_0}}\right)\left\{\mu^i_{c,\text{post}}(\tau_0) - \mu^i_{c,\text{pri}}(\tau_0)\right\} \tag{10}$$

with X representing a symbol for θ, μ, α, β, γ, F, B, and J.

Part 2: ***At iteration*** $l = 1$ ***and at time*** $t = \tau_0 + 1$, ***determine prior and posterior values of*** $\eta(t)$ ***and*** $\mu_c(t)$:

- **Stage 1**: Generate prior values for $\eta(t)$, $\phi(t)$, and $\mu_c(t)$.
  - **Step 1**: Generate n prior values for $\eta(\tau_0+1)$, denoted $\left\{\eta^i_{\text{pri},\tau_0+1} : i = 1, \ldots, n\right\}$, by setting $\eta^i_{\text{pri},\tau_0+1} = \eta^i_{\text{post},\tau_0}$ for $i = 1, \ldots, n$, where we denote $\eta^i_{\text{pri},\tau_0+1} = \big(\theta^i_{\text{pri},\tau_0+1}, \mu^i_{\text{pri},\tau_0+1}, \alpha^i_{\text{pri},\tau_0+1}, \beta^i_{\text{pri},\tau_0+1}, \gamma^i_{\text{pri},\tau_0+1}, F^i_{\text{pri},\tau_0+1}, B^i_{\text{pri},\tau_0+1}, J^i_{\text{pri},\tau_0+1}\big)^{\mathrm{T}}$ for each i.
  - **Step 2**: Using the RK4 method, generate n prior values for $\phi(\tau_0+1)$, denoted $\left\{\phi^i_{\text{pri}}(\tau_0+1) : i = 1, \ldots, n\right\}$, where for each i, $\phi^i_{\text{pri}}(\tau_0+1) = \big(S^i_{\text{pri}}(\tau_0+1), E^i_{\text{pri}}(\tau_0+1), I^i_{a_{\text{pri}}}(\tau_0+1), I^i_{s_{\text{pri}}}(\tau_0+1), A^i_{\text{pri}}(\tau_0+1), R^i_{\text{pri}}(\tau_0+1))^{\mathrm{T}}$. Specifically, let $k_1(t) = g\big(\phi(t), \eta^i_{\text{pri},t+1}\big)$, $k_2(t) = g\big(\phi(t) + \frac{k_1(t)}{2}, \eta^i_{\text{pri},t+1}\big)$, $k_3(t) = g\big(\phi(t) + \frac{k_2(t)}{2}, \eta^i_{\text{pri},t+1}\big)$, and $k_4(t) = g\big(\phi(t) + k_3(t), \eta^i_{\text{pri},t+1}\big)$. Then we set

$$\phi^i_{\text{pri}}(\tau_0+1) = \phi(\tau_0) + \frac{k_1(\tau_0) + 2k_2(\tau_0) + 2k_3(\tau_0) + k_4(\tau_0)}{6}. \tag{11}$$

  - **Step 3**: Generate n prior values for $\mu_c(\tau_0+1)$, denoted $\left\{\mu^i_{c,\text{pri}}(\tau_0+1) : i = 1, \ldots, n\right\}$, by setting $\mu^i_{c,\text{pri}}(\tau_0+1) = \frac{I^i_{s_{\text{pri}}}(\tau_0+1)}{F^i_{\text{pri},\tau_0+1}}$ for $i = 1, \ldots, n$. Then calculate

$$\sigma^2_{\text{pri},\tau_0+1} = \frac{1}{n-1}\sum_{i=1}^{n}\left\{\mu^i_{c,\text{pri}}(\tau_0+1) - \frac{\sum_{i=1}^{n}\mu^i_{c,\text{pri}}(\tau_0+1)}{n}\right\}^2,$$

together with the pairwise sample covariance

$$\sigma^{cov}_{X(\tau_0+1),\mu_c(\tau_0+1),\text{pri}} = \frac{1}{n-1}\sum_{i=1}^{n}\left[\left\{\mu^i_{c,\text{pri}}(\tau_0+1) - \frac{\sum_{i=1}^{n}\mu^i_{c,\text{pri}}(\tau_0+1)}{n}\right\} \times \left\{X^i_{\text{pri}}(\tau_0+1) - \frac{\sum_{i=1}^{n}X^i_{\text{pri}}(\tau_0+1)}{n}\right\}\right],$$

where X represents each of the symbols S, E, I_a, I_s, A, or R.

- **Stage 2**: Generate posterior values for $\eta(t)$, $\phi(t)$, and $\mu_c(t)$. This stage is similar to Stage 2 in Part 1.
  - **Step 1**: Generate n posterior values for $\mu_c(\tau_0+1)$ and $\eta(\tau_0+1)$, respectively, denoted $\left\{\mu^i_{c,\text{post}}(\tau_0+1) : i = 1, \ldots, n\right\}$ and $\left\{\eta^i_{\text{post},\tau_0+1} : i = 1, \ldots, n\right\}$.
  - **Step 2**: Generate n posterior values for $\phi(\tau_0+1)$, denoted $\left\{\phi^i_{\text{post}}(\tau_0+1) : i = 1, \ldots, n\right\}$, where for each i, $\phi^i_{\text{post}}(\tau_0+1) = \big(S^i_{\text{post}}(\tau_0+1), E^i_{\text{post}}(\tau_0+1), I^i_{a_{\text{post}}}(\tau_0+1), I^i_{s_{\text{post}}}(\tau_0+1), A^i_{\text{post}}(\tau_0+1), R^i_{\text{post}}(\tau_0+1)\big)^{\mathrm{T}}$, with each component of $\phi^i_{\text{post}}(\tau_0+1)$ given by

$$X^i_{\text{post}}(\tau_0+1) = X^i_{\text{pri}}(\tau_0+1) + \left(\frac{\sigma^{cov}_{X(\tau_0+1),\mu_c(\tau_0+1),\text{pri}}}{\sigma^2_{\text{pri},\tau_0+1}}\right)\left\{\mu^i_{c,\text{post}}(\tau_0+1) - \mu^i_{c,\text{pri}}(\tau_0+1)\right\}.$$

Here X represents each of the symbols S, E, I_a, I_s, A, or R.

Part 3: ***At iteration*** $l = 1$ ***and at time*** $t = \tau_0 + 2$, ***determine prior and posterior values of*** $\eta(t)$ ***and*** $\mu_c(t)$:

- **Stage 1**: Generate prior values for $\eta(t)$, $\phi(t)$, and $\mu_c(t)$.
  - **Step 1**: Generate n prior values for $\eta(\tau_0 + 2)$, denoted $\left\{\eta^i_{\text{pri},\tau_0+2} : i = 1, \dots, n\right\}$, by setting $\eta^i_{\text{pri},\tau_0+2} = \eta^i_{\text{post},\tau_0+1}$ for $i = 1, \dots, n$.
  - **Step 2**: Generate n prior values for $\phi(\tau_0 + 2)$, denoted $\left\{\phi^i_{\text{pri}}(\tau_0 + 2) : i = 1, \dots, n\right\}$, using the RK4 method where, like the result specified in Equation (11),

$$\phi^i_{\text{pri}}(\tau_0 + 2) = \phi^i_{\text{post}}(\tau_0 + 1) + \frac{k_1(\tau_0 + 1) + 2k_2(\tau_0 + 1) + 2k_3(\tau_0 + 1) + k_4(\tau_0 + 1)}{6}.$$

  - **Step 3**: Similar to Step 3 of Stage 1 in Part 2, generate n prior values for $\mu_c(\tau_0 + 2)$.
- **Stage 2**: Similar to Stage 2 in Part 2, generate n posterior values of $\eta(\tau_0 + 2)$, $\phi(\tau_0 + 2)$, and $\mu_c(\tau_0 + 2)$.

Part 4: ***Calculate the output of the first iteration***:

We repeat Part 3 for $t = \tau_0 + 3, \dots, \tau_0 + T$ and obtain a sequence of posterior values for $\eta(t)$, denoted $\left\{\eta^i_{\text{post},t} : i = 1, \dots, n; t \in \mathcal{T}\right\}$. Then, we calculate

$$\hat{\eta}^{(1)} = \frac{1}{n(T + 1)} \sum_{t=\tau_0}^{\tau_0+T} \sum_{i=1}^{n} \eta^i_{\text{post},t},$$

which is then adopted as the initial value of η at iteration $l = 1$.

Next, we describe the iterative procedures to update estimates of η for $l = 2, 3, \dots, L$. The iterations are similar to those sketched in the preceding Parts 1–4 except that the prior distribution π_η of $\eta(\tau_0)$ in Step 1 of Stage 1 of Part 1 is taken to be $\mathcal{N}(\hat{\eta}^{(l-1)}, a^{l-1}\Sigma)$, where $\hat{\eta}^{(l-1)}$ is the output of iteration $(l - 1)$, a is a discount factor representing a value in $(0, 1)$, and Σ is a user-specified $p \times p$ positive-definite matrix with p representing the dimension of parameter vector η (i.e., $p = 8$).

Let $\hat{\eta}^{(L)}$ denote the estimate of the model parameter η at the Lth iteration. To reduce the Monte Carlo error induced during the simulation procedure, we run the preceding algorithm repeatedly, say M times, and let $\hat{\eta}^{(L1)}, \dots, \hat{\eta}^{(LM)}$ denote the resulting estimates of η. Let $\hat{\eta} = \frac{1}{M}\sum_{j=1}^{M} \hat{\eta}^{(Lj)}$ be the final estimate of η, where M is often set to be a large number, for example, $M = 1000$, as in our numerical studies.

The implementation steps are summarized in Algorithm 1. We now comment on the specification of n, L, a, and Σ. While the choice of n may, in principle, be driven by the consideration of “the larger, the better”, our numerical experience suggests that a value in the range 100–500 usually works well in combination with suitable values of L and a. The iteration number L and the discount factor a are often specified in an ad hoc way by inspecting the evolution of the posterior distributions over iterations. The inclusion of the discount factor a ensures that $\hat{\eta}^{(L)}$ in Algorithm 1 will converge within a reasonable number of iterations. Li et al. (2020, Supplementary Material, p. 8) commented that a small value of a makes the algorithm “quench” too fast and miss the maximum likelihood estimate, while a value of a close to 1 seems to delay the prompt convergence of the algorithm. In applications, Li et al. (2020) suggested setting a to be a value in $(0.9, 0.99)$. Here, we specify Σ as a diagonal matrix with the ith diagonal element set as the variance of the uniform distribution over the range of the ith parameter in η, which we specified in Section 4.1.

Algorithm 1. IF-EAKF

```
Input: the sequence $\{y_t : t = \tau_0, \dots, \tau_0 + T\}$ of daily reported confirmed cases, the
    sequence $\{\sigma_t^2 : t = \tau_0, \dots, \tau_0 + T\}$, the covariance matrix $\Sigma$, a fixed discount
    factor $a \in (0, 1)$, and the number $L$ of iterations.
for $l = 1$ to $L$ do
    if $l = 1$ then
        Generate $n$ prior values $\left\{\eta^i_{\mathrm{pri},\tau_0} : i = 1, \dots, n\right\}$ for parameters at time $\tau_0$
          independently from distribution $\pi_\eta$;
    else
        Generate $n$ prior values, $\left\{\eta^i_{\mathrm{pri},\tau_0} : i = 1, \dots, n\right\}$, for parameters at time $\tau_0$
          independently from a multivariate Gaussian distribution $\mathcal{N}(\hat{\eta}^{(l-1)}, a^{(l-1)}\Sigma)$, where
          $\hat{\eta}^{(l-1)}$ is described below and is specified as an initial value for $l = 1$, and $a^{l-1}$
          represents a discount factor which may change with $l$.
    end
    Generate $n$ prior values $\left\{\mu^i_{c,\mathrm{pri}}(\tau_0) : i = 1, \dots, n\right\}$ for $\mu_c(\tau_0)$ based on
      $\left\{\eta^i_{\mathrm{pri},\tau_0} : i = 1, \dots, n\right\}$ and $\phi(\tau_0)$;
    Generate $n$ posterior values $\left\{\eta^i_{\mathrm{post},\tau_0} : i = 1, \dots, n\right\}$ for parameters and $n$ posterior
      values $\left\{\mu^i_{c,\mathrm{post}}(\tau_0) : i = 1, \dots, n\right\}$ for $\mu_c(\tau_0)$ based on their prior values, $\sigma^2_{\tau_0}$ and
      observation $y_{\tau_0}$;
    Generate $n$ prior values $\left\{\eta^i_{\mathrm{pri},\tau_0+1} : i = 1, \dots, n\right\}$ for parameters at time $\tau_0 + 1$ by
      setting $\eta^i_{\mathrm{pri},\tau_0+1} = \eta^i_{\mathrm{post},\tau_0}$ for $i = 1, \dots, n$. The RK4 method is used to generate $n$
      prior values $\left\{\phi^i_{\mathrm{pri}}(\tau_0 + 1) : i = 1, \dots, n\right\}$ for $\phi(\tau_0 + 1)$ based on Equations (1)–(6).
      Generate $n$ prior values $\left\{\mu^i_{c,\mathrm{pri}}(\tau_0 + 1) : i = 1, \dots, n\right\}$ for $\mu_c(\tau_0 + 1)$;
    for $t = \tau_0 + 1$ to $\tau_0 + T$ do
        Generate $n$ posterior values $\left\{\eta^i_{\mathrm{post},t} : i = 1, \dots, n\right\}$ for parameters at time $t$, $n$
          posterior values $\left\{\phi^i_{\mathrm{post}}(t) : i = 1, \dots, n\right\}$ for $\phi(t)$, and $n$ posterior values
          $\left\{\mu^i_{c,\mathrm{post}}(t) : i = 1, \dots, n\right\}$ for $\mu_c(t)$ based on their prior values, $\sigma_t^2$ and observation
          $y_t$;
        Generate $n$ prior values $\left\{\eta^i_{\mathrm{pri},t+1} : i = 1, \dots, n\right\}$ for parameters at time $t + 1$ by
          setting $\eta^i_{\mathrm{pri},t+1} = \eta^i_{\mathrm{post},t}$ for $i = 1, \dots, n$. The RK4 method is used to generate $n$
          prior values $\left\{\phi^i_{\mathrm{pri}}(t + 1) : i = 1, \dots, n\right\}$ for $\phi(t + 1)$. Generate $n$ prior values
          $\left\{\mu^i_{c,\mathrm{pri}}(t + 1) : i = 1, \dots, n\right\}$ for $\mu_c(t + 1)$;
    end
    Calculate the mean: $\hat{\eta}^{(l)} = \frac{1}{n(T+1)} \sum_{t=\tau_0}^{\tau_0+T} \sum_{i=1}^{n} \eta^i_{\mathrm{post},t}$.
end
Output: $\hat{\eta}^{(L)}$.
```

5. ANALYSIS OF THE QUEBEC COVID-19 DATA

As an illustration, we used our proposed model to analyze the observed data concerning the daily reported number of COVID-19 cases in the Canadian province of Quebec for the period between 2 April 2020 and 10 May 2020. To implement the procedure described in Section 4.2, we developed the R source code, which is available from the second author upon request. Using the notation in Section 3.1, the examination days are listed in $\mathcal{T} = \{\tau_0, \tau_0 + 1, \dots, \tau_0 + T\}$ with $T = 38$ and τ_0 being 2 April 2020. By time τ_0, the reported cumulative numbers of recoveries, of confirmed cases, and of deaths from COVID-19 are available at the website https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection.html, giving $\mathcal{D}_b = \{R_c, C_0, D_0\}$ with $R_c = 29$, $C_0 = 4611$, and $D_0 = 33$. In addition, the observed values $\mathcal{D}_c$ for the process $\{Y(t) : t > 0\}$ are available at https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection.html.

The data $\mathcal{D}_s$ for Quebec are not available. However, using other reported data for either Canada or Quebec, we may roughly approximate $Q(t)$ for $t < \tau_0$ and $t \in \mathcal{T}$. Let $n_{t,C}$ denote the number of COVID-19 patients with symptom onset on day t in Canada, which is available in the section labelled "Epidemic curve" of the website https://health-infobase.canada.ca/covid-19/epidemiological-summary-covid-19-cases.html?stat=num&measure=total_last7&map=pt\%23a4. Let $m_{t,C}$ and $m_{t,Q}$ denote the numbers of confirmed COVID-19 cases on day t in Canada and Quebec, respectively, which can be found at the website https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection.html. Assuming that the number of confirmed cases in Quebec and in Canada in a day appears in nearly the same ratio as the number of individuals with symptom onset in Quebec to that in Canada, $Q(t)$ is approximated by $n_{t,C} m_{t,Q} / m_{t,C}$ for $t < \tau_0$ and $t \in \mathcal{T}$. As $\mathcal{D}_s = \{Q(t) : t \in \mathcal{T}\}$ is only used to help describe the initial value $\phi(\tau_0)$, we hope this treatment of $Q(t)$ offers us a reasonable approximation, and we stress that this approximation yields a less accurate estimate of η than that obtained from the setting where $\mathcal{D}_s$ would have been available.

To assess the performance of the SEASAR model in terms of parameter estimation as well as prediction, we split the study period into two parts: the period from 2 April to 30 April 2020 (denoted $\{\tau_0, \tau_0 + 1, \dots, \tau_0 + T_1\}$ with $T_1 = 28$) and the period from 1 May to 10 May 2020 (denoted $\{\tau_0 + T_1 + 1, \dots, \tau_0 + T_1 + T_2\}$ with $T_2 = 10$). The data for the first period, called the training data, are used to estimate the model parameters by applying Algorithm 1. The data in the second period, called the test data, are used to assess the prediction performance of our proposed model.

5.1. Estimation and Prediction Results

We ran Algorithm 1 with (7) and (8) on the training data, where we set $n = 300$, $a = 0.9$, $L = 50$, and fixed the average latent period Z at 5.2 days, an estimate reported by Bai et al. (2020) and Hao et al. (2020). For the initial sizes of the six subpopulations discussed in Section 3.2, we took $r_1 = 2$, $r_2 = 1$, and $r_3 = 2$.

In Table 1, we reported the estimates (EST) of the model parameters in Equations (1)–(6) as well as the estimate of the *basic reproduction number* R_0. The estimate of R_0 suggests that the pandemic situation in Quebec for the period from 2 April 2020 to 30 April 2020 was not under control and the virus was spreading in the province.

Next, we evaluated the prediction performance using the test data. With η replaced by its estimate, we used the RK4 method to derive the estimates of $\phi(t)$ recursively from the SEASAR model for $t \in \mathcal{T}$ by using

$$\phi(t+1) = \phi(t) + \frac{k_1(t) + 2k_2(t) + 2k_3(t) + k_4(t)}{6}, \tag{12}$$

a form similar to that specified in Equation (11), where $k_j(t)$ is defined as indicated in Equation (11) with $\eta^i_{\mathrm{pri},t+1}$ replaced by η, and $j = 1, \ldots, 4$. Then we took the ratio of the resulting estimates of $I_s(t)$ and F as a predicted (or fitted) value $\hat{y}_t$ to y_t for $t \in \mathcal{T}$.

To evaluate the prediction performance, for $t = \tau_0 + T_1 + 1, \ldots, \tau_0 + T_1 + T_2$, we report the absolute prediction error (APE), defined $|\hat{y}^*_t - y^*_t|$, and the relative absolute prediction error (Rel.APE) in percent, defined as $\frac{|\hat{y}^*_t - y^*_t|}{y^*_t} \times 100$, where two settings are considered with y^*_t representing the daily net number of confirmed cases (i.e., $y^*_t = y_t$) or the daily cumulative number of confirmed cases (i.e., $y^*_t = \sum_{s \le t} y_s$), and we let $\hat{y}^*_t$ denote the corresponding predicted value of y^*_t. The results are reported in Table 2.

To visualize the results, we displayed in Figure 2 the mean estimates of the daily *net* number and of the daily *cumulative* number of cases for the period from 2 April 2020 to 30 April 2020 (in green), as well as the mean predicted daily *net* numbers and the predicted daily *cumulative* numbers of cases for the period from 1 May 2020 to 10 May 2020 (in blue), in comparison to the actual observed daily *net* numbers and the daily *cumulative* numbers of cases from 2 April 2020 to 10 May 2020 (in red). While the APE values vary greatly from day to day and it is difficult to quantify an acceptable range for the prediction error, examining the Rel.APE values gives us some insight into the prediction performance. The Rel.APE values for predicting a daily *net* number of confirmed cases seem to be acceptable and they are very small for predicting a daily *cumulative* number of confirmed cases. This suggests that prediction of a daily *cumulative* number of confirmed cases seems fairly acceptable.

TABLE 1: Analysis of the Quebec data: The estimates of the model parameters.

Parameter	θ	μ	α	β	γ	F	B	J	R_0
EST	0.88	0.10	0.48	0.01	0.97	2.36	39.58	37.56	1.06

TABLE 2: Analysis of the Quebec data using the proposed SEASAR model: Prediction performance for the daily net and daily cumulative numbers in 10 days for the period 1 May 2020 to 10 May 2020.

	Net number		Cumulative number	
	APE	Rel.APE	APE	Rel.APE
Day 1	241.07	21.72	36.96	0.13
Day 2	133.30	13.22	170.26	0.57
Day 3	11.52	1.29	181.78	0.60
Day 4	128.27	16.92	53.50	0.17
Day 5	98.08	12.35	44.57	0.14
Day 6	12.11	1.33	32.46	0.10
Day 7	7.29	0.80	25.17	0.07
Day 8	2.46	0.27	22.71	0.07
Day 9	79.38	0.95	102.08	0.29
Day 10	186.22	25.34	288.31	0.79

Note: APE and Rel.APE (in percent) represent the absolute prediction error and the relative absolute prediction error, respectively.

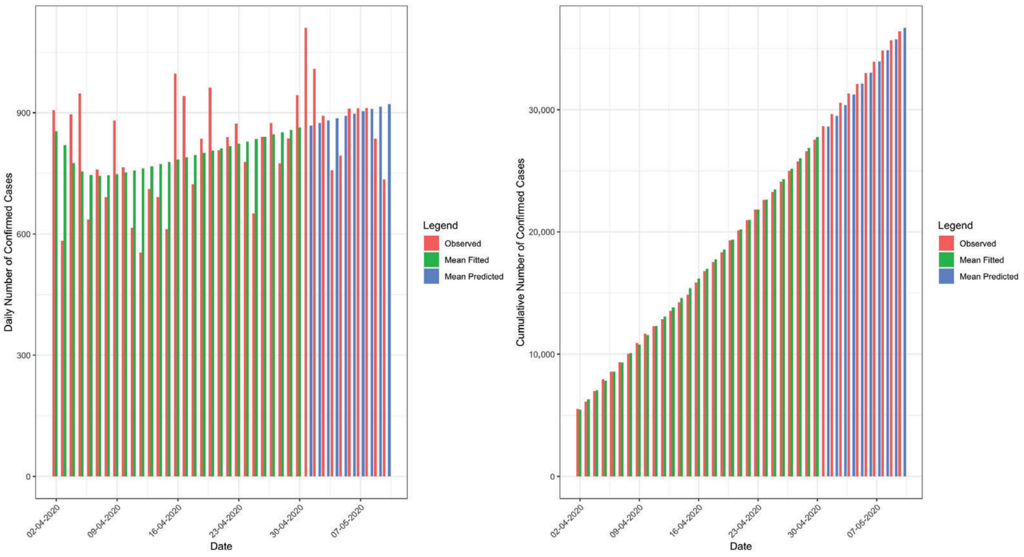

FIGURE 2: The model fitting to the number (in green) in the period 2 April 2020 to 30 April 2020, and the model prediction to the number (in blue) in the period 1 May 2020 to 10 May 2020, as opposed to the reported number (in red) in the period 2 April 2020 to 10 May 2020, where the number represents the daily net number of confirmed cases (left) and the daily cumulative number of confirmed cases (right).

Finally, to mitigate the error due to data aggregation that arises on weekends, we calculated APE and Rel.APE for the weekly net numbers of confirmed cases, which are 178.938 and 0.028, respectively. We also displayed these results in Figure S.3 of the Supplementary Material in a manner similar to that used in Figure 2.

5.2. A Comparison with the Neural Network, SIR, and SEIR Models

To further assess the performance of the SEASAR model, we compared its prediction performance to that of the neural network (NN), SIR, and SEIR models by examining prediction of the daily net, daily cumulative, and weekly net numbers of confirmed cases.

The NN method is commonly used by the machine learning community. Basically, it contains three elements: the input layer, the hidden layer(s) with a number of nodes, and the output layer. Here we take the input data as the observed time series $\mathcal{D}_c$ with $T = 28$ and τ_0 being 2 April 2020. We use the NN model with one hidden layer having three nodes, with the same implementation details as specified in Chen et al. (2021). The R function *nnetar* was used to fit the training data, and the R function *forecast* was used for prediction.

Next we considered the SIR model, which divides the target population into three subpopulations, respectively denoted S_{SIR}, I_{SIR}, and R_{SIR}, of susceptible cases, of infectious cases (i.e., those who are infected and are themselves infectious), and of removed cases (i.e., those cases who recover or die from COVID-19). Let $S_{\text{SIR}}(t)$, $I_{\text{SIR}}(t)$, and $R_{\text{SIR}}(t)$ denote the mean size of the corresponding subpopulation at time t, which are, by definition, related to the subpopulations classified by the SEASAR model via $I_{\text{SIR}}(t) = I_a(t) + I_s(t)$, $R_{\text{SIR}}(t) = R(t)$, and $S_{\text{SIR}}(t) = N - I_{\text{SIR}}(t) - R_{\text{SIR}}(t)$. Let λ denote the average transmission rate, defined as the average number of individuals infected by an infectious case per unit time, and let ξ denote the average removal rate, defined as the average rate of death or recovery. Under the same assumptions for the SEASAR model, the SIR model is characterized by the following ordinary differential equations:

$$\frac{dS_{\mathrm{SIR}}(t)}{dt} = -\frac{\lambda S_{\mathrm{SIR}}(t) I_{\mathrm{SIR}}(t)}{N}; \tag{13}$$

$$\frac{dI_{\mathrm{SIR}}(t)}{dt} = \frac{\lambda S_{\mathrm{SIR}}(t) I_{\mathrm{SIR}}(t)}{N} - \xi I_{\mathrm{SIR}}(t); \tag{14}$$

$$\frac{dR_{\mathrm{SIR}}(t)}{dt} = \xi I_{\mathrm{SIR}}(t). \tag{15}$$

Now we turn to the SEIR model that stratifies the population into four subpopulations. In addition to the three subpopulations considered in the SIR model, now denoted S_{SEIR}, I_{SEIR}, and R_{SEIR}, the SEIR model further considers the subpopulation of exposed cases (i.e., those who are infected but not yet infectious, and are still in the latent period), denoted E_{SEIR}. Let $S_{\mathrm{SEIR}}(t)$, $E_{\mathrm{SEIR}}(t)$, $I_{\mathrm{SIR}}(t)$, and $R_{\mathrm{SIR}}(t)$ denote the mean size of the corresponding subpopulation at time t, which, by definition, are related to the quantities in the SEASAR model by the corresponding $E_{\mathrm{SEIR}}(t) = E(t)$, $I_{\mathrm{SEIR}}(t) = I_a(t) + I_s(t)$, $R_{\mathrm{SEIR}}(t) = R(t)$, and $S_{\mathrm{SEIR}}(t) = N - E_{\mathrm{SEIR}}(t) - I_{\mathrm{SEIR}}(t) - R_{\mathrm{SEIR}}(t)$. Under the same assumptions that apply to the SEASAR model, the SEIR model is characterized by the ordinary differential equations

$$\frac{dS_{\mathrm{SEIR}}(t)}{dt} = -\frac{\lambda^* S_{\mathrm{SEIR}}(t) I_{\mathrm{SEIR}}(t)}{N}; \tag{16}$$

$$\frac{dE_{\mathrm{SEIR}}(t)}{dt} = \frac{\lambda^* S_{\mathrm{SEIR}}(t) I_{\mathrm{SEIR}}(t)}{N} - \frac{E_{\mathrm{SEIR}}(t)}{Z}; \tag{17}$$

$$\frac{dI_{\mathrm{SEIR}}(t)}{dt} = \frac{E_{\mathrm{SEIR}}(t)}{Z} - \xi^* I_{\mathrm{SEIR}}(t); \tag{18}$$

$$\frac{dR_{\mathrm{SEIR}}(t)}{dt} = \xi^* I_{\mathrm{SEIR}}(t); \tag{19}$$

where λ^* and ξ^* have meanings similar to those of λ and ξ in the SIR model, and Z is the latent period defined in Section 2.1 and is fixed at the value $Z = 5.2$ as in Section 5.1.

Modifying the estimation procedure that we outlined in Section 4.2, we obtained estimates of the parameters for the SIR and SEIR models, where the implementation details may be found in Appendices I and J in the Supplementary Material; we set $n = 300, a = 0.9, L = 50$, and $M = 1000$, as in Section 5.1. As we previously reported in Section 5.1, in Table 3 we summarized the APE and Rel.APE values for prediction of the daily net number and the daily cumulative number obtained from the fitted NN, SIR, and SEIR models for the same time period 1 May 2020 to 10 May 2020. For the prediction of the weekly net number, we estimated that the average APE values for the NN, SIR, and SEIR were 678.07, 1354.636, and 1863.149, respectively, and that the average Rel.APE values for the NN, SIR, and SEIR were 0.106, 0.212, and 0.292, respectively. In contrast, the proposed SEASAR model yielded a much smaller average APE and average Rel.APE, which were 178.938 and 0.028, respectively.

TABLE 3: Analysis of the Quebec data using the NN, SIR, and SEIR models: Prediction performance for the daily net and daily cumulative numbers in 10 days for the period 1 May 2020 to 10 May 2020.

	NN model				SIR model				SEIR model			
	Net number		Cumulative number		Net number		Cumulative number		Net number		Cumulative number	
Day	APE	Rel. APE	APE	Rel. APE	APE	Rel. APE	APE	Rel. APE	APE	Rel. APE	APE	Rel. APE
1	293.09	26.40	292.76	1.02	383.85	34.58	1939.29	6.77	31.91	2.87	6195.44	21.63
2	192.90	19.14	485.65	1.64	284.45	28.22	2223.75	7.50	145.76	14.46	6341.19	21.38
3	77.31	8.67	562.96	1.84	171.05	19.18	2394.80	7.84	273.72	30.69	6614.91	21.65
4	56.58	7.46	506.37	1.62	39.65	5.23	2434.46	7.78	419.80	55.38	7034.71	22.47
5	20.55	2.59	485.82	1.51	78.26	9.86	2512.72	7.83	395.98	49.87	7430.69	23.15
6	95.46	10.49	581.28	1.76	196.87	21.63	2709.59	8.21	292.29	32.12	7722.98	23.40
7	96.46	10.59	677.74	2.00	200.48	22.01	2910.08	8.58	303.70	33.33	8026.68	23.66
8	97.46	10.69	775.21	2.23	204.10	22.38	3114.17	8.94	315.23	34.56	8341.91	23.94
9	21.46	2.57	796.67	2.23	130.71	15.64	3244.88	9.10	403.87	48.31	8745.78	24.52
10	79.54	10.82	717.13	1.97	32.33	4.40	3277.21	9.00	517.63	70.42	9263.40	25.45

Note: APE and Rel.APE (in percent) represent the absolute prediction error and the relative absolute prediction error, respectively.

To visualize the results obtained from the SEASAR model (in verde olive), the NN model (in purple), the SIR model (in green), and the SEIR model (in blue), we displayed in Figure 3 the mean predicted daily net numbers, the mean predicted daily cumulative numbers, and the mean predicted weekly net numbers of confirmed cases for the period from 1 May 2020 to 10 May 2020, in comparison to the corresponding reported numbers for the same period. Our proposed SEASAR model appeared to outperform the SIR, SEIR, and NN models with respect to prediction during the period represented by the test data. The SEIR and SIR models performed quite differently; the SEIR model resulted in over-estimated values, whereas the SIR model produced under-estimated results, though the degree of bias varied considerably. It would be interesting to investigate what causes these systematic biases of the SEIR and SIR models, though such an investigation is beyond the scope of this article.

Finally, to see how various assumptions might affect the results, we conducted sensitivity analyses for 12 settings. For details, see Appendix H in the Supplementary Material.

6. SIMULATION STUDY

To assess the performance of our proposed method, we conducted simulation studies using the estimation procedure described in Section 4.2. In this section, we report the performance of the SEASAR model by assuming that the two required conditions are met: (1) the population is homogeneous and (2) the population is closed, i.e., there is no immigration or emigration. In Appendix K of the Supplementary Material, we assess the model performance when those conditions are violated.

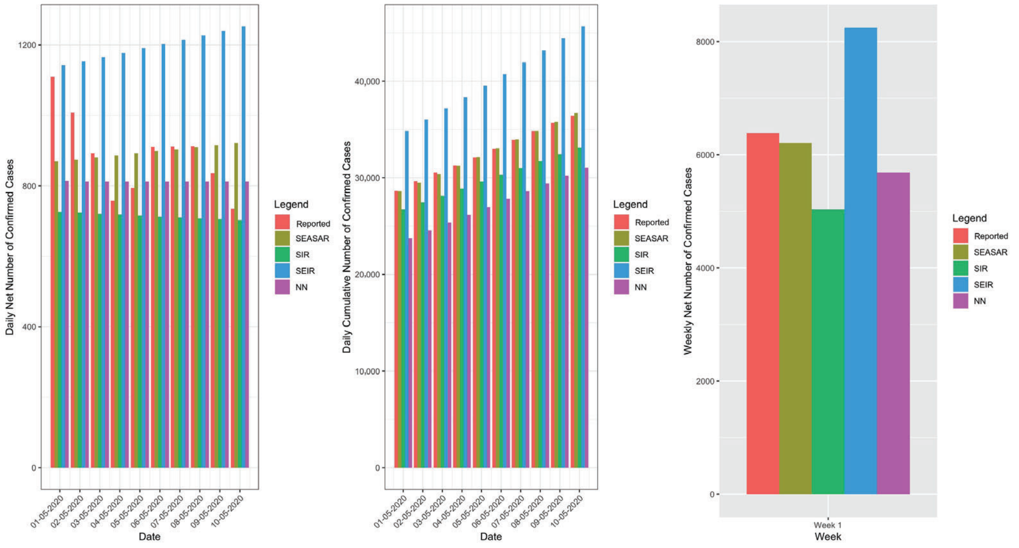

FIGURE 3: The average prediction to the number in the period 1 May 2020 to 10 May 2020 using the SEASAR model (in verde olive), SIR model (in green), SEIR model (in blue), and NN model (in purple), as opposed to the real number (in red) in the period 1 May 2020 to 10 May 2020, where the number represents the daily net number (left), daily cumulative number (middle), and weekly net number (right) of confirmed cases.

6.1. Data Generation

First, we generated data from the SEASAR model specified in Equations (1)–(6), together with (7) and (8), for a period of study days. We fixed $Z = 5.2$, $N = 8{,}433{,}301$, and $\eta = (0.2, 0.45, 0.18, 0.08, 0.7, 9, 24, 34)^{\mathrm{T}}$. Then we used Equation (12) to generate data with the initial value of $\phi(t)$ equal to $\phi(\tau_0) = (8{,}421, 149.295, 3475.073, 2018.316, 2018.316, 4549, 91)^{\mathrm{T}}$. To see what the trajectories of $\phi(t)$ look like, we plotted $\phi(t)$ in Figure S.6 in the Supplementary Material with T set to the value 200.

With the availability of $\{\phi(t) : t \in \mathcal{T}\}$, we calculated $\mu_c(t) = I_s(t)/F$ for $t \in \mathcal{T}$, and then used (7) and (8) to iteratively generate realizations y_t of $Y(t)$ for $t \in \mathcal{T}$, where we set $y_{\tau_0 - 1} = 449$, the number of confirmed cases in Quebec on 1 April 2020. Thus, we obtained the data $\mathcal{D}_c = \{y_t : t \in \mathcal{T}\}$. We repeated this data-generation process m times, and let $\mathcal{D}_c^j = \{y_t^j : t \in \mathcal{T}\}$ record each copy of $\mathcal{D}_c$ for $j = 1, 2, \ldots, m$, where we used $m = 1000$, and y_t^j represents a realized value of $Y(t)$ in the jth generated data sample.

6.2. Assessment of the Estimation Performance

Here we focused on evaluating estimation of the model parameter η as well as prediction of the daily numbers of confirmed cases. To this end, for $j = 1, 2, \ldots, m$, we divided $\mathcal{D}_c^j$ into two subsets, i.e., $\mathcal{D}_{CT}^j$ and $\mathcal{D}_{CP}^j$, with $\mathcal{D}_{CT}^j = \left\{y_t^j : t = \tau_0, \tau_0 + 1, \ldots, \tau_0 + T_1\right\}$ and $\mathcal{D}_{CP}^j = \left\{y_t^j : t = \tau_0 + T_1 + 1, \ldots, \tau_0 + T_1 + T_2\right\}$, where $T_1 + T_2 = T = 49$. We considered different values of T_1 to evaluate the model performance. Here $\mathcal{D}_{CT}^j$ represents the training data used to estimate the SEASAR model parameters, and $\mathcal{D}_{CP}^j$ denotes the test data used to evaluate the prediction performance.

When assessing the estimate of η, we considered three cases for the training data $\mathcal{D}_{CT}^j$ with $T_1 = 29$, 34, or 39, respectively, called "Case 1", "Case 2", and "Case 3". In each instance, we separately ran Algorithm 1 on each set of training data $\mathcal{D}_{CT}^j$ for $j = 1, \ldots, m$ to estimate the

TABLE 4: Simulation study in Section 6 with $a = 0.9$ and $L = 50$: Estimation results for the SEASAR model parameter; the entries with $*$ are the original values times 10^3.

$n = 100$

	Case 1			Case 2			Case 3		
Parameter	BIAS	RBIAS	SSD	BIAS	RBIAS	SSD	BIAS	RBIAS	SSD
θ	−0.01	−0.05	0.02	−0.01	−0.06	0.02	−0.01	−0.07	0.02
μ	0.15	0.34	0.11	0.19	0.41	0.11	0.20	0.45	0.11
α	−0.01	−0.08	0.02	−0.02	−0.09	0.02	−0.02	−0.10	0.02
β	1.59^*	0.02	0.02	1.60^*	0.02	0.02	1.49^*	0.02	0.02
γ	−0.06	−0.08	0.10	−0.07	−0.10	0.11	−0.08	−0.11	0.11
F	-8.47^*	-0.94^*	0.16	−0.01	-1.33^*	0.16	−0.02	-1.68^*	0.16
B	0.33	0.01	3.66	0.28	0.01	3.77	0.48	0.02	3.67
J	1.00	0.03	4.21	0.89	0.03	4.45	1.05	0.03	4.36

$n = 300$

	Case 1			Case 2			Case 3		
Parameter	BIAS	RBIAS	SSD	BIAS	RBIAS	SSD	BIAS	RBIAS	SSD
θ	−0.01	−0.06	0.02	−0.01	−0.07	0.02	−0.01	−0.07	0.02
μ	0.17	0.38	0.09	0.20	0.44	0.09	0.22	0.49	0.09
α	−0.02	−0.08	0.02	−0.02	−0.10	0.02	−0.02	−0.11	0.02
β	1.92^*	0.02	0.02	1.18^*	0.01	0.02	1.34^*	0.02	0.02
γ	−0.06	−0.09	0.10	−0.08	−0.11	0.10	−0.09	−0.12	0.11
F	-8.06^*	-0.90^*	0.16	−0.01	-1.45^*	0.16	−0.02	-1.80^*	0.16
B	0.46	0.02	2.13	0.56	0.02	2.19	0.40	0.02	2.25
J	1.06	0.03	2.52	0.87	0.03	2.59	0.96	0.03	2.49

$n = 500$

	Case 1			Case 2			Case 3		
Parameter	BIAS	RBIAS	SSD	BIAS	RBIAS	SSD	BIAS	RBIAS	SSD
θ	−0.01	−0.06	0.02	−0.01	−0.07	0.02	−0.02	−0.08	0.02
μ	0.18	0.39	0.08	0.21	0.46	0.09	0.23	0.50	0.09
α	−0.02	−0.09	0.02	−0.02	−0.10	0.02	−0.02	−0.11	0.02
β	1.90^*	0.02	0.02	1.35^*	0.02	0.02	1.25^*	0.02	0.02
γ	−0.07	−0.09	0.09	−0.08	−0.11	0.10	−0.09	−0.13	0.10
F	-8.83^*	-0.98^*	0.16	−0.01	-1.48^*	0.16	−0.02	-1.86^*	0.16
B	0.44	0.02	1.68	0.48	0.02	1.68	0.47	0.02	1.73
J	0.94	0.03	2.04	1.00	0.03	1.99	1.00	0.03	2.03

Note: BIAS and RBIAS represent the average bias of the estimates and the average relative bias of the estimates, respectively; and SSD stands for the sample standard deviation of the estimates.

TABLE 5: Simulation study in Section 6: Prediction performance of the proposed SEASAR model for Scenario-1-short ($T_1 = 29; T_2 = 5$), Scenario-1-long ($T_1 = 29; T_2 = 10$), Scenario-2-short ($T_1 = 39; T_2 = 5$), and Scenario-2-long ($T_1 = 39; T_2 = 10$).

	TAPE		TRAPE	
	ATAPE	SSD	ATRAPE	SSD
Setting 1				
Scenario-1-short	24.77	8.71	0.31	0.11
Scenario-1-long	49.07	15.03	0.67	0.21
Scenario-2-short	22.92	7.51	0.41	0.14
Scenario-2-long	45.57	10.72	0.90	0.22
Setting 2				
Scenario-1-short	105.32	77.35	0.02	0.02
Scenario-1-long	262.25	196.44	0.05	0.04
Scenario-2-short	103.72	80.48	0.02	0.01
Scenario-2-long	245.09	182.93	0.04	0.03

Note: T_1 and T_2 represent the size of the training and test data, respectively. ATAPE and ATRAPE represent the averages of total absolute prediction error and the total relative absolute prediction error, respectively.

model parameter η, where we considered different specifications of n, a, and L. Let $\hat{\eta}^j$ denote the resulting estimates for $j = 1, \dots, m$ for each configuration. We reported the average bias of the estimates (BIAS), which we calculated using $m^{-1}\sum_{j=1}^{m}\hat{\eta}_r^j - \eta_r$; the average relative estimate bias (RBIAS), which we calculated using $m^{-1}\sum_{j=1}^{m}\{\hat{\eta}_j r^j - \eta_r\}/\eta_r$; and the sample standard deviation (SSD), which we calculated using $\sqrt{(m-1)^{-1}\sum_{j=1}^{m}\left(\hat{\eta}_r^j - m^{-1}\sum_{j=1}^{m}\hat{\eta}_r^j\right)^2}$, where $\hat{\eta}_r^j$ and η_r denote the rth element of $\hat{\eta}^j$ and η, respectively, and $r = 1, \dots, 8$.

Table 4 summarizes the results for the settings with $a = 0.9$ and $L = 50$, where n is taken as 100, 300, or 500; additional results for $n = 50$ and 1000 may be found in Table S.6 in the Supplementary Material. While different choices of n lead to different degrees of estimation bias, as expected, the incurred bias or relative bias for those values of n fall in acceptable ranges in general. Interestingly, a larger value of n does not necessarily yield estimates with a smaller bias. On the contrary, increasing the value of n does help reduce the sample standard deviation. Overall, it seems that with suitable values of L and a, setting n to be a value between 100 and 500 may be plausible.

To evaluate the prediction performance of the SEASAR model, we used different training data to build a prediction model and then compared the performance over different prediction windows. We first fixed $T_1 = 29$ or 39 for the training data $\mathcal{D}_{CT}^j$ to build a prediction model in simulation j. Then for each scenario, we compared the prediction performance over a short and a relatively long time period by using the test data $\mathcal{D}_{CP}^j$ with $T_2 = 5$ and 10, respectively. For convenience, we used the labels "Scenario-1-short", "Scenario-1-long", "Scenario-2-short", and "Scenario-2-long", respectively, to indicate those settings with $(T_1, T_2) = (29, 5), (29, 10), (39, 5)$, and $(39, 10)$.

To assess the discrepancies between the predicted values and the corresponding values generated from the model for the test data, for $j = 1, \dots, m$, we calculated the total absolute prediction error (TAPE): $\text{TAPE}_j = \sum_{t=\tau_0+T_1+1}^{\tau_0+T_1+T_2}|\hat{y}_t^{*j} - y_t^{*j}|$, and the total relative absolute prediction error

(TRAPE): $\mathrm{TRAPE}_j = \sum_{t=\tau_0+T_1+1}^{\tau_0+T_1+T_2} \left| \frac{\hat{y}_t^{*j} - y_t^{*j}}{y_t^{*j}} \right|$, where we considered two settings for y_t^{*j}. In Setting 1, we let y_t^{*j} represent the daily net number of confirmed cases $y_t^j \in \mathcal{D}_{CP}^j$, and let $\hat{y}_t^{*j}$ denote its predicted value. In Setting 2, we let y_t^{*j} represent the daily cumulative number of confirmed cases $\sum_{s \le t} y_s^j$ with $y_s^j \in \mathcal{D}_{CP}^j$, and let $\hat{y}_t^{*j}$ denote the corresponding predicted value.

Figure S.7 of the Supplementary Material shows the boxplots of the m TAPEs and TRAPEs for the four scenarios under the two settings. To gain an overall sense of the prediction error, in Table 5 we reported the average TAPE (ATAPE) and the average of TRAPE (ATRAPE) over the m simulations, together with the associated sample standard deviations (SSD). As expected, with a given training dataset to build a SEASAR model for prediction, the ATAPE and ATRAPE for a shorter prediction window are smaller than the corresponding values for a longer prediction window. Furthermore, the former case incurs less variation than the latter one.

7. DISCUSSION

In this article, we introduced a new epidemic model, called the SEASAR model, to describe the transmission process for COVID-19, where the population is divided into six subpopulations, called *susceptible*, *exposed*, *asymptomatic*, *symptomatic*, *active*, and *removed*. While the proposed SEASAR model extends the SIR and SEIR models to accommodate the manifestations of COVID-19 related to asymptomatic infections and varying lag times between symptom onset and diagnosis, it has certain limitations as we outline below.

The model basically delineates settings that are reasonably characterized by time-invariant parameters in η. It does not, however, facilitate the investigation of other features of the data such as weekly cycles related to varying testing and reporting rates among weekdays and weekends. To gain further insight into transmission, one may prefer to use the SEASAR model to predict the number of cumulative cases rather than daily cases. Further, prediction over a short-term window tends to be more reliable than that over a longer period. In contrast, dividing the study period into five intervals to allow for interval-dependent model parameters, Hao et al. (2020) proposed a generalized SEIR model to describe the COVID-19 dynamic progression in Wuhan during the period from 1 January 2020 to 8 March 2020. With 10 unknown parameters, they focused on the estimation of only two parameters and replaced the other parameters with the estimates reported in the literature. To better describe the dynamic changes of the population, it is useful to develop a more flexible model with time-varying model parameters.

Consistent with the SIR and SEIR models, our proposed model requires two standard conditions: (1) the population is homogeneous and (2) the population size remains invariant over time. In applications, it is difficult to satisfy these conditions. By dropping the assumption of no immigration or emigration, Li et al. (2020) modified the SEIR model to delineate the COVID-19 transmission in 375 cities of China. Their method requires the availability of inter-city mobility data. It would be interesting to extend our proposed model with these two assumptions relaxed. For example, one could perhaps add immigration and emigration to the SEASAR model to reflect the population dynamics if the observed data include such information for estimation of the associated parameters. One may further stratify the six subpopulations by pandemic-related factors, such as age and medical conditions, to achieve more homogeneous subpopulations. Such a development typically requires rich information at the individual level; merely having data at the population level, such as $\mathcal{D}_b$, $\mathcal{D}_s$, and $\mathcal{D}_c$ in the estimation framework we have considered here, is not sufficient for building a more refined model than the SEASAR model.

As noted by a referee, the model parameter η cannot be well estimated based on the observed data $\mathcal{D}_b$, $\mathcal{D}_s$, and $\mathcal{D}_c$, even if it is combined with an additional distributional assumption such as the one identified in (7) and the availability assumption for Z and r_i with

$i = 1, 2, 3$. The use of Bayesian priors, as we have suggested following (S21) in Appendix D of Section S2 in the Supplementary Material, comes into play to help resolve the issues of nonidentifiability or nonestimability of the model parameters. While imposing prior information enables us to estimate η using the posterior distribution, this does not mean that the nonidentifiability issue is eliminated. We should note that estimation results based on the posterior distribution in such a circumstance may be greatly affected by the choice of priors.

The implementation of the IF-EAKF algorithm hinges on the availability of the variance σ_t^2 of $Y(t)$, as shown, for instance, in the expression indicated in Equation (9). Here σ_t^2 is assumed to be determined by the previously observed data as identified in Equation (8). While this scheme has been used in many studies of infectious diseases such as the study of influenza by Pei et al. (2018), the study of the West Nile virus by DeFelice et al. (2017), and the study of the respiratory syncytial virus by Reis & Shaman (2016), the reasonableness of the resulting analysis relies on how σ_t^2 is specified. It would be interesting to develop an estimation procedure that can also accommodate estimation of the unknown parameter σ_t^2. This research warrants in-depth studies, which extend beyond the scope of this article.

Another important aspect concerns the quality of the observed data, an issue that should not be overlooked (Yi, 2017). Under-reporting or over-reporting confirmed cases can occur on a daily basis due to reasons related to varying incubation times, insufficient test capacity, test errors, delay in data aggregation, and so on. Available COVID-19 data often involve measurement error (such as recall bias when reporting exposure to a COVID-19 infected case) and missing observations. It would be interesting to refine the proposed model to address those issues, and such research warrants a careful investigation.

ACKNOWLEDGEMENTS

The authors thank Professor Jerry Lawless and the review team for their helpful comments, which improved the presentation of this article. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Rapid Response Program COVID-19 of the Canadian Statistical Sciences Institute (CANSSI). Grace Y. Yi is a Canada Research Chair in Data Science (Tier 1). Her research was undertaken, in part, thanks to funding from the Canada Research Chairs Program.

REFERENCES

Bai, Y., Yao, L., Wei, T., Tian, F., Jin, D.-Y., Chen, L., & Wang, M. (2020). Presumed asymptomatic carrier transmission of COVID-19. *The Journal of the American Medical Association*, 323, 1406–1407.

Chen, L.-P., Zhang, Q., Yi, G. Y., & He, W. (2021). Model-based forecasting for Canadian COVID-19 data. *PLoS One*, 16, 1–18.

DeFelice, N. B., Little, E., Campbell, S. R., & Shaman, J. (2017). Ensemble forecast of human West Nile virus cases and mosquito infection rates. *Nature Communications*, 8, 1–6.

Duan, W., Fan, Z., Zhang, P., Guo, G., & Qiu, X. (2015). Mathematical and computational approaches to epidemic modeling: A comprehensive review. *Frontiers of Computer Science*, 9, 806–826.

Guan, W., Ni, Z., Hu, Y., Liang, W., Ou, C., He, J., Liu, L. et al. (2020). Clinical characteristics of coronavirus disease 2019 in China. *New England Journal of Medicine*, 382, 1708–1720.

Hao, X., Cheng, S., Wu, D., Wu, T., Lin, X., & Wang, C. (2020). Reconstruction of the full transmission dynamics of COVID-19 in Wuhan. *Nature*, 584, 420–424.

He, W., Yi, G. Y., & Zhu, Y. (2020). Estimation of the basic reproduction number, average incubation time, asymptomatic infection rate, and case fatality rate for COVID-19: Meta-analysis and sensitivity analysis. *Journal of Medical Virology*, 92, 2543–2550.

IHME. (2021). Modeling COVID-19 scenarios for the United States. *Nature Medicine*, 27, 94.

Ionides, E. L., Bretó, C., & King, A. A. (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 103, 18438–18443.

Kramer, M. (2020). *Epidemiological data from the nCoV-2019 outbreak: Early descriptions from publicly available data*. https://virological.org/t/epidemiological-data-from-the-ncov-2019-outbreak-early-descriptions-from-publicly-available-data/337

Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., & Shaman, J. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science*, 368, 489–493.

Mandal, S., Das, H., Deo, S., & Arinaminpathy, N. (2021). Combining serology with case-detection, to allow the easing of restrictions against SARS-CoV-2: A modelling-based study in India. *Scientific Reports*, 11, 1–9.

Ng, J. & Orav, E. J. (1990). A generalized chain binomial model with application to HIV infection. *Mathematical Biosciences*, 101, 99–119.

Ng, T. W., Turinici, G., & Danchin, A. (2003). A double epidemic model for the SARS propagation. *BMC Infectious Diseases*, 3, 1–16.

Osthus, D., Hickmann, K. S., Caragea, P. C., Higdon, D., & Del Valle, S. Y. (2017). Forecasting seasonal influenza with a state-space SIR model. *The Annals of Applied Statistics*, 11, 202–224.

Pei, S., Kandula, S., Yang, W., & Shaman, J. (2018). Forecasting the spatial transmission of influenza in the United States. *Proceedings of the National Academy of Sciences*, 115, 2752–2757.

Reis, J. & Shaman, J. (2016). Retrospective parameter estimation and forecast of respiratory syncytial virus in the United States. *PLoS Computational Biology*, 12, 1–15.

Shah, N. H. & Gupta, J. (2013). SEIR model and simulation for vector borne diseases. *Applied Mathematics*, 4, 13–17.

Süli, E. & Mayers, D. F. (2003). *An Introduction to Numerical Analysis*, Cambridge University Press, Cambridge, UK.

Tang, B., Wang, X., Li, Q., Bragazzi, N. L., Tang, S., Xiao, Y., & Wu, J. (2020). Estimation of the transmission risk of the 2019-nCoV and its implication for public health interventions. *Journal of Clinical Medicine*, 9, 462.

Tuite, A. R., Fisman, D. N., & Greer, A. L. (2020). Mathematical modelling of COVID-19 transmission and mitigation strategies in the population of Ontario, Canada. *CMAJ*, 192, E497–E505.

WHO (2020). *Report of the WHO-China joint mission on coronavirus disease 2019 (COVID-19)*. https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf

Yi, G. Y. (2017). Statistical Analysis with Measurement Error or Misclassification. Springer Science + Business Media LLC, New York.

Received 6 December 2020
Accepted 30 December 2021