Characterizing the Dynamic of COVID-19 with a New Epidemic Model: Susceptible-Exposed-Symptomatic-Asymptomatic-Active-Removed

Grace Y Yi,[12] Pingbo Hu,[1] Wenqing He[1]

[1]Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada

[2]Department of Computer Science, University of Western Ontario, London, Ontario, Canada

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 dynamic of COVID-19 to help understand the impact of the disease on public health and economy. While a number of epidemic models have been available to study infectious diseases, they are inadequate to describe the dynamic of COVID-19. In this paper, we develop a new epidemic model which utilizes a set of ordinary differential equations with unknown parameters to delineate the transmission process of COVID-19. Different from the traditional epidemic models, this model accounts for asymptomatic infections as well the lag between symptoms onset and the confirmation date of infection. We describe an estimation procedure for the unknown parameters in the proposed model by adapting the *iterated filter-ensemble adjustment Kalman filter* (IF-EAKF) algorithm to the reported number of confirmed cases. To assess the performance of our proposed model, we examine COVID-19 data in Quebec for the period of April 2, 2020 to May 10, 2020 and carry out sensitivity studies under a variety of assumptions. To reflect the transmission potential of an infected case, we derive the *basic reproduction number* from the proposed model. The estimated basic reproduction number suggests that the pandemic situation in Quebec for the period of April 2, 2020 to May 10, 2020 is not under control.

Key Words: Basic reproduction number, COVID-19, epidemic model, transmission, IF-EAKF algorithm

1 Introduction

The coronavirus disease 2019 (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has spread stealthily and has present a tremendous threat to the public health. On March 11, 2020, the World Health Organization (WHO) declared COVID-19 to be a global pandemic. The first COVID-19 case in Canada was identified in Toronto on January 25, 2020 (The Canadian Press, 2020). In the early stage, imported cases from other countries were the main source of the COVID-19 outbreak in Canada. On March 16, 2020, the closure of Canadian borders was announced to people who are not Canadian citizens or permanent residents (Vogel, 2020). Various measures and interventions have been taken by the federal government and provincial governments to mitigate the virus spread.

While those steps are important to help contain the virus transmissions, it is imperative to investigate the transmission dynamic from the quantitative perspectives. In the literature, a variety of epidemic models have been developed to study infectious diseases, including the *Susceptible-Infectious-Recovered* (SIR) model (e.g., Kermack and McKendrick, 1927), the *Susceptible-Infectious-Susceptible* (SIS) model (e.g., Duan et al., 2015), the *Susceptible-Exposed-Infectious-Recovered* (SEIR) model (e.g., Duan et al., 2015), the Reed-Frost model (e.g., Abbey, 1952), and their variants (e.g., Ng and Orav, 1990; Ng, Turinici, and Danchin, 2003). 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 and Gupta (2013) applied the SEIR model to examine transmission processes of vector borne diseases; Ng and Orav (1990) proposed a generalized Reed-Frost model to predict human immunodeficiency virus (HIV) incidence in San Francisco's homosexual population; and Ng et al. (2003) developed a modified SEIR model, called the *Susceptible-Exposed-Infectious-Recovered-Protection* (SEIRP) model, to study the outbreak of *Severe Acute Respiratory Syndrome* (SARS) in China.

While many epidemic models are available, they are inadequate to facilitate the unique epidemiological characteristics of COVID-19. In this paper, we propose a new epidemic model, called the *Susceptible-Exposed-Asymptomatic-Symptomatic-Active-Removed* (SEASAR) model, to delineate the COVID-19 transmission dynamic. This model generalizes the SIR and SEIR models by accounting for asymptomatic infections and the lag between symptoms onset and the diagnosis time, yet it does not require more assumptions required by the SIR and SEIR models. Consistent with many epidemic models, we make two rountine assumptions: (1) the population is homogeneous and well-mixed; and (2) there are no inbound and outbound travels. Similar to the SIR and SEIR models, the SEASAR model is a deterministic model which utilizes differential equations to describe the transmission dynamic of COVID-19. Furthermore, 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., Li et al., 2020), where resampling from Bayesian posterior distributions is basically invoked. We evaluate the performance of our transmission model and the estimation algorithm by analyzing the COVID-19 data in Quebec for the period of April 2, 2020 to May 10, 2020, where we focus on comparing differences between the predicted cumulative numbers of cases and the reported cumulative numbers of cases. We also conduct sensitivity analyses to assess how the estimation of the model parameters and the predicted results may change as the model assumptions are altered.

Analyzing the COVID-19 data in Quebec is driven by the following considerations. Quebec is the worst-hit province in Canada, and it is thereby interesting to examine the virus transmissions in this province. More importantly, it is imperative to ensure the required conditions to be met as much as possible when applying the model to analyze data. As pointed out earlier, the

validity of the proposed SEASAR model hinges on the assumption of no inbound and outbound travels, which is typically untrue in practice, but modeling data in a certain period is likely to make this assumption approximately true. For the period of April 2, 2020 to May 10, 2020, Quebec government set up checkpoints to block all non-essential travels into the province and people in Quebec were advised to stay home, thus, inbound and outbound travels in Quebec for this time window are perceived to be the least.

The remainder of this article is organized as follows. We develop the SEASAR model and elaborate on its rationale in Section 2. In Section 3, we present the initialization setup of the SEASAR model and describe the implementation procedure by adapting the IF-EAKF algorithm. In Section 4, we utilize the proposed SEASAR model to analyze the Quebec COVID-19 data for the period of April 2, 2020 to May 10, 2020. The article is concluded with a discussion presented in Section 5.

2 Model Framework

With a meta analysis, He et al. (2020) estimated that about 46% individuals with COVID-19 are asymptomatic, and they have the infectious ability to transmit the disease (e.g., Hao et al., 2020; Li et al., 2020). Due to the limited testing resources and the flu-like manifestation of COVID-19 as well as the incubation period, there is a time lag between symptoms onset and being confirmed for infected individuals (Kramer et al., 2020). To facilitate these features of COVID-19, we develop a new model, called the *Susceptible-Exposed-Asymptomatic-Symptomatic-Active-Removed* (SEASAR) model, under the assumptions of a well-mixed homogeneous population and of no inbound and outbound travels (i.e., the population size remains to be unchanged over time).

2.1 Illustration of the Proposed Model

To illustrate the ideas, we first consider a *static* framework by focusing at a given time point, say, on a given day. We divide the target population into six subpopulations with specific features, where 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., Peng et al., 2020), I_a stands for the subpopulation of *asymptomatic* infections (i.e., those with the infectious ability but showing no symptoms), I_s represents the subpopulation of *symptomatic* infections (i.e., those who show symptoms and have the infectious ability but are not confirmed yet), A is the subpopulation of *active* cases (i.e., those confirmed cases who do not recover or die), and R denotes the subpopulation of *removed* cases (i.e., those confirmed cases who recover or die from COVID-19).

Next, we introduce the parameters to facilitate the dynamic changes among the subpopulations. Let Z denote the *average latent period*, defined as the average time 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; He, Yi, and Zhu, 2020), so here we take Z as being available. Let θ denote the *symptomatic transmission rate*, defined as the average number of individuals infected by a symptomatic case per unit time (e.g., a day). Let the *asymptomatic transmission rate* be denoted as $\mu\theta$, which is 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 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 from symptoms onset to the time of being confirmed, let B denote the average time from being

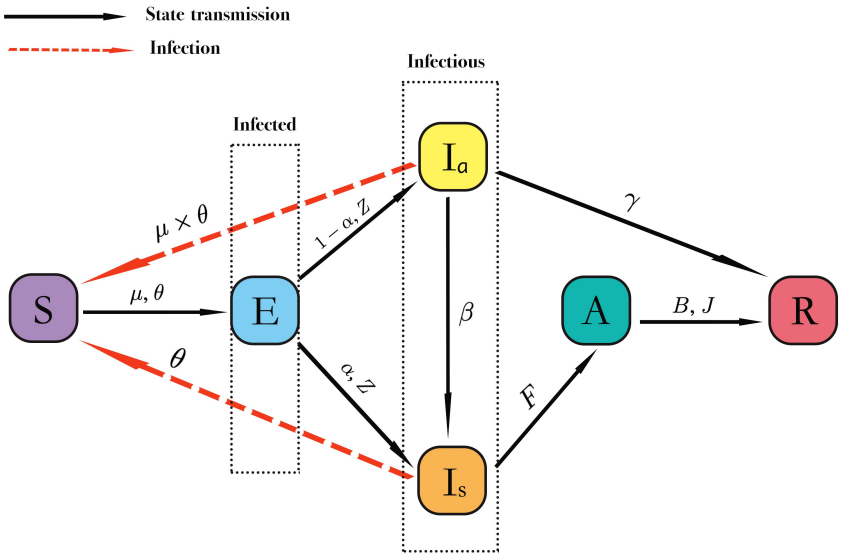

Figure 1: *Illustration of the SEASAR model. The population is divided into six compartments: S (susceptible), E (exposed), I_a (asymptomatic), I_s (symptomatic), A (active), and R (removed).*

confirmed to being recovered, and let J represent the average time from being confirmed to die.

Figure 1 is a flow chart showing the relationship among the six subpopulations. A black solid arrow between two subpopulations indicates the members of one subpopulation can move 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. Once COVID-19 cases are confirmed, they will be quarantined and lose the infectious ability, so the group S cannot be infected by the group A. Due to the limited test capacity, asymptomatic individuals are not being tested for COVID-19 (especially in the early stage of the outbreak). Thus, there is no transition from I_a to A. The parameters on black solid lines determine the number of people who change from one state to another state per unit time; and the parameters on red dashed lines determine the number of people infected by asymptomatic or symptomatic cases per unit time.

2.2 Dynamic Model

Figure 1 shows a static chart for the transmissions among the six subpopulations at a given time point. However, for any time period, the transmissions are not static but dynamic. To characterize the *dynamic* evolution of the subpopulations over time, we modify the six subpopulations by showing their dependence on time t, yielding the state components $\{S(t), E(t), I_a(t), I_s(t), A(t), R(t)\}$. For ease of notation, we also use the same symbols to represent the *sizes* of those state components. 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 subpopulation sizes at time t.

Given the setup, we now present the proposed SEASAR model, given by the following set of ordinary differential equations:

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

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

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

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

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

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

where N is the population size, which is time-invariant due to the assumptions of no outbound and inbound travels. Because of this assumption, $S(t)+E(t)+I_a(t)+I_s(t)+A(t)+R(t) = N$ for any time point t, and hence, any equation above is determined by other five equations.

For ease of referral of equations (1)-(6) in the later development, we express those equations in a compact form:

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

where $g(\cdot, \cdot)$ represents the vector function determined by the right hand side of (1)-(6). Fig-

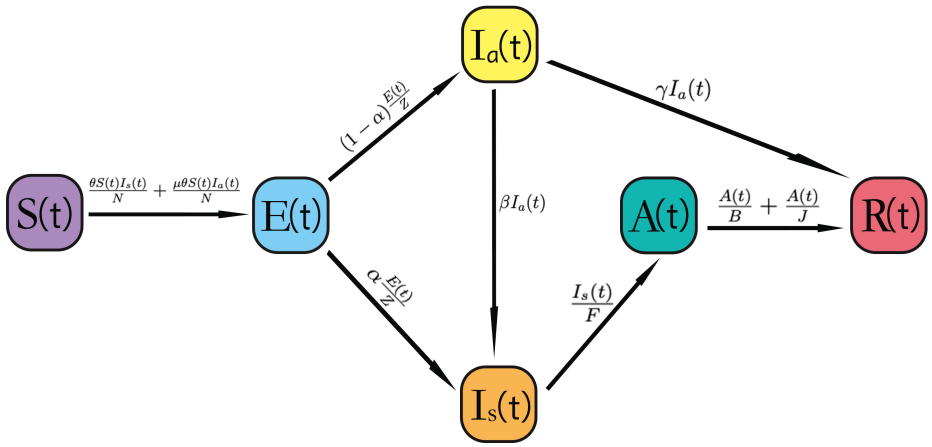

Figure 2: *Dynamic of the transmission among the six subpopulations at any time t*

ure 2 is a flowchart of the transmission dynamic for the six subpopulations together with the associated values.

The right hand side of each of equations (1)-(6) shows how the change rate of each subpopulation is related to the transmission rates as well as the associated parameters defined earlier. To see why, we first examine the right hand side of (1) by considering a small time interval $[t, t+\Delta t)$ of length Δt. Over the time interval $[t, t+\Delta t)$, because $\theta I_s(t)\Delta t$ represents the average number of people infected by symptomatic infections and $\frac{S(t)}{N}$ is the proportion of susceptible cases, so $\frac{\theta S(t)I_s(t)\Delta t}{N}$ equals the average number of *susceptible* cases who are infected by *symptomatic* infections; similarly, $\frac{\mu\theta S(t)I_a(t)\Delta t}{N}$ equals the average number of *susceptible* cases who are infected by *asymptomatic* infections. Thus, $\frac{\mu\theta S(t)I_a(t)\Delta t}{N} + \frac{\theta S(t)I_s(t)\Delta t}{N}$ records the average number of *susceptible* cases moving to the *exposed* state over the time interval $[t, t+\Delta t)$, i.e., the reduced number of *susceptible* cases over the time interval $[t, t+\Delta t)$ is $S(t+\Delta t) - S(t) = -\left\{\frac{\theta S(t)I_s(t)\Delta t}{N} + \frac{\mu\theta S(t)I_a(t)\Delta t}{N}\right\}$. Then dividing Δt on the both sides and

letting $\Delta t \to 0$ yields (1).

Equations (2)-(6) can be illustrated in an analogous way. Regarding (2), due to the changing from susceptible cases to being in the *exposed* state, the number of *exposed* cases is increased by $\frac{\mu\theta S(t)I_a(t)\Delta t}{N} + \frac{\theta S(t)I_s(t)\Delta t}{N}$ over the interval $[t, t+\Delta t)$; yet due to the latent period, the number of *exposed* cases is reduced by $\frac{E(t)\Delta t}{Z}$ over the interval $[t, t+\Delta t)$, yielding that the difference in the number of *exposed* cases over the time interval $[t, t+\Delta t)$ is $E(t+\Delta t) - E(t) = \frac{\mu\theta S(t)I_a(t)\Delta t}{N} + \frac{\theta S(t)I_s(t)\Delta t}{N} - \frac{E(t)\Delta t}{Z}$. Dividing Δt on the both sides and letting $\Delta t \to 0$ yields (2).

Regarding (3), when *exposed* cases pass the latent period, they become either *asymptomatic* or *symptomatic* infections. As the fraction of *asymptomatic* infections is $1-\alpha$, the number of asymptomatic infections is thus increased by $(1-\alpha)\frac{E(t)\Delta t}{Z}$ over the interval $[t, t+\Delta t)$. On the other hand, $\beta I_a(t)\Delta t$ is the average number of infections who initially are *asymptomatic* but later show symptoms over the interval $[t, t+\Delta t)$ (i.e., people in this group turn from the *asymptomatic* state to the *symptomatic* state), and $\gamma I_a(t)\Delta t$ is the average number of *asymptomatic* infections who recover from COVID-19 (i.e., people in this group move from the *asymptomatic* state to the *removed* state). Thus, the number of asymptomatic infections is reduced by $\beta I_a(t)\Delta t + \gamma I_a(t)\Delta t$ over the interval $[t, t+\Delta t)$. Therefore, the difference in the number of asymptomatic infections is $I_a(t+\Delta t) - I_a(t) = (1-\alpha)\frac{E(t)\Delta t}{Z} - \beta I_a(t)\Delta t - \gamma I_a(t)\Delta t$, yielding (3) by dividing Δt on the both sides and letting $\Delta t \to 0$.

Regarding (4), due to the change of individuals from the *exposed* and *asymptomatic* states to the *symptomatic* state, the number of symptomatic infections is increased by $\alpha\frac{E(t)\Delta t}{Z} + \beta I_a(t)\Delta t$ over the interval $[t, t+\Delta t)$. On the other hand, $\frac{I_s(t)\Delta t}{F}$ is the average number of *symptomatic* infections who are confirmed over the interval $[t, t+\Delta t)$ (i.e., people in this group turn from the *symptomatic* state to the *active* state); in other words, the reduced number of *symptomatic* infections over the interval $[t, t+\Delta t)$ is $\frac{I_s(t)\Delta t}{F}$. Thus, the difference in the number of *symptomatic* infections is $I_s(t+\Delta t) - I_s(t) = \alpha\frac{E(t)\Delta t}{Z} - \frac{I_s(t)\Delta t}{F} + \beta I_a(t)\Delta t$, leading to (4) if dividing

Δt on the both sides and letting $\Delta t \to 0$.

Regarding (5), due to the changing from the *symptomatic* state to the *active* state, the number of active cases is increased by $\frac{I_s(t)\Delta t}{F}$ over the interval $[t, t+\Delta t)$. Since $\frac{A(t)\Delta t}{B}$ is the average number of *active* cases who recover from COVID-19 and $\frac{A(t)\Delta t}{J}$ is the average number of *active* cases who die from COVID-19 over the interval $[t, t+\Delta t)$ (i.e., people in the two groups turn from the *active* state to the *removed* state), the number of active cases is reduced by $\frac{A(t)\Delta t}{B} + \frac{A(t)\Delta t}{J}$ over the interval $[t, t+\Delta t)$. Therefore, the difference in the number of active cases is $A(t+\Delta t) - A(t) = \frac{I_s(t)\Delta t}{F} - \frac{A(t)\Delta t}{B} - \frac{A(t)\Delta t}{J}$, yielding (5) if we divide Δt on the both sides and let $\Delta t \to 0$.

Regarding (6), due to the changing from the *active* and *asymptomatic* state to the *removed* state, the number of removed cases is increased by $\gamma I_a(t)\Delta t + \frac{A(t)\Delta t}{B} + \frac{A(t)\Delta t}{J}$ over the interval $[t, t+\Delta t)$, i.e., $R(t+\Delta t) - R(t) = \gamma I_a(t)\Delta t + \frac{A(t)\Delta t}{B} + \frac{A(t)\Delta t}{J}$. Thus, dividing Δt on the both sides and letting $\Delta t \to 0$ yields (6). Alternatively, (6) equals the negative sum of (1) to (5).

2.3 Basic Reproduction Number

Let $\eta = (\theta, \mu, \alpha, \beta, \gamma, F, B, J)^{\mathrm{T}}$ denote the vector of parameters of prime interest. Knowing the value of η allows us to describe the six subpopulations sizes using the models (1)-(6). Further, it enables us to describe the severity of the pandemic using some simple measure such as 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 more 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 the dying down situation. In Appendix A, we show that the mathematical expression of R_0 derived from the SEASAR model is

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

3 Estimation Procedure

In this section, we develop an estimation procedure for η by adapting the iterated filter-ensemble adjustment Kalman filter (IF-EAKF) algorithm (e.g., Li et al., 2020) in combination with the 4th order Runge-Kutta (RK4) method (e.g., Süli and Mayers, 2003, p.328).

3.1 Initial Sizes of Subpopulations

Let τ_0 denote the initial time point from which we start examining the data. By time τ_0, let R_c, C_0 and D_0 denote the reported cumulative number of recoveries, confirmed cases and deaths from COVID-19, respectively, which are available; and let R_a denote the cumulative number of recovered asymptomatic cases, which is unavailable. Let $Q(t)$ denote the number of reported cases with symptoms onset on day t.

To facilitate the relationship between the observed and unobserved values, let $r_1 = \frac{R_a}{R_c}$ denote the ratio of the unobserved R_a to the observed R_c, and let $r_2 = \frac{I_a(\tau_0)}{I_s(\tau_0)}$ represent the ratio of unobserved size $I_a(\tau_0)$ to the reported size $I_s(\tau_0)$. Motived by Hao et al. (2020), we express the size $E(\tau_0)$ in terms of its relative value to the total number of reported cases with symptoms onset over the time window of length Z, $E(\tau_0) = r_3 \times \sum_{t=\tau_0}^{[\tau_0+Z]} Q(t)$, where the function $[x]$ represents the biggest integer that is less than or equal to x, and r_3 is a positive value.

While the introduction of the ratios r_1, r_2 and r_3 does not give us a way to determine the unobserved values R_a, $I_a(\tau_0)$ and $E(\tau_0)$ using the observed data, these ratios offer us convenient measures to describe the pandemic situation in relative scales of the observed values at time τ_0. For instance, at the early stage of the pandemic, the testing kits are limited, so the number of recoveries from *confirmed* cases is likely to be a lot smaller than that from *asymptomatic* individuals, suggesting a large value of r_1. If r_2 is bigger than 1, then there are more *asymptomatic* infections than *symptomatic* infections.

Table 1 summarizes the initial sizes of the six subpopulations which are to be used in Section

Table 1: Initial state size for the SEASAR model

Variable	Meaning	Value
$S(\tau_0)$	The initial number of susceptible cases	$N-E(\tau_0)-I_a(\tau_0)-I_s(\tau_0)-A(\tau_0)-R(\tau_0)$
$E(\tau_0)$	The initial number of exposed cases	$r_3 \times \sum_{t=\tau_0}^{[\tau_0+Z]} Q(t)$
$I_a(\tau_0)$	The initial number of asymptomatic infections	$r_2\{\sum_{t\leq\tau_0} Q(t) - C_0\}$
$I_s(\tau_0)$	The initial number of symptomatic infections	$\sum_{t\leq\tau_0} Q(t) - C_0$
$A(\tau_0)$	The initial number of active cases	$C_0 - R_c - D_0$
$R(\tau_0)$	The initial number of removed cases	$D_0 + (1 + r_1)R_c$

3.3, where the values of r_1, r_2 and r_3 are specified. Write $\phi(\tau_0) = (S(\tau_0), E(\tau_0), I_a(\tau_0), I_s(\tau_0), A(\tau_0), R(\tau_0))^{\mathrm{T}}$.

3.2 Initialization of Model Parameters

Estimation of η, to be described in Section 3.3, is conducted by an iterative approximation procedure using the posterior distributions of η. Here we assume the prior information for the parameters to be non-informative except for constraining the parameters to certain ranges to reflect our a priori knowledge about them.

To be specific, the transmission rate θ of symptomatic infections is considered to be $0 \leq \theta \leq 7$ 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 $0.1 \leq \mu \leq 1$, the fraction α of symptomatic infections relative to all infections is restricted to be $0.1 \leq \alpha \leq 1$, the average

rate β of asymptomatic infections who develop symptoms per unit time is constrained to be $0.0002 \leq \beta \leq 0.8$, and the average recovery rate γ of asymptomatic infections per unit time is set as $0.1 \leq \gamma \leq 1$. The average time F from symptoms onset to being confirmed is considered to be between 1 and 10 days based on the study of Kramer et al. (2020). Based on the WHO report (WHO, 2020), the average time B from being confirmed to being recovered is considered to be in the range of 7 to 42 days, and the average time J from being confirmed to die is taken to change from 14 to 56 days.

3.3 IF-EAKF algorithm

With the setup in Sections 3.1 and 3.2, we describe an estimation procedure by adapting the iterated filtering (IF) algorithm. The IF approach basically produces the maximum likelihood estimates of model parameters and has been successfully applied to study many infectious diseases (Li et al., 2020), including cholera (e.g., King et al., 2008) and measles (e.g., He et al., 2010). An efficient IF approach roots in using the *ensemble adjustment Kalman filter* (EAKF) which can generate satisfactory results with only hundreds of samples (Li et al., 2020).

Since the number of confirmed cases is reported on a daily basis, we take the time unit as a day. Consider a population of interest. In contrast to the number $Q(t)$ of reported cases with symptoms onset on day t, defined in Section 3.1, we let $Y(t)$ denote the number of confirmed cases to be reported on day t, which is regarded as a random variable, and let y_t denote its realization. Let $\hat{Y}(t) = \frac{I_s(t)}{F}$ denote the number of confirmed cases on day t that is generated from the SEASAR model. We assume that $Y(t)$ and $\hat{Y}(t)$ are connected through the model

$$\log Y(t) = \log \hat{Y}(t) + \epsilon_t, \tag{7}$$

where the ϵ_t are independent of each other and of the $\hat{Y}(t)$, $\phi(t)$ and η; and ϵ_t follows a Gaussian distribution with mean 0 and time-dependent variance σ_t^2.

Being called the *observational error variance* (OEV) by Li et al. (2020), σ_t^2 is often estimated heuristically via an assumed function form of the observations. For example, in the analysis in Section 4.2, σ_t^2 is assumed to be

$$\sigma_t^2 = max\left(1, \frac{\log y_t}{4}\right) \quad \text{for} \quad t \geq \tau_0. \tag{8}$$

Similar forms of OEV were used in the studies of other infectious diseases, including influenza (e.g., Pei et al., 2018), Ebola (e.g., Shaman et al., 2014), West Nile virus (e.g., DeFelice et al., 2017), and respiratory syncytial virus (e.g., Reis and Shaman, 2016).

The model parameter η for (1)-(6), the subpopulation sizes $\phi(t)$ and $\hat{Y}(t)$ at time t for $t \geq \tau_0$ are unknown. One may consider the joint posterior distribution of η, $\log\ \hat{Y}(t)$ and $\log\ \phi(t)$, and then use the mean of the posterior distribution of η to estimate η, where $\log\ \phi(t) := (\log S(t), \log E(t), \log I_a(t), \log I_s(t), \log A(t), \log R(t))^{\mathrm{T}}$. To reduce computation costs, we adapt the discussion of (Anderson, 2001, p.2888) and describe a simple iterative procedure by pairing two members in $\{\eta, \log\ \hat{Y}(t), \log\ \phi(t)\}$, where we typically pair $\log \hat{Y}(t)$ with each component in $\{\eta, \log\ \phi(t)\}$. Let $\tau_0 + T - 1$ denote the time point by which we stop iterations. To see the main ideas, here we elaborate on the detail of the first iteration for the IF-EAKF algorithm which includes the following six stages, with similar details for other iterations omitted.

- **Stage 1:** At time $t = \tau_0$, generate prior values for η and $\hat{Y}(t)$:
  - **Step 1:** Let π_η denote a prior distribution for parameter η, which is taken as the uniform distribution over the ranges specified in Section 3.2, with the assumption that the parameter components in η are independent of each other.
  - **Step 2:** Specify a positive integer, say n. Then generate n values from π_η, denoted $\{\eta^i_{\text{pri},\tau_0} : i = 1, \cdots, n\}$, where for each i, $\eta^i_{\text{pri},\tau_0} = (\theta^i_{\text{pri},\tau_0}, \mu^i_{\text{pri},\tau_0}, \alpha^i_{\text{pri},\tau_0}, \beta^i_{\text{pri},\tau_0}, \gamma^i_{\text{pri},\tau_0}, F^i_{\text{pri},\tau_0}, B^i_{\text{pri},\tau_0}, J^i_{\text{pri},\tau_0})^{\mathrm{T}}$.

– **Step 3:** Using the initial size $I_s(\tau_0)$ of symptomatic infections, we generate n prior values for $\hat{Y}(\tau_0)$, denoted $\{\hat{Y}^i_{\text{pri}}(\tau_0) : i = 1, \cdots, n\}$, by setting $\hat{Y}^i_{\text{pri}}(\tau_0) = \frac{I_s(\tau_0)}{F^i_{\text{pri},\tau_0}}$ for $i = 1, \cdots, n$. Then calculate the sample mean and variance:

$$\overline{o}_{\text{pri},\tau_0} = \frac{\sum_{i=1}^{n} \log \hat{Y}^i_{\text{pri}}(\tau_0)}{n} \text{ and } \sigma^2_{\text{pri},\tau_0} = \frac{\sum_{i=1}^{n} \left\{ \log \hat{Y}^i_{\text{pri}}(\tau_0) - \overline{o}_{\text{pri},\tau_0} \right\}^2}{n-1},$$

together with the pairwise sample covariances:

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

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

- **Stage 2:** At time $t = \tau_0$, generate posterior values for η and $\hat{Y}(t)$:

We employ the following steps to generate n posterior values for $\hat{Y}(\tau_0)$ and η from their posterior distribution, derived in Appendix B.

– **Step 1:** Generate n posterior values for $\hat{Y}(\tau_0)$, denoted $\{\hat{Y}^i_{\text{post}}(\tau_0) : i = 1, \cdots, n\}$. For each i, $\hat{Y}^i_{\text{post}}(\tau_0)$ is determined by

$$\log \hat{Y}^i_{\text{post}}(\tau_0) = \frac{\sigma^2_{\tau_0}}{\sigma^2_{\tau_0} + \sigma^2_{\text{pri},\tau_0}} \overline{o}_{\text{pri},\tau_0} + \frac{\sigma^2_{\text{pri},\tau_0}}{\sigma^2_{\tau_0} + \sigma^2_{\text{pri},\tau_0}} \log y_{\tau_0} + \sqrt{\frac{\sigma^2_{\tau_0}}{\sigma^2_{\tau_0} + \sigma^2_{\text{pri},\tau_0}}} \left\{ \log \hat{Y}^i_{\text{pri}}(\tau_0) - \overline{o}_{\text{pri},\tau_0} \right\}, \quad (9)$$

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

– **Step 2:** Generate n posterior values for η, denoted $\{\eta^i_{\text{post},\tau_0} : i = 1, \cdots, n\}$, where for each i, $\eta^i_{\text{post},\tau_0} = (\theta^i_{\text{post},\tau_0}, \mu^i_{\text{post},\tau_0}, \alpha^i_{\text{post},\tau_0}, \beta^i_{\text{post},\tau_0}, \gamma^i_{\text{post},\tau_0}, F^i_{\text{post},\tau_0}, B^i_{\text{post},\tau_0}, J^i_{\text{post},\tau_0})^{\text{T}}$. Specifically, each component of $\eta^i_{\text{post},\tau_0}$ is given by

$$X^i_{\text{post},\tau_0} = X^i_{\text{pri},\tau_0} + \left(\frac{\sigma^{cov}_{X,\log \hat{Y}(\tau_0),\text{pri}}}{\sigma^2_{\text{pri},\tau_0}} \right) \{ \log \hat{Y}^i_{\text{post}}(\tau_0) - \log \hat{Y}^i_{\text{pri}}(\tau_0) \}, \quad (10)$$

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

- **Stage 3:** At time $t = \tau_0 + 1$, generate prior values for η, $\phi(t)$, and $\hat{Y}(t)$:

  - **Step 1:** Generate n prior values for η, denoted $\{\eta^i_{\text{pri},\tau_0+1} : i = 1, \cdots, n\}$, by setting $\eta^i_{\text{pri},\tau_0+1} = \eta^i_{\text{post},\tau_0}$ for $i = 1, \cdots, n$, where we denote $\eta^i_{\text{pri},\tau_0+1} = (\theta^i_{\text{pri},\tau_0+1}, \mu^i_{\text{pri},\tau_0+1}, \alpha^i_{\text{pri},\tau_0+1}, \beta^i_{\text{pri},\tau_0+1}, \gamma^i_{\text{pri},\tau_0+1}, F^i_{\text{pri},\tau_0+1}, B^i_{\text{pri},\tau_0+1}, J^i_{\text{pri},\tau_0+1})^{\text{T}}$ for each i.

  - **Step 2:** Using the RK4 method, we generate n prior values for $\phi(\tau_0 + 1)$, denoted $\{\phi^i_{\text{pri}}(\tau_0+1) : i = 1, \cdots, n\}$, where for each i, $\phi^i_{\text{pri}}(\tau_0+1) = (S^i_{\text{pri}}(\tau_0+1), E^i_{\text{pri}}(\tau_0+1), I^i_{a_{\text{pri}}}(\tau_0+1), I^i_{s_{\text{pri}}}(\tau_0+1), A^i_{\text{pri}}(\tau_0+1), R^i_{\text{pri}}(\tau_0+1))^{\text{T}}$. Specifically, let $k_1(t) = g(\phi(t), \eta^i_{\text{pri},t+1})$, $k_2(t) = g(\phi(t) + \frac{k_1(t)}{2}, \eta^i_{\text{pri},t+1})$, $k_3(t) = g(\phi(t) + \frac{k_2(t)}{2}, \eta^i_{\text{pri},t+1})$, and $k_4(t) = g(\phi(t) + k_3(t), \eta^i_{\text{pri},t+1})$. 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 $\hat{Y}(\tau_0+1)$, denoted $\{\hat{Y}^i_{\text{pri}}(\tau_0+1) : i = 1, \cdots, n\}$, by setting $\hat{Y}^i_{\text{pri}}(\tau_0 + 1) = \frac{I^i_{s_{\text{pri}}}(\tau_0+1)}{F^i_{\text{pri},\tau_0+1}}$ for $i = 1, \cdots, n$. Then we calculate

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

  together with the pairwise sample covariance

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

  where X is a symbol for S, E, I_a, I_s, A, and R.

- **Stage 4:** At time $t = \tau_0 + 1$, generate posterior values for η, $\phi(t)$, and $\hat{Y}(t)$. This stage is similar to Stage 2.

  - **Step 1:** Similar to Stage 2, we generate n posterior values for $\hat{Y}(\tau_0 + 1)$ and η, denoted $\{\hat{Y}^i_{\text{post}}(\tau_0 + 1) : i = 1, \cdots, n\}$ and $\{\eta^i_{\text{post},\tau_0+1} : i = 1, \cdots, n\}$, respectively.

– **Step 2:** Generate n posterior values for $\phi(\tau_0+1)$, denoted $\{\phi^i_{\text{post}}(\tau_0+1) : i = 1,\cdots,n\}$, where for each i, $\phi^i_{\text{post}}(\tau_0+1) = (S^i_{\text{post}}(\tau_0+1), E^i_{\text{post}}(\tau_0+1), I^i_{a_{\text{post}}}(\tau_0+1), I^i_{s_{\text{post}}}(\tau_0+1), A^i_{\text{post}}(\tau_0+1), R^i_{\text{post}}(\tau_0+1))^{\text{T}}$. Specifically, each component of $\phi^i_{\text{post}}(\tau_0+1)$ is given by

$$\begin{aligned}\log X^i_{\text{post}}(\tau_0+1) = \ & \log X^i_{\text{pri}}(\tau_0+1) \\ & + \left(\frac{\sigma^{cov}_{\log X(\tau_0+1),\log \hat{Y}(\tau_0+1),\text{pri}}}{\sigma^2_{\text{pri},\tau_0+1}}\right)\left\{\log \hat{Y}^i_{\text{post}}(\tau_0+1) - \log \hat{Y}^i_{\text{pri}}(\tau_0+1)\right\},\end{aligned}$$

where X is a symbol for S, E, I_a, I_s, A, and R.

- **Stage 5:** At time $t = \tau_0+2$, generate prior values for η, $\phi(t)$, and $\hat{Y}(t)$:

  – **Step 1:** Generate n prior values for η, denoted $\{\eta^i_{\text{pri},\tau_0+2} : i = 1,\cdots,n\}$, by setting $\eta^i_{\text{pri},\tau_0+2} = \eta^i_{\text{post},\tau_0+1}$ for $i = 1,\cdots,n$.

  – **Step 2:** Generate n prior values for $\phi(\tau_0+2)$, denoted $\{\phi^i_{\text{pri}}(\tau_0+2) : i = 1,\cdots,n\}$, using the RK4 method, where similar to (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 in Stage 3, we generate n prior values for $\hat{Y}(\tau_0+2)$.

- **Stage 6:** Similar to Stage 4, we generate n posterior values of η, $\phi(\tau_0+2)$ and $\hat{Y}(\tau_0+2)$.

We repeat Stages 5-6 for $t = \tau_0+3,\cdots,\tau_0+T-1$, and obtain a sequence of posterior values for η, denoted $\{\eta^i_{\text{post},t} : i = 1,\cdots,n; t = \tau_0,\tau_0+1,\cdots,\tau_0+T-1\}$. Then we calculate

$$\overline{x}_1 = \frac{1}{nT}\sum_{t=\tau_0}^{\tau_0+T-1}\sum_{i=1}^{n}\eta^i_{\text{post},t}$$

and use $\overline{x}_1$ for the next iteration.

The proceding descriptions show the steps for the first iteration; steps for subsequent iterations, similar to the first iteration, are outlined in Algorithm 1, where Σ is the covariance matrix

of the prior distribution π_η which is shrunk by a discount factor $a \in (0, 1)$, usually pre-specified to reduce the variance of the posterior distribution of the parameters as the iteration progresses. Using a discount factor is a standard procedure to ensure that $\overline{x}_L$ in Algorithm 1, the estimate at the Lth iteration, is nearly identical to the maximum likelihood estimate of η. If a is too small, the algorithm may "quench" too fast and fail to find the maximum likelihood estimate; if it is too close to 1, the algorithm may not converge in a reasonable time (see Li et al., 2020, Supplementary Materials, p.8). Practically, a can range between 0.9 to 0.99. The number L of iterations is set in an ad hoc way, often determined by inspecting the evolution of the posterior parameter distributions (see Li et al., 2020, Supplementary Materials, p.8). The scale of n is in hundreds, though in principle, the larger the better.

4 Data Analysis

4.1 Quebec COVID-19 Data

On April 1, 2020, Quebec provincial government in Canada set up checkpoints to block all non-essential travels into the province and advised people in Quebec to stay home. After May 10, 2020, Quebec and other provinces in Canada gradually reopened the economy, yielding inbound and outbound travels in Quebec. Therefore, it is reasonable to perceive inbound and outbound travels in the period of April 2, 2020 to May 10, 2020 in Quebec to be the least, and the assumption of no inbound and outbound travels required by the proposed SEASAR model is relatively feasible for the Quebec COVID-19 data in this period.

Driven by this, we apply the proposed model to analyze the daily reported number of COVID-19 cases in Quebec, Canada, in this period, available at https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection.html. We take the initial time point τ_0 as April 2, 2020 and split the study period into two parts: the period of April 1 to April 30, 2020, and the period of May 1 to May 10, 2020. The data for the first period are used to estimate

Algorithm 1: IF-EAKF

Input: the sequence $\{y_t : t = \tau_0, \cdots, \tau_0 + T - 1\}$ of daily reported confirmed cases, the sequence $\{\sigma_t^2 : t = \tau_0, \cdots, \tau_0 + T - 1\}$, the covariance matrix Σ, 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 $\{\eta^i_{\text{pri},\tau_0} : i = 1, \cdots, n\}$ for parameters at time τ_0 independently from distribution π_η;

  else

    Generate n prior values, $\{\eta^i_{\text{pri},\tau_0} : i = 1, \cdots, n\}$, for parameters at time τ_0 independently from a multivariate Gaussian distribution $\mathcal{N}(\overline{x}_{l-1}, a^{(l-1)}\Sigma)$, where $\overline{x}_{l-1}$ 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 $\{\hat{Y}^i_{\text{pri}}(\tau_0) : i = 1, \cdots, n\}$ for $\hat{Y}(\tau_0)$ based on $\{\eta^i_{\text{pri},\tau_0} : i = 1, \cdots, n\}$ and $\phi(\tau_0)$;

  Generate n posterior values $\{\eta^i_{\text{post},\tau_0} : i = 1, \cdots, n\}$ for parameters and n posterior values $\{\hat{Y}^i_{\text{post}}(\tau_0) : i = 1, \cdots, n\}$ for $\hat{Y}(\tau_0)$ based on their prior values, $\sigma^2_{\tau_0}$ and observation y_{τ_0};

  Generate n prior values $\{\eta^i_{\text{pri},\tau_0+1} : i = 1, \cdots, n\}$ for parameters at time $\tau_0 + 1$ by setting $\eta^i_{\text{pri},\tau_0+1} = \eta^i_{\text{post},\tau_0}$ for $i = 1, \cdots, n$. RK4 method is used to generate n prior values $\{\phi^i_{\text{pri}}(\tau_0 + 1) : i = 1, \cdots, n\}$ for $\phi(\tau_0 + 1)$ based on equations (1)-(6).

  Generate n prior values $\{\hat{Y}^i_{\text{pri}}(\tau_0 + 1) : i = 1, \cdots, n\}$ for $\hat{Y}(\tau_0 + 1)$;

  for $t = \tau_0 + 1$ *to* $\tau_0 + T - 1$ **do**

    Generate n posterior values $\{\eta^i_{\text{post},t} : i = 1, \cdots, n\}$ for parameters at time t, n posterior values $\{\phi^i_{\text{post}}(t) : i = 1, \cdots, n\}$ for $\phi(t)$, and n posterior values $\{\hat{Y}^i_{\text{post}}(t) : i = 1, \cdots, n\}$ for $\hat{Y}(t)$ based on their prior values, σ^2_t and observation y_t;

    Generate n prior values $\{\eta^i_{\text{pri},t+1} : i = 1, \cdots, n\}$ for parameters at time $t + 1$ by setting $\eta^i_{\text{pri},t+1} = \eta^i_{\text{post},t}$ for $i = 1, \cdots, n$. RK4 method is used to generate n prior values $\{\phi^i_{\text{pri}}(t + 1) : i = 1, \cdots, n\}$ for $\phi(t + 1)$. Generate n prior values $\{\hat{Y}^i_{\text{pri}}(t + 1) : i = 1, \cdots, n\}$ for $\hat{Y}(t + 1)$;

  end

  Calculate the mean: $\overline{x}_l = \frac{1}{nT} \sum_{t=\tau_0}^{\tau_0+T-1} \sum_{i=1}^{n} \eta^i_{\text{post},t}$.

end

Output: $\overline{x}_L$.

the model parameters by applying Algorithm 1, and the data in the second period are used to assess the prediction performance of our proposed model.

4.2 Estimation of Model Parameters

We run Algorithm 1 to the proposed SEASAR model, where we set $n = 300$, $a = 0.9$, $L = 50$, and the average latent period Z is taken as 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 3.1, we take $r_1 = 2$, $r_2 = 1$, and $r_3 = 2$. As described in Section 3.1, the initial sizes of subpopulations depend on the daily number of COVID-19 patients with symptoms onset before May 10, 2020, which is not available in Quebec. However, the daily number of COVID-19 patients with symptoms onset in Canada is collected by the Public Health Agency of Canada (https://health-infobase.canada.ca/covid-19/epidemiological-summary-covid-19-cases.html). Using this information, we approximate the daily number of COVID-19 patients with symptoms onset before May 10, 2020 in Quebec. Specifically, let $n_{t,C}$ denote the number of COVID-19 patients with symptoms onset on day t in Canada, and let $m_{t,C}$ and $m_{t,Q}$ denote the number of confirmed COVID-19 cases on day t in Canada and Quebec, respectively. We then take $\frac{n_{t,C}m_{t,Q}}{m_{t,C}}$ as an estimated number of COVID-19 patients with symptoms onset on day t in Quebec.

To account for stochastic effects, we run the IF-EAKF algorithm 1000 times and display in Figure 4 the histograms for those estimates. A 95% confidence interval (CI) for each parameter is determined by using the 2.5% and 97.5% percentiles of the 1000 estimates as its lower and upper bounds, respectively. Table 2 reports the average estimates and the associated 95% CIs for the parameters, together with the average of those 1000 estimated *basic reproduction numbers* and its associated 95% CI. The estimate of R_0 suggests that the pandemic situation in Quebec for the period of April 2, 2020 to April 30, 2020 is not under control and the virus spread continues.

Table 2: The average estimates and 95% CIs for the parameters

Parameter	Average Estimate	95% CI
θ	0.353	(0.307, 0.407)
μ	0.458	(0.390, 0.520)
α	0.276	(0.243, 0.313)
β	0.207	(0.161, 0.263)
γ	0.671	(0.611, 0.729)
F	6.210	(5.790, 6.660)
B	24.479	(21.820, 27.197)
J	34.975	(31.598, 38.196)
R_0	1.116	(1.053, 1.173)

4.3 Dynamic Projection of Subpopulation Sizes

To project possible trajectories of the pandemic in Quebec, Canada, we are interested in visualizing the estimated sizes for the six subpopulations in the next 1000 days. To this end, let $\hat{\eta}$ denote the vector of the parameter estimates reported in Table 2. Let $\hat{k}_1(t) = g(\phi(t), \hat{\eta})$, $\hat{k}_2(t) = g(\phi(t) + \frac{\hat{k}_1(t)}{2}, \hat{\eta})$, $\hat{k}_3(t) = g(\phi(t) + \frac{\hat{k}_2(t)}{2}, \hat{\eta})$, and $\hat{k}_4(t) = g(\phi(t) + \hat{k}_3(t), \hat{\eta})$. Then similar to (11), we apply the RK4 method to estimate $\phi(t)$ recursively for $t = \tau_0 + 1, \tau_0 + 2, \cdots, \tau_0 + 1000$:

$$\hat{\phi}(t+1) = \hat{\phi}(t) + \frac{\hat{k}_1(t) + 2\hat{k}_2(t) + 2\hat{k}_3(t) + \hat{k}_4(t)}{6},$$

where the initial value $\phi(\tau_0)$ is given in Section 3.1.

Figure 3 presents smooth lines connecting the estimates of an element of $\hat{\phi}(t)$ for the time points from $t = \tau_0$ to $t = \tau_0 + 1000$, where $S(\tau_0)$ is roughly equal to $N - \hat{E}(\tau_0) - \hat{I}_a(\tau_0) - \hat{I}_s(\tau_0) - \hat{A}(\tau_0) - \hat{R}(\tau_0)$, with $\hat{E}(\tau_0)$, $\hat{I}_a(\tau_0)$, $\hat{I}_s(\tau_0)$, $\hat{A}(\tau_0)$ and $\hat{R}(\tau_0)$ respectively denoting estimates for the corresponding subpopulation size and N being the population size of Quebec. Figure 3 shows the patterns that may be useful for us to project the pandemic evolution in Quebec in the next 1000 days if no interference measures are taken. The susceptible subpopulation size $S(t)$ decreases monotonically, suggesting that as time goes by, more people would move to the subpopulation of exposed cases. $E(t)$, $I_a(t)$, $I_s(t)$ and $A(t)$ have similar shape with a single

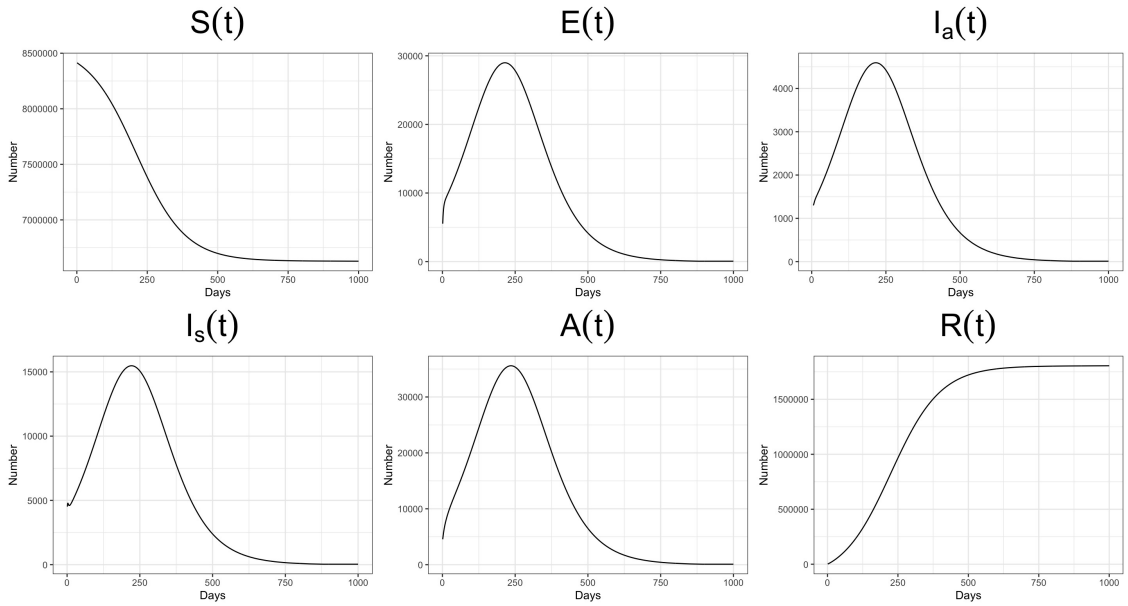

Figure 3: *The estimated curves of* $S(t)$, $E(t)$, $I_a(t)$, $I_s(t)$, $A(t)$, *and* $R(t)$.

peak before 250 days. The size $R(t)$ of removed cases is an increasing function of time t, and it becomes fairly flat after 600 days.

4.4 Prediction of Confirmed Cases

With the estimated parameters for the SEASAR model using the data from April 2, 2020 to April 30, 2020, we now predict the daily number of cases for the period of May 1, 2020 to May 10, 2020, and also compare them to the actually reported daily cases in this period. For each i, the ith estimate and prediction are generated as follows:

- **Step 1:** Let $\hat{\eta}_i = (\hat{\theta}_i, \hat{\mu}_i, \hat{\alpha}_i, \hat{\beta}_i, \hat{\gamma}_i, \hat{F}_i, \hat{B}_i, \hat{J}_i)^{\mathrm{T}}$ denote the ith estimate of parameters, then $\frac{I_s(\tau_0)}{\hat{F}_i}$ is the ith estimate of the number of cases on April 2, 2020.

- **Step 2:** Let $\hat{k}_{1i}(t) = g(\phi(t), \hat{\eta}_i)$, $\hat{k}_{2i}(t) = g(\phi(t) + \frac{\hat{k}_{1i}(t)}{2}, \hat{\eta}_i)$, $\hat{k}_{3i}(t) = g(\phi(t) + \frac{\hat{k}_{2i}(t)}{2}, \hat{\eta}_i)$, and $\hat{k}_{4i}(t) = g(\phi(t) + \hat{k}_{3i}(t), \hat{\eta}_i)$, then similar to (11), the ith estimate of the sizes of the six subpopulations at the beginning of April 3, 2020, denoted $\hat{\phi}_i(\tau_0 + 1) = (\hat{S}_i(\tau_0 +$

$1), \hat{E}_i(\tau_0+1), \hat{I}_{a_i}(\tau_0+1), \hat{I}_{s_i}(\tau_0+1), \hat{A}_i(\tau_0+1), \hat{R}_i(\tau_0+1))^{\mathsf{T}}$, is given by

$$\hat{\phi}_i(\tau_0+1) = \phi(\tau_0) + \frac{\hat{k}_{1i}(\tau_0) + 2\hat{k}_{2i}(\tau_0) + 2\hat{k}_{3i}(\tau_0) + \hat{k}_{4i}(\tau_0)}{6},$$

and thus, $\frac{\hat{I}_{s_i}(\tau_0+1)}{\hat{F}_i}$ is taken as the ith estimate of the number of cases on April 3, 2020.

Repeating Step 2 for $\tau_0+2, \cdots, \tau_0+28$, we obtain the ith estimate of the daily number of cases from April 2, 2020 to April 30, 2020. Further repeating Step 2 for $\tau_0+29, \cdots, \tau_0+38$, we obtain the ith prediction of the daily number of cases from May 1, 2020 to May 10, 2020.

To evaluate the differences between the predicted and reported daily number of cases from May 1 to May 10, we calculate the *mean absolute error* (MAE), $\frac{1}{1000}\sum_{i=1}^{1000}|\hat{y}_t^i - y_t|$, and the *relative mean absolute error* (RMAE), $\frac{1}{1000}\sum_{i=1}^{1000}\frac{|\hat{y}_t^i - y_t|}{y_t}$, for day t, where for $i = 1, \cdots, 1000$, $\hat{y}_t^i$ stands for the predicted number of cases on day t, and y_t is the reported number of cases on day t, defined in Section 3.3. The results are reported in Table 3. Furthermore, Figure 5 displays the mean estimates of the cumulative number of cases for the period of April 2, 2020 to April 30, 2020 (in green) and the mean predicted cumulative number of cases for the period of May 1, 2020 to May 10, 2020 (in blue), in comparison to the actually reported cumulative number of cases from April 2, 2020 to May 10, 2020 (in red). The fair agreement of the fitted or predicted values to the reported values suggest the good performance of the proposed SEASAR model.

4.5 Sensitivity Analysis

To further assess the performance of the proposed SEASAR model, we conduct ten sensitivity analyses to evaluate the sensitivity of the results to the specification of the initial values. We take the same setup for Section 4.2 except for changing one value as one of the following settings: (**S1**): The observational error variance (OEV) is taken as: $\sigma_t^2 = max\left(0.5, \frac{\sqrt{\log y_t}}{4}\right)$; (**S2**): The latent period is set as $Z = 4.0$ days (Guan et al., 2020); (**S3**): The latent period is taken as $Z = 5.08$ days (He et al., 2020); (**S4**): The latent period is set as $Z = 6.4$ days (Backer et al.,

Table 3: Differences between the predicted and reported numbers

Date	MAE	RMAE
May 1	235.273	0.212
May 2	125.569	0.125
May 3	39.142	0.044
May 4	140.526	0.185
May 5	113.034	0.142
May 6	45.432	0.050
May 7	48.103	0.053
May 8	51.530	0.057
May 9	105.473	0.126
May 10	212.360	0.289

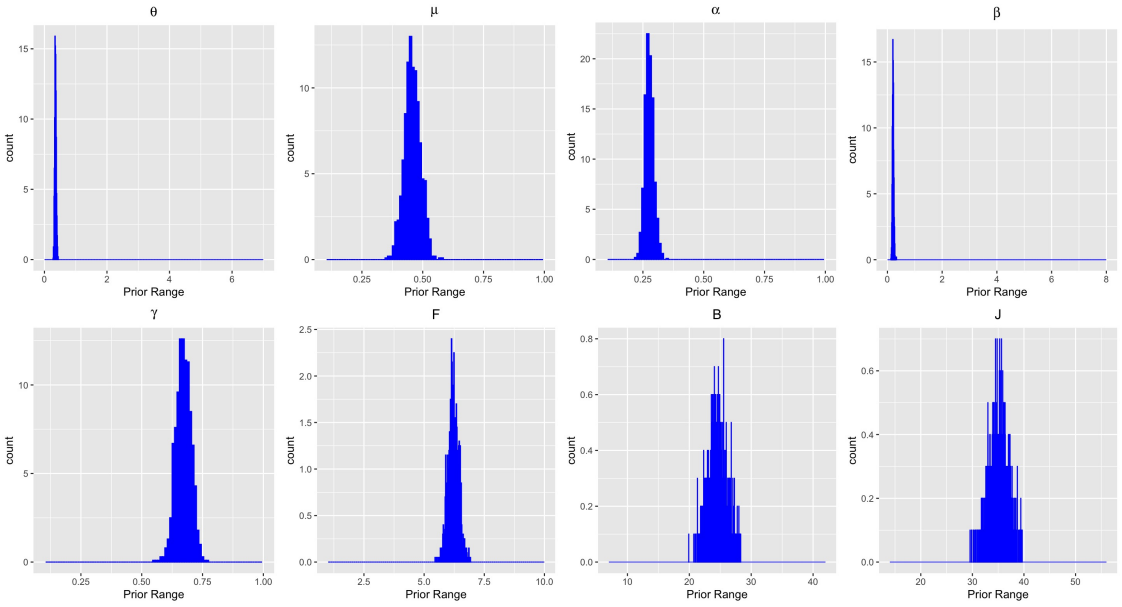

Figure 4: *The distribution of* 1000 *estimates for each parameter. The range of the x-axis for each subfigure is set as the initial range of the corresponding parameter.*

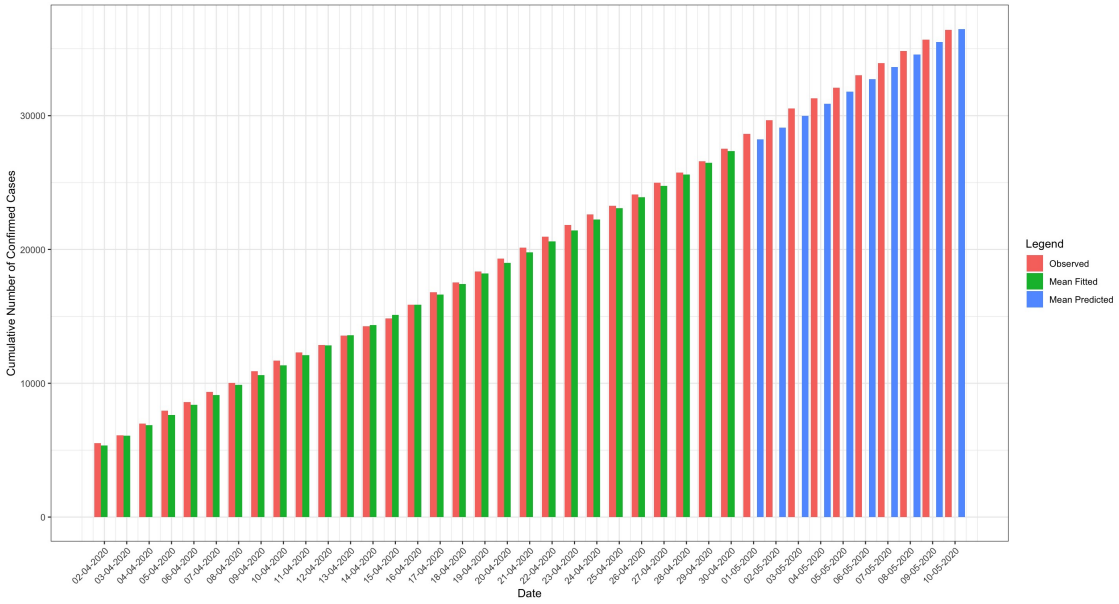

Figure 5: *The model fitting to the cumulative number of cases for each day (in green) in the period of April 2, 2020 to April 30, 2020, and the model prediction to the cumulative number of cases for each day (in blue) in the period of May 1, 2020 to May 10, 2020, as opposed to the reported cumulative number of cases for each day (in red) in the period of April 2, 2020 to May 10, 2020.*

2020); (**S5**): Set $r_3 = 1$; (**S6**): Set $r_3 = 3$; (**S7**): Set $r_2 = 2$; (**S8**): Set $r_2 = 0.5$; (**S9**): Set $r_1 = 0.1$; (**S10**): Set $r_1 = 3$.

The results of sensitivity analyses are reported in Figure 6. While the disparity of the fitted values or predicted values from the reported values varies from setting to setting, overall, those differences seem to be fairly small, suggesting reasonable robustness of the results to the specification of associated values in fairly realistic ranges. Further, we report in Table 4 the estimate and associated 95% CIs of the basic reproduction number R_0 obtained from those ten sensitivity analyses. All the estimates are greater than 1, and consequently, the proposed model suggests the ongoing virus spread under a variety of settings we consider.

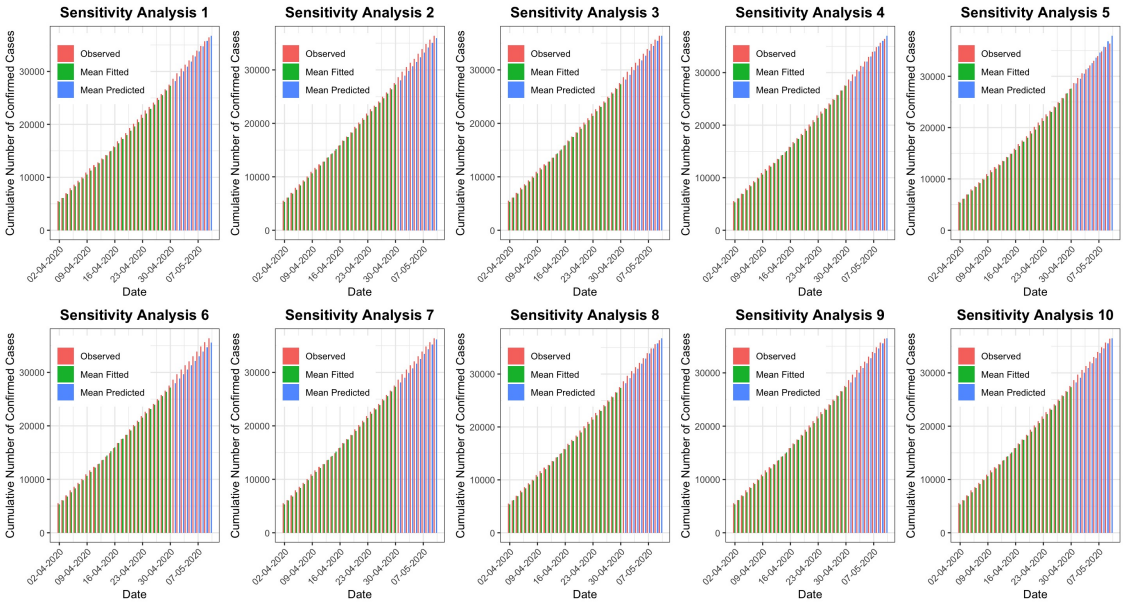

Figure 6: *The model fitting to the cumulative number of cases for each day (in green) in the period of April 2, 2020 to April 30, 2020, and the model prediction to the cumulative number of cases for each day (in blue) in the period of May 1, 2020 to May 10, 2020, as opposed to the reported cumulative number of cases for each day (in red) in the period of April 2, 2020 to May 10, 2020.*

Table 4: R_0 and the corresponding 95% CIs for 10 sensitivity analyses

Sensitivity Analysis	R_0	95% CI
Analysis 1	1.142	$(1.102, 1.176)$
Analysis 2	1.077	$(1.021, 1.125)$
Analysis 3	1.108	$(1.044, 1.165)$
Analysis 4	1.162	$(1.093, 1.223)$
Analysis 5	1.195	$(1.136, 1.249)$
Analysis 6	1.057	$(0.989, 1.115)$
Analysis 7	1.101	$(1.044, 1.160)$
Analysis 8	1.130	$(1.069, 1.183)$
Analysis 9	1.114	$(1.051, 1.169)$
Analysis 10	1.115	$(1.055, 1.170)$

5 Discussion

In this paper, we propose a new epidemic model, called the SEASAR model, to describe the transmission process of COVID-19. Consistent with many available epidemic models, our proposed model require two standard conditions: (1) the population is homogeneous and well-mixed; and (2) the population size remains invariant over time. To facilitate different states of the individuals in the population, we divide the population into six subpopulations (or compartments), respectively, called *susceptible*, *exposed*, *asymptomatic*, *symptomatic*, *active*, and *removed*. Such a classification allows us to accommodate the unique manifestations of COVID-19 which include asymptomatic infections and varying lag times between symptoms onset and diagnosis. The dynamic changes from compartment to compartment is delineated by ordinary differntial equations together with unknown parameters or transition rates.

To utilize the proposed model, we examine the COVID-19 data in Quebec, Canada, from April 2, 2020 to May 10, 2020. Our analysis is conducted under varying assumptions to reflect different possibilities due to the lack of the precise information. The sensitivity analyses reveal that the pandemic situation in Quebec, Canada, is not under control for the study period.

The proposed SEASAR model extends several existing epidemic models such as SIR and SEIR models. It would be interesting to develop more refined models along the same lines of the current development. For example, one may relax constant model parameters to be time-varying to gain a greater flexibility. Time-varying model parameters may be described by a parametric form or a weakly parametric form (e.g., piecewise constants over different time intervals). While the same principle can be applied to develop a procedure for estimating model parameters, technical details would be more notationally involved, and computation would be more costly.

Acknowledgements

This research is partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) as well as the Rapid Response Program COVID-19 of the Canadian Statistical Sciences Institute (CANSSI). Yi is Canada Research Chair in Data Science (Tier 1). Her research was undertaken, in part, thanks to funding from the Canada Research Chairs Program.

Appendix A: Expression of the Basic Reproduction Number

Following the principles of Diekmann et al. (2010), we derive the *basic reproduction number* R_0. The basic idea is to extract the information on the production of new infections from the model, and then calculate the expected number of new infections generated from a case in the population under the constraint $S(t) = N$.

Since a confirmed case must be quarantined, any individuals in the compartment of being active cannot be infectious. In addition, any individuals in the compartments of being susceptible or removed do not have infectious abilities. Therefore, only (2), (3) and (4) in the SEASAR model involves the information on new infections generated by infected cases who are not confirmed yet, thus not being quarantined. With the assumption that all individuals in the populations are susceptible (i.e., setting $S(t) = N$), those three equations become

$$\frac{dE(t)}{dt} = \theta I_s(t) + \mu\theta I_a(t) - \frac{E(t)}{Z}; \tag{12}$$

$$\frac{dI_a(t)}{dt} = (1-\alpha)\frac{E(t)}{Z} - \beta I_a(t) - \gamma I_a(t); \tag{13}$$

$$\frac{dI_s(t)}{dt} = \alpha\frac{E(t)}{Z} - \frac{I_s(t)}{F} + \beta I_a(t); \tag{14}$$

which can be equivalently written in a compact form:

$$\frac{d\phi_{sub}(t)}{dt} = G\phi_{sub}(t), \tag{15}$$

where $\phi_{sub}(t) = (E(t), I_a(t), I_s(t))^{\mathrm{T}}$ and $G = \begin{bmatrix} -\frac{1}{Z} & \mu\theta & \theta \\ \frac{1-\alpha}{Z} & -(\beta+\gamma) & 0 \\ \frac{\alpha}{Z} & \beta & -\frac{1}{F} \end{bmatrix}$.

Noting that only equation (12) includes the terms, $\theta I_s(t) + \mu\theta I_a(t)$, of the production of new infections, we re-write the matrix G as $G = G_1 + G_2$, where $G_1 = \begin{bmatrix} 0 & \mu\theta & \theta \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ reflects the coefficients in the expression $\theta I_s(t) + \mu\theta I_a(t)$, and $G_2 = \begin{bmatrix} -\frac{1}{Z} & 0 & 0 \\ \frac{1-\alpha}{Z} & -(\beta+\gamma) & 0 \\ \frac{\alpha}{Z} & \beta & -\frac{1}{F} \end{bmatrix}$. Although the definition of R_0 is conceptually intuitive, its determination is mathematically complicated (Delamater et al., 2019). Diekmann et al. (1990) took R_0 as the dominant eigenvalue of the next-generation matrix, which is derived as $-G_1 G_2^{-1}$ for the equations in (15) by adapting the arguments of Diekmann et al. (2010). Specifically, noting that $G_2^{-1} = \begin{bmatrix} -Z & 0 & 0 \\ \frac{-1+\alpha}{\beta+\gamma} & -\frac{1}{\beta+\gamma} & 0 \\ -\frac{(\alpha\gamma+\beta)F}{\beta+\gamma} & -\frac{\beta F}{\beta+\gamma} & -F \end{bmatrix}$, we obtain that $-G_1(G_2)^{-1} = \begin{bmatrix} -\frac{\mu\theta(\alpha-1)}{\beta+\gamma} + \frac{\theta(\alpha\gamma+\beta)F}{\beta+\gamma} & \frac{\mu\theta}{\beta+\gamma} + \frac{\theta\beta F}{\beta+\gamma} & \theta F \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$, whose dominant eigenvalue is $\frac{\theta(F\alpha\gamma+\beta F-\mu\alpha+\mu)}{\beta+\gamma}^{T}$. Therefore, $R_0 = \frac{\theta(F\alpha\gamma+\beta F-\mu\alpha+\mu)}{\beta+\gamma}$.

Appendix B: Derivations of the Expressions in Section 3.3

In this appendix, we present the derivations of the expressions used in Section 3.3, and describe the implementation of the IF-EAKF algorithm using the terminology of the "joint state-observation vector" (see Anderson, 2001, p.2886) which basically includes time-dependent variables.

For ease of exposition, we let $\log \phi(t)$ denote the vector $(\log S(t), \log E(t), \log I_a(t), \log I_s(t), \log A(t), \log R(t))^{\mathrm{T}}$. Let the joint state-observation vector at time t be $O(t) = (\eta^{\mathrm{T}}, (\log \phi(t))^{\mathrm{T}}, \log \hat{Y}(t))^{\mathrm{T}}$. Because $\phi(t)$ is regarded as deterministic at τ_0, the beginning of the study, so when $t = \tau_0$, $O(t)$ degenerates as $O(\tau_0) = (\eta^{\mathrm{T}}, \log \hat{Y}(\tau_0))^{\mathrm{T}}$. Here applying the logarithm to $\phi(t)$ and $\hat{Y}(t)$ aligns with the transformation involved with model (7).

The EAKF algorithm basically aims to work out the posterior distribution of the joint state-observation vector $O(t)$ for $t \geq \tau_0$. To reduce computation costs, we adopt the discussion of (Anderson, 2001, p.2888) and consider a simple implementation procedure by examining $O(t)$ via its paired subvectors separately, where we pair $\log \hat{Y}(t)$ with each of other elements in $O(t)$ and calculate the posterior distributions for $(\theta, \log \hat{Y}(t))^{\mathrm{T}}$, $(\mu, \log \hat{Y}(t))^{\mathrm{T}}$, $(\alpha, \log \hat{Y}(t))^{\mathrm{T}}$, $(\beta, \log \hat{Y}(t))^{\mathrm{T}}$, $(\gamma, \log \hat{Y}(t))^{\mathrm{T}}$, $(F, \log \hat{Y}(t))^{\mathrm{T}}$, $(B, \log \hat{Y}(t))^{\mathrm{T}}$, $(J, \log \hat{Y}(t))^{\mathrm{T}}$, $(\log S(t), \log \hat{Y}(t))^{\mathrm{T}}$, $(\log E(t), \log \hat{Y}(t))^{\mathrm{T}}$,$(\log I_a(t), \log \hat{Y}(t))^{\mathrm{T}}$, $(\log I_s(t), \log \hat{Y}(t))^{\mathrm{T}}$, $(\log A(t), \log \hat{Y}(t))^{\mathrm{T}}$, and $(\log R(t), \log \hat{Y}(t))^{\mathrm{T}}$ separately, which employs the same procedure in principle.

To show the ideas, we just describe the way of calculating the posterior distribution of $(\theta, \log \hat{Y}(t))^{\mathrm{T}}$ for $t \geq \tau_0$. Write $Z_t = (\theta, \log \hat{Y}(t))^{\mathrm{T}}$, and let $H = (0, 1)$ so that $HZ_t = \log \hat{Y}(t)$. Rather than attempt to find the analytic expression of the posterior distribution, the EAKF algorithm generates the posterior values of Z_t by using the prior values of Z_t and the observation y_t. To this end, let $\{\theta^i_{\mathrm{pri},t} : i = 1, \cdots, n\}$ and $\{\hat{Y}^i_{\mathrm{pri}}(t) : i = 1, \cdots, n\}$ denote sequences of prior values of θ and $\hat{Y}(t)$ at time t, respectively.

First, we calculate their sample means, sample covariance, and sample variances, which are, respectively, given by

$$\overline{\theta}_{\mathrm{pri},t} = \frac{\sum_{i=1}^{n} \theta^i_{\mathrm{pri},t}}{n}; \tag{16}$$

$$\overline{o}_{\mathrm{pri},t} = \frac{\sum_{i=1}^{n} \log \hat{Y}^i_{\mathrm{pri}}(t)}{n}; \tag{17}$$

$$\sigma^{cov}_{\theta,\log \hat{Y}(t),\mathrm{pri}} = \frac{1}{n-1} \sum_{i=1}^{n} \left\{ \log \hat{Y}^i_{\mathrm{pri}}(t) - \overline{o}_{\mathrm{pri},t} \right\} \left\{ \theta^i_{\mathrm{pri},t} - \overline{\theta}_{\mathrm{pri},t} \right\}; \tag{18}$$

$$\sigma^2_{\theta,\mathrm{pri},t} = \frac{\sum_{i=1}^{n} (\theta^i_{\mathrm{pri},t} - \overline{\theta}_{\mathrm{pri},t})^2}{n-1}; \tag{19}$$

$$\sigma^2_{\mathrm{pri},t} = \frac{\sum_{i=1}^{n} \left\{ \log \hat{Y}^i_{\mathrm{pri}}(t) - \overline{o}_{\mathrm{pri},t} \right\}^2}{n-1}. \tag{20}$$

Then we set the mean of prior values of Z_t as

$$\overline{z_t^p} = (\overline{\theta}_{\text{pri},t}, \overline{o}_{\text{pri},t})^{\text{T}} \tag{21}$$

and the covariance matrix of prior values of Z_t as

$$\Sigma_t^p = \begin{bmatrix} \sigma^2_{\theta,\text{pri},t} & \sigma^{cov}_{\theta,\log\hat{Y}(t),\text{pri}} \\ \sigma^{cov}_{\theta,\log\hat{Y}(t),\text{pri}} & \sigma^2_{\text{pri},t} \end{bmatrix}.$$

The EAKF algorithm assumes that the prior distribution of Z_t can be represented or reasonably approximated by a Gaussian distribution with the mean $\overline{z_t^p}$ and the covariance matrix Σ_t^p. Therefore, the density function of the prior distribution of Z_t is expressed as

$$f_{Z_t}(z_t) \propto \exp\Big\{ -\frac{1}{2}(z_t - \overline{z_t^p})^{\text{T}}(\Sigma_t^p)^{-1}(z_t - \overline{z_t^p})\Big\}. \tag{22}$$

In Appendix C, we show that the posterior distribution of Z_t is Gaussian with the mean

$$\overline{z_t^u} = \begin{bmatrix} \overline{\theta}_{\text{pri},t} + \frac{\log y_t \times \sigma^{cov}_{\theta,\log\hat{Y}(t),\text{pri}} - \overline{o}_{\text{pri},t} \times \sigma^{cov}_{\theta,\log\hat{Y}(t),\text{pri}}}{\sigma_t^2 + \sigma^2_{\text{pri},t}} \\ \frac{\overline{o}_{\text{pri},t} \times \sigma_t^2 + \log y_t \times \sigma^2_{\text{pri},t}}{\sigma_t^2 + \sigma^2_{\text{pri},t}} \end{bmatrix} \tag{23}$$

and covariance matrix

$$\Sigma_t^u = \frac{\sigma_t^2}{\sigma_t^2 + \sigma^2_{\text{pri},t}} \times \begin{bmatrix} \sigma^2_{\theta,\text{pri},t} + \frac{\sigma^2_{\theta,\text{pri},t} \times \sigma^2_{\text{pri},t} - \sigma^{cov}_{\theta,\log\hat{Y}(t),\text{pri}}}{\sigma_t^2} & \sigma^{cov}_{\theta,\log\hat{Y}(t),\text{pri}} \\ \sigma^{cov}_{\theta,\log\hat{Y}(t),\text{pri}} & \sigma^2_{\text{pri},t} \end{bmatrix}. \tag{24}$$

To describe the generation of the posterior values of θ and $\hat{Y}(t)$ at time t, for $i = 1, \cdots, n$, we let $z^u_{i,t} = (\theta^i_{\text{post},t}, \log\hat{Y}^i_{\text{post}}(t))^{\text{T}}$ denote posterior values of θ and $\log\hat{Y}(t)$ to be generated, and let $z^p_{i,t} = (\theta^i_{\text{pri},t}, \log\hat{Y}^i_{\text{pri}}(t))^{\text{T}}$ denote prior values of θ and $\log\hat{Y}(t)$. Then the EAKF algorithm (Anderson, 2001, p.2887) generates n posterior values of Z_t as follows:

$$z^u_{i,t} = D^{\text{T}}(z^p_{i,t} - \overline{z_t^p}) + \overline{z_t^u} \quad \text{for } i = 1, \cdots, n, \tag{25}$$

where $\overline{z_t^p}$ and $\overline{z_t^u}$ are given by (21) and (23), respectively, and D can be any matrix such that the covariance matrix of $\{z^u_{i,t} : i = 1, \cdots, n\}$ computed from (25) equals that computed by (24).

The existence of such a matrix D is justified in Appendix A of Anderson (2001). For instance, D can be taken as

$$\begin{bmatrix} \frac{\sigma^{cov}_{\theta,\log\hat{Y}(t),\text{pri}}}{\sigma^2_{\text{pri},t}} \times \left(\sqrt{\frac{\sigma_t^2}{\sigma_t^2+\sigma^2_{\text{pri},t}}} - 1\right) & \sqrt{\frac{\sigma_t^2}{\sigma_t^2+\sigma^2_{\text{pri},t}}} \end{bmatrix}, \tag{26}$$

as shown in Appendix D.

Substituting matrix D in (25) with (26), we obtain the following explicit expressions for posterior values of θ and $\log \hat{Y}(t)$ at time t for $i = 1, \cdots, n$:

$$\begin{aligned} \log \hat{Y}^i_{\text{post}}(t) = &\frac{\sigma_t^2}{\sigma_t^2+\sigma^2_{\text{pri},t}}\overline{o}_{\text{pri},t} + \frac{\sigma^2_{\text{pri},t}}{\sigma_t^2+\sigma^2_{\text{pri},t}}\log y_t \\ &+ \sqrt{\frac{\sigma_t^2}{\sigma_t^2+\sigma^2_{\text{pri},t}}}\left\{\log \hat{Y}^i_{\text{pri}}(t) - \overline{o}_{\text{pri},t}\right\}; \end{aligned} \tag{27}$$

$$\theta^i_{\text{post},t} = \theta^i_{\text{pri},t} + \frac{\sigma^{cov}_{\theta,\log\hat{Y}(t),\text{pri}}}{\sigma^2_{\text{pri},t}}\left\{\log \hat{Y}^i_{\text{post}}(t) - \log \hat{Y}^i_{\text{pri}}(t)\right\}. \tag{28}$$

Setting $t = \tau_0$, (27) and (28) yield (9) and (10), respectively.

Appendix C: Derivations of the Posterior Distribution of Z_t

For ease of notation, we write $U = \log Y(t)$ and $Z = Z_t$ for any given $t \geq \tau_0$. Let $H_1 = (1, 0)$, by the fact that $Z = (\theta, \log \hat{Y}(t))^{\text{T}}$ and $H = (0, 1)$, we have $\theta = H_1 Z$ and $\log \hat{Y}(t) = HZ$.

Now we show that U is conditionally independent of $H_1 Z$, given HZ. Indeed, for any $u, z_1, z_2 \in \mathbb{R}$, the conditional cumulative distribution function of U and $H_1 Z$, given $HZ = z_2$,

is

$$\begin{aligned} F_{U,H_1Z|HZ}(u, z_1|z_2) &= P(U \leq u, H_1Z \leq z_1|HZ = z_2) \\ &= P(HZ + \epsilon_t \leq u, H_1Z \leq z_1|HZ = z_2) \\ &= P(\epsilon_t \leq u - z_2, H_1Z \leq z_1|HZ = z_2) \\ &= P(\epsilon_t \leq u - z_2|HZ = z_2)P(H_1Z \leq z_1|HZ = z_2) \\ &= P(U \leq u|HZ = z_2)P(H_1Z \leq z_1|HZ = z_2) \\ &= F_{U|HZ}(u|z_2)F_{H_1Z|HZ}(z_1|z_2) \end{aligned}$$

where the second equality is due to (7), the fourth equality comes from the assumption that ϵ_t is independent of H_1Z, and the second last step is because that $P(\epsilon_t \leq u - z_2|HZ = z_2) = P(\epsilon_t \leq u - HZ|HZ = z_2) = P(HZ + \epsilon_t \leq u|HZ = z_2) = P(U \leq u|HZ = z_2)$ by (7). Therefore, we conclude that U is independent of H_1Z, given HZ.

Next, we find the conditional distribution of U given Z, $f_{U|Z}(u|z)$, where $z = (z_1, z_2)^{\mathrm{T}}$ with $z_1, z_2 \in \mathbb{R}$, and $u \in \mathbb{R}$. To do this, we use the definition,

$$f_{U|Z}(u|z) = \frac{f_{U,Z}(u, z)}{f_Z(z)} \tag{29}$$

and first determine the joint distribution of U and Z, $f_{U,Z}(u, z)$. By the definition of Z, we have

$$f_{U,Z}(u, z) = f_{U,H_1Z,HZ}(u, z_1, z_2)$$

which equals $f_{U,H_1Z|HZ}(u, z_1|z_2)f_{HZ}(z_2)$. Since U is conditionally independent of H_1Z, given HZ, therefore, we have that

$$\begin{aligned} f_{U,Z}(u, z) &= f_{U|HZ}(u|z_2)f_{H_1Z|HZ}(z_1|z_2)f_{HZ}(z_2) \\ &= f_{U|HZ}(u|z_2)f_{H_1Z,HZ}(z_1, z_2) \\ &= f_{U|HZ}(u|z_2)f_Z(z), \end{aligned} \tag{30}$$

where the last equality is due to the definition of $Z = (H_1 Z, HZ)^{\mathrm{T}}$. Combining (30) and (29) gives that

$$f_{U|Z}(u|z) = f_{U|HZ}(u|z_2), \tag{31}$$

which is determined by (7). Therefore, using the original notation, by (31), we obtain the conditional density function of $\log Y(t)$ given Z_t,

$$f_{\log Y(t)|Z_t}(\log y_t|z_t) \propto \exp\Big\{ -\frac{1}{2}(\log y_t - Hz_t)^{\mathrm{T}} \frac{1}{\sigma_t^2}(\log y_t - Hz_t)\Big\}. \tag{32}$$

Finally, we determine the posterior distribution of Z, i.e., the conditional distribution of Z given U, $f_{Z|U}(z|u)$. Combining (22) and (32) gives the density function of the posterior distribution of Z_t:

$$\begin{aligned}
f_{Z_t|\log Y(t)}(z_t|\log y_t) &\propto f_{Z_t}(z_t) \times f_{\log Y(t)|Z_t}(\log y_t|z_t) \\
&\propto \exp\Big[-\frac{1}{2}\Big\{ z_t^{\mathrm{T}}(\Sigma_t^p)^{-1} z_t - \overline{z_t^p}^{\mathrm{T}}(\Sigma_t^p)^{-1} z_t - z_t^{\mathrm{T}}(\Sigma_t^p)^{-1}\overline{z_t^p} \\
&\quad + z_t^{\mathrm{T}}\frac{H^{\mathrm{T}}H}{\sigma_t^2} z_t - (Hz_t)^{\mathrm{T}}\frac{1}{\sigma_t^2}\log y_t - (\log y_t)^{\mathrm{T}}\frac{1}{\sigma_t^2} Hz_t\Big\}\Big] \\
&\propto \exp\Bigg\{ -\frac{1}{2}\Big[z_t - \Big\{(\Sigma_t^p)^{-1} + \frac{H^{\mathrm{T}}H}{\sigma_t^2}\Big\}^{-1}\Big\{(\Sigma_t^p)^{-1}\overline{z_t^p} + \frac{H^{\mathrm{T}}\log y_t}{\sigma_t^2}\Big\}\Big]^{\mathrm{T}} \times \\
&\quad \Big[(\Sigma_t^p)^{-1} + \frac{H^{\mathrm{T}}H}{\sigma_t^2}\Big] \times \Big[z_t - \Big\{(\Sigma_t^p)^{-1} + \frac{H^{\mathrm{T}}H}{\sigma_t^2}\Big\}^{-1}\Big\{(\Sigma_t^p)^{-1}\overline{z_t^p} + \frac{H^{\mathrm{T}}\log y_t}{\sigma_t^2}\Big\}\Big]\Bigg\},
\end{aligned}$$

showing that the posterior distribution of Z_t is Gaussian with the mean

$$\overline{z_t^u} = \Big\{(\Sigma_t^p)^{-1} + \frac{H^{\mathrm{T}}H}{\sigma_t^2}\Big\}^{-1}\Big\{(\Sigma_t^p)^{-1}\overline{z_t^p} + \frac{H^{\mathrm{T}}\log y_t}{\sigma_t^2}\Big\} \tag{33}$$

and covariance matrix

$$\Sigma_t^u = \Big\{(\Sigma_t^p)^{-1} + \frac{H^{\mathrm{T}}H}{\sigma_t^2}\Big\}^{-1}. \tag{34}$$

Furthermore, we express (33) and (34) in terms of the observations and the prior values in (16), (17), (18), (19), and (20):

$$\overline{z_t^u} = \begin{bmatrix} \overline{\theta}_{\mathrm{pri},t} + \frac{\log y_t \times \sigma^{cov}_{\theta,\log \hat{Y}(t),\mathrm{pri}} - \overline{o}_{\mathrm{pri},t} \times \sigma^{cov}_{\theta,\log \hat{Y}(t),\mathrm{pri}}}{\sigma_t^2 + \sigma^2_{\mathrm{pri},t}} \\ \frac{\overline{o}_{\mathrm{pri},t} \times \sigma_t^2 + \log y_t \times \sigma^2_{\mathrm{pri},t}}{\sigma_t^2 + \sigma^2_{\mathrm{pri},t}} \end{bmatrix};$$

$$\Sigma_t^u = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_{\text{pri},t}^2} \times \begin{bmatrix} \sigma_{\theta,\text{pri},t}^2 + \frac{\sigma_{\theta,\text{pri},t}^2 \times \sigma_{\text{pri},t}^2 - \sigma_{\theta,\log \hat{Y}(t),\text{pri}}^{cov}}{\sigma_t^2} & \sigma_{\theta,\log \hat{Y}(t),\text{pri}}^{cov} \\ \sigma_{\theta,\log \hat{Y}(t),\text{pri}}^{cov} & \sigma_{\text{pri},t}^2. \end{bmatrix}.$$

Appendix D: Verification of the Condition Required from EAKF

Substituting the matrix D in (25) with (26) gives us

$$\begin{bmatrix} \theta_{\text{post},t}^i \\ \log \hat{Y}_{\text{post}}^i(t) \end{bmatrix} = \begin{bmatrix} 1 & \frac{\sigma_{\theta,\log \hat{Y}(t),\text{pri}}^{cov}}{\sigma_{\text{pri},t}^2} \times \left(\sqrt{\frac{\sigma_t^2}{\sigma_t^2 + \sigma_{\text{pri},t}^2}} - 1 \right) \\ 0 & \sqrt{\frac{\sigma_t^2}{\sigma_t^2 + \sigma_{\text{pri},t}^2}} \end{bmatrix} \times \begin{bmatrix} \theta_{\text{pri},t}^i - \overline{\theta}_{\text{pri},t} \\ \log \hat{Y}_{\text{pri}}^i(t) - \overline{o}_{\text{pri},t} \end{bmatrix}$$
$$+ \begin{bmatrix} \overline{\theta}_{\text{pri},t} + \frac{\log y_t \times \sigma_{\theta,\log \hat{Y}(t),\text{pri}}^{cov} - \overline{o}_{\text{pri},t} \times \sigma_{\theta,\log \hat{Y}(t),\text{pri}}^{cov}}{\sigma_t^2 + \sigma_{\text{pri},t}^2} \\ \frac{\overline{o}_{\text{pri},t} \times \sigma_t^2 + \log y_t \times \sigma_{\text{pri},t}^2}{\sigma_t^2 + \sigma_{\text{pri},t}^2} \end{bmatrix},$$

yielding that

$$\begin{aligned} \theta_{\text{post},t}^i = \theta_{\text{pri},t}^i - \overline{\theta}_{\text{pri},t} + \{\log \hat{Y}_{\text{pri}}^i(t) - \overline{o}_{\text{pri},t}\} \times \frac{\sigma_{\theta,\log \hat{Y}(t),\text{pri}}^{cov}}{\sigma_{\text{pri},t}^2} \times \left(\sqrt{\frac{\sigma_t^2}{\sigma_t^2 + \sigma_{\text{pri},t}^2}} - 1 \right) \\ + \overline{\theta}_{\text{pri},t} + \frac{\log y_t \times \sigma_{\theta,\log \hat{Y}(t),\text{pri}}^{cov} - \overline{o}_{\text{pri},t} \times \sigma_{\theta,\log \hat{Y}(t),\text{pri}}^{cov}}{\sigma_t^2 + \sigma_{\text{pri},t}^2} \end{aligned} \tag{35}$$

and

$$\begin{aligned} \log \hat{Y}_{\text{post}}^i(t) = & \left\{ \log \hat{Y}_{\text{pri}}^i(t) - \overline{o}_{\text{pri},t} \right\} \times \sqrt{\frac{\sigma_t^2}{\sigma_t^2 + \sigma_{\text{pri},t}^2}} \\ & + \frac{\overline{o}_{\text{pri},t} \times \sigma_t^2 + \log y_t \times \sigma_{\text{pri},t}^2}{\sigma_t^2 + \sigma_{\text{pri},t}^2}. \end{aligned} \tag{36}$$

Thus, (35) and (36) give us a sequence of posterior values of θ, $\{\theta_{\text{post},t}^i : i = 1, \cdots, n\}$, and a sequence of posterior values of $\log \hat{Y}(t)$, $\{\log \hat{Y}_{\text{post}}^i(t) : i = 1, \cdots, n\}$.

Next, we calculate the variance of the posterior values $\{\theta_{\text{post},t}^i : i = 1, \cdots, n\}$ using (35),

$$\begin{aligned} \sigma_{\theta,\text{post},t}^2 &= \frac{\sum_{i=1}^n \left(\theta_{\text{post},t}^i - \frac{\theta_{\text{post},t}^i}{n} \right)^2}{n-1} \\ &= \frac{\sigma_t^2}{\sigma_t^2 + \sigma_{\text{pri},t}^2} \times \left(\sigma_{\theta,\text{pri},t}^2 + \frac{\sigma_{\theta,\text{pri},t}^2 \times \sigma_{\text{pri},t}^2 - \sigma_{\theta,\log \hat{Y}(t),\text{pri}}^{cov}}{\sigma_t^2} \right), \end{aligned}$$

which is identical to the $(1,1)$ element of (24).

Using (36), we calculate the variance of the posterior values $\{\log \hat{Y}^i_{\text{post}}(t) : i = 1, \cdots, n\}$ of $\log \hat{Y}(t)$, given by

$$\begin{aligned}
\sigma^2_{\text{post},t} &= \frac{\sum_{i=1}^{n} \left\{ \log \hat{Y}^i_{\text{post}}(t) - \frac{\log \hat{Y}^i_{\text{post}}(t)}{n} \right\}^2}{n-1} \\
&= \frac{1}{n-1} \sum_{i=1}^{n} \{\log \hat{Y}^i_{\text{pri}}(t) - \overline{o}_{\text{pri},t}\}^2 \times \frac{\sigma_t^2}{\sigma_t^2 + \sigma^2_{\text{pri},t}} \\
&= \frac{\sigma_t^2 \times \sigma^2_{\text{pri},t}}{\sigma_t^2 + \sigma^2_{\text{pri},t}},
\end{aligned}$$

which is identical to the $(2,2)$ element of (24).

Furthermore, using (35) and (36), we calculate the covariance of $\{\theta^i_{\text{post},t} : i = 1, \cdots, n\}$ and $\{\log \hat{Y}^i_{\text{post}}(t) : i = 1, \cdots, n\}$,

$$\begin{aligned}
\sigma^{cov}_{\theta,\log \hat{Y}(t),\text{post}} &= \frac{\sum_{i=1}^{n} \left\{ \theta^i_{\text{post},t} - \frac{\sum_{i=1}^{n} \theta^i_{\text{post},t}}{n} \right\} \left\{ \log \hat{Y}^i_{\text{post}}(t) - \frac{\sum_{i=1}^{n} \log \hat{Y}^i_{\text{post}}(t)}{n} \right\}}{n-1} \\
&= \frac{1}{n-1} \sum_{i=1}^{n} (\theta^i_{\text{pri},t} - \overline{\theta}_{\text{pri},t}) \times \{\log \hat{Y}^i_{\text{pri}}(t) - \overline{o}_{\text{pri},t}\} \times \sqrt{\frac{\sigma_t^2}{\sigma_t^2 + \sigma^2_{\text{pri},t}}} \\
&\quad + \{\log \hat{Y}^i_{\text{pri}}(t) - \overline{o}_{\text{pri},t}\}^2 \times \frac{\sigma^{cov}_{\theta,\log \hat{Y}(t),\text{pri}}}{\sigma^2_{\text{pri},t}} \times \left(\frac{\sigma_t^2}{\sigma_t^2 + \sigma^2_{\text{pri},t}} - \sqrt{\frac{\sigma_t^2}{\sigma_t^2 + \sigma^2_{\text{pri},t}}} \right) \\
&= \frac{\sigma_t^2}{\sigma_t^2 + \sigma^2_{\text{pri},t}} \times \sigma^{cov}_{\theta,\log \hat{Y}(t),\text{pri}},
\end{aligned}$$

which is identical to the $(1,2)$ and $(2,1)$ elements of the matrix (24). Therefore, we verify that the matrix D satisfies the condition required by the EAKF algorithm.

References

Abbey, H. (1952). An examination of the reed-frost theory of epidemics. *Human Biology*, 24(3):201–233.

Anderson, J. L. (2001). An ensemble adjustment kalman filter for data assimilation. *Monthly Weather Review*, 129(12):2884–2903.

Backer, J. A., Klinkenberg, D., and Wallinga, J. (2020). Incubation period of 2019 novel coronavirus (2019-nCov) infections among travellers from Wuhan, China, 20-28 January 2020. *Eurosurveillance*, 25(5):1–6.

Bai, Y., Yao, L., Wei, T., Tian, F., Jin, D.-Y., Chen, L., and Wang, M. (2020). Presumed asymptomatic carrier transmission of COVID-19. *The Journal of the American Medical Association*, 323(14):1406–1407.

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

Delamater, P. L., Street, E. J., Leslie, T. F., Yang, Y. T., and Jacobsen, K. H. (2019). Complexity of the basic reproduction number (R_0). *Emerging Infectious Diseases*, 25(1):1–4.

Diekmann, O., Heesterbeek, J., and Roberts, M. G. (2010). The construction of next-generation matrices for compartmental epidemic models. *Journal of the Royal Society Interface*, 7(47):873–885.

Diekmann, O., Heesterbeek, J. A. P., and Metz, J. A. (1990). On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology*, 28(4):365–382.

Duan, W., Fan, Z., Zhang, P., Guo, G., and Qiu, X. (2015). Mathematical and computational approaches to epidemic modeling: A comprehensive review. *Frontiers of Computer Science*, 9(5):806–826.

Guan, W., Ni, Z., Hu, Y., Liang, W., Ou, C., He, J., Liu, L., Shan, H., Lei, C., Hui, D. S., et al. (2020). Clinical characteristics of coronavirus disease 2019 in China. *New England Journal of Medicine*, 382(18):1708–1720.

Hao, X., Cheng, S., Wu, D., Wu, T., Lin, X., and Wang, C. (2020). Reconstruction of the full transmission dynamics of COVID-19 in Wuhan. *Nature*, 584(7821):420–424.

He, D., Ionides, E. L., and King, A. A. (2010). Plug-and-Play inference for disease dynamics: Measles in large and small populations as a case study. *Journal of the Royal Society Interface*, 7(43):271–283.

He, W., Yi, G. Y., and 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(11):2543–2550.

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772):700–721.

King, A. A., Ionides, E. L., Pascual, M., and Bouma, M. J. (2008). Inapparent infections and cholera dynamics. *Nature*, 454(7206):877–880.

Kramer, M., Pigott, D., Xu, B., Hill, S., Gutierrez, B., and Pybus, O. (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-descri\
ptions-from-publicly-available-data/337.

Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., and Shaman, J. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). *Science*, 368(6490):489–493.

Ng, J. and Orav, E. J. (1990). A generalized chain binomial model with application to HIV infection. *Mathematical Biosciences*, 101(1):99–119.

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

Osthus, D., Hickmann, K. S., Caragea, P. C., Higdon, D., and Del Valle, S. Y. (2017). Forecasting seasonal influenza with a state-space SIR model. *The annals of applied statistics*, 11(1):202–224.

Pei, S., Kandula, S., Yang, W., and Shaman, J. (2018). Forecasting the spatial transmission of influenza in the United States. *Proceedings of the National Academy of Sciences*, 115(11):2752–2757.

Peng, L., Yang, W., Zhang, D., Zhuge, C., and Hong, L. (2020). Epidemic analysis of COVID-19 in China by dynamical modeling. *medRxiv*.

Reis, J. and Shaman, J. (2016). Retrospective parameter estimation and forecast of respiratory syncytial virus in the United States. *PLOS Computational Biology*, 12(10):1–15.

Shah, N. H. and Gupta, J. (2013). SEIR model and simulation for vector borne diseases. *Applied Mathematics*, 4(8A):13–17.

Shaman, J., Yang, W., and Kandula, S. (2014). Inference and forecast of the current West African Ebola outbreak in Guinea, Sierra Leone and Liberia. *PLOS Currents*, 6.

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

The Canadian Press (2020). Coronavirus: Here is a timeline of COVID-19 cases in Canada. https://globalnews.ca/news/6627505/coronavirus-covid-canada-timeline/.

Vogel, L. (2020). COVID-19: A timeline of Canada's first-wave response. https://cmajnews.com/2020/06/12/coronavirus-1095847/.

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.