



## OPEN ACCESS

## EDITED BY

Pierpaolo Ferrante,  
National Institute for Insurance Against  
Accidents at Work (INAIL), Italy

## REVIEWED BY

Arvind Ramanathan,  
Argonne National Laboratory (DOE),  
United States  
Youcef Mammeri,  
Laboratoire Amiénois de Mathématique  
Fondamentale et Appliquée (LAMFA), France

## \*CORRESPONDENCE

Patricio Cumsille  
✉ pcumsille@ubiobio.cl

## SPECIALTY SECTION

This article was submitted to  
Infectious Diseases: Epidemiology and  
Prevention,  
a section of the journal  
Frontiers in Public Health

RECEIVED 29 November 2022

ACCEPTED 02 March 2023

PUBLISHED 31 March 2023

## CITATION

Cumsille P, Rojas-Díaz O and Conca C (2023) A  
general modeling framework for quantitative  
tracking, accurate prediction of ICU, and  
assessing vaccination for COVID-19 in Chile.  
*Front. Public Health* 11:1111641.  
doi: 10.3389/fpubh.2023.1111641

## COPYRIGHT

© 2023 Cumsille, Rojas-Díaz and Conca. This is  
an open-access article distributed under the  
terms of the [Creative Commons Attribution  
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or  
reproduction in other forums is permitted,  
provided the original author(s) and the  
copyright owner(s) are credited and that the  
original publication in this journal is cited, in  
accordance with accepted academic practice.  
No use, distribution or reproduction is  
permitted which does not comply with these  
terms.

# A general modeling framework for quantitative tracking, accurate prediction of ICU, and assessing vaccination for COVID-19 in Chile

Patricio Cumsille<sup>1,2\*</sup>, Oscar Rojas-Díaz<sup>3</sup> and Carlos Conca<sup>2,4</sup>

<sup>1</sup>Department of Basic Sciences, Faculty of Sciences, University of Bío-Bío, Chillán, Chile, <sup>2</sup>Centre for Biotechnology and Bioengineering, University of Chile, Santiago, Chile, <sup>3</sup>Department of Mathematics and Computers Science, Faculty of Science, University of Santiago of Chile, Santiago, Chile, <sup>4</sup>Department of Mathematical Engineering and Center for Mathematical Modeling, University of Chile (UMI CNRS 2807), Santiago, Chile

**Background:** One of the main lessons of the COVID-19 pandemic is that we must prepare to face another pandemic like it. Consequently, this article aims to develop a general framework consisting of epidemiological modeling and a practical identifiability approach to assess combined vaccination and non-pharmaceutical intervention (NPI) strategies for the dynamics of any transmissible disease.

**Materials and methods:** Epidemiological modeling of the present work relies on delay differential equations describing time variation and transitions between suitable compartments. The practical identifiability approach relies on parameter optimization, a parametric bootstrap technique, and data processing. We implemented a careful parameter optimization algorithm by searching for suitable initialization according to each processed dataset. In addition, we implemented a parametric bootstrap technique to accurately predict the ICU curve trend in the medium term and assess vaccination.

**Results:** We show the framework's calibration capabilities for several processed COVID-19 datasets of different regions of Chile. We found a unique range of parameters that works well for every dataset and provides overall numerical stability and convergence for parameter optimization. Consequently, the framework produces outstanding results concerning quantitative tracking of COVID-19 dynamics. In addition, it allows us to accurately predict the ICU curve trend in the medium term and assess vaccination. Finally, it is reproducible since we provide open-source codes that consider parameter initialization standardized for every dataset.

**Conclusion:** This work attempts to implement a holistic and general modeling framework for quantitative tracking of the dynamics of any transmissible disease, focusing on accurately predicting the ICU curve trend in the medium term and assessing vaccination. The scientific community could adapt it to evaluate the impact of combined vaccination and NPIs strategies for COVID-19 or any transmissible disease in any country and help visualize the potential effects of implemented plans by policymakers. In future work, we want to improve the computational cost of the parametric bootstrap technique or use another more efficient technique. The aim would be to reconstruct epidemiological curves to predict the combined NPIs and vaccination policies' impact on the ICU curve trend in real-time, providing scientific evidence to help anticipate policymakers' decisions.

## KEYWORDS

COVID-19, predictive modeling, epidemiological modeling, time delays, vaccination, practical identifiability, parameter optimization, parametric bootstrap

## 1. Introduction

The COVID-19 pandemic has induced a significant research effort for tracking, prediction, and control. In Chile, which is no stranger to the above, health authorities initiated vaccination in the summer of 2021, gradually reducing overall ICU patients and death by COVID-19 while suspending non-pharmaceutical interventions (NPIs) such as lockdowns (partial or total). One of the main lessons is that we must be prepared to face another pandemic like it. Consequently, this article aims to develop a general modeling framework consisting of epidemiological modeling generalization and devising a practical identifiability approach to assess combined vaccination and NPIs strategies for the dynamics of any transmissible disease. To validate the framework, we applied it to track COVID-19 dynamics, accurately predict the ICU curve trend in the medium term, and assess vaccination in Chile.

The literature on COVID-19 modeling is vast. A search in the Web of Science (WOS) with the terms “COVID-19”, “modeling”, and “time delays”, refined by WOS categories related to STEM disciplines, resulted in 95 articles (in January 2023). Therefore, we only review some of those that we believe are important for their applications. For example, Al-Tuwairqi and Al-Harbi (1) proposed a model to investigate the effects of time delay in vaccine production on COVID-19 spread. In addition, Zhenzhen et al. (2) studied a model with “long memory” to describe the multi-wave peaks of the COVID-19 dynamics, where “long memory” allows for predicting this last using non-local terms, which means that one can include an arbitrary long history of the disease. Indeed, for a particular non-local term, the authors obtained a model with time delays. Furthermore, the authors modeled vaccination as an impulsive term that translates into decreased susceptibility.

Moreover, Ghosh et al. (3) derived a model with time delay, where the last is the disease duration, i.e., the average time in which infected individuals recover or die. Zhai et al. (4) investigated a SEIR-type model with time delay and vaccination control. The first parameter is similar to that introduced in our previous work (5) but is considered in the exposed population equation. They simulated vaccination as a control that decreases susceptibility, similar to the generalization we propose in the present work. Finally, our current work relies on generalizing the model developed by (5), which has common elements with some cited works here. Indeed, in (5), we introduced the same time delay that models the average time to recover or die, as in (3). At the same time, we could interpret our previous model (5) as one incorporating a long memory effect in the sense that it allows the reproduction of multi-wave peaks depending on the parameter values, as we showed.

The general goal is to implement a hybrid approach (6), in this case, a holistic combination of mathematical modeling with a practical identifiability approach to reconstruct and predict epidemiological curves based on careful optimization, synthetic data, automatic data scanning, and calibration. Precisely, the scientific novelty of the article relies on developing a general modeling framework that could contribute to anticipating epidemiological scenarios, evaluating the impact of combined

vaccination and NPI strategies for any transmissible disease, and helping to visualize the potential effects of implemented plans by policymakers.

To achieve the general goal, we rely on our previous work that forecasted COVID-19's second wave in May 2021 in Chile, calibrating data between March and September 2020 (before vaccination began) through suitable epidemiological modeling (5). Our specific goals are:

1. To generalize our previously developed epidemiological modeling to describe vaccination and assess combined vaccination and NPI strategies for the dynamics of any transmissible disease and, as a study case, of COVID-19.
2. To improve the practical identifiability approach to calibrate the generalized epidemiological modeling with different datasets representing different regions of Chile and accurately forecast the ICU curve trend in the medium term in any stage of the COVID-19 pandemic.
3. To provide open-source codes that implement our general epidemiological modeling framework with standardized parameter initialization for every dataset for reproducibility.

To implement the general modeling framework, we processed and used the official COVID-19 datasets provided by the Chilean government. The framework consists of two main parts: epidemiological modeling generalization and devising a practical identifiability approach. The first consists of a non-linear delay differential equations (DDE) system describing time variation and transitions between the compartments of susceptible, infected, recovered and the sum of ICU plus dead. The second relies on parameter optimization (5, 7, 8), a parametric bootstrap technique (9), and data processing. A novelty of this work is the implementation of a careful parameter optimization algorithm by searching for suitable initializations according to each processed dataset. In addition, we implemented a parametric bootstrap technique to accurately predict the ICU curve trend in the medium term and assess vaccination.

We have organized the article as follows: we describe the general modeling framework in Section 2. More precisely, based on our previous work (5), we describe the epidemiological modeling generalization in Section 2.1. Then, in Section 2.2, we detail the careful parameter optimization algorithm implementation, the parametric bootstrap technique, and data processing, among other methodology pieces. Then, we provide the modeling results to validate the framework with several datasets representing COVID-19 dynamics in different regions of Chile in Section 3. Finally, in Section 4, we discuss the results and give some conclusions in Section 5.

## 2. General modeling framework' description

In this section, we describe the general modeling framework. To do so, we have split it into two main subsections that generalize previously developed epidemiological modeling and devise the practical identifiability approach.

## 2.1. Epidemiological modeling generalization

In the present section, we describe the epidemiological modeling introduced by (5), then present and provide a complete description of each part of the modeling generalization.

### 2.1.1. Previous work

To derive the epidemiological modeling generalization, we rely on the *generalized SIR model with constant time delays* or *generalized SIR model* previously devised by (5), which describes the NPIs' effect through variations in the rate of disease transmission. We remark here that the NPIs' impact consists of social distancing and that the generalized SIR model does not describe vaccination. To be precise, the model corresponds to the following non-linear DDE system:

$$\frac{dS}{dt}(t) = -\frac{\beta(t)}{N}S(t)I(t - \tau_1), \tag{1a}$$

$$\frac{dI}{dt}(t) = \frac{\beta(t)}{N}S(t)I(t - \tau_1) - \gamma_{IR}I(t - \tau_2), \tag{1b}$$

$$\frac{dR}{dt}(t) = \gamma_{IR}I(t - \tau_2). \tag{1c}$$

It is worth noting that the generalized SIR model can generate complex dynamics since, by contrast to the classical SIR model, it can simulate more than one local maximum for the infected. Then, it provides a way to explain several COVID-19 waves, depending on the parameters' values (5). In addition, the generalized SIR model would produce better prediction results than the classical SEIR model since no observation of the exposed population is available since, similarly to the asymptomatic population, those exposed are challenging to observe. Indeed, the Chilean government's COVID-19 database (10), apart from the symptomatic cases, counts the asymptomatic ones, and there is no way to know how many of these become symptomatic (5).

The parameters of the model (1) are as follows:  $\beta(t)$  corresponds to the mean rate of disease transmission,  $\gamma_{IR}$  is the mean removal rate,  $\tau_1$  is the mean incubation time of disease, and  $\tau_2$  is the mean time from onset to clinical recovery or death caused by disease, or the duration time of disease until recovery or death.

Model (1) follows the susceptible-infectious-removed (SIR) paradigm. Susceptible (S) individuals infected by SARS-CoV-2 undergo incubation during a mean time ( $\tau_1$ ) before becoming infected (I). The infected individuals are infected by the disease for a mean time ( $\tau_2$ ) until being removed (R) by clinical recovery or death.

The initial conditions have to satisfy  $S(t_0) + I(t_0) + R(t_0) = N$ , where  $N$  is the size of the population under study for a closed system and taking into account that  $(S + I + R)'(t) = 0$  for all  $t > 0$ , for a suitably chosen  $t_0$ .

In the case of COVID-19, following the discussion by (5), we assume that the number of infected reported with symptoms confirmed by Reverse Transcriptional Polymerase Chain Reaction (RT-PCR) tests, denoted by  $I_r(t)$ , is underestimated since it depends on the availability and application of RT-PCR tests. Consequently, we assume that  $I_r(t)$  is a fraction of the *actual number of infected*  $I(t)$ ,

$$I_r(t) = f(t)I(t), \tag{2}$$

where  $f(t)$  is the ratio of positive RT-PCR tests number (confirmed cases) over the actual infected cases for the day  $t$ , which accounts for the *real positivity rate*. It is worth noting that  $f(t)$  is not the same as the positivity rate of detected cases, but it is related to it, and it involves the asymptomatic infected (5). We modeled  $f(t)$  as an inverted Sigmoid-type function such that if  $I(t)$  is small enough, which occurred during the beginning of the outbreak, then an important fraction of the real infected cases are detected ( $I_r(t) \approx I(t)$ ). On the contrary, when  $I(t)$  is large enough, which occurred just before the quarantines were imposed, then only a small fraction  $0 < a < 1$  of the real infected cases are detected ( $I_r(t) \approx aI(t)$ ). Precisely,  $f(t)$  is defined as

$$f(t) = 1 + \frac{a - 1}{1 + e^{-k(I(t) - I_{thr})}} \tag{3}$$

whose parameters are  $a$ ,  $k$  and  $I_{thr}$ , where  $a$  and  $k$  represent the minimum and the decay rate of  $f(t)$ , respectively. On the other hand, to measure how large/small  $I(t)$  is, we introduce a threshold  $I_{thr}$  such that  $I(t) \ll I_{thr}$  implies  $f(t) \approx 1$ , and  $I(t) \gg I_{thr}$  implies  $f(t) \approx a \ll 1$ .

Model (1), describing time variation and transitions S-to-I-to-R plus Equations (2)–(3), representing the real positivity rate, is quite general since it models the dynamics of any transmissible disease, not considering vaccination. The model parameters, gathered in vector  $(\beta(t), \gamma_{IR}, \tau_1, \tau_2, a, k, I_{thr})$ , are unknown or inaccessible and have to be identified from time-series observations  $I_r$  corresponding to the number of infected reported with symptoms confirmed by RT-PCR tests. In (5), we devised a practical identifiability approach, i.e., a set of techniques to reliably estimate parameters with acceptable accuracy from noisy data (11, 12). In particular, we reproduced epidemiological scenarios considering  $\beta(t)$  varying in time to describe NPI strategies ranging from total relaxation to imposing strict social distancing (complete lockdown differed by municipalities) at different periods. As a result, we forecasted the second wave of May 2021 in Chile, calibrating data between March and September 2020 (before vaccination began).

### 2.1.2. Toward modeling's generalization

To make model (1) more complete and realistic, we developed a generalization to measure the hospital load on the healthcare system through the number of patients hospitalized in the ICU and to assess the vaccination. Equations (2)–(3) that model the real positivity rate remain unchanged. In the following, we provide a modeling description of both aspects.

#### 2.1.2.1. Vaccination description

According to (13), vaccination protects in four ways: against infection, symptoms, severe disease, and reducing onward transmission. However, even considering part of the vaccination's ways of protection, the model can become very complex, as in the work by (13). Moreover, since COVID-19 data has great uncertainty, among other issues discussed in the data processing section, any model will provide results accordingly, no matter how exact its representation of reality is. Consequently, our present

work is in the same spirit as our previous work (5) in that keeping a model simple is critical to reasonably carrying out a practical identifiability approach. Even so, it is still a challenge to do it in real-time to anticipate epidemiological scenarios to help predict the hospital load on the healthcare system, as required in the first year of the pandemic (before vaccination). In this regard, we assume vaccination protects against infection by diminishing susceptibility, as assumed, for example, by (2, 4), which translates into adding new terms to Equations (1a) and (1c) for those susceptible and those recovered and several meaningful parameters associated with vaccination.

According to the previous discussion, the new terms that model vaccination account for the transition from susceptible to recovered, its inverse, and their respective times of transition. Transition rates are denoted by  $\gamma_{SR}(t)$  and  $\gamma_{RS}(t)$ , while the times model as delays within the new terms added, designated by  $\tau_3$  and  $\tau_4$ , respectively. We allow transition rates to vary in time to describe them more realistically since they depend on several factors that may change over time, as explained below (e.g., the immunity of an individual without booster doses decays faster).

Parameter  $\gamma_{SR}(t)$  relates mainly to the vaccination uptake rate, while  $\gamma_{RS}(t)$  relates to the waning of immunity after vaccination. Moreover, parameters  $\tau_3$  and  $\tau_4$  depend on every delivered vaccine's effectiveness and immunity waning. However, the vaccination uptake rate, effectiveness, and immunity waning of vaccines are not well-determined since they depend on several factors such as prior infection, age, sex, T-cell response, and the periodicity of vaccine injections. In addition, only natural infection mounts a significant and lasting immune response (14). Therefore, parameters  $\gamma_{SR}(t)$ ,  $\gamma_{RS}(t)$ ,  $\tau_3$ , and  $\tau_4$  depend on many factors, which makes it difficult to estimate them and they are strongly available data-dependent, independent of how exact the model's representation of reality is. Thus, they are inaccessible and must be identified from time-series observations of COVID-19, as noted before.

The model equations that consider vaccination are:

$$\frac{dS}{dt}(t) = -\frac{\beta(t)}{N}S(t)I(t - \tau_1) - \gamma_{SR}(t)S(t - \tau_3) + \gamma_{RS}(t)S(t - \tau_4), \tag{4a}$$

$$\begin{aligned} \frac{dR}{dt}(t) &= \gamma_{IR}(t)I(t - \tau_2) \\ &+ \gamma_{SR}(t)S(t - \tau_3) - \gamma_{RS}(t)S(t - \tau_4) + \gamma_{UR}(t)U(t - \tau_6). \end{aligned} \tag{4b}$$

The second and third terms of Equations (4a)–(4b) model the transitions from the susceptible to recovered compartment, and conversely,  $\tau_3$  stands for the mean time delay for those susceptible to become immune after vaccination, and  $\gamma_{SR}(t)$  indicates how fast it happens. Similarly,  $\tau_4$  designates the mean time until an individual loses immunity, so  $\tau_4$  is the mean duration of immunity by vaccination, and  $\gamma_{RS}(t)$  measures how fast it happens. Finally, the last term in Equation (4b) pertains to the  $U$  compartment, which we explain below.

### 2.1.2.2. $U$ compartment description and modeling generalization summary

Finally, as mentioned before, we added the variable  $U$  to model (1), representing the sum of the patients in the ICU plus those confirmed dead due to COVID-19, the equation of which contains

the transitions from  $I$ -to- $U$  and  $U$ -to- $R$ . The variable  $R$  now describes the recovered, whereas  $R$  in the model (1) represented the removed, i.e., the sum of those who had recovered plus those who had died.

As before, the mentioned transitions encompass rates and time delays, denoted by  $\gamma_{IU}(t)$ ,  $\tau_5$ ,  $\gamma_{UR}(t)$ , and  $\tau_6$ . Precisely, we model the transition  $I$ -to- $U$  by the rate  $\gamma_{IU}(t)$  and the time  $\tau_5$  that those infected took to be admitted to the ICU. In addition, we represent the transition  $U$ -to- $R$  by the rate  $\gamma_{UR}(t)$  and the time  $\tau_6$  that patients took to recover in the ICU.

We summarize the modeling generalization as the following non-linear DDE system:

$$\frac{dS}{dt}(t) = -\frac{\beta(t)}{N}S(t)I(t - \tau_1) - \gamma_{SR}(t)S(t - \tau_3) + \gamma_{RS}(t)S(t - \tau_4), \tag{5a}$$

$$\frac{dI}{dt}(t) = \frac{\beta(t)}{N}S(t)I(t - \tau_1) - \gamma_{IR}(t)I(t - \tau_2) - \gamma_{IU}(t)I(t - \tau_5), \tag{5b}$$

$$\begin{aligned} \frac{dR}{dt}(t) &= \gamma_{IR}(t)I(t - \tau_2) + \gamma_{SR}(t)S(t - \tau_3) - \gamma_{RS}(t)S(t - \tau_4) \\ &+ \gamma_{UR}(t)U(t - \tau_6), \end{aligned} \tag{5c}$$

$$\frac{dU}{dt}(t) = \gamma_{IU}(t)I(t - \tau_5) - \gamma_{UR}(t)U(t - \tau_6). \tag{5d}$$

Again, the previous system is closed since the variables' sum equals  $N$ , the size of the targeted population. The DDE system (5) and Equations (2)–(3) will be named *general epidemiological modeling*, which is quite broad since it models the dynamics of any transmissible disease under any combination of NPIs and vaccination. To describe simply the in-time-variation of parameters  $\beta(t)$ ,  $\gamma_{IU}(t)$ , and  $\gamma_{UR}(t)$ , we assumed that these are piecewise linear functions. Then, the functions  $\beta(t)$ ,  $\gamma_{IU}(t)$ , and  $\gamma_{UR}(t)$  are represented by the vectors  $\beta$ ,  $\gamma_{IU}$  and  $\gamma_{UR}$  in  $\mathbb{R}^{n_\beta+1}$  that represent  $n_\beta$  straight lines approximating the respective functions. We gave the same description for the in-time-variation of parameters  $\gamma_{SR}(t)$ ,  $\gamma_{RS}(t)$ , and  $\gamma_{IR}(t)$ , i.e., they are represented by the vectors  $\gamma_{SR}$ ,  $\gamma_{RS}$ , and  $\gamma_{IR}$  in  $\mathbb{R}^{n_\gamma+1}$ .

Therefore, general epidemiological modeling depends on  $p := 3(n_\beta + n_\gamma) + 15$  parameters gathered in the vector  $\theta \in \mathbb{R}^p$  defined by:

$$\theta = (\theta_1, \theta_2) \in \mathbb{R}^p, \tag{6a}$$

$$\theta_1 = (\gamma_{IR}, \gamma_{SR}, \gamma_{RS}, \gamma_{IU}, \beta, \gamma_{UR}, \tau)^t \in \mathbb{R}^{3(n_\gamma+n_\beta)+12}, \tag{6b}$$

$$\gamma_{IR}, \gamma_{SR}, \gamma_{RS} \in \mathbb{R}^{n_\gamma+1}, \tag{6c}$$

$$\beta, \gamma_{IU}, \gamma_{UR} \in \mathbb{R}^{n_\beta+1}, \tag{6d}$$

$$\tau = (\tau_1, \tau_2, \tau_3, \tau_4, \tau_5, \tau_6)^t \in \mathbb{R}^6, \tag{6e}$$

$$\theta_2 = (a, k, I_{thr})^t \in \mathbb{R}^3, \tag{6f}$$

where  $n_\gamma$ ,  $n_\beta$  is the number of time intervals to reconstruct  $\gamma_{SR}(t)$ ,  $\gamma_{RS}(t)$ , and  $\gamma_{IR}(t)$ , and  $\beta(t)$ ,  $\gamma_{IU}(t)$ ,  $\gamma_{UR}(t)$  piecewise linearly with equally spaced intervals, respectively. For instance, using  $n_\beta = 9$ ,  $n_\gamma = 9$ , one has  $p = 3(n_\beta + n_\gamma) + 15 = 69$  parameters to estimate, i.e.,  $\theta \in \mathbb{R}^{69}$ .

Table 1 summarizes the parameters of general epidemiological modeling given by Equations (5), (2), and (3).



Table 1 Parameters of general epidemiological modeling.

Symbol	Description	Unit
$\tau_1$	Mean incubation time	days
$\tau_2$	Mean time to recover for mild cases	days
$\tau_3$	Mean time from susceptible to recovery (by vaccination immunity)	days
$\tau_4$	Mean duration of immunity (by vaccination)	days
$\tau_5$	Mean time from infected to ICU	days
$\tau_6$	Mean time from ICU to recover	days
$\gamma_{IR}$	Mean recovery rate for mild cases	days <sup>-1</sup>
$\beta$	Mean transmission rate	days <sup>-1</sup>
$\gamma_{IU}$	Mean transition rate from infected to ICU	days <sup>-1</sup>
$\gamma_{UR}$	Mean recovery rate for ICU patients	days <sup>-1</sup>
$\gamma_{SR}$	Mean transition rate from susceptible to recovered	days <sup>-1</sup>
$\gamma_{RS}$	Mean transition rate from recovered to susceptible	days <sup>-1</sup>
$a$	Minimum of the real positivity rate	-
$k$	Decay rate of the real positivity rate	inhabitants <sup>-1</sup>
$I_{thr}$	Infection threshold of the real positivity rate	inhabitants

## 2.2. A practical identifiability approach

We devised a practical identifiability approach that relies on parameter optimization, a parametric bootstrap technique, and data processing, for which computer implementation includes open-source data and code repository through GitHub (15). The approach relies on solving a parameter estimation problem through a careful optimization algorithm, numerical resolution of modeling equations, a parametric bootstrap technique, and data processing. For solving the modeling equations, we required the provision of reasonable bounds for the parameters, which is critical for achieving a stable numerical method. In addition, the parameter range is meaningful, at least regarding the time delays that describe relevant parameters from the epidemiology viewpoint.

Next, we describe each piece of the practical identifiability approach.

### 2.2.1. Parameter estimation problem description

To reproduce and predict COVID-19 dynamics in Chile, one has to solve the *parameter estimation problem*: given a dataset of the time-series observations of COVID-19 dynamics, identify the parameter vector  $\theta$  such that general epidemiological modeling fits them in the least-squares sense. The time-series observations we used are

$$\{(I_r)_j, (U_r)_j\} : j = 1, \dots, n\}.$$

$I_r$  corresponds to the number of infected reported with symptoms confirmed by RT-PCR tests [see Equation (2) and its respective explanation],  $U_r$  corresponds to the observations of variable  $U$ , i.e., the sum of the patients in the ICU and confirmed deaths due to COVID-19, and  $n$  is the number of data.

More precisely, we have to find the vector  $\theta \in \mathbb{R}^p$ , defined by Equation (6), that minimizes the sum of squares:

$$SS(\theta) := \|(Res_I, Res_U)^t\|^2 = \sum_{j=1}^n [(Res_I)_j^2 + (Res_U)_j^2] \quad (7)$$

where  $Res_I$  and  $Res_U \in \mathbb{R}^n$  stand for the relative residuals of variable  $I$  and  $U$ , respectively, defined by

$$(Res_I)_j := \frac{[(I_r)_j - f(t_j, \theta_2)I(t_j, \theta_1)]}{f(t_j, \theta_2)I(t_j, \theta_1)} \quad j = 1, \dots, n \quad (8a)$$

$$(Res_U)_j := \frac{[(U_r)_j - U(t_j, \theta_1)]}{U(t_j, \theta_1)} \quad j = 1, \dots, n. \quad (8b)$$

The objective function defined in (7) corresponds to the sum of squares of the residuals relative to the model observations,  $f(t_j, \theta_2)I(t_j, \theta_1)$  and  $U(t_j, \theta_1)$ . The choice of the relative residuals obeys to take into account the unequal quality of the observations (16). In this case, the patients in the ICU plus confirmed deaths due to COVID-19 (variable  $U$ ) is better observed than those infected (variable  $I$ ).

The components of  $Res_I$ , defined in (8a), correspond to the differences between the time-series observations  $(I_r)_j$ , and the model observations  $f(t_j, \theta_2)I(t_j, \theta_1)$  (see Equation 2). The components of  $Res_U$ , defined in (8b), are the differences between the time-series observations  $(U_r)_j$ , and the model observations  $U(t_j, \theta_1)$ . Both variables,  $I$  and  $U$ , correspond to the solution of the general epidemiological model (5), (2), and (3) calculated at  $(t_j, \theta)$  for  $j = 1, \dots, n$ , for a given parameter vector  $\theta = (\theta_1, \theta_2) \in \mathbb{R}^p$  defined by (6).

### 2.2.2. A careful optimization algorithm

To calibrate general epidemiological modeling, we implemented a careful optimization algorithm that combines data processing algorithms (cleaning, smoothing, and curve interpolation), initial parameter estimation generation, and a non-linear least-squares optimization method to minimize  $SS(\theta)$  [Equation (7)] for each processed dataset.

The argument minimum of the sum of squares designs as  $\hat{\theta}$  for every dataset. The vector  $\hat{\theta}$  is called the *non-linear least-squares estimator*, abbreviated as *non-linear LSE*.

Next, we provide details on the implementation.

#### 2.2.2.1. Implementation

We employed the *Trust-Region Interior Reflective* (TIR) method implemented in Matlab R2022b as the subroutine *lsqnonlin*, specially adapted for solving non-linear least-squares minimization problems. The convergence of the TIR method depends strongly on initial parameter estimation, which has to be relatively close to

the optimal solution (7, 8). We efficiently minimized the objective function by implementing a *percentage decrease technique* from the parameters' ranges to calculate suitable initial parameter vectors for every dataset in a standardized manner. According to our previous experience (5), the initial parameters that mainly influence the fitting results are  $\beta, \gamma_{IU}, \gamma_{UR} \in \mathbb{R}^{n_{\beta}+1}$ , which describe the mean rate of disease transmission and transitions from *I*-to-*U* and *U*-to-*R*. To avoid overfitting, we selected  $n_{\beta}$  equispaced intervals to accurately fit and predict after the final calibration time for every dataset.

Next, we explain the numerical resolution of general epidemiological modeling within a range for meaningful parameters and the percentage decrease technique, which are critical for implementing our careful optimization algorithm.

### 2.2.2.2. Numerical resolution of general epidemiological modeling

To evaluate the objective function  $SS(\theta)$ , we numerically solved the model (5) at  $t_j, j = 1, \dots, n$ , for different parameter vectors  $\theta$  chosen *ad-hoc* for each dataset. We carried out the numerical resolution by a Runge-Kutta type formula (17): the subroutine *dde23* implemented in Matlab, designed for solving non-linear DDE systems. In addition, we reconstructed the function of history (required for solving DDE instead of the initial condition for classical ordinary differential equations systems) for the model (5) by interpolating the data  $I_r(t), U_r(t)$ , and recovered for every studied dataset. We used a shape-preserving piecewise cubic interpolation as devised by the *interp1* Matlab subroutine with the option *pchip*.

### 2.2.2.3. Meaningful parameter bounds

We computed the model parameters  $\theta$  given in (6) using meaningful bounds from an epidemiological viewpoint for the time delays:  $1 \leq \tau_1 \leq 14, 1 \leq \tau_2 \leq 21, 14 \leq \tau_5 \leq 56$ , and  $21 \leq \tau_6 \leq 42$  days. We imposed these bounds because the incubation period ( $\tau_1$ ) ranges from 1 to 14 days (mean of 5–6 days), the median time from onset to clinical recovery for mild COVID-19 cases is approximately 2 weeks ( $\tau_2$ ), and is 3–6 weeks for patients with severe or critical symptoms ( $\tau_6$ ). In addition, among patients who died, the time from symptom onset to outcome ranged from 2 to 8 weeks ( $\tau_5$ ) (18). In addition, we assume that vaccination immunity duration,  $\tau_4$ , ranges from 1 to 240 days since immunity declines only at 6–8 months after natural infection (19). In contrast, no range is well-determined for the transition time from susceptible to recovered,  $\tau_3$ , however, one may expect that it is relatively small, so we assume  $1 \leq \tau_3 \leq 14$ .

For the real positivity fraction,  $f(t)$ , one has that

$$0 < a < 1, \quad \min\{(I_r)_j : j = 1, \dots, n\} \leq I_{thr} \leq \max\{(I_r)_j : j = 1, \dots, n\}.$$

Finally, all the rest of the parameters (transition and transmission rates) have to be within ranges to achieve stability of the careful optimization algorithm implementation, mainly related to general epidemiological modeling numerical resolution, as we explain next.

### 2.2.2.4. Percentage decrease technique for numerical stability

Convergence of the optimization algorithm, *lsqnonlin*, depends directly on that of the *dde23* solver, both implemented in Matlab. Through our experiments, we verified that the model's numerical solution calculated by *dde23* strongly depends on the derivatives of the first points evaluated, and its convergence relies on the closeness of the initial curves used to build the history function. To overcome this stability problem without intervening or designing new Matlab numerical libraries, we developed a simple but effective technique, called the *percentage decrease technique*, to produce initial curves contained in the feasible space of the official data curves.

The percentage decrease technique consists of multiplying the upper bounds of the parameter vector  $\theta$  (see Equation 6) by a fraction  $\omega \in [1e - 4, 1e - 2]$ , excluding time delays  $\tau$  (see Equation 6e). Concretely, we chose the initial parameter vector defined by  $\theta^{(0)} := \omega \cdot UB_{\theta}$ , where  $UB_{\theta}$  stands for the upper bound of  $\theta$ , described in Section 2.2.2.3.

We calibrated  $\omega$  for every studied dataset representing a characteristic epidemiological curve, with the magnitude or period of peaks' duration differentiated, which is mainly related to the density and mobility of the population. We chose a value of  $\omega$  inversely proportional to the population size within the interval  $[1e - 4, 1e - 2]$ . Therefore, one should use a small value, e.g.,  $\omega = 1e - 3$  or 0.1% of  $UB_{\theta}$  for large cities (millions of inhabitants), and an even smaller value, e.g.,  $\omega = 1e - 4$  or 0.01% of  $UB_{\theta}$  for towns with fewer inhabitants (less than one million). Generally,  $\omega = 1e - 3$  works well for all cases obtaining a similar minimum error. Still, we calibrated a suitable value of  $\omega$  for every dataset to speed up the execution time of the optimization subroutine *lsqnonlin*.

### 2.2.3. A parametric bootstrap technique

Once we calibrated a dataset, we applied a parametric bootstrap technique (PBT) to quantify the parameters' uncertainty and construct confidence intervals to achieve reliable and accurate forecasting performance, obtained by propagating the uncertainty (9). The PBT generates synthetic datasets repeatedly sampled from the least-squares curves:

$$F(t_j, \hat{\theta}) := [f(t_j, \hat{\theta}_2) I(t_j, \hat{\theta}_1), U(t_j, \hat{\theta}_1)]. \tag{9}$$

However, the PBT requires intensive computational resources and time since it generates simulated data from  $F(t_j, \hat{\theta})$  and calculates least-squares parameter estimates for each generated synthetic dataset. Therefore, we applied this technique to show the predictive power of general epidemiological modeling and assess vaccination only for one dataset, corresponding to days 280–530 that encompass the Metropolitan region's second wave, and considering the last 12 weeks to test the forecasting performance of the PBT.

In the PBT, one assumes that synthetic data follows a given probability distribution with an expected value equal to the least-squares curves  $F(t_j, \hat{\theta})$ . To be precise, we implemented the following algorithm:

1. We calculated the parameter estimates  $\hat{\theta}$  through least-squares fitting the model to the time-series data to obtain the best-fit model given by  $F(t_j, \hat{\theta})$  (see Equation 9).
2. Using the least-squares fitted model  $F(t_j, \hat{\theta})$ , we generated  $M$  replicated synthetic datasets denoted by  $F_1^{SD}(t_j, \hat{\theta}), F_2^{SD}(t_j, \hat{\theta}), \dots, F_M^{SD}(t_j, \hat{\theta})$ . We generated synthetic datasets as random vectors with a mean equal to  $F(t_j, \hat{\theta})$ :

$$F_k^{SD}(t_j, \hat{\theta}) \sim \text{Dist} \left[ F(t_j, \hat{\theta}) \right] \tag{10}$$

where **Dist** is a given probability distribution of mean equal to  $F(t_j, \hat{\theta})$ , and variance proportional to the mean magnitude or the covariance matrix of it. For instance, we used the normal, Poisson, and negative binomial distributions. The last is adequate to model data over-dispersion while controlling its magnitude (9).

3. We re-calculated the least-squares parameter estimates fitting the model to each of the  $M$ -simulated datasets realizations. Each parameter vector is denoted by  $\hat{\theta}_\ell$  for  $\ell = 1, 2, \dots, M$ .
4. Using the set of re-estimated parameters  $\hat{\theta}_\ell, \ell = 1, 2, \dots, M$ , we calculated a confidence interval at the level of 95%. The uncertainty around the least-squares model fit is given by  $F(t_j, \hat{\theta}_1), F(t_j, \hat{\theta}_2), \dots, F(t_j, \hat{\theta}_M)$ .

Typical values for the number of bootstrap samples  $M$  range from 50 to 200 for a proper standard error estimation; see p. 13–14 in (20). Indeed, by choosing  $M = 100$ , the standard error estimate provides reliable results for parameter estimation, as shown in Section 4. However, beyond choosing a good number of bootstrap samples  $M$ , careful implementation of the PBT is critical for obtaining  $\hat{\theta}_\ell$  corresponding to  $F(t, \hat{\theta}_\ell)$  so that the curves that are reproduced and predicted remain positive for all  $t$  within the time interval of calibration and prediction.

The calculated confidence interval is our prediction interval, denoted by PI, and defined by:

$$\begin{aligned} \text{PI} &:= [\text{LB}_{\hat{\theta}}, \text{UB}_{\hat{\theta}}] := \left[ \bar{\hat{\theta}} \pm t_{0.975, M-1} \text{SE} \right] \\ &= \bar{\hat{\theta}} \left[ 1 \pm t_{0.975, M-1} \text{NSE} \right], \end{aligned} \tag{11a}$$

$$\text{NSE} := \frac{\text{SE}}{\bar{\hat{\theta}}} = \frac{1}{\sqrt{M}} \frac{s_{\hat{\theta}}}{\bar{\hat{\theta}}}, \tag{11b}$$

where  $\bar{\hat{\theta}} \in \mathbb{R}^p$  is the mean of the  $M$  bootstrap parameters  $\hat{\theta}_\ell \in \mathbb{R}^p$ ,  $t_{0.975, M-1}$  is the 0.975 percentile of the  $t$ -student distribution with  $M - 1$  degrees of freedom,  $s_{\hat{\theta}} \in \mathbb{R}^p$  is the standard deviation of the  $M$  bootstrap parameters, and **NSE**  $\in \mathbb{R}^p$  is the normalized standard error (5) (the definitions of  $s_{\hat{\theta}}$  and **NSE** are understood component-by-component).

The resulting uncertainty around the least-squares model fit,  $F(t, \hat{\theta})$ , is quantified by the 95% confidence bounds [LB $_{\hat{\theta}}$  and UB $_{\hat{\theta}}$ , defined by (11)] (9). Since we are interested in computing a more accurate prediction for the ICU curve trend in the medium term, we define an error criterion that includes performance for fit and forecasting. More precisely, we define

$$E(\hat{\theta}_\ell) := 0.6\text{FP}_U(\hat{\theta}_\ell) + 0.2\text{FP}_I(\hat{\theta}_\ell) + 0.1[\text{RMSE}_U(\hat{\theta}_\ell) + \text{RMSE}_I(\hat{\theta}_\ell)], \ell = 1, \dots, M. \tag{12}$$

In Equation (12), the root mean squared error (RMSE) to measure the fit performance is defined by (16):

$$\begin{aligned} \text{RMSE}_I &= \left[ \frac{1}{n-p} \sum_{j=1}^n (\text{Res}_I)_j^2 \right]^{1/2}, \\ \text{RMSE}_U &= \left[ \frac{1}{n-p} \sum_{j=1}^n (\text{Res}_U)_j^2 \right]^{1/2}. \end{aligned} \tag{13}$$

We calculated the RMSE over the calibration period ( $n$  and  $p$  are the calibrated dataset size and the number of model parameters, respectively). A similar criterion was employed for the forecasting performance (FP), but the sum over the prediction period was computed and normalized by the number of predicted data points.

In Equation (12), we gave more weight (60%) to  $\text{FP}_U$  since variable  $U$  is better observed than variable  $I$ , followed by  $\text{FP}_I$  (20%) and the RMSE for both variables (10% each). Furthermore, it was more important to make more effort to follow up with the sum of the ICU patients and those who had died than those who were infected. Thus, we first focused on having better forecasting performance for ICU patients plus those who died and then on the number of infected persons reported. We gave less importance to the fitting performance, so the model did not overfit to actual data, which have much uncertainty, among other problems discussed in Section 3.1.

From the error criterion given in (12), we define the best parameter vector, denoted by  $\hat{\theta}_{\hat{k}}$ , among the vectors  $\hat{\theta}_\ell$  for  $\ell = 1, \dots, M$  that minimizes  $E(\hat{\theta}_\ell)$ , i.e.,

$$E(\hat{\theta}_{\hat{k}}) := \min_{1 \leq \ell \leq M} E(\hat{\theta}_\ell) \tag{14}$$

The best parameter vector defined in this way realizes the minimum error of the PBT we implemented and induces the *best model curves*, which privilege the forecasting performance. So, the best model curves are those evaluated at the best parameter vector,  $F(t, \hat{\theta}_{\bar{k}})$ . In addition, the prediction interval PI produces an envelope of model curves  $F(t, \hat{\theta}_{\ell})$  for  $\ell = 1, \dots, M$ , quantifying uncertainty around the best model curves  $F(t, \hat{\theta}_{\bar{k}})$ .

Finally, to assess the impact of vaccination, we compared different immunity durations, the parameter  $\tau_4$ , from 60 to 240 days, spaced every 30 days. For that, we calculated a weighted mean between the least PBT error  $E(\hat{\theta}_{\bar{k}})$  given in (14) with the mean of the NSE given in (11b), denoted by  $\overline{\text{NSE}}$ . We evaluated the best  $\tau_4$  as the value minimizing the weighted error, WE, defined by

$$\text{WE} := \frac{2}{3}E(\hat{\theta}_{\bar{k}}) + \frac{1}{3}\overline{\text{NSE}}. \tag{15}$$

We gave more weight (66.67%) to the minimum PBT error and the least weight (only 33.33%) to the mean of the NSE to privilege the FP in (14) over the uncertainty represented by the NSE in (11b), which was within reasonable bounds.

### 2.2.4. Data processing and actual data limitations

The datasets' sources correspond to the Chilean government's COVID-19 database at the regional level, which we used to track and predict COVID-19 dynamics (10). We used datasets representing different regions of the north, south, and center, including the Metropolitan Region (MR), the major city of which is the country's capital, Santiago de Chile. They correspond to reported infected persons with symptoms confirmed by RT-PCR tests ( $I_r$ ), recovered cases ( $R$ ), patients in the ICU, those confirmed dead due to COVID-19 ( $D$ ), and the size of targeted populations ( $N$ ). We applied mobile averages with different window sizes to deal with data that were not always reported daily and to smooth out the data.

#### 2.2.4.1. Available actual data limitations

A common problem in processing and modeling epidemiological curve data is the time lag of the information reported, which was not daily but accumulated every 3–5 days in the Chilean case as observed in Figure 1, which depicts various sub-peaks in the official Chilean data curves, specifically in the data of those infected ( $I$ ), recovered ( $R$ ), and deceased ( $D$ ). In addition, the relevant information is only detailed at the national level,

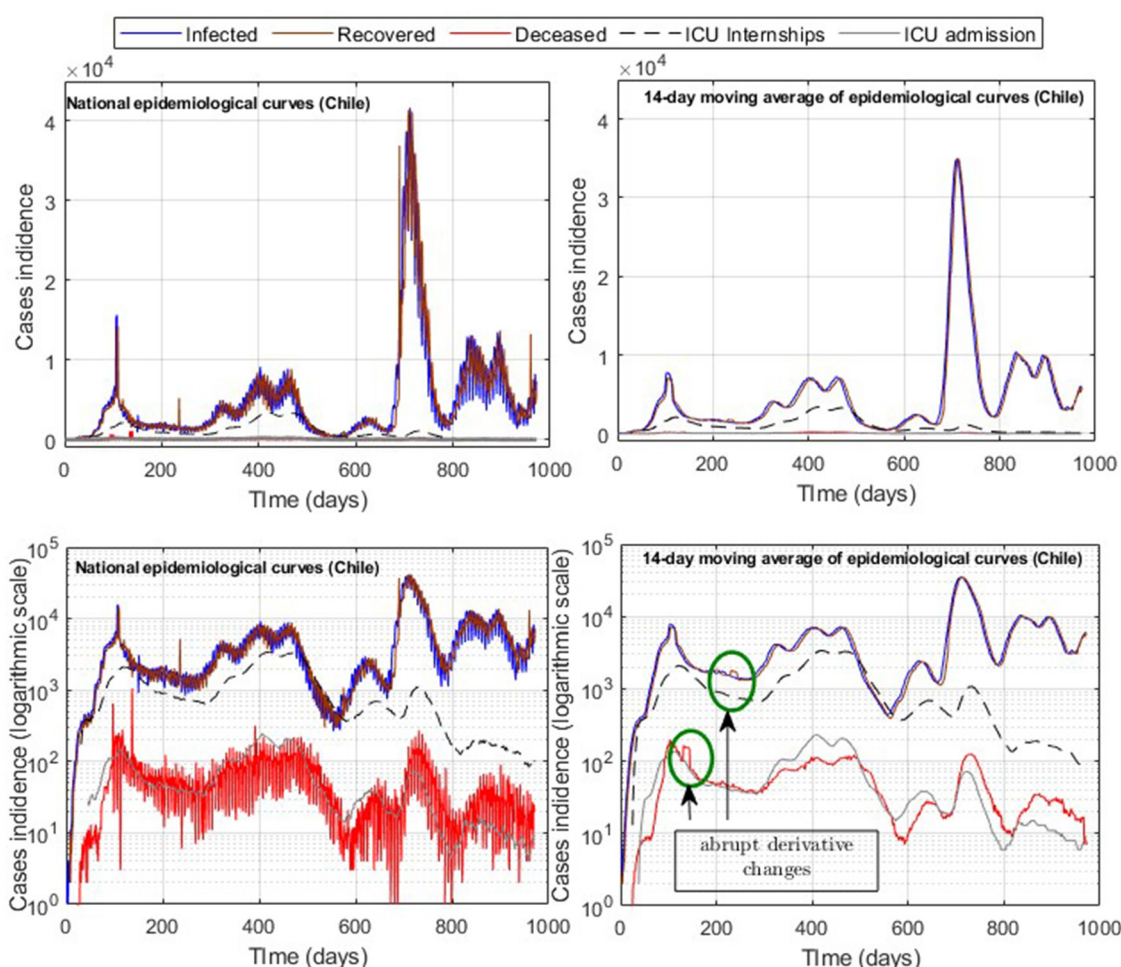


FIGURE 1 Official COVID-19 incidence curves (Chile); raw data and processing with moving averages.



such as daily admissions to the ICU and COVID-19 reinfections according to the vaccination scheme, which was accumulated weekly, and provisional since it was being validated [(10), product 90]. Therefore, using data in constant validation does not ensure a precise replication of the results, unlike *I*, *R*, and *ICU*, which have been maintained without significant variation over time. In addition, in the case of deaths in Chile at the beginning of 2022, the probable cases (without a prior confirmation test for COVID-19) were added to the official data (more than 10,000 cases) for the daily death count. Since the daily distribution of such cases is unknown, in our study, we only considered confirmed death due to COVID-19 reported daily.

Although different epidemiological datasets are available in the official Chilean repository (10), most are designed for statistical studies rather than for modeling studies. The most consistent data (without repetition of cases) is deceased persons, which has a high correlation (Pearson coefficient of 0.8) with hospitalized ICU patients (Figure 2). Therefore, in addition to being relevant data to determine the hospital load, hospitalized ICU patients are of high value in terms of the quality of the available data, from which one can conduct modeling studies at the regional level for the Chilean case. Furthermore, the interest in approximating regional curves and not only the national ones is based on the fact that it allows us to validate our general modeling framework with several datasets and analyze the epidemic in different geographical areas. The previous consideration is relevant to Chile, the longest country in the world, with different climates from north to south, and therefore a good case study for the present work. Finally, we applied our general modeling framework to study the COVID-19 data available (*I*, *R*, *ICU*, and *D*) and, in the future, we will extend and adapt to any transmissible disease in any country. It is feasible given that in most of the world, it is more viable to track critical cases (those hospitalized in ICUs), deaths, and, to a lesser extent, those who are infected/recovered (the actual total is never reached).

2.2.4.2. Hardware, software, and parallel computing

We used a data science workstation for the careful optimization algorithm implementation with the following features: Intel Core i9

7900x, 10 Cores/20 threads, 128 Gb memory, NVIDIA Titan RTX 24 Gb, and two mobile laptops with Intel Core i7s, 4 Cores/8 threads with 16 Gb memory. We implemented all our calibration codes by using the software Matlab R2022b. In addition, we used the Matlab parallel computing toolbox to speed up the computation with the parpool (“Processes,” 20) option on the workstation.

For implementing the PBT (forecasting), we used two mobile laptops equipped with an Intel Core i7 processor with 8 and 12 cores, respectively, with 16 Gb of RAM without parallel computing.

### 3. Results

#### 3.1. General modeling framework calibration results

##### 3.1.1. Data processing results

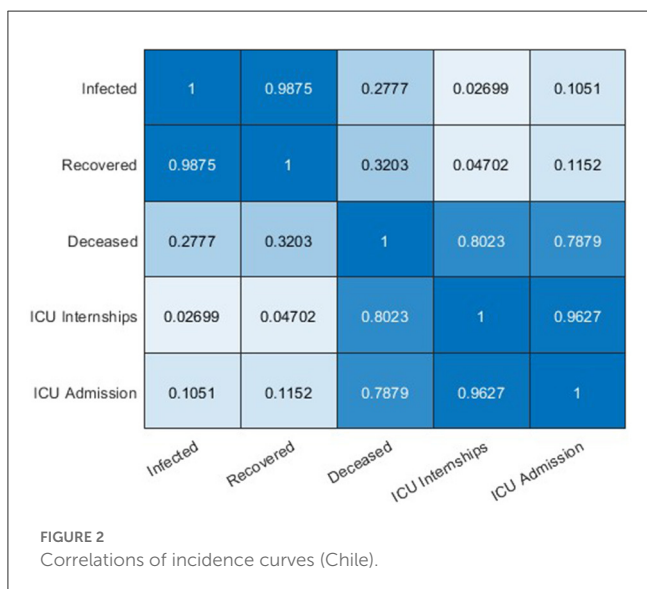
Data processing shows that the highest correlated variables are *I* and *R*, followed by ICU admission with ICU hospitalized and deceased with ICU hospitalized (Figure 3). The two latter are the most relevant indicators of the impact of the pandemic on the health system, which justifies our choice of privileging the forecasting performance in *U*, as made in Equation (12).

In addition, the optimization results strongly depend on the parameters’ initial values and the fitted data quality. By applying moving averages of 14 days, we reduced abrupt slope changes of the actual epidemiological curves (Figure 1), improving results with information loss of less than 2% of incidences for both infected and ICU. In addition, using moving windows combined with the cumulative distribution curves (Figure 3) reduces the number of steps of the Matlab dde23 solver (for non-linear DDE systems) because the slopes are always positive and not as steep as for daily curves. In this way, our method allows for calibrating parameters with accumulated and daily curves, where the cumulative data is helpful in efficiently fitting from the pandemic’s beginning to any subsequent location and day to perform block tracking.

##### 3.1.2. Calibration results

The percentage decrease technique, described in Section 2.2.2.4, ensures the overall method stability and convergence by preventing the NaN appearance (NaN means “Not a Number”) when overflow occurs (computational numerical limit exceeded). NaNs are propagated in the model’s history function, making the optimization solver lsqnonlin require more time to find new feasible points and often diverge. Using different  $\omega$  values in  $\theta^{(0)} := \omega \cdot UB_{\theta}$ , one may generate suitable initial curves for *I* and *U* below the epidemiological curves. For small  $\omega$  values ( $\omega \in [1e-4, 1e-2]$ ), the initial curve will approximately be a straight line, obtained with government data, and greater than zero. By applying this simple but effective technique, our careful optimization algorithm implementation ensures stability and convergence for the dde23 solver and the lsqnonlin optimization subroutine within the range of suitable parameters, as described in Section 2.2.2.3.

Furthermore, to speed up the computation, we performed a one-step initial optimization defining the real positivity rate  $f(t) = 1$ . This was the same as taking  $I(t) = I_r(t)$ , i.e.,



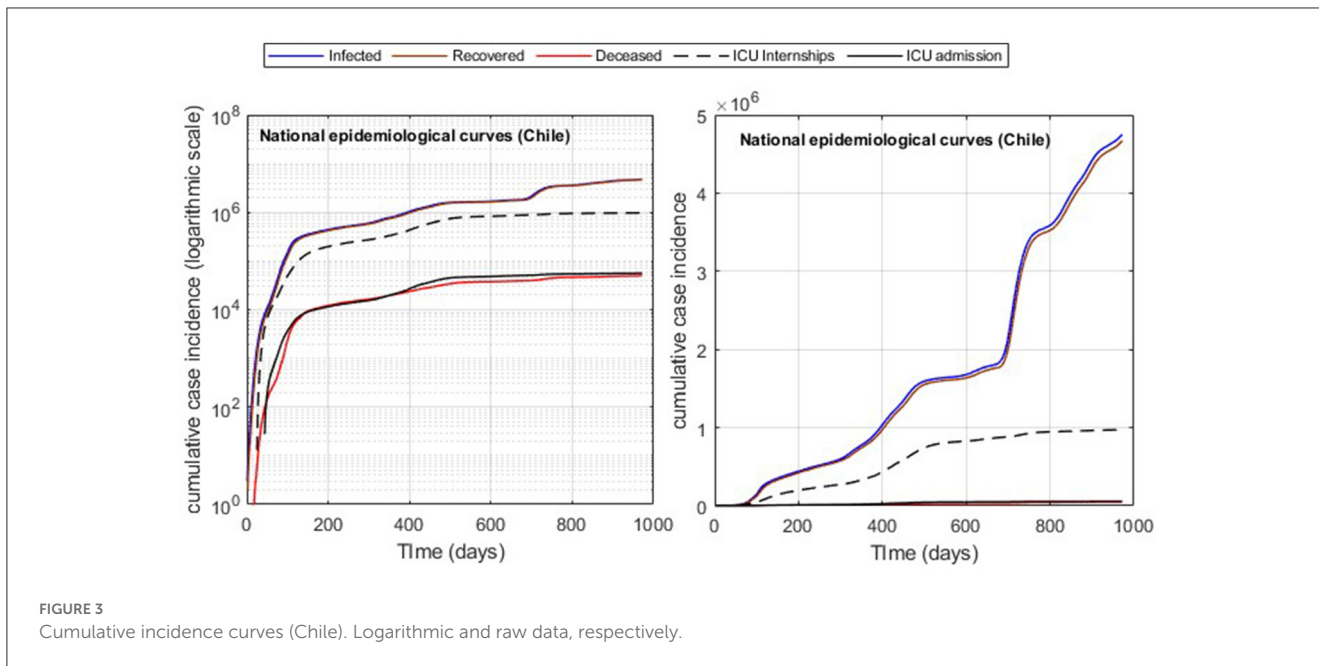


FIGURE 3 Cumulative incidence curves (Chile). Logarithmic and raw data, respectively.

Table 2 Model results for different  $\omega$ .

Id	Fraction $\omega$	Curve/window	$\overline{RE}_I$	$\overline{RE}_R$	$\overline{RE}_{ICU}$	$n_\beta + 1$	$t$ (days)	$UB_{\tau_4}$
MR.1, Figure 4	1e-03	cum/2 weeks	3.2808e-01	4.1021e-01	1.8342e-01	20	[30, 950]	240
MR.2, Figure 5	1e-03	cum/2 weeks	8.2663e-02	1.4144e-01	1.99823e-02	20	[30, 250]	240
MR.3, Figure 6	1e-03	cum/2 weeks	4.5210e-02	1.6722e-01	9.4964e-03	20	[30, 250]	50
VAL.1, Figure 7	1e-03	cum/3 weeks	1.0394e+00	6.9195e-01	7.9057e-01	20	[250, 950]	240
ANT.1, Figures 8C, D	3.5e-03	daily/3 weeks	1.8747e-01	1.0747e+00	8.2074e-02	20	[250, 600]	240
MAG.1, Figure 9	2.5e-04	cum/2 weeks	9.9075e-01	3.2812e+00	3.4139e+00	20	[100, 950]	240

all the actual infected are reported on the entire curve and storing the resulting  $\omega$  as a checkpoint. Then, when required to optimize the curve over any time interval, the obtained  $\omega$  was used as the initial value, starting the computation closer to the optimum.

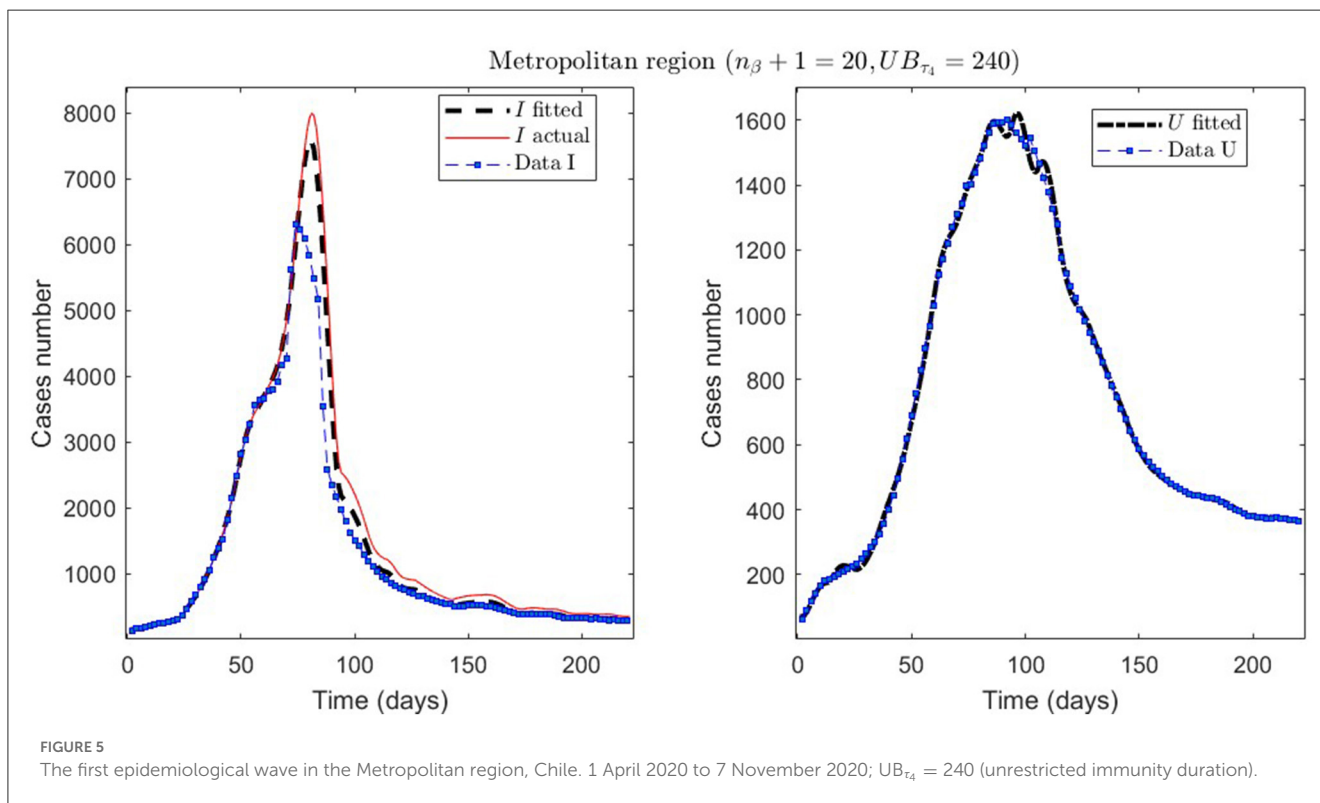
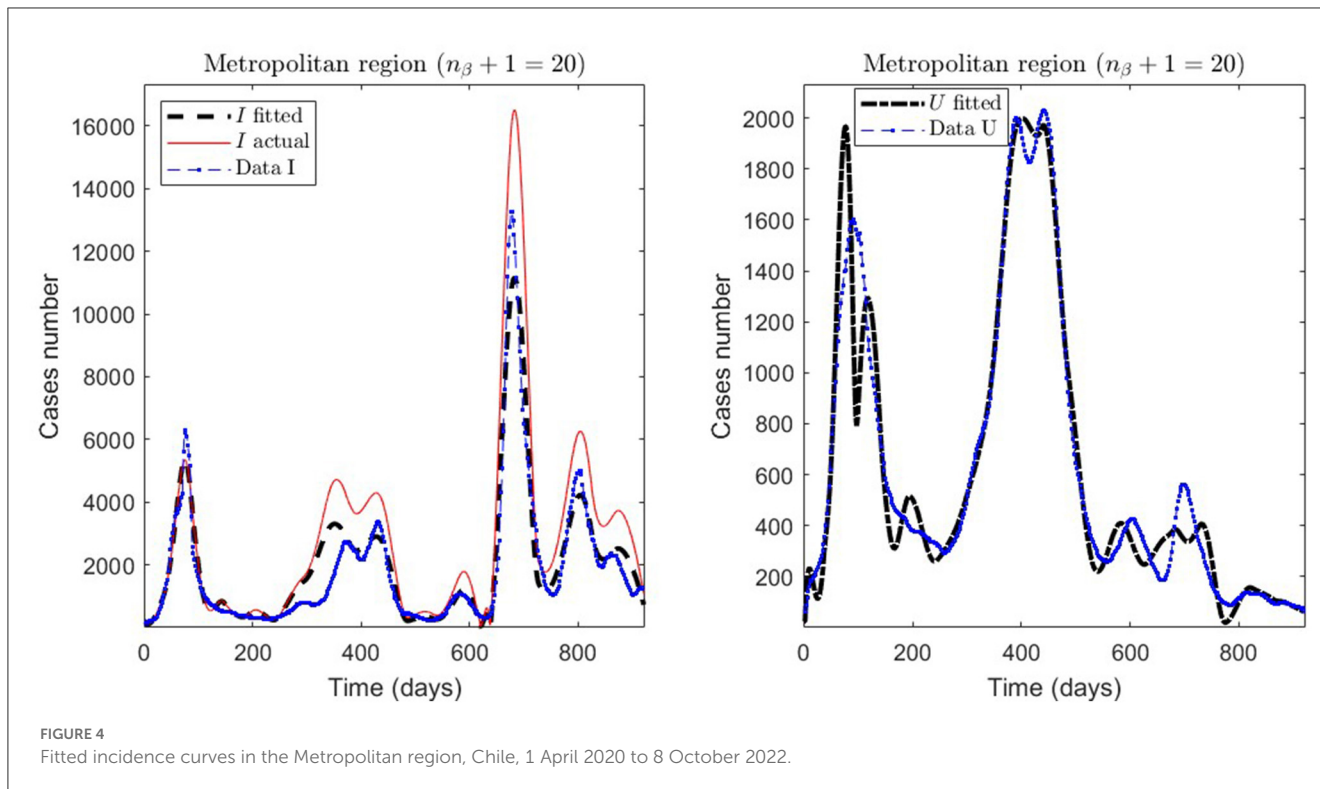
In Table 2, we present the optimization results for several datasets located throughout all of Chile: the Metropolitan region, Valparaíso region (center), Antofagasta region (northern), and Magallanes region (southern), designated by the identifiers MR, VAL, ANT, and MAG in the Id column. We plotted the corresponding curves in Figures 4–9. In addition, Table 2 shows the results concerning a study on the immunity duration in different time intervals to assess vaccination. We obtained the calibration results by varying the upper bound  $UB_{\tau_4}$  of the parameter  $\tau_4$  for a dataset that contains the MR’s first wave; see rows MR.2–MR.3 in Table 2 and Figures 5, 6. The interpretation was that the least mean relative error for  $I$  and  $U$ ,  $\overline{RE}_I$  and  $\overline{RE}_U$ , implied that the corresponding  $\tau_4$  value was the most probable for the respective dataset. This  $\tau_4$  variation is helpful for the analysis, simulation, and evaluation of epidemiological scenarios where, for example, a better fit for  $\tau_4$  larger means a high vaccine immunity duration.

From Table 2, for example, by decreasing the upper bound of  $\tau_4$ , comparing the results between MR.2 and MR.3

(through  $\overline{RE}_I$  and  $\overline{RE}_U$ ), we observe that  $\tau_4$  is small, which could be interpreted as correct, since in the MR’s first wave there was no vaccination and therefore no immunity due to it.

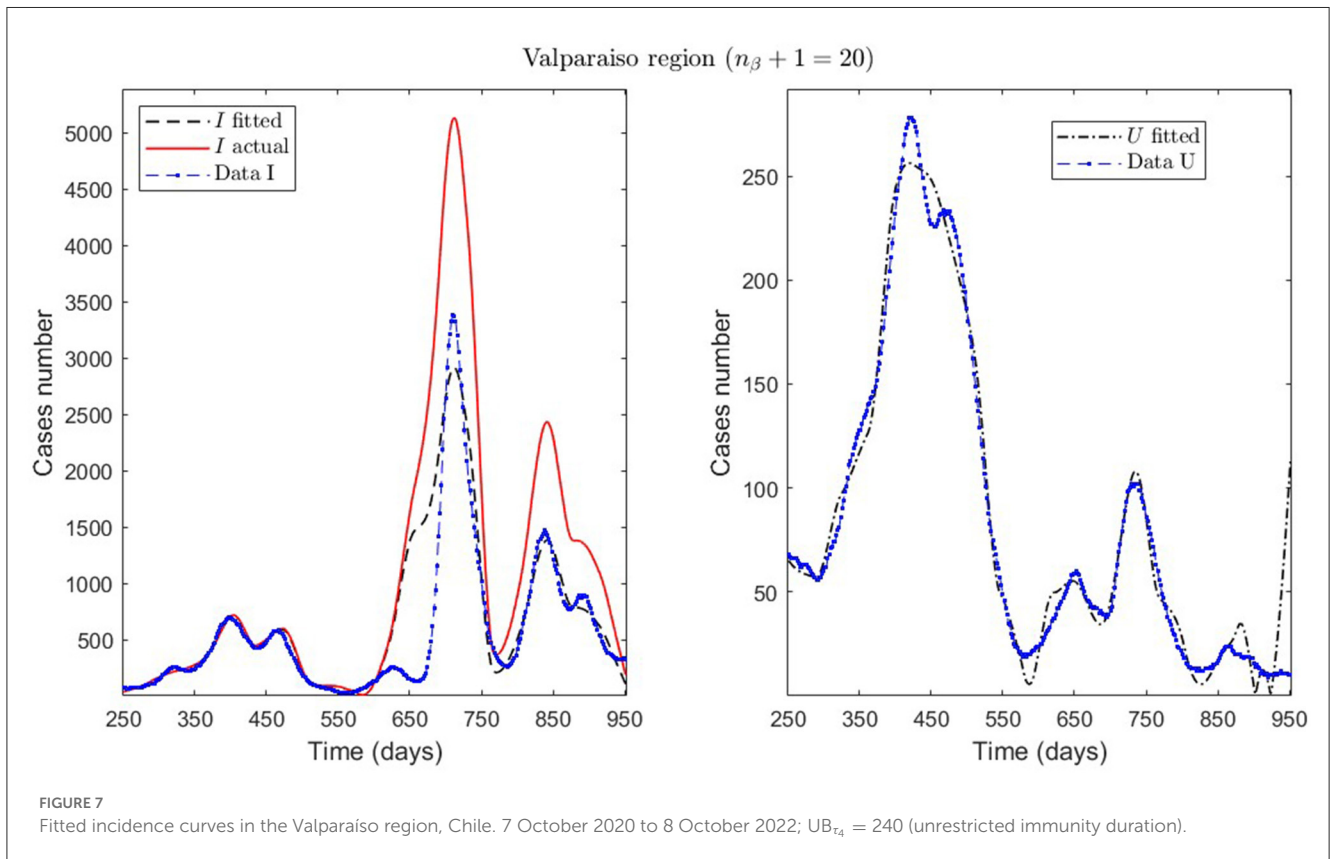
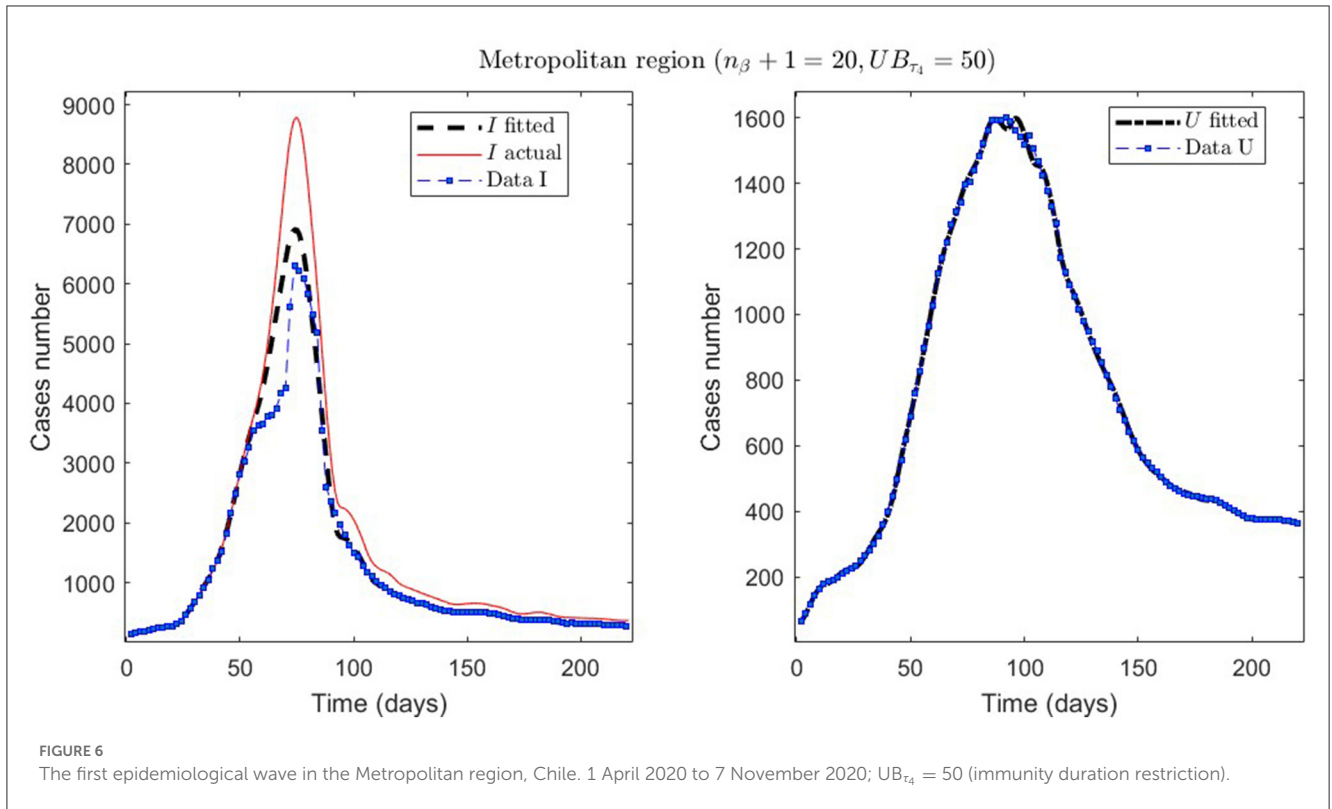
General epidemiological modeling with cumulative data allows us to fit the complete epidemiological curve for analytical purposes (MR.1 in Table 2) with slightly less precision. In some cases, it was necessary to increase the moving window size to handle abrupt changes of derivative in the curves or optimize with other fractions of the upper bounds (different values of  $\omega$ ). Although the error obtained may increase, these approximations (Val.1 in Table 2) should be considered since they may indicate a data problem or a change in the epidemiological scenario. The fact that these cases of higher error occur only for the data from those infected is another reason to focus more on the number of ICU patients plus deceased.

An example of the cases explained above is the Antofagasta region, the results from which are quantified with the ANT.1 identifier in the Id column in Table 2 and depicted in Figure 8A. We observed a wave not reported in the data for  $t \in [250, 950]$ . In addition, Figure 8C depicts that the model fits data for  $t \in [250, 600]$ , with low error for  $I$  and  $U$ , even if the actual infected (calculated by the model) was much higher than reported.

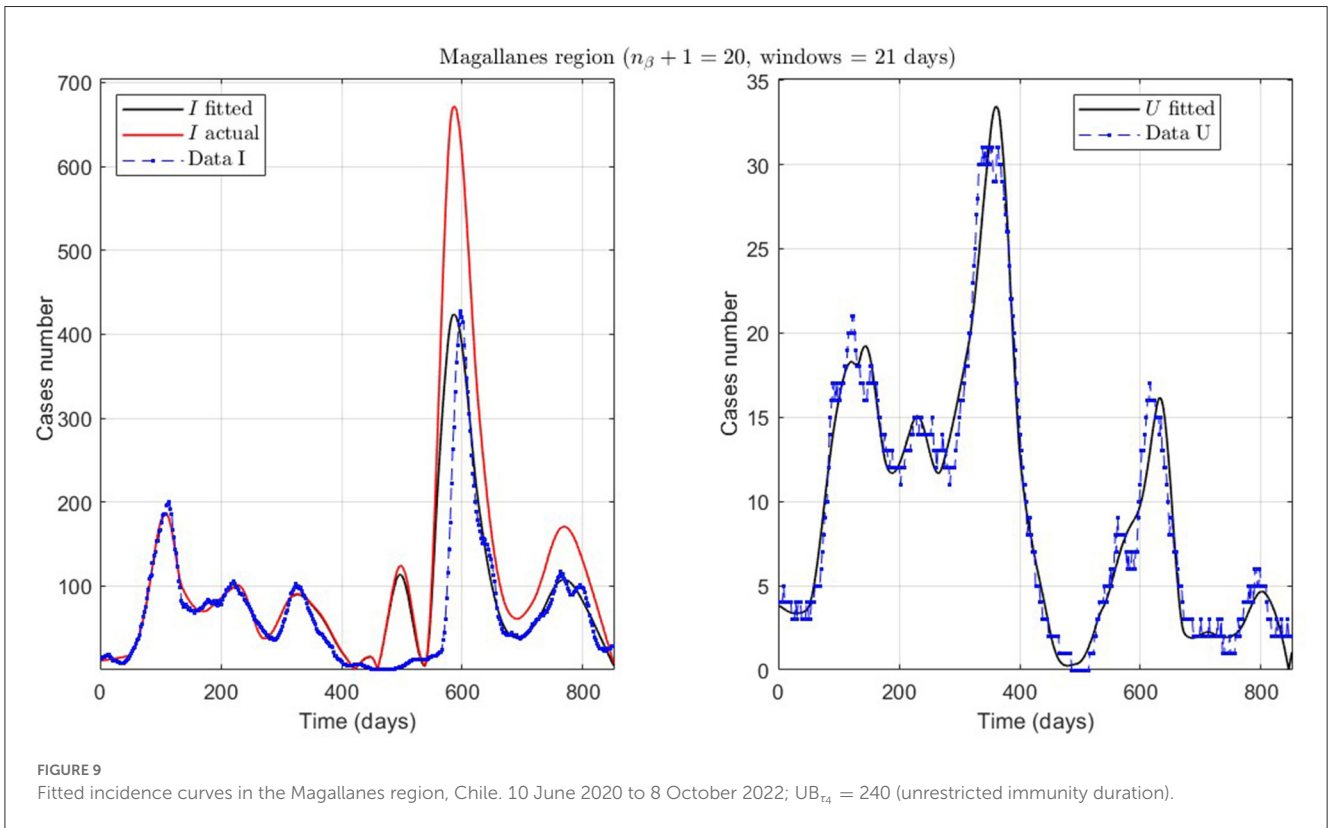
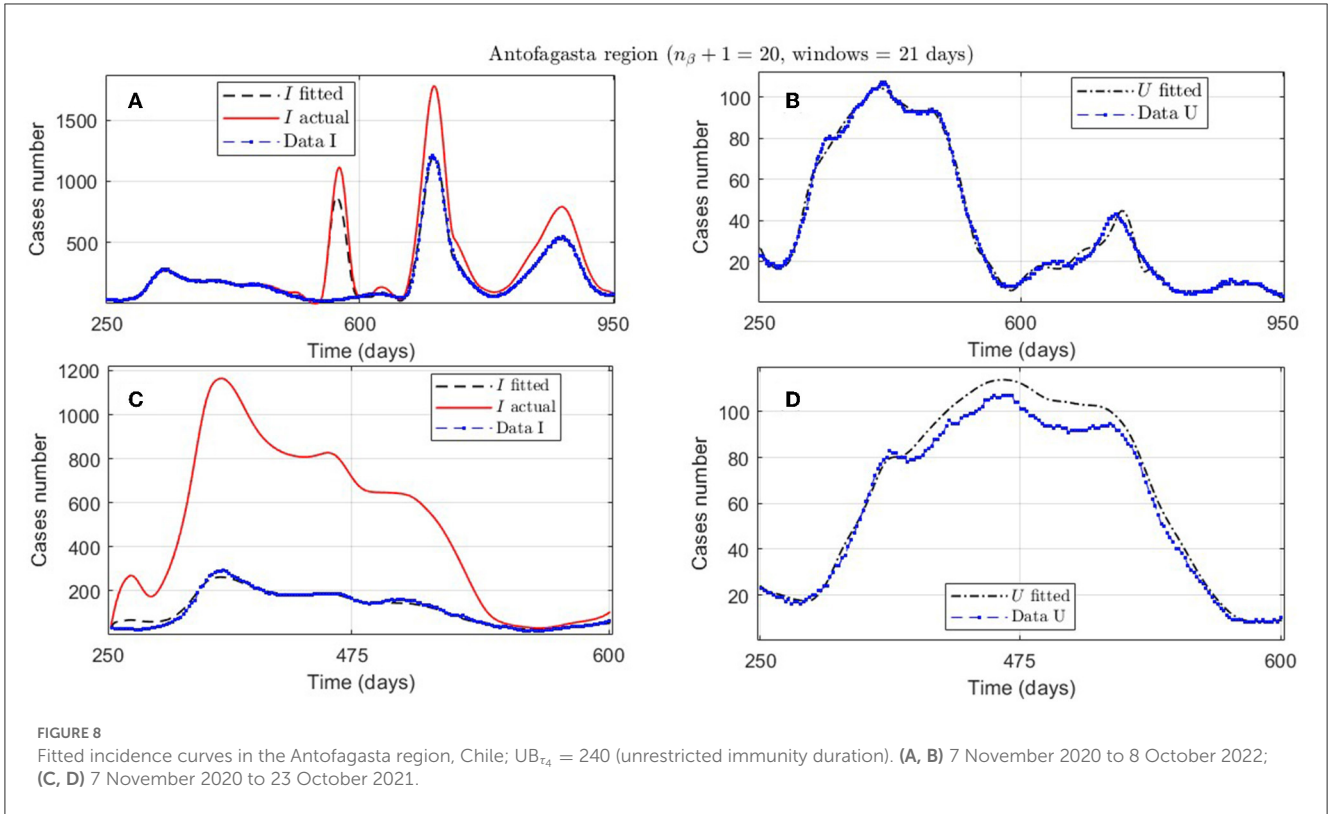


Finally, we present the case of the Magallanes region (MAG.1 in Table 2), a zone located in the extreme south of Chile with very few inhabitants and, therefore, little data. Despite

this, the model also fits curve  $U$  reasonably, despite a more significant error, where the relevance is to interpret the curve's trend.







### 3.2. Parametric bootstrap technique results

In this section, we want to show the predictive capabilities of our general modeling framework. For that, we applied the PBT for the 280–530 time interval in the Metropolitan region (MR), which encompasses the MR’s second wave, and the prediction period was 12 weeks.

First, we demonstrate the results obtained using the careful parameter optimization for fitting general epidemiological modeling to the actual data by minimizing the sum of squares (7) (step 1 of the algorithm described in Section 2.2.3). A code run with this method with 100 iterations took around 14 min in an Intel Core i7 processor with eight cores and 16 Gb of RAM.

We summarize the numerical results in Table 3 and depict the least-squares fitted model curves  $F(t_j, \hat{\theta})$  (Equation 9) in Figure 10. The corresponding error evaluated in the non-linear LSE  $\hat{\theta}$ , according to Equation (12), is  $E(\hat{\theta}) = 4.3387e - 01$ .

Second, we show the model results obtained through the PBT to achieve a better fitting and forecasting performance (the entire algorithm is described in Section 2.2.3). The PBT is computationally intensive since the code runs took around

24/16 h on one/two high-performance laptops, as described in Section 2.2.4.2. Using the two laptops, we calculated 50 parameter estimates to obtain the  $M = 100$  bootstrap parameter realizations for fitting general epidemiological modeling to every synthetic dataset generated by a normal distribution with moderate variance relative to the least-squares curves  $F(t_j, \hat{\theta})$  [Equation (9)], as explained in Section 2.2.3. Then, we constructed the best parameter estimate  $[\hat{\theta}_{\bar{k}}$  defined in (14)] and the PI  $[LB_{\hat{\theta}}$  and  $UB_{\hat{\theta}}$  defined in (11)]. With these estimates, we plotted the respective curves for infected persons  $I_r$  and the sum of ICU patients and those who were reported dead  $U_r$ .

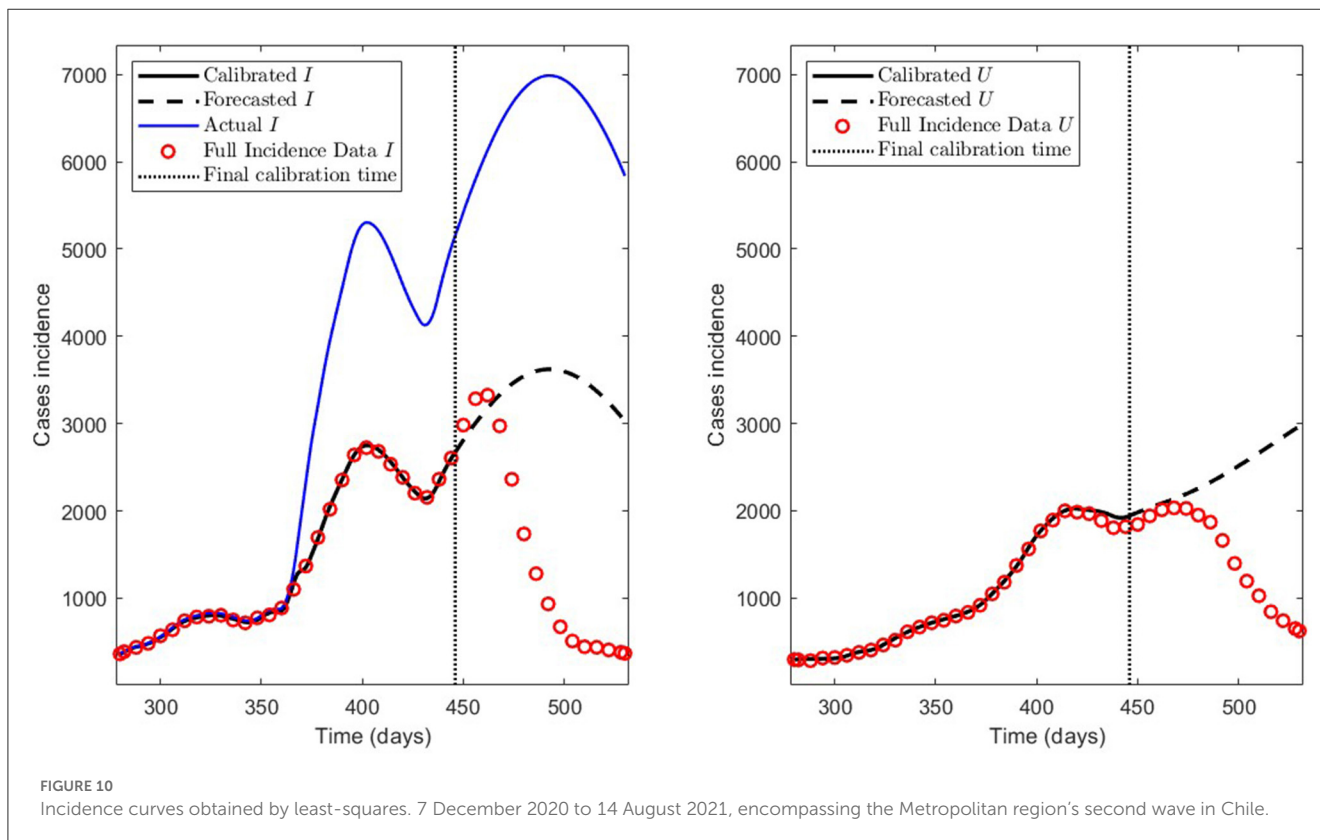
Figure 11 depicts the model results for synthetic datasets constructed as explained above. The minimum PBT error was  $E(\hat{\theta}_{\bar{k}}) = 2.8775e - 01$ , calculated according to Equation (14). Finally, the parameter estimates uncertainty, corresponding to the curves plotted in Figure 11, ranged between 0.02 and 18.92% with a mean of 4.61% measured in normalized standard error (NSE) in percent, as defined in Equation (11b).

### 3.3. Assessing the impact of vaccination

We compared different values of the parameter  $\tau_4$  in the range of 60–240 days to assess the impact of vaccination on immunity. For  $\tau_4$  values spaced every 30 days, and ranging between 60 and 240 days, we obtained the results summarized in Table 4, which shows the minimum PBT error  $E(\hat{\theta}_{\bar{k}})$  given in (14), the mean of the NSE ( $\bar{NSE}$ ) given in (11b), and the weighted error WE given in (15).

Table 3 Least-squares model results.

Criterion \ Variable	<i>I</i>	<i>U</i>
FP	6.3983e-01	4.3173e-01
RMSE	4.3539e-01	3.3227e-02



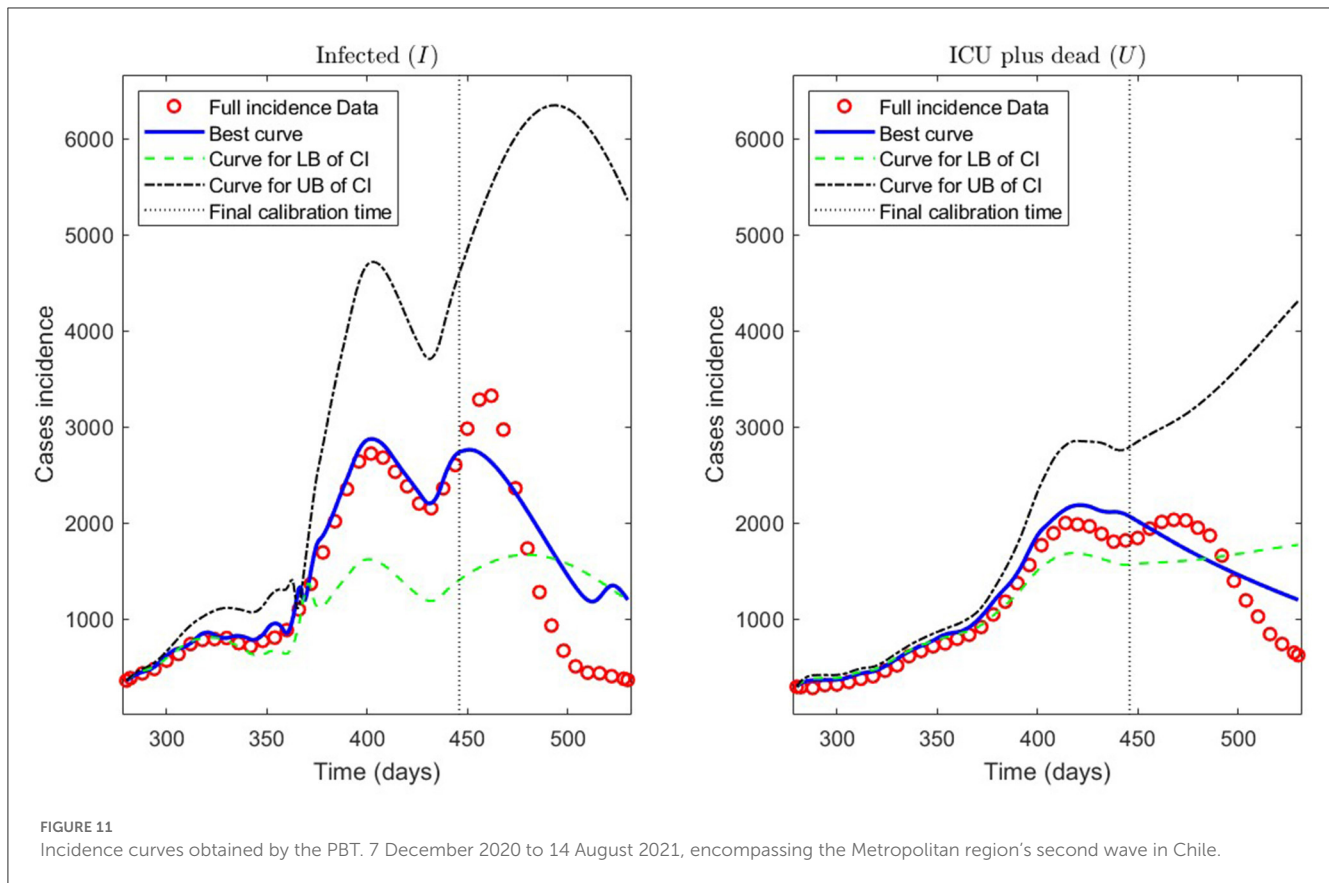


FIGURE 11 Incidence curves obtained by the PBT. 7 December 2020 to 14 August 2021, encompassing the Metropolitan region’s second wave in Chile.

From Table 4, we observe that the best value for  $\tau_4$  was  $\tau_4 = 60$  days since it yielded the least weighted error WE.

### 4. Discussion

Concerning quantitative tracking of COVID-19 dynamics, we can calibrate any dataset with our general modeling framework. In effect, we show the framework’s calibration capabilities through several examples for different regions of Chile; see Table 2 and the corresponding plotted curves in Figures 4–9. In addition, we corroborated that immunity duration was short ( $\tau_4 \leq 50$  days, as shown in row MR.3 in Table 2) during the Metropolitan region’s first wave (when there was no vaccination yet). Consequently, our general modeling framework provides a flexible tool for studying the dynamics of any transmissible disease and assessing vaccination, despite the vaccination deficiencies and data limitations discussed in Sections 2.1.2.1, 2.2.4.1. To do so, it suffices to have time-series observations for the number of infected persons and ICU patients and to find a suitable range of initial parameters meaningful from the epidemiology viewpoint. Then, the percentage decrease technique allows us to find a unique range of parameters proper for every studied dataset, which provides overall method numerical stability and convergence, as shown in this article.

Concerning the framework’s predictive capabilities, we applied the PBT to a dataset encompassing the Metropolitan region’s second wave. The results shown in Table 3 and depicted in

Table 4 Model results for different values of  $\tau_4$ .

$\tau_4$	$E(\hat{\theta}_{\tau_4})$	NSE	WE
60	2.7309e-01	4.9037e-02	1.9840e-01
90	2.8199e-01	4.9866e-02	2.0462e-01
120	2.9898e-01	5.3788e-02	2.1725e-01
150	2.9882e-01	6.2228e-02	2.1995e-01
180	2.8646e-01	5.4880e-02	2.0926e-01
210	3.4228e-01	5.1622e-02	2.4539e-01
240	3.5367e-01	6.2833e-02	2.5672e-01

$\tau_4$  is the mean immunity duration by vaccination; see Table 1.  $E(\hat{\theta}_{\tau_4})$  is the minimum error of the parametric bootstrap technique (PBT); see Equation (14). NSE is the mean normalized standard error (NSE); see Equation (11b), and WE is the weighted error; see Equation (15).

Figure 10, show that fitting the model to processed data by least-squares produces a forecasting performance for variables  $I$  and  $U$  that could be better (even more for  $U$ ), despite their respective fitting performance (RMSE) being excellent. From the PBT implementation, we found that the parameter uncertainty range, evaluated through the percentage NSE, needed to have reliable parameter estimates (between 0.02 and 18.92% with a mean of 4.61%), so the number of bootstrap samples  $M = 100$  was suitable. In addition, Figure 11 shows an outstanding forecasting performance for variables  $I$  and  $U$  (even more for  $U$ ) with an error of 1.5 times less than for the usual least-squares method ( $E(\hat{\theta}) \approx 1.5E(\hat{\theta}_{\tau_4})$ ). Therefore, our general epidemiological

framework can accurately predict the ICU curve trend in the medium term (12 weeks). However, a limitation to achieving such a good prediction performance is that the PBT is costly from a computational viewpoint.

In [Figure 11](#), we observe that the segmented curves (green lines) are more significant than the filled curves (blue lines) in specific time intervals, despite that the parameter vector of the firsts [evaluated at the lower bound of the PI; Equation (11a)] is smaller than the one of the seconds [evaluated at the best parameter vector; Equation (14)]. It may happen since, according to Equations (5b) and (5d), the curves for variables  $I$  and  $U$  will be significant if the  $\gamma_{IR}(t)$ ,  $\gamma_{IU}(t)$ , and  $\gamma_{UR}(t)$  rates are close to zero in some time intervals (which is the case of the segmented curves).

Concerning vaccination assessment relative to its immunity duration, from [Table 4](#), we infer that  $\tau_4 > 60$  values do not adapt to the dataset encompassing the MR's second wave. Therefore, the immunity duration should be less than 60 days. This exciting result implies a short immunity duration during the MR's second wave, which is not surprising since vaccination at that time was not yet widespread. In addition, this result would corroborate that the vaccination effect is not as significant as the immunity provided by natural infection, as discussed in Section 2.1.2.1. Indeed, the third wave magnitude (the most prolonged and steepest so far) shows that vaccination was relatively ineffective regarding protection against infection before it. However, one may think that vaccination manifests a positive effect by the time of the third wave, which we can infer from the fact that the ICU patient and death data are low compared to the infected data during that wave.

Our results rely on the piecewise linear reconstruction of time-varying parameters  $\gamma_{IR}$ ,  $\gamma_{SR}$ ,  $\gamma_{RS}$ ,  $\gamma_{IU}$ ,  $\beta$ , and  $\gamma_{UR}$ . We could improve this arbitrary choice by assuming the mechanical laws of the transmission rate ( $\beta$ ) or other rates as, for example, in (21).

## 5. Conclusions

This work attempts to implement a holistic and general modeling framework for quantitative tracking of the dynamics of any transmissible disease, focusing on accurately predicting the ICU curve trend in the medium term and assessing vaccination.

Implementing a careful optimization algorithm, we obtained outstanding results concerning quantitative tracking of COVID-19 dynamics for several processed datasets representing different regions of Chile and assessing vaccination. In addition, a parametric bootstrap technique allowed us to predict the ICU curve trend in the medium term accurately and assess vaccination. As a result, the scientific community could adapt our general modeling framework to evaluate the impact of combined vaccination and NPI strategies for COVID-19 or any transmissible disease in any country and help visualize the potential effects of implemented plans by policymakers.

In conclusion, the two main lessons are that we must be prepared to face another pandemic like COVID-19 and that it is more important to make more effort to follow up the ICU patients, which are highly correlated with dead confirmed by COVID-19.

To tackle the first lesson, in future work, we want to improve the computational cost of the parametric bootstrap technique or use another technique more efficiently. The aim would be to reconstruct epidemiological curves to predict the combined NPIs and vaccination policies' impact on the ICU curve trend in real time, providing scientific evidence to help anticipate policymakers' decisions.

## Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: <https://github.com/MinCiencia/Datos-COVID19>.

## Author contributions

PC contributed to conceptualization, formal analysis, investigation, methodology, project administration, software, supervision, validation, writing—original draft, review, and editing. OR-D contributed to conceptualization, formal analysis, investigation, methodology, software, validation, and writing—review and editing. CC contributed to theoretical analysis, supervision, and writing—review and editing. All authors contributed to the article and approved the submitted version.

## Funding

The Centre for Biotechnology and Bioengineering (CeBiB), Grant Number ANID PFBasal-01, partially supported PC and CC. In addition, PC was partially supported by the regular research project DIUBB 2120432 IF/R. Finally, CC was partially supported by the ACE210010 and CMM PIA AFB170001 projects. The funders had no role in study design, data collection, analysis, publication decision, or manuscript preparation.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.



## References

- Al-Tuwairqi SM, Al-Harbi SK. A time-delayed model for the spread of COVID-19 with vaccination. *Sci Rep.* (2022) 12:19435. doi: 10.1038/s41598-022-23822-5
- Zhenzhen L, Yongguang Y, YangQuan C, Guojian R, Conghui X, Shuhui W. Stability analysis of a nonlocal SIHRDP epidemic model with memory effects. *Nonlinear Dyn.* (2022) 109:121–41. doi: 10.1007/s11071-022-07286-w
- Ghosh S, Volpert V, Banerjee M. An epidemic model with time delay determined by the disease duration. *Mathematics.* (2022) 10:2561. doi: 10.3390/math10152561
- Zhai S, Luo G, Huang T, Wang X, Tao J, Zhou P. Vaccination control of an epidemic model with time delay and its application to COVID-19. *Nonlinear Dyn.* (2021) 106:1279–92. doi: 10.1007/s11071-021-06533-w
- Cumsille P, Oscar Rojas-Díaz, de Espanas PM, Verdugo-Hernández P. Forecasting COVID-19 Chile' second outbreak by a generalized SIR model with constant time delays and a fitted positivity rate. *Math Comput Simul.* (2022) 193:1–18. doi: 10.1016/j.matcom.2021.09.016
- Cumsille P, Coronel A, Conca C, Quiñinao C, Escudero C. Proposal of a hybrid approach for tumor progression and tumor-induced angiogenesis. *Theoret Biol Med Modell.* (2015) 12:13. doi: 10.1186/s12976-015-0009-y
- Badillo G, Cumsille P, Segura-Ponce L, Pataro G, Ferrari G. An efficient optimization methodology of respiration rate parameters coupled with transport properties in mass balances to describe modified atmosphere packaging systems. *Inverse Prob Sci Eng.* (2020) 28:1361–83. doi: 10.1080/17415977.2020.1717488
- Cumsille P, Godoy M, Gerdtzen ZP, Conca C. Parameter estimation and mathematical modeling for the quantitative description of therapy failure due to drug resistance in gastrointestinal stromal tumor metastasis to the liver. *PLoS ONE.* (2019) 14:e0217332. doi: 10.1371/journal.pone.0217332
- Chowell G. Fitting dynamic models to epidemic outbreaks with quantified uncertainty: a primer for parameter uncertainty, identifiability, and forecasts. *Infect Dis Modell.* (2017) 2:379–98. doi: 10.1016/j.idm.2017.08.001
- Cornejo S, Arancibia D, Molina F, Ministerio de Ciencia, Tecnología, Conocimiento e Innovación, Frías IA. *Official Chilean Repository of COVID19 Data.* (2022). Available online at: <https://github.com/MinCiencia/Datos-COVID19>
- Miao H, Xia X, Perelson AS, Wu H. On Identifiability of nonlinear ODE models and applications in viral dynamics. *SIAM Rev.* (2011) 53:3–39. doi: 10.1137/090757009
- Müller S, Lu J, Kügler P, Engl H. *Parameter Identification in Systems Biology: Solving ill-Posed Inverse Problems Using Regularization.* Johann Radon Institute for Computational and Applied Mathematics (RICAM); Austrian Academy of Sciences (ÖAW) (2008). Available online at: <https://www.ricam.oeaw.ac.at/files/reports/08/rep08-25.pdf>
- Moore S, Hill EM, Tildesley MJ, Dyson L, Keeling MJ. Vaccination and non-pharmaceutical interventions for COVID-19: a mathematical modelling study. *Lancet Infect Dis.* (2021) 21:793–802. doi: 10.1016/S1473-3099(21)00143-2
- Pérez-Alós L, Armenteros JJA, Madsen JR, Hansen CB, Jarlhelt I, Hamm SR, et al. Modeling of waning immunity after SARS-CoV-2 vaccination and influencing factors. *Nat Commun.* (2022) 13:1614. doi: 10.1038/s41467-022-29225-4
- GSM-covid19CL. *GitHub Repository Code Paper* (2022). Available online at: <https://github.com/neurovisionhub/GSM-covid19CL>
- Banks HT, Davidian M, Samuels JR, Sutton KL. An inverse problem statistical methodology summary. In: Chowell G, Hyman JM, Bettencourt LMA, Castillo-Chavez C, editors. *Mathematical and Statistical Estimation Approaches in Epidemiology.* Dordrecht: Springer (2009). p. 249–302. doi: 10.1007/978-90-481-2313-1\_11
- Shampine LF, Thompson S. Solving DDEs in Matlab. *Appl Num Math.* (2001) 37:441–58. doi: 10.1016/s0168-9274(00)00055-6
- WHO. *Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19)* (2019). Available online at: [https://www.who.int/publications/i/item/report-of-the-who-china-joint-mission-on-coronavirus-disease-2019-\(covid-19\)](https://www.who.int/publications/i/item/report-of-the-who-china-joint-mission-on-coronavirus-disease-2019-(covid-19))
- Dan JM, Mateus J, Kato Y, Hastie KM, Yu ED, Faliti CE, et al. Immunological memory to SARS-CoV-2 assessed for up to 8 months after infection. *Science.* (2021) 371:eabf4063. doi: 10.1126/science.aabf4063
- Efron B, Tibshirani RJ. *An Introduction to the Bootstrap.* 1st ed. New York, NY: Chapman and Hall/CRC (1994). doi: 10.1201/9780429246593
- Córdova-Lepe F, Vogt-Geisse K. Adding a reaction-restoration type transmission rate dynamic-law to the basic SEIR COVID-19 model. *PLoS ONE.* (2022) 17:e0269843. doi: 10.1371/journal.pone.0269843