

Research



Cite this article: Albani VVL, Albani RAS, Massad E, Zubelli JP. 2022 Nowcasting and forecasting COVID-19 waves: the recursive and stochastic nature of transmission. *R. Soc. Open Sci.* **9**: 220489.
<https://doi.org/10.1098/rsos.220489>

Received: 14 April 2022

Accepted: 2 August 2022

Subject Category:

Mathematics

Subject Areas:

biomathematics/applied mathematics

Keywords:

epidemiological models, nowcasting, forecasting, COVID-19, model calibration

Author for correspondence:

J. P. Zubelli

e-mail: jorge.zubelli@ku.ac.ae

Nowcasting and forecasting COVID-19 waves: the recursive and stochastic nature of transmission

V. V. L. Albani¹, R. A. S. Albani², E. Massad^{3,4} and J. P. Zubelli⁵

¹Department of Mathematics, Federal University of Santa Catarina, Florianópolis, Brazil

²Instituto Politécnico do Rio de Janeiro, Rio de Janeiro State University, Nova Friburgo, Brazil

³School of Medicine, University of São Paulo and LIM01-HCFMUSP, São Paulo, Brazil

⁴School of Applied Mathematics, Fundação Getúlio Vargas, Rio de Janeiro, Brazil

⁵Mathematics Department, Khalifa University, Abu Dhabi, UAE

VVLA, 0000-0002-5751-5895; RASA, 0000-0002-2091-194X; EM, 0000-0002-7200-2916; JPZ, 0000-0003-1581-9505

We propose a parsimonious, yet effective, susceptible–exposed–infected–removed-type model that incorporates the time change in the transmission and death rates. The model is calibrated by Tikhonov-type regularization from official reports from New York City (NYC), Chicago, the State of São Paulo, in Brazil and British Columbia, in Canada. To forecast, we propose different ways to extend the transmission parameter, considering its estimated values. The forecast accuracy is then evaluated using real data from the above referred places. All the techniques accurately provided forecast scenarios for periods 15 days long. One of the models effectively predicted the magnitude of the four waves of infections in NYC, including the one caused by the Omicron variant for periods of 45 days using out-of-sample data.

1. Introduction

Since the seminal work of Kermack & McKendrick [1], susceptible–infected–removed-type and susceptible–exposed–infected–removed-type (SEIR-type) models have played a crucial role in understanding the dynamics of infectious diseases through nowcasting and helping public health authorities to address disease outbreaks by forecasting [2]. In the pandemic of COVID-19, that started in the end of 2019 and was still ongoing in the beginning of 2022, SEIR-type models were again recurrently used to nowcast the SARS-CoV-2 spread dynamics and to forecasting [3–12].

Two of the main challenges the COVID-19 pandemic posed in the modelling of infectious diseases were to provide accurate nowcast and forecasts [9,13,14]. Nowcasting is fundamental to understanding the seriousness of the present situation based on the observed reports of infections and deaths, as well as the hospital capacity. Forecasting is the main tool used to inform how much intervention will be needed in the future to control or contain the pathogen spread. Thus, to address both tasks, it is necessary to access good quality data and to design accurate models that incorporate the time evolution of the transmission pattern as well as the disease severity. Moreover, it is necessary to use model calibration techniques that incorporate the uncertainties in data acquisition.

The way a pathogen spreads changes with time since it can be affected, for example, by the adoption of spread control measures and changes in the pathogen (e.g. mutation). However, the way such changes in the pathogen occur is in general unknown and hard to predict. In other words, transmission can be defined as a stochastic process. Thus, to design trustworthy and accurate scenarios, a model must incorporate such variability and randomness. Moreover, based on the past information, the model must be able to predict the future pattern of transmission, presenting realistic potential numbers of infections and deaths. A model that fails to address the latter appropriately can, for example, give arguments to make the public opinion skeptical about or even to deny the severity of an emerging disease. This seemed to be an important issue in the beginning of the COVID-19 pandemic when a series of predictions made with over-simplistic models and calibrated with poor data provided unrealistic catastrophic scenarios [13,15]. During secondary outbreaks, the implementation of non-pharmaceutical contention measures became more difficult and anti-lockdown protests emerged worldwide [16–18].

Due to these issues in model predictions, forecast accuracy was recurrently discussed during the COVID-19 pandemic. For example, Ioannidis *et al.* [13] recalls a series of unrealistic scenarios designed in the beginning of the pandemic and proposed some ways to improve model forecast. Gnanvi *et al.* [19] made a systematic review of models used to predict scenarios of infections and deaths related to COVID-19. They classified their results by comparing predictions with reports and found that reports were inside the 95% confidence intervals (CIs) of 75% of the considered models. Other works, proposed to refine the modelling of the dynamics of COVID-19 by incorporating, for example, the spatial distribution of infections, mobility information and different levels of disease severity depending on age and sex [3,10,20–23].

To address the issues raised above, the present article proposes a methodology to appropriately model, calibrate and forecast scenarios using SEIR-type models. More precisely, we show that even simpler SEIR-type models with appropriately calibrated time-dependent transmission parameters are still suitable to forecast scenarios accurately. Moreover, we compare the results of a series of techniques used to define the future pattern of transmission based on past information. All the results are based on real data from Chicago, New York City (NYC), the Canadian province of British Columbia and the Brazilian state of São Paulo. Some of the techniques were able to predict accurately four large outbreaks in NYC, including those caused by the Delta and Omicron variants of the SARS-CoV-2 using out-of-sample data. Another implication of the numerical results is that considering the pathogen transmission as a stochastic process can lead to more accurate scenarios.

2. Methods

As pointed in a series of recent works [3,5,24,25], the estimated values of the transmission parameter usually denoted by β change with time and presents some sort of randomness. To make accurate predictions, such variability must be incorporated into the model. Thus, the techniques proposed in what follows are designed to accurately predict the future pattern of transmission by accounting for past calibrated values of the time-dependent β . It is worth mentioning that using time-dependent transmission parameters also allows a remarkable adherence to reported data [3,5,24,25].

2.1. Describing the disease spread dynamics

To describe the spread dynamics of the virus SARS-CoV-2, we consider a SEIR-type model that, for simplicity, does not consider natural deaths and births. The model accounts for the following five

compartments, susceptible (S), exposed (E), infected (I), recovered (R) and deceased (D). The progress between the compartments is defined by the dynamical system of ordinary differential equations below [26]:

$$\frac{dS}{dt} = -\beta(t)SI, \quad (2.1)$$

$$\frac{dE}{dt} = \beta(t)SI - \sigma E, \quad (2.2)$$

$$\frac{dI}{dt} = \sigma E - (\gamma(t) + \delta(t))I, \quad (2.3)$$

$$\frac{dR}{dt} = \gamma(t)I \quad (2.4)$$

and
$$\frac{dD}{dt} = \delta(t)I. \quad (2.5)$$

The time-dependent transmission parameter is denoted by $\beta(t)$ and the infection to onset mean time is represented by $1/\sigma$, which is set to 5.1 days [27]. The quantity γ denotes the recovery rate and $\delta(t)$ is the death rate. Following Grasselli *et al.* [28], we set $\gamma = 1/14 \text{ days}^{-1}$. For each day t , the death rate $\delta(t)$ is defined as follows, it is the total number of registered deaths on day t divided by the number of reported infections on the day $t - 12$. So it is dimensionless. The 12 days is the mean time that someone takes from being infected to die by COVID-19 [3,24,25].

Natural death and birth could be included in the model, but such terms would make little difference in the analysis that follows, as the time interval considered here is small. So, to keep the model as simple as possible, we decided to not include such terms.

2.2. Model calibration

The estimation of the transmission parameter β from daily reports of COVID-19 infections is performed by the minimization of the Tikhonov-type functional below for each day t , as in [3]:

$$\mathcal{F}(\beta(t)) = I_{\text{REP}}(t) \log(\Omega(t)) - \Omega(t) - \log(I_{\text{REP}}(t)!) + \alpha(\beta(t) - \beta(t-1))^2, \quad (2.6)$$

where α is the regularization parameter, which has dimension days^2 , to make the second term in the right-hand side of equation (2.6) dimensionless, and $I_{\text{REP}}(t)$ denotes the reported COVID-19 infections on the day t . The term $\log(I_{\text{REP}}(t)!)$ is approximated by the Stirling's formula:

$$\log(I_{\text{REP}}(t)!) \approx \frac{1}{2} \log(2\pi I_{\text{REP}}(t)) + I_{\text{REP}}(t) \log(I_{\text{REP}}(t)) - I_{\text{REP}}(t).$$

The quantity $\Omega(t) = \int_{t-1}^t \sigma E(s) ds$ in equation (2.6) represents the model predicted infections for the day t . It depends on β , since it is part of the solution of the system of equations in equations (2.1)–(2.5). Therefore, the $\beta(t)$ that minimizes equation (2.6) is the value that makes the model to predict correctly the infections reported in the day t . The first part of the functional in equation (2.6) accounts for the discrepancy between the model predictions and the reports of infections. The second part penalizes the first to avoid overfitting, by imposing that the discrepancy between $\beta(t)$ and the already known value $\beta(t-1)$ must also be minimized. To avoid the introduction of bias in the calibration process, the parameter α must be appropriately chosen. In the present set of examples, we set $\alpha = 10^{-3}/2$ accordingly to an *a posteriori* rule [29].

2.3. Model forecast techniques

The transmission parameter β , as we shall see, changes with time, presenting an uncertain pattern. Since we do not have any formulation to predict how β changes with time, we propose five different techniques to evaluate future values of β allowing the SEIR-type model to make predictions.

2.3.1. Average method (AM)

The simplest technique to extend β by setting its future values as the average of its estimated values. More precisely, for any $t > t_n$, set $\beta(t) = \bar{\beta}$, where

$$\bar{\beta} = \frac{\beta(t_n) + \beta(t_{n-1}) + \cdots + \beta(t_{n-m+1})}{m},$$

$t_{n-m+1}, \dots, t_n$ are time values where β is calibrated, and $m < n$.

Since the trend of β at the end of the dataset is the most relevant information to predict the number of infections in the following days, we set $m = 5$ days. If m is large, it is possible to corrupt predictions with outdated information. Moreover, keeping the number of unknowns considerably smaller than the size of 'observed' quantities can help to reduce the effect of noise in the predictions.

2.3.2. Linear regression (LR)

Another simple technique to extend β is to fit a linear model. In other words, for any $t > t_n$, we set $\beta(t) = a \cdot t + b$, with the scalar parameters a and b fitted to the set of calibrated values of $\beta \{\beta(t_{n-m+1}), \beta(t_{n-m+2}), \dots, \beta(t_n)\}$. The calibration is performed by least squares.

In this case, since there are two unknowns, we set $m = 10$ days also to avoid outdated information and to keep the number of unknowns considerably smaller than the number of observations.

2.3.3. Truncated Fourier series

A more elaborated technique is to calibrate a truncated Fourier series (TFS) [30] that approximates a dataset. More precisely, assume that β can be approximated by a TFS in the time interval $[t_{n-m+1}, t_n]$,

$$\beta(t) \approx a_0 + \sum_{j=1}^N \left[a_j \cos\left(\frac{2\pi j t}{m}\right) + b_j \sin\left(\frac{2\pi j t}{m}\right) \right]. \quad (2.7)$$

For m and N fixed, the parameters a_0 , a_j and b_j ($j = 1, \dots, N$), which have the same dimension of β , are calibrated by least squares from the dataset $\{\beta(t_{n-m+1}), \beta(t_{n-m+2}), \dots, \beta(t_n)\}$. Thus, for each time $t > t_n$, $\beta(t)$ is given by the periodic function defined in equation (2.7) with calibrated parameters.

In this case, choosing the length of the time interval where the series is defined and the number of terms in the series is tricky. After testing different values and configurations, we found that setting the length of the time interval $m = 30$ days and the number of terms in the series $N = 3$ gives a computational affordable technique that fits well the data.

2.3.4. Truncated Fourier series with noise (TFSWN)

In order to add some uncertainty to the predicted values of β , we add to the TFS described above a Gaussian noise with mean zero and standard deviation given as follows:

- (i) Evaluate the difference between the values given by the TFS and the dataset $\{\beta(t_{n-m+1}), \beta(t_{n-m+2}), \dots, \beta(t_n)\}$.
- (ii) Evaluate the empirical standard deviation of the difference.

For each time $t > t_n$, $\beta(t)$ is given by the function in equation (2.7) plus a Gaussian random variable with mean zero and standard deviation given as above. In this case, we use the same values for m and N , as well as the parameters of the TFS inequation (2.7) found for the noiseless case.

2.3.5. A mean-reverting model for the transmission parameter

As an alternative to the methods described above, we propose a mean-reverting (MR) model for the dynamics of β . Since the time evolution of β seems to return to mean values, after random changes, it is natural to test such a model. So, inspired by the volatility of the Heston model [31] from Quantitative Finance, we assume that, for $t > t_n$, the values of $\beta(t)$ are given by the following stochastic process

$$\tilde{\beta}(t) = \beta(0) + \kappa \int_0^t (\theta - \tilde{\beta}(s)) ds + \xi \int_0^t \sqrt{\tilde{\beta}(s)} dW(s), \quad (2.8)$$

where $\{W(s) : s > 0\}$ represents the Wiener process also known as Brownian motion. Thus, set $\tilde{\beta}(0) = \beta(t_n)$, and $\beta(t) = \tilde{\beta}(t - t_n)$.

Based on the Euler–Maruyama scheme to solve the stochastic differential equation in equation (2.8), we have,

$$\tilde{\beta}_{j+1} = \tilde{\beta}_j + \kappa(\theta - \tilde{\beta}_j)\Delta t + \xi\sqrt{\tilde{\beta}_j}\varepsilon, \quad (2.9)$$

where ε has Gaussian distribution with mean zero and variance Δt .

Since the estimated values for β are given daily, we set $\Delta t = 1/365$. The parameters κ , θ and ξ are obtained from the set of estimated values for β , $\{\beta(t_{n-m+1}), \beta(t_{n-m+2}), \dots, \beta(t_n)\}$, by minimizing the functional

$$f(\kappa, \theta, \xi) = \sum_{j=1}^m [\beta(t_{n-j+1})^2 - 2\beta(t_{n-j+1})\mathbb{E}(\tilde{\beta}_j) + \mathbb{E}(\tilde{\beta}_j^2)], \quad (2.10)$$

with respect to the parameters κ , θ , ξ , where $\tilde{\beta}_j$ ($j = 1, \dots, m$) is a function of κ , θ , ξ . It is worth noticing that, in the calibration, $\tilde{\beta}_0 = \beta(t_{n-m})$.

The quantities $\mathbb{E}(\tilde{\beta}_j)$ and $\mathbb{E}(\tilde{\beta}_j^2)$ are deterministic and can be easily evaluated by the following recursive formulae:

$$\mathbb{E}(\tilde{\beta}_j) = \mathbb{E}(\tilde{\beta}_{j-1}) + \kappa[\theta - \mathbb{E}(\tilde{\beta}_{j-1})]\Delta t,$$

and

$$\begin{aligned} \mathbb{E}(\tilde{\beta}_j^2) &= \mathbb{E}(\tilde{\beta}_{j-1}^2) + \kappa^2[\theta^2 - 2\theta\mathbb{E}(\tilde{\beta}_{j-1}) + \mathbb{E}(\tilde{\beta}_{j-1}^2)]\Delta t^2 \\ &\quad + \xi^2\mathbb{E}(\tilde{\beta}_{j-1}^2)\Delta t + 2\kappa[\theta\mathbb{E}(\tilde{\beta}_{j-1}) - \mathbb{E}(\tilde{\beta}_{j-1}^2)]\Delta t. \end{aligned}$$

In the results, we shall use $m = 30$ days, which keeps the number of unknowns much smaller than the number of observations, and, in principle, allows the model to incorporate the most recent data.

3. Results

Next, we perform a series of numerical tests to compare the accuracy of the predicted accumulated numbers of infections obtained using the techniques described in §2.

Firstly, the SEIR-type model in equations (2.1)–(2.5) is calibrated using 7-day moving average time-series of the daily reports of COVID-19 infections and deaths from Chicago and New York City (NYC) in the USA, the Canadian province of British Columbia (BC) and the Brazilian state of São Paulo (SP). The datasets start mainly at the beginning of 2020 and end at the end of 2021. To stress-test the proposed methodology, we have chosen to use data from different places since there may be differences in the protocols of data collection and testing. More precisely, we seek the prediction performance from different data sources and sites with various characteristics, such as population size and other demographic factors. In addition, the data from these locations are publicly available from stable databases that were periodically updated and corrected up to the time of the writing.

We consider predictions of infections for 10, 15, 30, 45, 60 and 90 days after the period used in the calibration. Predictions are made as follows: we calibrate the model until some date t_n using the reported infections until the date t_n , then, for any $t_j > t_n$ we evaluate the model predictions of infections using the five techniques proposed in §2. These predictions are updated daily until the end of the dataset. We compare the reported and predicted accumulated numbers of infections, evaluating the normalized error in the out-of-sample predictions as follows, denote by I_{REP} the accumulated number of reported infections during the time interval of forecast and by I_{PRED} the accumulated predicted infections for the same period, then,

$$\text{ERROR} = \frac{|I_{\text{REP}} - I_{\text{PRED}}|}{I_{\text{REP}}}. \quad (3.1)$$

The error is evaluated for the daily out-of-sample predictions and classified by the number of days ahead and the technique used to generate the prediction. By out-of-sample we mean that the model calibration was performed with a dataset different from the one used to compare the model predictions. The daily

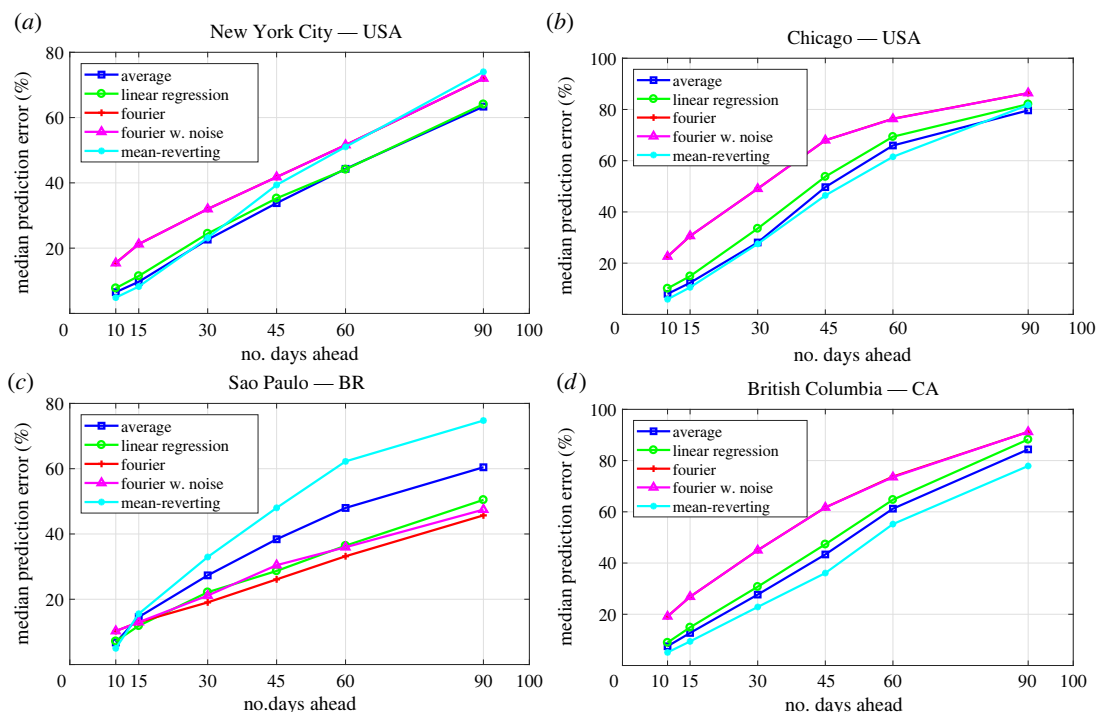


Figure 1. Median values of the percentage error in the out-of-sample predicted accumulated numbers of infections. Predictions are based on calibrated data and evaluated from 10 to 90 days ahead using the methods of extending the transmission parameter β described in §2.3. The reports are different from those used in the calibration.

errors are then used to evaluate median values and 70% CIs. The comparison between the errors' median values can be found in figure 1. Table 1 presents the errors' median values and their 70% CI.

To evaluate the errors of the stochastic techniques, namely TFSWN and MR, we used the median values of 5000 samples of predicted infections.

As figure 1 and table 1 show, all the methods presented similar error values for predicting 10–15 days ahead. Major differences started to appear from 30 to 90 days ahead predictions.

Predictions made with the AM presented small median error values with moderate 70% CI even for prediction with larger time intervals. For NYC, Chicago and BC, it presented one of the smallest median error values for all prediction time intervals. For SP, this technique also presented accurate results, but for longer prediction time intervals, other techniques performed considerably better. Although it is the simplest technique considered in this work, its performance is remarkably accurate.

The LR presented results similar to the AM technique. However, its errors were slightly larger than those for the AM using the datasets from NYC, Chicago and BC. For SP, it presented a better performance than the AM. It is worth mentioning that such a technique is also simple to implement and a nice alternative to the AM.

The TFS and TFSWN presented similar results since their graphs matched for almost all places. They both performed better for SP, but all the other techniques outperformed both. The main reason for such results must be that TFS fits the data quite well, which is useful for adherence to data, but may not be the case for model predictions. In other words, it seems that these techniques tend to overfit the data. However, these techniques seems to perform better for longer periods, since, in some sense, they capture better the log-term behaviour of the transmission parameter.

For Chicago and BC, the MR outperformed all the other techniques. For NYC it presented one of the best results for predictions until 30 days. For SP, it outperformed the other techniques only for predictions until 10 days. In general, for longer periods, like 60–90 days, the predictions 70% CIs were wider than those obtained with other techniques. Although these results may indicate that such a technique can be less accurate for larger prediction times, it was effective in predicting the magnitude of new waves of infections for NYC.

In the second example, we focus on the forecast of large outbreaks that happened in New York City during 2020 and 2021, including those caused by the emergence of the Delta (third wave) and Omicron (fourth wave) variants of the SARS-CoV-2. The four periods considered in this example

Table 1. Median values and their corresponding 70% confidence intervals of the percentage error in the out-of-sample predicted accumulated numbers of infections. Predictions are based on calibrated data and evaluated from 10 to 90 days ahead using the methods of extending the transmission parameter β described in §2.3. The reports are different from those used in the calibration.

		10 days	15 days	30 days	45 days	60 days	90 days
New York City — USA							
AM		6.51 (1.90–14.3)	9.70 (2.44–24.6)	22.6 (4.98–49.7)	33.8 (9.23–76.2)	44.3 (12.3–112)	63.4 (22.8–215)
LR		7.74 (2.12–16.8)	11.5 (2.87–26.5)	24.5 (6.93–52.3)	35.3 (9.53–79.8)	44.1 (14.2–119)	64.1 (19.2–244)
TFS		15.5 (4.77–35.3)	21.3 (7.21–48.6)	32.0 (10.8–73.8)	41.8 (13.9–97.5)	51.6 (13.4–145)	72.0 (19.7–247)
TFSWN		15.4 (4.77–35.3)	21.2 (7.16–48.6)	32.0 (10.6–73.8)	41.8 (13.8–97.3)	51.6 (13.5–144)	72.0 (19.8–247)
MR		4.84 (1.39–11.2)	8.23 (2.32–19.5)	23.2 (6.47–52.0)	39.4 (13.7–92.4)	51.0 (20.6–167)	74.0 (33.1–523)
Chicago — USA							
AM		7.99 (2.24–18.09)	12.34 (3.13–27.6)	28.1 (7.62–67.8)	49.7 (17.5–110)	65.9 (24.7–210)	79.6 (28.2–588)
LR		10.13 (2.79–20.53)	14.92 (4.52–31.0)	33.5 (10.3–73.7)	53.8 (22.2–129)	69.4 (24.7–242)	82.1 (28.7–620)
TFS		22.58 (6.45–46.91)	30.57 (9.78–66.9)	49.1 (17.7–140)	68.0 (25.5–269)	76.4 (31.2–422)	86.4 (25.3–930)
TFSWN		22.58 (6.54–46.90)	30.63 (9.79–66.7)	49.0 (17.8–140)	68.0 (25.6–268)	76.3 (31.3–420)	86.4 (25.1–929)
MR		5.89 (1.86–13.86)	10.56 (3.14–24.7)	27.5 (7.67–66.5)	46.4 (14.6–144)	61.5 (21.8–310)	81.8 (37.4–1068)
State of Sao Paulo — Brazil							
AM		6.76 (1.73–21.9)	14.6 (3.94–48.6)	27.3 (7.12–74.7)	38.4 (10.6–92.8)	48.0 (14.8–136)	60.4 (22.8–205)
LR		7.15 (1.80–20.7)	11.9 (3.36–38.8)	22.2 (5.79–67.0)	28.7 (8.90–84.3)	36.4 (11.3–104)	50.4 (14.7–206)
TFS		10.2 (2.02–31.2)	13.0 (2.45–36.8)	19.0 (5.04–49.9)	26.1 (7.45–62.6)	33.2 (8.71–82.3)	45.7 (13.2–131)
TFSWN		10.3 (2.00–30.7)	13.0 (2.57–35.5)	21.1 (5.67–64.8)	30.4 (8.48–80.9)	35.9 (11.2–89.9)	47.5 (14.3–138)
MR		4.98 (1.01–13.7)	15.6 (4.15–54.2)	32.9 (9.58–90.0)	48.0 (16.4–139)	62.2 (20.9–303)	74.8 (28.4–748)
British Columbia — Canada							
AM		7.52 (2.36–20.4)	12.7 (3.66–32.2)	27.7 (6.19–68.9)	43.4 (11.9–117)	61.2 (16.9–219)	84.4 (31.8–606)
LR		8.97 (2.74–24.4)	14.9 (4.61–37.4)	30.8 (6.66–75.6)	47.3 (14.1–131)	64.7 (15.4–248)	88.3 (28.8–639)
TFS		19.2 (5.16–42.8)	26.8 (6.87–60.8)	45.0 (12.2–89.5)	61.7 (14.7–185)	73.8 (17.5–311)	91.2 (30.9–768)
TFSWN		19.2 (5.23–42.8)	26.9 (6.84–60.8)	45.0 (12.1–89.5)	61.7 (14.7–184)	73.5 (17.4–311)	91.2 (30.9–769)
MR		5.11 (1.42–14.1)	9.41 (2.29–27.9)	22.8 (5.22–67.2)	36.1 (10.6–136)	55.2 (15.3–286)	77.9 (27.9–1479)

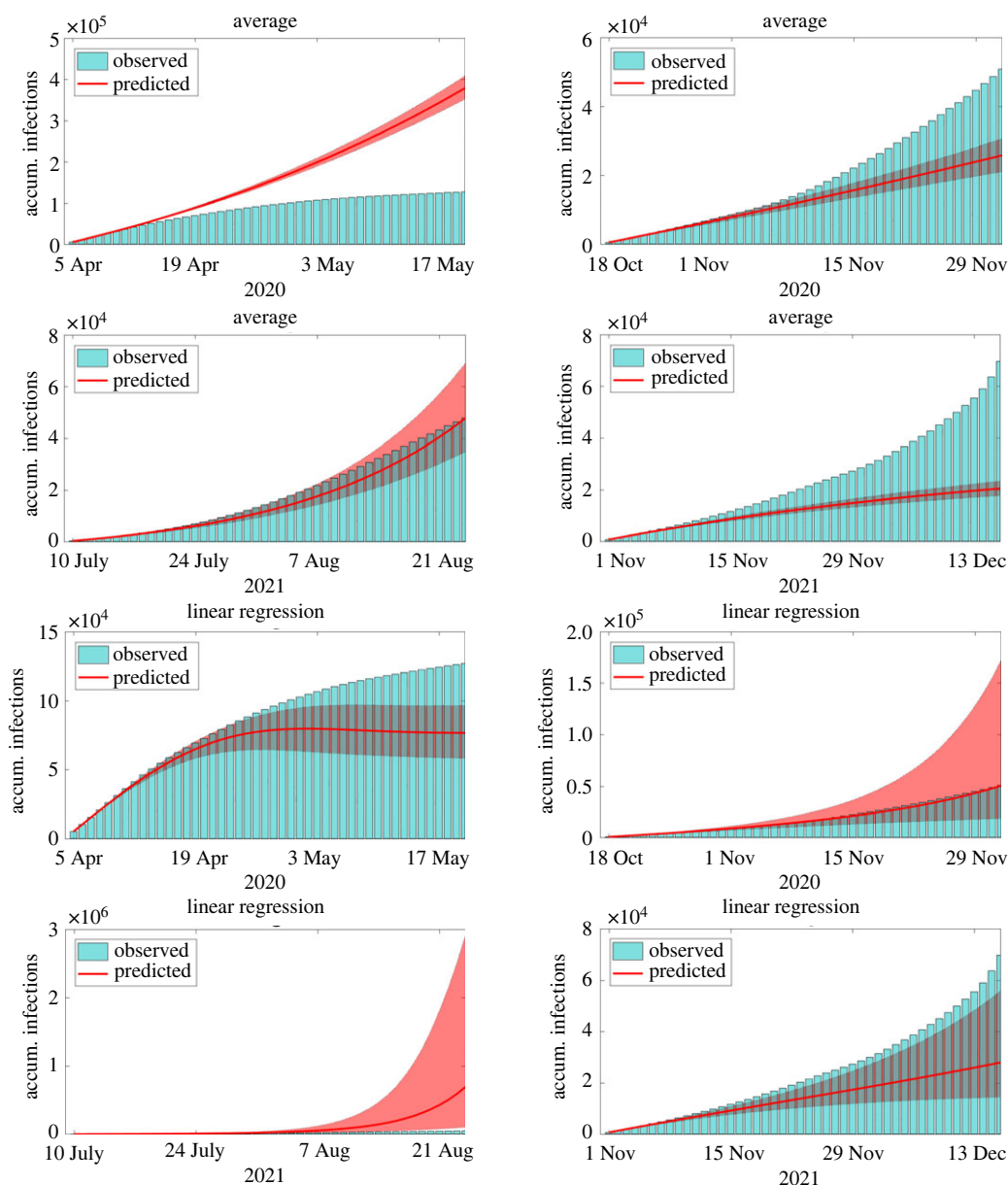


Figure 2. Comparison between reports and model predictions of accumulated COVID-19 infections during 45 consecutive days. Predictions are made just before four outbreaks of COVID-19 in NYC, including those triggered by the Delta and Omicron variants. Bars are the reports, solid lines and the filled areas are the median values and the 90% CI of model predictions, respectively. Predictions are evaluated using the methods for extending the transmission parameter β described in §2.3.

are from 5 April to 17 May 2020, from 18 October to 29 November 2020, from 10 July to 21 August 2021 and from 1 November to 13 December 2021, all them with roughly 45 days. In all cases, the calibration uses data until the day before the period of prediction, namely, 4 April 2020, 17 October 2020, 9 July 2021 and 9 November 2021. Predictions are compared with the accumulated number of reported infections and confidence intervals are generated with bootstrapping [32]. For MR and TFSWN, we use median values of infections. The comparison between reported and predicted accumulated infections during the periods of interest can be found in figures 2–4. Each column of panels refers to one outbreak.

In the first wave of infections, the LR and MR presented the best forecast performance with median values close to the observed number of infections. The LR started to underpredict the magnitude of the wave after 10–15 days of forecast, whereas the MR presented a wide 90% CI containing the reports. The AM presented accurate predictions until 10–15 days ahead. After this period, the method over predicted

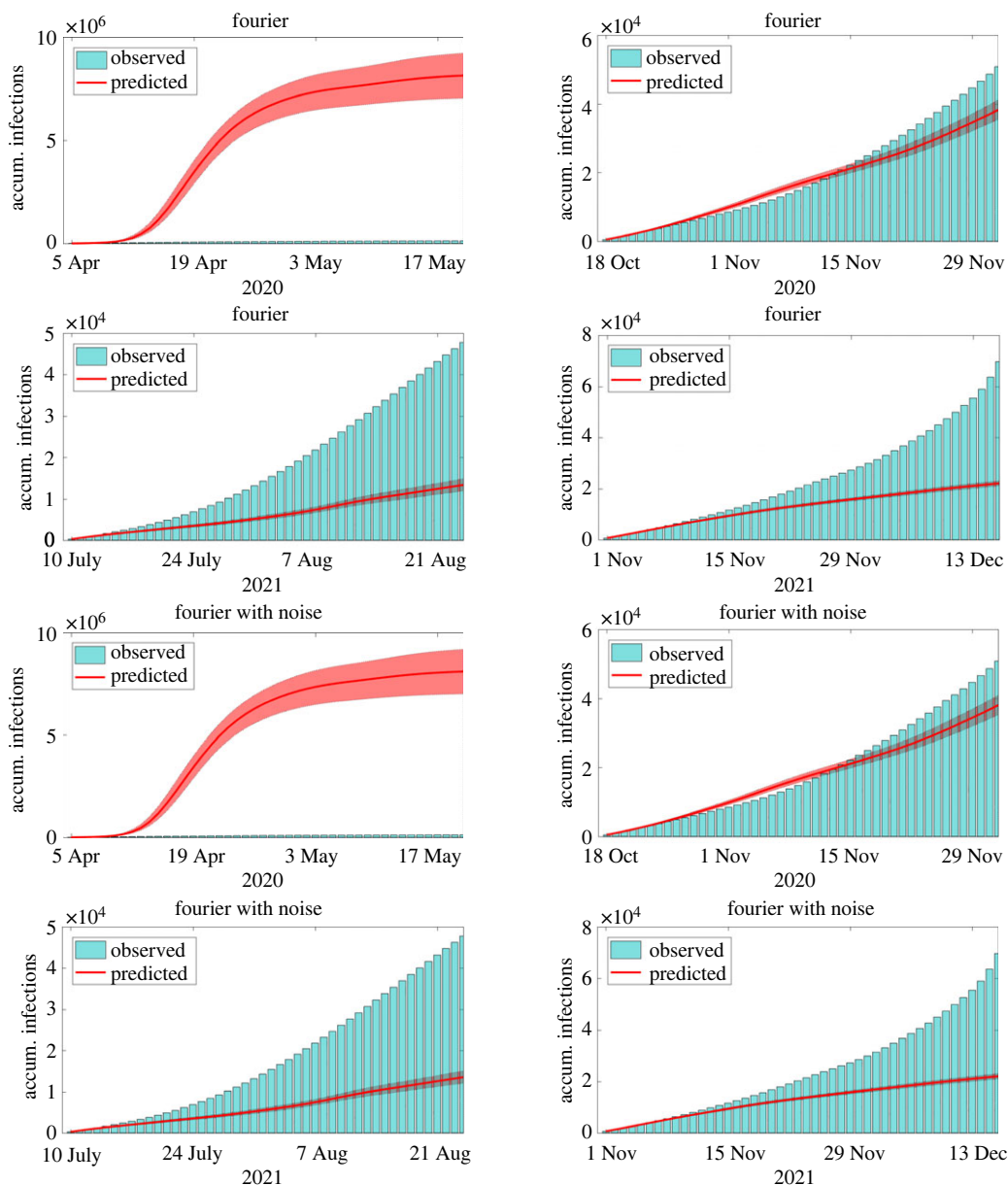


Figure 3. Comparison between reports and model predictions of accumulated COVID-19 infections during 45 consecutive days. Predictions are made just before four outbreaks of COVID-19 in NYC, including those triggered by the Delta and Omicron variants. Bars are the reports, solid lines and the filled areas are the median values and the 90% CI of model predictions, respectively. Predictions are evaluated using the methods for extending the transmission parameter β described in §2.3.

considerably the reported infections. The TFS and TFSWN forecast poorly the first outbreak probably due to its tendency for overfitting. It is worth noticing that, during the first wave, data were poorer.

In the other three outbreaks, all the models presented similar results. The AM forecast well the magnitude of the three waves, despite its 90% CI, for the second and fourth waves, did not contain the reported values for periods longer than 20–30 days. The LR presented more accurate predictions than the AM. It only overpredicted the third wave. The TFS and TFSWN predicted well the magnitude of the second wave but underpredicted the other two. The MR predicted accurately all the three waves, with its median values matching the reports of the third wave. The 90% CI was not too wide and contained the reports.

Overall, all the techniques performed well for shorter periods, i.e. from 10 to 15 days. For periods until 45 days, the AM, LR and MR seem to perform considerably better than the TFS and TFSWN. However, for larger periods of forecast, predictions become less reliable and can be only considered qualitatively.

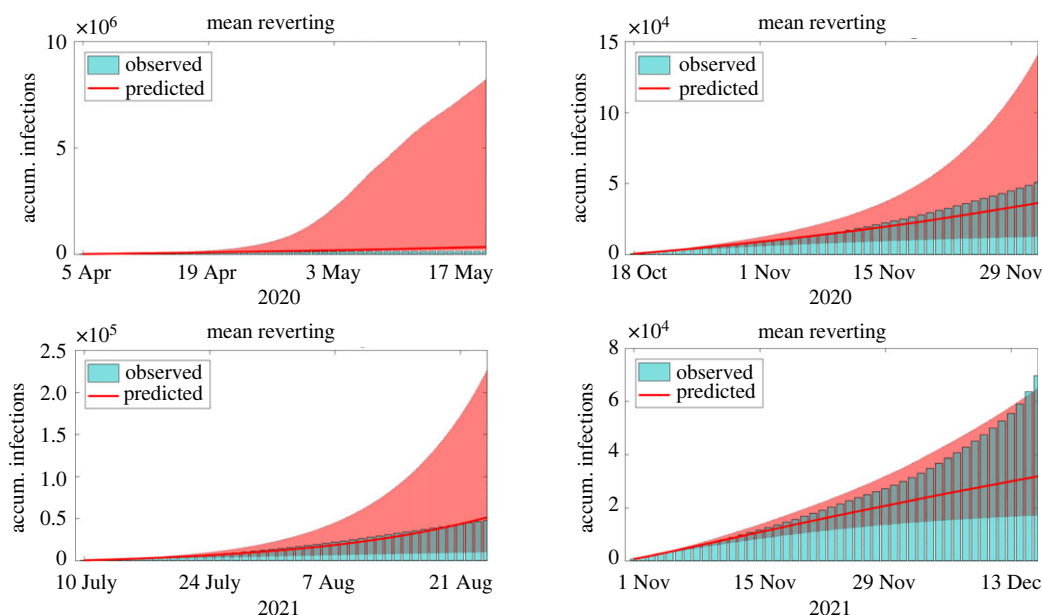


Figure 4. Comparison between reports and model predictions of accumulated COVID-19 infections during 45 consecutive days. Predictions are made just before four outbreaks of COVID-19 in NYC, including those triggered by the Delta and Omicron variants. Bars are the reports, solid lines and the filled areas are the median values and the 90% CI of model predictions, respectively. Predictions are evaluated using the methods for extending the transmission parameter β described in §2.3.

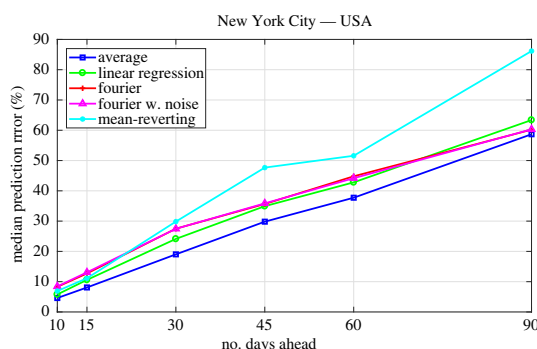


Figure 5. Median values of the percentage error in the predicted accumulated numbers of infections. Predictions are based on calibrated data and evaluated from 10 to 90 days ahead using the methods of extending the transmission parameter β described in §2.3. The epidemiological model in Albani *et al.* [3] was used.

Although the results of the AM, the LR and the MR are comparable, the MR has the advantage of incorporating the randomness of the transmission, offering a series of possible future scenarios. Based on these examples we suggest that, for shorter periods of forecast, i.e. until 45 days, the AM and MR are reliable choices. For longer periods, their results must be compared with those given by the TFS or the TFSWN. The advantage of the TFSWN over the TFS is the incorporation of uncertainty, which makes it easy to provide predictions naturally inside confidence intervals.

To illustrate that the results above can be extended to a larger class of SEIR-type models, we repeat the first experiment with NYC data and the model proposed by Albani *et al.* [3]. The version of the model used here accounts for different levels of disease severity and their dependency on age-range. Figure 5 shows the evolution of the median values of the normalized forecast errors. The same values with their 70% CI can be found in table 2. As figure 5 and table 2 show, by including more characteristics of the disease in the SEIR-type model, the overall forecast accuracy did not improve. We used all the forecast techniques from §2.3. All of them presented similar forecast errors. Such a pattern may be linked to the increase of the uncertainties in model calibration, since more sophisticated models have larger sets of unknowns for the same experimental datasets, since NYC does not have daily reports of COVID-19 cases, hospitalizations and deaths classified by age-range.

Table 2. Median values and their corresponding 70% confidence intervals of the percentage error in the predicted accumulated numbers of infections. Predictions are based on calibrated data and evaluated from 10 to 90 days ahead using the methods of extending the transmission parameter β described in S2.3.

New York City — USA						
	10 days	15 days	30 days	45 days	60 days	90 days
AM	4.60 (1.89–8.73)	8.09 (4.27–18.6)	19.0 (7.69–56.0)	29.8 (11.0–70.5)	37.7 (11.8–124)	58.7 (11.2–344)
LR	5.78 (2.34–9.35)	10.6 (4.63–16.5)	24.1 (11.7–43.1)	34.9 (11.8–70.0)	42.8 (10.7–125)	63.4 (21.6–211)
TFS	8.29 (3.30–15.0)	12.7 (5.16–22.1)	27.5 (4.50–53.7)	35.7 (11.0–75.0)	44.8 (23.7–114)	60.2 (32.5–191)
TFSWN	8.43 (3.39–15.0)	13.1 (5.18–22.3)	27.5 (4.69–53.9)	35.9 (10.5–75.0)	44.2 (23.3–114)	60.3 (34.2–191)
MR	6.91 (1.76–14.2)	11.1 (5.14–23.1)	29.8 (5.07–51.5)	47.7 (17.1–115)	51.6 (20.6–448)	86.2 (36.1–1532)

4. Discussion

Although the model used to test different prediction techniques may seem to be simple, by considering time varying transmission parameter and death rate, it incorporates naturally stochastic fluctuations in the pathogen transmission, like the introduction of new variants, changes in the disease virulence and mortality. In other words, providing an accurate nowcast. More sophisticated models are harder to calibrate and to forecast scenarios within longer periods, it is necessary to predict the evolution of other quantities, like mobility information or the disease severity dependency on age. It is also difficult to estimate the causality relations between transmission and mobility [33,34]. Moreover, as figure 5 shows, the performance of more sophisticated models may not be considerably better. Thus, if the main interest is in scenarios regarding the general population, it is better to use simpler, but robust and accurate, models.

Using the same dataset, predictions with different techniques presented considerably different performances in forecast. For shorter periods, the AM, the LR and the MR outperformed the TFS and the TFSWN. It is also interesting that the MR predicted accurately the four waves of infections in NYC, as illustrated by figure 4. In all cases, the model prediction 90% CI contained the reported infections. This indicates that the MR incorporates better the uncertain nature of the virus transmission and shows that transmission parameter β must be modelled as a stochastic process.

The reliability of model forecast decreases considerably with respect to the length of the prediction time interval, as tables 1–2 show. This is due to unpredictable regime changes in the transmission dynamics, like the emergence of new strains of the pathogen, vaccination or yet effectiveness of non-pharmaceutical measures like lockdowns. Thus, for periods longer than 45 days, predictions must be considered only qualitatively.

Finally, the present study has some limitations since we considered only the class of SEIR-type models. There are a number of other ways to model the dynamics of infectious diseases [2], but accounting for them is beyond the scope of this article that is intended to enlighten the stochastic nature of transmission. Moreover, we did not perform a rigorous statistical analysis of the transmission parameter and other rates defining the SEIR-type model in equations (2.1)–(2.5) since the calibration techniques considered in this work are deterministic. The statistical analysis of SEIR-type models and their parameters will be the subject of a future work.

Data accessibility. Data and relevant code for this research work are stored in GitHub: <https://github.com/viniciusalbani/NowcastForecastCovid> and have been archived within the Zenodo repository: <https://doi.org/10.5281/zenodo.6949649>. The datasets used in this study are also publicly available from [35–38].

Authors' contributions. V.V.L.A.: conceptualization, formal analysis, investigation, methodology, software, writing—original draft, writing—review and editing; R.A.S.A.: conceptualization, formal analysis, investigation, methodology, software, writing—original draft, writing—review and editing; E.M.: conceptualization, formal analysis, investigation, methodology, software, writing—original draft, writing—review and editing; J.P.Z.: conceptualization, formal analysis, investigation, methodology, software, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein. Conflict of interest declaration. We declare we have no competing interest.

Funding. V.V.L.A. acknowledges the financial support from Fundação Butantan and Fundação de Amparo à Pesquisa e Inovação do Estado de Santa Catarina through the grant nos. 01/2020 and 00002847/2021, respectively. R.A.S.A. acknowledges the financial support from Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) through the grant nos. E-26/202.932/2019 and E-26/202.933/2019. E.M. acknowledges the financial support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação Butantan through the grant nos. 305544/2011-0 and 01/2020, respectively. J.P.Z. acknowledges the financial support from Khalifa University, CNPq and Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro through the grant nos. FSU-2020-09, 307873/2013-7 and E-26/202.927/2017, respectively.

References

1. Kermack WO, McKendrick AG. 1927 A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. A* **115**, 700–721. (doi:10.1098/rspa.1927.0118)
2. Keeling MJ, Rohani R. 2008 *Modeling infectious diseases in humans and animals*. Princeton, NJ: Princeton University Press.
3. Albani VVL, Velho RM, Zubelli JP. 2021 Estimating, monitoring, and forecasting the Covid-19 epidemics: a Spatio-temporal approach applied to NYC Data. *Sci. Rep.* **11**, 9089. (doi:10.1038/s41598-021-88281-w)
4. Anastassopoulou C, Russo L, Tsakris A, Siettos C. 2020 Data-based analysis, modelling and forecasting of the COVID-19 outbreak. *PLoS ONE* **15**, e0230405. (doi:10.1371/journal.pone.0230405)
5. Calvetti D, Hoover AP, Rose J, Somersalo E. 2020 Metapopulation network models for understanding, predicting, and managing the Coronavirus disease COVID-19. *Front. Phys.* **8**, 261. (doi:10.3389/fphys.2020.00261)

6. Dehning J, Zierenberg J, Spitzner FP, Wibral M, Neto JP, Wilczek M, Priesemann V. 2020 Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science* **369**, eabb9789. (doi:10.1126/science.abb9789)
7. Gatto M, Bertuzzo E, Mari L, Miccoli S, Carraro L, Casagrandi R, Rinaldo A. 2020 Spread and dynamics of the COVID-19 epidemic in Italy: effects of emergency containment measures. *Proc. Natl Acad. Sci. USA* **117**, 10 484–10 491. (doi:10.1073/pnas.2004978117)
8. Wu JT *et al.* 2020 Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nat. Med.* **26**, 506–510. (doi:10.1038/s41591-020-0822-7)
9. Wu JT, Leung K, Leung GM. 2020 Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *Lancet* **395**, 689–697. (doi:10.1016/S0140-6736(20)30260-9)
10. Campos EL, Cysne RP, Madureira AL, Mendes GL. 2021 Multi-generational SIR modeling: determination of parameters, epidemiological forecasting and age-dependent vaccination policies. *Infect. Dis. Modell.* **6**, 751–765.
11. Goyal A, Reeves DB, Schiffer JT. 2022 Multi-scale modelling reveals that early super-spreader events are a likely contributor to novel variant predominance. *J. R. Soc. Interface* **19**, 20210811. (doi:10.1098/rsif.2021.0811)
12. Gray A, Greenhalgh D, Hu L, Mao X, Pan J. 2011 A stochastic differential equation SIS epidemic model. *SIAM J. Appl. Math.* **71**, 876–902. (doi:10.1137/10081856X)
13. Ioannidis JPA, Cripps S, Tanner MA. 2020 Forecasting for COVID-19 has failed. *Int. J. Forecast.* **38**, 423–438. (doi:10.1016/j.ijforecast.2020.08.004)
14. Lipton A, Lopez de Prado M. 2020 Three quant lessons from COVID-19. *Risk Magazine*, April 2020. (<https://ssrn.com/abstract=3580185> or <http://dx.doi.org/10.2139/ssrn.3580185>)
15. Aizenman N. 2020 New Global Coronavirus Death Forecast is Chilling – and Controversial. Accessed on 09-01-2022. Goats and Soda STORIES OF LIFE IN A CHANGING WORLD. Available from: <https://www.npr.org/sections/goatsandsoda/2020/09/04/909783162/new-global-coronavirus-death-forecast-is-chilling-and-controversial>.
16. Trian N. 2020. A year of protests and civil disobedience. Accessed on 09-01-2022. France 24. Available from: <https://www.france24.com/en/europe/20201228-2020-a-year-of-dissent-and-civil-disobedience>.
17. Vogel KP, Rutenberg J, Lerer L. 2020 The Quiet Hand of Conservative Groups in the Anti-Lockdown Protests. Accessed on 09-01-2022. The New York Times. Available from: <https://www.nytimes.com/2020/04/21/us/politics/coronavirus-protests-trump.html>.
18. Thomas MT. 2021 Protests during the 2020 COVID-19 lockdown in South Africa. Molab Inventory of Mobilities and Socioeconomic Changes. Halle/Saale: Max Planck Institute for Social Anthropology. (doi:10.48509/molab.8077)
19. Gnanvi J, Salako KV, Kotanmi B, Kakai RG. 2021 On the reliability of predictions on Covid-19 dynamics: a systematic and critical review of modelling techniques. *Infect. Dis. Modell.* **6**, 258–272. (doi:10.1016/j.idm.2020.12.008)
20. Candido DS *et al.* 2020 Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science* **369**, 1255–1260. (doi:10.1126/science.abb2161)
21. Kishore N, Kahn R, Martinez PP, De Salazar PM, Mahmud AS, Buckee CO. 2021 Lockdowns result in changes in human mobility which may impact the epidemiologic dynamics of SARS-CoV-2. *Sci. Rep.* **11**, 6995. (doi:10.1038/s41598-021-86297-w)
22. Alleman TW *et al.* 2021 Assessing the effects of non-pharmaceutical interventions on SARS-CoV-2 transmission in Belgium by means of an extended SEIQRD model and public mobility data. *Epidemics* **37**, 100505. (doi:10.1016/j.epidem.2021.100505)
23. Coelho FC *et al.* 2020 Assessing the spread of COVID-19 in Brazil: mobility, morbidity and social vulnerability. *PLoS ONE* **15**, e0238214. (doi:10.1371/journal.pone.0238214)
24. Albani VVL, Loria J, Massad E, Zubelli JP. 2021 COVID-19 underreporting and its impact on vaccination strategies. *BMC Infect. Dis.* **21**, 1111. (doi:10.1186/s12879-021-06780-7)
25. Albani V, Loria J, Massad E, Zubelli JP. 2021 The impact of COVID-19 vaccination delay: a data-driven modelling analysis for Chicago and New York City. *Vaccine* **39**, 6088–6094. (doi:10.1016/j.vaccine.2021.08.098)
26. Albani V, Albani R, Bobko N, Massad E, Zubelli J. 2021 On the role of financial support programs in mitigating the Sars-CoV-2 spread in Brazil. *MedRxiv preprint*.
27. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, Azman AS, Reich NG, Lessler J. 2020 The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Ann. Intern. Med.* **172**, 577–583. (doi:10.7326/M20-0504)
28. Grasselli G *et al.* 2020 Baseline characteristics and outcomes of 1591 patients infected with SARS-CoV-2 admitted to ICUs of the Lombardy region, Italy. *J. Am. Med. Assoc.* **323**, 1574–1581. (doi:10.1001/jama.2020.5394)
29. Albani V, De Cezaro A. 2019 A connection between uniqueness of minimizers and Morozov-like discrepancy principles in Tikhonov-type regularization. *Inverse Probl. Imaging* **13**, 211–229. (doi:10.3934/ipi.2019012)
30. Hamilton JD. 2020 *Time series analysis*. Princeton, NJ: Princeton University Press.
31. Gatheral J. 2006 *The volatility surface: a practitioner's Guide*. Hoboken, NJ: John Wiley & Sons.
32. Chowell G. 2017 Fitting dynamic models to epidemic outbreaks with quantified uncertainty: a primer for parameter uncertainty, identifiability, and forecasts. *Infect. Dis. Modell.* **2**, 379–398. (doi:10.1016/j.idm.2017.08.001)
33. Larrosa JM. 2021 SARS-CoV-2 in Argentina: lockdown, mobility, and contagion. *J. Med. Virol.* **93**, 2252–2261. (doi:10.1002/jmv.26659)
34. Tomori DV *et al.* 2021 Individual social contact data and population mobility data as early markers of SARS-CoV-2 transmission dynamics during the first wave in Germany – an analysis based on the COVIMOD study. *BMC Med.* **19**, 271. (doi:10.1186/s12916-021-02139-6)
35. NYC. Covid-19 data from NYC. Available from: <https://www1.nyc.gov/site/doh/covid/covid-19-data.page>.
36. Chicago. COVID-19 data from Chicago. Available from: <https://www.chicago.gov/city/en/sites/covid-19/home.html>.
37. Canada. COVID-19 data from Canada. Available from: <https://resources-covid19canada.hub.arcgis.com/datasets/provincial-daily-totals/explore>.
38. SEADE. COVID-19 data from São Paulo. Available from: <https://www.seade.gov.br/coronavirus/>.