Hybrid Modeling of Regional COVID-19 Transmission Dynamics in the U.S.

Yue Bai, Abolfazl Safikhani, and George Michailidis, *Member, IEEE*

***Abstract*—The fast transmission rate of COVID-19 worldwide has made this virus the most important challenge of year 2020. Many mitigation policies have been imposed by the governments at different regional levels (country, state, county, and city) to stop the spread of this virus. Quantifying the effect of such mitigation strategies on the transmission and recovery rates, and predicting the rate of new daily cases are two crucial tasks. In this paper, we propose a hybrid modeling framework which not only accounts for such policies but also utilizes the spatial and temporal information to characterize the pattern of COVID-19 progression. Specifically, a piecewise susceptible-infected-recovered (SIR) model is developed while the dates at which the transmission/recover rates change significantly are defined as "break points" in this model. A novel and data-driven algorithm is designed to locate the break points using ideas from fused lasso and thresholding. In order to enhance the forecasting power and to describe additional temporal dependence among the daily number of cases, this model is further coupled with spatial smoothing covariates and vector auto-regressive (VAR) model. The proposed model is applied to several U.S. states and counties, and the results confirm the effect of "stay-at-home orders" and some states' early "re-openings" by detecting break points close to such events. Further, the model provided satisfactory short-term forecasts of the number of new daily cases at regional levels by utilizing the estimated spatio-temporal covariance structures. They were also better or on par with other proposed models in the literature, including flexible deep learning ones. Finally, selected theoretical results and empirical performance of the proposed methodology on synthetic data are reported which justify the good performance of the proposed method.**

***Index Terms*—Break point detection, COVID-19, short-term forecast, spatio-temporal model.**

I. Introduction

SINCE the first officially reported case in China in late December 2019, the SARS-CoV-2 virus spread worldwide within weeks. As of late May 2021, there have been ∼34 million confirmed cases of COVID-19 in the United States alone and more than 170 million worldwide. In response to the rapid growth of confirmed cases, followed by hospitalizations and fatalities, initially in the Hubei Province and in particular its capital Wuhan in China, and subsequently in Northern Italy, Spain and the tri-state area of New York, New Jersey and Connecticut, various mitigation strategies were put rapidly in place with the most stringent one being "stay-at-home" orders. The key purpose of such strategies was to reduce the virus transmission rate and consequently pressure on public health infrastructure [1]. To that end, the California governor issued a "stay-at-home" order on March 19, 2020, that was quickly followed by another 42 states by early April. All states with such orders proceeded with multi-phase reopening plans starting in early May, allowing various non-essential business to operate, possibly at reduced capacity levels to enforce social distancing guidelines. In addition, mask wearing mandates also came into effect [2] as emerging evidence from clinical and laboratory studies showed that masks reduce the spread [3]. However, these reopening plans led to a substantial increase in the number of confirmed COVID-19 cases in many US states, followed by increased number of fatalities throughout the summer of 2020, concentrated primarily in the Southern US states. Different states and local communities adopted and implemented different non-pharmaceutical interventions to reduce infections, but a "3 rd wave" emerged in the fall of 2020 with cooling temperatures and people spending more time indoors. Further, during late fall of 2020, various variants of concerns started emerging around the world, characterized by higher transmission capabilities and potentially increased severity based on hospitalizations and fatalities [4]. Variants that exhibited a certain degree of spread in the US include B.1.1.7 (first detected in the United Kingdom), B.1.351 (first detected in South Africa), B.1.427 and B.1.429 (first detected in California) and P.1 (first detected in Brazil). The B.1.1.7 variant went on to become the dominant one in the US by March 2021, displacing the original dominant strain B.1.2, while the B.1.427/429 ones represented about 15% of the total based on genomic surveillance studies.[1]

The emergence of the COVID-19 pandemic led to the development of many data science and signal processing modeling approaches addressing diverse issues, including forecasting progress of the disease, impact of non-pharmaceutical intervention strategies [5], methods to estimate the Infection and Case Fatality Rates (IFR/CFR) [6], pre-existing conditions and

Manuscript received May 29, 2021; revised November 14, 2021; accepted December 20, 2021. Date of publication January 5, 2022; date of current version April 18, 2022. The work of George Michailidis was supported by the National Science Foundation (NSF) under Grant DMS-1821220. This work was supported by UF Informatics Institute COVID-19 SEED Fund. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Bjoern Schuller. *(Corresponding author: Abolfazl Safikhani.)*

Yue Bai is with the Department of Statistics, University of Florida, Gainesville, FL 32611 USA (e-mail: baiyue@ufl.edu).

Abolfazl Safikhani and George Michailidis are with the Department of Statistics and the Informatics Institute, University of Florida, Gainesville, FL 32611 USA (e-mail: a.safikhani@ufl.edu; gmichail@ufl.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/JSTSP.2022.3140703, provided by the authors.

Digital Object Identifier 10.1109/JSTSP.2022.3140703

[1][Online]. Available: https://covid.cdc.gov/covid-data-tracker/##variant-proportions

clinical factors that impact the CFR [7], computer audition for diagnosing COVID-19 [8], image analysis for COVID-19 [9], phylogenetic network analysis of Covid-19 genomes [10], impact of aerosol transmission to public health [11], guidelines for reopening critical social activities such as schools [12]. Note that as Covid-19 progressed together with our knowledge about it, the range of topics addressed significantly expanded, while the focus also exhibited certain shifts. As an example, early on (March 2020) it was believed that the virus can spread through contaminated surfaces, known as fomites, and this informed both the World Health Organization (WHO) and the Center of Disease Control (CDC) recommendations[2] on surface cleaning and disinfection. However, subsequent studies and investigations of outbreaks pointed that the majority of transmissions occur through droplets and aerosols that led to a revision of recommendations by the WHO and the CDC.[3] It also led to new research on aerosol dispersion models and on the role of ventilation to mitigate transmission [13]. Nevertheless, forecasting the spread of the epidemic throughout its course (initially with the imposition of various mitigation strategies, and more recently through the emergence of more transmissible variants and the increased pace of vaccination campaigns around the world) has remained a key task and a number of signal processing approaches have been developed as briefly summarized next.

A. Related Work

A number of epidemic models have been developed to analyze and predict COVID-19 transmission dynamics. Mathematical models, such as the class of susceptible-infectious-recovered (SIR) models are widely used to model and forecast epidemic spreads. [14] proposed a time-dependent SIR model and tracked transmission and recovery rates at each time point by employing ridge regression while [15] proposed a discrete-time susceptible-infectious-recovered-dead (SIRD) model and provided estimations of the basic reproduction number (R_0), and the infection mortality and recovery rates by least squares method. Moreover, [16] and [17] built an extended SIR model with time-varying transmission rates and implemented a Markov Chain Monte Carlo algorithm (MCMC) to obtain posterior estimates and credible intervals of the model parameters.

A number of models focused on identifying a change in the parameters of the underlying model employed. For example, [18] combined the widely used SIR model (see Section II) with Bayesian parameter inference through MCMC algorithms, assuming a time-dependent transmission rate. Instead of directly estimating a change point in the transmission rate and the other parameters in the SIR model, they assumed a fixed value on the number of the change points, and imposed informative prior distributions on their locations, as well as the transmission rate based on information from intervention policies. Further, [19] proposed to model the time series of the log-scaled cumulative confirmed cases and deaths of each country via a piecewise linear trend model. They combined the self-normalization (SN) change-point test with the narrowest-over-threshold (NOT) algorithm [20] to achieve multiple change-point estimation. Moreover, [21] and [22] analyzed the effect of social distancing measures adopted in Europe and the United States, respectively, using an interrupted time series (ITS) analysis of the confirmed case counts. Their work aim to find a change point in the time series data of confirmed cases counts for which there is a significant change in the growth rate. In [21]'s paper, the change points were determined by linear threshold regression models of the logarithm of daily cases while [22] used an algorithm developed in [23], based on an L_0 penalty on changes in slope to identify the change points. Finally, [24] utilized a branching process for modeling and forecasting the spread of COVID-19.

Another line of work employed spatio-temporal models for parameter estimation and forecasting the spread of COVID-19. For example, [25] introduced an additive varying coefficient model and coupled it with a non-parametric approach for modeling the data, to study spatio-temporal patterns in the spread of COVID-19 at the county level. Further, [26] proposed a heterogeneous infection rate model with human mobility from multiple regions and trained it using weighted least squares at regional levels while [27] fitted a generalized additive model (GAM) to quantify the province-specific associations between meteorological variables and the daily cases of COVID-19 during the period under consideration.

In addition to mathematical methods, many machine learning/deep learning methods were applied for forecasting of COVID-19 transmission. For example, [28] and [29] employed Artificial Neural Networks (ANN) and Long Short-Term Memory (LSTM) type deep neural networks to forecast future COVID-19 cases in Iran and Canada, respectively, while [30] developed a modified stacked auto-encoder for modeling the transmission dynamics of the epidemic. Review paper [31] presents a summary of recent COVID-19 forecasting models.

The previous brief overview of the literature indicates that there are two streams of models, the first mechanistic and the second statistical in nature. The former (SIR/SIRD) describe key components of the transmission chain and its dynamics and have proved useful in assessing scenarios of the evolution of a contagious disease, by altering the values of key model parameters. However, they are macroscopic in nature and can not easily incorporate additional information provided either by mitigation strategies or other features, such as movement of people assessed through cell phone data. Statistical models can easily utilize such information in the form of covariates to improve their forecasting power. However, they primarily leverage correlation patterns in the available data, that may be noisy, especially at more granular spatio-temporal scales (e.g., county or city level) that are of primary interest to public health officials and policy makers.

To that end, this paper aims to develop an interpretable *hybrid* model that combines a mechanistic and a statistical model, that respects the theoretical transmission dynamics of the former, but also incorporates additional spatio-temporal characteristics resulting in improved forecasting capabilities at fairly granular spatio-temporal scales.

[2][Online]. Available: https://www.who.int/news-room/commentaries/detail/transmission-of-sars-cov-2-implications-for-infection-prevention-precautions

[3][Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/science/science-briefs/sars-cov-2-transmission.html

Specifically, we analyze confirmed cases and deaths related to COVID-19 from several states and counties/cities in the United States from March 1st, 2020 to March 31st, 2021. In the absence of non-pharmaceutical interventions, the spread of COVID-19 can be modeled by a SIR model with fixed transmission and recovery rates. One of the main reasons to select that model as the building block for the proposed methodology is that the transmission and recovery rates are easy to interpret and hence can be used in policy decision making. However, many diverse mitigation policies were put in place at different regional levels in the U.S. Thus, the simple SIR model may not be a good fit for the data. Instead, we propose a piecewise stationary SIR model (Model 1), i.e. the SIR model parameters may change at certain (unknown) time points. Such time points are defined as "break (change) points". Unlike some other methods discussed in the literature review, in our modeling framework, the number of change points and their locations are assumed to be *unknown* and must be inferred from the data. Such flexibility on the modeling front allows inferring potentially different temporal patterns across different regions (states or counties), and yields a data-driven segmentation of the data which subsequently improves the fit (see more details in Section IV), but also complicates the model fitting procedures. To that end, a novel data-driven algorithm is developed to detect all break points, and to estimate the model parameters within each stationary segment. Specifically, we define certain time blocks and assume the model parameters are fixed during each block of time points. Then, a fused lasso penalty is used to estimate all model parameters [32]. This procedure is further coupled with hard-thresholding and exhaustive search steps to estimate the number and location of change points (details provided in Section II). To enhance the forecasting power of the model and to capture additional spatial and temporal dependence not explained through the SIR model, the piecewise constant SIR model is coupled with spatial smoothing (Model 2) and time series components (Model 3). The former is accomplished through the addition of a spatial effect term which accounts for the effect of neighboring regions, while the latter through a Vector Auto-Regressive (VAR) component to capture additional auto-correlations among new daily cases and deaths. Capturing the spatio-temporal dependence through Model 3 aids in reducing the prediction error significantly (sometimes around 80%) compared to the piecewise SIR model which confirms the usefulness of a hybrid modeling framework (for more details see Section IV). To verify the applicability of the proposed methodology to other data sets with similar characteristics, the developed algorithm is tested over several simulation settings and exhibits very satisfactory performance (details in Section III) and some theoretical properties of the proposed method (prediction consistency, as well as detection accuracy) are established in Appendix B.

The remainder of the paper is organized as follows. In Section II, proposed statistical models are introduced and data-driven algorithms are described to estimate their parameters. The proposed algorithms are tested on various simulation settings and the results are reported in Section III. The proposed models are applied to several U.S. states and counties and the results are described in Section IV. Finally, some concluding remarks are drawn in Section V.

II. A Family of Spatio-Temporal Heterogeneous SIR Models

The proposed class of hybrid models leverages the framework of the SIR model, which is presented next to set up key concepts.

A. The Standard SIR Model With Fixed Transmission and Recovery Rates

The standard SIR model [33] is a mechanistic model, wherein the total population is divided into the following three compartments: susceptible (uninfected), infected, and recovered (healed or dead). It is assumed that each infected individual, on average, infects β other individuals per unit time, and each infected individual recovers at rate γ. The two key model parameters, the transmission rate β and recovery rate γ, are assumed to be fixed over time. The temporal evolution of the SIR model is governed by the following system of three ordinary differential equations:

$$\frac{dS}{dt} = -\beta\frac{SI_f}{N}, \quad \frac{dI_f}{dt} = \beta\frac{SI_f}{N} - \gamma I_f, \quad \frac{dR}{dt} = \gamma I_f, \quad (1)$$

where S, I_f and R represent the individuals in the population in the susceptible, infected and recovered stages, respectively. Note that the variables S, I_f and R always satisfy $S + I_f + R = N$, where N is the total population size. In this formulation, we ignore the change in the total population, so that N remains constant over time. Due to the fact that COVID-19 records are discrete in time ($\Delta t = 1$ day), we consider the discrete-time version of SIR model, so that for each $t = 1, \ldots, T-1$, the system comprises of the following three difference equations

$$S(t+1) - S(t) = -\beta\frac{S(t)I_f(t)}{N}, \quad (2)$$

$$I_f(t+1) - I_f(t) = \beta\frac{S(t)I_f(t)}{N} - \gamma I_f(t), \quad (3)$$

$$R(t+1) - R(t) = \gamma I_f(t), \quad (4)$$

where $S(t)$ stand for the number of susceptible individuals at time t, $I_f(t)$ for the number of infected ones and finally $R(t)$ for those recovered. Note that these three variables $S(t)$, $I_f(t)$ and $R(t)$ still satisfy the constraint $S(t) + I_f(t) + R(t) = N$.

Notice that the number of infected cases $I_f(t)$ is not observable. Specifically, confirmed COVID-19 case counts may not capture the total infected cases due to limited testing availability/capacity, especially at the beginning of the pandemic (testing has been primarily restricted to individuals with moderate to severe symptoms). For example, in the United States, over 90% of COVID-19 infections were not identified/reported at the beginning of the pandemic [34], [35]. To that end, we define a relationship between the true infected cases and observed/recorded infected cases through an under-reporting function. Specifically, define $\Delta I(t) = \Delta I_f(t) \times (1 - u(t+1))$ for $t = 1, \ldots, T-1$, where $\Delta I(t) = I(t+1) - I(t)$, $\Delta I_f(t) = I_f(t+1) - I_f(t)$, $I_f(t)$ is the true infected cases, $I(t)$ is the observed/recorded

infected cases, and finally $u(t)$ is the under-reporting function. We consider a parametric model for the function $u(t)$ (see more details in Section III). Note that we could also use non-parametric method to solve the under-reporting issue, but since the sample size of the time series is limited, in this paper, we consider a parametric method instead. Given the observed infected cases $I(t)$ and under-reporting function $u(t)$, one can transform the data back to $I_f(t)$ by

$$I_f(1) = \frac{I(1)}{1-u(1)}, \tag{5}$$

$$I_f(t) = \frac{\Delta I(t-1)}{1-u(t)} + I_f(t-1) = \sum_{i=2}^{t} \frac{\Delta I(i-1)}{1-u(i)} + \frac{I(1)}{1-u(1)}, \tag{6}$$

for $t = 2, \ldots, T$.

Combining the difference (2) normalized by the total population with the transformations stated in (5) yield to the following simple linear equations:

$$\underbrace{\begin{pmatrix} \frac{\Delta I(t)}{N(1-u(t+1))} \\ \frac{\Delta R(t)}{N} \end{pmatrix}}_{Y_t} = \underbrace{\begin{pmatrix} \frac{S(t)}{N^2} I_f(t) & -\frac{1}{N} I_f(t) \\ 0 & \frac{1}{N} I_f(t) \end{pmatrix}}_{X_t} \underbrace{\begin{pmatrix} \beta \\ \gamma \end{pmatrix}}_{B}, \tag{7}$$

for each $t = 1, \ldots, T-1$ where $\Delta I(t) = I(t+1) - I(t)$ and $\Delta R(t) = R(t+1) - R(t)$.

Next, we extend the standard SIR model to accommodate temporal and spatial heterogeneity as well as to include stochastic temporal components. The former is achieved, by allowing the transmission and recovery rates to vary over time and through the inclusion of an additional term in (7) that captures spatial effects while the latter is achieved through adding a vector auto-regressive component [36].

B. Modeling Framework: A Stochastic Piecewise Stationary SIR Model With Spatial Heterogeneity

Compared to the standard SIR model, the proposed modeling framework makes three major changes/modifications. First, the assumptions underlying the transmission and recovery rates of the standard SIR model are stringent. Both environmental factors and changes in population behavior can lead to time varying behavior and this has been the case for Covid-19; see, e.g., discussion in [37]. Variants of the SIR model with time varying parameters have been proposed in the literature [38]. For our application, we assume that the transmission and recovery rates are *piecewise constant* over time, reflecting the fact that their temporal evolution is impacted by intervention strategies and environmental factors (Model 1). Second, the standard SIR model and its piecewise stationary counterpart do not account for any influence due to inter-region mobility and travel activity. We incorporate such inter-region information by considering the influence exerted by its few neighboring regions such as cities, counties or states (Model 2); see also [26]. Third, the standard homogeneous SIR model, previously discussed, is deterministic; hence, the output of the model is fully determined by the parameter values of the transmission and recovery rates and the initial conditions. Its stochastic counterpart [39], [40], possesses some inherent randomness. Alternatively, the general stochastic epidemic model can be approximated by the stochastic differential equation (see e.g. [41]):

$$dX(t) = f(X(t))dt + G(X(t))dW(t), \tag{8}$$

where the random variables $S(t)$ and $I_f(t)$ are continuous,

$$X(t) = \begin{pmatrix} S(t) \\ [2pt] I_f(t) \end{pmatrix}, \; f = \begin{pmatrix} -\beta \frac{SI_f}{N} \\ [2pt] \beta \frac{SI_f}{N} - \gamma I_f \end{pmatrix},$$

$$G = \begin{pmatrix} -\sqrt{\beta \frac{SI_f}{N}} & 0 \\ [2pt] \sqrt{\beta \frac{SI_f}{N}} & -\sqrt{\gamma I_f} \end{pmatrix}, \tag{9}$$

and $W = (W_1, W_2)'$ is a vector of two independent Wiener processes, i.e., $W_i(t) \sim \mathcal{N}(0, t)$. Given (8), the stochastic SIR model can be written as:

$$Y_t = X_t B + \epsilon_t, \tag{10}$$

where

$$Y_t = \begin{pmatrix} \frac{\Delta I(t)}{1-u(t+1)} \\ \Delta R(t) \end{pmatrix}, B = \begin{pmatrix} \beta \\ \gamma \end{pmatrix},$$

$$X_t = \begin{pmatrix} \frac{S(t)}{N} I_f(t) & -I_f(t) \\ 0 & I_f(t) \end{pmatrix},$$

$$\epsilon_t = \begin{pmatrix} \sqrt{\frac{\beta S(t)}{N} I_f(t)} \Delta W_1(t) - \sqrt{\gamma I_f(t)} \Delta W_2(t) \\ \sqrt{\gamma I_f(t)} \Delta W_2(t) \end{pmatrix},$$

with the increments $\Delta W_1(t)$ and $\Delta W_2(t)$ being two independent normal random variables, i.e., $\Delta W_i(t) \sim \mathcal{N}(0, \Delta t)$. It can be seen that the resulting regression model, based on the discrete analogue of (10), will have an error term exhibiting temporal correlation, driven by $I_f(t)$ and $R(t)$. An examination of the temporal correlation patterns in COVID-19 data (see left two panels in Figs. 6 and 14 in the supplementary material) supports this finding. To that end, we model the error process as a Vector Auto-Regressive (VAR) with lag p (Model 3). The corresponding temporal correlation plots for the residuals after inclusion of the VAR(p) component are depicted in the right two panels in Figs. 6 and 14 in the supplementary material and clearly show the importance of considering such an error structure. The piecewise stationary SIR model with spatial heterogeneity and a VAR(p) error process is given by

$$Y_t = \sum_{j=1}^{m_0+1} X_t B^{(j)} \mathbb{1}_{\{t_{j-1} \le t < t_j\}} + \alpha Z_t + \phi_1 \varepsilon_{t-1} + \ldots + \phi_p \varepsilon_{t-p} + e_t, t = 1, \ldots, T-1, \tag{11}$$

where $\{t_1, \ldots, t_{m_0}\}$ are unknown m_0 "change points" such that the transmission and recovery rates exhibit a change from $B^{(j)} = (\beta^{(j)}, \gamma^{(j)})'$ to $B^{(j+1)} = (\beta^{(j+1)}, \gamma^{(j+1)})'$ at time point t_j, while it remains fixed until the next break point. Hence, these break points divide the time series data into stationary

segments. Moreover, $Z_t = \sum_{j=1}^{q} \omega_j (\frac{\Delta I^j(t-1)}{N^j(1-u(t))}, \frac{\Delta R^j(t-1)}{N^j})'$ is the weighted average of spatial effect over the neighboring regions at time t where N^j denotes the total population in neighboring region j; α is a spatial effect parameter; ω_j's are spatial weights such that $\sum_{j=1}^{q} \omega_j = 1$. The latter two parameters capture inter-region mobility patterns. Finally, e_t is white noise with mean 0 and variance σ^2, and $\phi_1, \ldots, \phi_p$ are the corresponding autoregressive parameters. In the sequel, model (11) with $\alpha = \phi_1 = \cdots = \phi_p = 0$ is considered as Model 1 (piecewise SIR), the case of only restricting $\phi_1 = \cdots = \phi_p = 0$ is considered as Model 2 (piecewise SIR with spatial effects) while the full model with no constrains is considered as Model 3. Notice that the number of change points m_0 and their locations are unknown and must be estimated from the data together with all other model parameters including $B^{(j)}$'s, α, and $\phi_1, \ldots, \phi_p$. A brief discussion of the proposed algorithm to perform all such estimations is presented next.

C. Algorithm

The estimation of the model parameters is accomplished in the following three steps: *Step 1:* Fit Model 1 for each region of interest to obtain the change points; *Step 2:* Obtain the transmission and recovery rates and spatial effect as in Model 2. *Step 3:* Compute the residuals ($\widehat{\epsilon}_t$) from Step 2 and fit a VAR model to them, see [36]. The rationale behind this step-wise algorithm is that assuming that the influence of the spatial effect component Z_t is small, we can use Model 1 for each region of interest to estimate both the change points and the corresponding transmission and recovery rates. Denoting the final estimated change points by Model 1 as $\widetilde{\mathcal{A}}_n^f = \{\widetilde{t}_1^f, \ldots, \widetilde{t}_{\widetilde{m}^f}^f\}$, segment-specific transmission and recovery rates combined with an overall spatial effect can be readily estimated using least squares applied to augmented linear model which includes all segments concatenated to each other at time points $\widetilde{t}_j^f$'s and the spatial effect. Finally, the residuals of this augmented linear model utilizing the least squares estimates can be computed and additional least squares estimates on the residuals with its previous values in the design matrix can yield to estimates of autoregressive parameters. The difficult part of the algorithm is to estimate the number and locations of break points. Details of the algorithm are presented in the Appendix A while a brief summary is provided next. The first step of the algorithm aims to select candidate change points $\widehat{A}_n$ among blocks by solving a block fused lasso problem. The estimated change points obtained by the block fused lasso step includes all points with non-zero estimated parameters, which leads to overestimating the number of the true change points in the model. Nevertheless, the block fused lasso parameter estimates enjoy a prediction consistency property, which implies that the prediction error converges to zero with high probability as $n \to +\infty$. This result is stated and proved (see Theorem 1 in the Appendix B) under some mild conditions on the behaviour of the tail distributions of error terms. A hard-thresholding step is then added to reduce the over-selection problem from the fused lasso step by "thinning out" redundant change points exhibiting small changes in the estimated coefficients. After the hard-thresholding step, those candidate change points located far from any true change points will be eliminated when the block size is appropriate. On the other hand, there may be more than one selected change points remaining in small neighborhoods of each true change point. To remedy this issue, the remaining estimated change points are clustered while in each cluster, an exhaustive search examines every time point inside the neighborhood search region based on the cluster of candidate change points and selects the best time point as the final estimated change point.

III. Simulation Studies

We evaluate the performance of the proposed models on their predictive accuracy, change point detection and parameter estimation. We consider three simulation scenarios (see additional simulation settings in Section II in the supplementary material). The details of the simulation settings for each scenario are explained in Section III-A. All results are averaged over 100 random replicates.

We assess the results for the three models presented: Model 1, the piecewise stationary SIR model; Model 2, the piecewise stationary SIR model with spatial effect; and Model 3, the piecewise stationary SIR model with spatial effect and a VAR(p) error process. The out-of-sample Mean Relative Prediction Error (MRPE) is used as the performance criterion defined as:

$$\text{MRPE}(I) = \frac{1}{n^{test}} \sum_{t=T+1}^{T+n^{test}} \left| \frac{\widehat{I}(t) - I(t)}{I(t)} \right|, \tag{12}$$

where n^{test} is the number of time points for prediction, $\widehat{I}(t)$ is the predicted count of infected cases at time t, and $I(t)$ the observed one. The MRPE of $R(t)$ can be obtained by respectively replacing the $\widehat{I}(t)$ and $I(t)$ with $\widehat{R}(t)$ and $R(t)$. The predicted number of infected cases and recovered cases are defined as

$$\begin{aligned} \widehat{I}(t) &= I(t-1) + \widehat{\Delta I}(t-1), \\ \widehat{R}(t) &= R(t-1) + \widehat{\Delta R}(t-1), \end{aligned} \tag{13}$$

for all $t = T+1, \ldots, T+n^{test}$. For change point detection, we report the locations of the estimated change points and the percentage of replicates that correctly identifies the change point. This percentage is calculated as the proportion of replicates, where the estimated change points are close to each of the true break points. Specifically, to compute the selection rate, a selected break point is counted as a "success" for the j-th true break point, t_j, if it falls in the interval $[t_j - \frac{t_j - t_{j-1}}{5}, t_j + \frac{t_{j+1} - t_j}{5}]$, $j = 1, \ldots, m_0$. We also report the mean and standard deviation of estimated parameters for each models. All results are reported in Table I.

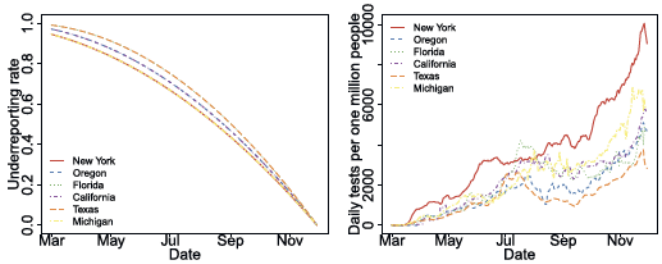
A. Simulation Scenarios

We consider three different simulation settings. The SIR model's coefficients and under-reporting functions are depicted in Fig. 1 in the Section II.

1) Simulation Scenario A (Model 3 with no under-reporting): The data are generated based on (11) with piecewise constant transmission and recovery rates. We set the number of time points $T = 200$, $m_0 = 1$, the change point $t_1 = \lfloor \frac{T}{2} \rfloor = 100$,

TABLE I
SIMULATION RESULTS INCLUDING SELECTION RATE, ESTIMATED PARAMETERS, AND OUT-OF-SAMPLE MEAN RELATIVE PREDICTION ERROR (MRPE). NOTE THAT IR INCLUDES BOTH $I(t)$ AND $R(t)$

	change point	truth	mean	std	selection rate
Scenario A	1	0.5	0.5	0	1
Scenario B	1	0.4	0.4	0	1
	2	0.8	0.8	4e-04	1
Scenario C	1	0.5	0.5	0	1
	parameter	true value	mean	std	
Scenario A	β_1	0.1	0.1	3e-04	
	β_2	0.05	0.05	1e-04	
	γ_1	0.04	0.04	1e-04	
	γ_2	0.04	0.04	1e-04	
	α	1	0.9945	0.0318	
Scenario B	β_1	0.1	0.1001	0.0078	
	β_2	0.05	0.0488	0.0106	
	β_3	0.1	0.0986	0.0087	
	γ_1	0.04	0.0392	0.0074	
	γ_2	0.06	0.0589	0.0114	
	γ_3	0.04	0.0391	0.0054	
	a	0.05	0.0515	0.016	
Scenario C	β_1	0.1	0.0904	0.016	
	β_2	0.05	0.0427	0.0123	
	γ_1	0.04	0.0336	0.0108	
	γ_2	0.04	0.0341	0.0099	
	α	1	1.84	1.6373	
	a	0.5	0.3905	0.1977	
	Model	MRPE(IR)	MRPE($I(t)$)	MRPE($R(t)$)	
Scenario A	Model 1	0.000252	0.000362	0.000142	
	Model 2	2.6e-05	3.9e-05	1.3e-05	
	Model 3	2.4e-05	3.7e-05	1.2e-05	
Scenario B	Model 1	0.00415	0.005888	0.002413	
Scenario C	Model 3 (Transformed by $u(t)$)	0.000162	0.000175	0.000149	
	Model 3 (Not transformed)	0.000911	0.000829	0.000993	



(a) under-reporting function (b) daily testing

Fig. 1. (left panel) Estimated under-reporting rate functions for all six states; (right panel) 7-day moving average of daily testing for all six states. (a) under-reporting function (b) daily testing.

$\beta^{(1)} = 0.10$, $\beta^{(2)} = 0.05$, $\gamma^{(1)} = 0.04$ and $\gamma^{(2)} = 0.04$. For the spatial effect, we set $\alpha = 1$, $\beta_s(t) = 0.10 - \frac{0.05t}{T-1}$, $\gamma_s(t) = 0.04$, $t = 1, \ldots, T-1$. We first generate the spatial effect data from SIR model in (11) with parameter $\beta_s(t)$ and $\gamma_s(t)$ and generate the error term by VAR(1) model with the covariance matrix of the noise process $\Sigma_\varepsilon = 0.1\mathbb{I}_2$, where $\mathbb{I}_2$ is the two-dimensional identity matrix. By plugging in the spatial effect data and error term data, we generate the dataset of the response variable Y_t from (11). The autoregressive coefficient matrix has entries 0.8, 0, 0.2, 0.7 from top left to bottom right. We assume no under-reporting issue in this scenario, i.e., $u(t) = 0$, hence $\Delta I(t) = \Delta I_f(t)$, for $t = 1, \ldots, T-1$.

2) Simulation Scenario B (Model 1 with exponentially decreasing under-reporting rate): In this scenario, we set the number of time points $T = 250$, $m_0 = 2$, the change points $t_1 = 100$ and $t_2 = 200$. We choose $\beta^{(1)} = 0.10$, $\beta^{(2)} = 0.05$, $\beta^{(3)} = 0.10$, $\gamma^{(1)} = 0.04$, $\gamma^{(2)} = 0.06$, $\gamma^{(3)} = 0.04$. Results are based on data generated from the SIR model in (11) with $\beta(t) \sim \text{Lognormal}(\sum_{j=1}^{m_0+1} \beta^{(j)} \mathbb{1}_{\{t_{j-1} \le t < t_j\}}, 0.005)$ and $\gamma(t) \sim \text{Lognormal}(\sum_{j=1}^{m_0+1} \gamma^{(j)} \mathbb{1}_{\{t_{j-1} \le t < t_j\}}, 0.005)$. The under-reporting rate is chosen to change over time. Specifically, we set the under-reporting rate $u(t) = 1 - \frac{1}{1+be^{-a(t-1)}}$, $t = 1, \ldots, T$, with $a = 0.05$ and $b = 10$.

3) Simulation Scenario C (Model 3 with quadratically decreasing under-reporting rate): The data are generated based on (11) with piecewise constant transmission and recovery rates. All the settings are exactly the same as those in scenario A except for the under-reporting rate. In this scenario, we set under-reporting rate $u(t) = 1 - (\frac{t+aT}{(1+a)T})^2$, $t = 1, \ldots, T$, with $a = 0.5$.

B. Simulation Results

The mean and standard deviation of the location of selected change point, relative to the the number of time points T – i.e., $\widehat{t}_1^f/T$ – for all simulation scenarios are summarized in Table I. The results clearly indicate that, in the piecewise constant setting, our procedure accurately detects the location of change points. The results of the estimated transmission rate $\widehat{\beta}$, recovery rate $\widehat{\gamma}$, spatial effect $\widehat{\alpha}$ and parameter of the under-reporting rate function $\widehat{a}$ suggest that our procedure produces accurate estimates of the parameters, under the various under-reporting function settings (b is assumed to be known). We generate additional 20 days worth of data to measure the prediction performance. The MRPE results for

$I(t)$ and $R(t)$ are provided in Table I. The results in scenario A indicate that adding the spatial effect can significantly improve the prediction, when the spatial component influences the individual data series. The results in scenario C indicate that adding the under-reporting function $u(t)$ can significantly improve the prediction, when there is under-reporting in the data series.

IV. Application to State and County Level COVID-19 Data in the U.S.

A. Data Description

The COVID-19 data used in this study are obtained from [42]. The curated data and code used in the analysis are available at the authors' GitHub repository.[4] The analysis is performed both at the state and county level and the raw data include both cases and deaths, as reported by state and local health departments and compiled by the NY Times. However, due to lack of complete information on recovered individuals (which is an important covariate in the models considered, but the daily number of recovered cases is only reported at the national level [43]), we calculate the number of recovered cases for each region (state/county) as follows: the number of deaths in the region, multiplied by the nationwide cumulative recovered cases and divided by the nationwide deaths. Specifically, we assume that the recovery versus deceased ratio for each state/county is fixed, and can be well approximated by the nationwide recovery-to-death ratio. As coronavirus infections increase, while laboratory testing faces capacity constraints, reporting only confirmed cases and deaths leads to (possibly severe) under-estimation of the disease's impact. On April 14, 2020, CDC advised states to count both confirmed and probable cases and deaths. As more states and localities did so since then, in this study we focus on the combined cases, which include both confirmed and probable cases. The populations of states and counties are obtained from [44]. Further, to decide which neighboring states/counties to include in Model 2, their distance to the target state/county of interest is used. The latter is obtained from the [45]. In the results presented, we define regions within 500 miles for states and 100 miles for counties/cities as neighboring ones in Models 2 and 3. For those areas with a large number of neighboring regions, such as New York state, we only consider the top five regions with the smallest distances. We assume the probability rate of becoming susceptible again after having recovered from the infection to be 0.5%. The reinfection rate in the short run ($\sim$6 months) is believed to be very low. Some evidence from health care workers (median age 38 years) estimates it at 0.2% [46]. Hence, the selected one seems a reasonable upper bound for the task at hand. Note that small variation of this rate did not impact the results.

We analyzed the daily count of COVID-19 cases at the state-level from March 1, 2020, to March 31, 2021 for the selected five states (NY, OR, FL, CA, TX), and at the county-level from the first day after March 1, 2020, when the region records at least one positive COVID-19 case to March 31, 2021. In addition, we analyzed the COVID-19 cases in the state of Michigan from March 1, 2020, to May 15, 2021. In particular, the sample size n for the five states presented next (NY, OR, FL, CA, TX) is $n = 395$ while for state of Michigan, Riverside County (CA) and Santa Barbara County (CA), $n = 432, 389, 381$, respectively. For the under-reporting rate function $u(t)$, both quadratic function -$u(t) = 1 - (\frac{t+aT}{(1+a)T})^2$- and exponential function -$u(t) = 1 - \frac{1}{1+be^{-a(t-1)}}$- are considered. The quadratic function achieved better performance in change point selection in the real data application (change point detection results using the exponential function are presented in Table XI in the Supplement). Therefore, all presented results in this Section are based on the quadratic function. Finally, to estimate the under-reporting parameter a, we perform a grid search within the interval [0.1,0.3]. The main reason for selecting this interval is that it matches with around 90% of COVID-19 under-reporting rate at the beginning of the pandemic as investigated and reported in [34], [35]. Note that we also assume that the COVID-19 under-reporting rate after December is very low as most of the regions built-up their testing capacity. Therefore, we set $u(t) = 1$ after December 2020.

Most of the states were selected due to being severely affected for a certain period of time during the course of COVID-19. The remaining regions illustrate interesting patterns gleaned from the proposed models. Let $I(t)$ and $R(t)$ denote the number of infected and recovered individuals (cases) on day t. Day 1 refers to the first day after March for which the region records at least one positive COVID-19 case. Fig. 2 in the Supplement depicts the actual case numbers $I(t)$ and $R(t)$ in the six states considered.

B. Results for Selected U.S. States

The various models considered are applied on scaled versions (divided by their standard deviations) of the predictors matrix X_t and the response vector Y_t. For Model 2, we consider four different types of weights: equal weights (Model 2.1), distance-based weights (Model 2.2), similarity-based weights (Models 2.3 and 2.4). In Models 2.1 and 2.2, the neighboring regions are selected based on distance. For states, a threshold of 500 miles is used and the resulting neighbors are displayed in Table VII in the supplementary material. When the spatial resolution is high (county level aggregated data), neighboring regions may exhibit similar patterns in terms of the evolution of transmission and recovery rates. Therefore, constructing weight matrices based on distance is meaningful as it is a common practice in spatial statistics [47]. Such spatial smoothing through proper weight matrices is especially helpful in increasing the statistical power through increasing the sample size, thus yielding more accurate predictions.

However, the evolution of COVID-19 may exhibit different patterns across neighboring states due to the coarse spatial resolution. Thus, defining weight matrices based on distance may not be ideal. Hence, Models 2.3 and 2.4 select regions based on similarities of infected/recovered cases. Similarity between the region of interest and the j-th potential similar region is

[4][Online]. Available: https://github.com/ybai69/COVID-19-Change-Point-Detection

TABLE II
NEIGHBORING STATES AND CITIES/COUNTIES BY SIMILARITY SCORE (FOR MODEL 2.3). (STATES: SELECTED FROM ALL STATES IN THE COUNTRY; COUNTIES: IN THE SAME STATE)

Region	Neighboring Regions
New York	Massachusetts, New Jersey, D.C., Pennsylvania, Michigan
Oregon	Maine, Washington, Vermont, West Virginia, New Hampshire
Florida	South Carolina, Nevada, Texas, Alabama, Mississippi
California	North Carolina, Texas, Nevada, West Virginia, Mississippi
Texas	California, North Carolina, South Carolina, Nevada, Alabama
Riverside	Orange, Los Angeles, Ventura, San Diego, Monterey
Santa Barbara	San Diego, Ventura, Orange, San Francisco, Contra Costa

defined as

$$s_j = \sqrt{\sum_{t=1}^{T-1}\left(\frac{\Delta I_f(t)}{N} - \frac{\Delta I_f^j(t)}{N^j}\right)^2 + \sum_{t=1}^{T-1}\left(\frac{\Delta R(t)}{N} - 1\frac{\Delta R^j(t)}{N^j}\right)^2}, \tag{14}$$

where $\Delta I_f(t) = \frac{\Delta I(t)}{(1-u(t+1))}$ and $\Delta I_f^j(t) = \frac{\Delta I^j(t)}{(1-u(t+1))}$. Model 2.3 uses the top five regions with the smallest similarity score, while Model 2.4 uses all states in the country. For states, the resulting neighbors for Model 2.3 are displayed in Table II.

In the equal weight setting (Model 2.1), $\omega_j = 1/q$ for any $j = 1, \ldots, q$. In both distance-based weight and similarity-based weight settings, power distance weights are used, wherein the weight is a function of the distance/similarity to the neighboring region $\omega_j = \frac{1}{d_j}$, $\omega_j = \frac{1}{s_j}$ where d_j is the distance score and s_j is the similarity score for the j-th region. Under the constraint that $\sum_{j=1}^{q} \omega_j = 1$, we obtain the normalized weights as $\omega_j = \frac{d_j^{-1}}{\sum_{k=1}^{q} d_k^{-1}}$, $\omega_j = \frac{s_j^{-1}}{\sum_{k=1}^{q} s_k^{-1}}$.

Before applying Model 3, we compare the in-sample and out-of-sample MRPEs (defined in Section III) in all four variants of Model 2 and select the parameter values estimated by the best-performing model. In subsequent analysis, the results from Model 2.3 are reported, since it proved to be the best performing one.

As expected, change points detected for state data are related to "stay-at-home" orders, or phased reopening dates issued by state governments. We define the reopening date as the time when either the "stay-at-home" order expired or state governments explicitly lifted orders and allowed (selected or even all) businesses to reopen [48]. The "stay-at-home" and reopening dates for all states are shown in Table III.

In Model 1, a change point is detected from March to April for all five states: New York, Oregon, Florida, California and Texas. These change points coincide with the onset of "stay-at-home" orders and correspond to a significant decrease in the transmission rate. The first change points detected are around two weeks after the state's Governors signed a statewide "stay-at-home" order (three weeks for CA), which is consistent with the fact that COVID-19 symptoms develop 2 days to 2 weeks following exposure to the virus. As can be seen from Fig. 4 in the supplementary material, in addition to the downward trend after lockdowns have been put in place, Oregon, Florida and Texas have a clear upward trend after state reopenings began in May. The model detects a change point in June for these states, which relate to their reopenings.

Note that the restriction in either phase 2 reopening plan in Florida and the phase 3 reopening plan in Texas are quite similar in terms of restaurants, bars, and entertainment businesses. Starting June 5, restaurants and bars in Florida could increase their indoor seating to 50% capacity. Movie theaters, concert venues, arcades, and other entertainment businesses could also open at 50% capacity. Starting June 3, all businesses in Texas could expand their occupancy to 50% with certain exceptions. Moreover, bars could increase their capacity to 50% as long as patrons are seated.

Assuming that Florida had not begun the phase 2 reopening plan on June 5, our model predicts 36,626 infected cases by June 12 while the actual number of infected cases is 39,327 (7.3% higher). Similarly, by June 19, our model predicts 41,728 infected cases, while the actual number of infected cases is 55,607 (33.3% higher). Similarly, suppose that Texas had not begun the phase 3 reopening plan on June 3, then by June 17, our model predicts 74,204 infected cases, while the actual number of infected cases is 76,377 (1% higher).

In July, many states paused plans to reopen, amid rising infected case counts.[5] These pausing actions effectively slowed the spread of COVID-19, as can be clearly seen in the downward trend of the transmission rate in July and August in Oregon, Florida, California and Texas. Interestingly, a change point in July is detected in Oregon, Florida and Texas, and a change point in August is detected in California, mainly related to this pausing.

The left panel of Fig. 1 depicts the estimated function $u(t)$ while the right panel displays the performed daily test normalized by the populations for the six states under consideration. Comparing these two plots indicate that states at which the testing capacity has been limited, the under-reporting rate was estimated higher and vise versa. For example, New York had more daily tests based on its population compared with the other states, which coincides with its lower estimated under-reporting rate at the beginning while the estimated under-reporting rate is higher for Texas and Florida which could be due to limited capacity in daily testing at the beginning of pandemic for these states. Note that the under-reporting rate is estimated around 90% in March for all states. Such high under-reporting rates at the beginning of the pandemic are reported in other countries as well, such as China [49].

The observed and fitted number of infected cases are displayed in Fig. 2. Note that the fitted number of infected cases and recovered cases are defined as

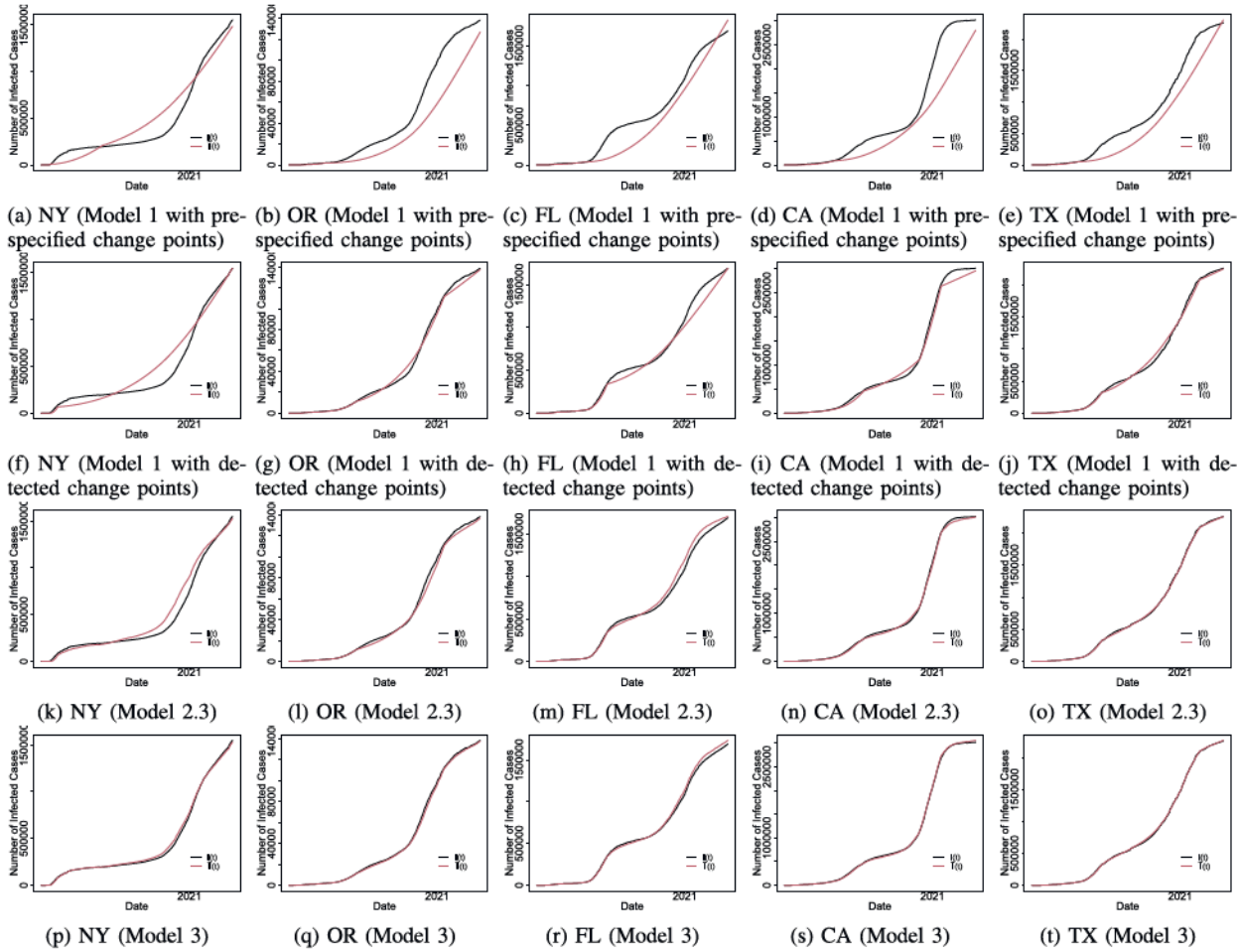
$$\widetilde{I}(t) = I(1) + \sum_{k=1}^{t-1} \widehat{\Delta I}(k), \quad \widetilde{R}(t) = R(1) + \sum_{k=1}^{t-1} \widehat{\Delta R}(k), \tag{15}$$

for all $t = 2, \ldots, T$. In summary, the piecewise constant SIR model (Model 1) with detected change points significantly improves the performance in prediction of the number of infected

[5][Online]. Available: https://www.usatoday.com/story/news/nation/2020/06/30/covid-cases-states-pausing-reopening-plans-list/3284513001/

TABLE III
STATEWIDE "STAY-AT-HOME" PLAN AND REOPENING PLAN BEGIN DATES, ALONG WITH THE DETECTED CHANGE POINTS (CPS) IN STATES AND COUNTIES/CITIES

Region	"Stay-at-home" plan	Reopening plan (statewide)	Detected change points
New York	March 22	July 6 (Phase 3)	April 04 2020
Oregon	March 23	May 15 (Phase 1)	April 06 2020, June 06 2020, July 18 2020, Jan 16 2021
Florida	April 3	June 5 (Phase 2)	April 13 2020, June 17 2020, July 25 2020
California	March 19	June 12 (Phase 2)	April 11 2020, April 29 2020, Aug 12 2020, Dec 06 2020, Jan 19 2021
Texas	April 2	June 3 (Phase 3)	April 18 2020, June 03 2020, June 15 2020, July 23 2020, Sep 20 2020, Jan 03 2021, Feb 10 2021
Michigan	March 24	June 8 (phase 5)	April 11 2020, June 04 2020, Nov 01 2020, Nov 21 2020, Dec 11 2020, March 21 2021
Riverside (CA)	March 19	June 12 (Phase 2)	April 17 2020, June 20 2020, July 26 2020, Nov 28 2020
Santa Barbara (CA)	March 19	June 12 (Phase 2)	April 15 2020, May 11 2020, June 10 2020, July 25 2020, Dec 27 2020



(a) NY (Model 1 with pre-specified change points) (b) OR (Model 1 with pre-specified change points) (c) FL (Model 1 with pre-specified change points) (d) CA (Model 1 with pre-specified change points) (e) TX (Model 1 with pre-specified change points)

(f) NY (Model 1 with detected change points) (g) OR (Model 1 with detected change points) (h) FL (Model 1 with detected change points) (i) CA (Model 1 with detected change points) (j) TX (Model 1 with detected change points)

(k) NY (Model 2.3) (l) OR (Model 2.3) (m) FL (Model 2.3) (n) CA (Model 2.3) (o) TX (Model 2.3)

(p) NY (Model 3) (q) OR (Model 3) (r) FL (Model 3) (s) CA (Model 3) (t) TX (Model 3)

Fig. 2. Observed (black) and fitted (red) number of infected cases estimated by three models in selected five states. The pre-specified change points are two fixed change points derived from the statewide lockdown date and reopening date. (a) NY (Model 1 with pre-specified change points). (b) OR (Model 1 with pre-specified change points). (c) FL (Model 1 with pre-specified change points). (d) CA (Model 1 with pre-specified change points). (e) TX (Model 1 with pre-specified change points). (f) NY (Model 1 with detected change points). (g) OR (Model 1 with detected change points). (h) FL (Model 1 with detected change points). (i) CA (Model 1 with detected change points). (j) TX (Model 1 with detected change points). (k) NY (Model 2.3). (l) OR (Model 2.3). (m) FL (Model 2.3). (n) CA (Model 2.3). (o) TX (Model 2.3). (p) NY (Model 3). (q) OR (Model 3). (r) FL (Model 3). (s) CA (Model 3). (t) TX (Model 3).

cases compared with the piecewise constant SIR model with pre-specified change points. Pre-specified change points are defined as stay-at-home or reopening order dates for each region. It can be seen from the first two rows in Fig. 2 that estimating when break points occurred using the developed algorithm improves the fit significantly, especially for states in which multiple change points are selected such as Oregon, California and Texas. Model 2.3 further improves the fit of the data for Florida and California, due to the addition of a spatial smoothing effect. Finally, Model 3 provides further improvements, in particular for New York, which justifies empirically the use of hybrid modeling to analyze regional transmission dynamics of COVID-19. To determine the significance level of the spatial effect in Model 2, we provide the estimate, p-value, and 95% confidence intervals for the parameter α in Table VIII in the Supplement. We find that the influence of the infected or recovered cases in adjacent states is statistically significant (p-value < 0.05) for all states.

Next, we assess the prediction performance for the three proposed models. The out-of-sample MRPE is used as the performance measurement, given by (12). We set the last two

TABLE IV
OUT-OF-SAMPLE MEAN RELATIVE PREDICTION ERROR (MRPE) OF $I(t)$. THE NUMBER IN BETWEEN THE BRACKETS STANDS FOR THE STANDARD DEVIATION OF THE RELATIVE PREDICTION ERROR

	NY MRPE(I)	OR MRPE(I)	FL MRPE(I)	CA MRPE(I)	TX MRPE(I)
Model 1	0.0011(7e-04)	9e-04(7e-04)	0.0016(0.001)	9e-04(2e-04)	7e-04(4e-04)
Model 2.1	0.0011(7e-04)	0.001(8e-04)	0.0018(7e-04)	5e-04(2e-04)	8e-04(6e-04)
Model 2.2	0.0011(7e-04)	0.001(7e-04)	0.0017(7e-04)	6e-04(2e-04)	9e-04(8e-04)
Model 2.3	**4e-04**(5e-04)	**7e-04**(6e-04)	0.002(8e-04)	**4e-04**(3e-04)	**6e-04**(4e-04)
Model 2.4	0.0016(8e-04)	0.001(8e-04)	0.001(7e-04)	7e-04(4e-04)	8e-04(5e-04)
Model 3	**4e-04**(4e-04)	**7e-04**(6e-04)	0.001(8e-04)	**4e-04**(3e-04)	**6e-04**(4e-04)
eSIR	0.0062(9e-04)	0.0076(0.0013)	0.0066(0.001)	0.0106(6e-04)	0.0095(0.001)
ANN	5e-04(4e-04)	0.0013(0.0013)	8e-04(8e-04)	**4e-04**(3e-04)	8e-04(4e-04)
LSTM	7e-04(4e-04)	0.0012(0.001)	**7e-04**(8e-04)	6e-04(8e-04)	**6e-04**(5e-04)
LSTM architecture (layer,neurons)	(1, 10)	(3, 10)	(1, 50)	(1, 100)	(3, 50)

TABLE V
OUT-OF-SAMPLE MEAN RELATIVE PREDICTION ERROR (MRPE) OF $R(t)$. THE NUMBER IN BETWEEN THE BRACKETS STANDS FOR THE STANDARD DEVIATION OF THE RELATIVE PREDICTION ERROR

	NY MRPE(R)	OR MRPE(R)	FL MRPE(R)	CA MRPE(R)	TX MRPE(R)
Model 1	0.0019(3e-04)	0.0028(0.0011)	0.0028(7e-04)	0.0042(8e-04)	0.002(7e-04)
Model 2.1	4e-04(3e-04)	**0.0017**(0.002)	0.001(6e-04)	0.0029(7e-04)	0.0025(0.0029)
Model 2.2	4e-04(3e-04)	0.0018(0.0019)	0.001(6e-04)	0.0029(7e-04)	0.0028(0.0042)
Model 2.3	4e-04(2e-04)	0.0022(0.0019)	0.0012(8e-04)	0.0022(7e-04)	0.001(7e-04)
Model 2.4	5e-04(8e-04)	0.002(0.0024)	9e-04(0.001)	0.0029(0.001)	0.0015(7e-04)
Model 3	4e-04(2e-04)	0.0022(0.0018)	0.001(6e-04)	0.0022(7e-04)	0.001(7e-04)
eSIR	0.013(0.0011)	0.014(0.0031)	0.0103(0.0012)	0.0117(0.0012)	0.0113(0.001)
ANN	**3e-04**(2e-04)	**0.0017**(0.0021)	8e-04(4e-04)	**0.0012**(0.001)	8e-04(5e-04)
LSTM	**3e-04**(3e-04)	0.0025(0.0021)	**6e-04**(4e-04)	0.0027(0.0036)	**6e-04**(6e-04)
LSTM architecture (layer,neurons)	(1, 10)	(3, 10)	(1, 50)	(1, 100)	(3, 50)

TABLE VI
OUT-OF-SAMPLE MEAN RELATIVE PREDICTION ERROR (MRPE) OF $I(t)$ AND $R(t)$ FOR SELECTED REGIONS

	New York		California		Riverside		Santa Barbara	
	I	R	I	R	I	R	I	R
Model 1	0.0011	0.0019	9e-04	0.0042	0.0056	0.0036	0.0037	0.0071
Model 2.1	0.0011	4e-04	5e-04	0.0029	0.0043	0.0028	0.0016	0.0039
Model 2.2	0.0011	4e-04	6e-04	0.0029	0.0045	0.0027	0.0014	0.0032
Model 2.3	4e-04	4e-04	4e-04	0.0022	0.0014	0.0025	0.0014	0.0028
Model 2.4	0.0016	5e-04	7e-04	0.0029	0.0013	0.0027	0.0017	0.0033
Model 3	4e-04	4e-04	4e-04	0.0022	0.0012	0.0016	0.0014	0.0028

weeks of our observation period as the testing period and use the remaining time points for training the model. Note that predicted number of infected and recovered cases are defined by (13).

The results of out-of-sample MRPE of $I(t)$ and $R(t)$ in the selected regions are reported in Table VI. The calculated MRPEs of $I(t)$ show that Model 3 which includes spatial effects and VAR temporal component outperforms the other models. Spatial smoothing itself (Model 2) reduces the prediction error significantly in some states. For example, in New York, the spatial smoothing reduced the MRPE (I) by 64% when using Model 2.3 (similarity-based weight). Finally, the reduction in MRPEs using Model 3 justifies the presence of the VAR component in the modeling framework. Additional results related to the VAR component (including the estimated auto-regressive parameters) are reported in Section V in the Supplement.

We also compare the developed model and associated methodology with the extended SIR, the ANN and the LSTM based models that were proposed in [17], [28], and [29], respectively. The extended SIR model was trained using the "eSIR" package in the `R` programming language. The transmission rate modifier $\pi(t)$ was specified according to actual interventions at different times and regions as described in [17]. The ANN was trained using the "nnfor" package in the `R` programming language. It comprises of 3 hidden layers with 10 nodes in each and a linear output activation function, which is the exact architecture in [28]. The number of repetitions for this algorithm was set to be 20 for $I(t)$ and 10 for $R(t)$. The LSTM architecture was implemented in `PyTorch`. The [29] did not specify the network architecture setting (number of layers, number of neurons). Thus, we performed a grid search over number of layers ranging from 1 to 3 with number of neurons as 10, 50, and 100. The best architecture in terms of minimizing the prediction error in the validation data is selected as the optimal LSTM architecture. Note that the prediction results with different network architecture in the LSTM were very similar. The selected number of layers and number of neurons based on grid search and corresponding prediction error for each region are also provided in Tables IV and V.

The results of out-of-sample MRPE for $I(t)$ and $R(t)$ in the five states under consideration together with their sample standard deviations in the bracket are reported in Tables IV and V, respectively. The proposed method clearly outperforms the extended SIR (eSIR) model across all five states for both $I(t)$ and $R(t)$. Further, it broadly matches the performance of ANN and LSTM for most states. Further, note that the proposed model is easy to interpret since its key parameters (infection and recovery rates) are routinely used by policy makers (see also the discussion in Section IV-C).

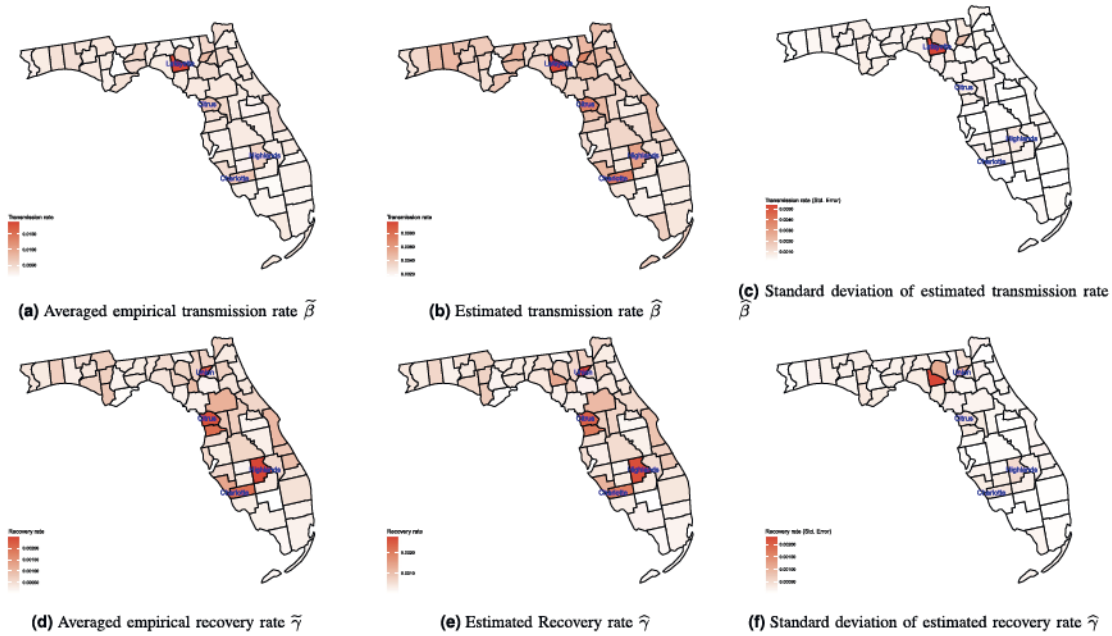
C. Results for Selected U.S. Counties

We worked on nine counties/cities. Due to limited space, results for two counties in the state of California (Riverside and Santa Barbara) are presented here while rest are described in Section IV in the supplementary materials. For determining their neighbors, a threshold of 100 miles is used. Model 2.3 uses the top five counties in the corresponding state with the smallest similarity score, while Model 2.4 uses all counties in the given state. The resulting neighbors for Model 2.3 are displayed in Table II. The statewide and countywide policy start dates and the detected change points are shown in Table III.

In December, Southern California was experiencing a fast and sustained outbreak, believed to be driven by a new strain designated as CAL.20 C.[6] To that end, we analyzed the daily count of cases in Riverside and Santa Barbara counties in CA. As seen from Fig. 11 in the supplementary material, three of the detected change points occurred on April 17 2020, July 20 2020, and July 26 2020 in Riverside County. The first one can be related to the decreased transmission rate in April mainly caused by the statewide lockdown, while the July ones could be due to the pause of reopening to halt the spread of COVID-19. Similarly, five change points are detected in Santa Barbara County. The first four change points can also be related to the lockdown, reopening and pause of reopening. An additional change point in Riverside County and Santa Barbara County is detected on November 28 and December 27, respectively, which may be driven by the new CAL.20 C variant.

We also provide the out-of-sample MRPE of $I(t)$ and $R(t)$ of selected counties in Table VI. The MRPE of $I(t)$ results show that adding the spatial effect can significantly improve the MRPE of $I(t)$ in both the Riverside County and Santa Barbara County.

[6][Online]. Available: https://www.newswise.com/coronavirus/local-covid-19-strain-found-in-over-one-third-of-los-angeles-patients2/



(a) Averaged empirical transmission rate $\widetilde{\beta}$ **(b)** Estimated transmission rate $\widehat{\beta}$ **(c)** Standard deviation of estimated transmission rate $\widehat{\beta}$

(d) Averaged empirical recovery rate $\widetilde{\gamma}$ **(e)** Estimated Recovery rate $\widehat{\gamma}$ **(f)** Standard deviation of estimated recovery rate $\widehat{\gamma}$

Fig. 3. Heatmap of transmission rate and recovery rate in Florida (county-level). (a) Averaged empirical transmission rate $\widetilde{\beta}$ (b) Estimated transmission rate $\widehat{\beta}$ (c) Standard deviation of estimated transmission rate $\widehat{\beta}$ (d) Averaged empirical recovery rate $\widetilde{\gamma}$ (e) Estimated Recovery rate $\widehat{\gamma}$ (f) Standard deviation of estimated recovery rate $\widehat{\gamma}$.

Adding the VAR(p) (Model 3) performs the best in Riverside County. In Riverside County, the spatial smoothing reduced the MRPE (I) by 76% when using Model 2.4 (similarity-based weight) while in Santa Barbara County, the spatial smoothing reduced the MRPE (I) by approximately 62% when using Model 2.3 (similarity-based weight). The MRPE of $R(t)$ results show that the piecewise constant model with spatial effect (Model 2) performs the best in Santa Barbara County while Adding the VAR(p) (Model 3) performs the best in Riverside County.

In Fig. 3, we provide heatmaps of the transmission and recovery rates in 67 counties in Florida, based on data from August 1st 2020 to December 1st 2020, a stable period in that state with no change point detected. The top left plot depicts the averaged empirical transmission rate $\widetilde{\beta}$, wherein it is calculated as the average $\widetilde{\beta}(t)$ within the given time interval, i.e., $\widetilde{\beta}(t) = (\frac{\Delta I(t)}{(1-u(t+1))} + \Delta R(t))/(\frac{S(t)}{N} I_f(t))$; the top middle plot, the heatmeap of the estimated transmission rate $\widehat{\beta}$ and finally the top right plot the heatmeap of the standard error of $\widehat{\beta}$ using linear regression analysis (for more details, see Section IV-A in the supplementary material). The bottom plots are the corresponding heatmaps of the recovery rate, wherein the bottom left plot depicts the averaged empirical recovery rate $\widetilde{\gamma}$, which is the average $\widetilde{\gamma}(t)$ within the given time interval, and $\widetilde{\gamma}(t) = \Delta R(t)/I_f(t)$. As shown in Fig. 3, the estimated rates are very close to the averaged empirical transmission and recovery rates. This point confirms the interpretability of the proposed hybrid modeling framework. During this time period, Lafayette County has the highest transmission rate, while Union County has the highest recovery rate. Moreover, Citrus, Charlotte and Highlands Counties have both (significantly) high transmission and recovery rates based on standard errors of the estimated parameters.

V. Concluding Remarks

COVID-19 has posed a number of challenges for modellers, both due to the lack of adequate data (especially early on in the course of the pandemic) and its characteristics (relative long period before emergence of symptoms compared to SARS and other respiratory viruses). A plethora of models -a number of them briefly summarized in the introductory section- were developed, most aiming to provide short and long term predictions of the spread of COVID-19. This work contributes to that goal by developing a hybrid model that enhances a piecewise stationary (mechanistic) SIR model with neighboring effects and temporal dependence to model the spread of COVID-19 at both state-level and county-level in the United States. The reasonable forecasts of Model 3 (including spatial effects and the VAR component) confirm the existence of spatial and temporal dependence among new daily cases which can not be accounted by the homogeneous deterministic SIR model. Further, the detection of change points in neighboring counties can provide insights into how the spread of COVID-19 impacted different communities at different points in time and also that of mitigation policies adopted by county (state) health administrators.

Appendix A
Algorithm Details

The key steps in our proposed detection strategy are summarized next while a summary is outlined in Algorithm 1 in the supplementary material. First, few notations are defined.

Notation: Denote the indicator function of a subset A as $\mathbb{1}_A$. $\mathbb{R}$ denotes the set of real numbers. For any vector $v \in \mathbb{R}^p$, we use $\|v\|_1$, $\|v\|_2$ and $\|v\|_\infty$ to denote $\sum_{i=1}^p |v_i|$, $\sqrt{\sum_{i=1}^p |v_i|^2}$ and $\max_{1\le i\le p}|v_i|$, respectively. The transpose of a matrix $\boldsymbol{A}$ is denoted by $\boldsymbol{A}'$.

Steps of the Proposed Algorithm.

Block Fused Lasso: the objective is to partition the observations into blocks of size b_n, wherein the model parameter B_t remains fixed within each block and select only those blocks for which the corresponding change in the parameter vector is much larger than the others. Specifically, let $n = T-1$ be the number of the times points for the response data Y_t and define a sequence of time points $1 = r_0 < r_1 < \ldots < r_{k_n} = n+1$ for block segmentation, such that $r_i - r_{i-1} = b_n$ for $i = 1, \ldots, k_n - 1, b_n \le r_{k_n} - r_{k_n-1} < 2b_n$, where $k_n = \lfloor \frac{n}{b_n} \rfloor$ is the total number of blocks. For ease of presentation, it is further assumed that n is divisible by b_n such that $r_i - r_{i-1} = b_n$ for all $i = 1, \ldots, k_n$. By partitioning the observations into blocks of size b_n and fixing the model parameters within each block, we set $\theta_1 = B^{(1)}$ and

$$\theta_i = \begin{cases} B^{(j+1)} - B^{(j)}, & \text{when } t_j \in [r_{i-1}, r_i) \text{ for some } j \\ 0, & \text{otherwise,} \end{cases}$$

for $i = 2, 3, \ldots, k_n$. Note that $\theta_i \neq 0$ for $i \ge 2$ implies that θ_i has at least one non-zero entry and hence a change in the parameters. Next, we formulate the following linear regression model in terms of $\Theta(k_n) = (\theta_1', \ldots, \theta_{k_n}')'$:

$$\underbrace{\begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \vdots \\ \boldsymbol{Y}_{k_n} \end{pmatrix}}_{\mathcal{Y}} = \underbrace{\begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{0} & \ldots & \ldots & \boldsymbol{0} \\ \boldsymbol{X}_2 & \boldsymbol{X}_2 & \ldots & \ldots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \boldsymbol{X}_{k_n} & \boldsymbol{X}_{k_n} & \ldots & \ldots & \boldsymbol{X}_{k_n} \end{pmatrix}}_{\mathcal{X}} \underbrace{\begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_{k_n} \end{pmatrix}}_{\Theta(k_n)} + \underbrace{\begin{pmatrix} \boldsymbol{\mathcal{E}}_1 \\ \boldsymbol{\mathcal{E}}_2 \\ \vdots \\ \boldsymbol{\mathcal{E}}_{k_n} \end{pmatrix}}_{E},$$

where $\boldsymbol{Y}_i = (Y_{r_{i-1}}, \ldots, Y_{r_i-1})'$, $\boldsymbol{X}_i = (X_{r_{i-1}}, \ldots, X_{r_i-1})'$, $\boldsymbol{\mathcal{E}}_i = (\varepsilon_{r_{i-1}}, \ldots, \varepsilon_{r_i-1})'$, $i = 1, \ldots, k_n$. $\mathcal{Y} \in \mathbb{R}^{2n}$, $\mathcal{X} \in \mathbb{R}^{2n \times 2k_n}$, $\Theta(k_n) \in \mathbb{R}^{2k_n}$ and $E \in \mathbb{R}^{2n}$.

A simple estimate of parameters $\Theta(k_n)$ can be obtained by using an ℓ_1-penalized least squares regression of the form

$$\widehat{\Theta}(k_n) = \underset{\Theta \in \mathbb{R}^{2k_n}}{\arg\min} \left\{ \frac{1}{2n} \|\mathcal{Y} - \mathcal{X}\Theta\|_2^2 + \lambda_n \|\Theta\|_1 \right\}, \quad (16)$$

which uses a fused lasso penalty to control the number of change points in the model. This penalty term encourages the parameters across consecutive time blocks to be similar or even identical; hence, only large changes are registered, thus aiding in identifying the change points. Further, a hard-thresholding procedure is added to cluster the jumps into two sets: large and small ones, so that those redundant change points with small changes in the estimated parameters can be removed. We only declare that there is a change point at the end point of a block, when associated with large jump of the model parameters.

Hard Thresholding: is based on a data-driven procedure for selecting the threshold η. The idea is to combine the K-means clustering method [50] with the BIC criterion [51] to cluster the changes in the parameter matrix into two subgroups. The main steps are:

- Step 1 (initial state): Denote the jumps for each block by $v_k = \|\widehat{\theta}_k\|_2^2$, $k = 2, \ldots, k_n$ and let $v_1 = 0$. Denote the set of selected blocks with large jumps as J (initially, this is an empty set) and set $\text{BIC}^{old} = \infty$.
- Step 2 (recursion state): Apply K-means clustering to the jump vector $V = (v_1, v_2, \ldots, v_{k_n})$ with two centers. Denote the sub-vector with a smaller center as the small subgroup, V_S, and the other sub-vector as the large subgroup, V_L. Add the corresponding blocks in the large subgroup into J. Compute the BIC by using the estimated parameters $\widehat{\Theta}$ after setting $\widehat{\theta}_i = 0$ for each block $i \notin J$ and denote it by BIC^{new}. Compute the difference $\text{BIC}^{\text{diff}} = \text{BIC}^{new} - \text{BIC}^{old}$and update $\text{BIC}^{old} = \text{BIC}^{new}$. Repeat this step until $\text{BIC}^{\text{diff}} \ge 0$.

Block clustering: the Gap statistic [52] is applied to determine the number of clusters of the candidate change points. The basic idea is to run a clustering method (here, K-means is selected) over a grid of possible number of clusters, and to pick the optimal one by comparing the changes in within-cluster dispersion with that expected under an appropriate reference null distribution (for more details, see Section III in [52]).

Exhaustive search: Define $l_i = (\min(C_i) - b_n)\mathbb{1}_{\{|C_i|=1\}} + \min(C_i)\mathbb{1}_{\{|C_i|>1\}}$ and $u_i = (\max(C_i) + b_n)\mathbb{1}_{\{|C_i|=1\}} + \max(C_i)\mathbb{1}_{\{|C_i|>1\}}$, where C_i's are the subsets of candidates blocks by block clustering procedure. Denote the subset of corresponding block indices by J_i. Define the following local coefficient parameter estimates:

$$\widehat{\boldsymbol{B}}_i = \sum_{k=1}^{\frac{1}{2}(\max(J_{i-1})+\min(J_i))} \widehat{\theta}_k, \text{ for } i = 1, \ldots, \widetilde{m}^f + 1, \quad (17)$$

where $\widetilde{m}^f$ is the number of clusters obtained in the block clustering procedure, $J_0 = \{1\}$ and $J_{\widetilde{m}^f+1} = \{k_n\}$.

Now, given a subset C_i, we apply the exhaustive search method for each time point s in the interval (l_i, u_i) to the data set truncated by the two end points in time, $\min(C_i) - b_n$ and $\max(C_i) + b_n$, i.e. only consider the data within the interval $[\min(C_i) - b_n, \max(C_i) + b_n)$. Specifically, define the final estimated change point $\widetilde{t}_i^f$ as

$$\widetilde{t}_i^f = \underset{s \in (l_i, u_i)}{\arg\min} \left\{ \sum_{t=\min(C_i)-b_n}^{s-1} \left\| Y_t - X_t \widehat{\boldsymbol{B}}_i \right\|_2^2 + \sum_{t=s}^{\max(C_i)+b_n-1} \left\| Y_t - X_t \widehat{\boldsymbol{B}}_{i+1} \right\|_2^2 \right\}, \quad (18)$$

for $i = 1, \ldots, \widetilde{m}^f$, where $\widehat{\boldsymbol{B}}_i$'s are the local coefficient parameter estimates based on the first step by block fused lasso. Denote the set of final estimated change points from (18) by $\widetilde{\mathcal{A}}_n^f = \{\widetilde{t}_1^f, \ldots, \widetilde{t}_{\widetilde{m}^f}^f\}$.

Remark: An alternative approach for detection of break points is to run a full exhaustive search procedure for both single and multiple change point problems. Such procedures are computationally expensive, and not scalable for large data sets. Simple

fused lasso is (block of size one) another method, which although computationally fast, it leads to over-estimating the number of break points; hence, it requires additional "screening" steps to remove redundant break points found using the fused lasso algorithm [53], [54]. Such screening steps usually include tuning several hyper-parameters. This task not only slows down the detection method, but is also not robust. The proposed approach (block fused lasso coupled with hard-thresholding) aims to solve this issue by choosing appropriate block sizes, while it only needs a single tuning parameter (the threshold) to be estimated.

Estimation of Infection and Recovery rates: Once the locations of break points are obtained, one can estimate the model parameters by running a separate regression for each identified stationary segment of the time series data. The work of [54] shows that this strategy yields consistent model parameter estimates.

Grid search of parameter in the under-reporting function: We use grid search to estimate the parameter a in the under-reporting function $u(t)$. Given a parameter grid of a, we transform the observed infected data $I(t)$ by $I_f(t) = \Delta I(t)/(1 - u(t)) + I_f(t-1)$, then apply the transformed data to the above method and compute the in-sample mean relative prediction error (MRPE) of $\Delta I_f(t)$. Choose the value a that minimizes the in-sample MRPE of $\Delta I_f(t)$.

Appendix B
Theoretical Properties

In this section, we establish the prediction consistency of the estimator from (16). To establish predictionfollowing assumptions are needed:

A.1) (Deviation bound) There exist constants $c_i > 0$ such that with probability at least $1 - c_1\exp(-c_2(\log 2n))$, we have

$$\left\| \frac{\mathcal{X}'E}{2n} \right\|_\infty \leq c_3\sqrt{\frac{\log 2n}{2n}}. \tag{19}$$

A.2) There exists a positive constant $M_B > 0$ such that

$$\max_{1\leq j\leq m_0+1} \|B^{(j)}\|_\infty \leq M_B.$$

Theorem 1: 1 Suppose A1-A2 hold. Choose $\lambda_n = 2C_1\sqrt{\frac{\log 2n}{2n}}$ for some large constant $C_1 > 0$, and assume $m_0 \leq m_n$ with $m_n = o(\lambda_n^{-1})$. Then with high probability approaching to 1 and $n \to +\infty$, the following holds:

$$\frac{1}{2n}\left\|\mathcal{X}(\widehat{\Theta} - \Theta)\right\|_2^2 \leq 8M_B\lambda_n m_0. \tag{20}$$

Proof of Theorem 1: By the definition of $\widehat{\Theta}$ in (16), the value of the function in (16) is minimized at $\widehat{\Theta}$. Therefore, we have

$$\frac{1}{2n}\left\|\mathcal{Y} - \mathcal{X}\widehat{\Theta}\right\|_2^2 + \lambda_n\left\|\widehat{\Theta}\right\|_1 \leq \frac{1}{2n}\|\mathcal{Y} - \mathcal{X}\Theta\|_2^2 + \lambda_n\|\Theta\|_1. \tag{21}$$

Denoting $\mathcal{A} = \{t_1, t_2, \ldots, t_{m_0}\}$ as the set of true change points, we have

$$\begin{aligned}
&\frac{1}{2n}\left\|\mathcal{X}\left(\widehat{\Theta} - \Theta\right)\right\|_2^2 \\
&\leq \frac{1}{n}\left(\widehat{\Theta} - \Theta\right)'\mathcal{X}'E + \lambda_n\left(\|\Theta\|_1 - \left\|\widehat{\Theta}\right\|_1\right) \\
&\leq \frac{1}{n}\left(\widehat{\Theta} - \Theta\right)'\mathcal{X}'E + \lambda_n\left(\sum_{i=1}^{k_n}\|\theta_i\|_1 - \sum_{i=1}^{k_n}\|\widehat{\theta}_i\|_1\right) \\
&\leq 2\sum_{i=1}^{k_n}\left\|\widehat{\theta}_i - \theta_i\right\|_1\left\|\frac{\mathcal{X}'E}{2n}\right\|_\infty + \lambda_n\sum_{i\in\mathcal{A}}\left(\|\theta_i\|_1 - \left\|\widehat{\theta}_i\right\|_1\right) \\
&\qquad - \lambda_n\sum_{i\notin\mathcal{A}}\|\widehat{\theta}_i\|_1 \\
&\leq \lambda_n\sum_{i\in\mathcal{A}}\left\|\widehat{\theta}_i - \theta_i\right\|_1 + \lambda_n\sum_{i\in\mathcal{A}}\left(\|\theta_i\|_1 - \left\|\widehat{\theta}_i\right\|_1\right) \\
&\leq 2\lambda_n\sum_{i\in\mathcal{A}}\|\theta_i\|_1 \\
&\leq 2\lambda_n m_0\max_{1\leq j\leq m_0}\left\|B^{(j+1)} - B^{(j)}\right\|_1 \\
&\leq 8M_B\lambda_n m_0,
\end{aligned}$$

with high probability approaching to deviation bound in (19). This completes the proof.

Theoretical properties of lasso have been have been studied by several authors [55]–[58]. In controlling the statistical error, a suitable deviation conditions on $\mathcal{X}'E/2n$ is needed. The deviation bound conditions (e.g. the assumption A1) are known to hold with high probability under several mild conditions. Under the condition that the error term $E \sim \mathcal{N}(0, \sigma^2 I_{2n})$, the deviation bound condition holds with high probability by Lemme 3.1 in [56]. Given that the p (the number of time series components) is small and fixed, we have $n \gg \log p$, therefore, in the case where the $\mathcal{X}$ is a zero-mean sub-Gaussian matrix with parameters (Σ_x, σ_x^2), and the error term E is a zero-mean sub-Gaussian matrix with parameters (Σ_e, σ_e^2), the deviation bound condition holds with high probability by Lemme 14 in [57].

Detection Accuracy: When the block size is large enough, such that $\log n/b_n$ remains small, if the selected change point $\widehat{t}_j$ is close to a true change point, the estimated $\widehat{\Theta}_j$ will be large (asymptotically similar to the true jump size in the model parameters); if the selected change point $\widehat{t}_j$ is far away from all the true change points, the estimated $\widehat{\Theta}_j$ will be quite small (converges to zero as sample size tend to infinity). Therefore, after the hard-thresholding, the candidate change points that are located far from any true change points will be eliminated. In other words, for any selected change point $\widetilde{t}_j \in \widetilde{\mathcal{A}}_n$, there would exist a true change point $t_{j'} \in \mathcal{A}_n$ close by, with the distance being at most b_n. Thus, the number of clusters (by radius b_n) seems to be a reasonable estimate for the true number of break points in the model.

On the other hand, since the set of true change points $\mathcal{A}_n$ has cardinality less than or equal to the cardinality of the set of selected change points $\widetilde{\mathcal{A}}_n$, i.e., $m_0 \leq \widetilde{m}$, there may be more than one selected change points remaining in the set $\widetilde{\mathcal{A}}_n$ in b_n-neighborhoods of each true change point. For a set A, define cluster (A, x) to be the minimal partition of A, where the diameter for each subset is at most x. Denote the subset in cluster$(\widetilde{\mathcal{A}}_n, b_n)$

by cluster$(\widetilde{\mathcal{A}}_n, b_n) = \{C_1, C_2, \ldots, C_{m_0}\}$, where each subset C_i has a diameter at most b_n, i.e., $\max_{a,b \in C_i} |a - b| \leq b_n$. Then with high probability converging to one, the number of subsets in cluster$(\widetilde{\mathcal{A}}_n, b_n)$ is exactly m_0. All candidate change points in $\widetilde{A}_n$ are within a b_n-neighborhood of at least one true change point and therefore, with high probability converging to one, there is a true change point t_i within the interval $(C_i - b_n, C_i + b_n)$. The distance between the estimated change point and the true change point will be less then $2b_n$. Therefore, by selecting $b_n = c \log n$ for a large enough constant $c > 0$, one can conclude that the proposed detection algorithm locates the true break points with an error bounded by the order $\mathcal{O}(\log n)$.

ACKNOWLEDGMENT

The authors would like to thank the Editor Bjoern Schuller and two anonymous referees for many constructive comments and suggestions.

REFERENCES

[1] R. M. Anderson *et al.*, "How will country-based mitigation measures influence the course of the COVID-19 epidemic?" *Lancet*, vol. 395, no. 10228, pp. 931–934, 2020.
[2] Centers for Disease Control and Prevention, "Considerations for wearing masks," 2020. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/cloth-face-cover-guidance.html
[3] N. H. Leung *et al.*, "Respiratory virus shedding in exhaled breath and efficacy of face masks," *Nature Med.*, vol. 26, no. 5, pp. 676–680, 2020.
[4] P. Horby *et al.*, "NERVTAG note on B. 1.1. 7 severity," *NERVTAG*, vol. 1, no. 7, pp. 1–3, 2021.
[5] S. Flaxman *et al.*, "Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe," *Nature*, vol. 584, no. 7820, pp. 257–261, 2020.
[6] G. Meyerowitz-Katz and L. Merone, "A systematic review and meta-analysis of published research data on COVID-19 infection-fatality rates," *Int. J. Infect. Dis.*, vol. 101, pp. 138–148, 2020.
[7] E. J. Williamson *et al.*, "Factors associated with COVID-19-related death using opensafely," *Nature*, vol. 584, no. 7821, pp. 430–436, 2020.
[8] K. Qian, B. W. Schuller, and Y. Yamamoto, "Recent advances in computer audition for diagnosing COVID-19: An overview," in *Proc. IEEE 3rd Global Conf. Life Sci. Technol.*, 2021, pp. 181–182.
[9] D. Yaron *et al.*, "Point of care image analysis for COVID-19," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 8153–8157.
[10] P. Forster, L. Forster, C. Renfrew, and M. Forster, "Phylogenetic network analysis of SARS-CoV-2 genomes," *Proc. Nat. Acad. Sci.*, vol. 117, no. 17, pp. 9241–9243, 2020.
[11] E. L. Anderson, P. Turnham, J. R. Griffin, and C. C. Clarke, "Consideration of the aerosol transmission for COVID-19 and public health," *Risk Anal.*, vol. 40, no. 5, pp. 902–907, 2020.
[12] M. A. Honein, L. C. Barrios, and J. T. Brooks, "Data and policy to guide opening schools safely to limit the spread of SARS-CoV-2 infection," *Jama*, vol. 325, no. 9, pp. 823–824, 2021.
[13] M. Z. Bazant and J. W. Bush, "A guideline to limit indoor airborne transmission of COVID-19," *Proc. Nat. Acad. Sci.*, vol. 118, no. 17, 2021, Art. no. e2018995118.
[14] Y.-C. Chen, P.-E. Lu, C.-S. Chang, and T.-H. Liu, "A time-dependent SIR model for COVID-19 with undetectable infected persons," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 4, pp. 3279–3294, Oct.–Dec. 2020.
[15] C. Anastassopoulou *et al.*, "Data-based analysis, modelling and forecasting of the COVID-19 outbreak," *PLoS One*, vol. 15, no. 3, 2020, Art. no. e0230405.
[16] L. Wang *et al.*, "An epidemiological forecast model and software assessing interventions on the COVID-19 epidemic in China," *J. Data Sci.*, vol. 18, no. 3, pp. 409–432, 2020.
[17] J. Wangping *et al.*, "Extended sir prediction of the epidemics trend of COVID-19 in Italy and compared with Hunan, China," *Front. Med.*, vol. 7, 2020, Art. no. 169.
[18] J. Dehning *et al.*, "Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions," *Science*, vol. 369, 2020, Art. no. eabb9789.
[19] F. Jiang, Z. Zhao, and X. Shao, "Time series analysis of COVID-19 infection curve: A change-point perspective," *J. Econometrics*, 2020.
[20] R. Baranowski, Y. Chen, and P. Fryzlewicz, "Narrowest-over-threshold detection of multiple change points and change-point-like features," *J. Roy. Stat. Soc.: Ser. B. (Statist. Methodol.)*, vol. 81, no. 3, pp. 649–672, 2019.
[21] Z. Vokó and J. G. Pitter, "The effect of social distance measures on COVID-19 epidemics in Europe: An interrupted time series analysis," *GeroScience*, vol. 42, pp. 1075–1082, 2020.
[22] A. B. *et al.*, "Social distancing merely stabilized COVID-19 in the United States," *Stat*, vol. 9, no. 1, 2020, Art. no. e302. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/sta4.302
[23] P. Fearnhead, R. Maidstone, and A. Letchford, "Detecting changes in slope with an L_0 penalty," *J. Comput. Graphical Statist.*, vol. 28, no. 2, pp. 265–275, 2019.
[24] A. L. Bertozzi, E. Franco, G. Mohler, M. B. Short, and D. Sledge, "The challenges of modeling and forecasting the spread of COVID-19," *Proc. Nat. Acad. Sci.*, vol. 117, no. 29, pp. 16732–16738, 2020.
[25] L. Wang *et al.*, "Spatiotemporal dynamics, nowcasting and forecasting of COVID-19 in the United States," 2020, *arXiv:2004.14103*.
[26] A. Srivastava and V. K. Prasanna, "Learning to forecast and forecasting to learn from the COVID-19 pandemic," 2020, *arXiv:2004.11372*.
[27] H. Qi *et al.*, "COVID-19 transmission in mainland China is associated with temperature and humidity: A time-series analysis," *Sci. Total Environ.*, vol. 728, 2020, Art. no. 138778.
[28] L. Moftakhar, S. Mozhgan, and M. S. Safe, "Exponentially increasing trend of infected patients with COVID-19 in Iran: A comparison of neural network and ARIMA forecasting models," *Iranian J. Public Health*, vol. 49, no. Supple 1, pp. 92–100, 2020.
[29] V. K. R. Chimmula and L. Zhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks," *Chaos, Solitons Fractals*, vol. 135, 2020, Art. no. 109864.
[30] Z. Hu, Q. Ge, L. Jin, and M. Xiong, "Artificial intelligence forecasting of COVID-19 in China," 2020, *arXiv:2002.07112*.
[31] I. Rahimi, F. Chen, and A. H. Gandomi, "A review on COVID-19 forecasting models," *Neural Comput. Appl.*, pp. 1–11, 2021.
[32] R. Tibshirani *et al.*, "Sparsity and smoothness via the fused lasso," *J. Roy. Stat. Soc.: Ser. B. (Statist. Methodol.)*, vol. 67, no. 1, pp. 91–108, 2005.
[33] W. O. Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epidemics," *Proc. Roy. Soc. London. Ser. A*, vol. 115, no. 772, pp. 700–721, 1927.
[34] S. L. Wu, A. N. Mertens, and Y. S. Crider, "Substantial underestimation of SARS-CoV-2 infection in the United States," *Nature Commun.*, vol. 11, no. 1, pp. 1–10, 2020.
[35] H. Lau *et al.*, "Evaluating the massive underreporting and undertesting of COVID-19 cases in multiple global epicenters," *Pulmonology*, vol. 27, pp. 110–115, 2021.
[36] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. Berlin, Germany: Springer Science & Business Media, 2005.
[37] G. Giordano *et al.*, "Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy," *Nature Med.*, vol. 26, pp. 855–860, 2020.
[38] X. Liu and P. Stechlinski, "Infectious disease models with time-varying parameters and general nonlinear incidence rate," *Appl. Math. Modeling*, vol. 36, no. 5, pp. 1974–1994, 2012.
[39] M. Bartlett, "Some evolutionary stochastic processes," *J. Roy. Stat. Soc.. Ser. B. (Methodological)*, vol. 11, no. 2, pp. 211–229, 1949.
[40] N. T. Bailey, "The total size of a general stochastic epidemic," *Biometrika*, vol. 40, pp. 177–185, 1953.
[41] P. E. Greenwood and L. F. Gordillo, "Stochastic epidemic modeling," in *Mathematical and Statistical Estimation Approaches in Epidemiology*. Berlin, Germany: Springer, 2009, pp. 31–52.
[42] *The New York Times*, "Coronavirus (COVID-19) data in the United States," 2020. [Online]. Available: https://github.com/nytimes/covid-19-data
[43] G. Wang *et al.*, "Comparing and integrating US COVID-19 daily data from multiple sources: A county-level dataset with local characteristics," 2020, *arXiv:2006.01333*.
[44] *U. S. Census Bureau*, "Population and housing unit estimates datasets," 2019. [Online]. Available: https://www.census.gov/programs-surveys/popest/data/data-sets.html

[45] National Bureau of Economic Research, "Population and housing unit estimates datasets," 2010. [Online]. Available: http://data.nber.org/distance/2010/sf1/
[46] S. F. Lumley *et al.*, "Antibody status and incidence of SARS-CoV-2 infection in health care workers," *New England J. Med.*, vol. 384, pp. 533–540, 2021.
[47] N. Cressie, *Statistics for Spatial Data*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2015.
[48] S. Mervosh *et al.*, "See how all 50 states are reopening," 2020. [Online]. Available: https://www.nytimes.com/interactive/2020/us/states-reopen-map-coronavirus.html
[49] R. Li *et al.*, "Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2)," *Sci.*, vol. 368, no. 6490, pp. 489–493, 2020.
[50] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *J. Roy. Stat. Soc. Ser. C (Appl. Statist.)*, vol. 28, no. 1, pp. 100–108, 1979.
[51] G. Schwarz *et al.*, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.
[52] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J. Roy. Stat. Soc.: Ser. B. (Statist. Methodol.)*, vol. 63, no. 2, pp. 411–423, 2001.
[53] Z. Harchaoui and C. Lévy-Leduc, "Multiple change-point estimation with a total variation penalty," *J. Amer. Stat. Assoc.*, vol. 105, no. 492, pp. 1480–1493, 2010.
[54] A. Safikhani and A. Shojaie, "Joint structural break detection and parameter estimation in high-dimensional non-stationary VAR models," *J. Amer. Stat. Assoc.*, pp. 1–26, 2020.
[55] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of lasso and dantzig selector," *Ann. Statist.*, vol. 37, no. 4, pp. 1705–1732, 2009.
[56] S. van de Geer, P. Bühlmann, and S. Zhou, "The adaptive and the thresholded lasso for potentially misspecified models (and a lower bound for the lasso)," *Electron. J. Statist.*, vol. 5, pp. 688–749, 2011.
[57] P.-L. Loh and M. J. Wainwright, "High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity," *Ann. Statist.*, vol. 40, no. 3, pp. 1637–1664, Jun. 2012.
[58] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Stat. Sci.*, vol. 27, no. 4, pp. 538–557, 2012.

Yue Bai received the B.Sc. degree in statistics from the Renmin University of China, Beijing, China, in July 2016, and the M.Sc. degree in statistics from the University of Wisconsin–Madison, Madison, WI, USA, in May 2017. She is currently working toward the Ph.D. degree in statistics with the University of Florida, Gainesville, FL, USA. Her research interests include high-dimensional time series, change point detection, and machine learning.

Abolfazl Safikhani received the B.Sc. and M.Sc. degrees in mathematics from the Sharif University of Technology, Tehran, Iran, in 2008 and 2010, respectively, and the Ph.D. degree in statistics from Michigan State University, East Lansing, MI, USA, in 2015. After that he joined Columbia University, New York, NY, USA, as a term Assistant Professor. In 2019, he joined the Statistics Department, University of Florida (UF), Gainesville, FL, USA, as an Assistant Professor and he is also affiliated with the Informatics Institute, UF, Gainesville, FL, USA. His research interests include time series models, network modeling, high-dimensional statistics, and unsupervised learning.

George Michailidis (Member, IEEE) received the B.S. degree in economics from the University of Athens, Athens, Greece, in 1987, the M.A. degrees in economics and mathematics and the Ph.D. degree in mathematics from the University of California, Los Angeles, CA, USA. The Postdoc in operations research with Stanford University, Stanford, CA. In 1998, he joined the Department of Statistics, University of Michigan, Ann Arbor, MI, USA, where he became a Full Professor in 2008. In 2015, he joined the University of Florida, Gainesville, FL, USA, as the Founding Director of the Informatics Institute. He is a Fellow of the American Statistical Association, Institute of Mathematical Statistics, and International Statistical Institute. His research interests include network analysis, queueing theory, stochastic control and optimization, applied probability, and machine learning.