Modeling and Forecasting COVID-19

Li Wang, Guannan Wang, Xinyi Li, Shan Yu, Myungjin Kim, Yueying Wang, Zhiling Gu, and Lei Gao

Background

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) outbreak started in December 2019, and it has expanded to impact nearly every corner of the world. On New Year's Eve of 2019, the World Health Organization (WHO) was informed of mysterious pneumonia cases in Wuhan, China. On January 3, 2020, 44 cases were reported to WHO, among whom 11 were severely ill. By February 4, confirmed cases had been reported in 24 countries outside China. The US confirmed its first case in Washington State on January 21—a man who returned to the US from Wuhan. On January 30, the US Centers for Disease Control and Prevention (CDC) confirmed the person-to-person spread of coronavirus in the US. On the same day, the WHO declared the coronavirus outbreak as a Public Health Emergency of International Concern.[1]

Many nonpharmaceutical interventions (NPIs) were implemented to prevent the spread of COVID-19. For example, Wuhan implemented screening measures for travelers leaving the city at airports, railway stations, and other passenger terminals, and eventually closed off Wuhan City on January 22. The US started screening at twenty airports at the end of January. On February 25, San Francisco became the first US city to declare a state of emergency over COVID-19, followed by the states of Washington and Florida. On March 13, President Donald Trump declared a national COVID-19 emergency, and sixteen states announced school closures by then. On March 19, California issued a "stay-at-home" order for all of its 40 million residents, and within two weeks, the majority of the states had taken similar actions. By the end of March, more than 91% of the world's population lived in countries with restrictions for nonresident travelers from abroad.

Even with all the control measures taken in place, the spread of COVID-19 is still dramatic. From February 7 to 14, 2020, both the total cases confirmed and deaths worldwide almost doubled within one week. Meanwhile, COVID-19 kept spreading globally, and over 50 countries reported confirmed cases by the end of February. During mid-March, COVID-19 presented in all 50 states in the US. Starting from March 26, the US led the world in COVID-19 cases. On May 28, the US COVID-19 death count passed one hundred thousand, and then in the middle of June,

Li Wang is a professor of statistics at Iowa State University. Her email address is lilywang@iastate.edu.

Guannan Wang is an assistant professor of mathematics at the College of William & Mary. Her email address is gwang01@wm.edu.

Xinyi Li is an assistant professor of mathematical and statistical sciences at Clemson University. Her email address is lixinyi@clemson.edu.

Shan Yu is an assistant professor of statistics at the University of Virginia. Her email address is sy5jx@virginia.edu.

Myungjin Kim is a PhD candidate at Iowa State University. His email address is mjkim@iastate.edu.

Yueying Wang is a PhD candidate at Iowa State University. Her email address is yueyingw@iastate.edu.

Zhiling Gu is a PhD student at Iowa State University. Her email address is zlgu@iastate.edu.

Lei Gao is an assistant professor of finance at Iowa State University. His email address is lgao@iastate.edu.

Communicated by Notices *Associate Editor Richard Levine.*

For permission to reprint this article, please contact: reprint-permission@ams.org.

DOI: https://doi.org/10.1090/noti2263

[1] *The number of references is limited to twenty. Please refer to the background and introduction section of* [WWG+20] *for the detailed references.*

the number of confirmed cases of COVID-19 hit two million in the US. The confirmed and death cases kept increasing rapidly in the following months. By the beginning of September, the US surpassed six million confirmed cases and hit seven million on September 25. On October 16, the US surpassed eight million confirmed cases and 218 thousand deaths.

The effect of COVID-19 is profound. The World Bank estimated that the coronavirus pandemic could push an additional 16 million people into extreme poverty. On April 14, 2020, the International Monetary Fund warned that the world is facing its worst economic downturn as coronavirus lockdowns continue to wreak havoc on the global economy. Due to the immense pressure of the crippled economy and an anxious public amid a pandemic, the US started to loosen lockdown measures in late April and backtracked after reopening for a few weeks and seeing a surge in cases. As the pandemic progresses, there are broader interests from the public in answering questions such as how far the SARS-CoV-2 virus will spread, how many lives it will eventually claim, how effective intervention strategies will be, as well as whether and when the pandemic will resurge.

Epidemic modeling is an essential scientific tool to answer these questions by aiding people to understand the pandemic data, make predictions, and help the medical professionals and decision-makers allocate resources and design/evaluate intervention strategies to fight against COVID-19. In this paper, we first discuss two main modeling frameworks in epidemiology studies: mathematical and statistical modeling. Next, we overview the difficulties and challenges of forecasting COVID-19 under the enormous uncertainty and introduce some forecasting methods. Borrowing the strength of mathematical and statistical modeling, we propose a novel space-time epidemic modeling framework to study the spatial-temporal pattern in the spread of COVID-19. Based on this framework, we provide both short-term and long-term forecasts of the infected and death counts at the county level in the US. We also discuss various new perspectives on how to select an appropriate model for COVID-19 data analyses.

Epidemic Modeling: Mathematical and Statistical Perspectives

Epidemic modeling has three main aims [DG99]: (1) to understand better the mechanisms by which diseases spread; (2) to identify which factors contribute to the spread of the epidemic, and therefore how we may control it; (3) to predict the future course of the epidemic. Although there are many epidemic modeling methods, mathematical and statistical models have played important roles in COVID-19 studies. As illustrated in Figure 1,

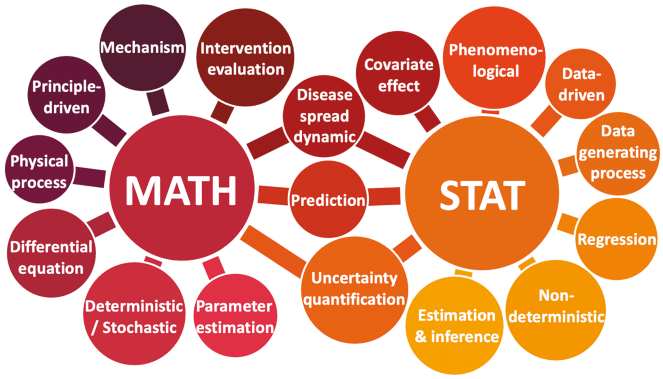

Figure 1. Mathematical and statistical perspectives on epidemic modeling.

mathematical and statistical approaches are complementary, but their starting points are different, and the corresponding models tend to incorporate different details.

As mentioned above, the fundamental concept of infectious disease epidemiology is investigating how the diseases spread. Mathematical models are undeniably useful in understanding the dynamics of infectious disease spread (e.g., when the peak will occur and whether resurgence will happen) and the effects of control measures [KR08]. An essential type of mathematical model is the class of mechanistic models such as the Susceptible-Infectious-Removed (SIR) compartmental model or the Susceptible-Exposed-Infectious-Recovered model (SEIR) as illustrated in Figure 2; see details in [BCCF19, LBHU16]. Mechanistic models make explicit hypotheses about the biological mechanisms that drive the dynamics of infection, and they function well if the aim is to evaluate the effectiveness of hypothetical NPIs in controlling disease spread [LC16].

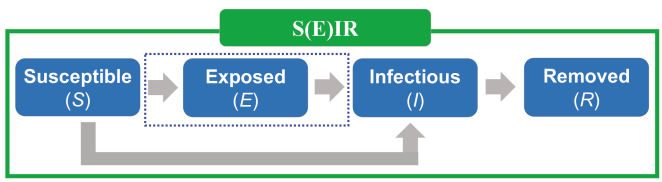

Figure 2. An illustration of SIR and SEIR models.

In the literature, statistical modeling has given the scientific field many successes in analyzing data and getting information about the mechanisms producing the data. Statistical modeling is a powerful tool for extracting information about the disease spread in epidemic studies [HHOW20]. Statistics starts with data, and statistical modeling allows data to speak for themselves. There are two cultures in statistical modeling [Bre01]: the data modeling culture and the algorithmic modeling culture. The first one assumes that the data are generated by a given stochastic

data model, and it is usually designed for inference about the relationships between variables whilst also catering to prediction. Algorithmic models treat the data mechanism as unknown and are usually designed to make the most accurate predictions possible.

When analyzing the spread of infectious diseases, other factors, such as demographic characteristics, socioeconomic status, and control policies, may also be responsible for temporal or spatial patterns. For example, the spread of the disease varies considerably across different geographical regions. Local area-features, like socioeconomic factors and demographic conditions, can dramatically influence the course of the epidemic. These data are usually supplemented with the population information at the county level. Moreover, the capacity of the health care system and control measures, such as government-mandated social distancing, also have a significant impact on the spread of the epidemic. Regression is a widely used statistical modeling method in epidemic studies because it produces a combination of the variables with weights indicating the impact of the variable [Jew03]. It can help determine which factors matter most, which can be ignored, and how those factors interact with each other. The benefit of regression analysis is that it can be used to understand different patterns in data. These insights may often be very valuable in understanding which factors contribute to the spread of COVID-19.

Predicting the spread speed and severity of COVID-19 is crucial to resource management, developing strategies to deal with the epidemic, and ultimately assisting in prevention efforts. Mathematical models are able to mimic the way disease spreads and can be used to project or simulate future transmission scenarios under various assumptions. Statistical models are more oriented towards predictions [HHOW20]. In fact, predictions are at the heart of statistical modeling. For example, time series analysis, one commonly used statistical forecasting approach, works by taking a series of historical observations and extrapolating the patterns into the future. Machine learning makes predictions based on known properties learned from the training data. However, purely statistical models only describe the observed data and give little information about the mechanism since they do not account for how transmission occurs. Therefore, they are generally not well suited for long-term predictions, and a few weeks is usually close to being the ultimate prediction limit. Another advantage of statistical modeling is its ability to quantify uncertainty in the prediction, especially at an early phase of an epidemic with limited data. For example, statistical models can provide a prediction interval to understand the uncertainty surrounding the forecast [BD16]. See more discussion in the following section.

In summary, mathematical models are usually constructed in a more principle-driven manner, while statistical models are more data-driven. Although both mathematical and statistical models can be used to study the effect of NPIs and make predictions, the implementation details are different, and an understanding of the corresponding limitations is crucial. For maximal effectiveness, researchers working to advance epidemic modeling will need to appreciate and exploit the complementary strengths of mathematical and statistical models.

Forecasting COVID-19 under Enormous Uncertainty

Several quantities are of interest in COVID-19 forecasting, such as the timing of and incidence in the peak week, cumulative incidence, and weekly incidence. The policy/decision-makers are also interested in evaluating outbreak size and duration, and employing the epidemic curve to identify the mode of transmission of the disease and measure its prevalence of the disease.

Forecasting goals can also be classified as long-term or short-term forecasts. Long-term disease forecasts can predict COVID-19 peak or severity, while short-term forecasts can be used to guide resource allocation in the short term by local agencies or to anticipate the case burden by hospitals in the coming week; see [ABD+20]. The projection can be made at different resolution levels, for example, national, regional, or local. National-level or state-level long-term forecasts are of interest to policymakers regulating intervention strategies and deciding how much funding to allocate for resources. Prediction models with a finer resolution are needed to assess the local risk of COVID-19. Knowing more about the vulnerable communities and the reasons for those communities that are more likely to be infected are crucial for the policy and decision-makers to assist in prevention efforts [ABD+20, WWG+20].

Challenges of forecasting COVID-19. The difficulty depends on the forecasting target. The short-term forecast is relatively easy since we observe a clear time series trend. Moreover, there is less uncertainty about what is observed. By capturing underlying intricate patterns and relationships, many statistical methods can be used for short-term forecasts, such as time series analysis and machine learning methods. The long-term forecast is much more difficult than many people think. There are three contributing factors that may affect the accuracy of forecasts [HA18]: (1) how well we understand the factors that contribute to the forecasting target; (2) how much high-quality data are available; (3) whether the outcome of forecasts can affect the spread of COVID-19.

Firstly, the accuracy of the forecast is constrained by what we know about the disease. With an emerging

new disease such as COVID-19, many biologic features of transmission are hard to measure and remain unknown. More work is necessary to better understand the risk factors for severe illness or complications, for example, age, race/ethnicity, gender, and medical conditions. Compared to the meteorologic forecasting method often used in weather forecasting, epidemic forecasting is still at an early stage of development, and the human component makes it particularly challenging [MFG+16].

Secondly, we don't have a wholly accurate picture of how widespread COVID-19 is. In forecasting, we have to deal with incomplete and inaccurate data. The number of confirmed cases might be vastly underreported due to the limited availability of testing. Moreover, the size of underreported cases varies enormously by countries or regions. COVID-19 hospitalizations and deaths data might be more reliable; however, the official numbers of people who have died of COVID-19 are not consistent with the number of observed fatal cases on the front lines. There might be a lag in reporting in some cases due to delays and possible breakdowns in logging positive tests and making them public. The lack of reliable data sources becomes a severe problem for forecasting the dynamic course of the crisis and where resources are most urgently needed.

Lastly, the forecast may influence NPIs and human social behavior because the latter are likely to be based on the outcome of the forecast. The changes in human social behavior can further affect the transmission dynamics and shape the future of an epidemic; see Figure 3. Thus, to accurately forecast an emerging epidemic, we need to foresee individuals' behavior, potential changes in the pathogen, as well as their interactions as they relate to the transmission dynamics.

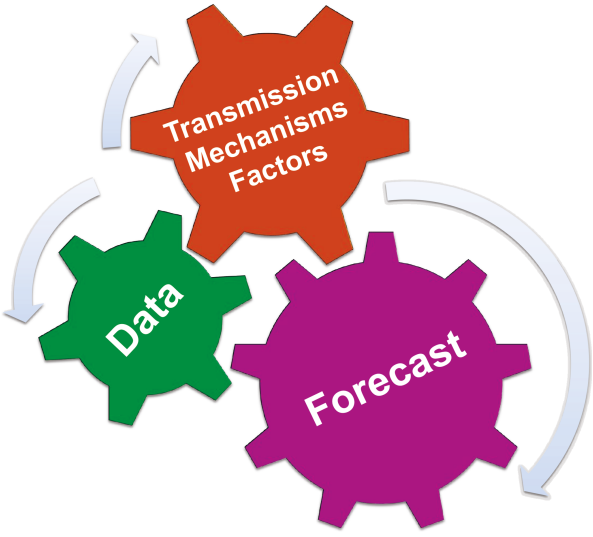

Figure 3. COVID-19 forecast.

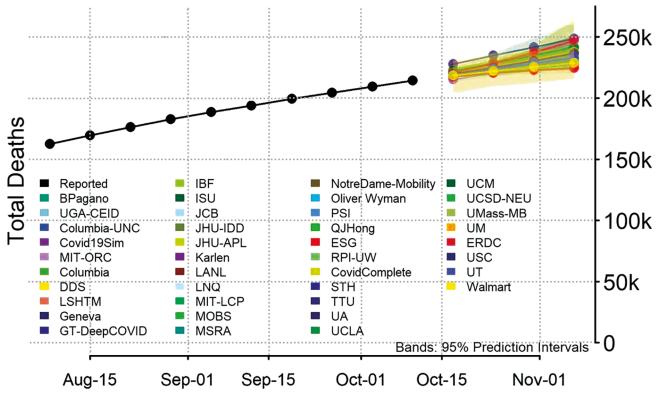

Figure 4. National total fatal cases forecast. (Source with full name for each group: `https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html`.)

Furthermore, the forecast of COVID-19 is amid a significant amount of uncertainty. There are at least four sources of uncertainty. Firstly, there is uncertainty due to the lack of knowledge of the biological features of transmission. Most medical parameters are unknown or enormously uncertain; for example, due to mutation of the virus, some parameters might change during the development of the disease. It may take years to understand the complexity of the spread of SARS-CoV-2 fully. Secondly, the aforementioned missing or incomplete data is another source of uncertainty in model forecasting. Thirdly, health-policy support for COVID-19 requires some knowledge of social patterns not only under physical distancing policies but also in various reopening scenarios. Last but not least, the source of uncertainty may also come from the measurement error, modeling procedure, and the sampling error.

As mentioned above, there are many types of factors potentially influencing the evolution of an epidemic. A single point prediction is usually not enough, and the uncertainty of that prediction must also be estimated, especially when a forecast is made at an early phase of an epidemic with sparse data. Furthermore, accurate quantification of the uncertainty is essential when determining how much emphasis to put on them, for instance, when making policy decisions. Measuring uncertainty is usually an integral part of statistical models, where the uncertainty of the prediction is generally presented as prediction intervals around a prediction; see an example of projection band for the national-level or state-level COVID-19 death count in the next four months by a team from the Institute for Health Metrics and Evaluation (IHME) [M+20]. There are also many useful uncertainty analysis techniques for mathematical models, such as sensitivity analysis, separate simulations, and many others.

How can we forecast COVID-19? The rise of COVID-19 has triggered novel forecasting methods. Since the beginning of the pandemic, several models have been released, including works from the IHME, the Los Alamos National Laboratory, the Massachusetts Institute of Technology, and Iowa State University. Instead of relying on just a single forecast, the CDC works with partners to bring together several forecasts for COVID-19 infected and death counts in one place; see Figure 4. These forecasts have been developed independently from different teams and shared publicly. The complete set of forecasts is referred to as the ensemble and individual forecasts within it as ensemble members. Collecting and combining forecasts of cumulative deaths for US jurisdictions in standardized, probabilistic formats one to four weeks ahead [R+20] generates real-time, publicly available ensemble forecasts. The ensemble forecast is constructed as an equally-weighted average of forecasts from all eligible models. Ensemble forecasts are provided by Nicholas Reich and coworkers at the ReichLab (`https://viz.covid19forecasthub.org/`).

In Table 1, we summarize the teams who are forecasting the spread of COVID-19 and the methods they are using. Many teams start with SIR and SEIR models and develop new models to analyze COVID-19 data. For instance, FRED is open-source software for modeling infectious diseases. The COVID-19 Simulator uses a validated compartment model to simulate the trajectory of COVID-19 at the state level in the US. Statistical tools and machine learning methods are also considered by some teams. For example, Columbia-UNC considers the forecast of COVID-19 in the survival data analysis framework; and DDS utilizes time series analysis to predict the spread of COVID-19. Several other teams develop hybrid methods by combining mathematical models and statistical models. In Google-HSPH, the authors consider the standard SEIR model and use their end-to-end modeling framework to infer meaningful estimates for undocumented cases. ISU-STEM [WWG+20] incorporates the underlying mechanism of disease spread in mathematical models and nonparametric statistical tools to study the spatiotemporal structure and the effects of covariates as well as future prediction and uncertainty quantification.

Spatiotemporal Epidemic Models

It is well known that the S(E)IR models with random mixing assumptions can overestimate the health service needed by not taking into account the behavioral change and government-mandated action. Spatiotemporal models are able to bring in more information to the epidemic study [LBHU16]. Borrowing the mechanistic rules from the compartment models, we develop a class of new epidemic models based on the flexible nonparametric

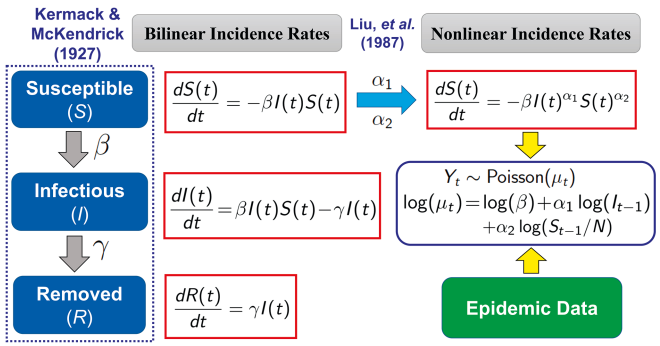

Figure 5. An illustration of the infection model based on SIR.

techniques to reconstruct the spatiotemporal dynamics of the disease transmission. Below, we introduce a novel spatiotemporal epidemic model (STEM) to model and predict the spread and severity of COVID-19 at the area level. For a simple illustration, we describe the STEM-based on the parsimonious SIR models, but it can be extended to the SEIR models with an "exposed" compartment for infected but not infectious individuals.

Modeling the number of incident cases. As illustrated in Figure 2, the SIR model consists of three compartments: susceptible individuals, infectious individuals, and removed (and immune) or deceased individuals. Let $S(t)$, $I(t)$, and $R(t)$ represent the number of individuals in each compartment at time t. Traditional SIR models [KM27] capture the dynamical mechanism of the disease spread under the assumption that the rates of people getting infected are proportional to the number of infectious people and susceptible people in a given population, specifically, $dS(t)/dt = -\beta I(t)S(t)$. Between $I(t)$ and $R(t)$, the transition rate is assumed to be proportional to the number of infectious individuals, which is $\gamma I(t)$. We refer to this type of models as the *SIR model with bilinear incidence rates*; see Figure 5.

However, this assumption may not necessarily hold in reality. For example, given the strict social distancing and self-quarantine policies, the number of effective contacts between infectious individuals and susceptible individuals may decrease at a high infection level. Then, the incidence rate will be lower than a linear rate. Therefore, in our study, we consider a more general SIR model introduced by [LHL87] with two additional parameters α_1 and α_2. We refer to this type of models as the *SIR model with nonlinear incidence rates*. Let Y_t be the new cases at t, and let N be the total population. In mathematics, this system is typically solved by ordinary differential equations. From a statistical point of view, we can treat this system as a generalized linear regression problem and solve it using the maximum likelihood approach.

Group	Category	Spatial Scale	Covariates Included	Unreported Cases	Control Measure
Columbia University (Columbia)	Hybrid	N,S,C	✓		✓
Columbia University and University of North Carolina at Chapel Hill (Columbia-UNC)	Statistical	N,S			✓
COVID-19 Simulator (Covid19Sim)	Mathematical	N,S			✓
COVIDhub - Ensemble (Ensemble)	Hybrid	N,S,C	✓	✓	✓
Discrete Dynamical System (DDS)	Statistical	N,S			
Framework for Reconstructing Epidemiological Dynamics (FRED)	Agent-based	N,S,C	✓		✓
Georgia Institute of Technology (GT-DeepCOVID)	Machine Learning	N,S	✓		
Google and Harvard School of Public Health (Google-HSPH)	Hybrid	N,S,C	✓	✓	✓
Institute for Health Metrics and Evaluation (IHME)	Statistical	N,S			✓
Iowa State Univeristy (ISU)	Hybrid	N,S,C	✓		✓
Johns Hopkins University Applied Physics Lab (JHU-APL)	Mathematical	N,S,C	✓		
Los Alamos National Lab (LANL)	Statistical	N,S,C	✓		
Massachusetts Institute of Technology, Laboratory of Computational Physiology (MIT-LCP)	Machine Learning	N,S,C	✓		
Massachusetts Institute of Technology, Operations Research Center (MIT-ORC)	Mathematical	N,S		✓	✓
Northeastern University, Modeling of Biological and Socio-technical System Lab (MOBS)	Hybrid	N,S			
Oliver Wyman	Mathematical	N,S,C		✓	✓
Rensselaer Polytechnic Institute and University of Washington (RPI-UW)	Mathematical	N,S	✓	✓	
Texas Tech University, Hussain Lab (TTU)	Hybrid	N,S	✓	✓	✓
University of California, Los Angeles (UCLA)	Hybrid	N,S,C		✓	
University of California Merced MESA Lab (UCM)	Mathematical	N,S			✓
University of Michigan (UM)	Hybrid	N,S		✓	
University of South California (USC)	Hybrid	N,S,C		✓	
University of Texas COVID-19 Modeling Consortium (UT)	Mathematical	N,S		✓	
Youyang Gu (YYG)	Hybrid	N,S	✓		✓

Note. Spatial Scale: Prediction is available at National (N), State (S), and County (C) level. (Source: `https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html`.)

Table 1. A subset of COVID-19 models used in the CDC's ensemble forecast.

Now we consider multiple areas. For area i, let Y_{it} be the number of new cases, and let I_{it}, D_{it}, R_{it}, and S_{it} be the number of accumulated active infectious cases, accumulated death cases, accumulated recovered cases, and susceptible population at time t, respectively. Let N_i be total population for the ith area, and denote $Z_{it} = \log(S_{it}/N_i)$. Note that the data observed are heterogeneous and time-varying. A simple "global" model cannot explain the relationships well due to the local features, referred to as spatial nonstationarity. To address such nonstationarity or variability, we need to have sufficiently flexible models to reflect the spatially varying structure within the data. We assume that the determinants of the daily new cases of a particular area can be explained not only by the features of that area but also by the characteristics of the surrounding areas. We use deterministic smooth surface functions in our models to describe the variations and connections among values at different locations. Let $X_{i1}, \ldots, X_{ip}$ be a set of covariates of area i, such as socioeconomic factors, health service resources, and demographic conditions as illustrated in Figure 6.

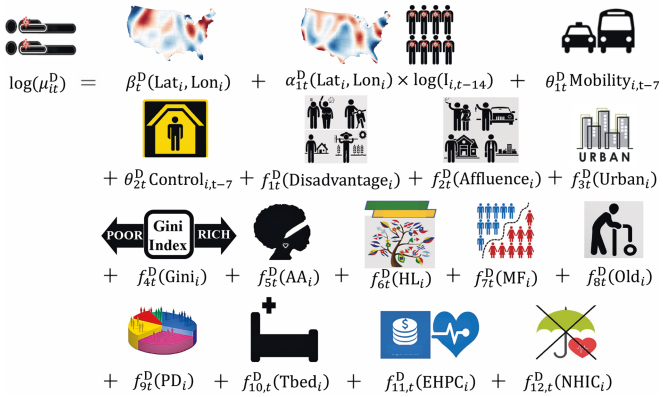

Figure 6. An illustration of the infection model with endemic components.

Let A_{ijt} be the jth dummy variable of actions or measures taken for area i at time t and further denote $\mathbf{U}_i = (\mathrm{Lon}_i, \mathrm{Lat}_i)$ be the GPS coordinates of the geographic center of area i. With the rich data released every day, we can consider the nonparametric method to model the covariates and coefficient functions. The nonparametric

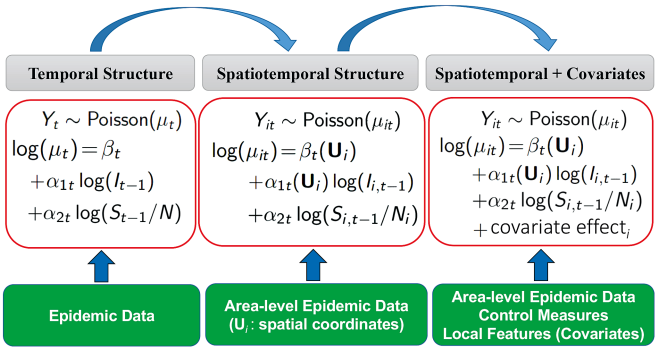

Figure 7. An illustration of the evolution of the spatiotemporal model.

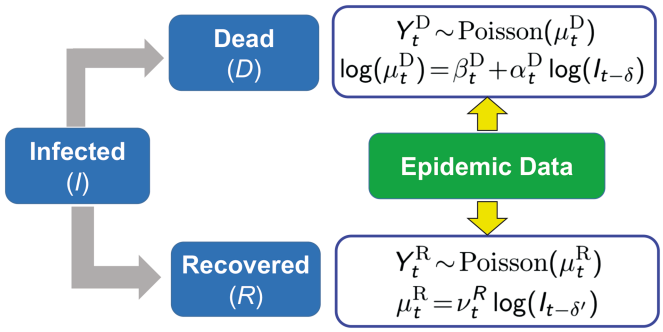

Figure 8. An illustration of the death model and recovery model. Y_t^{D} is the number of new fatal cases and Y_t^{R} is the number of new recovered cases.

structure offers flexibility in assessing the dynamics of the spread at different time points and locations and avoid model misspecification. We assume that the conditional mean value of daily new incident cases μ_{it}^{I} in an area can be modeled as follows:

$$\log(\mu_{it}^{\mathrm{I}}) = \beta_t^{\mathrm{I}}(\mathbf{U}_i) + \alpha_{1t}^{\mathrm{I}}(\mathbf{U}_i)\log(I_{i,t-1}) + \alpha_{2t}^{\mathrm{I}} Z_{i,t-1} + \sum_{j=1}^{p} \theta_{jt}^{\mathrm{I}} A_{ij,t-r} + \sum_{k=1}^{q} f_{kt}^{\mathrm{I}}(X_{ik}), \tag{1}$$

where the θ_{jt}^{I}'s and α_{2t}^{I} are unknown constant coefficients. The bivariate surface functions, $\beta_t^{\mathrm{I}}(\cdot)$ and $\alpha_{1t}^{\mathrm{I}}(\cdot)$, are unknown coefficient functions, which are used to describe the variations and connections among different locations $\mathbf{U}_i$. The univariate functions, $f_{kt}^{\mathrm{I}}(\cdot)$, $k = 1, \ldots, q$, are used to describe the effect of explanatory variables on the new cases, and the parameter r in the $A_{ij,t-r}$'s denotes a small delay time allowing for the control measure to be effective (here we take $r = 7$). For model identifiability, we assume $\mathrm{E}(f_{kt}^{\mathrm{I}}) = 0$, $k = 1, \ldots, q$. See Figure 7 for an evolution of the modeling process.

In model (1), $\exp\{\beta_t^{\mathrm{I}}(\mathbf{U}_i)\}$ represents the transmission rate, and $\alpha_{1t}^{\mathrm{I}}(\mathbf{U}_i)$ and α_{2t}^{I} are the mixing parameters of the contact process at location $\mathbf{U}_i$ and time point t. By including spatially varying coefficients, the determinants of the daily new cases of a certain area involve both the features of this area and the characteristics of its surrounding areas. The above-proposed epidemic model incorporates the nonlinear incidence rates and represents a much wider range of dynamical behavior than the models with bilinear incidence rates [LHL87]. These dynamical behaviors are determined mainly by $\beta_t^{\mathrm{I}}(\cdot)$ and $\alpha_{1t}^{\mathrm{I}}(\cdot)$. When $\alpha_{1t}^{\mathrm{I}}(\cdot)$ and α_{2t}^{I} are both 1, it corresponds to the standard assumption of homogeneous mixing in [DJDH95].

Modeling the number of fatal and recovered cases. One obstacle in the fitting of the model (1) is the lack of direct observations for the number of active cases (I_{it}). Instead, the most commonly reported number is the count of total confirmed cases (C_{it}) and fatal cases (D_{it}). Some public health organizations also release information on recovered cases (R_{it}). Note that $I_{it} = C_{it} - R_{it} - D_{it}$, so we can obtain the information of I_{it} from the data of reported cases. We propose to model D_{it}, R_{it}, and Y_{it} alternatively in a system. As illustrated in Figure 8, there are two outcomes for infected patients: recovery or death. According to the CDC (`https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html/`), the median number of days from symptom onset to death is around 13 $\sim$ 17 days. Therefore, to model the death count, we borrow the information of the previous active cases $I_{i,t-\delta}$, where δ (here $\delta = 14$) is the time delay between illness and death; see Figure 8. We can also introduce the covariates of local features to the model to improve the accuracy. For modeling the number of recovery cases, ideally, if sufficient data for recovered cases can be collected from each area, a similar model can be fitted to explain the growth of the recovered cases.

Although there have been regional, national, and global data on confirmed cases and deaths, not much has been officially reported on recovery. Furthermore, for the states that regularly update the number of recovered patients, the counts can seldom be mapped to counties. Due to the lack of data, we are no longer able to use all the explanatory variables discussed above to model daily new recovered cases; see Figure 9. Instead, we mimic the relationship between the number of recovered and active cases from some compartmental models in epidemiology [SR13].

Therefore, we assume that the conditional mean value of daily fatal cases (μ_{it}^{D}) and recovery (μ_{it}^{R}) in an area can be modeled as follows:

$$\log(\mu_{it}^{\mathrm{D}}) = \beta_t^{\mathrm{D}}(\mathbf{U}_i) + \alpha_t^{\mathrm{D}}(\mathbf{U}_i)\log(I_{i,t-\delta}) + \sum_{j=1}^{p} \theta_{jt}^{\mathrm{D}} A_{ij,t-r'} + \sum_{k=1}^{q} f_{kt}^{\mathrm{D}}(X_{ik}), \tag{2}$$

$$\mu_{it}^{\mathrm{R}} = \nu_t^R I_{i,t-\delta'}, \tag{3}$$

where the θ_{jt}^{D}'s and ν_t^{R} are unknown constant coefficients, $\beta_t^{\mathrm{D}}(\cdot)$ and $\alpha_t^{\mathrm{D}}(\cdot)$ are unknown bivariate coefficient functions, $f_{kt}^{\mathrm{D}}(\cdot)$, $k = 1, \ldots, q$, are univariate functions to be estimated, and δ and δ' are the time delay between illness and death or recovery. For model identifiability, we assume $\mathrm{E}(f_{kt}^{\mathrm{D}}) = 0$, $k = 1, \ldots, q$. The recovery rate ν_t^{R} enables us to make reasonable predictions for future recovered patients counts and provide researchers with the foresight of when the epidemic will end.

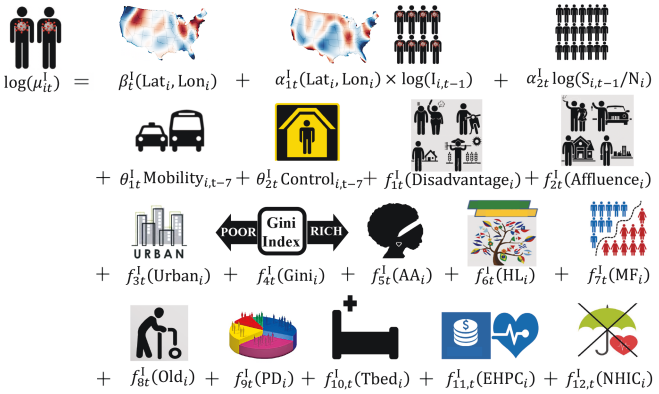

Figure 9. An illustration of the death model with endemic components.

The above equations (1), (2), and (3) together form our STEM, which is based on the foundation of epidemic modeling, but it is able to provide a rich characterization of different types of covariates of local features. Moreover, it accounts for both spatiotemporal nonstationarity and area-level local features simultaneously. It also offers more flexibility in assessing the dynamics of the spread at different times and locations than various parametric models in the literature.

We refer the interested readers to [WWG+20] for the details about how to fit the STEM. Furthermore, the STEM approach enables one to examine the effect of county-level predictors on the spread of COVID-19. In this analysis, we consider the integrated data from 3,104 counties from the 48 mainland US states and the District of Columbia. The epidemic component of the data contains infected, death, and recovered cases from January 21 to September 3, 2020. Our analysis shows that, after controlling for social-economic factors, the percent of persons under 65 years without health insurance has a significant impact on the COVID-19 breakout in the community. We can observe a significant positive relationship between the non-healthy-coverage rate and the infection rate of COVID-19. An under-covered population is much easier to be infected with the virus. Meanwhile, the population density is often considered to have a linear relationship with COVID-19 infection cases in most studies and news reports. Our results are consistent with the intuition. The local healthcare expenditure has a similar impact on COVID-19 infections.

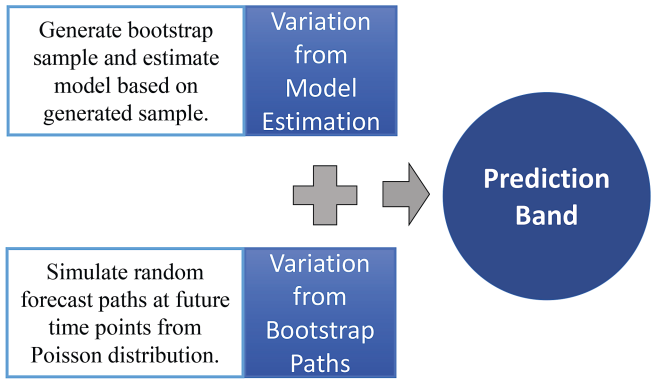

Figure 10. An illustration of prediction band.

STEM-based Prediction

Based on the STEM, we propose an h-step ahead prediction method. The basic idea is that we alternatively update the daily incidence and the total number of cases in each compartment using the proposed STEM method. The computation algorithms are given in [WWG+20]. Specifically, we start with an initial estimation of models (1) and (2). Based on the current number of infected and susceptible people, we can predict the number of new incidence cases, death, and recovery the next day. Then, we update the number of susceptible, dead, recovered, and infectious people. Repeating these steps for h times, we can obtain our h-step ahead prediction.

To evaluate the projection uncertainty, we use the bootstrap method to establish the prediction band. As illustrated in Figure 10, the prediction uncertainty comes from two parts: the estimation variation and the variance from bootstrap paths. In the first part, we generate a bootstrap sample from our estimated model at time points up to t. Based on this sample, we obtain the bootstrap estimator. After repeating for B times, say 1000 times, we could obtain the bias-corrected estimator. In the second part of the bootstrap, we simulate an h-step ahead path based on the bootstrap estimators. The daily incidences follow from Poisson distribution. After the first two steps, we have B paths. By leaving αB paths out, we can obtain the $100(1-\alpha)\%$ prediction band.

Using the methods proposed in [WWG+20], we provide the long-term forecast for both the infection and death count, and this forecast can help us predict the timing of the outbreak peak and the number of health resources required at the peak. Given the lack of reliable recovered data, we treated the daily recovery rate as another variable and considered the values from 0.05 to 0.15. To better

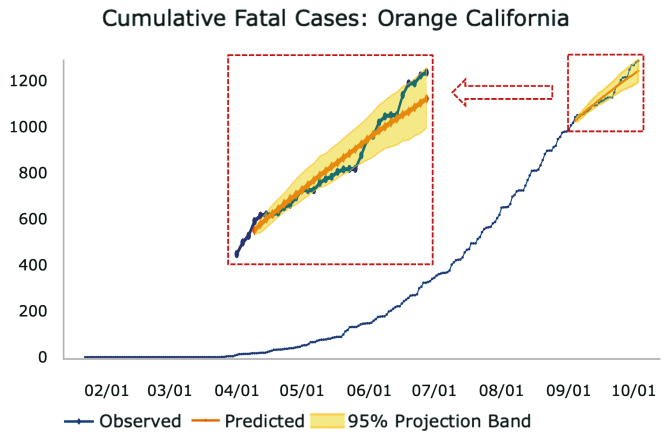

Figure 11. The number of observed and predicted cumulative fatal cases in Orange County, California.

illustrate our research findings, we launched a COVID-19 dashboard (`https://covid19.stat.iastate.edu/`) with multiple R shiny apps embedded. The main map provides the county, state, and national level 7-day forecast together with the related risk analysis. In the Long-term Project tab of our dashboard, we illustrate a long-term forecast of COVID-19 infected and death count up to the county level. Figure 11 shows the reported cumulative fatal cases of COVID-19 and the corresponding one-month-ahead prediction for Orange County, California. The forecast for other states or counties can be found on our dashboard, which is updated constantly. The performance of the proposed forecasting methods is evaluated in [WWG+20].

Conclusions and Discussion

Epidemic modeling holds the key to understanding the course of the epidemic. This paper selectively overviews two main epidemic modeling approaches: mathematical and statistical modeling. Besides, we discuss the challenges of forecasting COVID-19 and discuss some forecasting methods. We also present a novel spatiotemporal epidemic model to quantify the accuracy of forecasts and uncertainty in forecasts.

Nowadays, the number of epidemic models is overwhelming, so how to choose the most appropriate model becomes an essential question, especially under such a pandemic. Unfortunately, there is no one model to work for all the problems, and the model selection can be very complicated, as illustrated in Figure 12. We would like to close this paper with some discussions on how to choose the right model.

Mathematical models or statistical models? The two modeling approaches are complementary but with different starting points and implementation details. The choice of a model is intimately tied to the specific research goal.

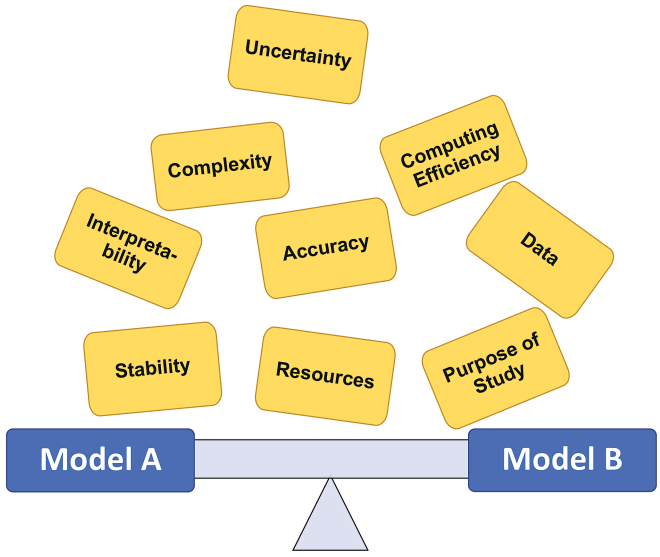

Figure 12. Model selection.

The two approaches can go hand in hand in epidemic modeling for maximum effectiveness. In COVID-19 studies, we have seen a growing number of hybrid methods combining characteristics of traditional mathematical and statistical models.

A more complex or simpler model? One crucial factor that can significantly affect the accuracy of predictions is the complexity of the model to describe the epidemic. It is reasonable to choose models with different complexities during different periods of the pandemic. For example, to model the disease spread at the early stage of the pandemic, a simple model is usually preferred due to the sparsity of the cases. As the disease processes, we gain more knowledge of the transmission and disease progression. There is also confounding heterogeneity in the spread, such as spatiotemporal variation and age-specific risk of severe disease. A complex model with a significant amount of flexibility can capture the heterogeneities and complexity of the underlying process; therefore, it may show advantages by incorporating more information about the disease transmission and local features that might affect the disease's spread. Even within the same modeling framework, we can adjust the complexity of the model used. For example, we can consider different types (i.e., varying or constant) of intercept and slope in the proposed STEM. However, a complex model includes more parameters than a simpler one, and the estimation of many unknown parameters can lead to a greater degree of uncertainty in model predictions.

Stability. The stability (robustness, or in a broad sense includes replicability and repeatability) principle requires that each step in estimation and prediction is stable concerning appropriate perturbations, such as small changes

in the model or data [Yu13]. In epidemiology studies, stability has been shown to be essential to draw reliable conclusions when interpreting results from models. Stability can help evaluate interpretation methods and is mandatory for reliable interpretations.

Interpretability. More complex models can embed more features to capture exciting patterns in the dataset if trained appropriately. Meanwhile, this usually makes them convoluted and more challenging to explain. This issue is usually due to their "black-box" nature, i.e., people do not know how or why the model came up with a particular output. It is challenging to understand what caused it to arrive at this prediction. A good balance between the model's interpretation and underlying trend extraction is another key to picking a model. Governments have been using statistical models to assist in health system adjustment as the virus spreads in the communities. Model interpretability is crucial for them to understand the underlying process. Complex models without clear interpretation make it hard for the decision-makers to extract useful information and take effective NPIs.

Resources. Different models require different types of resources that are available. For example, modern complex machine learning techniques are usually more computationally expensive than traditional models. To realize successful training of the models, the state of the art machine learning algorithm may require a computer cluster, and it may take several weeks to train entirely from scratch. By contrast, traditional simple models can be implemented on a traditional personal computer (PC), and they usually take much less time to train. Therefore, knowing the type of device we could deploy the models, such as the distributed system over the cloud, PCs, or mobile devices, can be crucial in choosing a suitable model. To provide timely updated COVID-19 data analysis results and predictions, we need accurate and (computationally) efficient methods.

Data. Data has always been important to modeling. Some epidemic models make particular assumptions about the structure of the data or the desired results. The choice of the model also depends on the size of good quality data. In general, machine learning models typically need more data than mathematical and statistical models to perform well. To achieve the desired prediction accuracy, neural networks and random forests usually require thousands or millions of observations. In contrast, statistical models often allow inference and make decent predictions on dozens or hundreds of observations. However, if there are very few observations, inference from statistical models can be problematic as well. For example, when analyzing COVID-19, early in a disease outbreak, infected and death cases are rare, and thus, a simple exponential growth curve may be more accurate and stable than many complex models. As the epidemic evolves, surveillance data become abundant and have a higher spatiotemporal resolution. A simple model using the available data might be misleading unless it can incorporate the various steps being taken to slow transmission. A more complex model, such as the spatiotemporal models, can benefit understanding the spread of disease and improve the prediction accuracy.

While epidemic models can be useful tools for tracking and forecasting COVID-19, they also have limitations. Models are only as accurate as of the input assumptions, which depend on continually changing data. Many unexpected scenarios will or may happen in the future. For evidence-based decision making, it is essential to understand the basic model assumptions and limitations before drawing conclusions. The performance of the model is constrained by our knowledge of the virus. We hope the discussions in this paper stimulate new methodological developments in epidemic modeling and forecasting as the pandemic progresses.

ACKNOWLEDGMENTS. The authors would like to thank the editor, associate editor, and two anonymous referees for their constructive comments and suggestions. This work was supported in part by National Science Foundation awards DMS-1916204, CCF-1934884, the Iowa State University Plant Sciences Institute Scholars Program, and the Laurence H. Baker Center for Bioinformatics and Biological Statistics.

References

[ABD+20] Nick Altieri, Rebecca L. Barter, James Duncan, Raaz Dwivedi, Karl Kumbier, Xiao Li, Robert Netzorg, Briton Park, Chandan Singh, Yan Shuo Tan, Tiffany Tang, Yu Wang, and Bin Yu, *Curating a COVID-19 data repository and forecasting county-level death counts in the united states*, arXiv (2020), available at `https://arxiv.org/pdf/2005.07882.pdf`.

[BCCF19] Fred Brauer, Carlos Castillo-Chavez, and Zhilan Feng, *Mathematical models in epidemiology*, Texts in Applied Mathematics, vol. 69, Springer, New York, 2019. With a foreword by Simon Levin. MR3969982

[BD16] Peter J. Brockwell and Richard A. Davis, *Introduction to time series and forecasting*, 3rd ed., Springer Texts in Statistics, Springer, [Cham], 2016. MR3526245

[Bre01] Leo Breiman, *Statistical modeling: the two cultures*, Statist. Sci. **16** (2001), no. 3, 199–231, DOI 10.1214/ss/1009213726. With comments and a rejoinder by the author. MR1874152

[DG99] D. J. Daley and J. Gani, *Epidemic modelling: an introduction*, Cambridge Studies in Mathematical Biology,