SELECT ANONYMOUS REVIEWER REPORT and AUTHOR RESPONSE TO REVIEWERS

A Spatiotemporal Epidemiological Prediction Model to Inform County-Level COVID-19 Risk in the United States

Yiwang Zhou, Lili Wang, Leyao Zhang, Lan Shi, Kangping Yang, Jie He, Bangyao Zhao, William Overton, Soumik Purkayastha, and Peter Song

DOI 10.1162/99608f92.79e1f45e

First Revision

Point-to-Point Responses to Editor

We are grateful to you for your clear, realistic, and constructive guidelines on the revision of our manuscript. Below are our point-to-point responses, beginning with your comment in italics.

Editor's comment: *Please discuss the data limitations, and how you would deal with them if there is no time constraint, e.g., what will be the ideal approaches you would take?*

Our response: We appreciate these guidelines that help us address major critiques on data limitations and data quality when data from public databases are used in the analysis. We agree with the reviewers that both the numbers of infections and deaths are under-reported in any publicly available databases due to the limited capacities of data collection.

One important insight on the under-reporting of infected cases is that the current health surveillance system has difficulty in capturing asymptomatic individuals with light symptoms and/or providing sufficient resources for the COVID-19 diagnostic tests. It is fortunate that after the submission of our manuscript, several states in the US have released the results of serological test surveys for herd immunity, including NY, CA and MS. Although these surveys are of small scale and are limited in some isolated counties, the results are relevant to the underreporting issue and may provide useful information for us to correct the under-reporting of infections. Thus, in this revision we introduce a new compartment of antibody to extend the existing eSIR model, termed as eSAIR model, which is used to estimate the state-level probabilities of an individual being susceptible, self-immunized, infected (prevalence), and removed on the date of the last observed data available. This state-level model generates the initial values that are later utilized by our proposed CA-eSAIR model to project county-level risk. If there are no constraints of time and resources, extensive serological surveys on the proportion of the population with antibody against the COVID-19 at county-level would help our CA-eSAIR make better community-level risk prediction. In this revision, due to these limitations, given the fact that NY has had antibody test surveys with the highest coverage within the state of NY so far, we focus on the county-level prediction in great details in New York state as an illustration for the application of our CA-eSAIR model.

In regard to the under-reporting of deaths or fatalities from the COVID-19, this is not only a public health surveillance issue but also a fundamental issue of ascertainment in epidemiology. In fact, many ordinary deaths cannot be attributed to a defining cause with high certainty. This problem becomes even more complicated in the case of deaths from the coronavirus. Since the medical diagnosis of "COVID-19 disease symptoms" is not fully clinically well-defined, the number of deaths is always a questionable figure. In addition, there are some anecdotally reported cases of at-home deaths potentially linked to the coronavirus, which are however believed to be rather isolated incidences in the US and should not really affect the analysis results much.

In summary, to address the issue of data quality linked particularly to under-reporting, we expand our model by adding a compartment of antibody, and the subsequent analysis is done with the utility of serological survey results from COVID-19 tests in state NY. We did not deal with the issue of death ascertainment and certification, if or not being attributed to the coronavirus. Being a pathological problem, the latter is somewhat beyond the scope of this paper.

Editor's comment: *Given the time constraints, what are the corners you cut and what are the likely consequences of such cutting (e.g., under-predicting, under assessing uncertainties)?*

Our response: We appreciate the flexibility very much allowed for the revision. Despite the significant time constraints, we are still committed to addressing all major critiques in the review. Here is the list of improvements that we have made in this round of revisions.

- We expanded the classical SIR model to a new eSAIR with an addition of a new compartment of antibody. This eSAIR model has never been studied in the literature of infectious diseases before and is a brand-new contribution to address the unique situation of the COVID-19 pandemic in the US. This extension with an antibody compartment is viable because of the coronavirus test surveys, and more of such surveys may come in the future. The novelty is clear: it is the first stochastic infectious disease model that integrates public health survey data into the modeling of infectious disease dynamics and prediction.
- We expanded the SIR model to incorporate social distancing as a transmission rate modifier. The novelty in this expansion lies in the use of time-varying efficacy of social distancing evaluated by real-time data captured using mobile devices. Thanks to several

research institutes in the US that provide timely and data-driven estimates for various social distancing policies across different states.

- We developed an improved procedure to determine inter-county mobility and connectivity. Again, we used some data relevant to personal mobility such as the percent of people having out-country trips and the nearest neighboring airport, instead of only using a simple geo-distance in the previous version of the paper.
- We addressed the uncertainty by propagating the uncertainty of the estimated model parameters into the prediction. Based on the MCMC method, estimation uncertainties, including those associated with the estimated prevalence and estimated proportion of people with antibody, can be easily assessed. We demonstrated how such uncertainty may lead to uncertainties in the prediction.

As seen, we have tried our best to address all major concerns in the review and tried to cut our corners as less as possible. Nevertheless, we did take some approximations and assumptions in the proposed method due to data limitations and time constraints. Here is a list of things that we could do better.

- In our model we assume that an infected person who recovers from his/her infection is immune to the coronavirus within the period of time considered for risk prediction. This assumption is very likely to be true but has not been justified yet.
- Our prediction presented in this paper is based on the limited serological survey data from state New York, which can be improved greatly in the near future when more and more states conduct similar surveys for the antibody against the coronavirus. Nevertheless, our CA-eSAIR model provides a toolbox to incorporate such important results.
- In addition to the self-immunization rate $\alpha_c(t)$, there are two other coefficients that need to be specified, including a temporal transmission rate modifier $\pi_c(t)$ and a spatial inter-county connectivity coefficient $\omega_{cc'}(t)$. These two coefficients are specified by the findings from some other research institutes through mobile device data. Much room exists for future improvements on these two coefficients.

Editor's comment: *Provide uncertainty assessments whenever you can, especially uncertainties in the model selections.*

Our response: We agree with you that using limited data from public surveillance databases to learn a complex spatiotemporal dynamic system of the COVID-19 infection is subject to much uncertainty, and addressing such uncertainty is of great importance. In this proposed system, uncertainty comes from two major sources.

- The first kind is the specification of the transmission rate modifier $\pi_c(t)$ due to social distancing, self-immunization rate $\alpha_c(t)$ based on the limited antibody test survey results, and inter-county connectivity coefficient $\omega_{cc'}(t)$ estimated by mobile devices data. These quantities are not estimated but rather specified from external sources of information, which are subject to some uncertainty. In general, the model selection on these three functions are difficult due to a lack of adequate data. We do consider a tuning step in the form of inter-county connectivity function by minimizing a one-step ahead prediction error.
- The second kind is the MCMC estimation uncertainty for the model parameters in the proposed eSAIR model. The model parameters include $\beta, \gamma, \theta_t^S, \theta_t^I, \theta_t^A$ and θ_t^R. In the MCMC framework, we can calculate the 95% credible intervals for these parameters from 200,000 MCMC draws. In principle, when there are no time constraints, we could project 200,000 risk scores from the CA-eSAIR model, from which we could assess prediction uncertainty. Since this prediction is done at a county-level, a factor of 3109 counties on 200,000 predictions/per county will lead to an extremely high computational cost. To simplify this calculation, we let the propagation of estimation uncertainty into risk prediction in the way that the 95% credible intervals of $\theta^S, \theta^A, \theta^I, \theta^R$ carry over those for the projected risk. This is just a quick and dirty solution that manifests the uncertainty in the risk projection.

Editor's comment: *Be as explicit as you can about limitations and cautions in communicating your results and findings; this is not just about covering your neck (recall most of the predictions are eventually variable), but most importantly to reduce potential harm or unintended consequences because policy makers take your results too literally.*

Our response: We appreciate very much for your advice about the potential limitations and caveats in our methodology and findings. In the revision, we have paid extra caution in our

conclusions and discussion. Essentially, this paper is to provide a spatiotemporal prediction model, an analytic toolbox that may be used by practitioners to perform their own analyses. In addition, we explicitly state assumptions and specifications of model components throughout the paper.

Point-to-Point Responses to Associated Editor

We are very thankful to your insightful comments on our work, which have helped us improve the manuscript. Below we provide our point-to-point responses to each of your comments, beginning with yours in italics.

AE's comment: *The issue of unreported cases seems to throw a wrench in the whole SIR construction.*

Our response: We appreciate your critique. Indeed, we fully agree with the reviewers that both the numbers of infections and deaths are underreported in publicly available databases due to the limited capacities of data collection. In the revision we have tried to tackle this problem by accounting for self-immunization via the development of personal antibodies to the coronavirus.

One important insight on the under-reporting of infected cases is that the current health surveillance system has difficulty in capturing asymptomatic individuals with light symptoms and/or providing sufficient resources for the COVID-19 diagnostic tests (RT-PCT). It is fortunate that after the submission of our manuscript, several states in the US released the results of serological test surveys for herd immunity, including NY, CA and MS. Although these serological surveys are of small scale and are limited in some isolated counties, these results provide useful information for us to correct the under-reporting of infections. Thus, in this revision we introduce a new compartment of antibody to extend the existing eSIR model, termed as eSAIR model, which is used to estimate the state-level probabilities of an individual susceptible, self-immunized, infected (prevalence), and removed on the date of the last observed data available; these serve as the initial values being utilized by our proposed CA-eSAIR model to project county-level risk. If there are no constraints of time and sources, extensive surveys on the proportion of the population with COVID-19 antibody at county-level would help our CA-eSAIR make better community-level risk prediction. In this revision, due to these limitations, given the fact that NY has had an antibody test survey with the highest coverage so far, we focus on the county-level prediction in great detail in New York state as an illustration for the application of our CA-eSAIR model.

In regard to the under-reporting of deaths from COVID-19, this is not only a public health surveillance issue but also a fundamental ascertainment issue in epidemiology. In fact, many ordinary deaths cannot be attributed to a defining cause with high certainty. This problem becomes even more complicated in the case of deaths from coronavirus. Since the medical diagnosis of "COVID-19 disease" is not clinically well-defined, the number of deaths is always a questionable figure. In addition, there are some anecdotally reported cases of at-home deaths potentially linked to the coronavirus, which are however believed to be rather isolated incidences in the US and should not really affect the analysis results much.

In summary, to address the issue of data quality linked particularly to under-reporting, we expand our model by adding a compartment of antibody, and the subsequent analysis is done with the utility of survey results from COVID-19 tests in NY. We did not deal with the issue of death ascertainment and certification, if or not being attributed to the coronavirus. Being a pathological problem, the latter is somewhat beyond the scope of this paper.

AE's comment: *Control measures change the regime of the disease propagation, but this does not seem to be taken into account.*

Our response: Thanks for pointing out this important issue. Indeed, in our original development of the eSIR model, we already noticed the importance of varying regimes of the infection dynamics due to different preventive measures. In the revision, similar to our previous work, we included a transmission rate modifier $\pi_c(t)$ in our CA-eSAIR model to take into account the public health interventions. This function decreases the population-level transmission rate due to the main public health intervention (i.e. social distancing) in the US. The state-level effectiveness of social distancing is obtained from the published values by the Transportation Institute at the University of Maryland (https://data.covid.umd.edu/) derived from the cell phone mobile data. Our CA-eSAIR model allows county-level value of social distancing. But without access to the high-resolution data from the UMD webpage, we have to use a state-level value. This may be improved in the future with higher-resolution data available.

AE's comment**:** *Similarly, the transmission between counties is taken to depend only on the distance, but other factors are likely important.*

Our response: Thanks for pointing out another important issue in our prediction model. Indeed, it is difficult, in general, to determine an objective connectivity coefficient $\omega_{cc'}(t)$ due to our limited knowledge and limited data source available. This problem itself seems to define a research area, and we would welcome any new ideas and approaches to improve the specification of this inter-county connectivity function. Being said, in this revision, we have tried our best capacity to improve the previous geo-distance-based connectivity. This improvement lies in the inclusion of two additional factors in the function, including 1) the percentage decrease in encounters density compared to national baseline, which is obtained from the social distancing scoreboard of the Unacast company (https://www.unacast.com/covid19/social-distancing-scoreboard) based on human mobility data, and 2) the information of airports in the US (e.g. annual enplanements) and their accessibility to each county. In addition, we consider a factor η to tune the scale of the travel distance $r(c, c')$ by minimizing the one-day ahead prediction error. In the future, we hope to collaborate with experts in this field for a further improvement in defining the connectivity coefficient $\omega_{cc'}(t)$.

AE's comment: *Goodness-of-fit is not really assessed. Looking at past data, how accurate are the predictions?*

Our response: We did a quick check on the eSAIR model for its goodness of fit on individual state-level data, where MCMC worked reasonably well with observed numbers of infected and removed cases falling in the 95% credit intervals of the in-sample prediction. More importantly, in the revision we focused on examining the one-day ahead prediction accuracy based on the sum-of-squared prediction errors (SSPE), where the error is the difference between the predicted number of infections and the corresponding observed number of infections on day $t_0 + 1$ in a county. This prediction accuracy is examined over 39 continental states and Washington DC for which the MCMC passed convergence diagnosis. For other continental states where the MCMC failed to convergence due to inadequate data (e.g. very low number of deaths), the national average estimates of the model parameters are used as the initial values for prediction.

AE's comment: *The article requires some serious editing.*

Our response: We have incorporated all suggested changes from the reviewers. Also, a native speaker has edited the language throughout this paper.

Point-to-Point Response to Referee 1

We appreciate your review and constructive comments that helped us improve our manuscript. In what follows we present our point-to-point response to each of your comments listed in italics.

Referee's comment: *In the epidemic time series you have the observed and the underlying latent process θ_t, where $E[Y_t] = \theta_t$, from the previous paper. The problem is that observed infections are only a fraction of actual infections, many of which are unreported. The classic SIR model formulation (here the series θ_t) refers to actual, not reported infections. The problem of estimating the under-reporting rate is fairly difficult in epidemiology and hasn't been tackled here at all. Thus, the model, as presented, is too simplistic and the results may be wrong.*

Our response: Indeed, the problem of estimating the under-reporting rate is difficult. In the revision we have tried to tackle this problem by accounting for self-immunization via the development of personal antibody to the coronavirus under the assumption that the substantial majority of infections with no hospitalization (and thus not recorded in the database) are self-recovered with antibodies. Along with this line, we extend our model by adding a new compartment of antibody, termed as eSAIR model that is used for the prediction.

One important insight on the under-reporting of infected cases is that the current health surveillance system has difficulty in capturing asymptomatic individuals with light symptoms and/or providing sufficient resources for the COVID-19 diagnostic tests. It is fortunate that after the submission of our manuscript, several states in the US released the results of serological surveys for herd immunity, including NY, CA and MS. Although these surveys are of small scale and are limited in some isolated counties, these survey results provide useful information for us to correct the under-reporting of infections. Thus, in this revision we introduce a new compartment of antibody to extend the existing eSIR model, termed as eSAIR model, which is used to estimate the state-level probabilities of an individual susceptible, self-immunized, infected (prevalence), and removed on the date of the last observed data available, and these estimates serve as the initial values being utilized by our proposed CA-eSAIR model to project county-level risk. If there are no constraints of time and sources, extensive surveys on the proportion of the population with COVID-19 antibody at county-level would help our CA-eSAIR make better community-level risk

prediction. In this revision, due to these limitations, given the fact that NY has had an antibody test survey with the highest coverage so far, we focus on the county-level prediction in great detail in New York state as an illustration for the application of our CA-eSAIR model.

In regard to the under-reporting of deaths from COVID-19, this is not only a public health surveillance issue but also a fundamental issue in epidemiology. In fact, many ordinary deaths cannot be attributed to a defining cause with high certainty. This problem becomes even more complicated in the case of deaths from coronavirus. Since the medical diagnosis of "COVID-19 disease" is not clinically well-defined, the number of deaths is always a questionable figure. In addition, there are some anecdotally reported cases of at-home deaths potentially linked to the coronavirus, which are however believed to be rather isolated incidences in the US and should not really affect the analysis results much.

In summary, to address the issue of data quality linked particularly to under-reporting, we expand our model by adding a compartment of antibody, and the subsequent analysis is done with the utility of survey results from COVID-19 tests in NY. We did not deal with the issue of death certification, if or not being attributed to the coronavirus. Being a pathological problem, the latter is somewhat beyond the scope of this paper.

Referee's comment: *The model makes no provision for the introduction of control measures. These would change the basic parameter β (and hence R_0). In particular for the data application I would like to see either a time-varying β, e.g., that switches between two regimes, or fitting the model as it is now to two different time segments, one before control measures were introduced largely, and one after their introduction (maybe mid-March?). Intuitively, you should get a smaller β after their introduction. The first solution is probably more involved, so I won't require it. But I think the second one is doable.*

Our response: Thanks for pointing out this important extension that leads to an improvement. Indeed, in our previous development of the eSIR model, we already noticed the importance of varying regimes of the infection dynamics due to different preventive measures. In the revision, similar to our previous work, we included a transmission rate modifier $\pi_c(t)$ in our CA-eSAIR model to take into account public health interventions. This function decreases the population-

level transmission rate due to the main public health intervention (i.e. social distancing) in the US. The state-level effectiveness of social distancing is obtained from the published values by the Transportation Institute at the University of Maryland (https://data.covid.umd.edu/) derived from the cell phone mobile data. Our CA-eSAIR model allows county-level value of social distancing. But unfortunately, we do not have access to the high-resolution data from the UMD webpage. So, we have to use a state-level value. This may be improved in the future with better county-level data available.

Referee's comment: *In the first paragraph of Section 2.1, you define θ_t as a probability, later as prevalence in equation (1). This terminology can be confusing since people often think of probability as being a parameter of the model. However here, I think you mean the fraction of individuals infected. You could also make it clear from the beginning that θ has more interpretations throughout.*

Our response: In the proposed state-space eSAIR model, θ_t is a probability (or a population-level proportion/fraction) of being susceptible(S)/self-immunized(A)/infected(I)/removed(R) at a given day *t*, and they follow a Markov process over time. Being the latent states, these probabilities can be estimated from the MCMC method, and their estimates are used as initial values to make risk prediction through the CA-eSAIR model. The θ_t can be understood as the prevalence of being susceptible/self-immunized/infected/removed. We have made clear in the paper that θ has more interpretations.

Referee's comment: *Figure 1 is taken from your previous paper, but the caption has to explain more of what's going on here. For text below Figure 1, again, you borrow heavily from the previous paper. But the reader needs to know what's important for the understanding of this paper, and what isn't. If something can be easily introduced in an in-line equation, do that, if relevant.*

Our response: We have added a thorough review of the eSIR model in Section 2.1 before we introduce our new eSAIR model in this revision. Consequently, Figure 1 and its caption are revised with the inclusion of antibody compartment.

Referee's comment: *In the end of section 2.2, normally more justification would be needed for the equations, like for instance, in theory θ^I could become larger than one. Does your data fitting check for these cases?*

Our response: Thanks for pointing out this issue. In theory, it is possible that θ^A, θ^I or θ^R become larger than one and θ^S becomes smaller than zero, which may occur especially in the case of a very long-term risk prediction, *say*, a half-year ahead prediction, since we have a substantially large number of terms in the summation. Given the fact that the pandemic of COVID-19 evolves so fast with constantly varying regimes of the disease, we would not like to consider risk prediction longer than one month. So, practically this technical issue is very unlikely to occur. Nevertheless, to make it technically correct, we confine $\theta^S, \theta^A, \theta^I, \theta^R$ within [0,1] in our software package.

Referee's comment: *Provide a justification for equations (4) & (5). Is there an interpretation for this risk score or is it just a heuristic measure?*

Our response: We consider an intuitive way to define the risk of infection as a cumulative chance during the prediction period as P(infection at or before day *t*) = P(infection at day 1) + P(infection at day 2 | not infection at day 1) + …+ P(infection at day *t* | no infection before day *t*-1). We have made a remark in the paper about the intuition behind this definition.

Referee's comment: *For section 3.1 that related to my second concern, there is a gap in the literature when it comes to the estimation of R_0 after controls were introduced. If you could obtain and report such an estimate, I think it would be very useful.*

Our response: As pointed out above, in the proposed eSAIR model we have introduced a transmission rate modifier function $\pi(t)$ to adjust the transmission rate β using the mobile device data. Therefore, the estimate of the basic reproduction number R_0 is obtained by adjusting control measures/human interventions. A detailed introduction of the new eSAIR model with a time-varying transmission rate modifier $\pi(t)$ is given in Section 2.1.

Point-to-Point Response to Referee 2

We would like to express our gratitude for your review and careful reading of our paper. Your insightful comments in the review report helped us improve greatly on our manuscript. Below are the listed point-to-point responses to each of your comments presented in italics.

Referee's comment: *If I have understood everything right, the CA-SIR model seems to incorporate transmission from neighboring counties in an entirely arbitrary way, which is not driven by data, and even worse, doesn't even seem to be tunable. At its most basic, we end up with:* $\omega_{cc'}(t) = exp\{-d(c,c')\}$*, where* $d(c,c')$ *is the geo-distance between* c *and* c'*. But there is no parameter that allows for the between county transmission effect to vary. My understanding is that the analysis is conditional, with the eSIR model being fitted first to data, giving the estimates of* β *and* γ*, and then these estimates are being plugged into the CA-SIR model, so at no point does data have any effect on what the* ω *effect is. At least if the model was jointly estimated the spatial effect would have some bearing on* β *and* γ *(although this would be a very poor way of incorporating the spatial effect). Doesn't this mean that the predictions made from the CA-SIR model are completely arbitrary? If I simply change the scale of the geo-distance I am measuring, don't the results change? If this is the case, I would have very little faith in any of the predictions/forecasts made by the CA-SIR model. Why cannot the CA model be calibrated to the county-level time series, even if done conditional on the results from fitting the eSIR model?*

Our response: Thanks for your insightful critique that arises several important points, which will be addressed one by one below.

- The CA-eSAIR model (formerly CA-SIR model) is a model for risk prediction based on the results obtained from fitting the state-level eSAIR model. In principle, fitting the eSAIR model can be performed with county-level data if such data are available in good quality. In reality, the county-level data in most counties are sparse and less reliable (e.g. an infectious person living in a county died in a hospital that is located in another county). Thus, it is more robust to run an epidemiological model at a relatively large population like in a state where the preventive measures are supposed to be homogeneous across the counties within the specific state.

- We totally agree with you that there is a certain level of subjectivity in the specification of the inter-county connectivity coefficients $\omega_{cc'}(t)$, which requires knowledge beyond the domain of public health. However, a good specification of $\omega_{cc'}(t)$ is needed in the county-level prediction using the spatiotemporal CA-eSAIR model. Indisputably, the better quantification of inter-county connectivity, the better prediction of county-level infection risk. There may not exist the best quantification for this coefficient. But some better versions exist. One of the major focal areas for improvement in this revision is to improve the definition of $\omega_{cc'}(t)$ in order to reduce the subjectivity (or arbitrariness).
- Our improvement is made in the following three domains: 1) We borrow information of inter-county mobility derived from the score of the percentage decrease in encounters density compared to the national baseline, obtained from the social distancing scoreboard of the Unacast company (https://www.unacast.com/covid19/social-distancing-scoreboard) based on human mobility data; 2) we incorporate the location and the annual enplanements of over 300 commercial airports in the continental US into the definition of inter-county range, which combines both the information of geo-distance and air-distance; and 3) we add a tuning parameter to obtain an optimal scaling of the inter-county range by minimizing the one-day ahead prediction error. Therefore, we have greatly reduced the arbitrariness on the formation of the inter-county connectivity. We also comment that a more desirable data-driven specification of $\omega_{cc'}(t)$ itself presents a big research topic, which is beyond the scope of this paper.

Referee's comment: *The standard of the writing is really rather poor. I started correcting this, but I realized they were just too many errors – I think this manuscript really needs to go to a copy editor before it is published.*

Our response: We appreciate this suggestion. We would appreciate the help of a copy editor to improve the writing.

Referee's comment: *In the introduction the authors state that COVID-19 is one of the most lethal communicable infectious diseases– this may be true in terms of a fully susceptible population, but I don't think it's true on an individual level.*

Our response: Thanks for this good point. We have revised the sentence as "Being one lethal communicable infectious disease, ..."

Referee's comment: *In section 2.1 t (time) seems to be continuous, and in section 2.2 it seems to be discrete. This difference should be emphasized, making it clearer in the text. Is this why t is a subscript in the eSIR model and as the function input in parenthesis in the CA part? If so, please make it clear.*

Our response: In the mathematical/statistical model of eSAIR in Section 2.1, θ_t represents a continuous-time dynamic system that emits observed data of the daily number of infections and removed cases captured by the public database. Since our prediction is performed on a daily basis, namely discrete time points in the unit of day, we express the CA-eSAIR model in a form of discrete-time stochastic process for convenience. Indeed, this CA-eSAIR may also be written as a continuous-time dynamic system. We have clarified this point in Section 2.2.

Referee's comment: *The fact that the authors are using observed daily time series of confirmed infected cases and removed cases to calibrate their SIR seems to imply an underlying assumption that testing rates are homogeneous over space and time. We know this is not true, and it should be made clear or this could lead to misleading results.*

Our response: Yes, it is an important implicit assumption associated with data collection. We have added this assumption in the introduction of the eSAIR model in the fourth paragraph of Section 2.2.

Referee's comment: *I did not understand the relevance of the transmission modifier $\pi(t)$ and $\phi(t)$ to the model here. Are they being used at all in the modelling in this manuscript? And if so, how are they introduced to the model?*

Our response: To make the paper self-contained, we have added more details of the eSAIR model in Section 2.1, where both the transmission rate modifier $\pi(t)$ and the self-immunization rate $\alpha(t)$ are discussed. Similarly, the CA-eSAIR model is nuanced in Section 2.2. Factor $\phi(t)$ is the rate of quarantine initially introduced in our eSIR paper to address the super stringent stay-home policy in China, which is not applicable in the US. Thus, in this paper we do not consider quarantine ($\phi(t)$ is not included in this paper). However, to address the under-reporting issue in the US public

database, we add a new compartment of antibody to reflect the real situation of COVID-19 outbreak in the US.

Referee's comment: *MCMC is being used here to estimate the SIR model. Is this within a Bayesian framework? What priors are being used? And in the cellular automata model is there no information about the uncertainty of the SIR parameters being used? – is this presumably extremely straightforward to obtain via the MCMC?*

Our response: You raised an interesting question.

- First, yes, we use a Bayesian framework for the implementation of MCMC to fit the eSAIR model. One could run MCMC in the spatiotemporal CA-eSAIR model for a total of 3109 counties, but we did not do it because, as in our response to your first point, the county-level data in most of the counties are sparse and less reliable. Thus, it is more robust to run an epidemiological model at a relatively large population utilizing state-level data.
- Second, we agree with you that the quantification of prediction uncertainty is of great importance, which has been added in the revision. In the MCMC framework, we can calculate the 95% credible intervals for these parameters from 200,000 MCMC draws. In principle, when there are no time constraints, we can project 200,000 risk scores using the CA-eSAIR model, each from one MCMC draw from the eSAIR model. Consequently, we can assess the prediction uncertainty. Since this prediction is done at a county-level where each projection involves 3109 counties, multiplying a factor of 200,000 will lead to an extremely high computational cost. To simplify this calculation, we demonstrate the propagation of estimation uncertainty into risk prediction in the way that the 95% credible intervals of $\theta^S, \theta^A, \theta^I, \theta^R$ carry over those uncertainties for the projected risk. This is just a simple solution that manifests the uncertainty in the risk projection.

Referee's comment: *I couldn't find anywhere where the "geo-distance" d is defined for the model. Is it the Euclidean distance between county centroids? The minimum distance between counties?*

Our response: The geo-distance we calculated between county c and c' is the geodesic distance, which is the shortest distance between two points on the surface of an ellipsoidal model of the earth. The default algorithm used for the calculation is given by Karney (2013)[1]. We have added this reference in the revision.

1. Karney, C. F. (2013). Algorithms for geodesics. *Journal of Geodesy*, *87*(1), 43-55.

Referee's comment: *I think the issue of whether joint modelling of the CA and eSIR model is feasible/desirable should be addressed in the discussion section. What are the downsides to the conditional approach taken here (I assume the downsides relate primarily to computational burden)?*

Our response: We totally agree with you on the exercise of caution on reporting our prediction results, which are indeed obtained under various assumptions, in addition to the issue of data quality. Two major limitations that are discussed in the discussion section are given as follows. We did take some approximations and assumptions in the proposed method due to the limited data and time constraints. Here is a list of things that we could do better.

- In our model, we assume that an infected person who recovers from his/her infection is immune to the coronavirus within the period of time considered for risk prediction. This assumption is very likely to be true but has not been justified yet. Our prediction presented in this paper is based on the limited serological survey data from state New York, which can be improved greatly in the near future when more and more states conduct similar surveys for the coronavirus tests. Nevertheless, our CA-eSAIR model provides a toolbox to incorporate such important results.
- In addition to the self-immunization rate $\alpha_c\,(t)$, there are two other coefficients that need to be specified, including a temporal transmission rate modifier $\pi_c\,(t)$ and a spatial inter-county connectivity coefficient $\omega_{cc'}\,(t)$. These two coefficients are specified by the findings from some research institutes through mobile device data. Much room exists for future improvements on these two coefficients.

Referee's comment: *What are the estimates used for the basic reproduction number, β and γ? Are they posterior means? Why is there no uncertainty interval reported?*

Our response: Yes, parameters of β and γ are estimated by the posterior means. Then, we use their posterior means to calculate the basic reproduction number R_0. The uncertainty for R_0 is given in Table 1, where the uncertainty for the estimates of β and γ are also listed.

Point-to-Point Response to Referee 3

We would like to express our gratitude for your review and careful reading of our paper. Your insightful comments in the review report helped us improve greatly on our manuscript. Below are the listed point-to-point responses to each of your comments presented in italics.

Referee's comment: *The methods proposed appear sound, however there is no real discussion of data quality which greatly undermines any results from the model. The proposed eSIR model that is used as a basis for this approach was developed for use in Hubei, where it is plausible that -- although socioeconomic and other factors may differ across the province -- the testing strategy might be uniform. This is certainly not the case in the US, where testing strategies and rates vary dramatically across states. I cannot see how the proposed model accounts for differential sampling across the different measurement locations and cannot understand how results can be valid if the error or sampling variability or selection bias varies across sampling units and even within a sampling unit over time.*

For example, while it is almost surely the case that New York state has the greatest number of cases, to say that it has "33.4% of the confirmed infections of the country" without mentioning that because of their high rate of infection, testing has been much more aggressive there than elsewhere is a serious omission. Even death counts are problematic and less comparable, as, for example, some places attribute nursing home deaths or probable (but unconfirmed) cases as COVID-19 deaths, whereas others do not. Deaths at home may also be counted in different ways in different locations or be updated at different speeds.

Our response: Thanks for these insightful remarks on various issues related to data quality. We agree with you that the issue of under-reporting is closely related to the testing strategies and rates. More than two months later after the first reported COVID-19 case in the US, the testing selection bias has been greatly mitigated by better availability of tests via driving through testing stations and other walk-in testing clinics, although such issue with no doubt remains part of the sampling bias. Below we list a few points that we did in the revision to address your critique.

- We extend the eSIR model to an eSAIR model by including a new compartment of antibody. This extension is appealing to incorporate the serological testing survey results into the infection dynamics system to account for the proportion of people who were

infected but did not get chance for the coronavirus RT-PCR diagnostic test, and now are self-immunized with the COVID-19 antibody. We believe that this new addition of antibody compartment can greatly address the under-reporting issue raised in your remarks.

- Indeed, we fit the eSAIR model state by state to estimate state-specific model parameters, under the assumption that the testing rate is homogeneous within a state. Although there exists heterogeneity of the test rate across counties in a state, such inter-county differences are deemed much smaller than the inter-state differences. We added some comments in the revision (see the fourth paragraph of Section 2.2) to address this issue.

Referee's comment: *It would have been of interest to see how well the model actually performs, using predictions from, say March 15 or March 30 and then comparing predicted with observed counts.*

Our response: Thanks for this good suggestion. In this revision, we report the sum-of-squared prediction errors (SSPE), which is an error being the difference between one-day ahead predicted and observed proportions of infections. See and Figure A3. See additional correspondences with the Dataviz Editor.

Referee's comment: *The grammar errors and typos.*

Our response: We have tried our best to identify and correct the grammar errors and typos, including those mentioned in the review reports by the other referees. The final version of the revised manuscript has been proofread by a native speaker.

Point-to-Point Response to Referee 4

We would like to express our gratitude for your review and careful reading of our paper. Your insightful comments in the review report helped us improve greatly on our manuscript. Below are the listed point-to-point responses to each of your comments presented in italics.

Referee's comment: *One weakness of the paper is the ad-hoc method of parameter estimation. Non-spatial SIR models are fit separately for each state with some adjustment made for counties above the average infection level in the state. No assessment is made of parameter uncertainty and the effect it will have on predictions.*

Our response: We appreciate your critique, which helps us clarify our analysis strategies used in the paper.

- Our parameter estimation is done systematically and consistently using the classical state-space model and the standard MCMC algorithm. This Bayesian estimation method in the framework of state-space models has been extensively developed and applied widely in the literature since the early 1990s. We did not propose new parameter estimation methods.
- We have clarified the reasons behind the choice of running state-specific eSAIR model in the fourth paragraph of Section 2.2. They are (i) the testing strategies and rates are rather different over states; (ii) fitting a county-level spatiotemporal model is challenging due to insufficient county-level data in some counties. Given potential data quality issues, it turns out that the state-level analysis provides a more reliable estimation of the model parameters.
- Using the Bayesian framework, we can get the 95% credible interval to quantify the estimation uncertainty from 200,000 MCMC draws. In principle, when there are no time constraints, we can project 200,000 risk scores using the CA-eSAIR model, each from one MCMC draw from the eSAIR model. Consequently, we can assess prediction uncertainty. Since this prediction is done at a county-level where each projection involves 3109 counties, multiplying a factor of 200,000 will lead to an extremely high computational cost. To simplify this calculation, we demonstrate the propagation of estimation uncertainty into risk prediction in the way that the 95% credible intervals of the $\theta^S, \theta^A, \theta^I, \theta^R$ carry over those uncertainties for the projected risk. This is just a simple solution that manifests the uncertainty in the risk projection.

Referee's comment: *A second weakness of the paper is the lack of assessment of the predictions made. The map in Figure 4 has abrupt changes in infection risk at each state border. Is it true that all of the counties to the east of the Texas-Louisiana border have very high risk and all of the counties neighboring them to the west have low risk? Or is this simply an artifact of the way parameters estimated?*

Our response: Thanks for this good question. Unfortunately, we do not know if the discrepancies over neighboring counties on the state borders are artifact. In the meanwhile, we do not know if a certain smoothing technique should be used on counties along state borders as control measures and testing rates are state-specific. As a technical improvement in this revision, we check the prediction accuracy using the sum-of-squared prediction errors (SSPE), one error being the difference between one-day ahead predicted and observed proportions of infections. See page 6.

Referee's comment: *Are the predictions made on 14 April close to the infection counts observed since then?*

Our response: We add the prediction accuracy assessment in the revision based on the SSPE between one-day ahead predicted infection proportions and observed proportions. In the submitted revision, one-day ahead risk prediction of May 3rd is performed with the data collected up to May 2nd. See the third paragraph of Section 2.2. Additional comments may be found in our correspondences with the Dataviz Editor.

Referee's comment: *Is this spatiotemporal model more accurate than fitting individual SIR models to each state?*

Our response: We did fit these two scenarios and found that the spatiotemporal CA-eSAIR model gives SSPE = 4.13e-08 at the state level, while the state-level eSAIR model has SSPE = 1.59e-06 at the state level. The former is 38.5 times more accurate than the latter in reducing the prediction error. See some additional correspondences with the Dataviz Editor.

Second Revision – May 25, 2020

Referee's comment: *The results produced in Figure 4 do not correspond closely to the raw rates mapped in https://coronavirus.jhu.edu/us-map. This paper's risk estimates change abruptly at state borders, most notably in Idaho and Nebraska, whereas the observed rates on https://coronavirus.jhu.edu/us-map show more changes within states than at state borders. The north-east of New York state has observed fairly low incidence, whereas the risk in Figure 4 is in the highest bin. In their response, the authors state "Unfortunately, we do not know if the discrepancies over neighboring counties on the state borders are artifact." I must conclude that the biggest signal the model is producing, large state-level changes, is indeed an artifact and the results produced in this paper must be viewed with suspicion.*

Our response: We appreciate your point of view on the issue of non-smooth projected risk scores over some counties along the state borders. Thinking more about a potential solution overcoming abrupt changes across state borders, we realize that this may not be that simple and is certainly beyond the capacity of our current methodology. We regard this as a limitation in this paper. This limitation pertains to the initial values generated from a state-level eSAIR analysis in that we assume both control measures and testing policies/strategies are state-specific. Such within-state homogeneity is also used in the prediction. Consequently, the resulting intrastate projected risks seem to be more homogeneous than the inter-state projected risks, and some counties at state boarders appear to have noticeable discrepancies in their projected risks. It is of interest in a future work to discern the true inter-state differences from artifacts in the risk prediction over the border counties. To do this, we may fit the eSAIR model for the border counties in addition to the current analysis with within-state counties. Alternatively, we may perform a local smoothing (e.g. spatial moving average) for the risk scores of counties at state borders if the need of this procedure is deemed necessary, judged by a certain objective criterion. This deserves a further exploration. We have added this additional discussion in the conclusion section of the paper (see the second last paragraph) to address the inter-state differences in the projected risks.

Dataviz Editor's comment: *Am I just being ultra-picky or did your four referees have so much else to do that they did not look at the figures? I searched the response letter for "fig" and found nothing relevant, but did come across this interesting exchange: "Referee's comment: It would*

have been of interest to see how well the model actually performs, using predictions from, say March 15 or March 30 and then comparing predicted with observed counts. Our response: Thanks for this good suggestion. In this revision, we report the sum-of-squared prediction errors (SSPE), which is an error being the difference between one-day ahead predicted and observed proportions of infections. See page 15 and Figure A5." That's a really good comment of the referee, but the answer is strange, since there is no Figure A5. I'm also puzzled by the SSPE value. It seems low, until you remember that it is a sum of squared small values and only for a 1-day ahead forecast. I would need more information to judge whether this is good or not (a map of the differences would be good). Since the paper discusses 7-day ahead forecasts they should also carry out a 7-day ahead test.

Our response: This is a typo, which has been corrected. It is actually Figure A3. We have included a new figure (Figure A3b) in the revision to show the squared error values for the one-day ahead risk prediction for each county in state New York. This is an example to illustrate the tuning results. We appreciate your suggestion of carrying out a 7-day ahead test. Although it is possible to carry out a 7-day-ahead test, we did not choose to do so. Instead, we did a one-day ahead test in Figure A3b, for the consideration of the prediction uncertainty. In our view, validation based on a single prediction value may not be very rigorous. Given the pandemic in the US evolves in a fast pace, with a lot of uncertainty and heterogeneity in such a process, a prediction interval with the quantification of prediction uncertainty is deemed a better validation approach in this context. It, however, requires substantially more computing resources to calculate a posterior prediction interval, which will be studied in our future projects.

Third Revision – May 30, 2020

Editor's comment: *As for (4), it does seem that your response got the logic reversed - if a single test is not reliable, then you should do more than not doing it at all. Is this test very time consuming? It does seem to me that some model testing/validation is needed. Speaking of reliability, I noticed that you have an MCMC diagnosis test in place, but I cannot find what it is, and what criterion's you use to determine the pass - Gelman-Rubin's R? Did you run multiple chains? I was a bit worried about your use of 200,000 (on page 6), without saying how you choose it. I learned this in a hard way many years ago when I was still in Chicago. A student ran an MCMC for 100,000, and all looked fine. Somehow, I had a gut feeling that's not enough and hence kept asking him to run more. By the time it hit 370,000, the chain suddenly changed to an entirely different mode! Whereas there is always a danger of running MCMC, the last thing you want is that someone else uses exactly your code and produces very different results. So please exercise your ultimate due diligence to avoid that (e.g., by running multiple chains from really different starting points, including some extreme ones).*

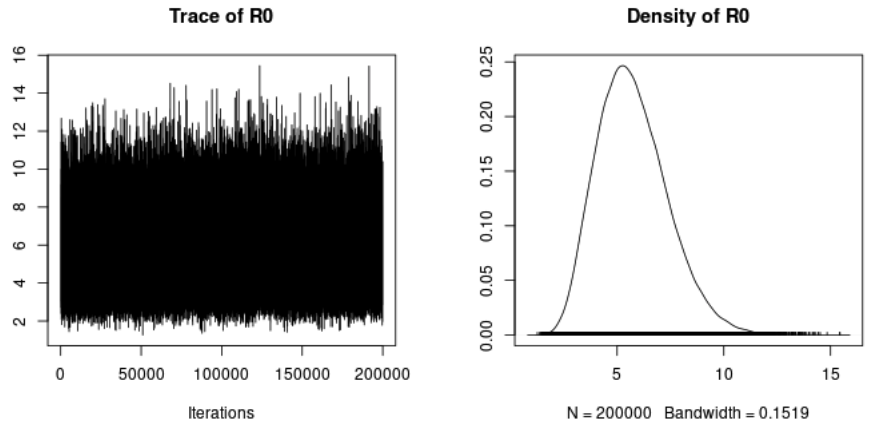
Our response: Thanks for these valuable suggestions. We have incorporated them (in fact two major points) in the revision.

- In regard to the first suggestion on the prediction error, we have included test results for the 1 to 7-day ahead prediction accuracy of the estimated infection prevalence at county-level. The following table gives the averaged squared error of the predicted county-level infection prevalence for each day from May 3 to May 9. The prediction error is all at the rate of 10^{-5}, namely one case difference for the cumulative number of infections per 10,000 people in a county. For an example with an average county of 100,000 people (in fact 97,118 in our data), with the data up to May 2, the predicted total number of infections on May 3 is about 20 cases more or less than the actual observed number of confirmed infections, and in most of the cases our predicted numbers are larger than the reported. The prediction error increases due to the increased uncertainty over time. The prediction errors within the first 3 days are very close. This prediction error should be interpreted with caution, due to the underreporting issue related to the number of confirmed cases in a county. Since our prediction model takes the rate of antibody in both estimation and prediction of prevalence, the predicted number of infections includes both symptomatic

and asymptomatic cases in the prediction. In contrast, the observed data available in the public database only contains the number of confirmed symptomatic cases (with no asymptomatic cases). So, we expect that the error rate would be even smaller had the number of asymptomatic cases available in the test data. We added the above discussion on the top of page 13 (above Figure 6).

Date	May 3	May 4	May 5	May 6	May 7	May 8	May 9
SSPE$\times 10^{-5}$	2.02	2.12	2.40	2.90	3.73	4.88	6.42

- To ensure the convergence of the MCMC, we set the adaptation number to be 10^4, thinned the chain by keeping one draw from every 10 random draws to reduce autocorrelation, set a burn-in period of 2×10^5 draws to let the chain stabilize, and start from 4 separate chains of length 5×10^5. Thus, in total, we have had 2×10^5 effective draws in that about 2×10^6 draws were discarded. Moreover, we monitored trace plots of all the relevant model parameters to check the quality of the mixing for the MCMC draws and algorithm convergence. The draws after the burn-in were also exported to run additional checking using the CODA R package. Below, we provide an example of the trace plot and density estimate for the basic reproduction number R_0 for your perusal. In our view, the MCMC draws look good. We added a short description of the practice above in the revision (the paragraph below Table 1).



Dataviz Editor's comment: *The authors' response on their test confuses me. Why do they not want to carry out a 7-day test? They write "In our view, validation based on a single prediction value may not be very rigorous." I agree, but that implies they should do more testing. If you propose a*

model to make 7-day forecasts and include your predictions, readers would like to know that you have done all you can to test your model.

Our response: Thanks for these comments that have been raised by the editor, too. We copied our answer to the editor here for the sake of your convenience. We have included test results for the 1 to 7-day ahead prediction accuracy of the estimated infection prevalence at county-level. The following table gives the averaged squared error of the predicted county-level infection prevalence for each day from May 3 to May 9. The prediction error is all at the rate of 10^{-5}, namely one case difference for the cumulative number of infections per 10,000 people in a county. For an example with an average county of 100,000 people (in fact 97,118 in our data), with the data up to May 2, the predicted total number of infections on May 3 is about 20 cases more or less than the actual observed number of confirmed infections, and in most of the cases our predicted numbers are larger than the reported. The prediction error increases due to the increased uncertainty over time. The prediction errors within the first 3 days are very close. This prediction error should be interpreted with caution, due to the underreporting issue related to the number of confirmed cases in a county. Since our prediction model takes the rate of antibody in both estimation and prediction of prevalence, the predicted number of infections includes both symptomatic and asymptomatic cases in the prediction. In contrast, the observed data available in the public database only contains the number of confirmed symptomatic cases (with no asymptomatic cases). So, we expect that the error rate would be even smaller had the number of asymptomatic cases available in the test data.

Date	May 3	May 4	May 5	May 6	May 7	May 8	May 9
SSPE$\times 10^{-5}$	2.02	2.12	2.40	2.90	3.73	4.88	6.42

Fifth Revision – June 7, 2020

Editor's comment: *I think comment 1) is a very good one, and you need to add a bit of explanation of why you didn't use the one weighted by the county sizes. At this late stage I'd hate to ask you to do both, but maybe you can also provide a map of predictive errors themselves (on its original scale, instead of the squared ones) together with a map of the county size? This would show, for example, if there is a spike around zero, and also give readers a much better sense of where the typical errors are, since the average error might be a really bad one if the histogram of the errors are skewed toward one direction or another (to see this clearly, I would also include a histogram of the prediction errors).*

Our response: We did the following revisions.

1) Per the suggestion by the dataviz editor, in this round of revision we have adopted the weighted absolute prediction error (WAPE) to replace the previous (unweighted) average squared prediction error. The weight is a ratio of county population size over the population size of all counties. We agree with you that this WAPE, adjusted by county population size, is more appropriate in its magnitude and easier to interpret. See the highlighted changes on page 19.
2) To see how the change of the prediction error metric may affect the selection of tuning parameter η, below we included Table 5 (not included here but available in the original letter to the editor) that showed the performances of three different types of one-day ahead prediction errors, including the weighted absolute prediction error (WAPE), the (unweighted) average squared prediction error (ASPE), and the weighted squared prediction error (WSPE). Note that ASPE is the one used in the previous version, and the WAPE is the new one used in the current version, and WSPE is the case regarding what if the previous ASPE were weighted by county population size. Overall, the optimal tuning parameter selection can be carried out with each selection criterion with stable numerical performances. We did not include this table in the paper as it did not seem to be an essential piece of analysis in the empirical study and can avoid confusion. We would like to stay with WAPE for the ease of interpretation.

3) To illustrate the prediction accuracy, per your suggestion, we included Figure 8 that showed the nationwide 7-day ahead WAPEs (panel A) along with the county population size (Panel B).
4) In addition, per your suggestion, we added Figure 9 to illustrate the densities (similar to histograms) of WAPEs over the period of 7 days considered in the prediction, namely May 3-9.

Editor's comment: *And regarding the notion of "error", since you are checking against the confirmed cases, which we know are extremely problematic, you may want to emphasize that "small" errors here don't necessarily mean more accurate, though then a reader may ask why did you use minimal SSPE to start with. To answer such questions in a more nuanced way, you may want to take a look of the excellent discussion in this forthcoming article in the special issue, especially it's Section 2: https://www.dropbox.com/s/r15vmvbthhb3qmp/Angelopoulos-etal-final.pdf?dl=0.*

Our response: We have added this reference in our paper to facilitate the discussion on the prediction error. We agree with you about that "small" errors don't necessarily mean more accurate; rather the WAPE is only a relative metric, which should be interpreted with caution. This is mainly because of biases in surveillance data collection and infection case ascertainment. See more discussions in the closing sentences of the second paragraph on page 19.

Dataviz editor's comment: *If SSPE is the average squared prediction error, then what kind of average is it? There are over 3000 counties in the US with widely varying populations. Would the errors be weighted by population? Many counties might have a prediction of zero and a prediction error of zero.*

Our response: Thanks for your good suggestion. In this revision, we adopted the average absolute prediction error weighted by county population, called weighted absolute prediction error (WAPE) in the paper. The weighted average is based on all the counties within the 39 states that passed the MCMC convergence diagnosis; these states have experienced severer covid-19 pandemic so that their data are relatively abundant to fit the model well. The initial values of the other states are given by the national average estimates in the risk prediction. Per your suggestion, we included Figures 8 and 9 to illustrate the prediction accuracy. Figure 8 shows the nationwide 7-day ahead

WAPEs (panel A) along with the county population size (panel B). Suggested by the editor, Figure 9 illustrates the densities (similar to histograms) of WAPEs over the period of 7 days (May 3-9).