Peer Review File

Manuscript Title: Burden and characteristics of COVID-19 in the United States during 2020

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

This is an extremely interesting study that attempts to model the spread of COVID across the entire USA in 2020 using a discrete time spatial model fitted at the county level. Its scale and ambition are very suitable for publication in the journal Nature. Overall, I like the structure of the model, which has day and night periods and is informed by commuting data. However, there are some small areas that fall short of the required scientific quality that I hope that the authors can address before publication.

The main statistical methodology used by the paper is the Ensemble Adjustment Kalman Filter (EAKF). The authors note that this method has been used successfully to infer epidemiological parameters for a range of diseases, but fail to provide us with the limitations of the method. For instance, the state variables and parameter rates must all be positive and therefore cannot completely follow a Gaussian distribution. Does this have any implications on the analysis? Especially in periods when the infection rate is low and I(t) is close to zero in some locations.

Is uncertainty well calibrated in the analysis? Figure 1a shows that almost all of the observations fall well within the 95% confidence interval, which suggests not.

Some of the main epidemiological parameters (Z, D, mu and theta) are fixed in the analysis based on data through to 13th of March. Surely, this is only a tiny fraction of the epidemic data? Can we really not learn a great deal more information about these parameters by looking at the whole dataset rather than just the first 7 weeks? Or was this early data better quality somehow?
Theta is the rate of random population exchange and appears to be fixed at 0.15. What is the interpretation (and the units) of this parameter? This is a key parameter that governs the spatial correlation between locations and could be explored in more detail, if not inferred. What proportion of the travel between locations comes from the random component and what proportion from the commuting?

If the model is adding value to the data, it should be able to provide some forecasts into the future. Can you predict which areas will have the most cases in the first few months of 2021? How long is the time horizon in which the forecasts are sensible?

Where are the graphs of the alphas and betas? They appear to be temporally independent in the prior at each location. Does this limit the model's ability to provide precise forecasts as there is nothing on which to base the next beta except for the prior? Speaking of the prior, what is it? The SI does not name a distribution and I do not understand how the uncertainty is specified. The betas are the main infection rate parameters that drive the epidemic and so the usefulness of the model rests on the believability of their estimated values. Are they smooth in time or volatile? Are they ever estimated to be negative? What is the range of beta values?

Considering that this is a spatial analysis over the whole of the USA, the maps presented in main text are very small. What conclusions can be drawn about the spatial distribution of infections and

susceptibility? Figure 2b shows a high degree of spatial correlation which disappears in 2c. What does the model tell us about the distribution of cases that looking at the raw data (e.g. cumulative number of cases per county) does not? How does the susceptibility map compare to the commuting data map? Are areas with more incoming commuters at higher risk than those with outgoing commuters? Why show results for 1st of December in Figure 2c (which is not the end of the analysis), and then go on to discuss results up to 31st of December in the following paragraphs?

Overplotting in Figure 4a makes it of little use, and Figures 4b and 4c are in incomprehensible units. I'm not sure that dividing the data into locations with increasing or decreasing cases and then fitting a straight line to the Rt plots makes any sense. Since the number of locations in each group is frequently changing it could be sensitive to Simpson's paradox at the very least. Are there not simpler and more direct ways to identify that there is a negative correlation between the direction of the observed cases and the direction of Rt? Is this even a useful finding?

The definition of Rt needs much more discussion. I notice that Rt does not account for the susceptibility in the population, so if 50% of people are susceptible and Rt is 1.5 then each person is successfully infecting 0.75 people on average and the epidemic declines. Much greater discussion of why this definition of Rt was chosen could be included.

Validation on the antibody data: I don't understand why you adjust the data to match the model. The model contains information about how many of the cases are recent (prior to antibody waning), so why not compare the model-based estimates of the proportion of people who would have recent enough infections to be antibody positive with the actual observed data?

Referee #2 (Remarks to the Author):

In this paper, the authors use county level COVID19 case data from 2020 to build a transmission model of SARS-CoV-2 in the US. This model allows the authors to capture the underlying level of infection in the community. As they also have death data – they then use this to track underlying IFR and CFR. The detailed data available in the US makes this a very attractive dataset. In addition, the questions the authors attempt to address (underlying level of susceptibility in space and time, IFR/CFR trends, impact of interventions) – are clearly critical to the response. Having said that I am sceptical of arguably the most important and politically sensitive estimate, which is the IFR.

I was surprised with several features of the IFR estimates. Firstly, the overall mean IFR, while not presented, appears to be around 0.3%-0.4% (assuming a weighted average of when infections occur). This seems very low given what has been estimated for the US and other Western countries previously (0.6-1.1%) (Brazeau et al., Imperial report 34 2020, O'Driscoll Nature 2020, Meyerowitz-Katz G, Merone L IJID 2020). There is no attempt to reconcile these differences. Secondly, there are very significant long terms trends in IFR – going from 1% at the start, down to 0.3% by the end of 2020. Thirdly, there are large differences across locations in IFR despite very similar age distributions. Finally, there are substantial short-term fluctuations in IFR. None of these key patterns in IFR are explored in any depth. I believe the way it is currently presented could lead to confusing and potential harmful narratives emerging (e.g., you're up to five times more likely to die if you get infected in Miami than Chicago).

I think it is very difficult to believe that IFR has dropped by 70% since the start of the epidemic. The study cites earlier case detection and improved clinical care as drivers of this decrease. The citations they point to back up these hypotheses is the use of 'Awake Proning' and the use of Remdisivir. The impact of these treatments on mortality remains unclear. For example, the WHO's

Solidarity trial in 11k patients found no difference (NEJM 2020). It seems very unlikely that these treatments can explain such a large drop in IFR. Changing age distribution of cases and deaths have also been unable to explain changes in CFR elsewhere (Lefrancq et al., 2021).

The timing of the fluctuations in IFR within individual locations are also unexpected. For example, in Miami the lowest IFR is reached in July 2020 when it reaches a nadir of around 0.2% with highs of over 1% in the months immediately prior and post this date. July represents the peak of cases in the city, when most strain was being placed on hospitals. A similar pattern is inferred in Los Angeles and Phoenix. Other studies that have tracked CFR over time have consistently found the opposite effect – ie the probability of death (among hospitalised patients at least) goes up when cases go up and there is more strain on the hospital system (Braata et al., 2021, Taccone et al., 2021, Lefrancq et al., 2021, Wilde et al., 2021). It is difficult to think of credible hypotheses to explain a 5 times difference in IFR over such short time intervals and for the lowest levels in IFR to be reached when there were most strains on the healthcare system.

I believe that some of these issues in the IFR come from the way case ascertainment and deaths were considered. The authors use a simple fixed 9 day delay between confirmation and death based on a single study in New York City. This is a substantial oversimplification as the time between infection and death is highly dispersed - ranging from a few days to months after infection and also depends on age and whether the individuals were in ICU (Lefrancq et al., 2021). Epidemic curves of deaths are always a lot more smooth than epidemic curves of cases as a result. If the authors deconvoluted the timing of deaths using a distribution of the time from infection to death (e.g., Wu Nature Medicine 2020) that could remove some of these effects and could also help tease apart the likely substantial spatial heterogeneity in the time between case ascertainment and death.

The way case ascertainment is considered is a little confusing. The alpha (case ascertainment) in equation 9 only seems to differ by county but this may be a typo as it seems to also vary by month (Figure 1C).

The assays used by the CDC in their seroprevalence studies were a combination of the Roche and the Abbott assays – with different assays used in different locations. The Abbott assay appears to have a much more significant effect of waning than the Roche one (Peluso et al., medRxiv 2021).

I could not see a definition of mu – it seems to be a relative transmissibility of documented and undocumented infectivity with a fixed value of 0.64. It would be useful to set out the sources for the fixed input parameters.

Referee #3 (Remarks to the Author):

Pei et al. develop a spatiotemporal model for COVID-19 epidemic dynamics in the United States and fit this model to county-level reported case data. The model is further calibrated with human migration data (commuter travel between counties) and compared to state-level and metro-level seroprevalence estimates.

This is a nice analysis which in comparison to other studies incorporates greater detail and realism in spatial structure and human migration. The application of commuter traffic data is a real strength of the analysis although the practical importance of using this data isn't demonstrated. The authors present original model-based estimates of attack rates, case-ascertainment, IFRs and CFRs.
The paper is somewhat lacking in establishing the unique knowledge added by this modelling approach. Most of what the paper presents is already general knowledge, but original model-based estimates are useful. The paper could be written in a way to put the results in context of what is

already known and delineate what is original.

Figure 3: It is remarkable how complex & non-monotonic some of the IFR estimates are in certain locations. For me, this is one of the most interesting figures and deserves more discussion. The authors put forward an explanation that the monotonic national trend is driven by improving quality of care. That is undoubtedly the most important factor. But how would they explain what occurred in LA & Miami? The 2nd peak in LA could potentially be related to growth of a variant (452R) but that would not explain why the IFR subsequently decreased. The 2nd peak in Miami is very hard to explain. Is this related to the age profile of patients? Hospital occupancy? Socioeconomic status of counties impacted in different waves?

A model that incorporates detailed information about human mobility/migration could be useful for prediction. It would be helpful if the authors demonstrated the practical importance of incorporating human migration data. How does that compare to a model fitted to county-level data independently or in some sort of hierarchical framework, e.g. with a correlation structure in Rt between neighbouring counties? Unfortunately, it is not clear to me how incorporation of this migration data improved the main outputs of the analysis (ascertainment, IFRs…).

The authors have made most if not all of their code available on GitHub. The code appears lightly documented. It would be useful to have a brief document explaining how to run the code and reproduce results.

Minor Issues

L12: Citation needed. Was the US epidemic actually an outlier in terms of severity compared to the rest of the Americas?

L17: Reporting an overall figure for the ascertainment rate is not very useful since it has varied greatly over time. Reporting values for different periods would be useful.

L28-37: This level of background is not needed; the details are commonly known.

L115: What is meant by ‘up to 0.96’? This looks like the central estimate. CIs should specify type (credible intervals?) and quantiles.

L126-140 & Figure 4: I find this collection of results very confusing. It is not sufficiently clear what we are supposed to observe in Figure 4 or how the authors arrive at their interpretation. How the analysis is carried out is briefly described in the SI but is also not sufficiently clear (comments below). Would it not be better to explicitly incorporate the time of different NPIs or a NPI stringency index into the model? For example: https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker
This has been incorporated in models such as this one: https://doi.org/10.1098/rsos.201726

L160: Reference needed. Is this statement based on reported cases & deaths or seasonally-adjusted excess mortality?

Figure 2: Use unambiguous date formats

Model equations L357: It says that the model is a Markov process and that these equations ‘depict’ the model. More rigour would be appreciated at this point. I think that these equations define the expected value of the Markov process and that stochastic terms are introduced later. But L439 it says that some variables are updated deterministically. The model equations do not show that beta is time dependent. Then how does Rt vary through time? Surely not purely by changes in susceptibles. If this is a variable that is sampled and given dynamic variation by the kalman filter, this needs to be defined explicitly. Is that what the authors mean when they say the model is a

Markov process? It would be helpful to explicitly define deterministic & stochastic terms in this model.

The calculation of IFRs (L596) is a little simplistic. It would be better to integrate over the time-to-event distribution in the denominator. Alternatively, deaths & IFRs could be incorporated in the model and fitted to observed deaths directly. I worry that results are sensitive to the survival distribution from New York, and there is a long skew in the survival times. If the authors have access to linelisted data giving outcome of individual cases it would be far better to estimate IFRs with that. At a minimum, the authors can investigate sensitivity to Td & the width of the moving window (7 days).

L608: Define these variables. Again, it is unclear to me if Rt is something that is being estimated with the kalman filter or if this is a post-hoc analysis after the beta's have been estimated.

Referee #4 (Remarks to the Author):

In this article authors employ a metapopulation model to describe the transmission dynamics of COVID-19 at the county level in the USA. Authors then fitted the model time series of reported cases to quantify reproduction numbers, severity, prevalence and ascertainment rate. However, I have serious concerns about their approach to "fitting" the model to the case curves. Specifically:

1) The model is overparameterized. A huge number of parameters were estimated which explains the nice fits shown in the figures. For each data point, the transmission rate and ascertainment rate are estimated across counties. This means that hundreds of parameters were estimated for each county. So, what is the point of modeling then? Estimating a few parameters per county is needed to ensure that the parameters are identifiable from the data (a set of parameters can be uniquely estimated from a given model and data set). Lack of identifiability, or non-identifiability, occurs when multiple sets of parameter values yield a very similar model fit to the data, which is the biggest concern with your parameter estimation approach. It could be useful to evaluate parameter identifiability using simulated data prior to fitting the model to the real dataset.
2) Authors assume that documented cases are immobile and do not participate in human movement. This strong assumption seems odd. Could authors instead estimate a parameter to quantify the extent of the mobility for these individuals?
3) Some of the features reported (reproduction numbers, geographic transmission patterns or waves) have been reported elsewhere by the New York Times and others in formal (preprints, etc) or informal ways (designated websites). I am not sure about the novelty of some of the findings reported.

Author Rebuttals to Initial Comments:

Referee #1 (Remarks to the Author):

This is an extremely interesting study that attempts to model the spread of COVID across the entire USA in 2020 using a discrete time spatial model fitted at the county level. Its scale and ambition are very suitable for publication in the journal Nature. Overall, I like the structure of the model, which has day and night periods and is informed by commuting data. However, there are some small areas that fall short of the required scientific quality that I hope that the authors can address before publication.

We appreciate the referee's positive feedback. In the revision, we have performed extensive analyses to address the comments raised by the referee. Our responses to these comments are listed below.

The main statistical methodology used by the paper is the Ensemble Adjustment Kalman Filter (EAKF). The authors note that this method has been used successfully to infer epidemiological parameters for a range of diseases, but fail to provide us with the limitations of the method. For instance, the state variables and parameter rates must all be positive and therefore cannot completely follow a Gaussian distribution. Does this have any implications on the analysis? Especially in periods when the infection rate is low and I(t) is close to zero in some locations.

The EAKF is a deterministic filter that only adjusts the first two moments of the ensemble of simulations. As a consequence of this approach, the EAKF is biased for nonlinear systems, unlike, say, a particle filter; however, particle filters encounter functional difficulties working with higher dimension systems, whereas the EAKF is well-suited for such complexity (it is routinely applied to numerical weather models with 10^9 state variables). This efficiency is important for the metapopulation model used here with ~10^4 state variables and parameters. Further, in a comprehensive comparison of different filtering techniques for inference and forecasting of influenza, the EAKF algorithm demonstrated performance similar to other Bayesian inference algorithms specifically developed to handle nonlinear dynamics and non-Gaussian distributions[1]. To handle violations of physicality, e.g. adjustment of a positive definite variable or parameter to a negative number, we impose bounds on variables/parameters and resample out-of-bound values among ensemble members to ensure these variables/parameters remain non-negative. In the revised SI, we discuss these issues and the limitations of the EAKF in lines 152-169.

Is uncertainty well calibrated in the analysis? Figure 1a shows that almost all of the observations fall well within the 95% confidence interval, which suggests not.

To examine model calibration, we show the coverage of the 25%, 50%, 75% and 95% CIs for daily cases at county and national levels in a reliability plot (Fig. S4 in the revised SI). The coverage of CIs lies on the diagonal line, indicating the posterior estimate is well calibrated in the analysis. We report this additional analysis in lines 252-261 and in Fig. S4 of the revised SI.

Some of the main epidemiological parameters (Z, D, mu and theta) are fixed in the analysis based on data through to 13th of March. Surely, this is only a tiny fraction of the epidemic data? Can we really not learn a great deal more information about these parameters by looking at the whole dataset rather than just the first 7 weeks? Or was this early data better quality somehow?

We fixed *Z, D,* μ and θ to values estimated during the first wave of the pandemic in order to facilitate better estimation of β and α, which vary in both time and space. The rationale for using case data through March 13th is that non-pharmaceutical interventions (NPIs) had not widely been implemented in the US by that time. As a result, inter-county mobility and transmission rates in different locations have less spatial variation, which facilitates a better identification of those parameters. However, per the reviewer's suggestion, we explored whether our findings are sensitive to the period used for estimating *Z, D,* μ and θ. Specifically, we performed a separate inference of parameters using case data reported during the 6-week period from February 21 to April 2, 2020. The posterior distributions obtained from this ensemble are: $Z = 3.38$ (2.92 – 5.07), $D = 4.12$ (2.85 – 4.63), $\mu = 0.66$ (0.51 – 0.90), and $\theta = 0.18$ (0.06– 0.26). Model inference using this set of fixed parameters yielded similar results for model fitting, ascertainment rate and population susceptibility (Fig. S5 in the revised SI). We include this sensitivity analysis in lines 299-306 of the revised SI and report it on lines 47-49 of the revised manuscript.

Theta is the rate of random population exchange and appears to be fixed at 0.15. What is the interpretation (and the units) of this parameter? This is a key parameter that governs the spatial correlation between locations and could be explored in more detail, if not inferred. What proportion of the travel between locations comes from the random component and what proportion from the commuting?

In the transmission model, we assume the volume of random movement is proportional to the number of commuters between two counties. The parameter θ represents the ratio of random movement to commuting (e.g. $\theta = 0.15$ means the number of random visitors is 15% of the number of commuters). For $\theta = 0.15$, 13% of the travel between locations comes from the random component and the remaining 87% from the commuting. θ is estimated using case data from the early phase of the pandemic and fixed during the year. To test sensitivity to this parameter value, we repeated the inference using $\theta = 0.18$, which yielded similar

results (Fig. S5 in the revised SI). We clarify these issues and findings in the revised SI in lines 63-65.

If the model is adding value to the data, it should be able to provide some forecasts into the future. Can you predict which areas will have the most cases in the first few months of 2021? How long is the time horizon in which the forecasts are sensible?

We very much appreciate the reviewer's comment. We have used forms of this model since March 2020 to routinely generate projections of future cases, deaths, hospitalizations and ICU demand on a weekly to twice-weekly basis. All these projections are posted and archived (see https://github.com/shaman-lab/COVID-19Projection, https://cuepi.shinyapps.io/COVID-19/, https://reichlab.io/covid19-forecast-hub/, https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html). An analysis of the projections, their utility, and accuracy is beyond the scope of the present work, but a number of multi-group white papers and manuscripts are forthcoming, e.g.[2]. We want to note that our projections of future COVID activity are conditional probabilities, i.e. the projections depend on assumptions as to human behavior and policies in the coming weeks, which are themselves difficult to predict. As a consequence, we consider these efforts projections of possible future outcomes rather than forecasts. Among our scenarios, we have routinely generated an 'as is' projection scenario, which simply uses the most recent state variable and parameter estimates as the starting point for integrating forward 6 weeks into the future.

Where are the graphs of the alphas and betas? They appear to be temporally independent in the prior at each location. Does this limit the model's ability to provide precise forecasts as there is nothing on which to base the next beta except for the prior? Speaking of the prior, what is it? The SI does not name a distribution and I do not understand how the uncertainty is specified. The betas are the main infection rate parameters that drive the epidemic and so the usefulness of the model rests on the believability of their estimated values. Are they smooth in time or volatile? Are they ever estimated to be negative? What is the range of beta values?

The EAKF algorithm updates model parameters sequentially at each time step, using the posterior estimate from the last time step as the prior for the current time step. As a result, the estimated time-varying parameters for each location are autocorrelated in time (i.e. parameter priors are not temporally independent), which regularizes the inference and smooths the curve of estimated parameters over time. For system initialization, the initial prior distributions of parameters were set using the empirical distributions (obtained directly from the ensemble) estimated from case data prior to 13 March 2020. These empirical distributions, provided in the code, are represented by an ensemble with 100 members, which determines the uncertainty of these parameters. Key statistics are reported in the Supplementary Information on lines 175-176, and we provide details on the EAKF in the revised SI on lines 200-206. The estimated α and β for five metropolitan areas are shown in Fig. S3. The initial ranges of α and β were set as $\alpha \in [0.05, 1]$ and $\beta \in [0.2, 2.5]$ in the inference.

Considering that this is a spatial analysis over the whole of the USA, the maps presented in main text are very small. What conclusions can be drawn about the spatial distribution of infections and susceptibility? Figure 2b shows a high degree of spatial correlation which disappears in 2c. What does the model tell us about the distribution of cases that looking at the raw data (e.g. cumulative number of cases per county) does not? How does the susceptibility map compare to the commuting data map? Are areas with more incoming commuters at higher risk than those with outgoing commuters? Why show results for 1st of December in Figure 2c (which is not the end of the analysis), and then go on to discuss results up to 31st of December in the following paragraphs?

We have enlarged the maps in Fig. 2 and now present susceptibility for 31 December, rather than 1 December. The enlarged maps are presented in Fig. S13. The high degree of spatial correlation in Fig. 2b is an artifact of the fact that the majority of counties had low attack rates by August 2020. Fig. 2c shows the more considerable spatial variation of attack rates by year-end.

The model inference adds information to the raw reported case data by estimating the true attack rates (both reported and unreported infections) in different locations as the raw data underestimate the true burden of infection due to underreporting rates that vary over space and time. For instance, the New York and Los Angeles metro areas reported 677K and 952K cases in 2020. However, due to the lower ascertainment rate in

New York during the early pandemic, we estimate that these two areas experienced a similar number of total infections (5.2M and 5.5M), despite reported cases in LA being 40% higher.

We analyzed the correlation between daily outgoing and incoming commuters (per capita, averaged over 2020) in US counties and the fraction of population infected by the end of the year. On a logarithmic scale, the number of incoming commuters is better correlated with the infected fraction of the population (Fig. S14 in the revised SI), though the explanatory power of the commuting data is limited. This analysis is now reported on lines 420-425 in the revised SI and lines 92-93 of the revised manuscript.

Overplotting in Figure 4a makes it of little use, and Figures 4b and 4c are in incomprehensible units. I'm not sure that dividing the data into locations with increasing or decreasing cases and then fitting a straight line to the Rt plots makes any sense. Since the number of locations in each group is frequently changing it could be sensitive to Simpson's paradox at the very least. Are there not simpler and more direct ways to identify that there is a negative correlation between the direction of the observed cases and the direction of Rt? Is this even a useful finding?

Per the reviewer comment, we have performed a new, different analysis to demonstrate how changes in Rt, reflecting community response to growing or declining case numbers, changed over the course of the pandemic. For this analysis, we identified counties that experienced both increasing and decreasing reported cases during all three waves in 2020 and examined the distributions of weekly R_t change in each county (Fig. 4 in the revised manuscript). In total, we included 1052 counties across five US regions (71, 438, 371, 88, and 84 counties in each of 5 regions, respectively). As the set of counties analyzed in each region is consistent across all three waves, the analysis should not be sensitive to Simpson's paradox for each region. We used paired Wilcoxon signed-rank test to determine if the weekly R_t change significantly increased over successive pandemic waves. The results again indicate that community adjustments of Rt in the presence of growing case numbers slowed with each successive pandemic wave. These changes are evident by region and demonstrate statistically significant differences in community control response over time. While it is difficult to infer causality, these associations suggest a normalization of the pandemic and an abandonment or fatigue for implementing NPI measures.

The definition of Rt needs much more discussion. I notice that Rt does not account for the susceptibility in the population, so if 50% of people are susceptible and Rt is 1.5 then each person is successfully infecting 0.75 people on average and the epidemic declines. Much greater discussion of why this definition of Rt was chosen could be included.

We chose the time-varying basic reproduction number, Rt, rather than the effective reproduction number, Re, in order to segregate the effects of population susceptibility from the structural (e.g. population density) and behavioral (e.g. NPIs) determinants of the time-varying reproduction number. While Re is the truest measure of whether cases will grow in a locality, in evaluating community response to growing or declining case numbers, we wanted to focus on behavioral changes due to policy changes and population compliance with NPIs. Structural effects are assumed not to vary appreciably. By utilizing Rt rather than Re, we remove the effects of changing population susceptibility within a wave and fully focus on the changes to transmissibility due to behavior. We now present a fuller explanation of this choice in lines 137-139 in the revised manuscript and lines 554-562 in the revised SI.

Validation on the antibody data: I don't understand why you adjust the data to match the model. The model contains information about how many of the cases are recent (prior to antibody waning), so why not compare the model-based estimates of the proportion of people who would have recent enough infections to be antibody positive with the actual observed data?

This is a good suggestion. We now convolve antibody waning on the model posterior estimates of infection and compare the model-generated seroprevalence with the observed data. The match between model estimated antibody prevalence and seroassay measures is similarly high (as measured by Pearson correlation and MAE). Details are provided on lines 59-60 of the revised manuscript and lines 410-416 of the revised SI.

Referee #2 (Remarks to the Author):

In this paper, the authors use county level COVID19 case data from 2020 to build a transmission model of SARS-CoV-2 in the US. This model allows the authors to capture the underlying level of infection in the community. As they also have death data – they then use this to track underlying IFR and CFR. The detailed data available in the US makes this a very attractive dataset. In addition, the questions the authors attempt to address (underlying level of susceptibility in space and time, IFR/CFR trends, impact of interventions) – are clearly critical to the response. Having said that I am sceptical of arguably the most important and politically sensitive estimate, which is the IFR.

I was surprised with several features of the IFR estimates. Firstly, the overall mean IFR, while not presented, appears to be around 0.3%-0.4% (assuming a weighted average of when infections occur). This seems very low given what has been estimated for the US and other Western countries previously (0.6-1.1%) (Brazeau et al., Imperial report 34 2020, O'Driscoll Nature 2020, Meyerowitz-Katz G, Merone L IJID 2020). There is no attempt to reconcile these differences. Secondly, there are very significant long terms trends in IFR – going from 1% at the start, down to 0.3% by the end of 2020. Thirdly, there are large differences across locations in IFR despite very similar age distributions. Finally, there are substantial short-term fluctuations in IFR. None of these key patterns in IFR are explored in any depth. I believe the way it is currently presented could lead to confusing and potential harmful narratives emerging (e.g., you're up to five times more likely to die if you get infected in Miami than Chicago).

We thank the reviewer for these comments. The previous estimates of IFR by Brazeau et al., Imperial report 34 2020, O'Driscoll Nature 2020, Meyerowitz-Katz G, Merone L IJID 2020 were all made for activity recorded during the first 6 months of 2020. These estimates, which range from 0.6-1.1% are consistent with the estimates we present in this work—a cumulative IFR of 0.5-1.1% during June 2020 (Fig. S20 in the revised SI). We have now added discussion of these earlier works, the consistency of our own estimates with these previous estimates, and the level of spatial heterogeneity we find for IFR in the US for this time period. See details in lines 515-545 in the revised SI.

We believe the time trend of IFR is correct. Firstly, the CFR, now pulled entirely from observations using line-list data released by the US CDC[3], shows a similar time trend (Fig. S17 in the revised SI). Secondly, three additional factors should contribute to a decline of IFR over time: 1) the average age of reported cases decreased indicating a lower proportion of elderly among cases (Fig. S18 in the revised SI); IFR has a pronounced dependency on age and increases by several orders of magnitude moving from children <10 to the elderly (see Yang et al. 2021, LID[4]); 2) testing availability improved over the year such that infected individuals were likely diagnosed earlier during their course of infection, which may have improved chances of survival; and 3) clinical treatment did improve both in terms of general patient management and the availability of drugs such as monoclonal antibody therapies. While some of the drugs, such as Remdisivir, may have nominal effects, in aggregate, clinical management and care of COVID-19 patients has improved. For instance, the use of dexamethasone lowered 28-day mortality among patients receiving respiratory support (rate ratio 0.64 among patients receiving invasive mechanical ventilation, 0.82 among patients receiving only oxygen)[5]. Further, Brazeau et al. selected data up to August 17, 2020 to estimate IFR and explicitly stated that "The August 17, 2020 cutoff was selected as one-month after the release of the RECOVERY trial results under the assumption that changes in COVID-19 clinical practice would alter the IFR compared to earlier in the pandemic"[6]. In another study, a reduced mortality rate in hospital from the spring to the summer was reported in the New York City[7] - adjusted mortality among hospitalizations dropped from 25.6% in March to 7.6% in August, after adjusting for patient characteristics. These facts together suggest that the IFR in the US likely decreased substantially during the 2020. We have added further discussion of these issues to the revised manuscript and SI (lines 450-462, 526-540).

Regarding the spatial variation of IFR, despite similar population age distributions, the age structure of infection and hospitalization varied across locations (Fig. S18 in the revised SI). In addition, the mortality rate among confirmed cases and hospitalized patients also exhibits considerable variation. For instance, the hospitalization mortality rate in New York is much higher than other locations during the early pandemic (Fig. S17b).

We agree that the short-term fluctuations of IFR required further interrogation. IFR appears to surge when cases begin to infect the elderly again (Fig. S18b in the revised SI), which doesn't happen in every community

and appears to partially explain the spatial differences in IFR among metropolitan areas. However, in addition, we now have access to line-list data[3], which were not available at the time of initial submission. This line-list dataset includes date related to illness or specimen collection (or the date received by CDC), county of residence, demographic information and health outcomes (hospitalization, ICU admission and death) for the majority of cases reported from the beginning of the pandemic through March 31, 2021. Using this dataset, we were able to estimate the distribution of time from case confirmation to death. We then deconvolved the death time series using these estimated delay distributions (lines 464-497, Fig. S19 in the revised SI) and re-calculated the IFR using this deconvolved time series of deaths. New results show fewer short-term fluctuations (Fig. 3c in the revised manuscript), which generally match the age distribution of infections and hospitalizations. For instance, the dip of IFR in Phoenix in August is contemporary with increased hospitalization of young patients in the 18-49 age-group (Fig. S18b). We now provide a detailed discussion of these issues in the revised SI.

I think it is very difficult to believe that IFR has dropped by 70% since the start of the epidemic. The study cites earlier case detection and improved clinical care as drivers of this decrease. The citations they point to back up these hypotheses is the use of 'Awake Proning' and the use of Remdisivir. The impact of these treatments on mortality remains unclear. For example, the WHO's Solidarity trial in 11k patients found no difference (NEJM 2020). It seems very unlikely that these treatments can explain such a large drop in IFR. Changing age distribution of cases and deaths have also been unable to explain changes in CFR elsewhere (Lefrancq et al., 2021).

As discussed above, the CFR, which is pulled entirely from observations using line-list data shows a similar time trend. While this is partly explained by an increasing ascertainment rate (i.e. increasing the numbers of confirmed cases per actual infection and thus the denominator of CFR), the drop in CFR during 2020 (Fig. S17a) far exceeds the rise in ascertainment rate. This indicates improved survival for cases. Such improved survival implies a decrease of IFR, though less extreme, as the rise of the ascertainment rate does not affect IFR.

Also, as noted above, improved patient care and treatment has been reported. Although repurposed antiviral drugs (e.g. remdesivir, hydroxychloroquine, lopinavir, and interferon regimens) have had little or no effect on hospitalized patients with COVID-19, the use of dexamethasone has been shown to substantially reduce the 28-day mortality rate among patients receiving invasive mechanical ventilation (rate ratio 0.64 compared with the usual care group) and oxygen (rate ratio 0.82 compared with the usual care group)[5]. After the publication of this study in July 2020, the mortality rate among hospitalized patients with severe symptoms appears to have decreased. Indeed, nationally, the crude hospitalization fatality rate (HFR) obtained using the line-list data declined from July to September (Fig. S17b), even though the proportion of hospitalizations from the 65+ age-group increased (Fig. S18b). Other studies have shown that in New York City, the adjusted mortality among hospitalizations dropped from 25.6% in March to 7.6% in August, after adjusting for patient characteristics[7]. The estimated IFR, as well as the crude HFR increased during the fall/winter wave, when there was more strain on the healthcare system (especially in Los Angeles). However, although hospitalization during the fall/winter wave was much higher, the IFR, crude CFR and crude HFR (Fig. S17) still remained below the levels that occurred during the spring wave nationally. Discussion of these issues has been added to the revised SI (lines 450-462, 526-540).

The timing of the fluctuations in IFR within individual locations are also unexpected. For example, in Miami the lowest IFR is reached in July 2020 when it reaches a nadir of around 0.2% with highs of over 1% in the months immediately prior and post this date. July represents the peak of cases in the city, when most strain was being placed on hospitals. A similar pattern is inferred in Los Angeles and Phoenix. Other studies that have tracked CFR over time have consistently found the opposite effect – ie the probability of death (among hospitalised patients at least) goes up when cases go up and there is more strain on the hospital system (Braata et al., 2021, Taccone et al., 2021, Lefrancq et al., 2021, Wilde et al., 2021). It is difficult to think of credible hypotheses to explain a 5 times difference in IFR over such short time intervals and for the lowest levels in IFR to be reached when there were most strains on the healthcare system.

We thank the reviewer for pointing us to these relevant studies. We now use a more sophisticated method to compute IFR. Specifically, we estimated the distribution of the delay from case confirmation to death using line-list data recently released by the US CDC (lines 464-497, Fig. S19 in the revised SI). We re-calculated the IFR using deconvolved death time series adjusted for these estimated delay distributions (Fig. 3c in the revised

manuscript). The new results show fewer short-term fluctuations (Fig. 3c in the revised manuscript). Remaining fluctuations generally match the age distribution of infections and hospitalizations. For example, the increased hospitalization of young patients in the 18-49 age-group (Fig. S18b) could explain the decrease of IFR in Phoenix in August.

In the new results, we also observe increased IFR and CFR during the fall/winter wave, when there was more strain on the healthcare system (especially in Los Angeles) (Fig. 3c). The line-list data indicate that the HFR increased from October to December (Fig. S17b), in agreement with the studies pointed out by the reviewer. We include discussion of this issue in the revised manuscript (lines 116-117).

I believe that some of these issues in the IFR come from the way case ascertainment and deaths were considered. The authors use a simple fixed 9 day delay between confirmation and death based on a single study in New York City. This is a substantial oversimplification as the time between infection and death is highly dispersed - ranging from a few days to months after infection and also depends on age and whether the individuals were in ICU (Lefrancq et al., 2021). Epidemic curves of deaths are always a lot more smooth than epidemic curves of cases as a result. If the authors deconvoluted the timing of deaths using a distribution of the time from infection to death (e.g., Wu Nature Medicine 2020) that could remove some of these effects and could also help tease apart the likely substantial spatial heterogeneity in the time between case ascertainment and death.

We appreciate the reviewer's constructive suggestions. We agree that the use of a fixed time-to-death was an oversimplification. We now have access to line-list data, which were not available at the time of initial submission. To estimate the IFR, we need to account for the lag from case confirmation to death, which varied over space and time. Unfortunately, the line-list data do not include time of events, and other publicly available data are not complete enough to support an analysis for the entire US during 2020. To address this difficulty, we developed a grid search method to estimate time-to-event distributions from case confirmation to death in each county (see details in lines 464-487 in the revised SI). We assume the delay follows a gamma distribution with a long tail. To reflect improved healthcare later in the year, for each county, we fitted two delay distributions during two periods (February 21 to June 30 and July 1 to December 31, results are robust to the choice of separation date)—see Fig. S19 for examples for Maricopa County AZ (Phoenix) and Miami-Dade County FL (Miami). In most counties, the delay in the later period is longer than the delay early in the pandemic. For instance, the estimated mean delay for New York County, NY is 8 days before July 1, which agrees with a previous estimate of around 9 days[4], and 36 days after July 1. Using the estimated delay distributions, we performed a Richardson-Lucy-Type deconvolution to recover the number of deaths among patients confirmed on each day (see details in lines 489-497 in the revised SI). This new analysis did remove some of the short-term fluctuations and tease apart the spatial heterogeneity in the time between case ascertainment and death.

The way case ascertainment is considered is a little confusing. The alpha (case ascertainment) in equation 9 only seems to differ by county but this may be a typo as it seems to also vary by month (Figure 1C).

The reviewer is correct, alpha varies by county and through time. We have corrected this notation.

The assays used by the CDC in their seroprevalence studies were a combination of the Roche and the Abbott assays – with different assays used in different locations. The Abbott assay appears to have a much more significant effect of waning than the Roche one (Peluso et al., medRxiv 2021).

Indeed, assays used in the CDC seroprevalence studies may perform differently in terms of sensitivity and specificity. To test whether this has an effect on the fitting to seroprevalence data, we repeated the analysis using a lower sensitivity (85.1%, sensitivity of the Abbott assay during 61-80 days after infection[8]) and specificity (99.2%) and found the results to be robust. Please find details in Fig. S12 and lines 410-416 of the revised SI.

I could not see a definition of mu – it seems to be a relative transmissibility of documented and undocumented infectivity with a fixed value of 0.64. It would be useful to set out the sources for the fixed input parameters.

Yes. The parameter μ is the relative transmissibility of undocumented infections compared with documented infections. Studies from several countries indicate that asymptomatic cases of COVID-19, which are typically unreported, are less contagious than symptomatic cases[9–12]. This parameter is estimated using case data through 13 March 2020, as in our previous study[13]. For a sensitivity analysis, we also performed another estimate using case data reported during the 6-week period from February 21 to April 2, 2020. The estimate of $\mu = 0.66$ is robust and inference using this new parameter remains similar. We now provide a fuller description of the model parameters, including those that are fixed, in the revised SI (lines 57-65).

Referee #3 (Remarks to the Author):

Pei et al. develop a spatiotemporal model for COVID-19 epidemic dynamics in the United States and fit this model to county-level reported case data. The model is further calibrated with human migration data (commuter travel between counties) and compared to state-level and metro-level seroprevalence estimates.

This is a nice analysis which in comparison to other studies incorporates greater detail and realism in spatial structure and human migration. The application of commuter traffic data is a real strength of the analysis although the practical importance of using this data isn't demonstrated. The authors present original model-based estimates of attack rates, case-ascertainment, IFRs and CFRs.
The paper is somewhat lacking in establishing the unique knowledge added by this modelling approach. Most of what the paper presents is already general knowledge, but original model-based estimates are useful. The paper could be written in a way to put the results in context of what is already known and delineate what is original.

We thank the reviewer for these comments. We believe the findings provide detailed spatial-temporal estimates of the ascertainment rate, IFR, and cumulative attack rates during 2020, which deliver benchmark insights into the progression of the virus in the US during the year. The results corroborate more limited assessments of IFR from early in the pandemic, match serosurvey estimates of attack rates, show trends and spatial heterogeneity in case reporting rates and population susceptibility, and indicate how community response to increasing case numbers—estimated by reductions of Rt—diminished during the year, consistent with a growing normalization of the pandemic and fatigue with control measures. In doing so, the model-inference system provides a comprehensive estimate of conditions throughout the US at county scale.

The commuter traffic data are vital for conducting inference. In previous work on the pandemic in China[14], we showed how a more limited number of locations compromises the identifiability of the model-inference system. By including all 3142 counties in the US—and linking them with observed travel tendencies—the model is better able to depict the introduction and possible re-introductions that instigate community transmission (though once community transmission is established and persists, these linkages matter less). In the revision, we performed a synthetic analysis to explore the ability of the inference system to track the temporal variation of transmission rates and ascertainment rates in major metropolitan areas under different scenarios. The inference system can generally capture the time variation of these parameters in different metropolitan areas (Fig. S1 in the revised SI), even when the change of parameters is initiated at distinct times with varying magnitudes. This synthetic test indicates that the inclusion of mobility data enables a better identification of key epidemiological parameters. The unique capability of this model-inference system allows us to explore fine-grained characteristics of COVID-19 spread in the US, which were only reported previously in a single location or aggregated at the national level.

Figure 3: It is remarkable how complex & non-monotonic some of the IFR estimates are in certain locations. For me, this is one of the most interesting figures and deserves more discussion. The authors put forward an explanation that the monotonic national trend is driven by improving quality of care. That is undoubtedly the most important factor. But how would they explain what occurred in LA & Miami? The 2nd peak in LA could potentially be related to growth of a variant (452R) but that would not explain why the IFR subsequently decreased. The 2nd peak in Miami is very hard to explain. Is this related to the age profile of patients? Hospital occupancy? Socioeconomic status of counties impacted in different waves?

The fluctuations in the IFR estimates were indeed an interesting feature. Per comments from reviewer 2, we have re-visited these estimates by changing our handling of the time delay from case reporting to death. Rather than using a fixed time delay, we instead have leveraged recently released line-list data to estimate

time-to-event distributions of this process (see details in lines 464-487 in the revised SI). We now use these distribution estimates to model deaths from case numbers. In doing so, the amplitude of the oscillations has lessened, though this feature still persists (Fig. 3c).

The remaining fluctuations of IFR may be partly explained by changes to the age distribution of cases and hospitalizations (Figs. S17-S18). For instance, the dip of IFR in Phoenix in August could be attributed to a decrease in the age profile of infections, as suggested by the increase in hospitalizations of young patients in the 18-49 age-group (Fig. S18b). We also now point to evidence explaining the drop of IFR during the year and find that the IFR increased during the fall/winter wave when healthcare systems were strained. These details are now added to the revised SI (lines 526-545).

A model that incorporates detailed information about human mobility/migration could be useful for prediction. It would be helpful if the authors demonstrated the practical importance of incorporating human migration data. How does that compare to a model fitted to county-level data independently or in some sort of hierarchical framework, e.g. with a correlation structure in Rt between neighbouring counties? Unfortunately, it is not clear to me how incorporation of this migration data improved the main outputs of the analysis (ascertainment, IFRs…).

We have used a similar model with human mobility to predict the spatial transmission of influenza and other respiratory viruses (Pei et al. PNAS 2018[15]; Pei et al. Nat Comm 2021[16]). Compared with a local model, the metapopulation model can better predict the spatial movement of infectious diseases. Another benefit of including human mobility is the improvement of parameter identifiability. In a previous study on the pandemic in China[14], we showed that coupling disease transmission in different locations in a metapopulation model leads to more accurate identification of epidemiological parameters. Here, we performed inference using model-simulated outbreaks and found that the inference system can generally capture the time variation of parameters in major metropolitan areas (see details in lines 226-250 in the revised SI).

We also have used forms of this metapopulation model since March 2020 to routinely generate projections of future cases, deaths, hospitalizations and ICU demand on a weekly to twice-weekly basis. All these projections are posted publicly and archived (see https://github.com/shaman-lab/COVID-19Projection, https://cuepi.shinyapps.io/COVID-19/, https://reichlab.io/covid19-forecast-hub/, https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html). An analysis of the projections, their utility, and accuracy is beyond the scope of the present work, but a number of multi-group white papers and manuscripts are forthcoming. We want to note that our projections of future COVID activity are conditional probabilities, i.e. the projections depend on assumptions as to human behavior and policies in the coming weeks, which are themselves difficult to predict. As a consequence, we consider these efforts projections of possible future outcomes rather than forecasts. Among our scenarios, we have routinely generated an 'as is' projection scenario, which simply uses the most recent state variable and parameter estimates as the starting point for integrating conditions 6 weeks in the future.

The authors have made most if not all of their code available on GitHub. The code appears lightly documented. It would be useful to have a brief document explaining how to run the code and reproduce results.

We have now updated the GitHub site to include all the necessary code and have created a ReadMe file detailing how to run the code and reproduce the results.

Minor Issues

L12: Citation needed. Was the US epidemic actually an outlier in terms of severity compared to the rest of the Americas?

The US had most confirmed COVID-19 cases and deaths of any country in the world during 2020, though not the most per capita (WHO). We have amended the text to "…was particularly devastating for the United States, which experienced the highest numbers of reported cases and deaths during 2020."

L17: Reporting an overall figure for the ascertainment rate is not very useful since it has varied greatly over time. Reporting values for different periods would be useful.

We have changed the abstract to now describe the progression of the ascertainment rate over time. Specifically, we write, "The pandemic in the US during 2020 was characterized by national ascertainment rates that increased from 11.3% during March (95% credible interval (CI):8.3 – 15.9%) to 24.5% (18.6 – 32.3%) during December."

L28-37: This level of background is not needed; the details are commonly known.

We have deleted this paragraph.

L115: What is meant by 'up to 0.96'? This looks like the central estimate. CIs should specify type (credible intervals?) and quantiles.

We have removed "up to" and specified the type of CI, i.e. credible intervals, at its first appearance in lines 18 and 67 in the revised manuscript.

L126-140 & Figure 4: I find this collection of results very confusing. It is not sufficiently clear what we are supposed to observe in Figure 4 or how the authors arrive at their interpretation. How the analysis is carried out is briefly described in the SI but is also not sufficiently clear (comments below). Would it not be better to explicitly incorporate the time of different NPIs or a NPI stringency index into the model? For example:https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker This has been incorporated in models such as this one: https://doi.org/10.1098/rsos.201726

We appreciate the reviewer's comment. In response to this and comments from Reviewer 1, we performed a new, different analysis to demonstrate how changes in Rt, reflecting community response to growing or declining case numbers, changed over the course of the pandemic. We identified counties that experienced both increasing and decreasing reported cases during all three waves in 2020 in five US regions, and examined the distributions of weekly R_t change in each county (Fig. 4 in the revised manuscript). An increased weekly Rt change indicates a more muted response to case growth. In total, we included 1052 counties across five regions (71, 438, 371, 88, and 84 counties in the Northeast, Southeast, Midwest, Southwest, and West Regions, respectively). We used paired Wilcoxon signed-rank test to determine if the change of weekly R_t significantly increased over successive pandemic waves. The results again indicate that community adjustments of Rt in the presence of growing case numbers slowed with each successive pandemic wave. These changes are evident by region and demonstrate statistically significant differences in community control response over time. While it is difficult to infer causality, these associations suggest a normalization of the pandemic and an abandonment or fatigue for implementing NPI measures.

There are many studies analyzing the effects of various NPIs on SARS-CoV-2 transmission and COVID-19 outcomes. It is beyond the scope of this paper to address this important and very challenging analysis. Rather, we use our estimates of daily Rt by county to quantify whether reductions of virus transmissibility—produced by both policies and compliance with mask wearing and social distancing—improved or slackened during 2020.

L160: Reference needed. Is this statement based on reported cases & deaths or seasonally-adjusted excess mortality?

The US reported the most COVID-19 cases and deaths in the world during 2020. We have clarified this statement in the revised manuscript.

Figure 2: Use unambiguous date formats

We have changed the date formatting.

Model equations L357: It says that the model is a Markov process and that these equations 'depict' the model. More rigour would be appreciated at this point. I think that these equations define the expected value of the Markov process and that stochastic terms are introduced later. But L439 it says that some variables are updated deterministically. The model equations do not show that beta is time dependent. Then how does Rt

vary through time? Surely not purely by changes in susceptibles. If this is a variable that is sampled and given dynamic variation by the kalman filter, this needs to be defined explicitly. Is that what the authors mean when they say the model is a Markov process? It would be helpful to explicitly define deterministic & stochastic terms in this model.

We have updated our presentation of the model equations, their description, and model integration. Equations (1-10) define the expected value of the Markov process; we generate stochastic dynamics using Poisson distributions. Specifically, we use random sampling from Poisson distributions to quantify the numbers of new infections, new recoveries, etc. The deterministic and stochastic terms in the model are defined explicitly in the revised SI. The model is integrated stochastically; however, the EAKF update of the ensemble of simulations is deterministic[17], so that the higher moments of the distributions are preserved during ensemble update.

Both β and α are time dependent; both variables contribute to variations of Rt through $R_t = \beta(t)D[\alpha(t) + \mu(1-\alpha(t))])$, which is derived using the next-generation matrix approach[18].

The calculation of IFRs (L596) is a little simplistic. It would be better to integrate over the time-to-event distribution in the denominator. Alternatively, deaths & IFRs could be incorporated in the model and fitted to observed deaths directly. I worry that results are sensitive to the survival distribution from New York, and there is a long skew in the survival times. If the authors have access to linelisted data giving outcome of individual cases it would be far better to estimate IFRs with that. At a minimum, the authors can investigate sensitivity to Td & the width of the moving window (7 days).

Per the reviewer comment and comments from reviewer 2, we have re-visited these estimates by changing our handling of the time from case reporting to death. As the reviewer suggests, rather than using a fixed time delay, we instead have leveraged recently released line-list data to establish empirical time-to-event distributions of this process. We now use these distribution estimates to model deaths from case numbers. Specifically, we developed a method to estimate the distribution of the delay from case confirmation to death in each county (see details in lines 464-487 in the revised SI). We assume the delay follows a gamma distribution with a long tail. To reflect improving healthcare in the latter half of 2020, for each county, we fitted separate delay distributions for two time periods (February 21 to June 30 and July 1 to December 31, results are robust to the choice of separation date)—see Fig. S19 for examples for Maricopa County AZ (Phoenix) and Miami-Dade County FL (Miami). In most counties, the delay in the second period is longer than that in the first. For instance, the estimated mean delay for New York County NY is 8 days before July 1, which agrees with a previous estimate of around 9 days, and 36 days after July 1. Using the estimated delay distributions, we performed a Richardson-Lucy-Type deconvolution to recover the number of deaths among patients confirmed on each day (see details in lines 489-497 in the revised SI).

L608: Define these variables. Again, it is unclear to me if Rt is something that is being estimated with the kalman filter or if this is a post-hoc analysis after the beta's have been estimated.

R_t is computed using estimated model parameters via $R_t = \beta(t)D[\alpha(t) + \mu(1-\alpha(t))])$. Here $\beta(t)$ is the local transmission rate on day t, D is the mean infectious period, $\alpha(t)$ is the ascertainment rate on day t, and μ is the relative transmissibility of undocumented infections. The equation for R_t was derived using the next-generation matrix approach (see details in the SI of Li et al. Science 2020[14]). For each county, R_t can be computed using estimated model parameters. We clarify the definition of R_t in the revised SI (lines 549-553).

Referee #4 (Remarks to the Author):

In this article authors employ a metapopulation model to describe the transmission dynamics of COVID-19 at the county level in the USA. Authors then fitted the model time series of reported cases to quantify reproduction numbers, severity, prevalence and ascertainment rate. However, I have serious concerns about their approach to "fitting" the model to the case curves. Specifically:

1) The model is overparameterized. A huge number of parameters were estimated which explains the nice fits shown in the figures. For each data point, the transmission rate and ascertainment rate are estimated across counties. This means that hundreds of parameters were estimated for each county. So, what is the point of

modeling then? Estimating a few parameters per county is needed to ensure that the parameters are identifiable from the data (a set of parameters can be uniquely estimated from a given model and data set). Lack of identifiability, or non-identifiability, occurs when multiple sets of parameter values yield a very similar model fit to the data, which is the biggest concern with your parameter estimation approach. It could be useful to evaluate parameter identifiability using simulated data prior to fitting the model to the real dataset.

We agree with the reviewer that we are attempting to estimate many parameters with this model-inference system. In our first COVID-19 work (Li et al., Science 2020[14]), we build a similar meta-population model for 375 cities in China. We used this model in conjunction with daily case observations and data assimilation to estimate just 6 parameters over a 14-day period. We invested considerable effort showing that this system, which included undocumented infections and separate levels of contagiousness for documented and undocumented infections, was identifiable and additionally investigated the circumstances (fewer cities in the network) in which identifiability would be compromised.

However, upon transferring this model to the 3142 counties of the US and recognizing that NPIs and virus transmission vary in both time and space, we realized a more realistic model structure would allow both the ascertainment and the basic contact rate to vary through time and by county. This, of course, greatly increases the model parameters and brings issues of identifiability to the fore. There are several issues that instill confidence in our findings: 1) we fix the shared parameters, Z, D, μ and θ, at values estimated from the early stages of the pandemic in the US; the results are insensitive to the duration of the pandemic used to find the estimates for these four parameters (see response to reviewer 1 above) and match estimates we obtained with the model-inference system and a single β and α for all counties; 2) as demonstrated in our previous work on COVID-19 spread in China[14], coupling transmission dynamics across different locations in the metapopulation model can greatly improve the parameter identifiability; 3) the sequential update of the EAKF algorithm imposes temporal autocorrelation on parameter estimates, which regularizes the inference and smooths the curve of estimated parameters over time; 4) we focus on estimates for metropolitan areas for which COVID-19 case levels are quite high; and 5) we out-of-sample validate the model using seroprevalence measures, thus comparing our estimates of cumulative attack rates with surveys of seroconversion adjusted for waning antibody titers.

Still, the reviewer's point is valid; we therefore performed some synthetic tests of identifiability. For these we generated outbreaks of COVID-19 cases using the metapopulation model in free simulation with prescribed time-varying patterns for β_i and α_i. We included scenarios in which these time-varying patterns differed by county or metropolitan area. These synthetic outbreaks of cases were then used with the full model-inference system to determine whether the parameter estimates nationally and for key metropolitan areas could be inferred. We found that the inference system generally captures the time variation of parameters in different metropolitan areas (Fig. S1), even when the change of parameters is initiated at distinct times with varying magnitudes. This synthetic test indicates that the model-inference system can reasonably recover key time-varying parameters for major metropolitan areas. In the revised SI, we provide a more detailed discussion on the subject of identifiability on lines 208-250.

2) Authors assume that documented cases are immobile and do not participate in human movement. This strong assumption seems odd. Could authors instead estimate a parameter to quantify the extent of the mobility for these individuals?

The reviewer makes an interesting point. Rather than try to estimate the mobility of documented infections, which may be difficult to identify and hard to validate, we instead performed two sensitivity analyses in which 25% and 50% of documented infections were allowed to participate in human movement on any given day. Estimates of critical epidemiological features (susceptibility, ascertainment rate, IFR) were insensitive to these increases of mobility (see details in Figs S6-S7 in the revised SI). This insensitivity appears to stem from the reduced intercounty movement observed during much of 2020, the limited fraction documented contagious persons (relative to total population) comprise within a county at any given time, and the fact that community transmission was robust in many counties throughout the year limiting the impact of further (re)introduction of the virus. These issues appear to minimize the effect of assuming documented cases are immobile.

3) Some of the features reported (reproduction numbers, geographic transmission patterns or waves) have

been reported elsewhere by the New York Times and others in formal (preprints, etc) or informal ways (designated websites). I am not sure about the novelty of some of the findings reported.

We believe the results are novel and that the methods used to produce these results are reliable. Estimates of the IFR in the US during the early part of the pandemic are available, but not produced for the entire year, which essentially represents the period prior to any impacts of vaccination, nor distributed by geography to allow assessment of differences by state or major metropolitan areas. Further, estimates of the ascertainment rate and population susceptibility are quite limited. Our group has produced a number of preprints, postings and projections that we have not felt necessary to follow-up with peer-reviewed journal publication; however, in this instance, there are a range of novel findings—spatial and temporal varying estimates of susceptibility, IFR, and the ascertainment rate during 2020—verified by posterior fitting to case observations and out-of-sample validation with seroprevalence survey, such that we felt it important to report these results formally.

References

1. Yang, W., Karspeck, A. & Shaman, J. Comparison of Filtering Methods for the Modeling and Retrospective Forecasting of Influenza Epidemics. *PLOS Comput. Biol.* **10**, e1003583 (2014).

2. Zebrowski, A. *et al.* A Spatiotemporal Tool to Project Hospital Critical Care Capacity and Mortality From COVID-19 in US Counties. *Am. J. Public Health* e1–e10 (2021) doi:10.2105/AJPH.2021.306220.

3. U.S. Centers for Disease Control and Prevention. COVID-19 Case Surveillance Public Use Data with Geography. https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4 (2021).

4. Yang, W. *et al.* Estimating the infection-fatality risk of SARS-CoV-2 in New York City during the spring 2020 pandemic wave: a model-based analysis. *Lancet Infect. Dis.* (2020) doi:10.1016/S1473-3099(20)30769-6.

5. Dexamethasone in Hospitalized Patients with Covid-19. *N. Engl. J. Med.* **384**, 693–704 (2021).

6. Brazeau, N. *et al. Report 34: COVID-19 infection fatality ratio: estimates from seroprevalence*. http://spiral.imperial.ac.uk/handle/10044/1/83545 (2020) doi:10.25561/83545.

7. Horwitz, L. I. *et al.* Trends in COVID-19 Risk-Adjusted Mortality Rates. *J. Hosp. Med.* **16**, 90–92 (2021).

8. Muecksch, F. *et al.* Longitudinal analysis of clinical serology assay performance and neutralising antibody levels in COVID19 convalescents. *medRxiv* 2020.08.05.20169128 (2020) doi:10.1101/2020.08.05.20169128.

9. Sayampanathan, A. A. *et al.* Infectivity of asymptomatic versus symptomatic COVID-19. *The Lancet* **397**, 93–94 (2021).

10. Bi, Q. *et al.* Household Transmission of SARS-CoV-2: Insights from a Population-based Serological Survey. *medRxiv* 2020.11.04.20225573 (2021) doi:10.1101/2020.11.04.20225573.

11. Gao, M. *et al.* A study on infectivity of asymptomatic SARS-CoV-2 carriers. *Respir. Med.* **169**, 106026 (2020).

12. Byambasuren, O. *et al.* Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: Systematic review and meta-analysis. *Off. J. Assoc. Med. Microbiol. Infect. Dis. Can.* **5**, 223–234 (2020).

13. Pei, S. & Shaman, J. Initial Simulation of SARS-CoV2 Spread and Intervention Effects in the Continental US. *medRxiv* 2020.03.21.20040303 (2020) doi:10.1101/2020.03.21.20040303.

14. Li, R. *et al.* Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* **368**, 489–493 (2020).

15. Pei, S., Kandula, S., Yang, W. & Shaman, J. Forecasting the spatial transmission of influenza in the United States. *Proc. Natl. Acad. Sci.* **115**, 2752–2757 (2018).

16. Pei, S., Teng, X., Lewis, P. & Shaman, J. Optimizing respiratory virus surveillance networks using uncertainty propagation. *Nat. Commun.* **12**, 222 (2021).

17. Anderson, J. L. An Ensemble Adjustment Kalman Filter for Data Assimilation. *Mon. Weather Rev.* **129**, 2884–2903 (2001).

18. Diekmann, O., Heesterbeek, J. a. P. & Roberts, M. G. The construction of next-generation matrices for compartmental epidemic models. *J. R. Soc. Interface* **7**, 873–885 (2010).

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

This manuscript is much improved from the previous submission, but there are still two issues remaining.

1. I still don't see that Figure 4 supports the conclusion drawn in the paper, that there was "a lessening ability or willingness to control SARS-CoV-2 in the US as the year progressed". To me Figure 4 might say that R_t was found to increase over time showing an inability of the model to correctly capture the decline in susceptibility in the population. There are several other equally plausible explanations too (temporal confounding etc.). The authors have been too keen to jump to an epidemiological conclusion without feeling the need to mention the limitations or alternative explanations that I would expect to see in high quality science. I don't see how the heavy subsetting in use in this analysis makes the conclusion any clearer, in fact the opposite applies.

The associated paragraph in the text is even worse: "During the spring wave, Rt decreased (negative k) in most counties during periods of case growth, particularly in the Northeast US". Looking at the figure, I see all of the medians are almost exactly zero, except for the Northeast US. This description appears to misrepresent the figure.

2. Figure 1b looks nice at face value, but the more I think about it then the less useful it appears.

We want to see evidence that the spatially explicit model can predict seropositivity levels *over and above* the general country trend in seropositivity over time. The top 5 points look great, but the rest of the figure suffers from overplotting. Really, I want to see something like the difference between the observed and predicted sero-positivity levels on the y-axis (ie the residual) and time on the x-axis. I think there is actually a much stronger validation of the model than is evident from the figure, which is not very clear.

Referee #2 (Remarks to the Author):

I found the manuscript much improved. In particular the use of the deconvolution for the death data is a major improvement.

I remain a little puzzled by the mu parameter, which represents relative transmissibility of detected vs undetected infections. This is a fixed term based on cases early on in the epidemic. However, the authors allow for case ascertainment to increase over the epidemic (it seems to roughly double). As parametrized, I believe this is saying that while a greater proportion of infections were being detected, the relative infectivity of those detected stayed the same. This seems odd and I imagine could have a marked impact on the dynamics of the epidemic in the early part of 2020. The authors may want to consider allowing mu to change in time to compensate for increased case ascertainment.

It is worth noting that many COVID-19 deaths may have been missed from official figures, this would increase the estimated IFR, eg., Weinberger et al., 2020 estimated a third were missed, more recent estimates suggest that the number missed may be even greater (http://www.healthdata.org/special-analysis/estimation-excess-mortality-due-covid-19-and-scalars-reported-covid-19-deaths).

In the context of reinfection risk, especially with different variants, the authors may want to consider relabelling '% susceptible' in their plots to something like '% uninfected'.

Referee #4 (Remarks to the Author):

It is fundamentally wrong to estimate more parameters than the number of data points available to infer model parameters. Authors have confirmed that they are estimating a total of 4 + n*m parameters where n is the number spatial units (states) and m is the number of days of the epidemic. Thus, for 52 states and 365 days, authors are estimating a total of 4+18980 = 18984 parameters while the dataset is comprised of a total of n*m=18980 data points. It would not be difficult to show that it would be impossible to uniquely distinguish whether the transmission rate or the ascertainment rate is responsible for ascending or descending phases in the trajectory of epidemics. In other words, an ascending or descending phase in the epidemic's trajectory could result from an increasing/decreasing transmission rate or an increasing/decreasing ascertainment rate, indicating lack of parameter identifiability has implications on the interpretation of the results. The main issue here is methodological in that the statistical framework set up for parameter inference is not well posed. Again, this explains why the model appears to "fit" the data so well.

Author Rebuttals to First Revision:

Reviewer Response

We thank the reviewers for their comments, and the editor for the opportunity to revise the manuscript. Per the request to keep the manuscript within journal standard length limit, we have moved the revised Figure 4 to Supplemental Information and now only present 3 figures—presenting the inference, the central findings of the paper. The summary paragraph and body text now total less than 2500 words, and the title is less than 75 characters. All other presentation requirements are also met.

Ref 1:

This manuscript is much improved from the previous submission, but there are still two issues remaining.

1. I still don't see that Figure 4 supports the conclusion drawn in the paper, that there was "a lessening ability or willingness to control SARS-CoV-2 in the US as the year progressed". To me Figure 4 might say that R_t was found to increase over time showing an inability of the model to correctly capture the decline in susceptibility in the population. There are several other equally plausible explanations too (temporal confounding etc.). The authors have been too keen to jump to an epidemiological conclusion without feeling the need to mention the limitations or alternative explanations that I would expect to see in high quality science. I don't see how the heavy subsetting in use in this analysis makes the conclusion any clearer, in fact the opposite applies.

The associated paragraph in the text is even worse: "During the spring wave, Rt decreased (negative k) in most counties during periods of case growth, particularly in the Northeast US". Looking at the figure, I see all of the medians are almost exactly zero, except for the Northeast US. This description appears to misrepresent the figure.

Response: We agree with the reviewer that the result presented in Figure 4 of the previous version of the manuscript is tentative. Unlike the central findings of the paper—the inference of the burden of disease and epidemiological characteristics (e.g. the IFR, the ascertainment rate, etc.)—this analysis uses the inference output to examine whether community transmission rates changed in response to growing or declining cases and if those transmission rate tendencies changed over successive waves. The results are at best suggestive, so we have moved a modified version of the analysis and findings to Supplementary Information and shortened its discussion in the main text. We have also included an improved discussion of the limitations of the analysis, along with other plausible explanations for our findings.

In the analysis, we examine R_t i.e. the time-varying reproduction number, which is exclusive of susceptibility, not the effective reproductive number, which is $R_e = R_t * S$. Susceptibility, S, has been reasonably well constrained by our model-inference approach, as demonstrated by the out-of-sample validation with seroprevalence data (Figures 1B, S12, S13 and S14). Note that in most studies using statistical methods to estimate reproduction numbers, the

effective reproductive number is the quantity estimated (sometimes also denoted by "R_t", which is different from our definition), as population susceptibility cannot be disentangled. Here, the dynamical model enables the estimation of susceptibility so we can remove the effect of declining susceptibility and focus on the variations of the time-varying reproduction number due to shifting levels of non-pharmaceutical interventions and seasonal changes to virus transmissibility.

In the revised manuscript, we analyze the association of the weekly change of R_t with the change of weekly cases per 100,000 people at the county level. The analysis is performed for five US regions (Northeast, Southeast, Midwest, Southwest, West) during the spring (Feb 21 – May 31), summer (Jun 1 – Sep 15), and fall/winter (Sep 16 – Dec 31) waves. In these five regions, 116, 162, 126, 45 and 54 counties that reported cumulative cases greater than 100 per 100,000 people during all three waves and had a population over 100,000 were included in the analysis. The inclusion criterion was used to focus on counties that experienced high community transmission. Results are robust to the choice of threshold value. A positive/negative change of weekly cases indicates increased/decreased community prevalence of COVID-19. We then fitted the weekly R_t change to the change of weekly cases per 100,000 people using a linear function and quantified R_t changes in response to local COVID-19 infection levels using the slope of the linear fit. Across all regions and pandemic waves, the slopes are consistently negative (except in Southwest during the spring wave when there was limited transmission, two-sided t test), suggesting that the community actively responded to the growth of local infections (Fig. S23). Within each region, in general, the absolute magnitude of the slope decreased over successive waves. While not definitive, the results are suggestive that communities responded to growing cases by tightening controls—again note that with the removal of susceptibility, our R_t merely reflects the potential in a fully susceptible population and is not slowing due to depletion of population susceptibility. Further, the magnitude of that tightening lessened, generally, over successive waves.

In the revised manuscript (main text), we discuss this analysis, though the details and figure are now relegated to the supplement.

*2. Figure 1b looks nice at face value, but the more I think about it then the less useful it appears. We want to see evidence that the spatially explicit model can predict seropositivity levels *over and above* the general country trend in seropositivity over time. The top 5 points look great, but the rest of the figure suffers from overplotting. Really, I want to see something like the difference between the observed and predicted sero-positivity levels on the y-axis (ie the residual) and time on the x-axis. I think there is actually a much stronger validation of the model than is evident from the figure, which is not very clear.*

Response: We have provided the requested plot as an inset to Figure 1b.

Ref 2:

I found the manuscript much improved. In particular the use of the deconvolution for the death data is a major improvement.

I remain a little puzzled by the mu parameter, which represents relative transmissibility of detected vs undetected infections. This is a fixed term based on cases early on in the epidemic. However, the authors allow for case ascertainment to increase over the epidemic (it seems to roughly double). As parametrized, I believe this is saying that while a greater proportion of infections were being detected, the relative infectivity of those detected stayed the same. This seems odd and I imagine could have a marked impact on the dynamics of the epidemic in the early part of 2020. The authors may want to consider allowing mu to change in time to compensate for increased case ascertainment.

Response: We now have performed a sensitivity analysis in which we repeated the inference allowing mu to vary over time. The findings remain similar: 1) the estimate of mu through time is similar to the fixed value inferred in the main analysis using data from early spring 2020; and 2) the estimates of the other parameters and conditions are similar. These findings are presented in the revised Supplementary Information (Figures S8 and S9).

It is worth noting that many COVID-19 deaths may have been missed from official figures, this would increase the estimated IFR, eg., Weinberger et al., 2020 estimated a third were missed, more recent estimates suggest that the number missed may be even greater (http://www.healthdata.org/special-analysis/estimation-excess-mortality-due-covid-19-and-scalars-reported-covid-19-deaths).

Response: We have added mention of this potential bias to the revised manuscript, i.e. that the IFR estimates are a lower bound.

In the context of reinfection risk, especially with different variants, the authors may want to consider relabelling '% susceptible' in their plots to something like '% uninfected'.

Response: We thank the reviewer for this suggestion; however, on reflection, we prefer to and have kept the legend as % susceptible, which we believe is clearer.

Ref 4:

*It is fundamentally wrong to estimate more parameters than the number of data points available to infer model parameters. Authors have confirmed that they are estimating a total of 4 + n*m parameters where n is the number spatial units (states) and m is the number of days of the epidemic. Thus, for 52 states and 365 days, authors are estimating a total of 4+18980 = 18984 parameters*

*while the dataset is comprised of a total of n*m=18980 data points. It would not be difficult to show that it would be impossible to uniquely distinguish whether the transmission rate or the ascertainment rate is responsible for ascending or descending phases in the trajectory of epidemics. In other words, an ascending or descending phase in the epidemic's trajectory could result from an increasing/decreasing transmission rate or an increasing/decreasing ascertainment rate, indicating lack of parameter identifiability has implications on the interpretation of the results. The main issue here is methodological in that the statistical framework set up for parameter inference is not well posed. Again, this explains why the model appears to "fit" the data so well.*

Response: The EAKF, the inference method employed in this study, is routinely used in conjunction with numerical weather models possessing upwards of 10^9 state variables, but many fewer observations. Dynamical linkage of the system state in both space and time—i.e. spatial and temporal autocorrelation—enables accurate estimation of these systems with relatively sparse observations. The massive state space of these numerical weather models is typically estimated every 6 hours, yet despite sparse observations these models can still generate accurate weather predictions days in the future.

The metapopulation model we developed for this study, like numerical weather models, is also dynamically linked in space and time—a linkage that facilitates the estimation of system parameters. Within this model, four global parameters (i.e. shared by all counties) are fixed based on inference conducted early during the pandemic, μ, θ, D, Z. For the remaining parameters and state space estimation, the metapopulation model is then run at county scale (3142 counties) from March 1, 2020 – December 31, 2020 (306 days). For each county, we estimate 2 parameters, $\alpha_{i,t}, \beta_{i,t}$, and 4 state variables $S_{i,t}, E_{i,t}, {I^r}_{i,t}, {I^u}_{i,t}$. This amounts to 6*3142*306 posterior estimates using 3142*306 observations of county-day cases. The EAKF updates both the state variables and the parameters with each observation assimilation.

While there are 6 times as many estimates as observations, the estimates are not independent. Both the parameters and the state variables are highly autocorrelated. For the state variables, this manifests from the dynamics of the system. For the parameter estimates, the temporal autocorrelation naturally arises from the use of the EAKF, a sequential Monte Carlo method, in which today's posterior estimate is tomorrow's prior. Indeed, for such methods, Bayes' rule provides a target for the update of the state given an observation: $p(Z_t|y_t, y_{t-1}, \dots y_1) \propto p(y_t|Z_t)p(Z_t|y_{t-1}, \dots y_1)$. Here, Z_t is the system state at time *t* (all parameters and state variables) and y_t are the observations on day *t.* The first term on the right-hand side, $p(y_t|Z_t)$, is the likelihood of observing the data given the system state, and the second term is the prior distribution of the system state, here the posterior from the previous day. The updated distribution, $p(Z_t|y_t, y_{t-1}, \dots y_1)$, is the new posterior and depends on all previously assimilated observations, i.e. $y_t, y_{t-1}, \dots y_1$.

Physically, we expect the parameters $\alpha_{i,t}, \beta_{i,t}$ to be highly temporally autocorrelated at county population scales, and not to vary erratically and randomly (though some events, such as lockdowns, may create jumps in contact rates). In this sense, the parameters, while not dynamically evolved by the model equations, are treated as state variables. Estimation is possible because of the dynamical linkage between the parameters and state variables.

In addition, the model parameters are also subject to strong spatial constraints imposed by human mobility across county boundaries. In using the EAKF, the parameters in one county are updated using case data from both the county itself as well as other counties connected in the mobility network (each county population mixes with a mean of 18 other counties). As a result, on average, the parameter estimates in a single county are informed by case data from 19 counties. Such spatial constraint also reduces the effect of county-specific biases and compensation between the transmission rate and reporting rate, e.g. if the reporting rate in one county is underestimated, the imported infections to other counties would be overestimated but corrected during data assimilation.

While we cannot precisely estimate the effective number of independent parameters being estimated without circularly relying on our model-inference estimates, we do posit that temporal autocorrelation alone effectively reduces the number of parameter estimates by a factor 30 or more (i.e. the decorrelation time scale for $\alpha_{i,t}, \beta_{i,t}$ is greater than one month). This would indicate we have at least 5 times as many observations as effective, independent parameters to estimate.

We have shown in synthetic tests (Figures S1 of the manuscript) that the inference system can identify time-varying parameters reasonably well for major metropolitan areas. We can further affirm that the distribution of residuals at the county level in these synthetic tests is small, providing additional evidence of sound parameter inference. Further, as discussed in our reviewer responses previously, in this work we focus on the results from large metropolitan areas for which COVID-19 case levels are quite high, and we expect the estimates for such areas to be more robust. Lastly, the out-of-sample validation of the model using seroprevalence measures shows that our estimates of cumulative attack rates correspond well with surveys of seroconversion adjusted for waning antibody titers, further verifying the model inference.

Reviewer Reports on the Second Revision:

Referees' comments:

Referee #2 (Remarks to the Author):

I previously reviewed this paper – my previous concerns with the paper have been addressed. In this review I have now also focused on the concerns of reviewer 1, who had two remaining issues with the paper.

There was a concern that the discussion regarding Figure 4 was overly speculative. This particular analysis explored whether subsequent changes in local transmission rates changed in responses to the dynamics of the local epidemic at the time. This is clearly a critical question relating to individual and community behaviours. In response to this concern, the authors have moved this figure to the supplement and lessened the speculative nature of the discussion around this figure. I found the author response adequate to this concern.

The reviewer also requested an additional analysis that explores the predicted versus observed serology results over time. The results from this additional analysis are compelling and the authors have included the results as an inset in the main figure.