science.sciencemag.org/cgi/content/full/science.abb3221/DC1

Supplementary Materials for

Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2)

Ruiyun Li*, Sen Pei*†, Bin Chen*, Yimeng Song,
Tao Zhang, Wan Yang, Jeffrey Shaman†

*These authors contributed equally to this work.
†Corresponding author. Email: sp3449@cumc.columbia.edu (S.P.); jls106@cumc.columbia.edu (J.S.)

Published 16 March 2020 on *Science* First Release
DOI: 10.1126/science.abb3221

This PDF file includes:

Other Supplementary Material for this manuscript includes the following:
(available at science.sciencemag.org/cgi/content/full/science.abb3221/DC1)

MDAR Reproducibility Checklist (.pdf)
Data S1 (.zip)

Materials and Methods

1. Model Configuration and Initialization

The transmission model incorporates information on human movement within the following metapopulation structure:

$$\frac{dS_i}{dt} = -\frac{\beta S_i I_i^r}{N_i} - \frac{\mu\beta S_i I_i^u}{N_i} + \theta\sum_j \frac{M_{ij}S_j}{N_j - I_j^r} - \theta\sum_j \frac{M_{ji}S_i}{N_i - I_i^r} \quad [1]$$

$$\frac{dE_i}{dt} = \frac{\beta S_i I_i^r}{N_i} + \frac{\mu\beta S_i I_i^u}{N_i} - \frac{E_i}{Z} + \theta\sum_j \frac{M_{ij}E_j}{N_j - I_j^r} - \theta\sum_j \frac{M_{ji}E_i}{N_i - I_i^r} \quad [2]$$

$$\frac{dI_i^r}{dt} = \alpha\frac{E_i}{Z} - \frac{I_i^r}{D} \quad [3]$$

$$\frac{dI_i^u}{dt} = (1-\alpha)\frac{E_i}{Z} - \frac{I_i^u}{D} + \theta\sum_j \frac{M_{ij}I_j^u}{N_j - I_j^r} - \theta\sum_j \frac{M_{ji}I_i^u}{N_i - I_i^r} \quad [4]$$

$$N_i = N_i + \theta\sum_j M_{ij} - \theta\sum_j M_{ji} \quad [5]$$

where S_i, E_i, I_i^r, I_i^u and N_i are the susceptible, exposed, documented infected, undocumented infected and total population in city *i*. Note that we define patients with symptoms severe enough to be confirmed as documented infected individuals; whereas other infected persons are defined as undocumented infected individuals. We provide a rate parameter, β, for the transmission rate due to documented infected individuals. The transmission rate due to undocumented individuals is reduced by a factor μ. In addition, α is the fraction of documented infections, *Z* is the average latency period and *D* is the average duration of infection. The effective reproduction number (R_e) is calculated as $R_e = \alpha\beta D + (1-\alpha)\mu\beta D$ (see Section 6 below for details). Spatial coupling within the model is represented by the daily number of people traveling from city *j* to city *i* (M_{ij}) and a multiplicative factor, θ, which is greater than 1 to reflect underreporting of human movement. We assume that individuals in the I_i^r group do not move between cities, though these individuals can move between cities during the latency period. A similar metapopulation model has been used to forecast the spatial transmission of influenza in the United States (*20*).

The core model structure (Equations 1-5) was integrated stochastically using a 4th order Runge-Kutta (RK4) scheme. Specifically, for each step of the RK4 scheme, each unique term on the righthand side (rhs) of Equations 1-4 was determined using a random sample from a Poisson distribution, i.e.

$$\mathrm{U1} = \mathrm{Pois}\left(\frac{\beta S_i I_i^r}{N_i}\right)$$

$$\mathrm{U2} = \mathrm{Pois}\left(\frac{\mu\beta S_i I_i^u}{N_i}\right)$$

$$\mathrm{U3} = \mathrm{Pois}\left(\theta\sum_j \frac{M_{ij}S_j}{N_j - I_j^r}\right)$$

$$\mathrm{U4} = \mathrm{Pois}\left(\theta \sum_j \frac{M_{ji} S_i}{N_i - I_i^r}\right)$$

$$\mathrm{U5} = \mathrm{Pois}\left(\alpha \frac{E_i}{Z}\right)$$

$$\mathrm{U6} = \mathrm{Pois}\left((1-\alpha) \frac{E_i}{Z}\right)$$

$$\mathrm{U7} = \mathrm{Pois}\left(\theta \sum_j \frac{M_{ij} E_j}{N_j - I_j^r}\right)$$

$$\mathrm{U8} = \mathrm{Pois}\left(\theta \sum_j \frac{M_{ji} E_i}{N_i - I_i^r}\right)$$

$$\mathrm{U9} = \mathrm{Pois}\left(\frac{I_i^r}{D}\right)$$

$$\mathrm{U10} = \mathrm{Pois}\left(\frac{I_i^u}{D}\right)$$

$$\mathrm{U11} = \mathrm{Pois}\left(\theta \sum_j \frac{M_{ij} I_j^u}{N_j - I_j^r}\right)$$

$$\mathrm{U12} = \mathrm{Pois}\left(\theta \sum_j \frac{M_{ji} I_i^u}{N_i - I_i^r}\right)$$

The solutions for Equations 1-4 were then calculated as:

$$\frac{dS_i}{dt} = -\mathrm{U1} - \mathrm{U2} + \mathrm{U3} - \mathrm{U4}$$

$$\frac{dE_i}{dt} = \mathrm{U1} + \mathrm{U2} - \mathrm{U5} - \mathrm{U6} + \mathrm{U7} - \mathrm{U8}$$

$$\frac{dI_i^r}{dt} = \mathrm{U5} - \mathrm{U9}$$

$$\frac{dI_i^u}{dt} = \mathrm{U6} - \mathrm{U10} + \mathrm{U11} - \mathrm{U12}$$

Equation 5 was solved deterministically.

The initial prior ranges of the parameters for the model were drawn using Latin hypercube sampling from uniform distributions with the following prior ranges:

- β: the transmission rate of symptomatically infected patients. $0.8 \leq \beta \leq 1.5$
- μ: the multiplicative factor reducing the transmission rate of unreported infected patients. $0.2 \leq \mu \leq 1$.

- θ: the multiplicative factor to adjust mobility data estimates of human movement between cities. $1 \leq \theta \leq 1.75$.
- Z: the mean latency period. $2\ days \leq Z \leq 5\ days$.
- α: the fraction of infections that develop severe symptoms. $0.02 \leq \alpha \leq 1.0$.
- D: the average duration of infection for infected patients. $2\ days \leq D \leq 5\ days$.

The initial prior ranges for α and μ were chosen to cover most possible values, i.e. [0,1]; the initial prior ranges for the latency and infection periods, *Z* and *D,* were provided in order to produce an aggregate period range of 4-10 days spanning previous estimates for the total infection period and generation time (*6, 14, 15*), as well as previously published estimates for other coronaviruses (summarized in Table 1 of *13*); the initial prior range for β was set to enable a broad initial range for R_e (i.e. [0.35, 7.5], based on $R_e = \alpha\beta D + (1-\alpha)\mu\beta D$); and the initial prior range for θ was set to capture the discrepancy between recorded Tencent travel volume and reported travel (*1*). Note that the Ensemble Adjustment Kalman Filter (EAKF, described in Section 5 below) is not constrained by the initial priors and can migrate outside these ranges to obtain system solutions.

For the outbreak origin, Wuhan city, the initial exposed population, E_{wuhan}, and initial undocumented infected population, I^u_{wuhan}, were drawn from a uniform distribution $[0, Seed_{max}]$. A single emergence of SARS-CoV2 has been estimated to have occurred in late 2019. Specifically, the most recent common ancestor (TMRCA) is estimated for 17 November 2019; [95% CI: 27 August – 19 December] (*21*). Assuming Wuhan as the epicenter with emergence on November 17 and an estimated doubling time of 6.4 days [95% CI: 5.8-7.1 days] (13) implies about 400 cumulative infections (most still latent or still infectious) [95% CI: 180-715] on January 10, the start date of our transmission model simulations. If we use a November 1 emergence—still well within the TMRCA 95% CI—these numbers rise considerably to about 2000 cumulative infections [95% CI: 1000-4000]. We used this last upper estimate to set $Seed_{max}$ (2000 initial latent (E_0) and undocumented infections (I_0^u) on January 10). Note $Seed_{max}$ only sets an upper bound: each model simulation randomly selects a number from 0 to $Seed_{max}$ (i.e. $0 \leq E_0 \leq Seed_{max}$; $0 \leq I_0^u \leq Seed_{max}$). Also, although infections were documented prior to January 10, these cases were sporadic and the EAKF adjustment can account for the effects of these early infections (by selecting elevated exposed and unreported infection levels). For other cities, we defined C_i as the number of travelers from Wuhan to city i on the first day of Chunyun. The initial exposed, documented infected and undocumented infected populations were set to $E_i = 3C_iE_{wuhan}/N_{wuhan}$, $I_i^r = 0$ and $I_i^u = 3C_iI^u_{wuhan}/N_{wuhan}$.

2. Observations of Confirmed COVID-19 Cases

Daily numbers of confirmed cases for 375 Chinese cities (Fig. S1) were collected from official reports on the website of National and Provincial Health Commissions in China (*1*). Confirmed cases were defined as a suspected case with a positive test result for viral nucleic acid (*22*). These data were compiled for periods before (January 10-23, 2020) and after (January 24 – February 8, 2020) the January 23 implementation of travel restrictions. These data are included in Other Supplementary Materials (Data S1) and posted at (*23*).

Suspected cases were diagnosed based on clinical symptoms and exposure, where exposure was indicated if an individual had resident history in Wuhan, travel history to Wuhan, or contact with

individuals from Wuhan who had experienced fever and respiratory symptoms. Additionally, the city of a confirmed case was the city where an individual was confirmed rather than home residence. Thus, individuals might acquire infection in one location, but become symptomatic and be confirmed in another. These definitions were consistent for the entire study period (January 10 – February 8, 2020).

3. Reporting Delay

Our transmission model (Eqs. 1-5) does not explicitly represent the process of infection confirmation. Thus, we mapped simulated documented infections to confirmed cases using a separate observational delay model. In this delay model, we account for the time interval between a person transitioning from latent to contagious (i.e. $E \rightarrow I_i^r$) and observational confirmation of that individual infection. To estimate this delay period, T_d, we examined line-list data from early confirmed cases in China (*17*). Prior to January 23, 2020, the time-to-event distribution of the interval (in days) from symptom onset to confirmation is well-fit by a Gamma distribution ($a = 1.85, b = 3.57, LL = -252.24$) [the Gamma distribution provides a better fit than a Weibull distribution ($A = 7.29, B = 1.41, LL = -255.17$)] (Fig. S2). Consequently, we adopted a Gamma distribution to model T_d. After January 23, this distribution is also well-fit by a Gamma distribution ($a = 2.34, b = 2.59, LL = -1251.94$) [Weibull distribution ($A = 6.78, B = 1.63, LL = -1255.38$)].

In practice during transmission model simulation, for each new documented infection transitioning from *E* to I_i^r, a random number t_d was drawn from the Gamma distribution $G(a, T_d/a)$. This new case was 'reported' as a confirmed infection t_d days after the transition from $E \rightarrow I_i^r$. The reported cases on a given day were then accumulated as the model integrated forward in time. Because infected patients may shed SARS-CoV2 before the onset of symptoms (*18, 24*), we considered longer mean times for the reporting delay than those estimated directly from the line-list data. [Note that the transition from *E* to I_i^r represents the onset of contagious shedding, not symptom onset.] Specifically, we tested a number of Gamma distributions in which we fixed the shape parameter, $a = 1.85$, but varied the distribution mean, $T_d = ab$, by increasing *b*. The best-fitting model, i.e. the best combination of posterior fit parameters (see Section 5 below), initial maximal seeding (see Section 1 above) and reporting delay Gamma distribution parameters, was identified by log likelihood.

4. Mobility Data

To capture individual movement among the 375 cities simulated in the metapopulation model, we use human mobility data from the Tencent location-based service (LBS) used in popular Tencent mobile phone applications (Apps), such as Wechat, QQ, and Maps. These data were collected using the application program interface (API) from the Tencent big data platform (*25*). No identifiable information is included in the dataset (Data S1, also posted on *23*).

The Tencent data platform stopped releasing human mobility information between paired cities after 2018. (Beginning 2019 only outflows and inflows to and from Beijing have been released.) Other data sources, such as the Mobility Index from Baidu (*16*), have continued to produce measures of mobility, but these do not provide counts of travel that are comparable among cities. Consequently, we used the 2018 Tencent data and assumed the travel patterns captured in 2018

during Chunyun are similar to those of the analogous time period during 2020, prior to January 23 travel restrictions. This assumption is fair, as previous studies have shown that similar travel patterns exist across years (*26*), other studies have also used prior year data to approximate travel patterns during 2020 (*6, 13*), and Baidu Mobility Index patterns are similar during 2019 and 2020 prior to January 23. Here, we used daily data from the first 14 days of Chunyun in 2018 (February 1 – 14, 2018) as proxy travel data for the first 14 days of Chunyun in 2020 (January 10 – 23, 2020).

Mobility for all age groups are included in the Tencent data. Although the usage of mobile differs among age groups, the constituent sources, particularly WeChat, have good penetration in older populations, as they are widely used for payments. Thus, we believe the data are representative of mobility patterns during Chunyun.

In the Tencent mobility data, for each day, the top 10 outflows from each of 375 Chinese cities were recorded. For city-to-city connections for which only some of the days in this two-week time period rank in the top 10, we linearly interpolated missing daily outflow values. In total, 92,248 inter-city travel records were used to represent travel during January 10-23.

Strict travel restrictions were implemented in several Chinese cities beginning January 23, 2020. As a result, the 2018 mobility data we use are likely not as representative of inter-city human movement after travel restrictions were implemented in Wuhan. Indeed, the 2020 Baidu Mobility Index data (*16*) indicate a precipitous decrease in travel to and from Wuhan occurs after January 23 (see Section 11, below). Inflow to Wuhan drops about 80%, and outflow from Wuhan drops about 98%. We use these data on relative travel patterns after January 23 to inform the model fitting for January 24-February 8 (see Section 11, below).

4.1 Gravity Model Test

Note that we tested fitting the mobility data to gravity models with different deterrence functions (i.e., power-law $M_{ij} = kp_i^{\tau_1}p_j^{\tau_2}/d_{ij}^{\rho}$, exponential $M_{ij} = kp_i^{\tau_1}p_j^{\tau_2}/\exp(d_{ij}/r)$, and truncated power-law $M_{ij} = kp_i^{\tau_1}p_j^{\tau_2}/[d_{ij}^{\rho}\exp(d_{ij}/r)]$), where p_i is the population of city i and d_{ij} is the geographical distance (km) between city i and j. The gravity models do not effectively reflect the observed mobility pattern (Fig. S3). As a consequence, we concluded that using a gravity model in the inference would not yield credible results and instead developed the metapopulation model presented in Section 1 above.

5. Model-inference framework

We infer model epidemiological parameters using an iterated filtering (IF) approach (*8-10*). The IF framework can be used to infer the maximum likelihood estimates of parameters in epidemic models and has been successfully applied to infectious diseases such as cholera (*9*) and measles (*27*). The IF framework is designed as follows: an ensemble of system states, which represent the distribution of parameters and variables, are repeatedly adjusted using filtering techniques in a series of iterations, during which the variance is gradually tuned down. In the process, the distribution of parameters is iteratively optimized per observations and converges to values that achieve maximum likelihood.

In its original implementation, the filtering technique used for IF was sequential Monte Carlo (i.e. a particle filter) (*28*). Here, due to the high-dimensionality of the metapopulation model, we used

a different efficient data assimilation algorithm - the Ensemble Adjustment Kalman Filter (EAKF) (*29*). Particle filters require a large number of particles (*30*); however, the EAKF can generate similar results using only hundreds of ensemble members (*20*). Originally developed for use in weather prediction, the EAKF assumes a Gaussian distribution of both the prior and likelihood and adjusts the prior distribution to a posterior using Bayes rule deterministically (*29*).

To represent the state-space distribution, the EAKF maintains an ensemble of system state vectors acting as samples from the distribution. In particular, the EAKF assumes that both the prior distribution and likelihood are Gaussian, and thus can be fully characterized by their first two moments (mean and variance). The update scheme for ensemble members is computed using Bayes rule (posterior $\propto$ prior $\times$ likelihood) via the convolution of the two Gaussian distributions. For observed state variables, the posterior of the ith ensemble member is updated through

$$o_{t,post}^{i} = \frac{\sigma_{t,obs}^{2}}{\sigma_{t,obs}^{2} + \sigma_{t,prior}^{2}} \bar{o}_{t,prior} + \frac{\sigma_{t,prior}^{2}}{\sigma_{t,obs}^{2} + \sigma_{t,prior}^{2}} o_t + \sqrt{\frac{\sigma_{t,obs}^{2}}{\sigma_{t,obs}^{2} + \sigma_{t,prior}^{2}}} \left(o_{t,prior}^{i} - \bar{o}_{t,prior}\right).$$

Here $o_{t,post}^{i}$ and $o_{t,prior}^{i}$ are the posterior and prior of the observed variable for the ith ensemble member at time t; $\bar{o}_{t,prior}$ is the mean of the prior observed variable; $\sigma_{t,obs}^{2}$ and $\sigma_{t,prior}^{2}$ are the variances of the observation and the prior observed variable; and o_t is the observation at time t. Unobserved variables and parameters are updated through their covariability with the observed variable, which can be computed directly from the ensemble. In particular, the ith ensemble member of unobserved variable or parameter x^i is updated by

$$x_{t,post}^{i} = x_{t,prior}^{i} + \frac{\sigma\left(\left\{x_{t,prior}\right\}_n, \left\{o_{t,prior}\right\}_n\right)}{\sigma_{t,prior}^{2}} \left(o_{t,post}^{i} - o_{t,prior}^{i}\right).$$

Here $x_{t,post}^{i}$ and $x_{t,prior}^{i}$ are the posterior and prior of the unobserved variable or parameter for the ith ensemble member at time t; and $\sigma\left(\left\{x_{t,prior}\right\}_n, \left\{o_{t,prior}\right\}_n\right)$ is the covariance between the prior of the unobserved variable or parameter $\left\{x_{t,prior}\right\}_n$ and the prior of the observed variable $\left\{o_{t,prior}\right\}_n$ at time t. In the EAKF, variables and parameters are updated deterministically so that the higher moments of the prior distribution are preserved in the posterior.

In applying the EAKF, we used the daily number of reported cases in city l on a given day t, y_l^t, as observations. For each y_l^t, we assume a heuristic observation error variance (OEV, denoted by $\sigma_{t,l}^2$):

$$\sigma_{t,l}^{2} = \max\left(4, \frac{(y_l^t)^2}{4}\right).$$

Similar forms of OEV have been successfully used for inference and forecasting for a range of infectious diseases including influenza (*20, 31*), Ebola (*32*), West Nile virus (*33*) and respiratory syncytial virus (*34*). We did test an alternate Poisson OEV with form $\sigma_{t,l}^2 = \max(4, y_l^t)$; the results remain similar (see Section 10 below for details).

The IF-EAKF algorithm proceeds per the pseudo-code in Algorithm 1.

Algorithm 1. IF-EAKF

Input: The metapopulation model $\mathcal{M}$, observations $\{y_l^t\}$ in T days and M locations, the observational error variance (OEV) $\{\sigma_{t,l}^2\}$, the initial system state $\bar{x}_0$, the initial covariance matrix Σ, a discount factor $a \in (0,1)$, and the number of iterations L.

for $l = 1$ to L **do**

Generate an ensemble of system state with n members using a multivariate Gaussian distribution: $\{\hat{x}_l^0\}_n \sim \mathcal{N}\left(\bar{x}_{l-1}, a^{2(l-1)}\Sigma\right)$.

for $t = 1$ to T **do**

Run model $\mathcal{M}$ with posterior $\{\hat{x}_l^{t-1}\}_n$ obtained from last update for one day, and return the ensemble of weekly incidence: $\{o_l^t\}_n = \mathcal{M}(\{\hat{x}_l^{t-1}\}_n)$.

Update the prior distribution $\{x_l^t\}_n \equiv \{\hat{x}_l^{t-1}\}_n$ to posterior $\{\hat{x}_l^t\}_n$ using the EAKF: $\{\hat{x}_l^t\}_n = EAKF(\{x_l^t\}_n, \{y_l^t\}, \{o_l^t\}_n, \{\sigma_{t,l}^2\})$.

end for

Calculate the ensemble mean of posterior over time as the input in next iteration: $\bar{x}_l = \sum_t E_n(\{\hat{x}_l^t\}_n)/T$, where E_n computes ensemble mean.

end for

Output: $\bar{x}_L$ as the maximum likelihood estimate of the system state.

In each iteration of the IF, the standard deviation of each parameter is shrunk by a factor $a \in (0,1)$. In practice, the discount factor a can range between 0.9 and 0.99. If a is too small, the algorithm may 'quench' too fast and fail to find the MLE; if it is too close to 1, the algorithm may not converge in a reasonable time interval. The number of iterations required for this convergence was determined by inspecting the evolution of posterior parameter distributions. In particular, the iteration time should be set to avoid divergence in the EAKF, in which the ensemble distribution collapses to a narrow range. In our implementation, we used $n = 300$ ensemble members, a shrinking parameter $a = 0.9$ and an iteration number $L = 10$.

Algorithm 1 returns the MLEs for parameters. In different runs, the MLEs are slightly different due to the stochasticity in the model and the initialization of the inference algorithm. In this study, we ran 1,000 independent realizations to generate the average MLEs of inferred parameters and their corresponding 95% CIs.

6. Calculation of R_e in Wuhan city

We calculated the reproductive number R_e in Wuhan city using the inferred parameters. Specifically, R_e is the largest eigenvalue of the next-generation matrix (NGM) (*35, 36*). Define $X = [E, I^r, I^u]^T$ and $Y = [S, R]^T$. The vectors for new infection and other rates are:

$$\mathcal{F} = \begin{bmatrix} \frac{\beta S I^r}{N} + \frac{\mu\beta S I^u}{N} \\ 0 \\ 0 \end{bmatrix}, \mathcal{V} = \begin{bmatrix} \frac{E}{Z} \\ \frac{I^r}{D} - \frac{\alpha E}{Z} \\ \frac{I^u}{D} - \frac{(1-\alpha)E}{Z} \end{bmatrix}.$$

The disease-free equilibrium is $x_0 = [0,0,0,N,0]^T$. We then have

$$F = \frac{\partial \mathcal{F}}{dX}\Big|_{x_0} = \begin{pmatrix} 0 & \beta & \mu\beta \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

and

$$V = \frac{\partial \mathcal{V}}{dX}\Big|_{x_0} = \begin{pmatrix} \frac{1}{Z} & 0 & 0 \\ -\frac{\alpha}{Z} & \frac{1}{D} & 0 \\ -\frac{1-\alpha}{Z} & 0 & \frac{1}{D} \end{pmatrix}.$$

The NGM is $K = FV^{-1}$. R_e is then computed as the leading eigenvalue of the NGM K, i.e.,

$$R_e = \alpha\beta D + (1-\alpha)\mu\beta D.$$

Of note, at the beginning of the epidemic (e.g., before January 23, 2020), R_e is equivalent to the basic reproductive number, R_0. As the epidemic unfolds, declines in population susceptibility reduce the effective reproductive number. However, here to assess the effectiveness of control measures, we used the same formula above for periods after January 23, 2020, without accounting for the relatively small decrease in population susceptibility due to infections during January 24 – February 8, 2020.

7. Synthetic testing

Before applying the model-inference framework to the observed COVID-19 incidence data, we tested the model-inference framework using model-generated outbreaks. Specifically, we generated a synthetic outbreak using a free simulation of the metapopulation model with a set of specified parameters. We then ran IF-EAKF inference using the daily cases for each city, as generated in stochastic free simulation, as observations. The aim is to determine whether the model-inference framework can ingest observations and recover the specified parameters. This assessment of the performance of the inference algorithm also allows inspection of the sensitivities of the inference results to model assumptions.

7.1. Accuracy of parameter estimation

We first generated a synthetic outbreak using the following parameter values: $\beta = 1.0$, $\mu = 0.8$, $\theta = 1.4$, $Z = 4$ days, $\alpha = 0.1$, $D = 4$ days, $T_d = 6$ days. For the IF, a 300-member ensemble was used. Priors of variables and parameters were drawn from the ranges reported in Section 1 above using a Latin Hypercubic Sampling algorithm. We used $Seed_{max} = 2{,}000$ in Wuhan city to initiate the outbreak. During inference, the seeding parameter $Seed_{max}$ was also set to 2,000. To account for stochastic effects, we applied the IF-EAKF inference algorithm 300 times and report the distributions of estimated parameters. All parameters were accurately estimated (Fig. S4) and the effective reproductive number R_e was recovered (Fig. S5). After each iteration, the variances of estimated parameters were reduced, and mean estimates converged to stable values (Fig. S6).

To further validate the inference approach, we also tested the system on another synthetic outbreak generated with a lower R_e ($\beta = 1.0$, $\mu = 0.6$, $\theta = 1.4$, $Z = 4$ days, $\alpha = 0.2$, $D = 4$

days, $T_d = 6$ days). Again, epidemiological parameters and R_e were captured by the estimated distributions (see Figs. S7-8).

Two additional synthetic outbreak were also tested: the first with a higher reporting rate ($\alpha = 0.5$, $\beta = 1.0$, $\mu = 0.6$, $\theta = 1.4$, $Z = 4$ days, $\alpha = 0.5$, $D = 4$ days, $T_d = 6$ days); the second with a still higher reporting rate and low transmission reduction factor for undocumented infections ($\alpha = 0.7$, $\beta = 1.0$, $\mu = 0.3$, $\theta = 1.4$, $Z = 4$ days, $\alpha = 0.5$, $D = 4$ days, $T_d = 6$ days). As shown in Figs. S9-12, these parameter combinations were also accurately estimated.

Together, the synthetic testing indicates that all metapopulation model parameters are identifiable, including both α and μ, which are the core focus of this study. In particular, we note the parameters are well estimated for outbreaks generated with a high transmission reduction factor for undocumented infections and a low fraction of documented infections (i.e. $\mu = 0.8$ and $\alpha = 0.1$, Figs. S4-5) and a low transmission reduction factor for undocumented infections and a high fraction of documented infections (i.e. $\mu = 0.3$ and $\alpha = 0.7$, Figs. S11-12).

7.2. Sensitivity of parameter estimation to seeding

As the numbers of exposed (E) and undocumented infected (I^u) individuals are unobserved, we estimated these state variables along with the other parameters/variables using the IF-EAKF approach. In particular, E and I^u may be sensitive to the imposed range of initial seeding ($Seed_{max}$). We examined the sensitivity of the overall parameter estimation to the seeding parameter $Seed_{max}$. To do this, we repeated the inference shown in Figs. S4-5 using a higher seeding parameter $Seed_{max} = 3{,}000$ (the true $Seed_{max}$ is 2,000). As shown in Figs. S13-14, with a mis-specified, higher prior for $Seed_{max}$, the estimation biases remain limited and unchanged, and R_e is identified.

7.3. System identifiability

Identification of model parameters is the central aim of this work. Our contention is that the assimilation and use of multiple streams of data from different cities, along with the movement data, enables parameter estimation.

To demonstrate that the utilization of multiple data streams in the inference system improves model parameter identifiability, we performed synthetic tests using four additional transmission models: 1) a reduction of the metapopulation model to a single location, *i*, i.e.

$$\frac{dS_i}{dt} = -\frac{\beta S_i I_i^r}{N_i} - \frac{\mu\beta S_i I_i^u}{N_i}$$
$$\frac{dE_i}{dt} = \frac{\beta S_i I_i^r}{N_i} + \frac{\mu\beta S_i I_i^u}{N_i} - \frac{E_i}{Z}$$
$$\frac{dI_i^r}{dt} = \alpha\frac{E_i}{Z} - \frac{I_i^r}{D^r}$$
$$\frac{dI_i^u}{dt} = (1-\alpha)\frac{E_i}{Z} - \frac{I_i^u}{D^u}$$

2) the metapopulation model representing only two cities (Wuhan and Xiaogan); 3) the metapopulation model representing only ten cities (Wuhan, Yichang, Xiangyang, Jinmen, Xiaogan, Huanggang, Xianning, Suizhou and Enshi), and 4) the metapopulation model

representing 50 cities. The distributions of estimated parameters and the actual parameters used to generate synthetic outbreaks are shown in Figs. S15. Several parameters in the single location model are not well identified, in particular the relative transmissibility μ, and the mean infection period, D; however, as more data streams and mobility data are used to constrain the system, parameter posterior credible intervals capture the truth. These simulation findings indicate that accurate constraint of system parameters is supported by the assimilation of observational data streams from multiple locations (375 cities in the full model, Figs. S4-14).

To demonstrate the importance of the Tencent mobility data and geographic interconnectedness for parameter estimation, we performed additional synthetic testing using the ten-city model. Specifically, we generated a synthetic outbreak using the metapopulation model with inter-city movement, but shut down this inter-city mobility during inference. The distributions of estimated parameters for this experiment are shown in Fig. S16. Without inter-city connectedness, the model parameters cannot be accurately estimated. Together, the results presented in Figs. S15-S16 indicate that the assimilation of data from multiple cities and the inclusion of movement information support constraint of all 6 model parameters, including the fraction of documented infections, α, and the relative transmissibility μ.

8. Inference using documented cases

We used the reported cases from 375 Chinese cities during January 10, 2020 to January 23, 2020 to infer model parameters. In total, 801 cases were reported, with 454 cases in Wuhan city.

We tested a range of reporting delays ($T_d = 6$, 7, 8, 9 and 10 days). For each combination of seeding and reporting delay parameters, we ran the inference 300 times. To validate the estimates, we generated 2000 outbreaks using the inferred mean parameters with seeding randomly drawn from a uniform distribution $U[0, Seed_{max} = 2{,}000]$, and then compared the distributions of simulated new cases in all cities with reported case observations. The goodness-of-fit was measured using log-likelihood (LL). The log-likelihood is computed using the posterior distribution of confirmed cases in each city. For each observation, we calculate the logarithmic value of the weight assigned to a +/-15% interval around the reported incidence (+/- 10%, 15% and 20% were also tested, and the results remained the same). We set the minimum logarithmic value as -20. LL is the sum of these values. Inference results for the best-fitting model with the maximum LL ($T_d = 9$ days) are shown in Table 1 of the main text, and model fitting for this inference solution is shown in Fig. 1 of the main text.

9. Spatial movement of COVID-19 in China

Using the best-fitting model ($Seed_{max} = 2{,}000$, $T_d = 9$ days), we generated 300 simulated outbreaks starting January 10th until January 23rd. We computed the daily number of cities with cumulative incidence ≥ 10 and compared these distributions with the reported numbers of invaded cities during the same period (Fig. S17). The observations and simulations are in good agreement. In particular, major cities outside epicenter and connected to Wuhan, specifically, Huanggang, Beijing, Shenzhen, Shanghai, Xiaogan and Chongqing, reached 10 cumulative cases on January 20, 21, 21, 22, 23, 23, respectively. In our simulated outbreaks (2000 free simulations using inferred parameters), Beijing, Xiaogan, Huanggang, Chongqing and Jinmen reached 10 cumulative cases on January 21, 22, 22, 22, and 23, respectively.

We also generated simulations with the best-fitting parameter estimates but using re-wired outflows from each city to other randomly selected cities (i.e. a scrambled mobility matrix). This disruption of the mobility matrix greatly altered the cities with cumulative incidence ≥10 before January 23. For instance, in one realization, Tonghua, Jieyang, Yutian and Shenyang reached 10 cumulative cases on Jan 22, 22, 22, and 23, respectively. This result is completely distinct from the observed spatial spread pattern.

10. Sensitivity of parameter estimates and identifiability

Distributions for estimated parameters and R_e for different settings of T_d ($T_d = 6, 7, 8, 9, 10$) and $Seed_{max}$ ($Seed_{max} = 1500, 2500$) are shown in Fig. S18 and Fig. S19. Estimations of β, μ, θ, Z and D are robust to different settings of T_d and $Seed_{max}$. The relative insensitivity to the value T_d suggests that as long as a random Gamma distributed reporting delay of sufficient mean length is imposed, the parameters can be identified based on the growth and intercity spread of confirmed cases. The insensitivity to $Seed_{max}$ is likely due to two factors: 1) initial infections are drawn uniformly from zero to $Seed_{max}$, not at $Seed_{max}$; and 2) the EAKF, by continually adjusting state variable estimates with each assimilation of data (i.e. observed confirmed cases), can quickly adjust spuriously high or low initial state values to more realistic levels.

To test the sensitivity of parameter estimates to the assumed OEV form, we performed another inference using a Poisson OEV, specifically, $\sigma_{t,l}^2 = \max(4, y_l^t)$. Estimation was robust to this OEV form (Fig. S20).

We further tested the sensitivity of the parameter estimates to the form of the prior distributions. In particular, instead of generating the initial priors using Latin hypercube sampling from uniform distributions, we used normal distributions with means set as the midpoint of the uniform prior ranges and standard deviations set to 30% of those respective mean values. The inference results remain similar (Fig. S21). Again, this robustness may be partly due to the capability of the EAKF to migrate posterior distributions toward true parameter values.

We also ran inference assuming that there was an additional reporting delay before January 17. Specifically, for documented infections occurring before January 17, 2020, we added two days to the randomly generated Gamma reporting delay. The results in Fig. S22 indicate that this additional delay does not affect the inference findings.

To further explore the identifiability of α and μ, we examined how marginal variations of these two parameters affect R_e (Fig. 1E). An increase of either parameter leads to a monotonic increase of R_e and the accumulation of more reported cases; however, R_e is more sensitive to changes to μ, the relative contagiousness of undocumented infections. Because multiple combinations of α and μ produce the same estimate for R_e, we also tested whether the combination identified by our model-inference framework (i.e. $\alpha = 0.14$, $\mu = 0.55$) is maximally likely. Indeed, this identified combination has the highest log-likelihood (Fig. 1F), indicating that the framework, given the abundance of observations (daily data from 375 cities), is able to discriminate among combinations of α and μ.

Lastly, to test whether the estimates for $\alpha, \beta, \mu, Z,$ and θ were sensitive to the use of a single average infection period, D, we fit the daily incidence data in 375 cities during January 10-23 to a metapopulation model in which documented and undocumented infections had separate average infection periods (denoted by D^r and D^u). Here the model is:

$$\frac{dS_i}{dt} = -\frac{\beta S_i I_i^r}{N_i} - \frac{\mu\beta S_i I_i^u}{N_i} + \theta \sum_j \frac{M_{ij} S_j}{N_j - I_j^r} - \theta \sum_j \frac{M_{ji} S_i}{N_i - I_i^r}$$
$$\frac{dE_i}{dt} = \frac{\beta S_i I_i^r}{N_i} + \frac{\mu\beta S_i I_i^u}{N_i} - \frac{E_i}{Z} + \theta \sum_j \frac{M_{ij} E_j}{N_j - I_j^r} - \theta \sum_j \frac{M_{ji} E_i}{N_i - I_i^r}$$
$$\frac{dI_i^r}{dt} = \alpha \frac{E_i}{Z} - \frac{I_i^r}{D^r}$$
$$\frac{dI_i^u}{dt} = (1-\alpha)\frac{E_i}{Z} - \frac{I_i^u}{D^u} + \theta \sum_j \frac{M_{ij} I_j^u}{N_j - I_j^r} - \theta \sum_j \frac{M_{ji} I_i^u}{N_i - I_i^r}$$
$$N_i = N_i + \theta \sum_j M_{ij} - \theta \sum_j M_{ji}$$

The best-fit model posterior estimates of key epidemiological parameters are reported in Table S1 and are similar with those obtained using the model presented in the main text (Table 1). Further, the estimates for D^r and D^u are similar to D.

11. Inference of model parameters after January 23, 2020

We modeled the transmission of SARS-CoV2 in China after implementation of control measures on January 23. These control measures included: travel restrictions imposed between major cities and Wuhan; self-quarantine and contact precautions advocated by the government; and more available rapid testing for infection confirmation (*11-12*). These control measures along with changes in care-seeking behavior due to increased awareness of the virus and increased personal protection behavior (e.g. wearing of facemasks, self-isolation when sick), likely altered the epidemiological characteristics of the outbreak after January 23. To quantify these differences, we re-estimated the system parameters using the metapopulation model-inference framework and city-level daily cases reported between January 24 and February 8. As the compliance of the inter-city travel restriction is unknown, we tested the following two scenarios:

1) Travel from and to Wuhan are reduced by 98%, and other inter-city travel is reduced by 80%.
2) All inter-city mobility is shut down.

The first scenario is informed by the reductions observed in the Baidu Mobility Index (Table S2). In addition, to represent reduced person-to-person contact and increased infection detection, we updated the initial priors for β to $[0.2, 1.2]$ for both extreme scenarios. We also tested a range of reporting delays, T_d, from 6 days to 10 days. As before, we used the daily reported cases in all cities to compute the log-likelihood.

In order to reflect the rapid change in control efforts, we inferred model parameters during two overlapping periods: January 24 to February 3 and January 24 to February 8. For these periods, the best-fitting models are shown in Figs. S23-S26. Estimated parameters, R_e and goodness-of-fit are reported in Table 2 of the main text and Table S3.

12. Independent model validation using infection rates among evacuees to other countries

A recent study (*37*) summarized infection rates in evacuees to Singapore, South Korea, Japan and Germany at the end of January. The average infection rate was reported as 1.39%. Based on this estimate, we performed two independent tests to corroborate the parameters inferred by the metapopulation model.

1). According to official report, around 5 million people left Wuhan city before January 23. The total population in Wuhan city after January 23 (when travel restrictions were imposed) is therefore around 6 million. A 1.39% infection rate suggests an estimated 83,400 infections prior to February 1 in Wuhan. Simulation with the metapopulation model using inferred parameters produces a total infected population of 48,420 (95% CI [10,849, 89,524]) before Feb 1, which in general matches the estimated 83,400 infections in magnitude.

2). In our model, infections occurring prior to February 1 will be documented with a reporting delay (as inferred, an average of 9 days before January 23 and 6 days after January 23). For simplicity, we assume the reporting delay for each person is constant. Based on this assumption, infections before February 1 should continue appearing until February 7. The cumulative confirmed cases for February 7 in Wuhan is 13,562, which suggests a *reported* infection rate of 13,562/6 millions = 0.22%. Compared with the infection rate 1.39% (*37*), the reporting rate in Wuhan should be 0.22%/1.39%=15.8%. This estimate generally agrees with our inferred reporting rate of 14% before January 23.

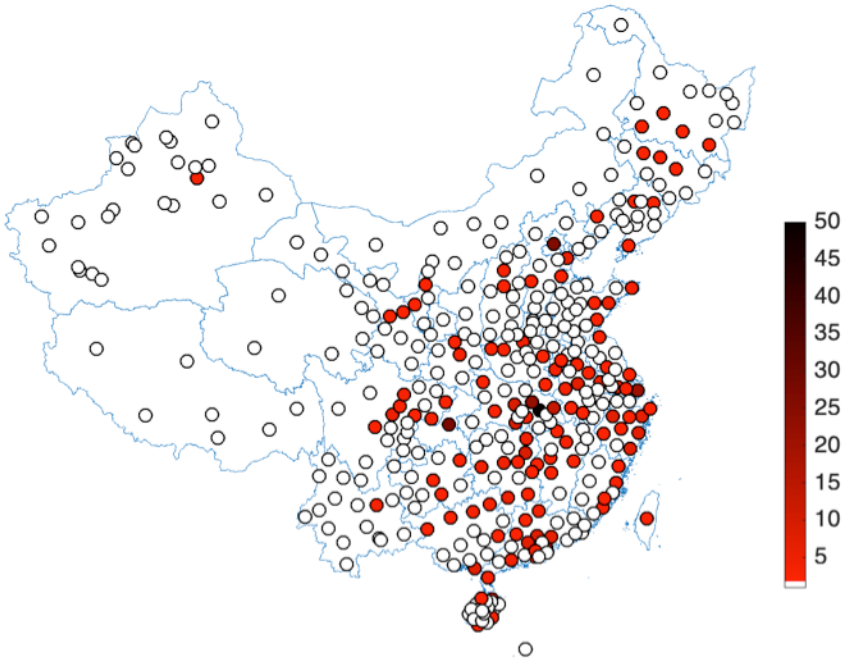

Fig. S1. Cumulative cases in 375 Chinese cities on January 23, 2020. Wuhan city had 454 cases. White circles indicate zero reported cases.

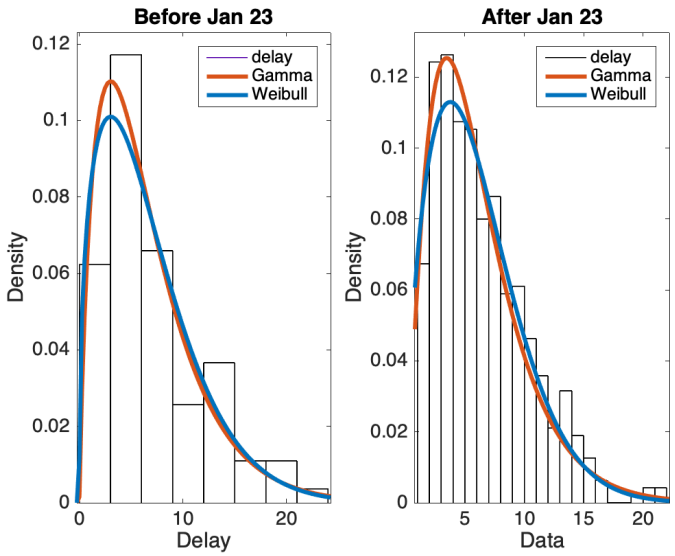

Fig. S2. Distribution of interval between onset of symptoms and confirmation (days) for cases confirmed before (left) and after (right) January 23, 2020 (*17*). The data prior to January 23 were better fitted with a Gamma distribution ($\boldsymbol{a = 1.85, b = 3.57, LL = -252.24}$) than a Weibull distribution ($\boldsymbol{A = 7.29, B = 1.41, LL = -255.17}$). After January 23, the data were better fitted with a Gamma distribution ($\boldsymbol{a = 2.24, b = 2.59, LL = -1251.94}$) than a Weibull distribution ($\boldsymbol{A = 6.78, B = 1.63, LL = -1255.38}$).

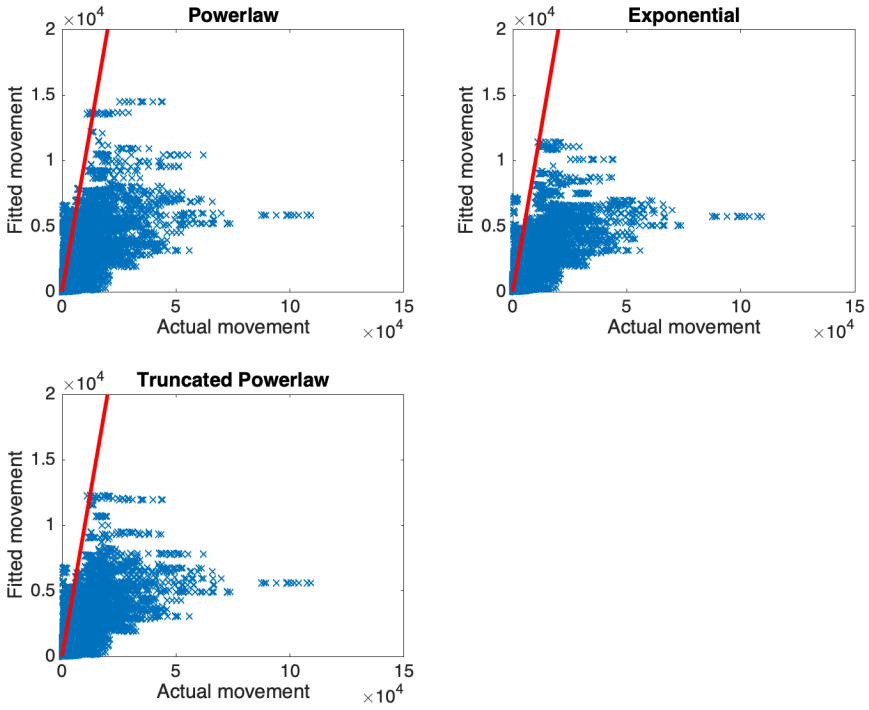

Fig. S3. Fitting the Tencent mobility data to gravity models with power-law, exponential and truncated power-law deterrence functions. The red line shows $\boldsymbol{y = x}$.

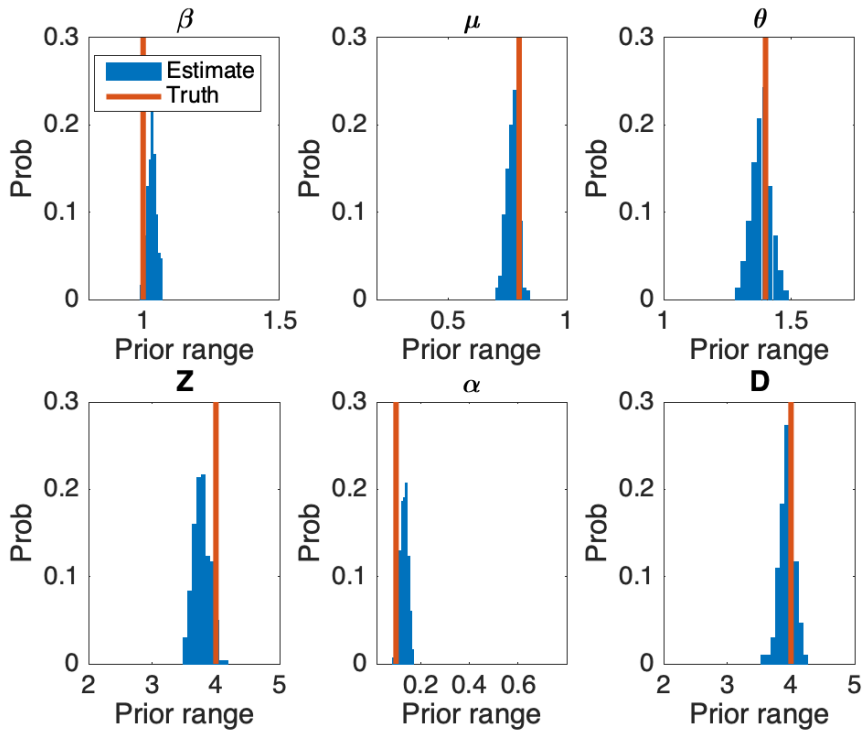

Fig. S4. Accuracy of parameter estimation. The actual parameters used in generating the synthetic outbreak are depicted by vertical red lines. Blue bars represent the distribution of the posterior parameter estimates. The ranges of the x-axis are set as the initial prior parameter ranges.

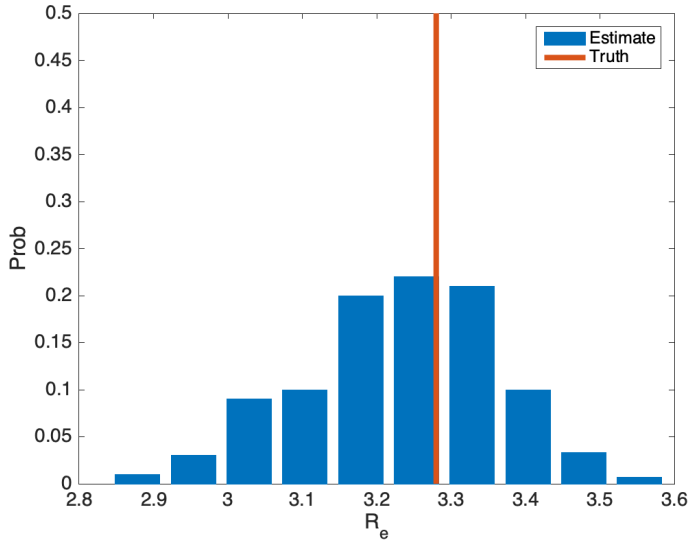

Fig. S5. Comparison of the actual R_e (vertical red line) and the distribution of estimated R_e (blue bars) for the fittings shown in Figure S3.

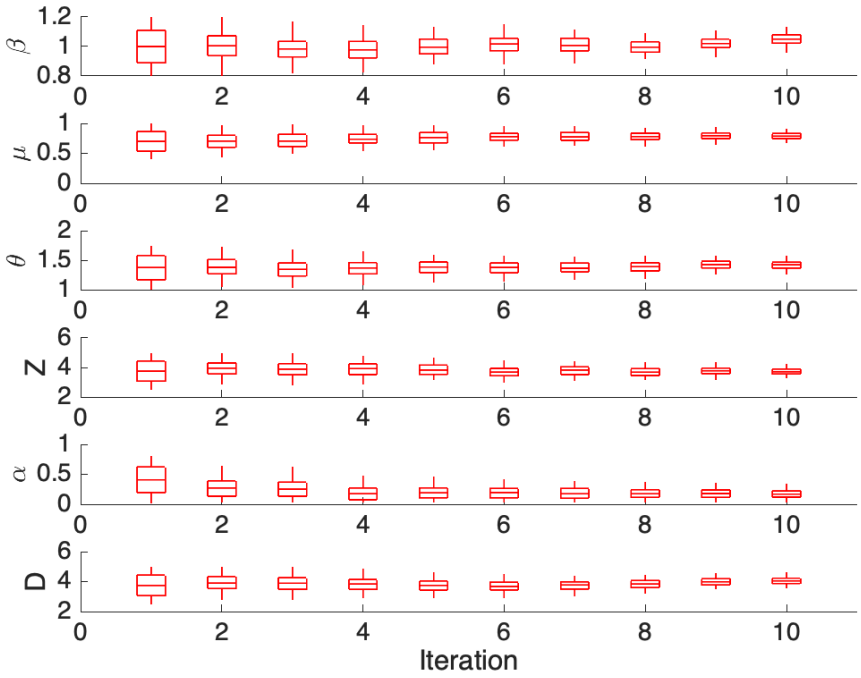

Fig. S6. Distributions of posterior estimates for all parameters at the beginning of each iteration (L = 10). The variance of the posterior estimates decreases, and the mean estimates converge to stable values.

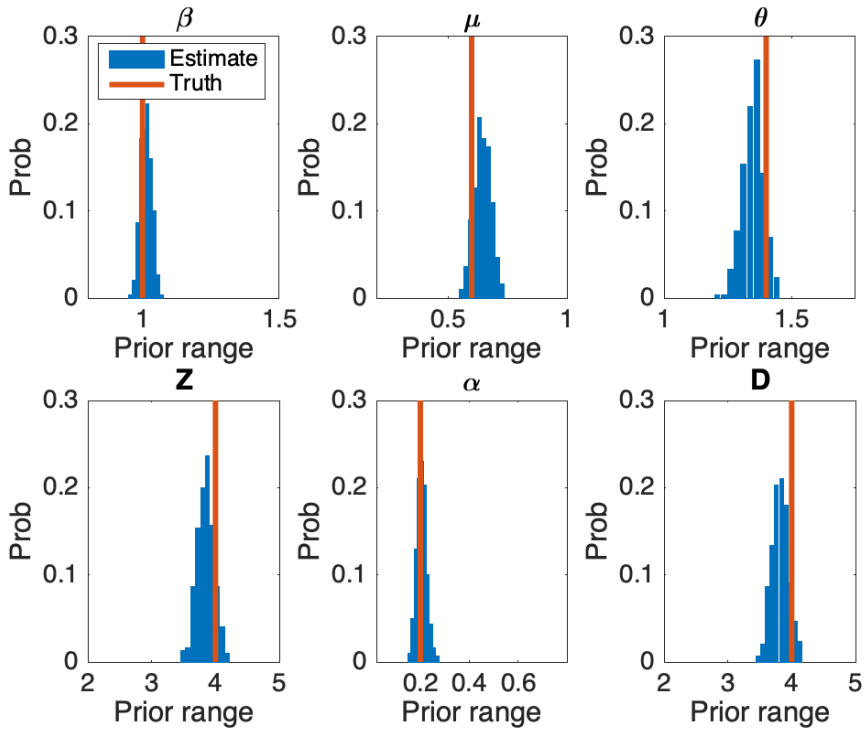

Fig. S7. Accuracy of parameter estimation. The actual parameters used in generating the synthetic outbreak are depicted by vertical red lines. Blue bars represent the distribution of the posterior parameter estimates. The ranges of the x-axis are set as the initial prior parameter ranges.

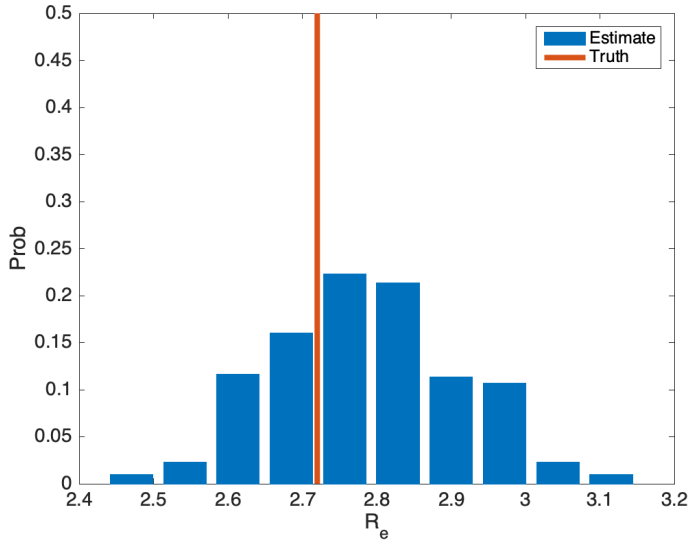

Fig. S8. Comparison of the actual $\boldsymbol{R_e}$ (vertical red line) and the distribution of estimated $\boldsymbol{R_e}$ (blue bars) for the fittings shown in Figure S6.

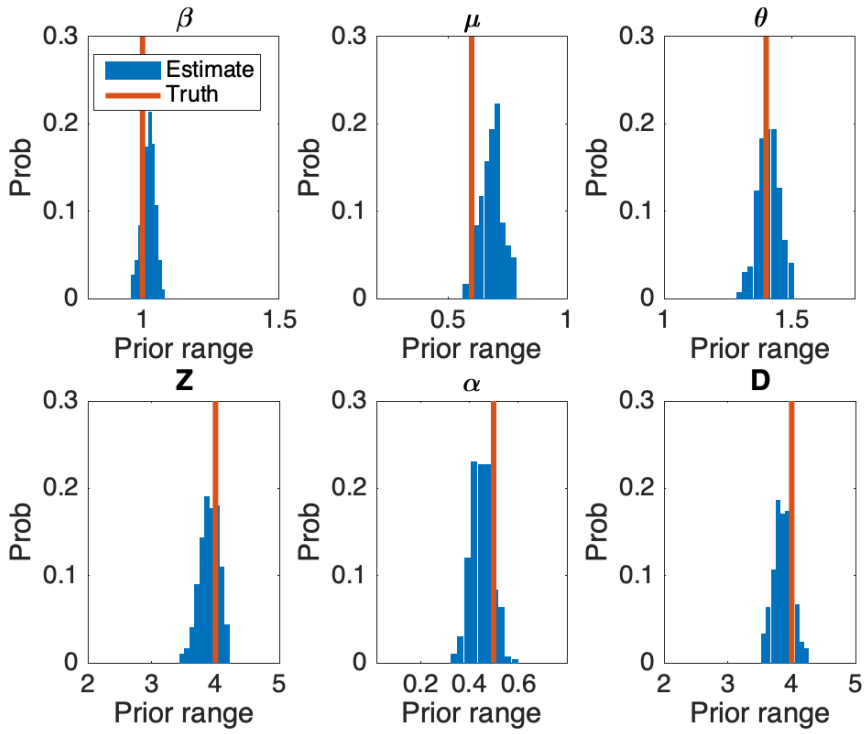

Fig. S9. Accuracy of parameter estimation. The actual parameters used in generating the synthetic outbreak are depicted by vertical red lines. Blue bars represent the distribution of the posterior parameter estimates. The ranges of the x-axis are set as the initial prior parameter ranges.

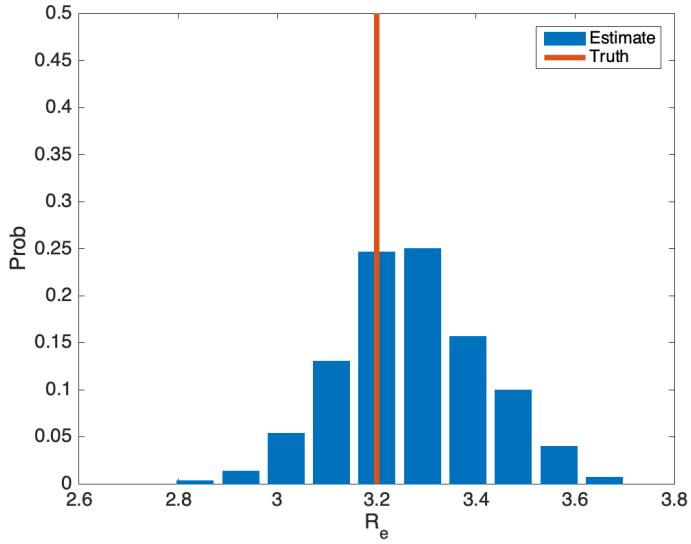

Fig. S10. Comparison of the actual R_e (vertical red line) and the distribution of estimated R_e (blue bars) for the fittings shown in Figure S8.

.

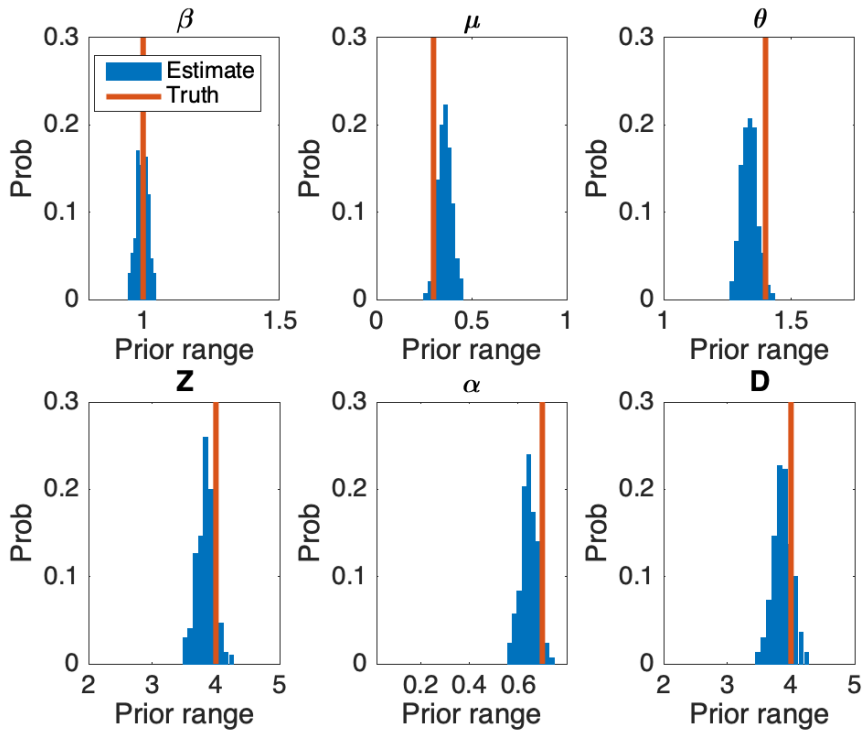

Fig. S11. Accuracy of parameter estimation for a truth with a high fraction of documented infections ($\boldsymbol{\alpha = 0.7}$) and low relative contagiousness of undocumented infections ($\boldsymbol{\mu = 0.3}$). The actual parameters used in generating the synthetic outbreak are depicted by vertical red lines. Blue bars represent the distribution of the posterior parameter estimates. The ranges of the x-axis are set as the initial prior parameter ranges.

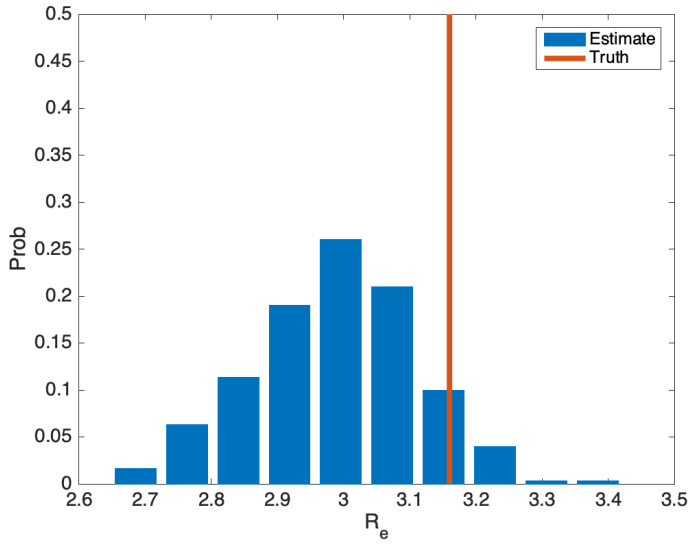

Fig. S12. Comparison of the actual $\boldsymbol{R_e}$ (vertical red line) and the distribution of estimated $\boldsymbol{R_e}$ (blue bars) for a truth with a high fraction of documented infections ($\boldsymbol{\alpha = 0.7}$) and low relative contagiousness of undocumented infections ($\boldsymbol{\mu = 0.3}$) (individual parameter estimates shown in Figure S10).

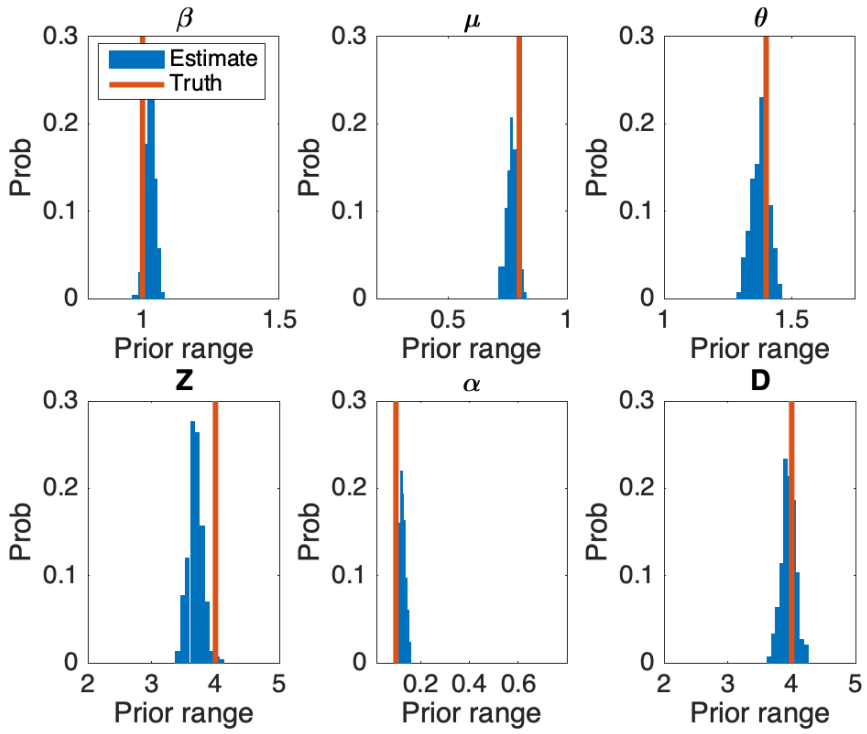

Fig. S13. Accuracy of parameter estimation. The actual parameters used in generating the synthetic outbreak are depicted by vertical red lines. Blue bars represent the distribution of the posterior parameter estimates. The ranges of the x-axis are set as the initial prior parameter ranges. The actual seeding parameter for the synthetic outbreak is $\boldsymbol{Seed_{max} = 2,000}$, while $\boldsymbol{Seed_{max}}$ was set to 3,000 during inference.

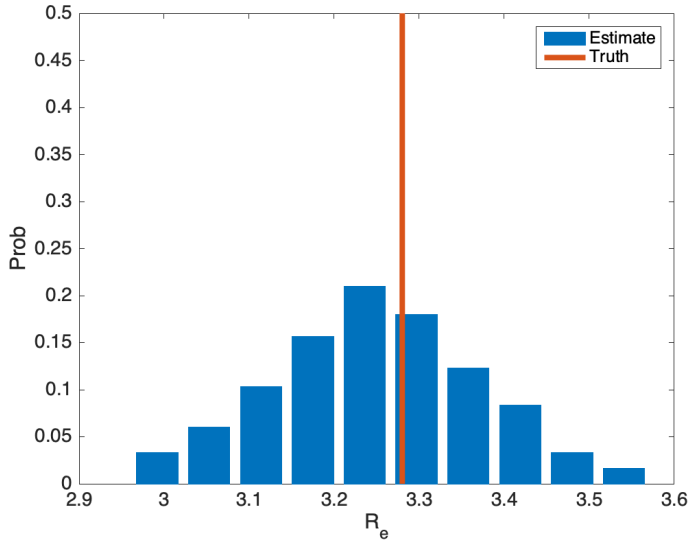

Fig. S14. Comparison of the actual $\boldsymbol{R_e}$ (vertical red line) and the distribution of estimated $\boldsymbol{R_e}$ (blue bars) for the fittings shown in Figure S12. The actual seeding parameter for the synthetic outbreak is $\boldsymbol{Seed_{max} = 2,000}$, while $\boldsymbol{Seed_{max}}$ was set to 3,000 during inference.

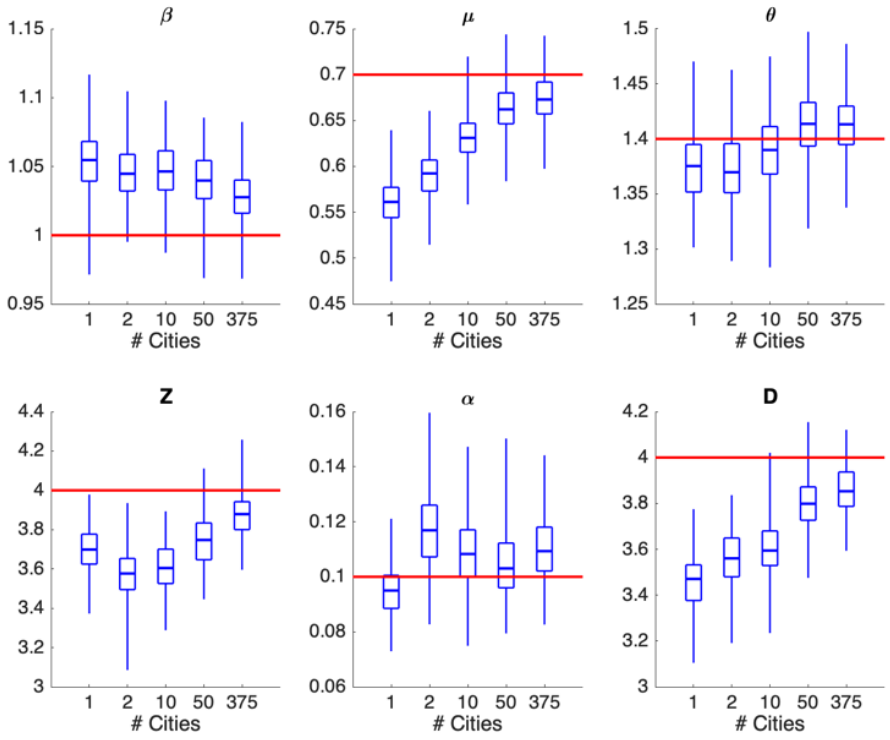

Fig. S15. Accuracy of parameter estimation. Estimation of system parameters using a single-location reduction of the metapopulation model and the metapopulation model implemented for 2, 10, 50 and 375 cities. The actual parameters used in generating the synthetic outbreak are depicted by horizontal red lines. Blue bars represent the distribution of the posterior parameter estimates.

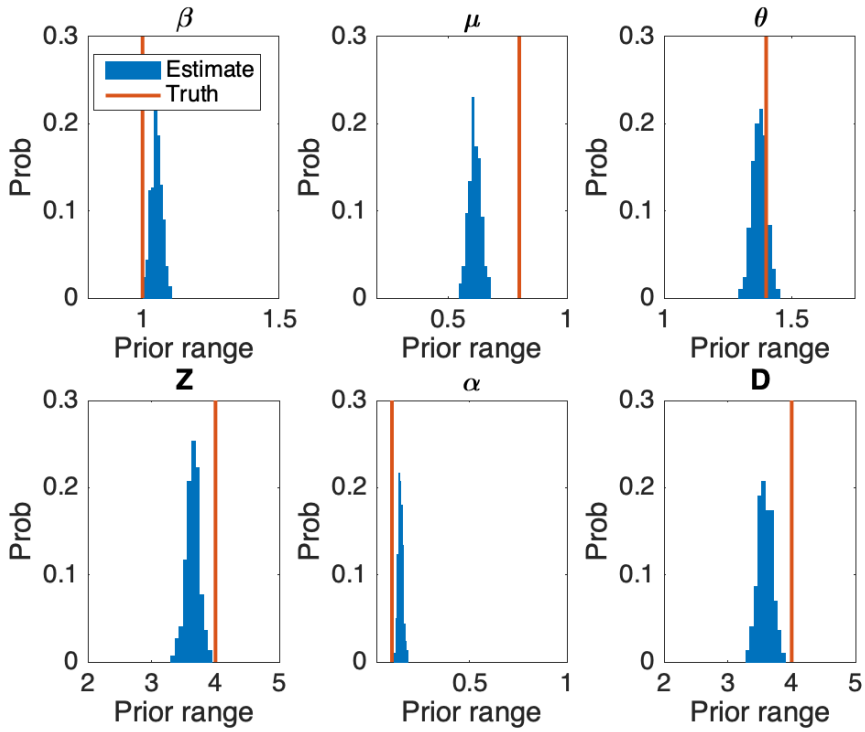

Fig. S16. Accuracy of parameter estimation. Estimation of system parameters using a metapopulation model consisting of ten cities: Wuhan, Yichang, Xiangyang, Jinmen, Xiaogan, Huanggang, Xianning, Suizhou and Enshi. Here, inter-city travel was shut down during inference. The actual parameters used in generating the synthetic outbreak are depicted by vertical red lines. Blue bars represent the distribution of the posterior parameter estimates. The ranges of the x-axis are set as the initial prior parameter ranges.

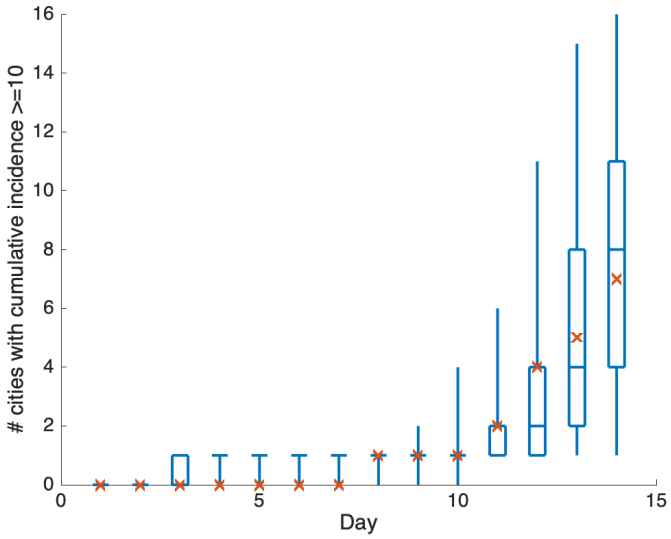

Fig. S17. Model-generated distributions of the number of cities with cumulative incidence ≥ 10 at each day from January 10 to January 23. Red crosses are reported numbers until January 23.

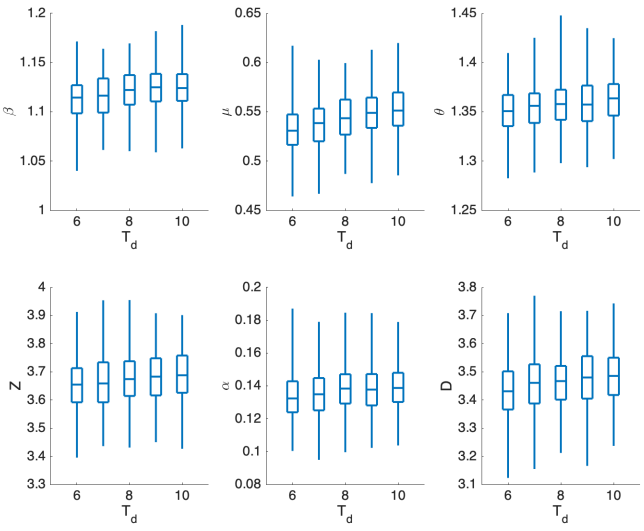

Fig. S18. Distributions of estimated parameters for different settings of $\boldsymbol{T_d}$. Boxes show median and interquartile values and whiskers indicate the 95% CIs.

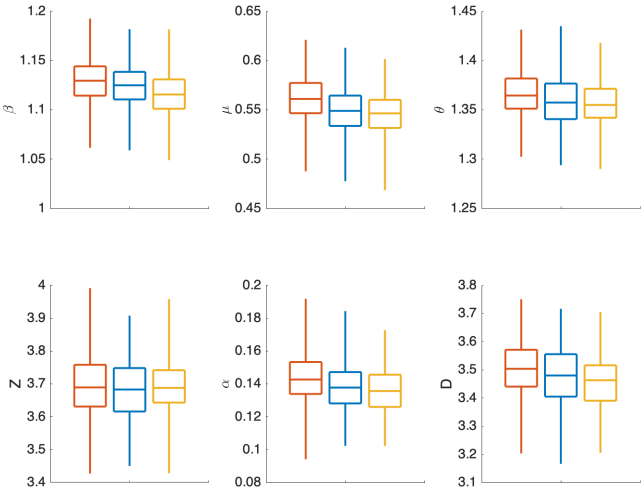

Fig. S19. Distributions of estimated parameters for different settings of $\boldsymbol{Seed_{max}}$. Boxes show median and interquartile values and whiskers indicate the 95% CIs. Red: $\boldsymbol{Seed_{max} = 1500}$, blue: $\boldsymbol{Seed_{max} = 2000}$, yellow: $\boldsymbol{Seed_{max} = 2500}$. We set $\boldsymbol{T_d = 9}$ days for all three inferences.

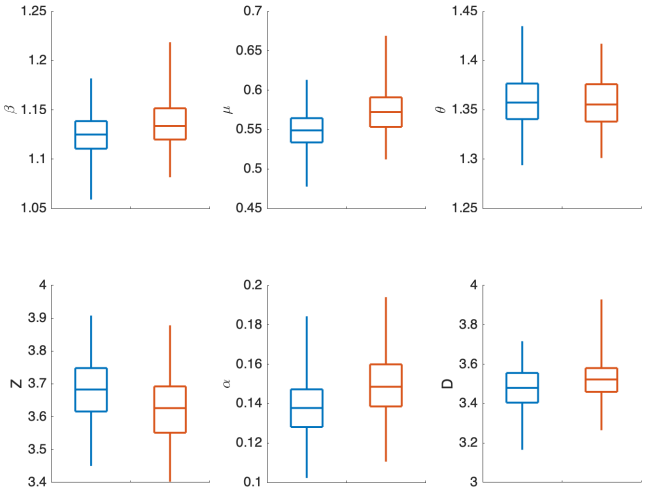

Fig. S20. Distributions of estimated parameters for different settings of OEV. Boxes show median and interquartile values and whiskers indicate the 95% CIs. Red: Poisson OEV $\boldsymbol{\sigma_{t,l}^2 = \max(4, y_l^t)}$, blue: $\boldsymbol{\sigma_{t,l}^2 = \max(4, (y_l^t)^2/4)}$. We set $\boldsymbol{Seed_{max} = 2000}$ and $\boldsymbol{T_d = 9}$ days for both inferences.

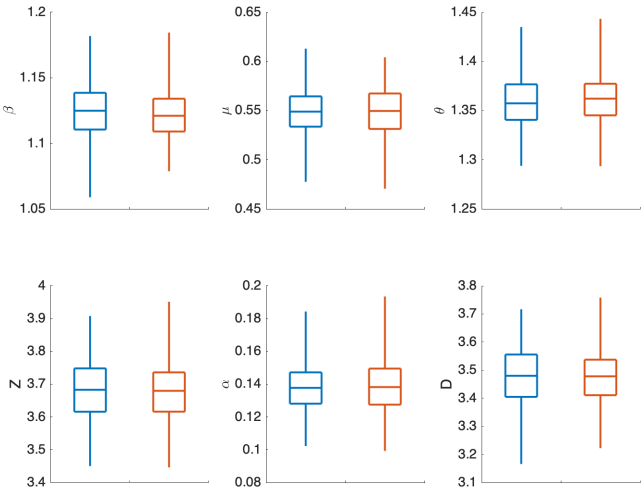

Fig. S21. Distributions of estimated parameters for different forms of prior. Boxes show median and interquartile values and whiskers indicate the 95% CIs. Red: Normal prior distribution with 30% standard deviations of mean values, blue: prior sampled using LHS. We set $\boldsymbol{Seed_{max}} = \mathbf{2000}$ and $\boldsymbol{T_d} = \mathbf{9}$ days for both inferences.

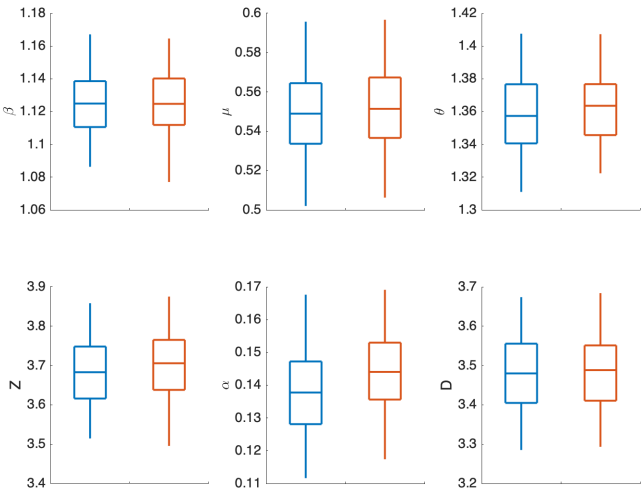

Fig. S22. Distributions of estimated parameters. Boxes show median and interquartile values and whiskers indicate the 95% CIs. Red: two-day additional reporting delay for infections occurred before January 23, blue: a same Gamma-distributed reporting delay across the two-week period. We set $\boldsymbol{Seed_{max} = 2000}$ and $\boldsymbol{T_d = 9}$ days for both inferences.

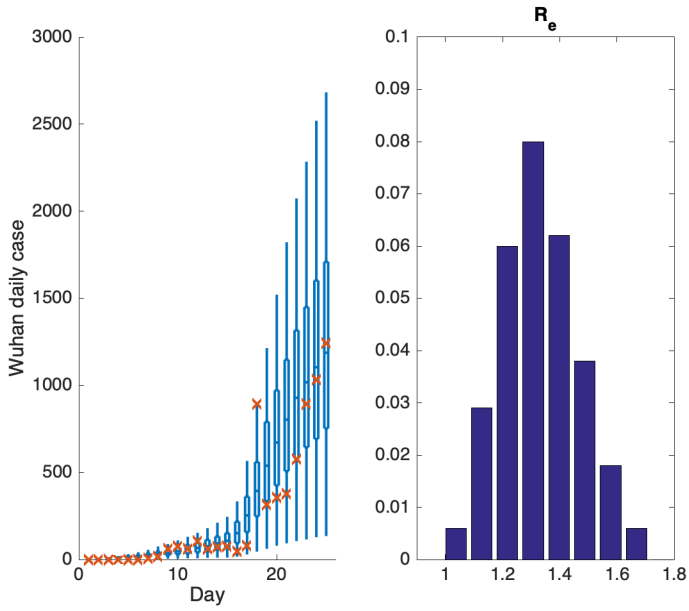

Fig. S23. Model fitting ($\boldsymbol{T_d = 6}$ days) to documented cases in Wuhan through February 3, 2020 with no travel between cities (left). The distribution of estimated $\boldsymbol{R_e}$ is shown in the right panel.

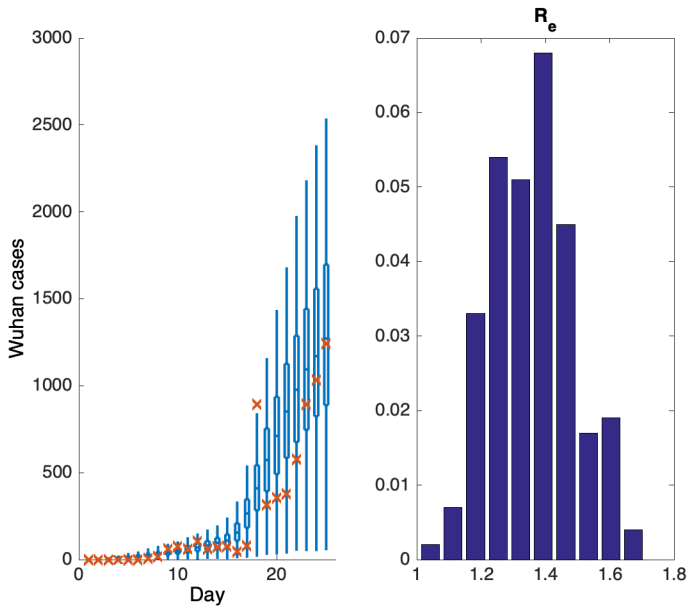

Fig. S24. Model fitting ($\boldsymbol{T_d = 6}$ days) to documented cases in Wuhan through February 3, 2020 (left). Travel to and from Wuhan is reduced by 98%, and other inter-city travel is reduced by 80%. The distribution of estimated $\boldsymbol{R_e}$ is shown in the right panel.

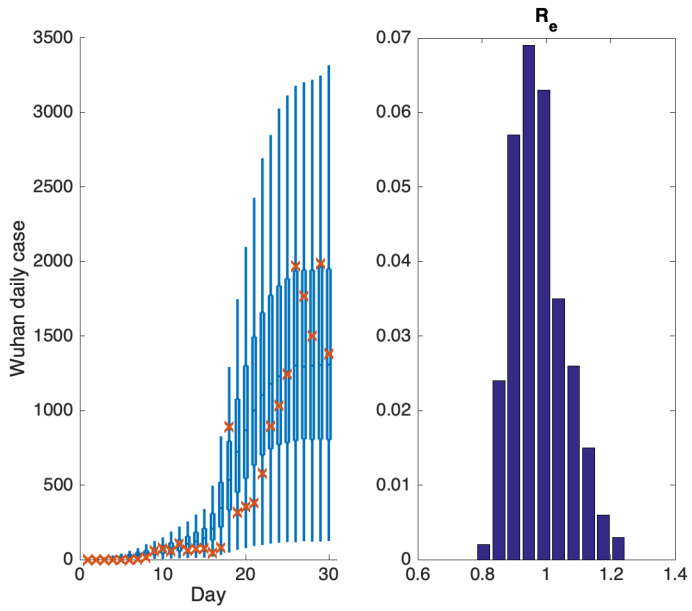

Fig. S25. Model fitting ($\boldsymbol{T_d = 6}$ days) to documented cases in Wuhan through February 8, 2020 with no travel between cities (left). The distribution of estimated $\boldsymbol{R_e}$ is shown in the right panel.

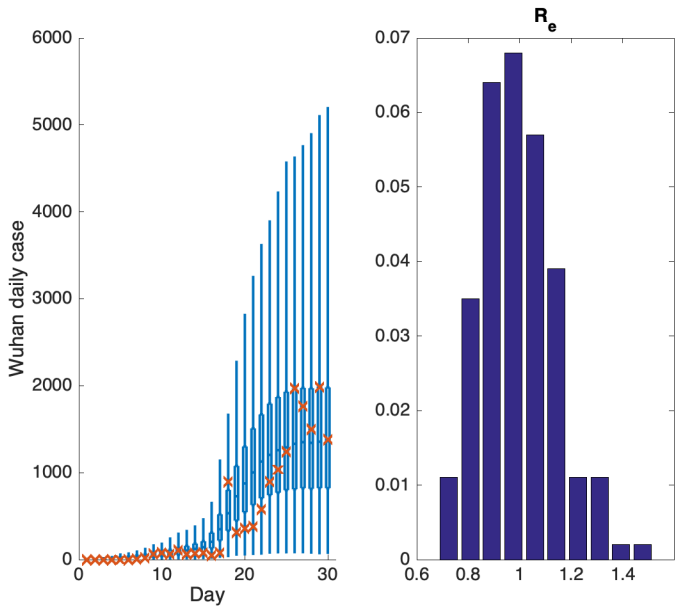

Fig. S26. Model fitting ($\boldsymbol{T_d = 6}$ days) to documented cases in Wuhan through February 8, 2020 (left). Travel to and from Wuhan is reduced by 98%, and other inter-city travel is ae reduced by 80%. The distribution of estimated $\boldsymbol{R_e}$ is shown in the right panel.

Table S1. Best-fit model posterior estimates of key epidemiological parameters for simulation during January 10-23, 2020 with the metapopulation model adjusted to include separate mean infectious periods for documented and undocumented infections ($Seed_{max} = 2000$, $T_d = 9$ days).

Parameter	Median (95% CIs)
Transmission rate (β, days^{-1})	1.12 (1.07, 1.17)
Relative transmission rate (μ)	0.55 (0.49, 0.60)
Latency period (Z, days)	3.68 (3.48, 3.90)
Infectious period for documented infections (D^r, days)	3.47 (3.26, 3.67)
Infectious period for undocumented infections (D^u, days)	3.45 (3.24, 3.70)
Reporting rate (α)	0.14 (0.10, 0.17)
Basic reproductive number (R_e)	2.38 (2.16, 2.59)
Mobility factor (θ)	1.36 (1.31, 1.42)

Table S2. Baidu Mobility Index travel to and from Wuhan during January 23 – February 8, 2020.

Date	Jan 23	24	25	26	27	28	29	30	31	Feb 1	2	3	4	5	6	7	8
To Wuhan	1.75	0.88	0.63	0.51	0.42	0.41	0.37	0.35	0.33	0.36	0.39	0.40	0.41	0.37	0.37	0.36	0.34
From Wuhan	11.14	3.89	1.30	0.66	0.43	0.32	0.26	0.24	0.24	0.24	0.46	0.21	0.23	0.28	0.28	0.27	0.28

Table S3. Best-fit model posterior estimates of key epidemiological parameters for simulation of the model with no travel between cities during January 24 – February 3 and January 24 – February 8 ($Seed_{max} = 2000$ on January 10, $T_d = 9$ days before January 24, $T_d = 6$ days between January 24 and February 8).

Parameter	January 24 – February 3 (Median (95% CIs))	January 24 - February 8 (Median (95% CIs))
Transmission rate (β, days^{-1})	0.51 (0.37, 0.68)	0.35 (0.30, 0.52)
Relative transmission rate (μ)	0.47 (0.36, 0.64)	0.42 (0.34, 0.61)
Latency period (Z, days)	3.62 (3.44, 3.87)	3.43 (3.30, 3.63)
Infectious period (D, days)	3.15 (2.62, 3.71)	3.32 (2.92, 4.04)
Reporting rate (α)	0.65 (0.60, 0.69)	0.69 (0.66, 0.71)
Effective reproductive number (R_e)	1.32 (1.07, 1.54)	0.96 (0.83, 1.16)

Data S1. (separate file) File (DataS1.zip) includes:

Data.zip:

Daily incidence by city, intercity movement, city coordinates, city populations (Incidence.csv; Mobility.csv; city_coordinates.csv; pop.csv)

Code.zip:

Matlab code and input files for running model-inference system (M.mat; SEIR.m; cities.mat; incidence.mat; inference.m initialize.m; pop.mat)

References and Notes

1. National Health Commission of the People's Republic of China, Update on the novel coronavirus pneumonia outbreak; http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml [accessed 8 February 2020].

2. World Health Organization, Coronavirus disease (COVID-2019) situation reports; https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/ [accessed 1 March 2020].

3. J. F. Chan, S. Yuan, K. H. Kok, K. K. To, H. Chu, J. Yang, F. Xing, J. Liu, C. C. Yip, R. W. Poon, H. W. Tsoi, S. K. Lo, K. H. Chan, V. K. Poon, W. M. Chan, J. D. Ip, J. P. Cai, V. C. Cheng, H. Chen, C. K. Hui, K. Y. Yuen, A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: A study of a family cluster. *Lancet* **395**, 514–523 (2020). doi:10.1016/S0140-6736(20)30154-9 Medline

4. P. Wu, X. Hao, E. H. Y. Lau, J. Y. Wong, K. S. M. Leung, J. T. Wu, B. J. Cowling, G. M. Leung, Real-time tentative assessment of the epidemiological characteristics of novel coronavirus infections in Wuhan, China, as at 22 January 2020. *Euro Surveill.* **25**, 2000044 (2020). doi:10.2807/1560-7917.ES.2020.25.3.2000044 Medline

5. V. J. Munster, M. Koopmans, N. van Doremalen, D. van Riel, E. de Wit, A novel coronavirus emerging in China - Key questions for impact assessment. *N. Engl. J. Med.* **382**, 692–694 (2020). doi:10.1056/NEJMp2000929 Medline

6. Z. Du, L. Wang, S. Cauchemez, X. Xu, X. Wang, B. J. Cowling, L. A. Meyers, Risk for transportation of 2019 novel coronavirus disease from Wuhan to other cities in China. *Emerg. Infect. Dis.* **26**, (2020). doi:10.3201/eid2605.200146 Medline

7. http://society.people.com.cn/n1/2018/0315/c1008-29869526.html [accessed 8 February 2020].

8. E. L. Ionides, C. Bretó, A. A. King, Inference for nonlinear dynamical systems. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 18438–18443 (2006). doi:10.1073/pnas.0603181103 Medline

9. A. A. King, E. L. Ionides, M. Pascual, M. J. Bouma, Inapparent infections and cholera dynamics. *Nature* **454**, 877–880 (2008). doi:10.1038/nature07084 Medline

10. S. Pei, F. Morone, F. Liljeros, H. Makse, J. L. Shaman, Inference and control of the nosocomial transmission of methicillin-resistant *Staphylococcus aureus*. *eLife* **7**, e40977 (2018). doi:10.7554/eLife.40977 Medline

11. Health Commission of Hubei Province, The 8th Press Conference on the Prevention and Control of COVID-19; http://wjw.hubei.gov.cn/fbjd/dtyw/202001/t20200130_2016544.shtml.

12. Health Commission of Hubei Province, The 9th Press Conference on the Prevention and Control of COVID-19; http://wjw.hubei.gov.cn/fbjd/dtyw/202001/t20200131_2017018.shtml.

13. J. T. Wu, K. Leung, G. M. Leung, Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A

modelling study. *Lancet* **395**, 689–697 (2020). doi:10.1016/S0140-6736(20)30260-9 Medline

14. J. Riou, C. L. Althaus, Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. *Euro Surveill.* **25**, 2000058 (2020). doi:10.2807/1560-7917.ES.2020.25.4.2000058 Medline

15. N. Imai, I. Dorigatti, A. Cori, C. Donnelly, S. Riley, N. M. Ferguson, Report 2: Estimating the potential total number of novel Coronavirus cases in Wuhan City, China (2020); https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/news--wuhan-coronavirus/.

16. Baidu Migration; https://qianxi.baidu.com/ [accessed 26 February 2020].

17. M. Kramer, D. Pigott, B. Xu, S. Hill, B. Gutierrez, O. Pybus, Epidemiological data from the nCoV-2019 Outbreak: Early Descriptions from Publicly Available Data; http://virological.org/t/epidemiological-data-from-the-ncov-2019-outbreak-early-descriptions-from-publicly-available-data/337) [accessed 24 February 2020].

18. C. Fraser, S. Riley, R. M. Anderson, N. M. Ferguson, Factors that make an infectious disease outbreak controllable. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 6146–6151 (2004). doi:10.1073/pnas.0307506101 Medline

19. S. Pei, SenPei-CU/COVID-19: COVID-19, Version 1, Zenodo (2020); http://doi.org/10.5281/zenodo.3699624.

20. S. Pei, S. Kandula, W. Yang, J. Shaman, Forecasting the spatial transmission of influenza in the United States. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2752–2757 (2018). doi:10.1073/pnas.1708856115 Medline

21. A. Rambaut, Phylodynamic Analysis | 129 genomes | 24 Feb 2020; http://virological.org/t/phylodynamic-analysis-129-genomes-24-feb-2020/356/ [accessed 25 February 2020].

22. China National Health Commission, Policy and regulatory documents; http://www.nhc.gov.cn/wjw/gfxwjj/list.shtml [accessed 14 February 2020].

23. SenPei-CU/COVID-19, Data and code posting; https://github.com/SenPei-CU/COVID-19.

24. Field Briefing, Diamond Princess COVID-19 Cases, 20 Feb Update; https://www.niid.go.jp/niid/en/2019-ncov-e.html [accessed 25 February 2020].

25. Tencent Big Data Platform; https://heat.qq.com [accessed 14 February 2020].

26. D. Zhu, Z. Huang, L. Shi, L. Wu, Y. Liu, Inferring spatial interaction patterns from sequential snapshots of spatial distributions. *Int. J. Geogr. Inf. Sci.* **32**, 783–805 (2018). doi:10.1080/13658816.2017.1413192

27. D. He, E. L. Ionides, A. A. King, Plug-and-play inference for disease dynamics: Measles in large and small populations as a case study. *J. R. Soc. Interface* **7**, 271–283 (2010). doi:10.1098/rsif.2009.0151 Medline

28. M. S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Trans. Signal Process.* **50**, 174–188 (2002). doi:10.1109/78.978374

29. J. L. Anderson, An ensemble adjustment Kalman filter for data assimilation. *Mon. Weather Rev.* **129**, 2884–2903 (2001). doi:10.1175/1520-0493(2001)129<2884:AEAKFF>2.0.CO;2

30. C. Snyder, T. Bengtsson, P. Bickel, J. Anderson, Obstacles to high-dimensional particle filtering. *Mon. Weather Rev.* **136**, 4629–4640 (2008). doi:10.1175/2008MWR2529.1

31. W. Yang, A. Karspeck, J. Shaman, Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. *PLOS Comput. Biol.* **10**, e1003583 (2014). doi:10.1371/journal.pcbi.1003583 Medline

32. J. Shaman, W. Yang, S. Kandula, Inference and forecast of the current West African ebola outbreak in Guinea, sierra leone and liberia. *PLOS Curr.* **6**, 10.1371/currents.outbreaks.3408774290b1a0f2dd7cae877c8b8ff6 (2014). Medline

33. N. B. DeFelice, E. Little, S. R. Campbell, J. Shaman, Ensemble forecast of human West Nile virus cases and mosquito infection rates. *Nat. Commun.* **8**, 14592 (2017). doi:10.1038/ncomms14592 Medline

34. J. Reis, J. Shaman, Retrospective parameter estimation and forecast of respiratory syncytial virus in the United States. *PLOS Comput. Biol.* **12**, e1005133 (2016). doi:10.1371/journal.pcbi.1005133 Medline

35. O. Diekmann, J. A. P. Heesterbeek, J. A. Metz, On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations. *J. Math. Biol.* **28**, 365–382 (1990). doi:10.1007/BF00178324 Medline

36. O. Diekmann, J. A. P. Heesterbeek, M. G. Roberts, The construction of next-generation matrices for compartmental epidemic models. *J. R. Soc. Interface* **7**, 873–885 (2010). doi:10.1098/rsif.2009.0386 Medline

37. S. Lai, I. Bogoch, N. Ruktanonchai, A. Watts, Y. Li, J. Yu, X. Lv, W. Yang, H. Yu, K. Khan, Z. Li, Assessing spread risk of Wuhan novel coronavirus within and beyond China, January-April 2020: a travel network-based modelling study. medRxiv 2020.02.04.20020479 [Preprint]. 5 February 2020. https://doi.org/10.1101/2020.02.04.20020479.