THE UNIVERSITY OF CHICAGO

INFERRING THE EXTENT AND IMPACT OF HETEROGENEITY DURING EMERGING VIRUS OUTBREAKS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF ECOLOGY AND EVOLUTION

BY
RAHUL SUBRAMANIAN

CHICAGO, ILLINOIS
DECEMBER 2021

To my grandmothers, Kamala Ramamoorthy and Sumathi Seshadri

TABLE OF CONTENTS

LIST OF FIGURES vi

LIST OF TABLES xxii

ACKNOWLEDGMENTS xxiii

ABSTRACT xxx

1 INTRODUCTION 1

2 QUANTIFYING ASYMPTOMATIC INFECTION AND TRANSMISSION OF COVID-19 IN NEW YORK CITY USING OBSERVED CASES, SEROLOGY AND TESTING CAPACITY 9
- 2.1 Introduction 9
- 2.2 Results 13
- 2.3 Methods 28
- 2.4 Funding 34
- 2.5 Acknowledgments 34
- 2.6 Supporting Information 34
  - 2.6.1 SEPIAR Model Details 34
  - 2.6.2 Model Fitting Techniques 37
  - 2.6.3 Testing Model 41
  - 2.6.4 Testing Priorities 41
  - 2.6.5 Syndrome Surveillance Estimates 50
  - 2.6.6 Overall Reproductive Number Derivation 54
  - 2.6.7 Supplemental Figures 56

3 PREDICTING RE-EMERGENCE TIMES OF DENGUE EPIDEMICS AT LOW REPRODUCTIVE NUMBERS: DENV1 IN RIO DE JANEIRO, 1986-1990. . . . 65
- 3.1 Introduction 65
- 3.2 Results 68
  - 3.2.1 Replenishment of Susceptible Individuals is Insufficient to Explain Re-Emergence 73
- 3.3 Discussion 80
- 3.4 Methods 85
  - 3.4.1 Data Description 85
  - 3.4.2 Basic Model Formulation 85
  - 3.4.3 Fitting the stochastic model 88
  - 3.4.4 Stochastic Simulation 89
  - 3.4.5 Sensitivity Analysis 90
  - 3.4.6 Comparison with Vector Model and literature R_0 90
- 3.5 Author Contributions 90
- 3.6 Funding 90

3.7 Acknowledgments . . . 91
3.8 Supporting information . . . 91
3.8.1 Supplemental Methods: Analytical Skip Derivation . . . 91
3.8.2 Supplemental Methods: Stochastic Model Description . . . 95
3.8.3 Supplemental Results: Sensitivity Analysis . . . 101
3.8.4 Supplemental Results: Vector Model Considerations . . . 102
3.9 Supplemental Figures . . . 105

4 CONNECTIVITY STRUCTURE AND POPULATION SIZE SHAPE THE SPREAD OF A NEW DENGUE SEROTYPE ACROSS A METROPOLITAN AREA . . . 125
4.1 Introduction . . . 125
4.2 Results . . . 127
4.3 Discussion . . . 138
4.3.1 Summary of Results . . . 138
4.3.2 Relationship to Previous Work . . . 139
4.3.3 Model Caveats . . . 141
4.3.4 Implications . . . 144
4.4 Methods . . . 145
4.4.1 Data . . . 145
4.4.2 Model Description . . . 146
4.4.3 ODE Equations . . . 147
4.4.4 Panel Model Movement Equations . . . 151
4.4.5 Discretizations . . . 157
4.4.6 Compartment Transitions . . . 158
4.4.7 Measurement Model Equations . . . 159
4.4.8 Model Fitting Strategy . . . 159
4.5 Funding . . . 163
4.6 Acknowledgments . . . 163
4.7 Supporting information . . . 164
4.7.1 Supplemental Figures . . . 164
4.7.2 Fully Coupled Model Equations . . . 164
4.8 Details for Coupled Model Equations . . . 164

5 CONCLUSION . . . 178
5.1 Concluding remarks . . . 178
5.2 Future directions . . . 179

REFERENCES . . . 183

LIST OF FIGURES

2.1 **Model diagrams. (A) The full SEPIAR model used for inference.** The model is an extension of an SEIR formulation that considers both pre-symptomatic transmission (from compartment P) and asymptomatic transmission (from compartment A). **B) When the strength of pre-symptomatic transmission b_{p} is set to 0, the SEPIAR model reduces to the SEIAR model.** Since we assume that $\phi_U = \phi_E$, when $b_p = 0$ the infectious pre-symptomatic compartment behaves like an additional exposed compartment. **C) When the strength of asymptomatic transmission b_{a} is set to 0, the SEPIAR model reduces to the SEPIR model.** Individuals in the asymptomatic infectious compartment (A) make no contribution to the force of infection, so asymptomatic individuals essentially recover after leaving the pre-symptomatic period (P). In all three panels, circular/elliptical compartments contribute to the force of infection, while rectangular compartments do not. The green ellipse denotes the point at which severe/hospitalized COVID patients are sampled and enter the testing queue for severe cases, while the red ellipse denotes the corresponding entry point for the queue for non-severe symptomatic cases. 12

2.2 **The probability of symptomatic infection. (A) Simulated vs. observed cases from the profile of the asymptomatic transmission strength (b_{a}) using the SEPIAR model.** The red line is the median from 100 simulations using the Maximum-Likelihood Estimates (MLE), while the red shaded region denotes the 2.5 to 97.5% quantiles across 100 simulations from all parameter combinations within 2 log-likelihood units of the profile MLE. Likelihoods here are with respect to case data. The observed daily case counts are denoted by the blue line. **B) Model Likelihood as a function of the proportion of cases that are symptomatic (p_S) for each parameter combination from panel A.** The y-axis shows the likelihood for that parameter combination with respect to serology data. All parameter combinations above the blue line have likelihoods within 2-log-likelihood units of the MLE (defined with respect to serology). This corresponds to a range of values for p_S of approximately 13 to 18%. **C) Comparison of observed vs. simulated estimates of herd immunity in the population from parameter combinations supported by both case and antibody data (all points above the blue line in panel B).** The red line denotes the median value of herd immunity (the proportion of the population that has recovered ($\frac{R}{N}$) at that point in time in 100 simulations from the MLE parameter combination. The red shaded region denotes the 2.5 to 97.5% quantiles for these simulations from all parameter combinations within 2-log-likelihood units of the MLE with respect to serology (all parameter combinations above the blue line in panel B). The blue line denotes estimates of herd immunity from a recent serological survey in New York City [1]. The blue shading denotes 95% confidence intervals for those serology estimates using the methods of [1]. 13

2.3 **Plots of the reproductive number of symptomatic individuals (R_0) (A) and the overall reproductive number ($R_{0_{\mathrm{NGM}}}$) (B), as a function of the relative strength of pre-symptomatic transmission (b_{p}) and the relative strength of asymptomatic transmission (b_{a}).** Each point represents one parameter combination within 2 log-likelihood units of the MLE (with respect to serology) from the b_{a} profile. **C) Plot of the overall reproductive number vs the reproductive number in symptomatic individuals for the same points colored by b_{a}.** The black arrows show the direction of increasing strength of asymptomatic transmission (b_{a}) and pre-symptomatic transmission (b_{p}). For this same plot except colored by the strength of pre-symptomatic transmission (b_{p}), see Fig. 2.11. For ease of plotting, we exclude two parameter combinations which had a very low relative rates of pre-symptomatic transmission (i.e. b_{p} was lower than 0.020). The two outlier combinations had high reproductive numbers ($R_0 = 17.77, R_{0_{\mathrm{NGM}}} = 3.95$ and $R_0 = 4.97, R_{0_{\mathrm{NGM}}} = 4.37$). These outliers are included in the Fig. 2.12. 16

2.4 **The contribution to the force of infection at the peak of the outbreak on April 14, 2020 from symptomatic, asymptomatic, and pre-symptomatic infections under different relative asymptomatic transmission rates b_{a}.** For each parameter combination from the fitted SEPIAR model supported by case and serology data (corresponding to the points in Figure 2.3), we simulate 100 trajectories and calculate the proportion of the overall force of infection on April 14,2020 that is due to asymptomatic, symptomatic, and pre-symptomatic infections. We pool trajectories from all parameter combinations that have the same value of b_{a}, and calculate the median, 2.5%, and 97.5% quantiles for each infection class and value of b_{a}. The colored bars represent for each infection class, the median proportion of its contribution to the force of infection (and hence may not sum exactly to 1). The error bars represent the corresponding 2.5%, and 97.5% quantiles. Versions of this plot calculated respectively 4 weeks before, and 4 weeks after, the peak can be found in the SI Appendix Fig. S9. We excluded two outlier parameter combinations that had extremely low relative rates of pre-symptomatic transmission (i.e. where b_{p} was less than 0.02). 17

2.5 **Comparison of daily COVID hospitalizations under the model with observed COVID hospitalizations in New York City and emergency department respiratory syndrome surveillance reports.** The red line represents the median daily case hospitalizations from 100 simulations from the parameter combination with the highest likelihood with respect to serology from the b_{a} profile. The red shading represents the bounds of the 2.5% and 97.5% quantiles across all parameter combinations from the b_{a} profile that are supported by case and serology data. The blue line shows observed COVID daily hospitalizations in New York City. The yellow line denotes daily reports of respiratory illness from syndrome surveillance in New York City emergency departments, while the pink line denotes anomalous respiratory surveillance reports compared to previous years. 19

2.6 **Diagram of the grid searches for the SEPIR (A) and SEIAR (B) models.** 38

2.7 **SEPIAR Profile Fitting Procedure**. This diagram summarizes how the Monte Carlo profile of b_a for the SEPIAR model was fit to case data and subsequently to serology. 40
2.8 **Diagram of the general testing framework described in Section 2.6.4**. The New York City-specific modifications described in Section 2.6.4 are not shown here. 45
2.9 **Diagram of the testing priorities described in Section 2.6.4.** 51
2.10 **Monte Carlo profile of the strength of transmission in asymptomatic cases relative to that of symptomatic cases (b_a).** Each point represents the parameter combination from the Monte Carlo profile for b_a with the highest log-likelihood (with respect to observed cases) for a given value of b_a. All points above the blue line are supported by the case data (i.e. they have likelihoods within 2 log likelihood units of the profile MLE). 57
2.11 **Additional plots of the overall reproductive number ($R_{0_{NGM}}$) vs the reproductive number in symptomatic individuals (R_0) from parameter combinations supported by case and serology data from the full SEPIAR (A) and SEIAR (B) models. A)** Each point represents one parameter combination within 2 log-likelihood units of the MLE (with respect to cases and serology) from the b_a profile using the full SEPIAR model. Each point is colored by the strength of pre-symptomatic transmission (b_p). **B)** Each point represents one parameter combination within 2 log-likelihood units of the MLE (with respect to serology) from the grid search of the SEIAR model (no pre-symptomatic transmission). Points are colored by the relative strength of asymptomatic transmission (b_a). For ease of plotting, in panel A) we exclude two parameter combination with very low pre-symptomatic transmission rates b_p. These outliers are also excluded from Figure 3 in the main manuscript , but are shown in Supporting Figure S2.12. 58
2.12 **Additional plots of the overall reproductive number ($R_{0_{NGM}}$) vs the reproductive number in symptomatic individuals (R_0) including the outlier parameter combination colored by the relative strength of asymptomatic transmission (b_a) (A) or the relative strength of pre-symptomatic transmission (b_p) (B).** Each point represents one parameter combination within 2 log-likelihood units of the MLE (with respect to cases and serology) from the b_a profile of the full SEPIAR model. We include here the outlier parameter combinations with very high symptomatic R_0 greater than 15 that were excluded from Figure 3 and Figure S2.11. 59
2.13 **Monte Carlo profile of the probability that an individual (who does not have COVID-19) who shows up to the emergency department with respiratory symptoms is severe enough to merit testing for COVID-19 (s).** Each point represents the parameter combination from the Monte Carlo profile for the scaling parameter s with the highest log-likelihood (with respect to observed cases) for a given value of s. All points above the blue line have likelihoods within 2-log likelihood units of the profile MLE). 60

2.14 **Contributions from pre-symptomatic, symptomatic, and asymptomatic infections to the overall force of infection 4 weeks before (A) and after the peak in reported cases (B).** The calculation of the contribution to the overall force of infection from simulated trajectories of parameter combinations from the SEPIAR model supported by case and serology data as described in Figure 4 of the main manuscript was replicated using time points 4 weeks before and after the peak of reported cases on April 14, 2020 instead of at the time of the peak. As in Figure, two parameter combinations with rates of pre-symptomatic transmission b_p below 0.02 were excluded from the plot. 61

2.15 **Plot of observed respiratory syndrome surveillance reports compared to simulations from fitted statistical model** The red line corresponds to weekly respiratory infections from syndrome surveillance reports in NYC hospitals in 2016, 2017 and 2019 that were used to fit the statistical model in Section 2.6.5. The blue line represents the median estimate for the number of expected syndrome surveillance reports ($G(w, y)$ for that week and year from 100 simulations from the fitted statistical model. The shaded light blue region represents the 2.5% and 97.5% quantiles from those 100 simulations. 62

2.16 **Validation analysis results from fitting model to simulated data. We are particularly interested here in verifying the ability of the inference pipeline to estimate the value of the probability of symptomatic infection p_{S} used in the simulations. A) Observed data and simulated trajectories used for fitting**. The green points denote observed daily reported case counts in New York City. The red points denote daily reported cases from a representative simulated trajectory from a "low" b_{a} parameter combination ($b_{\mathrm{a}} = 0.07$, $R_0 = 6.10$, $b_{\mathrm{q}} = 0.23$, $b_{\mathrm{p}} = 0.94$, $p_{\mathrm{S}} = 0.15$, $p_{\mathrm{H}} = 0.16$, $\gamma = 6.33$,$E_0 = 63566.34$ $z_0 = 13443$, and the overall reproductive number $R_{0_{\mathrm{NGM}}} = 2.27$), while the blue points denote daily reported cases for a representative trajectory from a "high" b_{a} parameter combination ($b_{\mathrm{a}} = 0.97$, $R_0 = 3.08$, $b_{\mathrm{q}} = 0.16$, $b_{\mathrm{p}} = 0.99$, $p_{\mathrm{S}} = 0.15$, $p_{\mathrm{H}} = 0.17$, $\gamma = 11.73$, $E_0 = 54806$, $z_0 = 11625$, and $R_{0_{\mathrm{NGM}}} = 3.50$). **B) Supported parameter ranges for the proportion of cases that are symptomatic (p_{S} for fits to simulated and observed data**. Red and blue dots represent respectively parameter combinations supported by the case data when the model was fit to the low b_{a} (red) or high b_{a} (blue) trajectories. For comparison, the green dots represent parameter combinations supported by the case when the model was fit to observed case data (as shown in panel B of Figure 2). All parameter combinations above the green line have likelihoods within 2-log-likelihood units of the MLE defined with respect to serology. Our approach recovers the value of p_{S} used in the simulations, and does so accurately. 63

2.17 **Comparison of infection fatality ratios (IFR) estimated from fitted model parameters under different testing strategies**. The red shaded region denotes a histogram of the infection fatality ratio calculated with respect to all cases, both symptomatic and asymptomatic. The IFR was calculated for each parameter combination from the SEPIAR model that was supported by both case and serology data. The proportion of hospitalized cases that result in deaths was estimated from observed confirmed COVID-19 hospitalisations and deaths in New York City during time period of the study. The red histogram shows the range of IFR values expected under the SEPIAR model if all cases (symptomatic and asymptomatic) are observed. Each count in the histogram represents the expected IFR for one parameter combination that is supported by the case and serology data. The blue histogram shows the expected IFR if all symptomatic cases are observed. The higher IFR obtained in the blue histogram compared to the red histogram demonstrates how different testing strategies can alter the IFR. The orange line denotes the observed IFR calculated by dividing the total number of confirmed COVID-19 deaths in NYC during the study period by the total number of confirmed cases. The gap between the orange line and the blue histogram illustrates how limited testing capacity can affect the IFR that is estimated, since not all symptomatic cases were tested due to limited testing capacity early in the outbreak. 64

3.1 **A) Graphical illustration of how the expected number of skips (n_c) is calculated.** The black dots represent the threshold fraction of the population susceptible at the time of prediction required for n skips to occur ($s_c(n)$). The plot shows ($s_c(n)$) as a function of n (the number of skips) obtained from Equation 1 with seasonality amplitude $\delta = 0.2$ (contacts per person per day) and reproductive number $R_0 = 1.4$. In this example, the red line represents the fraction of the population susceptible at the time of prediction (s_0). If s_0 is smaller than $s_c(n)$, at least n skips will occur. To find the expected number of skips (n_c), we identify the largest number of skips n such that s_0 is smaller than the susceptibility threshold required for those skips $s_c(n)$. In this example, the red line intersects the $s_c(n)$ curve between $s_c(n = 6)$ and $s_c(n = 7)$. Therefore, a critical skip number of $n_c = 6$ is obtained. **B) and C) The critical skip value n_c as a function of** R_0 for (B) different values of the amplitude of seasonal transmission δ with $s_0 = 0.7$ and (C) different values of the fraction of the population susceptible at the time of prediction (s_0) with $\delta = 0.70$. In all three panels, the frequency of transmission ω the population turnover rate μ and population growth rate r are fixed at respective values $\omega = \frac{2\pi}{365}\text{days}^{-1}$ corresponding to an annual periodicity, $\omega = 1/(74.46 * 365))\text{days}^{-1}$ corresponding to an average lifespan of 75 years, and $r = 1.55\mu\text{days}^{-1}$ consistent with the growth of the city of Rio de Janeiro. These values were chosen for the purpose of illustration, based on the inverse of the average life expectancy in Brazil in 2012 according to the 2010 census [2] , and the interpolation of population estimates for the resident population of the municipality of Rio de Janeiro from the 1991 [3] and 2000 [4] censuses assuming exponential growth. 71

3.2 **(A) Observed dengue case data.** Monthly reported dengue cases in the city of Rio de Janeiro, Brazil from April 1986-1995. The grey shaded region denotes observations that were included in the fitting of the stochastic model from May 1, 1986 to July 1, 1988 inclusive. Serotype DENV1 re-emerged in 1990. DENV2 was first detected in the state of Rio de Janeiro in 1990 but did not become dominant until 1991 [5, 6]. Both co-circulated afterwards. We focus on the invasion of DENV1 from 1986-1987 and its initial re-emergence in DENV1 in 1990 using a single serotype transmission model. This allows us to evaluate this transmission model in a region where only one serotype was circulating, where cross-immunity could not easily be invoked to explain the absence or reduction of dengue in a given year. **(B) Deterministic critical number of DENV1 skips n_c for Rio de Janeiro from September 1988.** Expected number of skips n_c with amplitude of seasonal transmission $\delta = 0.7$ and the fraction of the population susceptible after the first DENV1 invasion as of September 1, 1987 (s_0) calculated from the data (A). We use a reporting rate ρ of 3% when calculating s_0 , consistent with serological estimates from the literature [7]. For comparison purposes, we also include the expected number of skips n_c assuming a reporting rate of 10%. . . . 73

3.3 **A-C) Selected parameter profiles for the stochastic model.** Profiles of the mean annual transmission rate β_0 (A), seasonal transmission amplitude δ (B), and reporting rate ρ (C). The red curve is a polynomial fit to the subset of the profile points shown on the figure. The single dashed grey horizontal line represents the likelihood value 2 log likelihood units below the maximum likelihood estimate. This line provides an estimate of confidence intervals for the given parameter. The grey vertical line denotes the parameter value of the maximum likelihood estimate. The maximum likelihood estimate for the reporting rate in panel C is very close to the literature value obtained from serology (approximately 3 percent). [7] 74

3.4 **A-B) Comparison of simulated values with the fitted model and observed data on a log (A) and regular (B) scale.** Observed monthly cases from April 1986 to June 1988 are shown in blue. Median values from 100 simulations with the maximum likelihood parameter combination are shown in red. The shaded red region denotes the 2.5% and 97.5%th quantile boundaries from those simulations. **C) Estimates for $\mathbf{R_0(t)}$.** The black line denotes the trajectory of R_0(t) for the maximum likelihood estimate. The shaded grey region represents the 2.5% and 97.5%th quantile boundaries for trajectories from all parameter combinations within 2 log likelihood units of the maximum likelihood estimate. Each parameter combination has only one seasonal trajectory for $R_0(t)$ since $R_0(t)$ is a deterministic quantity. $R_0(t)$ for all parameter estimates ranges from 1.79-2.09 in the on season to 0.31-0.52 in the off-season. 76

3.5 **A) Expected number of skips (n_c) calculated using parameters obtained from the fitted stochastic model.** The open circles show the expected number of skips n_c from Equation 1 using parameters and the fraction of the population susceptible after the initial DENV1 invasion (s_0) estimated from the fitted stochastic model. Each circle corresponds to one parameter combination, and we included here all parameter combinations for the fitted model with a seasonal transmission amplitude (δ of 0.7 (contacts per person per day) and a likelihood value within two log-likelihood units of the maximum likelihood estimate (MLE). See Fig. 3.20 for expected skips from parameter combinations with different values of δ and Figure S10 for parameter combinations from the profile of the recovery rate, γ. For comparison purposes, the black line shows the expected number of skips for the deterministic skip calculation from panel B of Fig. 3.2 with the reporting rate ρ fixed at the literature value of 3%. **B) Probability of epidemic in 1990 under forward stochastic simulation of fitted model.** The fitted stochastic model was simulated forward in time from 1986-1990 with population growth. A pulse of 20 infected individuals were assumed to arrive each day in January 1990. Each parameter combination within 2 log likelihood units of the maximum likelihood estimate was simulated 100 times. The re-emergence probability was calculated by determining the number of simulations in which the susceptible population decreased in 1990. The plot shows re-emergence probability as a function of the process noise intensity σ_{P}. Each point represents a single parameter combination. The maximum likelihood estimate parameter combination is circled in red. 78

3.6 **Observed vs simulated cases from parameter combination for the stochastic SIR Cosine Model, fit to 2.5 years of DENV1 case data (fixed recovery rate)** Log of observed monthly cases from April 1986 to June 1988 are shown in blue. Simulated cases were estimated from 100 simulations for each parameter combination within 2 log-likelihood units of the highest likelihood parameter combination (the MLE). Median values from 100 simulations from the MLE the are shown in red. The range of simulation medians across all parameter combinations within 2 log-likelihood units is shaded pink. The shaded blue region denotes the 95% quantile boundaries across all 100 simulations from the MLE. The shaded grey region denotes 95% quantile boundaries from all simulations (across all parameter combinations within 2 log-likelihood units of the MLE). . . 106

3.7 **A-B Comparison of simulated values with the fitted model and observed data on a log (A) and regular (B) scale.)** Observed monthly cases from April 1986 to June 1988 are shown in blue. Median values from 100 simulations with the maximum likelihood parameter combination are shown in red. The shaded red region denotes the 2.5% and 97.5%th quantile boundaries from those simulations. **C) Estimates for R_0(t).** The black line denotes the trajectory of R_0(t) for the maximum likelihood estimate. The shaded grey region represents the 2.5% and 97.5%th quantile boundaries for trajectories from all parameter combinations within 2 log likelihood units of the maximum likelihood estimate. Each parameter combination has only one seasonal trajectory for R_0(t) since R_0(t) is a deterministic quantity. Results for three models are shown: the SIR Cosine Model (described in the main manuscript) as well as an SIR Spline and SEIR Spline Model. 107

3.8 **Comparison of all examined parameter combinations within 2 log likelihood units of the maximum likelihood estimate for each fitted model with observed dengue case counts for each of 3 models.** For each parameter combination, 100 independent stochastic simulations were conducted. The right hand panel shows cases on a standard scale, while the left hand panel is on a log scale. Log of observed monthly cases from April 1986 to June 1988 are shown in blue. Simulated cases were estimated from 100 simulations for each parameter combination within 2 log-likelihood units of the highest likelihood parameter combination (the MLE). Median values from 100 simulations from the MLE the are shown in red. The range of simulation medians across all parameter combinations within 2 log-likelihood units is shaded pink. The shaded blue region denotes the 95% quantile boundaries across all 100 simulations from the MLE. The shaded grey region denotes 95% quantile boundaries from all simulations (across all parameter combinations within 2 log-likelihood units of the MLE). . 108

3.9 **Profiles of reporting rate (ρ) for all three models.** The red curve is a polynomial fit to the subset of the profile points shown on the figure. The single dashed grey horizontal line represents the likelihood value 2 log likelihood units below the peak of the profile. This line provides an estimate of confidence intervals for the given parameter. The grey vertical line denotes the parameter value of the maximum likelihood estimate across all parameter profiles for that model. 109

3.10 **Profiles of the environmental process noise magnitude parameter ($\sigma_{\mathbf{P}}$) for all three models.** The red curve is a polynomial fit to the subset of the profile points shown on the figure. The single dashed grey horizontal line represents the likelihood value 2 log likelihood units below the peak of the profile. This line provides an estimate of confidence intervals for the given parameter. The grey vertical line denotes the parameter value of the maximum likelihood estimate across all parameter profiles for that model. 110

3.11 **Profiles of the measurement noise magnitude parameter ($\sigma_{\mathbf{M}}$) for all three models.** The red curve is a polynomial fit to the subset of the profile points shown on the figure. The single dashed grey horizontal line represents the likelihood value 2 log likelihood units below the peak of the profile. This line provides an estimate of confidence intervals for the given parameter. The grey vertical line denotes the parameter value of the maximum likelihood estimate across all parameter profiles for that model. 111
3.12 **Profiles of the initial number of infected people at the start of the simulation (I_0) for all three models.** The red curve is a polynomial fit to the subset of the profile points shown on the figure. The single dashed grey horizontal line represents the likelihood value 2 log likelihood units below the peak of the profile. This line provides an estimate of confidence intervals for the given parameter. The grey vertical line denotes the parameter value of the maximum likelihood estimate across all parameter profiles for that model. . . . 112
3.13 **Profiles of the phase parameter (ϕ) for the SIR Cosine Model.** The red curve is a polynomial fit to the subset of the profile points shown on the figure. The single dashed grey horizontal line represents the likelihood value 2 log likelihood units below the peak of the profile. This line provides an estimate of confidence intervals for the given parameter. The grey vertical line denotes the parameter value of the maximum likelihood estimate across all parameter profiles for that model. 113
3.14 **Profiles of the recovery rate (γ) for the SIR Cosine Model.** The red curve is a polynomial fit to the subset of the profile points shown on the figure. The single dashed grey horizontal line represents the likelihood value 2 log likelihood units below the peak of the profile. This line provides an estimate of confidence intervals for the given parameter. The grey vertical line denotes the value of gamma that was used for the original fit of the SIR Cosine Model (where the recovery rate was fixed) 114
3.15 **Expected number of skips from deterministic calculation given all combinations of R_0, seasonal transmission amplitude (δ), and reporting rate (ρ) from each parameter combination of the recovery rate (γ) profile within 2 log likelihood units of the parameter combination with the highest likelihood in the original model.** In Figure 5 panel A, expected numbers of deterministic skips were calculated for all parameter combinations within 2 log-likelihood units from the maximum likelihood estimate of the SIR Cosine Model. Parameter values from the profile of the recovery rate (γ) in the sensitivity analysis shown in Supplemental Figure 3.15 were not included in the skip calculations for Figure 5 Panel A. Here, we replicate those skip calculations for all parameter combinations from the γ sensitivity analysis within 2 log-likelihood units of the original maximum likelihood estimate. 115

3.16 **Comparison of reporting rate (rho) vs. reproductive number (R_0) and environmental process noise magnitude $\sigma_{\mathbf{P}}$) for each parameter combination of the γ profile of the SIR Cosine Model within 2 log-likelihood units of the parameter combination with the highest likelihood in the original model.** This figure compares the values of R_0 and reporting rate (ρ) and process noise magntiude (σ_{P}) for all parameter combinations from the gamma profile used for the deterministic skip calculations shown in Supplemental Figure 3.15. 116

3.17 **Comparison of filter means between maximum likelihood parameter combination and parameter combination with large process noise over third year of simulation.** Average of filter means for the number of monthly cases (C) at each observed data point from 10 runs of the Sequential Monte Carlo algorithm pfilter run at MLE parameter combination (in green) and at the parameter combination with the highest likelihood out of all parameter combination with the highest permissible amount of process noise ($\sigma_{\mathrm{P}} = 1$) shown in red. The observed cases are shown in blue for comparison. Only the third year of the fit (corresponding to the period from January 1988 through July 1988) is shown. Shaded ribbons show the average of the filter mean +/- 2 times the standard deviation of the filter means (across all 10 runs). 117

3.18 **Comparison of filter means between maximum likelihood parameter combination and parameter combination with large process noise over second year of simulation.** Average of filter means for the number of monthly cases (C) at each observed data point from 10 runs of the Sequential Monte Carlo algorithm pfilter run at MLE parameter combination (in green) and at the parameter combination with the highest likelihood out of all parameter combination with the highest permissible amount of process noise ($\sigma_{\mathrm{P}} = 1$) shown in red. The observed cases are shown in blue for comparison. Only the second year of the fit (corresponding to the period from January 1987 through December 1987) is shown. Shaded ribbons show the average of the filter mean +/- 2 times the standard deviation of the filter means (across all 10 runs). 118

3.19 **Probability of stochastic epidemic in 1990 vs Process Noise Intensity ($\sigma_{\mathbf{P}}$) under simulation from top 2LL parameter combinations of stochastic SIR Cosine Model.** The fitted stochastic model was simulated forward in time from 1986-1990 with population growth. Daily pulse rates of 2,5,10,20, 50, and 100 infected individuals per day in January 1990 were used. Each parameter combination within 2 log-likelihood units of the maximum likelihood estimate was simulated 100 times. The re-emergence probability was calculated by determining the number of simulations in which the susceptible population decreased in 1990. The plot shows re-emergence probability as a function of the process noise intensity σ_{P}. Each point represents a single parameter combination at a particular pulse rate. Points are colored by pulse rate. MLE parameter combination points are circled in red. 119

3.20 **Expected number of skips from deterministic calculation using parameter estimates from the fitted stochastic model.** The red crosses show the expected number of skips nc from Equation 1 of the main text using parameters and the fraction of the population susceptible after the initial DENV1 invasion (s_0) estimated from the fitted stochastic model. Each circle corresponds to one parameter combination, and we included here all parameter combinations for the fitted SIR Cosine model with different values of the reproductive number R_0, seasonal transmission amplitude δ, and reporting rate ρ. See Supplemental Figure 3.15 for the expected number of skips for all parameter combinations obtained from the profile of the recovery rate (γ). 120

3.21 **Different combinations of mean transmission rate β_0 and recovery rate γ that yield the same reproductive number R_0 value have different values of the fraction of the population susceptible ($\frac{S}{N}$) after 1 year.** For example, suppose that we have the parameter combinations $\delta = 0.5$, $\beta_0 = 0.3$, $\gamma = 0.2$ and $\delta = 0.5$, $\beta_0 = 0.15$, $\gamma = 0.1$. While both parameter combinations give an R_0 of 1.5, the first yield an $\frac{S}{N}$ after one outbreak of around 20%, while the second gives an $\frac{S}{N}$ of approximately 40%. The plot shows values of $\frac{S}{N}$ as a function of γ with β_0 modified to give different values of R_0. For all points in the plot, the amplitude of seasonal transmission $\delta = 0.5$, the initial number of infected individuals $I(t = 0) = 1$, and the frequency of the seasonality of transmission $\omega = \frac{2\pi}{365}$, corresponding to an annual periodicity. 121

3.22 Transmission rate considered by Stone et al [8] in (black), and in this work (red). 122

3.23 **A boxplot of R_0 for each month in Rio de Janeiro from 2010-2016 with monthly estimates from [9] .** The superimposed solid red line indicates the mean monthly R_0 obtained from our stochastic SIR cosine model with the parameter combination with the highest likelihood. 123

3.24 **A) Weekly temperature of the city of Rio de Janeiro.** The data is represented by black dots and cosine function by the solid red line. **B) Transmission rate re-scaled between 0 and 1.** The black line is the fitted sinusoidal transmission rate and the blue line the is the effective transmission rate shown on Eq.(3.40). 124

4.1 **Heatmaps of observed dengue cases and temperature anomalies for 20 municipalities surrounding the city of Rio. A) Heatmap of weekly observed dengue cases on a log scale from January 2012 thru June 2013.** Grey shaded values denote dates with zero observed cases. Municipalities are arranged in increasing order of population. **B) Heatmap of temperature anomaly for each municipality and date. We obtained mean daily temperature estimates for each municipality using ERA5 reanalysis data. Temperature measurements were smoothed using a 2-week moving average. For each date, we calculated the mean temperature across all municipalities for the smoothed temperature time series. The daily anomalies are shown here.** 128

4.2 **Plot of the log ratio of total cases during the peak epidemic seasons (January-June) in 2013 vs 2012 for each of the 20 municipalities as a function of their A) log total daily flux to the city of Rio de Janeiro and B) log average population density excluding areas with less than 50 people per square kilometer.** We can see here that cities which have larger flues to the city of Rio tend to have smaller peak ratios (more cases in the first year of the epidemic compared to the second year. Points are colored by their For cities with similar total flux to Rio (such as Queimados, Mesquita, and Nilopolis) , higher population density appears to be associated with a lower peak ratio. **B) Map of the 20 cities surrounding Rio used to fit the panel model.** Each municipality is shaded by the proportion of outbound flux from that municipality to all other municipalities in the state of Rio that has a destination within the 20-city region. The proportion of outbound flux is at least 80% in all municipalities, indicating that the selected region is relatively self-contained. 130

4.3 **Plots of observed and simulated cases. A) Plot of observed and simulated time series on a log scale.** The blue time series denotes the observed weekly cases for each city. The dark red line represents the median value for 100 simulations from the parameter combination with the highest likelihood from the grid search (the MLE). Red shading denotes the bounds for the 2.5% and 97.5% quantiles from the simulations. **B) Plot of total observed cases vs total simulated cases on a log scale from 100 simulations of the MLE.** We aggregate cases in each simulation trajectory across days 1-200 (first peak), 200-400 (trough) and 400-600 (second peak). The filled circles represent the median value of the total simulated cases across all 100 trajectories within each epidemic stage and municipality, while the error bars denote the 2.5% and 97.5% quantiles. The black line is the 1:1 line. The points are colored according to the proportion of flux in each municipality to or from Rio. This quantity is obtained by adding the total flux between that municipality and Rio (in both directions) and dividing by the total inbound and outbound flux from that municipality to other municipalities within the region (including Rio). For a version of this plot using the parameter combination with the second-highest log-likelihood, which was the only other parameter combination from the grid search within 2 log-likelihood units of the MLE, see Fig. 4.8. 131

4.4 **Coupling to Rio de Janeiro has a substantial impact on dengue dynamics. Contribution to the force of infection in Itaborai and Nilopolis from commuters who work in Rio de Janeiro during the day.** The force of infection was calculated for 100 trajectories from the MLE for both municipalities, and contributions from Rio were averaged across all trajectories for each month and municipality. Movement from Rio makes a substantial contribution to the force of infection in Nilopolis, which is located just north of Rio de Janeiro and has a large proportion of flux to Rio. Movement from Rio makes a smaller contribution to the force of infection in Itaborai, which is located at the eastern edge of the region, and has a low proportion of flux to Rio. Itaborai does have substantial commuter traffic to some of the suburbs of Rio such as Niteroi which is not captured in the panel model. 133

4.5 **Peak ratio plots as a function of total flux from each municipality to Rio de Janeiro colored by population density. Peak ratios were calculated using A) observed cases B) 100 simulations from the MLE and C) 100 simulations from the only other parameter combination with 2 log-likelihood units of the MLE.** For simulated peak ratios, filled points represent the median peak ratio across all trajectories. The median values of the simulations can somewhat capture the general trend of cities with higher flux having lower peak ratios (i.e. more cases in the first year). However, the simulations are generally unable to capture the differences in peak ratios between cities with the same commuter traffic to Rio but different population densities. 134

4.6 **Error Analysis for panel model. A) Map of log mean-squared error for normalized cases during second peak.** For each of 100 simulation trajectories from the MLE, we divided the simulated cases by the population of each municipality and calculated the squared error for this normalized value with respect to the observed cases divided by the population at the same time and city. We then calculated the average mean squared error across all 100 trajectories and time points within the interval of the second peak (days 400 thru 600 of the simulation). Note that many of the municipalities east of the city of Rio have higher MSE during the second peak. **B) Map of epidemic intensity in 2013.** We calculate the Shannon entropy of epidemics in each municipality in 2013 using observed case data, following the approach of Dalziel et al [10], excluding observations with zero cases. Cities with higher Shannon entropy values in 2013 had more intense epidemics, with a larger proportion of the total observed cases in 2013 concentrated within a short time interval. **C) Proportion of flux to and from Rio.** See Figure 3 for an explanation of how this quantity is calculated. Note that the municipalities east of Rio have a substantial proportion of intra-peri-urban fluxes that do not originate or end in Rio. **D) Plot of log MSE in second peak as a function of the proportion of flux to Rio**. Municipalities are colored by their epidemic intensity in 2013. Note that the municipalities east of Rio tend to have high MSE and a low proportion of flux to Rio, while municipalities with a high proportion of Rio flux can have a low or high MSE. . 135

4.7 **Slice of intra-peri-urban flux parameter**. Each black dot represents the log of the mean likelihood from 10 repetitions of the block particle filter for the fully coupled model parameterized using the same values as the MLE from the gird search of the panel model with the addition of the parameter κ_{suburb}, which represents the proportion of the workday spent at the work location for movement fluxes that do not start or end in the city of Rio. The size of this parameter is a proxy for the importance of non-Rio fluxes. The blue line is equal to two log-likelihood units below the peak of the slice. The non-zero location of the slice peak may indicate that non-Rio fluxes may play a role in the dynamics. However, since the panel and coupled models are not fully nested and this slice analysis is not a true profile, we cannot make this assumption. For a version of the slice that uses the parameter combination with the second highest log-likelihood from the panel model grid search instead of the MLE, see Fig. 4.10. 136

4.8 **Plots of observed and simulated cases. A) Plot of observed and simulated time series on a log scale.** The blue time series denotes the observed weekly cases for each city. The dark red line represents the median value for 100 simulations from the parameter combination with the second-highest likelihood from the grid search. Red shading denotes the bounds for the 2.5% and 97.5% quantiles from those simulations. **B) Plot of total observed cases vs total simulated cases on a log scale from 100 simulations of the parameter combination with the second highest log-likelihood.** We aggregate cases in each simulation trajectory across days 1-200 (first peak), 200-400 (trough) and 400-600 (second peak). The filled circle represent the median value of the total simulated cases across all 100 trajectories within each epidemic stage and municipality, while the error bars denote the 2.5% and 97.5% quantiles. The black line is a reference line along which total simulated cases are equal to the total observed cases. The points are colored according to the proportion of flux in each municipality to or from Rio. This quantity is obtained by adding the total flux between that municipality and Rio (in both directions) and dividing by the total inbound and outbound flux from that municipality to other municipalities within the region (including Rio). 165

4.9 **A) Map of log mean-squared error for normalized cases during second peak.** Here we use the parameter combination with the second-highest log-likelihood instead of the MLE. **B) Plot of log MSE in second peak from second-highest log-likelihood parameter combination as a function of the proportion of flux to Rio**. 166

4.10 **Slice of intra-peri-urban flux parameter**. Each black dot represents the mean likelihood from 10 repetitions of the block particle filter for the fully coupled model parameterized using the same values as the parameter combination with the second highest log-likelihood from the gird search of the panel model with the addition of the parameter κ_{suburb}, which represents the proportion of the workday spent at the work location for movement fluxes that do not start or end in the city of Rio. The size of this parameter is a proxy for the importance of non-Rio fluxes. The blue line is equal to two log-likelihood units below the peak of the slice. The fact that the slice peaks at a non-zero value indicates that non-Rio fluxes may play a role in the dynamics. However, since the panel and coupled models are not fully nested and this slice analysis is not a true profile, we cannot make this assumption. 167

4.11 **Small sigma P MLE**. Plot of observed vs simulated cases and MSE vs proportion of Rio flux for parameter combination that had the highest log-likelihood out of all parameter combinations with low environmental noise. This log-likelihood is several hundreds units lower than the overall MLE. 168

4.12 **Contribution to the force of infection in Itaborai and Nilopolis from commuters who work in Rio de Janeiro during the day.** The force of infection was calculated for 100 trajectories from the MLE for both municipalities, and contributions from Rio were averaged across all trajectories for each month and municipality. Movement from Rio makes a substantial contribution to the force of infection in Nilopolis, which is located just north of Rio de Janeiro and has a large proportion of flux to Rio. Movement from Rio makes a smaller contribution to the force of infection in Itaborai, which is located at the eastern edge of the region and has a low proportion of flux to Rio. Itaborai does have substantial commuter traffic to some of the suburbs of Rio such as Niteroi which is not captured in the panel model. 169
4.13 **Heatmap of the deterministic reproductive number R_0 from the MLE for each date and municipality assuming a fully closed SEIR model with no movement from Rio.** To estimate this quantity, we use the temperature-dependent transmission rate from the panel model at the MLE parameter values at each observation date for each location, ignoring environmental stochasticity. 170
4.14 **Heatmap of the deterministic reproductive number R_0 from the parameter combination with the second-highest log-likelihood from the grid search for each date and municipality assuming a fully closed SEIR model with no movement from Rio.** To estimate this quantity, we use the temperature-dependent transmission rate from the panel model at the second-highest loglikelihood parameter values at each observation date for each location, ignoring environmental stochasticity. 171
4.15 **Initial value parameters (the initial infected population I_0) for the parameter combination with the highest log-likelihood from the panel grid search (the MLE) on a A) regular and B) log scale.** 172
4.16 **Initial value parameters (the initial infected population I_0) for the parameter combination with the second-highest log-likelihood from the panel grid search on a A) regular and B) log scale.** 173
4.17 **Initial value parameters (the initial exposed population E_0) for the parameter combination with highest log-likelihood from the panel grid search (the MLE) on a A) regular and B) log scale.** 174
4.18 **Initial value parameters (the initial exposed population E_0) for the parameter combination with the second-highest log-likelihood from the panel grid search on a A) regular and B) log scale.** 175
4.19 **Maps of unit-specific parameters (reporting rate ρ and mosquito to human population ratio k) from the MLE parameter combination.** Note that the value of k shown here is the parameter value multiplied by 5, as is the case within the process model. This re-scaling in conjunction with a logit parameter transformation ensures that the mosquito to human population ratio is bounded between 0 and 5. 176

4.20 **Maps of unit-specific parameters (reporting rate ρ and mosquito to human population ratio k) from the panel grid search parameter combination with the second-highest log-likelihood.** Note that the value of k shown here is the parameter value multiplied by 5, as is the case within the process model. This re-scaling in conjunction with a logit parameter transformation ensures that the mosquito to human population ratio is bounded between 0 and 5. 177

LIST OF TABLES

2.1 Comparmtents in the SEPIAR epidemiological model 35
2.2 Comparison of MLE Likelihoods from SEPIR, SEIAR, and SEPIAR models with respect to serology data 40
2.3 Parameter values used for the non-COVID-19 severe cases estimate (rounded). . 54

4.1 Comparison of Likelihoods from versions of panel model with and without movement from Rio de Janeiro. The likelihood for the version with movement is obtained from the MLE of the grid search, while the likelihood for the version with no movement from Rio is obtained from the profile of the coupling paramaeter κ, specifically the parameter combination with the highest likelihood at which $\kappa = 0$. The version without movement is fully nested within the original version with movement. 132

ACKNOWLEDGMENTS

There are many people whom I would like to thank who have helped me tremendously during my time in grad school, and they are far too many people to list here. Secondly, for many of these individuals, no words that I can write here will be sufficient to capture how important their support as been for me. With those two caveats in mind, there are several individuals in particular that I would like to acknowledge.

First and foremost, I would like to acknowledge my thesis advisor, Mercedes Pascual. None of the opportunities that I've enjoyed at UChicago would have been possible without her care and support. Mercedes has been so incredibly kind, patient, and supportive, and has been an extraordinary mentor and role model for me, whether as a scientist, advisor, or human being. Graduate school can often feel like a very exciting roller coaster, and having her guidance and support at every step along the way has made all the difference. She has this incredible ability to make people feel like everything will be all right, and I always leave her office happier than when I entered. I have always felt like I could be myself, and sometimes I think that she knows me better than I do! I feel truly lucky to have had the chance to work with her, and will treasure every minute of my time in the lab. While I had always imagined that graduate school would be fun, I had no idea how amazing it would be to have an advisor like Mercedes.

Secondly, I would like to thank the members of my committee-Greg Dwyer, Stefano Allesina, and Cathy Pfister for being so supportive throughout my PhD. Their advice and guidance at each stage of my PhD has been invaluable, and I have learned so much from them about ecology, my research, pursuing post-graduate careers, and everything in between. I have always enjoyed meeting with them, and really appreciate their support.

Thirdly, I would like to acknowledge Audrey Aronowsky. Audrey has taken the time to advise and assist me at every stage of my PhD, particularly as I transitioned between research projects and phases of my research. Her support and guidance were crucial-I do not think that the experiences that I've enjoyed as a graduate student in Mercedes' lab would

have been possible without her help. She has always gone above and beyond to make sure that all of the students in our program are able to flourish, and I am very lucky to have had her as my graduate program coordinator.

Next I would like to acknowledge Victoria Romeo-Aznar, who has been my primary collaborator and a wonderful friend throughout my time in Mercedes' lab. I will always be grateful to Victoria for taking the time to teach me so much about dengue, the Aedes aegypti mosquito, and how to do good science in general. I can't count the number of times I have walked over to her desk with a question or to get her feedback on an idea, and those conversations are something that I will also deeply miss. I would also like to acknowledge Qixin He, who has been my primary COVID-19 collaborator, for all her help and support, both as a scientist, colleague, and friend. I will definitely miss our lunchtime conversations!

I would like to especially acknowledge Pamela Martinez, for all of her guidance and advice. I feel truly lucky to have had such a kind and caring mentor throughout my time in the lab. She has always taken the time out of her schedule to give me incredibly useful advice on issues big and small, whether it was joining Mercedes' lab, learning how to fit models with POMP, or navigating my post-graduate career. Without Pamela's guidance, I would not have had all of the opportunities that I've enjoyed in Mercedes' lab, and for that I am extremely grateful.

I would also like to acknowledge several other members of the Pascual lab who took the time to guide me during my arrival in the lab, including Mauricio, who taught me how to use R to make maps of disease spread, and Xiagjun, who showed me how to run POMP jobs on the Midway cluster. More broadly, I want to thank all of the members of the Pascual lab that I've had the pleasure of knowing over the years (Ruby, Shai, David, Sergio, Armun, Qi, Frederic, and Shuanger). Our lab is truly like a family and has truly felt like my home. I have enjoyed every minute that I've spent with all of you, in the lab, lunchroom, or out on various adventures and escapades. I have learned so much from all of you and will miss you considerably. I would like to thank my dengue collaborators in Ann Arbor, Ed Ionides,

Aaron King, and especially Kidus Asfaw, who developed the particle filtering algorithms that I use throughout this thesis. I had the chance to travel to Ann Arbor to work closely with them to learn more about how to use POMP for dengue data, and the training and guidance that I received from them was invaluable. I have learned so much from them, both regarding the specific methodology of POMP as well as underlying in statistical inference. This knowledge and perspective has been extremely useful throughout my thesis research. I have really enjoyed working with and learning from them, and am grateful to have had the opportunity to do so.

I would also like to thank my dengue collaborators in Rio de Janeiro: Marcelo Gomes, Flávio Coelho, Laís Freitas, Raquel Lana, Oswaldo Cruz, and especially Claudia Codeço. I have learned so much about dengue transmission in Rio de Janeiro from them, and they have always been happy to answer the many questions that I've had over the course of the dengue project. I really enjoyed having the opportunity to travel to Rio de Janeiro in the summer of 2018 to meet them in person, and am grateful to have the opportunity to collaborate with them.

I would also like to acknowledge Joy Bergelson for her advice and guidance at several stages of my graduate career, whether it was preparing fellowship applications, scheduling my general knowledge exam, or transitioning between research projects. She has always been so helpful and supportive throughout this process.

Early in my time at UChicago, I had the opportunity to do an extended rotation with Sarah Cobey. I'd like to thank Sarah and the members of the Cobey lab, especially Sylvia, for providing me with this opportunity and for taking the time to make me feel welcome in the lab. The skills and perspective that I gained through my rotation were invaluable as I progressed through my graduate career.

I would also like to acknowledge all of the faculty in Ecology and Evolution and in other departments whom I have interacted with a student, TA, or instructor for courses, bootcamps, seminar series and journal clubs, as well as several lab groups that I have had

the chance to interact closely with, including the Allesina and Pfister-Wootton labs. These experiences (especially Ecoshake) have been an integral part of my time here at UChicago, and I have learned so much from all of you. The collaborative, supportive nature of our department and cluster is something that I have treasured and really benefited from.

I want to thank the Ecology and Evolution and Darwinian cluster program staff, including Mary Johnson, Bonnie Brown, Jeff Wisnewski, Connie Homan, Mike Guerra, and Marcy Hochberg. They have always been so supportive and kind and were always ready to help whenever I needed it!

I would also like to acknowledge several funding sources that supported me during my time here. These included two training grants, the GANN program in quantitative ecology and the NSF-NRT-DSEERT training program, as well as the Steiner Graduate Fellowship, the Urban Doctoral Fellowship, and summer travel funding from the Department of Ecology and Evolution. Furthermore, this thesis was completed with resources and support provided by the University of Chicago's Research Computing Center.

I have been quite fortunate to have a very close-knit E and E cohort, and would like to thank them for all of their care and support. I would especially like to acknowledge John Park, Mark Bitter, and Caroline Oldstone-Jackson for taking the time to check in and support me, even at some of the most stressful moments of my PhD when I was rather buried in my own research! I am grateful for their support and kindness, and would also like to acknowledge the broader Darwinian student community for being so inclusive and considerate.

One of the perks of being here for a longer time is that I've had the chance to get to know people from various other programs and departments across the university, many of whom have played an important role in my graduate education.

First among these groups are my NRT colleagues. I would like to acknowledge Liz Moyer, Emily Padston, and my NRT Cohort (Jim, Ziwei, Jess, and Takuya) as well as the rest of the NRT fellows and affiliated folks whom I've overlapped with (Haynes, Katie, Daniel, Maria,

Amanda, Ivan, Yuqi, Harshil, Girogio, Rafeal, Ander, Cherry, Katerina, Jess, and Hunter) . Working with and learning from all of you has given me a deeper appreciation for both data science and science itself, and the inter-disciplinary skillset and perspective that I obtained through the NRT program has been extremely valuable as I navigate the next stage of my career. Furthermore, I've really enjoyed getting to know all of you through bootcamps, practicums, and weekly meetings, and will miss talking with all of you over Zoom!

I'd also like to acknowledge the Mansueto Institute, including Steven, Anne, and my Urban Doctoal Fellows cohort (Onur, Zach, Robert, Nisarg, Sadiq, Jordie, Ashley, and Joseph). Learning from all of you through our writing and review sessions greatly deepened my perspective as a social scientist, and opened my eyes to some of the many wonderful ways that one can view the dynamics of a large city and the people who live in it beyond the type of population modelling that I have gotten accustomed to using. Thank you so much for all of your insightful comments and advice!

I would also like to acknowledge several individuals at the Research Computing Center who have been extremely helpful at various stages of my research, including Kimberly Grasch, Yuxing Peng, Hossein Pourreza, Johnathan Skone, Peter Carbonetto, and Parmanand Sinha. I use RCC resources extensively for my particle filtering, and have really appreciated the care that they put into all of their computing workshops. The RCC has been incredibly supportive whenever I have needed help during my research, whether it was requesting additional computing resources to test out new particle filtering algorithms or their COVID-19 special allocation during the first wave of the pandemic, which was vital in helping me complete the first chapter of my thesis. I am very grateful for their support.

I would also like to acknowledge Julia Kochinsky and Angela Li at the Center for Spatial Data Science. I have learned so much from them on spatial statistics and mapping in R, whether through the Tuesday morning seminar series or Angela's R Spatial data classes and workshops!

I've been incredibly fortunate to work with several excellent mentors during my time in

graduate school, especially Mercedes. However, I would also like to take this opportunity to acknowledge some of the mentors in earlier stages of my life who opened my mind to the possibility of pursuing a career in infectious disease research and provided me with the opportunity to go to graduate school in the first place.

I'd like to thank my undergraduate research advisors, Andrea Graham, Bryan Grenfell, and Nimalan Arinaminpathy, for exposing me to the wonders of disease ecology, immunology, and mathematical modelling. Working with them gave me a glimpse of how much fun research in disease ecology can be, and I am grateful to them for providing me with this opportunity and for taking the time to guide and mentor me throughout my academic career.

I would like to thank my high school science teacher, Christine Schultz, and Andrea Ferrante at the University of Alaska, for kindling my interest in science. During my time in high school, I had the chance to participate in a high school protein modelling program under Ms. Schulz's guidance in which students were paired with a mentor studying a particular protein. Learning about Dr. Ferrante's work at the Blood Center of Wisconsin examining the interaction between T-cell receptors, MHC presentation proteins, and the matrix protein of the influenza virus sparked my interest in immunology. During my senior year, Ms. Schultz and Dr. Ferrante took the time to guide me during a small retrospective cohort study of influenza infection in 2009, which was my first exposure to epidemiology. I am so grateful to the two of them for taking the time to guide me and introduce me to these fields, for I had no idea at the time what I would eventually study!

Finally, I'd like to acknowledge my high school's gifted and talented coordinator, Robin Schlei, who sparked my passion for research and critical thinking. Ms. Schlei was so much more than a gifted and talented coordinator-she served as a de facto teacher, counselor, mentor, and friend to so many of her students, including me. She took the time to have lunch with students every week, and was my first research mentor. Those first projects examined the evolution of the two main American political parties over the 20th century, high-speed rail in the Midwest, and connectivity within the city of Chicago's public transportation

networks. While the projects we did were very different from the type of research I currently specialize in, the mindset and emphasis on critical thinking, experimentation, and empathy that she emphasized continue to shape my approach to research. She played a crucial role in the evolution of my interests from public policy towards public health, epidemiology, and science, and continues to be my role model. I would not be where I am today without her kindness and support.

Last but certainly not least, I want to acknowledge all the love and support that I've received from friends and family. I've been lucky to have a close group of friends whom I could count on for support, including Bryanna, Divya, Steve, Jenna, Franklin, Linda, Peppar, Jessica, and Melissa. I want to thank them for always taking the time to check in and see how I'm doing and being so supportive at every stage of my grad school experience.

I would like to thank my parents for all of their love, advice, and support throughout my time in grad school. Their support has been simply indispensable.

I would also especially acknowledge my grandmothers, Kamala Thathi and Sumathi Thathi, for all of the love, kindness, and guidance that they have given me throughout my life, and especially during graduate school. The two of them are definitely the most intellectually curious , intelligent, and analytical people that I know. They've played a crucial part in my education and my life as a whole, whether it was teaching me how to read, do arithmetic, traveling half way across the world because I didn't like daycare, or asking me insightful questions about my dengue research. Getting an education and being a good human being were the two most important values that guided their lives, and they sacrificed everything so that my parents and I would have access to opportunities that they never had. Their boundless sense of hope and optimism has always been my inspiration. I am truly grateful to them for everything, because none of the opportunities that I've been fortunate to have would have been possible without them.

Finally, I would like to acknowledge Spike, Stitch, Lilo, and Tyke for their affection, support, and unconditional love.

ABSTRACT

Emerging viruses such as COVID-19 and dengue pose substantial public health risks in large cities, but their impact on host populations can be quite heterogeneous and hard to quantify using traditional mathematical models. I quantify the extent and impact of heterogeneity in infection status and in underlying transmission for both diseases. In my first chapter, I use a model that incorporates daily changes in testing capacity to precisely quantify the proportion of COVID-19 cases in New York City that were symptomatic during the initial epidemic wave. In my second chapter, I demonstrate that susceptible depletion on a city-wide aggregate level cannot explain the rapid re-emergence of dengue serotype DENV1 in the 1980s in Rio de Janeiro, Brazil, and suggest that inter-annual variation in climate, spatial heterogeneity within a large city, and coupling between cities may play an important role in dengue dynamics. In my third chapter, I use a panel of mechanistic models driven by cases from Rio to show that connectivity to the city of Rio played a crucial role in the spread of DENV4 through the Rio metropolitan area in 2012-2013, and that secondary movement hubs in the suburbs east of Rio could also be very important in facilitating the virus' spread to more outlying areas. These essential fluxes could be incorporated into statistical models that can already capture inter-annual variation in climate and socio-economic variables or integrated with smaller-scale mechanistic models to provide a multi-scale understanding of dengue transmission and re-emergence.

CHAPTER 1

INTRODUCTION

Emerging viruses such as COVID-19 and dengue pose substantial challenges to public health. These pathogens can have a heterogenous impact on human populations, whether in terms of the severity of morbidity such asymptomatic, symptomatic, and severe infections as well as in terms of the underlying drivers of transmission such as heterogeneity in population density between neighborhoods in a large city or differences in connectivity between municipalities in a metropolitan area. Mathematical models of infectious disease transmission are an important tool for characterizing this heterogeneity and evaluating its consequences on a population level, particularly during the early stages of an outbreak. Questions on heterogeneity are intertwined with those of the scale of description in challenging ways. Moreover, emerging viral outbreaks pose unique challenges for epidemiological inference and require modification to existing approaches and model formulations.

Our theoretical understanding of infectious disease transmission as well as the inference framework that we use to obtain that understanding are based on the dynamics of more endemic childhood diseases [11, 12, 13, 14] and to some extent seasonal influenza. These diseases are characterized by multi-year time series in which demographic turnover and susceptible-depletion at city-level scales are the primary drivers of disease dynamics, along with human movement between cities [13, 12], and well-established surveillance systems. Substantial progress has been made in developing inference methods for these types of time-series data [15, 14, 16, 17]. These types of approaches have also been extended to seasonal influenza [10] and applied at the scale of large urban landscapes to evaluate heterogeneity in influenza transmission across U.S. cities.

However, emerging virus outbreaks such as COVID-19, Zika, and new dengue serotype invasions exhibit very different characteristics. Many are characterized by short time series, with invasions occurring over a 2–3-year period [18, 5, 19]. New viruses such as COVID-19 have substantial under-reporting [20], with reporting rates that increase in time due to

increases in testing capacity [21]. This increase in testing capacity can make it difficult for models to distinguish between cases that were underreported due to a lack of testing capacity or due to a lower severity of symptoms. This confounding may explain why many early models of COVID-19 fit to case data had difficulty estimating the proportion of cases that were symptomatic [20] and had to wait for detailed symptom data from transmission studies to become available much later in the course of the epidemic.

New model formulations that can incorporate changes in testing capacity and reporting can help evaluate heterogeneity in infection status during the early phases of COVID-19 status. These model formulations can leverage other alternate data sources that were available earlier in the epidemic, including testing capacity estimates [22], serology [1], and syndrome surveillance reports [23]. For my first chapter, I use a model which incorporates all these data sources and accounts for changes in daily testing capacity to precisely estimate the fraction of COVID-19 cases that are asymptomatic. I find that between 1 in 5 and 1 in 7 cases are symptomatic, and that non-symptomatic cases substantially contribute to community transmission. Furthermore, depending on the strength of asymptomatic transmission, either the overall or symptomatic reproductive number during the first wave was higher than often assumed.

Like COVID-19, dengue can have a heterogenous impact on human populations. Much of this variation may be shaped by differences in drivers such as temperature [24, 25], rainfall[26, 27, 28, 29], and human movement, as well as heterogeneity in population density and socioeconomic status. Mathematical models are frequently used to determine when previously circulating arbovirus serotypes may re-emerge and understand how epidemics spread from one location to another during the invasion of a new serotype.

Although dengue serotypes have been circulating for decades, the invasion of new serotypes can result in large outbreaks. The invasions of dengue serotypes DENV1 and DENV4 in southeastern Brazil in 1986 and 2012 [30, 18, 31] provide several illustrative examples. Unlike dengue epidemics in Thailand, where all four serotypes of dengue have been co-circulating

for decades [32], dengue dynamics in southeastern Brazil are characterized by sequential invasions of dengue serotypes [5, 18, 33], with each invasion taking place over a 2–3-year period in which the invading serotype is the dominant serotype, as well as occasional re-emergent outbreaks of serotypes that invaded in previous years[34, 30]. The most recent dengue invasion (DENV4) in 2012 [31, 30, 33, 18] was followed by the invasion of chikungunya in 2015 and Zika in 2016 [35].

Multiple factors can play a role in shaping dengue dynamics, including serotype interactions [32], antibody-dependent enhancement [36], climate drivers such as temperature [24, 25] and rainfall [26, 27, 28, 29, 37], socio-economic variables such as population density [38] and access to water sources [37], human movement patterns [39, 40, 41], as well as the depletion of susceptible hosts. Immunity to the same dengue serotype is believed to be long-lasting, with short term cross-immunity between serotypes [42]. However, intra-serotype immunity may not necessarily be long lasting, since neutralization studies have indicated that intra-subtype antigenic variation can be just as large as inter-subtype variation [43]. The dengue virus is transmitted by the Aedes aegypti mosquito, which is also the transmission vector for the chikungunya, Zika, and yellow fever viruses. Prior infection with Zika can increase a host's risk of severe dengue infection [44] . The Aedes aegypti development and biting rates are temperature dependent [24, 25], and these dependencies can be parameterized using results from laboratory experiments [25]. Rainfall can impact mosquito populations in several ways. The Aedes aegypti mosquito breeds in pools of standing water in which biomass is available. Increased rainfall can result in increased vector capacity by creating more pools of standing water [37]. However, flooding can also reduce mosquito population sizes by washing out mosquito larvae [26, 27]. Furthermore, periods of drought can also indirectly increase mosquito population sizes if the drought alters water storage patterns [37]. Within a large city, neighborhoods with different human population densities and socio-economic characteristics can have different mosquito population sizes, with denser areas experiencing higher transmission [45, 38]. Household to household transmission of dengue via human movement

is also an extremely important driver of dengue epidemics [39].

Some of these factors, such as susceptible depletion [13], serotype interactions [46, 32], and temperature dependence [25] can be represented into dynamic models, while others such as rainfall dependence are difficult to include mechanistically [37]. Stochastic versions of these models can also account for demographic stochasticity, measurement error, and environmental noise due to environmental variables that cannot be explicitly included in the model [47]. Human movement can also be incorporated into mechanistic dynamical models, although fitting stochastic meta-population models to case data can be technically challenging due to the high dimensionality of the state space [48]. Fitting these models often requires either additional vector capacity data [41] or fitting only a few cities in a region [40]. Dynamical models have difficulty mechanistically capturing the impact of rainfall on transmission. These models can be parameterized such that the ratio of mosquitoes to humans in a particular area is a function of population density, but these relationships have only been explored for case data at very fine spatial scales [38].

Spatiotemporal statistical models can easily capture lagged effects due to climate variables such as temperature and rainfall, as well as site-specific terms due to population density and socio-economic variables [49, 37, 50]. While they cannot account for the depletion of susceptibles mechanistically, these models can use year-specific terms to account for years with large epidemics [49, 37]. However, these models use case data aggregated at a micro-region scale rather than at the municipality scale [49, 37] to avoid dealing with many observations with zero observed cases in smaller municipalities in regions with non-endemic dynamics such as southeastern Brazil. Furthermore, these statistical models often have difficulty incorporating human movement fluxes between non-adjacent municipalities. [37, 51])

Dynamical mechanistic models are frequently used to make long-term re-emergence forecasts on the order of decades [52], while statistical models can be used to make short-term forecasts on the order of months [49]. Both mechanistic and statistical models have been used to understand the spread of new dengue serotypes across locations[32, 49], depending

on the factors which are assumed to be important at the scale and location being examined.

At a large scale, when analyzing epidemics at a regional, national, or continental level, climate drivers such as temperature and rainfall are often assumed to be the key drivers of arbovirus transmission along with population size. Re-emergence forecasts [52] or dengue "risk maps" [53] using mechanistic models at this scale frequently consider differences in seasonality in transmission due to temperature variation between locations, but do not consider inter-annual climate variation. At a national scale, mechanistic models that seek to explain the spread of dengue from one municipality to another may include movement patterns in some form [40].

Factors such as population density and other socio-economic variables are often emphasized at intra-city scales such as at the neighborhood scale or even smaller. Many detailed transmission studies of dengue focus on these scales [45]. While several approaches have been developed to incorporate the impact of heterogeneity in population density into transmission models in which susceptible depletion occurs at the sub-census tract level [38, 54], these parameterizations require fine-scale case data, and these effects are often not included in larger scale models which assume that susceptible depletion occurs at the city level.

For this analysis, we focus on understanding dengue re-emergence and spread at the level of a large city, specifically Rio de Janeiro, Brazil. This scale is finer than the region-wide scale used for re-emergence forecasts and risk maps, but coarser than the neighborhood scale used for transmission studies. This scale is particularly important for several reasons. First, large cities serve as key hubs for dengue transmission in a region [55, 50]. Secondly, public health surveillance and responses are often implemented on a local level[56]. Finally, while some cities may make finer-scale case data available in specific scenarios, city-wide case data are often the finest resolution of case data that are available to scientists during an emerging outbreak. Approaches that require very fine scale case data may be difficult to generalize and apply to multiple cities at once. Given this context, it is important to investigate the utility and limitations of the city-level transmission models most likely to be deployed during

an emerging outbreak or across large regions.

There are several challenges associated with modelling large cities, since cities in essence are collections of smaller neighborhoods, each with their own transmission characteristics. Treating the large city as a giant well-mixed population essentially ignores this underlying meta-population structure. Secondly, large cities are not closed systems and their municipal boundaries do not necessarily correspond to population boundaries. Large cities are becoming increasingly connected and integrated with the suburbs that surround them. Understanding how arboviruses spread in these environments may require mathematical approaches that consider the metropolitan region as a whole along with the movement fluxes between the municipalities that comprise them.

There are two inter-related questions which my work seeks to address. The first one concerns the extent to which susceptible depletion and heterogeneity population density at smaller scale explain dengue dynamics at larger scales. My second chapter begins to answer this question by first examining whether modelling susceptible depletion at a city-wide scale (essentially treating the city as a large well-mixed population) is sufficient to capture re-emergence dynamics. This is an assumption that has been made when making re-emergence forecasts for other arboviruses such as Zika [52]. In my second chapter, I investigate the extent to which susceptible depletion at the city-wide level can explain the re-emergence of DENV1 following its initial invasion in 1986 [5]. I show that a model fit to the initial invasion substantially under-estimates the time to re-emergence, and the expected time to re-emergence can be very sensitive to small changes in transmission parameters. These results suggest that treating the city as a well-mixed population may not be sufficient.

These results are consistent with intriguing empirical patterns for reported cases at a very fine scale within the city of Rio de Janeiro [54]. In particular, patterns for the ratio of successive epidemic waves indicate an important role of population density at the extremely fine scales of census tract (about a block or two). This ratio is of central interest because it reflects the interplay of herd immunity build-up and transmission seasonality. The patterns

further suggest a key role of population density through arrival time, and a dependence of the rate at which new "sparks" of infected cases arrive in units in the city during the 2012 DENV4 invasion as a function of the population density of that unit but also total prevalence at the level of the whole city [54]. This suggests a simpler metapopulation model than the typical formulation requiring the full coupling among all units, in which local population density is finely resolved but coupling is global. I discuss future directions concerning this model in the Conclusions.

My third chapter investigates the ways in which connectivity due to daily commuter movement may facilitate the spread of dengue epidemics spread between municipalities in a metropolitan area . Spatiotemporal statistical models are well-suited to investigate how dengue epidemics spread between municipalities [37, 51], but their inability to incorporate movement fluxes between municipalities is a major limitation of this approach [51]. Current research efforts have focused on ways to incorporate the "essential" aspects of the underlying movement structure into statistical models [51]. However, it is not clear what the epidemiologically "essential" aspects of this movement structure are in the first place, particularly for large metropolitan areas like Rio de Janeiro. There can be substantial heterogeneity in movement fluxes in the municipalities surrounding Rio, with some having large amounts of commuter traffic and others being more isolated. Furthermore, several "suburbs" of Rio are themselves rather large municipalities and may also be destinations for commuter traffic. In my third chapter, I use a panel dynamical model in which 20 surrounding municipalities are coupled to Rio via commuter traffic but are independent of each other and investigate the extent to which this model can capture observed dynamics of the DENV4 invasion. I then use newly developed particle-filtering techniques [57, 58] to evaluate the likelihood of a fully coupled version of this model, specifically the importance of peri-urban movement fluxes that do not start or end in the city of Rio. Early results indicate that movement to and from the city of Rio de Janeiro did play an important role in the spread of DENV4 from Rio to the surrounding municipalities and fluxes from the city of Rio were sufficient to reproduce some

of the observed dynamics. However, we also found that including peri-urban fluxes improved the performance of the model, especially in municipalities which had a high proportion of flux to or from cities other than Rio. Our results suggest that incorporating fluxes to and from large cities may be a reasonable approximation to the underlying movement structure when developing future extensions to statistical models as long as those large cities are the main sources of that flux, but that peri-urban fluxes can also be epidemiologically meaningful. More broadly, our results illustrate the importance of understanding the underlying movement structure of a metropolis, and that it should be possible to make considerably simpler approximations of that structure.

Overall, this thesis provides several examples for how existing modelling approaches can be modified to understand the heterogeneous impact of emerging viral outbreaks and the importance of making those modifications. All these approaches make use of city-level case data typically available during an emerging outbreak. In my first chapter, I use of models incorporating testing capacity to estimate the fraction of COVID-19 cases that are asymptomatic in New York City in spring 2020. In my second chapter, I demonstrate that models which assume that the depletion of susceptibles occurs at a city-wide level fail to predict the rapid re-emergence of DENV1 in Rio de Janeiro, Brazil, suggesting the potential importance of models that can incorporate small-scale heterogeneity in population density within a large city. Finally, in my third chapter, I develop a panel framework that incorporates movement fluxes between the city of Rio de Janeiro and each of 20 surrounding municipalities and use this model to evaluate which aspects of the heterogeneity and structure in commuter movement are epidemiologically meaningful.

CHAPTER 2

QUANTIFYING ASYMPTOMATIC INFECTION AND TRANSMISSION OF COVID-19 IN NEW YORK CITY USING OBSERVED CASES, SEROLOGY AND TESTING CAPACITY

[Originally published as: Subramanian, R., Q. He, M. Pascual. 2021. Quantifying Asymptomatic Infection and Transmission of COVID-19 in New York City using Observed Cases, Serology and Testing Capacity. Proceedings of the National Academy of Sciences 118(9): e2019716118]

2.1 Introduction

Since the emergence of the novel coronavirus in December 2019 [59], the COVID-19 pandemic has resulted in over 16 million cases and 600,000 deaths worldwide [60]. Schools and universities in the United States are gradually re-opening amid concerns that a second wave of the epidemic may re-emerge in the fall and winter of 2020.

As they craft testing policies and intervention strategies to mitigate a second wave, public health officials need to better understand the role that symptomatic and asymptomatic individuals play in the community transmission of COVID-19 and in the development of herd immunity to the disease. However, fundamental epidemiological questions remain poorly understood, including what fraction of cases are symptomatic and how well asymptomatic cases can transmit relative to symptomatic ones. These questions are especially urgent given ambiguity in recent CDC guidelines regarding the testing of asymptomatic individuals [61].

Answering these questions can also provide further insight on the basic reproductive number of SARS-CoV-2, and how the virus would spread in a population in the absence of interventions. This number known as R_0 is defined as the mean number of secondary cases arising from a primary case in the absence of immunity, and is estimated on the basis of a particular epidemiological model. Mathematical models for the population dynam-

ics of COVID-19 incorporate different features such as asymptomatic and pre-symptomatic transmission, super-spreading, or heterogeneity in susceptibility. A considerable range of R_0 estimates has been reported, ranging from at least 1.5 [62] to 5.7 [63] in Wuhan. A much narrower range between 2 and 3 is frequently cited in the popular press, or assumed when simulating models [64] or fitting these to data [65, 66] . This assumption may be based on the dynamics of COVID-19 in regions that implemented interventions early [20, 67, 68, 69, 70]. A more precise estimate of R_0 from a city where substantial transmission was occurring prior to intervention, such as New York City, would provide a relevant baseline. Furthermore, if "super-spreading" by a small fraction of symptomatic infections fuels COVID-19 transmission, a precise estimate of the mean number of secondary cases arising from such an individual, may be just as valuable. A model that precisely estimates the fraction of symptomatic cases may help epidemiologists discern if either the overall or symptomatic reproductive numbers are higher than assumed.

The probability that a COVID-19 infection is symptomatic is difficult to estimate [71] and a wide range of values have been suggested [71, 72, 73]. Estimates from cruise ship outbreaks [74], Wuhan evacuees [75], long term care facilities [76], and contact tracing of index cases [72] may not be representative of the general population. Increases in the testing capacity for COVID-19 over time [77, 21, 20] make population-level estimation of this probability difficult due to confounding with other parameters such as the reporting, hospitalization, and fatality rates. When the testing capacity is limited in the early stages of an outbreak, severe cases are more likely to be tested, which can bias estimates of the probability that an infection is symptomatic and the fatality rate. Changes in testing capacity over time also confound the definition itself of asymptomatic individuals in transmission models, when these are not differentiated from unreported cases. These changes can also bias the reported deaths attributed to COVID-19.

These challenges can be improved upon by explicitly incorporating changes in testing capacity into an epidemiological process model. While some early models of the COVID-

19 outbreak in Wuhan attempted to take into account changes in testing capacity [21] or differences in reporting rate during periods of the epidemic [20], the limited information on these trends in Wuhan meant that they had to be estimated on a coarse temporal scale (2-3 week intervals) and had to be inferred along with other parameters in the model. In the United States, many states and municipalities such as New York City [78, 22] have published daily estimates of the number of total COVID-19 tests conducted, together with the number of positive COVID-19 tests. While these data are often used by public health officials to gauge the spread of the COVID-19 outbreak, they have yet to be incorporated explicitly into epidemiological models.

We present an epidemiological model that incorporates RT-PCR testing as an integral process informed by empirical levels. The explicit consideration of testing allows us to clearly define asymptomatic individuals as those that will never transition to displaying symptoms, and to differentiate them from those who have been unreported because they were not tested. We fit the model to PCR-confirmed COVID-19 cases in New York City, using publicly available data provided by the New York State Department of Public Health [22]. The resulting model can clearly delineate symptomatic and asymptomatic infections independently from the reporting rate. We subsequently fit the model to estimates of prior exposure obtained from a recent serological study in New York City [1] to further constrain inference results.

Our model obtains a precise estimate for the symptomatic proportion of COVID-19 cases. We show that most COVID-19 infections are asymptomatic, and that these asymptomatic infections together with pre-symptomatic ones substantially drive community transmission, contributing 50% or more of the total force of infection. Furthermore, depending on the transmissibility of individual asymptomatic cases relative to symptomatic ones, either the overall reproductive number or the symptomatic reproductive number may be higher than typically assumed. Our results highlight the importance of testing and contact tracing of asymptomatic individuals, and of making these data publicly available as health officials

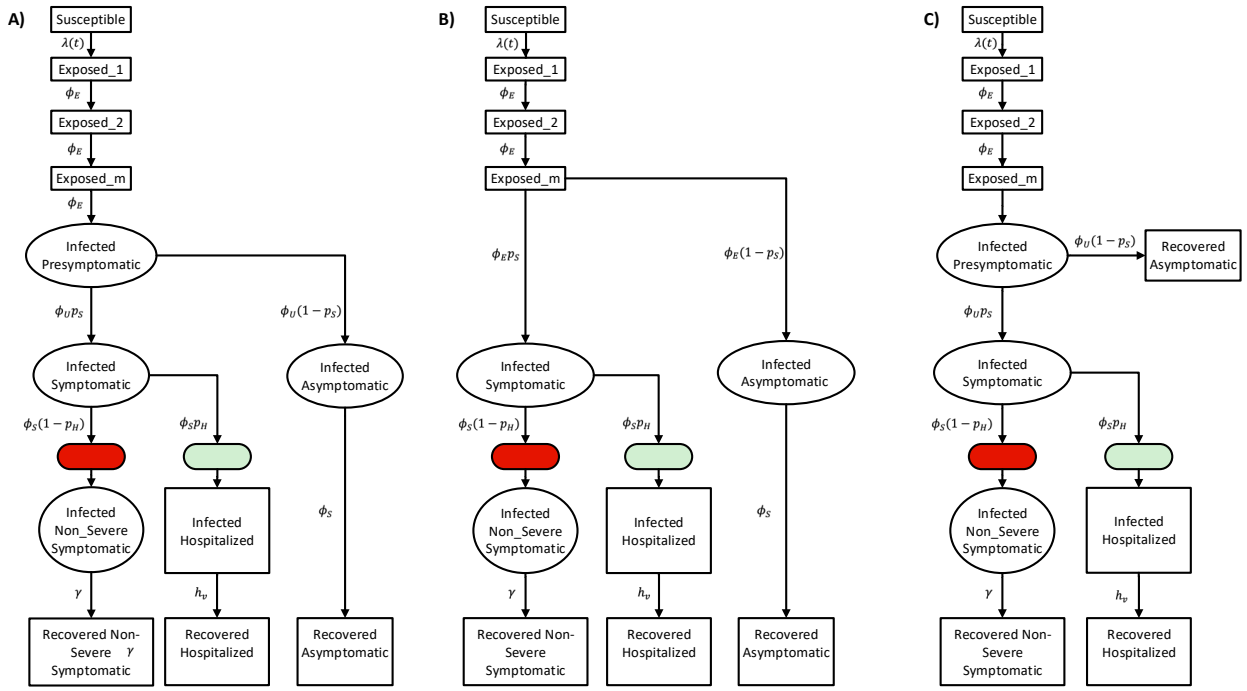

Figure 2.1: **Model diagrams. (A) The full SEPIAR model used for inference.** The model is an extension of an SEIR formulation that considers both pre-symptomatic transmission (from compartment P) and asymptomatic transmission (from compartment A). **B) When the strength of pre-symptomatic transmission b_{p} is set to 0, the SEPIAR model reduces to the SEIAR model.** Since we assume that $\phi_U = \phi_E$, when $b_p = 0$ the infectious pre-symptomatic compartment behaves like an additional exposed compartment. **C) When the strength of asymptomatic transmission b_{a} is set to 0, the SEPIAR model reduces to the SEPIR model.** Individuals in the asymptomatic infectious compartment (A) make no contribution to the force of infection, so asymptomatic individuals essentially recover after leaving the pre-symptomatic period (P). In all three panels, circular/elliptical compartments contribute to the force of infection, while rectangular compartments do not. The green ellipse denotes the point at which severe/hospitalized COVID patients are sampled and enter the testing queue for severe cases, while the red ellipse denotes the corresponding entry point for the queue for non-severe symptomatic cases.

prepare for and manage a second wave.

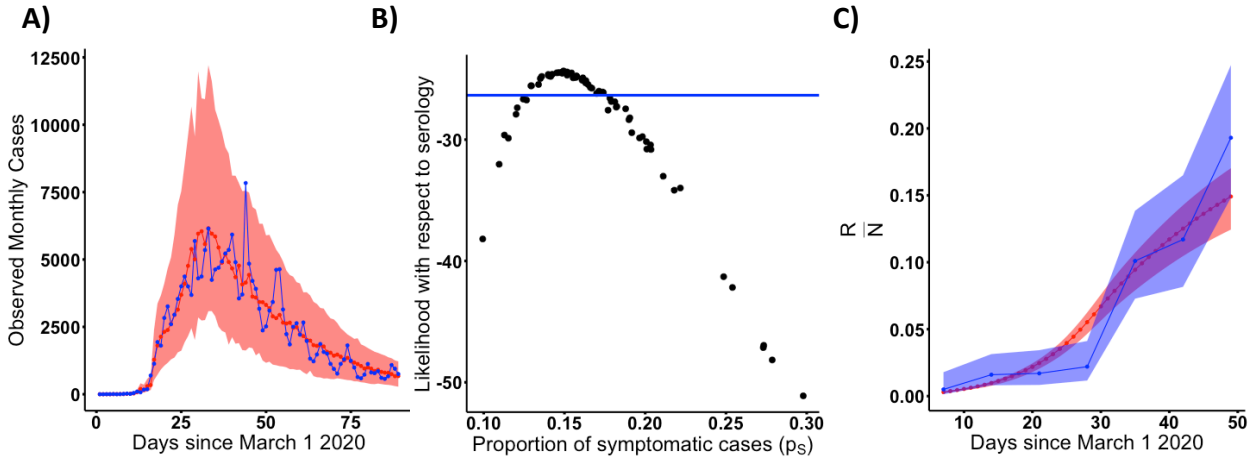

Figure 2.2: **The probability of symptomatic infection. (A) Simulated vs. observed cases from the profile of the asymptomatic transmission strength (b_a) using the SEPIAR model.** The red line is the median from 100 simulations using the Maximum-Likelihood Estimates (MLE), while the red shaded region denotes the 2.5 to 97.5% quantiles across 100 simulations from all parameter combinations within 2 log-likelihood units of the profile MLE. Likelihoods here are with respect to case data. The observed daily case counts are denoted by the blue line. **B) Model Likelihood as a function of the proportion of cases that are symptomatic (p_S) for each parameter combination from panel A.** The y-axis shows the likelihood for that parameter combination with respect to serology data. All parameter combinations above the blue line have likelihoods within 2-log-likelihood units of the MLE (defined with respect to serology). This corresponds to a range of values for p_S of approximately 13 to 18%. **C) Comparison of observed vs. simulated estimates of herd immunity in the population from parameter combinations supported by both case and antibody data (all points above the blue line in panel B).** The red line denotes the median value of herd immunity (the proportion of the population that has recovered ($\frac{R}{N}$) at that point in time in 100 simulations from the MLE parameter combination. The red shaded region denotes the 2.5 to 97.5% quantiles for these simulations from all parameter combinations within 2-log-likelihood units of the MLE with respect to serology (all parameter combinations above the blue line in panel B). The blue line denotes estimates of herd immunity from a recent serological survey in New York City [1]. The blue shading denotes 95% confidence intervals for those serology estimates using the methods of [1].

2.2 Results

We present a stochastic epidemiological model (Fig. 2.1) that explicitly incorporates daily changes in testing capacity and the lag between sampling and testing (see Methods). The

underlying model, referred to hereafter as the SEPIAR model (Fig. 2.1 A) has a susceptible-exposed-infectious-recovered structure with compartments for both severe (hospitalized) and non-severe symptomatic infections as well as pre-symptomatic (P) and asymptomatic (A) infections. We also consider two nested simplified versions: one with no pre-symptomatic transmission (the SEIAR model, Fig. 2.1B); and one with no asymptomatic transmission (the SEPIR model, Fig. 2.1C). By varying specific parameters weighting the transmission rate of P and A relative to that of symptomatic individuals, we can continuously move across these two extreme structures. Daily reports of the number of tests conducted in New York City are fed in as a co-variate in the testing sub-model (see SI Appendix). The model takes into account CDC priorities in sampling and testing: all hospitalized cases are sampled and eventually tested, while non-severe symptomatic individuals are sampled and tested only if excess capacity is available at the time of sampling. We also incorporate the re-testing of hospitalized individuals as they leave the hospital. This model is fit to observed cases in New York City from March 1,2020 to June 1, 2020 and serological estimates of herd immunity in New York City from March 8,2020 to April 19,2020 (see Methods and SI Appendix). We compare the full model with the two nested simplified versions. Although all three model structures are supported by the case data, the model with no asymptomatic transmission is not supported when these data are considered in conjunction with serology information (SI Appendix, Table 2.2).

To evaluate the strength of transmission in asymptomatic cases relative to symptomatic cases, we construct a Monte Carlo profile using the full SEPIAR model (SI Appendix, Fig. S5). We isolate parameter combinations from the profile that are supported by the case and serology data, and examine the values of those combinations. Particular parameters of interest that we focus on include the proportion of cases that are symptomatic, p_S, the ratio of the transmission rate of asymptomatic individuals to that of symptomatic individuals, b_{a} , and the reproductive numbers. We use R_0 to denote the symptomatic reproductive number (i.e. the mean number of secondary cases arising from each primary symptomatic case), and

$R_{0_{\text{NGM}}}$ to denote the overall reproductive number for the model (i.e. the mean number of cases arising from a primary infection, where the average considers all types of infections).

The proportion of COVID-19 cases that are symptomatic is well identified, with a confidence interval ranging from 12.9% to 17.4% (Figure 2). Although a wide range of parameter combinations for the proportion of symptomatic infection are supported by the case data on its own, a much narrower estimate is obtained when the case and serology data are considered together (Fig. 2.2A, B). Within this range, estimates of herd immunity are consistent with the dynamics of observed serology (Fig. 2.2C), in particular the rapid rise in seroprevalence over March and April 2020. We validated the inference pipeline by fitting the model to two simulated trajectories from two parameter combinations that are both supported by the case, serology, and testing data but correspond to regimes with strong or weak asymptomatic transmission. As shown in panel B of Fig. 2.16, models fit to both of these trajectories instead of observed cases are able to accurately estimate and recover the proportion of symptomatic cases used in the simulations.

The overall reproductive number or symptomatic reproductive number may be larger than is often assumed. From our profile of the relative asymptomatic transmission rate b_{a}, we identify two main regimes of transmission that are supported by both the case and serology data (Fig. 3), in which either R_0 $or R_{0_{\text{NGM}}}$ is higher than the 2-3 range often assumed for COVID-19. Notably, we find no parameter combinations in which both reproductive numbers are below 3 and fall within this range.

In the first regime, asymptomatic individuals transmit at almost the same rate as symptomatic individuals. That is, b_{a} is large, even close to 1 in some parameter combinations. The overall reproductive number takes on values between 3.2 and 4.4, and asymptomatic cases substantially contribute to the overall force of infection (Fig. 2.4).

In the second regime, asymptomatic individuals transmit at very low rates relative to symptomatic individuals, with estimates of b_{a} close to zero or in some parameter combinations even equal to zero. Concomitantly, the symptomatic reproductive number is much

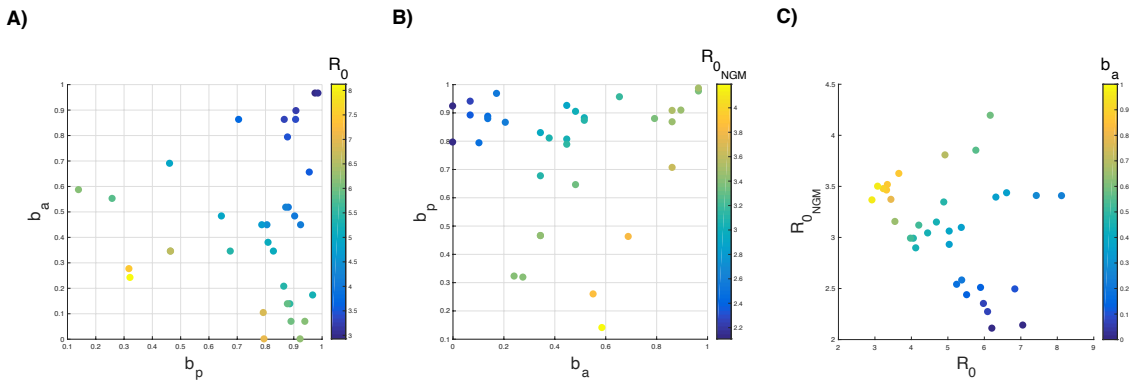

Figure 2.3: **Plots of the reproductive number of symptomatic individuals (R_0) (A) and the overall reproductive number ($R_{0_{\mathrm{NGM}}}$) (B), as a function of the relative strength of pre-symptomatic transmission (b_{p}) and the relative strength of asymptomatic transmission (b_{a}).**
Each point represents one parameter combination within 2 log-likelihood units of the MLE (with respect to serology) from the b_{a} profile. **C) Plot of the overall reproductive number vs the reproductive number in symptomatic individuals for the same points colored by b_{a}.** The black arrows show the direction of increasing strength of asymptomatic transmission (b_{a}) and pre-symptomatic transmission (b_{p}). For this same plot except colored by the strength of pre-symptomatic transmission (b_{p}), see Fig. 2.11. For ease of plotting, we exclude two parameter combinations which had a very low relative rates of pre-symptomatic transmission (i.e. b_{p} was lower than 0.020). The two outlier combinations had high reproductive numbers ($R_0 = 17.77, R_{0_{\mathrm{NGM}}} = 3.95$ and $R_0 = 4.97, R_{0_{\mathrm{NGM}}} = 4.37$). These outliers are included in the Fig. 2.12.

higher than frequently assumed, taking on values between 3.9 and 8.1. Nevertheless, even in this regime pre-symptomatic and asymptomatic infections together contribute at least 50% of the overall force of infection at the peak of the outbreak.

In both regimes, pre-symptomatic individuals transmit at almost the same rate as symptomatic individuals, with estimates of b_{p} close to 1, also making a substantial contribution to the overall force of infection (Fig. 2.4).

We also observe a third regime in which both reproductive numbers are higher than assumed, but in this regime pre-symptomatic individuals transmit at a very low rate, with b_{p} close to 0. Several combinations in this regime can be observed in the top right corner of Fig. 3 (C) and in Fig. 2.12. This is also the regime obtained in Fig. 2.11 if one uses the SEIAR model, which assumes that pre-symptomatic individuals do not transmit (i.e. b_{p} is

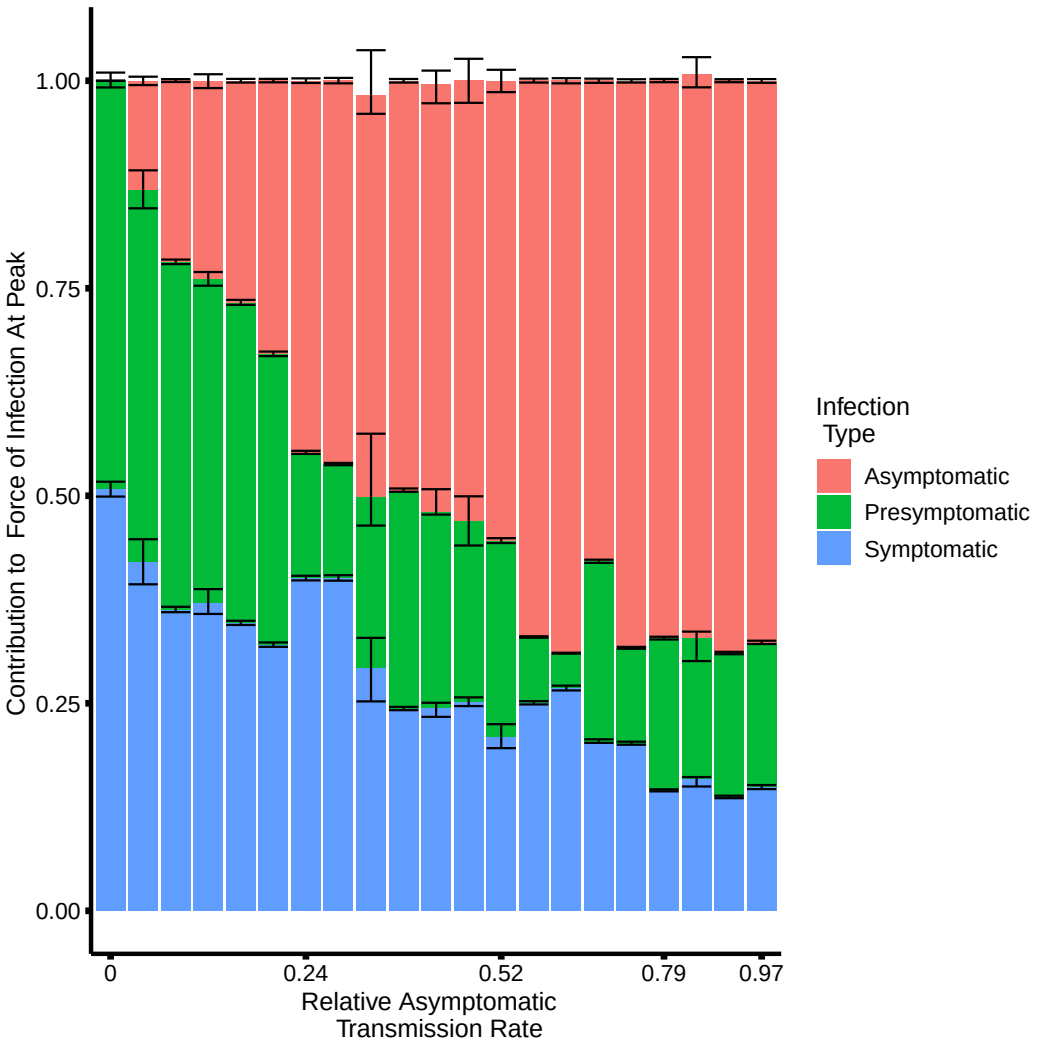

Figure 2.4: **The contribution to the force of infection at the peak of the outbreak on April 14, 2020 from symptomatic, asymptomatic, and pre-symptomatic infections under different relative asymptomatic transmission rates b_a.** For each parameter combination from the fitted SEPIAR model supported by case and serology data (corresponding to the points in Figure 2.3), we simulate 100 trajectories and calculate the proportion of the overall force of infection on April 14,2020 that is due to asymptomatic, symptomatic, and pre-symptomatic infections. We pool trajectories from all parameter combinations that have the same value of b_a, and calculate the median, 2.5%, and 97.5% quantiles for each infection class and value of b_a. The colored bars represent for each infection class, the median proportion of its contribution to the force of infection (and hence may not sum exactly to 1). The error bars represent the corresponding 2.5%, and 97.5% quantiles. Versions of this plot calculated respectively 4 weeks before, and 4 weeks after, the peak can be found in the SI Appendix Fig. S9. We excluded two outlier parameter combinations that had extremely low relative rates of pre-symptomatic transmission (i.e. where b_p was less than 0.02).

fixed at 0). Given previous evidence of pre-symptomatic transmission of COVID-19 [79, 80], we focus on the two regimes which incorporate substantial pre-symptomatic transmission.

In line with previous studies [81], we estimate a large value for the initial number of infected and incubating individuals with COVID-19 in New York City at the start of the simulation on March 1st. Parameter combinations that were supported by the case and serology data ranged from 9,000-18,000 initial infected individuals and 44,000-72,000 exposed individuals. A key question to consider when evaluating the plausibility of this magnitude of undetected infections is whether it is consistent with no signal of an anomalous number of hospitalizations. In other words, would this large rise in early infections result in a corresponding rise in COVID-hospitalizations that may not have been detected as COVID-related? We examine this question by comparing simulated daily hospitalizations from our fitted model with observed COVID-19 daily hospitalizations in New York City, as well as with syndrome surveillance reports of respiratory illness from emergency departments in New York City hospitals (Fig. 2.5), which we can use as an indicator for a rise in undetected hospitalizations. We show that a scenario with a large number of initial infections on March 1st is indeed consistent with the time at which observed COVID-19 hospitalizations peak, providing further support for this contention. We also find that the imposition of social distancing on March 17th and the stay at home order on March 22nd in New York City resulted in a substantial decrease in the initial transmission rate. Parameter estimates for the ratio of the post-intervention transmission rate to the pre-intervention transmission rate (b_{q}) ranged from 0.134 to 0.240, corresponding to a 75.98%-86.62% reduction in the strength of transmission after the intervention.

Testing strategies and capacity can substantially influence estimates of the infection fatality ratio, or IFR (Fig. 2.17). This metric of outbreak severity is generally defined as the total number of deaths divided by the total number of cases. In practice, this ratio is calculated by dividing the total number of confirmed deaths by the total number of observed cases. However, depending on the testing strategy used and the testing capacity available,

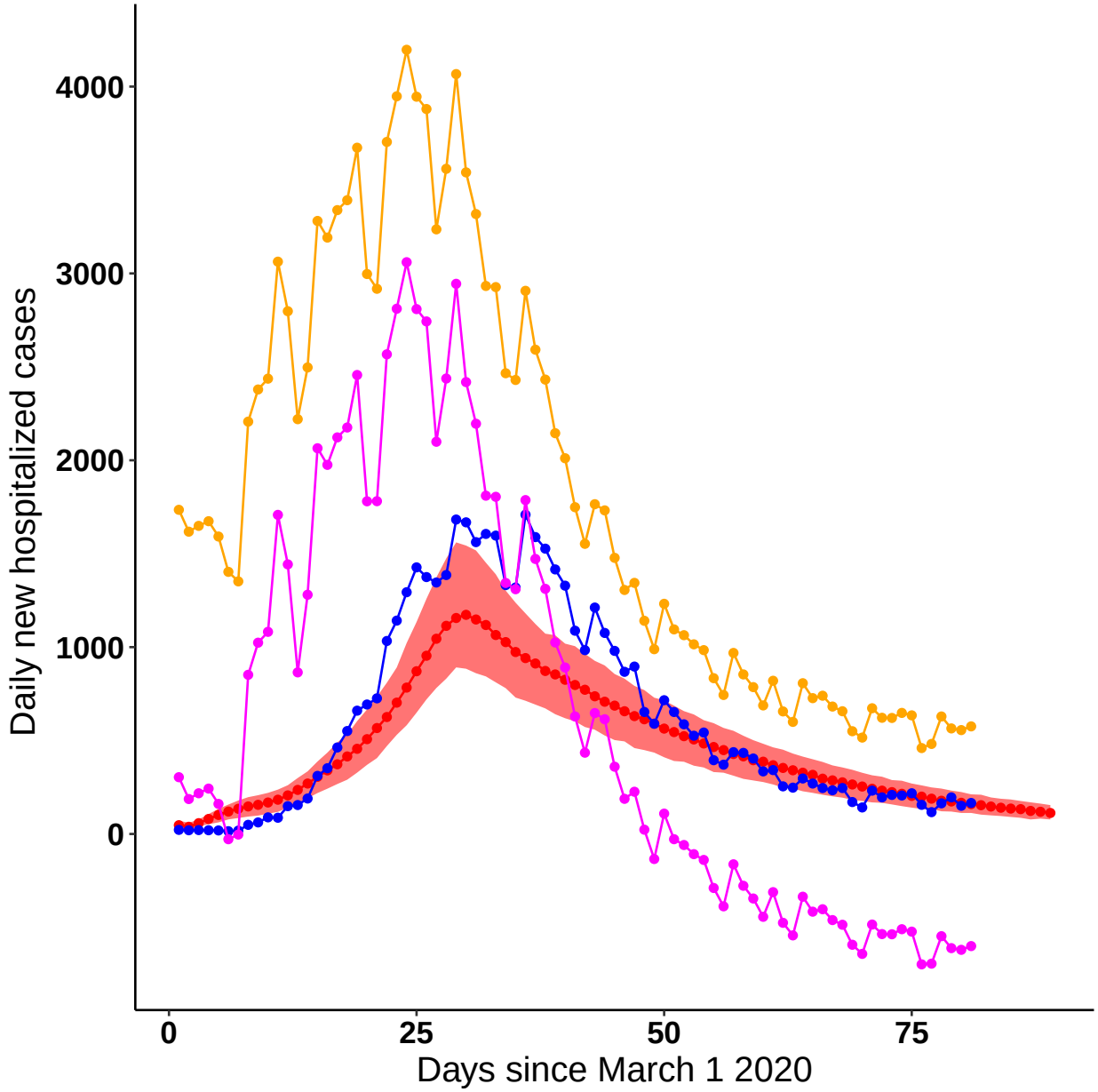

Figure 2.5: **Comparison of daily COVID hospitalizations under the model with observed COVID hospitalizations in New York City and emergency department respiratory syndrome surveillance reports.** The red line represents the median daily case hospitalizations from 100 simulations from the parameter combination with the highest likelihood with respect to serology from the b_{a} profile. The red shading represents the bounds of the 2.5% and 97.5% quantiles across all parameter combinations from the b_{a} profile that are supported by case and serology data. The blue line shows observed COVID daily hospitalizations in New York City. The yellow line denotes daily reports of respiratory illness from syndrome surveillance in New York City emergency departments, while the pink line denotes anomalous respiratory surveillance reports compared to previous years.

not all cases will be observed. Using parameters from the fitted SEPIAR model that are supported by case and serology data, we generate a range of infection fatality ratios that would be expected under two different testing strategies. Since we do not model deaths from COVID-19 hospitalized patients in our model, we estimate the proportion of hospitalizations that result in death using confirmed COVID-19 hospitalizations and deaths in New York City during the study period. In the first testing strategy, all cases are observed; in the second one, all symptomatic cases but no asymptomatic ones are observed (red and blue shaded histograms, respectively, in Fig. 2.17). Testing only symptomatic cases can result in at least a four-fold increase in the IFR that is calculated. Limitations in testing capacity may also impact the estimated IFR. If the testing capacity is limited at the start of the outbreak, the observed IFR measured during the epidemic (the orange vertical line in Fig. 2.17) will be higher than the IFR expected if all symptomatic cases were tested. Variation in individual model parameters within the range supported by the case and serology data does not result in substantial variation in the IFR calculated for each testing scenario.

Discussion

With a transmission model that incorporates daily changes in RT-PCR testing capacity and is fit to observed case data and serology, we estimate that the probability that an exposed individual develops symptoms is low. Since asymptomatic infections represent a large fraction of the infected population, they contribute substantially to community transmission in the aggregate together with pre-symptomatic cases, even when they individually transmit at a low per-capita rate. They also contribute substantially to building herd immunity.

We use testing information to estimate the probability that a new case will become symptomatic without the biases present in cruise-ship [74] and traveler studies [75], or the parameter confounding present in city-wide models. Early cruise-ship and evacuee studies found that most COVID-19 cases were symptomatic. However, given the small number of total infections [75, 82], evacuee studies may over-estimate the fraction of symptomatic

cases if infections in observed severe cases [83] last longer [84] than in asymptomatic ones. Cruise-ship studies may likewise over-estimate this parameter if asymptomatic cases, which were tested later than symptomatic cases [74], recover prior to testing. City-wide models, which avoid these biases, indicate that most COVID-19 cases are undetected [20]. They confound however the fraction of symptomatic cases with the reporting or hospitalization rate, as they neglect daily testing changes, and cannot distinguish between asymptomatic and undetected cases. The alternative approach of fitting the models to death data is not necessarily exempt from biases in parameter estimates, due to changes in hospital capacity over time [85, 86], co-morbidities in host populations [87, 88], and the long delay between the onset of infection and death [89]. Furthermore, the under-reporting of cases can also bias the assumed case fatality rate [86]. Our approach resolves these issues by incorporating daily testing capacity as part of the model when estimating parameters from serology and case data. Models without explicit consideration of this capacity have difficulty estimating the proportion of cases that are symptomatic from these data [90], suggesting that including testing is crucial.

If asymptomatic individuals transmit at a high rate, then the overall reproductive number pre-intervention in New York City is larger than the 2-3 range often assumed in models [64, 65, 66] and media reports [68, 91, 92, 93] based on early estimates from Wuhan [62, 94, 95]. Furthermore, we find no supported parameter combinations in which both the overall and symptomatic reproductive numbers fall within this range. Early Wuhan models may under-estimate R_0 by ignoring pre-symptomatic transmission and making restrictive assumptions, including that COVID-19 has the same incubation period and serial interval as SARS-CoV [62, 94, 95], or that most cases are symptomatic [96]. Early Wuhan case data may be insufficient to precisely estimate R_0 without making these assumptions [97, 98, 99]. Thus, models and intervention strategies should consider that the overall R_0 may be higher than 3 in certain locations [63, 100].

If asymptomatic individuals are unlikely to transmit and do so with low probability,

then the small fraction of cases that are symptomatic are transmitting at a high rate, in line with recently reported "super-spreading" events [101, 102]. Super-spreading events are instances in which a single infected individual infects a large number of people. These events can be hard to measure on a population level in the absence of detailed transmission data. In classic super-spreading dynamics, most primary cases do not result in many secondary cases, while a subset of primary cases result in a large number of secondary cases [103, 104, 66]. This heterogeneity in the reproductive number is indeed what we observe when asymptomatic individuals transmit poorly. Our model is admittedly a coarse description of this heterogeneity, since it incorporates only two different classes of infections, symptomatic or asymptomatic. Future models can build upon this framework with additional classes for age, socio-economic status, location or susceptibility [105] using fine-scale case data. These models could elucidate how infections in hospitals or home-care settings may be contributing to the high R_0 of symptomatic cases. However, our results also indicate that even when the symptomatic reproductive number is large, pre-symptomatic and asymptomatic infections contribute together to at least 50% of the overall force of infection.

It follows that community-wide interventions that account for non-symptomatic cases should be crucial for mitigating outbreaks. If asymptomatic cases transmit poorly, then concurrent additional interventions targeting super-spreading symptomatic infections may help reduce community transmission.

Resolving the non-identifiability of the relative strength of asymptomatic transmission (b_{a}), would require extensive community testing and contact tracing of asymptomatic cases. Community testing on its own can provide an estimate of the total proportion of cases that are asymptomatic, but it may not provide insight on whether those asymptomatic individuals can transmit and how well they can transmit. Symptomatic and asymptomatic individuals have similar viral loads [106], but a high viral load does not necessarily imply high transmissibility. One limitation of early contact tracing studies is that estimates of transmissibility may over-sample symptomatic index cases and contacts, particularly during

the early phase of an epidemic [72, 107]. In certain studies, only symptomatic contacts are further investigated. Ideally, one would use frequent systematic community testing for studies identifying both symptomatic and asymptomatic potential index cases for further contact tracing and testing of all contacts regardless of symptoms. Furthermore, fixing the probability that an infection becomes symptomatic based on the results of serology-informed models such as ours, could increase the precision with which contact tracing studies can estimate the strength of asymptomatic transmission. Colleges that are currently re-opening may be ideal test locations for this kind of combined approach, which may also help detect super-spreading events.

While it cannot capture all testing intricacies, our framework illustrates how transmission models can incorporate daily changes in testing capacity and identify parameters that were previously difficult to estimate such as the probability that an infection will become symptomatic. While we do not explicitly denote differences between labs, hospitals, or diagnostic tests, we account for this variation by including additional measurement noise after simulating the RT-PCR testing process. We also consider how sampling individuals without COVID-19 may deplete the daily testing capacity. In particular, hospitalized individuals with non-COVID-19 related severe respiratory disease may have a higher priority for testing than non-severe COVID-19 cases. Our model uses syndrome surveillance reports [108, 23, 109, 110] of respiratory illness from New York City hospitals in previous years, along with weekly influenza cases, to estimate the number of non-COVID-19 severe respiratory cases that were tested. The statistical model assumes that in every year, only a fraction of influenza cases are confirmed and that non-influenza respiratory cases exhibit seasonality. We use flu and syndrome surveillance estimates from previous years to estimate the fraction of influenza cases that are not confirmed and the shape of the seasonality on non-influenza related respiratory illness. During the 2020 epidemic, COVID-19 mitigation measures that reduced urban mobility may have also reduced transmission of other respiratory diseases such as influenza. The model captures some of this decrease, since the number of severe

non-COVID respiratory cases is a function of the number of confirmed flu cases in the same season. The model thus captures the impact of decreased flu cases in 2020 due to changes in mobility patterns. This framework could be used in conjunction with other epidemiological models, and extended to other municipalities or countries with location-specific testing priorities, re-testing procedures, or diagnostic tests.

In cities where mobility information is available, the statistical model may include overall population mobility as a co-variate. In other cities that report the daily number of hospitalized individuals with COVID-19 symptoms who were tested each day, one could subtract from this number the total COVID-19 hospitalizations estimated by the epidemiological model, to obtain the number of non-COVID-19 positive hospitalized cases that were tested. Depending on the information available for each location, future iterations of this framework could explicitly incorporate different diagnostic tests and their respective sensitivities and specificities. It could also be used to examine how altering testing strategies such as switching from symptom-based testing to community testing may improve transmission parameter inference and efficacy of control efforts. This may be an important consideration for countries that have limited testing capacity but are still in the midst of the first pandemic wave, such as India.

Given the potential role of population density and socioeconomic status on contact rates and access to care, there may be considerable heterogeneity in infection rates and seropositivity in different neighborhoods of New York City. While over-dispersion in measurement error can implicitly account for this variation in our implementation, future formulations could do so explicitly with a spatial model of transmission between neighborhoods and within specific settings such as hospitals and home-care networks. This level of resolution would require however observed cases, testing capacity and hospitalizations within each unit. Incorporating human movement estimates into the model could enable analysis of how the infectious period of the virus may impact the clustering of cases within particular neighborhoods.

Future studies can investigate the impact of including a testing sub-model on parameter

estimation and the level of detail required in such a sub-model. For example, one could compare the results of parameter estimation from fitting a given epidemiological model with a queue-based testing model to those that assume a fixed reporting rate and a delay in the reporting of cases. We expect the former to exhibit more uncertainty when informed by surveillance data from the beginning of the pandemic when little testing capacity is available, but to reduce this uncertainty as the time series is extended and this capacity changes. Models that assume a fixed reporting rate may under-estimate the range in uncertainty of epidemiological parameters that are heavily informed by the early part of the time series, and may even under-estimate the values of the parameters themselves. Models with a queue-based testing sub-model may obtain more precise estimates of parameters that impact the end of the outbreak, such as those related to the depletion of susceptible individuals, acquisition of immunity, or in our model, the impact of social distancing and stay-at-home orders on overall transmission. Even if including some form of testing model that takes into account changes in capacity is key to obtaining more precise parameter estimates, simpler versions of our implementation may be sufficient. For example, the more generalizable components such as the testing of hospitalized individuals may be more important than taking into account their re-sampling as they leave the hospital. Simplifying the testing model based on model selection analyses can facilitate wider adaption of the testing framework to other cities, countries, or time periods.

Future versions of the model could also capture heterogeneity in the severity of infection and the acquisition of immunity. When fitting the model, we treat seropositivity as a reasonable correlate of herd immunity. The assay used to measure seroprevalence in New York City [1] elicits neutralizing antibodies [111] and can detect seroconversion in severe, mild, and asymptomatic cases [111]. We assume that this herd immunity does not decay over the course of the study, given the short duration of time from March to June 2020 and the observation that antibody responses can persist for at least 5 months following the start of the pandemic in New York City [112]. The immune response to the SARS-Cov virus

consists of several components including antibody [113] and T-cell mediated [114] responses, and heterogeneity in particular pathways of the immune response can influence the severity of infections [115, 116]. The severity of infection may in turn impact the type, strength and duration of the immunity acquired [117, 118]. As future experimental studies determine how each immune response can mitigate infection, viral shedding, and transmission, relevant aspects of the dynamics of host immunity can be incorporated into the model and corroborated with data. Understanding how host heterogeneity in immune responses may impact the infection severity and herd immunity may be valuable when considering long-term vaccination policies.

While a model with explicit within-host dynamics would be challenging to fit using case data, relevant aspects could be incorporated in several ways. For example the model could include additional classes of infection corresponding to levels of severity. Alternatively, a distribution of susceptibility or immunity could be used to capture heterogeneity in the immune response between individuals. Finally, the model could incorporate functional forms for the acquisition and waning of immunity that are fixed based on experimental observations of serology and T-cell dynamics. Fitting these models to times series from multiple locations will improve inference, but the testing capacity and strategy in each location should be taken into account when doing so.

Our finding that many individuals were already infected by March 1st is consistent with earlier estimates that community transmission began in February in NYC [23, 119, 81]. Previous studies could not explain however why no substantial increase in COVID-19-like illness was observed prior to February 28th in syndrome surveillance data [23]. Our simulations show that the lag between infection onset and hospitalization can explain this discrepancy. Even when initialized with many infected cases on March 1st , simulated hospitalizations do not rise until several weeks later concurrent with observed COVID-19 hospitalizations (Fig 2.5) Most likely, the estimated initial conditions suggest multiple parallel foci of initiation of the epidemic with multiple importations of infections. Another suggested possibility

is a dosage-dependence effect, wherein the severity of an individual's infection depends on the size of the virus population that the person becomes infected with during one or more transmission events, and hence on the overall viral load of COVID-19 in the community. In this scenario, early COVID-19 cases in February and early March would be less severe. This would be consistent with the syndrome surveillance data, where we see a rise in early March of respiratory infection reports in the emergency departments of hospitals, but do not yet see a rise in COVID-19 hospitalizations. This phenomenon might also explain why our model slightly under-estimates the peak in daily hospitalizations, even though it correctly identifies the time and shape of that peak.

We show that testing capacity and strategy can substantially affect estimates of the infection fatality ratio, a quantity that is frequently used by public health officials in assessing the severity of an outbreak. Our model ignores several factors such as non-hospital deaths from COVID-19, which may increase the true IFR, and rising trends in hospital capacity and improved treatments, which may decrease it. Nevertheless, our results underscore the importance of considering testing strategy and capacity when interpreting literature estimates of the infection fatality ratio.

In conclusion, explicit consideration of changes in testing capacity allows us to infer with certainty from case and serology data that most new COVID-19 cases do not become symptomatic. We also inferred that the overall or symptomatic reproductive number may be larger than often assumed depending on how well asymptomatic cases can transmit. Despite this uncertainty, the strong consistent contribution to community transmission from cases without symptoms observed across scenarios supported by the data, should be considered when formulating public health intervention strategies. Making available detailed information on testing policy and data on testing capacity over time will strengthen the ability of epidemiological models to learn from the past and inform us about the future.

2.3 Methods

We examine three different model structures that have been used to characterize COVID-19 dynamics in previous studies (Fig 2.1). All models are modified versions of the traditional susceptible-exposed-infected-recovered (SEIR) model [120]. The first model, the SEPIR model [74, 121], is the most standard extension in which transitions are between a linear chain of compartments. Its formulation adds a compartment P for pre-symptomatic transmission. The second one, the SEIAR model [65, 20], differs conceptually in that it includes asymptomatic individuals rather than pre-symptomatic ones, and defines them as distinct, in the sense that they will never transition to exhibiting symptoms. This definition implicitly recognizes that there are essentially two classes of individuals in terms of susceptibility to disease and symptoms. The third structure for the SEPIAR model [80, 64] is a combination of the first two and includes them as nested, particular, cases.

All three models include a chain of m exposed classes to incorporate the total time between the onset of infection and the onset of symptoms as gamma distributed (with mean 5.5. days and standard deviation 2.25 days) [122]. Symptomatic individuals are subdivided into two sequential classes I_{S_1} and I_{S_2} for practical purposes, to follow their numbers before and after some of them transition to hospitalization. Individuals spend an average of $\frac{1}{\phi_S}$ days in I_{S_1} and $\frac{1}{\gamma}$ days in I_{S_2}.

The parameter R_0 represents the reproductive number experienced by symptomatic individuals. We define a baseline pre-intervention transmission rate in symptomatic individuals β_0 by dividing R_0 by the average total time that non-severe cases transmit with symptoms. We also define a post-intervention transmission rate β_1, which is equal to the pre-intervention transmission rate β_0 multiplied by a scaling factor b_{q}. Low values of b_{q} represent a substantial reduction in the transmission rate due to interventions. Social distancing guidelines were issued by New York City starting on March 17 [123, 56], and a stay-at home order was issued which took effect on the evening of March 22 [124]. Thus,prior to the imposition of social

distancing, the transmission rate of symptomatic individuals in our models, $\beta(t)$, is equal to β_0. From March 18th thru March 22nd, $\beta(t)$ decreases linearly from β_0 to β_1. From March 23rd onwards, $\beta(t)$ is equal to β_1.

In all models, a fraction p_S of exposed individuals E_m becomes symptomatic. After an average of 5 days of transmission, symptomatic cases are hospitalized with probability p_{H}. Symptomatic cases that are not severe enough to require hospitalization recover at rate γ. Hospitalized individuals recover at rate $h_v = \frac{1}{13}$ [84] and do not transmit while hospitalized. In practice, some individuals with severe COVID-19 symptoms that required hospitalization may not have been hospitalized due to barriers to care. However, their contribution to community transmission is unlikely to have been substantial. These individuals would remain isolated at home during the period of severe infection and avoid non-household contacts, while household contacts would have been exposed for a substantial time prior to the onset of severity. We assume a fixed population size for New York City of 8 million individuals [125].

Assumptions about which infected classes are infectious and how they contribute to the transmission rate allow us to reduce the SEPIAR model to the SEPIR or SEIAR models. Pre-symptomatic individuals transmit for an average of about a day (0.92 days [79]) at a transmission rate equal to the baseline transmission rate $\beta(t)$ multiplied by a scaling factor b=b_{p}. Asymptomatic infections transmit for an average of 5 days, equal to the average duration between the onset of symptoms and hospitalization in severe cases, at a transmission rate equal to the baseline rate $\beta(t)$ multiplied by scaling factor b_{a}.

The models are implemented numerically via an Euler approximation of the deterministic equations to which demographic stochasticity is added. Specifically, the number of individuals making state transitions from compartments with more than one exit is drawn from an Euler-multinomial distribution [17]. The number of individuals making state transitions from compartments with only one exit is drawn from a binomial distribution.

Description of Testing Model: The model takes into account daily changes in the

testing capacity using estimates of daily tests conducted in New York City from the New York State Department of Health [22], as well as the re-testing of severe and non-severe symptomatic cases prior to leaving the hospital or quarantine. We assume that there are two categories of cases-severe (hospitalized) cases and non-severe cases subject to different testing priorities [126]: the initial testing of new hospitalized COVID-19 cases (highest priority), the re-testing of those individuals when they leave the hospital, the testing of new non-severe symptomatic COVID-19 cases, and finally the re-testing of those symptomatic cases (lowest priority). All severe COVID-19 cases after March 1st are sampled when they enter the hospital and eventually tested once enough capacity is available. We assume that symptomatic non-severe cases are sampled at the same time in the course of their infection as severe cases. However, we assume that they are not tested if they recover before enough testing capacity is available. During the early stages of the epidemic, the CDC recommended test-based strategies to determine when to conclude home isolation or hospitalization [61]. Accordingly, we assume that hospitalized cases are re-tested twice (over a 24 hour period) after the average length of time in the hospital (13 days), while non-severe cases are likewise re-tested twice after the end of a 14-day quarantine period.

We also take into account the potential for non-COVID-19 severe respiratory cases to be sampled in hospitals and tested (with the same priority as hospitalized COVID-19 cases). We use confirmed influenza cases [127] and syndrome surveillance reports of respiratory disease from emergency departments in New York City hospitals in previous years [128] to estimate the number of non-COVID-19 severe respiratory cases that may have been sampled (see SI Appendix). We assume that the RT-PCR testing has a sensitivity of 90% [129], that testing takes 48 hours [130], and that there is an additional negative-binomial distributed dispersion after the RT-PCR testing with standard deviation σ_{M}. This dispersion takes into account variation in sampling and testing protocols across laboratories and hospitals, as well as variation in the sensitivity and time required for different PCR assays.

Overview of the model fitting and inference strategy. Unless otherwise mentioned,

we fit the following parameters: the recovery rate for non-severe symptomatic infections (γ), the scaling factors for asymptomatic, pre-symptomatic, and post-intervention transmission (b_{a}, b_{p}, and b_{q}), the symptomatic probability (p_{S}) and the hospitalization probability (p_{H}), the reproductive number for symptomatic cases (R_0), the dispersion parameter for RT-PCR testing (σ_M), and the initial number of infected (I_0) and exposed (E_0) individuals at the start of the simulation on March 1, 2020. We use the iterated filtering algorithm MIF [131] within the R-package POMP (for partially observed Markov process models) to fit parameter combinations by likelihood maximization. The iterated filtering algorithm is specifically designed for fitting stochastic and nonlinear models with hidden variables in the presence of both process and measurement error. We apply the Sequential Monte Carlo algorithm pfilter [47] to evaluate the likelihood of the final parameter combinations obtained with the computational search. Likelihoods are estimated by simulating state variables at each observation time from an underlying Markov process model, and then calculating the likelihood of each observation given the simulated value of the state variable and a measurement model. For the analysis of the full SEPIAR model, we generate a Monte Carlo profile [132] for the relative strength of asymptomatic transmission (b_{a}).

For all resulting parameter combinations within 2 log–likelihood units of the MLE, we then calculate the likelihood with respect to serology using seroprevalence data previously published by [1] from a screening group representative of the general population using plasma samples from patients at Mount Sinai Hospital in New York City. In the Mount Sinai study, random, de-identified, and cross-sectional samples were obtained over the course of the outbreak from patients at OBGYN visits and deliveries, oncology-related visits, as well as hospitalizations due elected or planned surgeries, transplant surgeries, pre-operative medical assessments and related outpatient visits, cardiology office visits, or other regular office or treatment visits whose purpose was unrelated to COVID-19 [1]. The assay used to measure seroprevalence [1] elicits neutralizing antibodies [111] and can detect seroconversion in severe, mild, and asymptomatic cases [111]. We treat the seroprevalence measurement at each

time point as a measure of short-term herd immunity in the population, specifically of the proportion of the population that has already recovered from COVID-19 infection. We assume that this herd immunity does not decay over the course of the study, since antibodies have been shown to persist for at least 5 months [112]. We compare the seroprevalence at each time point from the serology data to the recovered fraction of the population $\frac{R}{N}$ from simulated trajectories of the epidemiological model. When calculating the likelihood of each trajectory at each observation time with respect to the seroprevalence data, we assume that the number of people who seroconverted in the Mount Sinai study is drawn from a Binomial distribution with p equal to the value of $\frac{R}{N}$ in the simulated trajectory at that time, and N equal to the total number of people sampled. We sum the log-likelihoods across all observation time points and then average over all trajectories using the logmeanexp function in the R package POMP [47] to obtain a log-likelihood for each parameter combination with respect to the serology data. We isolate all combinations supported by the serology data that have log-likelihoods within 2 units of the MLE.

For each combination, we examine the proportion of cases that are symptomatic p_S, the reproductive number in symptomatic individuals R_0, and the overall reproductive number for the model $R_{0_{\text{NGM}}}$. We derive the following expression for $R_{0_{\text{NGM}}}$ using the Next Generation Matrix [133] :

$$R_{0_{\text{NGM}}} = \frac{\beta * b_p}{\phi_U} + \frac{\beta * b_a(1-p_S)}{\phi_S} + \frac{\beta p_S}{\phi_S} + \frac{\beta(1-p_{\text{H}})p_S}{\gamma} \tag{2.1}$$

Calculation of the Infection Fatality Ratio (IFR). The Infection Fatality Ratio or IFR is frequently defined as the ratio of deaths to cases. Let IFR_{all} represent the IFR with respect to all cases:

$$\text{IFR}_{\text{all}} = \frac{\text{Confirmed deaths}}{\text{All cases}} \tag{2.2}$$

This is equivalent to the proportion of all cases that result in death. We can estimate this quantity using the parameters from the fitted SEPIAR model. Recall that p_{S} is the probability that a case is symptomatic, and p_{H} is the probability that a symptomatic case becomes hospitalized. These quantities are equivalent to the proportion of all cases that are symptomatic, and the proportion of symptomatic cases that are hospitalized. Let p_F represent the proportion of all hospitalizations that are fatal. Since this parameter is not fit in the SEPIAR model, we estimate it from observed data by dividing the total number of confirmed COVID-19 deaths by the total number of confirmed COVID-19 hospitalizations in New York City during the study period. We use data updated on June 15, 2020 from the New York City Health Department COVID-19 Data Portal [78], and obtain an estimate of $p_F = 0.33$. We can then write an expression for $\mathrm{IFR}_{\mathrm{all}}$ using the parameters estimated with the fitted SEPIAR model:

$$\mathrm{IFR}_{\mathrm{all}} = p_{\mathrm{S}} p_{\mathrm{H}} p_{\mathrm{F}}, \tag{2.3}$$

and obtain this quantity for each parameter combination supported by case and serology data (red histogram in SI-Appendix Fig. 12).

Let $\mathrm{IFR}_{\mathrm{symp}}$ represent the IFR that would be estimated if all symptomatic cases were observed but no asymptomatic ones were observed. This is equivalent to the proportion of all symptomatic cases that result in death:

$$\mathrm{IFR}_{\mathrm{symp}} = \frac{\text{Confirmed deaths}}{\text{All symptomatic cases}} = p_{\mathrm{H}} p_{\mathrm{F}} \tag{2.4}$$

(shown in the blue histogram of Fig. 17).

We compare these two quantities with the observed IFR, calculated by dividing the total deaths by the total number of PCR confirmed COVID cases in NYC during the study period

(the orange line in Fig. 17).

Additional details: Further details of the SEPIAR equations, testing model, Monte Carlo Profile of the SEPIAR model, initial grid searches and model comparison of the SEPIR and SEIAR models, and derivation of the overall reproductive number $R_{0_{\mathrm{NGM}}}$, are provided in the SI Appendix.

2.4 Funding

R.S. was supported by a National Science Foundation Research Traineeship (no. 1735359: NRT-INFEWS: Computational data science to advance research at the energy environment nexus).

2.5 Acknowledgments

The authors would like to thank Aaron King for his insightful discussions. This work was completed with resources and support provided by the University of Chicago's Research Computing Center.

2.6 Supporting Information

2.6.1 SEPIAR Model Details

SEPIAR Model Compartments

ODE Equations

$$\frac{dS}{dt} = -(\lambda_{\mathrm{FOI}}(t))S(t) \tag{2.5}$$

For exposed compartment E_m where $m = 1$:

$$\frac{dE_1}{dt} = (\lambda_{\mathrm{FOI}}(t))S(t) - \phi_{\mathrm{E}}E_1(t) \tag{2.6}$$

Compartment	**Infection Status**
$S(t)$	Susceptible
$E_m(t)$	Exposed in compartment m
$P(t)$	Pre-symptomatic
$A(t)$	Asymptomatic
$I_{\mathrm{S}_1}(t)$	Infected symptomatic not yet sampled
$I_{\mathrm{S}_2}(t)$	Infected symptomatic (non-severe)
$H(t)$	Hospitalized (severe infected)
$R_{\mathrm{A}}(t)$	Recovered (Asymptomatic)
$R_{\mathrm{F}}(t)$	Recovered (Symptomatic Non-Severe)
$R_{\mathrm{H}}(t)$	Recovered (Severe)

Table 2.1: Comparmtents in the SEPIAR epidemiological model

For exposed compartment E_m where $1 < m <= M$:

$$\frac{dE_m}{dt} = \phi_{\mathrm{E}} E_{m-1}(t) - \phi_{\mathrm{E}} E_m(t) \tag{2.7}$$

$$\frac{dP}{dt} = \phi_{\mathrm{E}} E_M(t) - \phi_{\mathrm{U}} P(t) \tag{2.8}$$

$$\frac{dI_{\mathrm{S}_1}}{dt} = p_{\mathrm{S}} \phi_{\mathrm{U}} P(t) - \phi_{\mathrm{S}} I_{\mathrm{S}_1}(t) \tag{2.9}$$

$$\frac{dH}{dt} = p_{\mathrm{H}} \phi_{\mathrm{S}} I_{\mathrm{S}_1}(t) - h_{\mathrm{V}} H(t) \tag{2.10}$$

$$\frac{dI_{\mathrm{S}_2}}{dt} = (1 - p_{\mathrm{H}}) \phi_{\mathrm{S}} I_{\mathrm{S}_1}(t) - \gamma I_{\mathrm{S}_2}(t) \tag{2.11}$$

$$\frac{dA}{dt} = (1 - p_{\mathrm{S}})\phi_{\mathrm{U}} P(t) - \phi_{\mathrm{S}} A \tag{2.12}$$

$$\frac{dR_{\mathrm{A}}}{dt} = \phi_{\mathrm{S}} A \tag{2.13}$$

$$\frac{dR_{\mathrm{F}}}{dt} = \gamma I_{\mathrm{S}_2}(t) \tag{2.14}$$

$$\frac{dR_{\mathrm{H}}}{dt} = h_{\mathrm{V}} H(t) \tag{2.15}$$

$$\lambda_{\mathrm{FOI}}(t) = \frac{\beta(t)[(I_{\mathrm{S}_1}(t)) + (I_{\mathrm{S}_2}(t))] + \beta_{\mathrm{a}}(t)[A(t)] + \beta_{\mathrm{p}}(t)[P(t)]}{N} \tag{2.16}$$

Accumulator Variables

Let C_{Q1} represent the total number of individuals with severe COVID-19 cases who enter the hospital over a single-day period. In the SEPIAR model, this is the number of people moving from compartment I_{S_1} to H in a single day.

We assume that non-severe COVID-19 cases are sampled at the same time in the course of their infection as severe cases, provided that sufficient testing capacity is available. Let C_{Q3} represent the total number of people who move from compartment I_{S_1} to compartment I_{S_2} over a single day in the SEPIAR model. These people represents symptomatic COVID-19 cases that do not become severe.

The quantities C_{Q1} and C_{Q3} generated from the epidemiological model are used as inputs

for the testing model.

2.6.2 Model Fitting Techniques

Fitted Parameters

Unless otherwise mentioned, we fit the recovery rate for non-severe symptomatic infections (γ), the scaling factors for asymptomatic and pre-symptomatic transmission (b_{a} and b_{p}), the scaling factor for post-intervention transmission (b_{q}), the proportion of new infected cases that will become symptomatic (p_{S}), the proportion of symptomatic cases that are severe enough to require hospitalization (p_{H}), the reproductive number for symptomatic cases (R_0) the dispersion parameter for RT-PCR testing (σ_{M}), and the initial number of infected (I_0) and exposed (E_0) individuals at the start of the simulation on March 1, 2020. We constrain the fitting algorithm to explore only positive values for all fitted parameters and only values between 0 and 1 for p_{S}, p_{H}, b_{a}, b_{p}, and b_{q}.

Initial Grid Searches of SEPIR and SEIAR Models

We first fit the SEPIR model, which does not have asymptomatic transmission, and the SEIAR model, which does not have pre-sympamatic tranmission (Figure 12.6). For each model, we generate a grid of 25,000 initial parameter combinations using Latin Hypercule Sampling. For each initial combination, the given model is fit to observed case data for two sets of 50 iterations using the iterated filtering algorithm MIF2 [131] using 50,000 particles. The final likelihood of each parameter combination with respect to observed case data is then estimated using the sequential Monte Carlo algorithm pfilter [47]. We then isolate all parameter combinations within 2 log-likelihood units of the parameter combination with the highest likelihood (the maximum likelihood estimate or MLE), and calculate the likelihood of each combination with respect to the serology data. We then isolate the parameter combination with the highest likelihood with respect to the serology data.

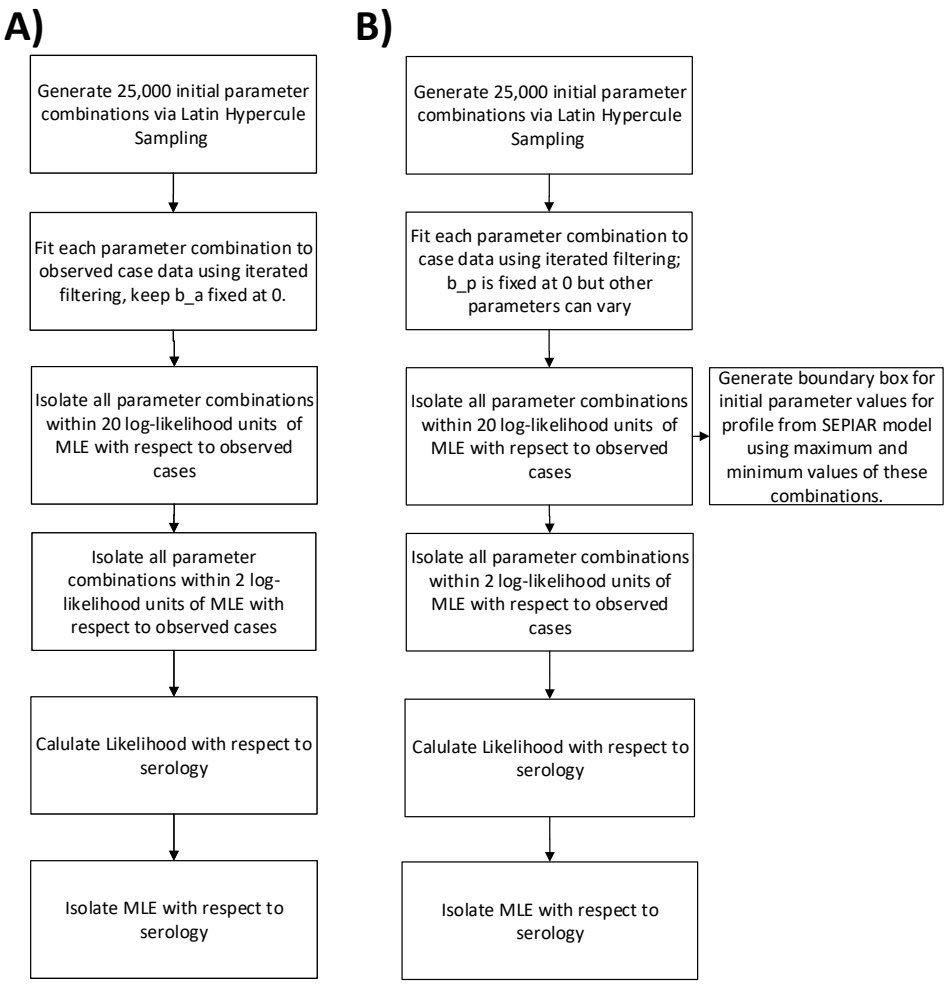

Figure 2.6: **Diagram of the grid searches for the SEPIR (A) and SEIAR (B) models.**

Monte Carlo Profile of SEPIAR Model

For the analysis of the full SEPIAR model, which includes both pre-symptomatic and asymptomatic transmission, we construct a Monte Carlo Profile[132] of the relative strength of asymptomatic transmission (b_{a}). As Figure 2.7 illustrates, we generate a set of starting points at 30 different evenly spaced values for b_{a} between 0 and 1. For each of those 30 starting points, we create 40 different initial sampling points with the same value of b_{a} but different values for the other parameters being fitted. Initial values for the relative strength of pre-symptomatic transmission b_p were drawn from a uniform distribution between 0 and 1. The values for all of the other fitted parameters were uniformly drawn from the boundaries of all parameter combinations obtained from fitting the SEIAR model that had likelihoods with respect to case data within 20 log-likelihood units of the MLE. This yielded a total of 1200 starting points. Each profile starting point was then fit to case data using the iterated filtering algorithm MIF2 [131] and the Sequential Monte Carlo algorithm pfilter [47], with MIF2 constrained to keep b_{a} unchanged. For all parameter combinations that were supported by observed case data (i.e. that had log-likelihoods within 2 units of the MLE), we then calculated the likelihood with respect to serology. All parameter combinations from the full SEPIAR model with serology likelihoods within 2-log-likelihood units of the MLE were used in subsequent analyses of the proportion of cases that are symptomatic (p_{S}), the reproductive number in symptomatic individuals (R_0), and the overall reproductive number for the model ($R_{0_{\mathrm{NGM}}}$) which was calculated using the Next Generation Matrix [133].

Model Comparison

We compare the likelihoods with respect to the serology data of all the maximum likelihood estimates from the SEPIR, SEIAR, and SEPIAR models via the Likelihood Ratio Test. Recall that when calculating the likelihood with respect to case data, all three models had maximum likelihoods that were not statistically different.

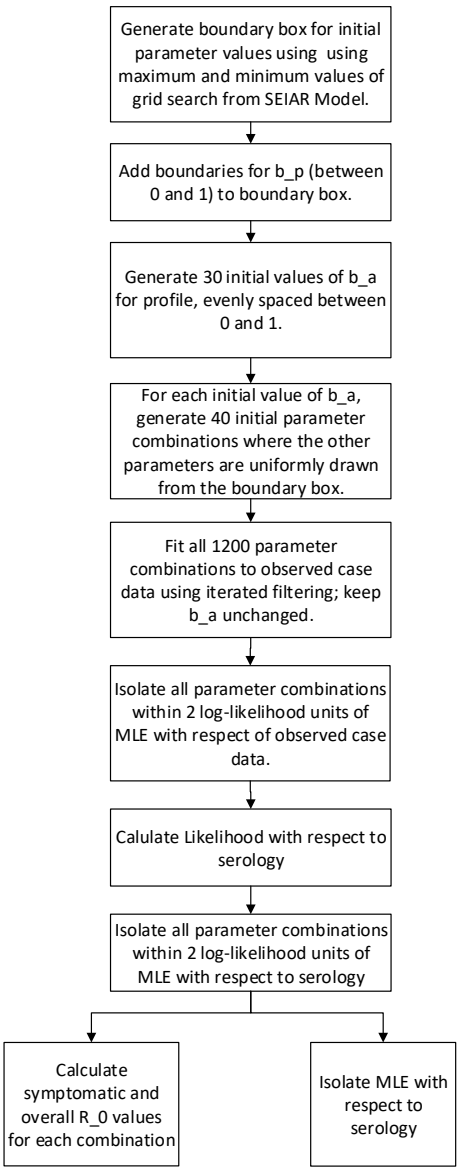

Figure 2.7: **SEPIAR Profile Fitting Procedure**. This diagram summarizes how the Monte Carlo profile of b_{a} for the SEPIAR model was fit to case data and subsequently to serology.

Model	**Log-Likelihood**
SEPIAR	-24.34239**
SEIAR	-24.6274**
SEPIR	-29.17705

Table 2.2: Comparison of MLE Likelihoods from SEPIR, SEIAR, and SEPIAR models with respect to serology data

2.6.3 Testing Model

The testing model is implemented with discrete time steps of a day, denoted hereafter by $\mathbf{t}$.

Testing Capacity Data

We use the total number of RT-PCR tests conducted each day across the five counties compromising New York City as measured by the New York State Health Department's COVID tracker as the daily testing capacity for the city. This capacity is fed into the model as a co-variate ($L_{\text{reported}}(\mathbf{t})$).

Let $L_0(\mathbf{t})$ represent the initial testing capacity on day $\mathbf{t}$. Recall that we assume it took 2 days to conduct a RT-PCR Test. Therefore, we advance the testing capacity by 2 days from the observed value, since the testing capacity on day $\mathbf{t}$ will correspond to the number of tests conducted on day $\mathbf{t}+2$.

$$L_0(\mathbf{t}) = L_{\text{reported}}(\mathbf{t}+2) \tag{2.17}$$

2.6.4 Testing Priorities

Given a finite daily testing capacity, we assume that tests are used in the following order during any given day until the testing capacity is exhausted:

1. **Initial testing of hospitalized patients.** Hospitalized patients include patients who have severe cases of COVID-19 (Queue 1) as well as individuals who do not have COVID-19 but are hospitalized with respiratory symptoms (Queue NC). Patients are added to queue 1 as they enter compartment H in the epidemiological model.

2. **Re-testing hospitalized patients when they leave the hospital (Queue 2).** Patients hospitalized with COVID-19 are re-tested twice over a 24 hour period as they exit the hospital.

3. **Initial testing of non-severe symptomatic cases (Queue 3)**.Patients are added to queue 1 as they enter compartment I_{S_2} in the epidemiological model.

4. **Re-testing of non-severe symptomatic cases 14 days after they are first sampled (Queue 4).**

We assume that any remaining unused testing capacity remaining after all four queues have been tested is used to test individuals without COVID-19. Thus, this unused capacity does not roll over to the next day.

Supporting Figure S2.9 illustrates these testing priorities.

General framework for incorporating changes in testing capacity

Below we provide the detailed steps of the general testing framework which can be applied to other locations where testing capacity is changing over time. A diagram of this framework is shown in Supporting Figure S2.8. We illustrate the steps by focusing on the testing of hospitalized severe COVID-19 cases. In subsection 2.6.4, we describe several modifications to this framework that we make to take into account additional queues and modifications that are specific to New York City in the early stage of the epidemic.

We use the variable Q_1 to represent all COVID-19 hospitalized cases that have been sampled but have not been tested yet. We first add the hospitalized COVID-19 cases that have just been sampled on day $\mathbf{t}$ ($C_{Q1}(\mathbf{t})$) to Q_1.

$$Q_{1_1}(\mathbf{t}) = Q_{1_1}(\mathbf{t}) + C_{Q1}(\mathbf{t}) \tag{2.18}$$

Let the variable T_{Q1} represent all of the people in Queue 1 who will be tested. If $L_0(\mathbf{t})$ is bigger than the Queue 1, then everyone in Queue 1 can be tested and

$$T_{\mathrm{Q1}}(\mathbf{t}) = Q_1(\mathbf{t}) \tag{2.19}$$

Let $L_{\mathrm{unused}}(\mathbf{t})$ represent the testing capacity left over after Q1 has been tested:

$$L_{\mathrm{unused}}(\mathbf{t}) = L_0(\mathbf{t}) - T_{\mathrm{Q1}}(\mathbf{t}) \tag{2.20}$$

There are then no individuals left in Q1 who need to be tested.

$$Q_1(\mathbf{t}) = 0 \tag{2.21}$$

However, suppose that on a given day there is not enough testing capacity to test everyone in Q1 (i.e. at the start of that day, $L_0(\mathbf{t}) < Q_1(\mathbf{t})$).

In this case, we assume that all of the testing capacity is used to test Q1:

$$T_{\mathrm{Q1}}(\mathbf{t}) = L_0(\mathbf{t}) \tag{2.22}$$

We therefore decrease Queue 1 by the number of people who were tested.

$$Q_1(\mathbf{t}) = Q_1(\mathbf{t}) - T_{\mathrm{Q1}}(\mathbf{t}) \tag{2.23}$$

In this case, there is no unused testing capacity:

$$L_{\mathrm{unused}}(\mathbf{t}) = 0 \tag{2.24}$$

Let $Y_{\mathrm{Q1}}(\mathbf{t})$ represent the number of people who tested positive on day $\mathbf{t}$, and c the sensitivity of the RT-PCR assay.

We simulate RT-PCR testing as a draw from a Binomial distribution where

$N = T_{\mathrm{Q1}}(\mathbf{t})$ and $p = c$.

$$Y_{\mathrm{Q1}}(\mathbf{t}) \sim Binomial(T_{\mathrm{Q1}}(\mathbf{t}), c) \tag{2.25}$$

If there is unused testing capacity after everyone in Queue 1 has been tested (i.e. $L_{\mathrm{unused}}(\mathbf{t}) > 0$, then this capacity can be used to test individuals in other queues with a lower priority; see Section 2.6.4).

If testing of any of the individuals in these later queues results in additional new positive cases, these are added to $Y_{\mathrm{Q1}}(\mathbf{t})$ to obtain the total number of expected new COVID-19 cases during day $\mathbf{t}$, denoted by $Y_{\mathrm{sum}}(\mathbf{t})$.

NYC Specific Modifications and Additional Queues

There are several aspects of the testing model that are particular to New York City and the initial wave of the epidemic in the U.S. We summarize those aspects here. With sufficient description of the specific testing implementation, similar modifications could be considered for other locations.

- **Re-Sampling of Hospitalized Cases-** We assume that hospitalized COVID-19 cases will be re-sampled twice over a 24-hour period as they leave the hospital, to make sure that they have recovered. We take this into account with a second queue, Queue 2 ($Q_2(\mathbf{t})$), which represents patients who tested positive who had a severe COVID-19 infection, were re-sampled as they left the hospital, and need to be tested again. We do not simulate RT-PCR testing, keep track of results of testing, or deal with lags for Queue 2, since the RT-PCR results of this re-sampling are not relevant for our model.

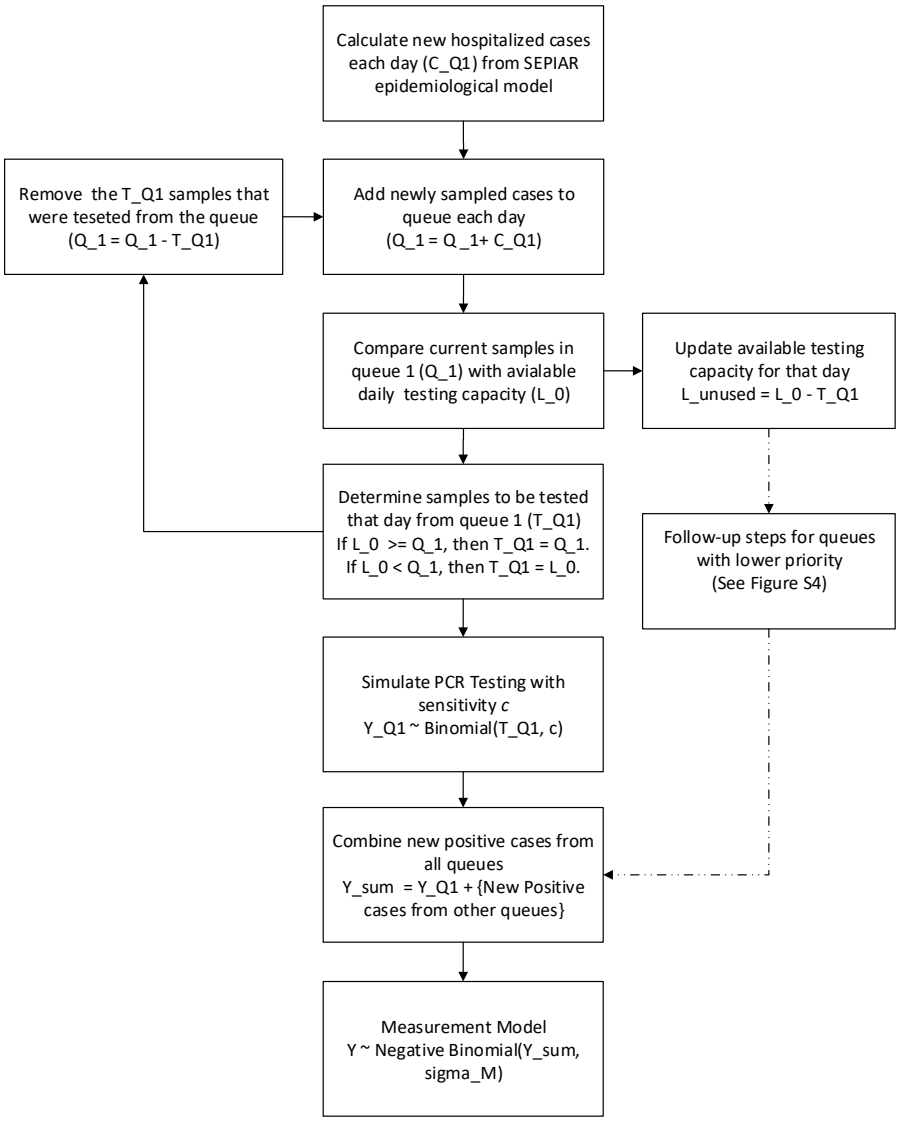

Figure 2.8: **Diagram of the general testing framework described in Section 2.6.4**. The New York City-specific modifications described in Section 2.6.4 are not shown here.

What is relevant is that at least some of the available testing capacity $L_{\text{unused}}(\mathbf{t})$ may be used up re-testing hospitalized patients as they leave the hospital before other groups such as non-severe symptomatic patients can be tested.

In our epidemiological model, we assume that individuals spend an average of 13 days in the hospital. To be consistent, in our testing model, we assume that individuals who test positive for COVID-19 will be re-sampled once 13 days after they enter Queue 1, and then re-sampled a second time 1 day later. To keep track of the days since each sample entered Queue 1, we modify our implementation of Queue 1 by introducing initial sampling cohorts. There are thirteen cohorts, representing people who entered Queue 1 up to 13 days before time $\mathbf{t}$. Let $T_{\text{Q1}_v}(\mathbf{t})$ represent the number of people who were sampled v days before day $\mathbf{t}$ who will be tested on day $\mathbf{t}$. We calculate $T_{\text{Q1}_v}(\mathbf{t})$ as we loop through each cohort v in Queue 1, Q_{1_v}, starting with the oldest (Q_{1_V}) and ending with the most recent (Q_{1_1}).

For example, if there is sufficient capacity to test everyone in the oldest cohort (i.e. $L_0 > Q_{1_V}(\mathbf{t})$), then we essentially follow a similar procedure to equations 2.19 and 2.20:

$$T_{\text{Q1}_V}(\mathbf{t}) = Q_{1_V}(\mathbf{t}) \tag{2.26}$$

$$L_{\text{unused}}(\mathbf{t}) = L_0(\mathbf{t}) - T_{\text{Q1}_V}(\mathbf{t}) \tag{2.27}$$

We similarly decrease $L_{\text{unused}}(\mathbf{t})$ as the capacity is used up from testing each subsequent cohort. For example, suppose that there is enough capacity to test a subsequent

cohort v (i.e. $L_{\text{unused}}(\mathbf{t}) > Q_{1_v}(\mathbf{t})$). Then:

$$T_{\text{Q1}_v}(\mathbf{t}) = Q_{1_v}(\mathbf{t}) \tag{2.28}$$

$$L_{\text{unused}}(\mathbf{t}) = L_{\text{unused}}(\mathbf{t}) - T_{\text{Q1}_v}(\mathbf{t}) \tag{2.29}$$

Alternatively, when we do not have enough testing capacity to test all of the cohort, then:

$$T_{\text{Q1}_v}(\mathbf{t}) = L_{\text{unused}}(\mathbf{t}) \tag{2.30}$$

We loop through all cohorts in Queue 1 until either all the people in the queue have been tested or until the unused testing capacity is exhausted. When simulating RT-PCR testing, we again loop over all sampling cohorts v from $1 : V$. Equation 2.25 is modified accordingly:

$$Y_{\text{Q1}_v}(\mathbf{t}) \sim Binomial(T_{\text{Q1}_v}(\mathbf{t}), c) \tag{2.31}$$

For the first re-sampling, we need to keep track of those individuals who tested positive when they first entered the hospital. Because different cases form the same cohort may be tested on different days, we need a variable to accumulate those cases that tested positive and belong to the same cohort. Let $P_{\text{Q1}_v}(\mathbf{t})$ represent all cases from Queue 1 initially sampled $v - 1$ days before time $\mathbf{t}$ who have tested positive so far.

We accumulate the total number of people in initial sampling cohort v who have tested

positive so far by adding the number of people from that cohort who tested positive on day $\mathbf{t}$ ($Y_{\mathrm{Q1}_v}(\mathbf{t})$) to $P_{\mathrm{Q1}_v}(\mathbf{t})$:

$$P_{\mathrm{Q1}_v}(\mathbf{t}) = P_{\mathrm{Q1}_v}(\mathbf{t}) + Y_{\mathrm{Q1}_v}(\mathbf{t}) \tag{2.32}$$

For the first re-sampling, the oldest cohort is entered into $Q_2(\mathbf{t})$:

$$Q_2(\mathbf{t}) = Q_2(\mathbf{t}) + P_{\mathrm{Q1}_V}(\mathbf{t}) \tag{2.33}$$

For the second re-sampling, we keep track of the individuals in this oldest cohort and enter them into Q2 again one day later (on day $t + 1$).

At the end of each simulation day, all initial sampling cohorts are advanced by 1 day.

- **2-Day Lag in Testing-** We incorporate a 2-day lag between when tests are conducted and results are reported to take into account the 48 hour testing time of early RT-PCR tests. We do not update cohorts during this lag, so this effectively adds another 2 days between the initial sampling and the re-sampling beyond the 13 days spent in the hospital. If this framework is applied to other locations and time periods, this modification may not be needed, particularly if rapid diagnosis RT-PCR tests are in use.

- **Testing of non-symptomatic severe cases-**Queue 3, which records the number of non-severe symptomatic cases that need to be tested, is implemented identically to Queue 1, except that individuals can exit the queue at rate γ as they recover, even if they have not yet been tested. Sampling cohorts are used in this queue as well.

- **Re-testing of non-symptomatic severe cases-** Early CDC guidelines [61]) recom-

mended a 14-day quarantine for non-hospitalized symptomatic individuals, and that these individuals be re-tested at the end of the quarantine. Queue 4, which records the re-sampling of non-severe symptomatic cases as they exit quarantine, is implemented identically to Queue 2, except that cases are re-sampled 14 days after they enter Queue 3, rather than 13 days, to match the length of the quarantine period.

- **Queues are numbered in order of priority-** Any unused testing capacity after Queue 1 is empty is applied to Queue 2, and subsequently Queues 3 and 4.

- **Severe non-COVID hospitalized cases-** Queue NC, which represents severe non-COVID-19 respiratory cases, is implemented in a similar fashion as Queue 1. Let $G_{\text{severe}}(w, y)$ represent our estimate of the weekly expected severe non-COVID-19 respiratory cases that would be sampled for testing in the hospital, where w denotes the week, and y, the year. This estimate is obtained from syndrome surveillance data as described later in Section (2.6.5). The number of daily new cases that enter Queue NC, $C_{\text{Q}_{\text{NC}}}$, is therefore:

$$C_{\text{Q}_{\text{NC}}} = \frac{G_{\text{severe}}(w, y)}{7} \tag{2.34}$$

Queue NC has the same priority as Queue 1, since both groups of individuals will present with severe respiratory symptoms at the time of sampling. Both groups are hence sampled at once. For the situation where testing capacity is greater than the total samples in a given sampling cohort in Queue 1 and Queue NC, we simply test all samples from that cohort in both queues. For the situation where the testing capacity is limiting, we simulate a draw from a hypergeometric distribution. For example, in the model, we modify equation 2.30 as follows:

$$T_{\text{Q1}_v}(\mathbf{t}) \sim hypergeom(Q_{1_v}(\mathbf{t}), Q_{\text{NC}_v}(\mathbf{t}), L_{\text{unused}}(\mathbf{t})) \tag{2.35}$$

We assume that the RT-PCR test is 100% specific, so all cases in Queue NC will test negative since they do not have COVID-19. We thus do not simulate RT-PCR testing for Queue NC. The main importance of individuals in Queue NC is that they deplete some of the testing capacity that would otherwise be used for Queue 1.

Measurement Model

Let Y_{sum} represent the total number of new positive tests from all of the queues.

We assume that there is additional negative-binomial distributed dispersion after the RT-PCR testing with standard deviation σ_{M}. Thus, if Y is the observed number number of daily cases, we simulate Y from a negative binomial distribution with mean equal to Y_{sum} and and variance $Y_{\text{sum}} + \frac{\sigma_{\text{M}}^2}{{Y_{\text{sum}}}^2}$

2.6.5 Syndrome Surveillance Estimates

We estimate the number of non-COVID-19 severe respiratory cases that may have presented each week of the epidemic in NYC hospitals using syndrome surveillance data from NYC hospital emergency departments and observed influenza cases in NYC in previous years. The estimate we seek here represents the typical number of non-COVID-19 respiratory cases one would expect in a given week of the year given the seasonal pattern of influenza cases and that of other respiratory ailments that present to NYC hospitals.

Description of Data

Syndrome Surveillance for Respiratory Disease Weekly respiratory syndrome surveillance reports for all emergency departments in New York City hospitals from 2016-2020 were obtained from the New York Health Department's web portal [128]. These reports include all cases in which the chief complaint mentioned bronchitis, chest cold, chest congestion, chest pain, cough, difficulty breathing, pneumonia, shortness of breath, or upper respiratory infection. For this analysis, we use weeks 2-20 from 2016,2017 and 2019.

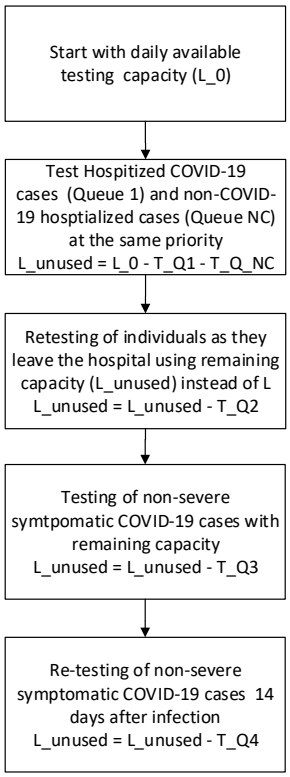

Figure 2.9: **Diagram of the testing priorities described in Section 2.6.4.**

Confirmed Flu Cases Confirmed influenza cases from all New York Counties from 2016-2020 for all counties in New York State were obtained from the New York State public health portal [127]. We used data from the five countries comprising New York City that correspond to the same time period as the syndrome surveillance data. We excluded 2018 from this analysis since 2018 was an anomalous, severe, influenza season [134].

Description of Statistical Model

Recall that $G_{\text{severe}}(w, y)$ is the number of non-COVID-19 respiratory infections during week w of year y that were severe enough to be sampled for COVID-19 testing. This was the quantity added to Queue NC in Equation 2.34 of Section 2.6.4. We assume that these cases are a fixed fraction s of the total non-COVID respiratory syndrome cases presenting in the emergency departments of hospitals in NYC in week w of year y, which we denote as $(G(w, y))$:

$$G_{\text{severe}}(w, y) = G(w, y) * s \tag{2.36}$$

We assume that in the absence of COVID-19, a portion of observed respiratory syndrome surveillance reports are associated with influenza, and that the non-influenza associated reports have a fixed seasonality.

Therefore, we consider that our estimate for the non-COVID-19 weekly respiratory syndrome surveillance cases $(G(w, y))$ varies linearly with confirmed influenza cases $F(w, y)$ in NYC and presents additional weekly variability whose effect is represented non-parametrically with a polynomial dependency as follows:

$$G(w, y) = g_0 + g_{\text{F}} F(w, y) + b_3 w^3 + b_2 w^2 + b_1 w + b_0 + \epsilon \tag{2.37}$$

where

$$\epsilon \sim rnorm(0, \sigma_\epsilon) \tag{2.38}$$

Model Fitting and Simulation We estimate the intercept g_0 and influenza coefficient g_{F} via a linear regression of respiratory syndrome surveillance reports and confirmed cases in New York City in 2016,2017 and 2019. We then fit the residuals from this linear regression to a third-order polynomial seasonality function to obtain estimates of coefficients b_i. We estimate the error term σ_ϵ by measuring the residual standard error from the polynomial fit.

When fitting the epidemiological model, we simulate values for $G(w, y)$ using observed weekly influenza cases in 2020 as the co-variate $F(w, y)$ obtained from the same source [127]. A plot of the fitted model is shown in Figure S2.15.

Estimation of Proportion Sampled for COVID-19 testing

Recall that the scaling parameter s represents the probability that an individual who shows up to the emergency department with respiratory symptoms is severe enough to merit testing for COVID-19. As a proxy for this value, we use the ratio of individuals aged 65 or older who were hospitalized for influenza to the number of individuals aged 65 or older who had a medical visit for influenza during the 2018-2019 influenza season [135].

The value we use for this scaling parameter s for the fitting of all three models is 0.16.

We profile over this parameter value as a sensitivity analysis(see Figure S2.13). Allowing this parameter to vary does not result in a higher likelihood with respect to serology data compared to the MLE parameter combination from the main analysis.

Table of Fitted Parameters

Parameter	Value
g_{F}	0.12
g_0	1183
b_3	0.012
b_2	0.981
b_1	-37.2
b_0	229
σ_ϵ	109
s	0.16

Table 2.3: Parameter values used for the non-COVID-19 severe cases estimate (rounded).

Detection of anomalies in 2020 syndrome surveillance

From our model of non COVID-19 respiratory cases, we can obtain an estimate of the expected number of syndrome surveillance reports $G_{w,y}$ in 2020 in the absence of COVID-19. Note that we do not use the scaling parameter s in this calculation, unlike when fitting the epidemiological model.

If we subtract this value from the observed respiratory syndrome surveillance reports in 2020, we obtain a metric for anomalous respiratory syndrome surveillance reports related to COVID-19. This is the pink line in Figure 5 of the main manuscript.

2.6.6 Overall Reproductive Number Derivation

Following [133], we derive $R_{0_{\mathrm{NGM}}}$ as the leading eigenvalue of the following matrix:

$$K = -T\Sigma^{-1}, \tag{2.39}$$

which is composed of two other matrices, T and Σ^{-1}, defined below.

The Transmission Matrix is given by

$$
T = \begin{matrix}
0 & 0 & 0 & 0 & 0 & \beta_P & \beta_A & \beta & \beta \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{matrix}
$$

The Transition Matrix is given by

$$
\Sigma = \begin{matrix}
-\phi_E & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\phi_E & -\phi_E & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \phi_E & -\phi_E & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \phi_E & -\phi_E & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \phi_E & -\phi_E & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \phi_E & -\phi_U & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & (1-p_S)\phi_U & -\phi_S & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & p_S\phi_U & 0 & -\phi_S & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & (1-p_{\mathrm{H}})\phi_S & -\gamma
\end{matrix}
$$

Then, the first five columns of the inverse of the transition matrix are:

$$\Sigma^{-1} = \begin{matrix} \frac{-1}{\phi_E} & 0 & 0 & 0 & 0 \\ \frac{-1}{\phi_E} & \frac{-1}{\phi_E} & 0 & 0 & 0 \\ \frac{-1}{\phi_E} & \frac{-1}{\phi_E} & \frac{-1}{\phi_E} & 0 & 0 \\ \frac{-1}{\phi_E} & \frac{-1}{\phi_E} & \frac{-1}{\phi_E} & \frac{-1}{\phi_E} & 0 \\ \frac{-1}{\phi_E} & \frac{-1}{\phi_E} & \frac{-1}{\phi_E} & \frac{-1}{\phi_E} & \frac{-1}{\phi_E} \\ \frac{-1}{\phi_U} & \frac{-1}{\phi_U} & \frac{-1}{\phi_U} & \frac{-1}{\phi_U} & \frac{-1}{\phi_U} \\ \frac{-(1-p_S)}{\phi_S} & \frac{-(1-p_S)}{\phi_S} & \frac{-(1-p_S)}{\phi_S} & \frac{-(1-p_S)}{\phi_S} & \frac{-(1-p_S)}{\phi_S} \\ \frac{-p_S}{\phi_S} & \frac{-p_S}{\phi_S} & \frac{-p_S}{\phi_S} & \frac{-p_S}{\phi_S} & \frac{-p_S}{\phi_S} \\ \frac{-(1-p_{\mathrm{H}})p_S}{\gamma} & \frac{-(1-p_{\mathrm{H}})p_S}{\gamma} & \frac{-(1-p_{\mathrm{H}})p_S}{\gamma} & \frac{-(1-p_{\mathrm{H}})p_S}{\gamma} & \frac{-(1-p_{\mathrm{H}})p_S}{\gamma} \end{matrix}$$

The last four columns of the inverse of the transition matrix are:

$$\Sigma^{-1} = \begin{matrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \frac{-1}{\phi_U} & 0 & 0 & 0 \\ \frac{-(1-p_S)}{\phi_S} & \frac{-1}{\phi_S} & 0 & 0 \\ \frac{-p_S}{\phi_S} & 0 & \frac{-1}{\phi_S} & 0 \\ \frac{-(1-p_{\mathrm{H}})p_S}{\gamma} & 0 & \frac{-(1-p_{\mathrm{H}})}{\gamma} & \frac{-1}{\gamma} \end{matrix}$$

From the leading eignvalue of the resulting matrix K, we finally obtain:

$$R_{0_{NGM}} = \frac{\beta_P}{\phi_U} + \frac{\beta_A(1-p_S)}{\phi_S} + \frac{\beta p_S}{\phi_S} + \frac{\beta(1-p_{\mathrm{H}})p_S}{\gamma} \tag{2.40}$$

2.6.7 Supplemental Figures

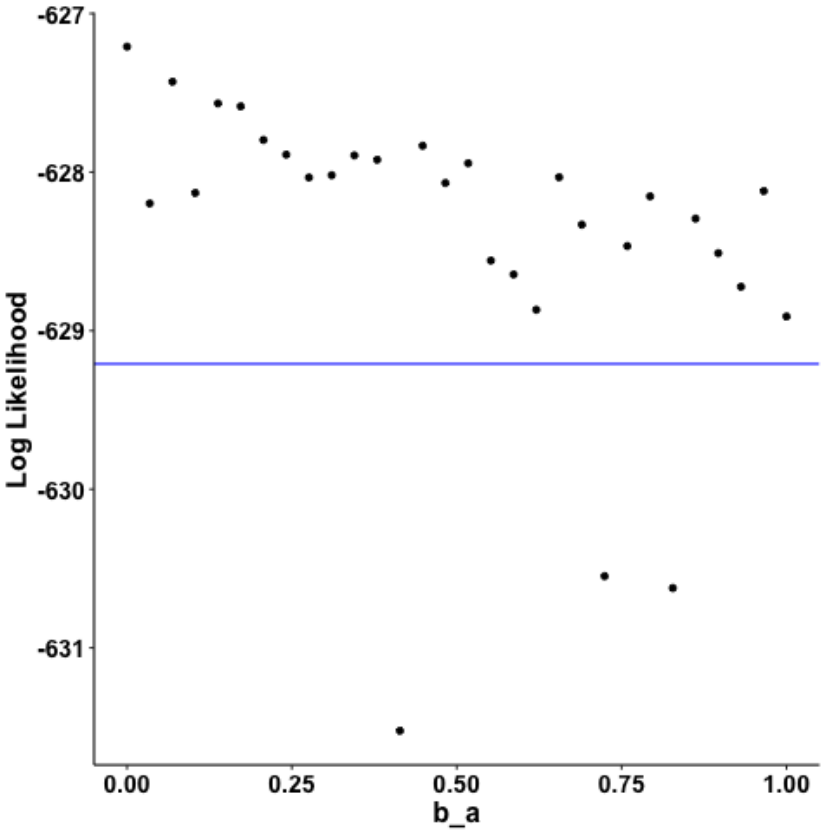

Figure 2.10: **Monte Carlo profile of the strength of transmission in asymptomatic cases relative to that of symptomatic cases (b_a).** Each point represents the parameter combination from the Monte Carlo profile for b_a with the highest log-likelihood (with respect to observed cases) for a given value of b_a. All points above the blue line are supported by the case data (i.e. they have likelihoods within 2 log likelihood units of the profile MLE).

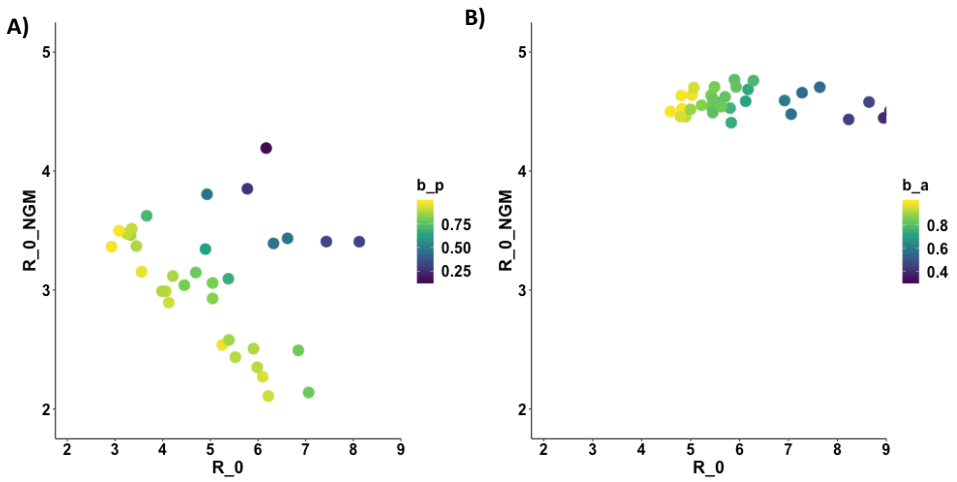

Figure 2.11: **Additional plots of the overall reproductive number ($R_{0_{\mathrm{NGM}}}$) vs the reproductive number in symptomatic individuals (R_0) from parameter combinations supported by case and serology data from the full SEPIAR (A) and SEIAR (B) models. A)** Each point represents one parameter combination within 2 log-likelihood units of the MLE (with respect to cases and serology) from the b_{a} profile using the full SEPIAR model. Each point is colored by the strength of pre-symptomatic transmission (b_{p}). **B)** Each point represents one parameter combination within 2 log-likelihood units of the MLE (with respect to serology) from the grid search of the SEIAR model (no pre-symptomatic transmission). Points are colored by the relative strength of asymptomatic transmission (b_{a}). For ease of plotting, in panel A) we exclude two parameter combination with very low pre-symptomatic transmission rates b_p. These outliers are also excluded from Figure 3 in the main manuscript , but are shown in Supporting Figure S2.12.

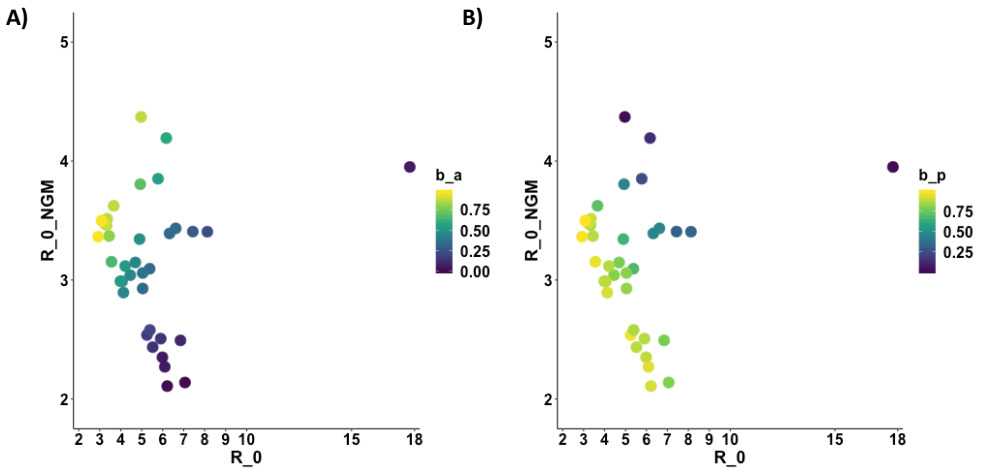

Figure 2.12: **Additional plots of the overall reproductive number ($R_{0_{\mathrm{NGM}}}$) vs the reproductive number in symptomatic individuals (R_0) including the outlier parameter combination colored by the relative strength of asymptomatic transmission (b_{a}) (A) or the relative strength of pre-symptomatic transmission (b_{p}) (B).** Each point represents one parameter combination within 2 log-likelihood units of the MLE (with respect to cases and serology) from the b_{a} profile of the full SEPIAR model. We include here the outlier parameter combinations with very high symptomatic R_0 greater than 15 that were excluded from Figure 3 and Figure S2.11.

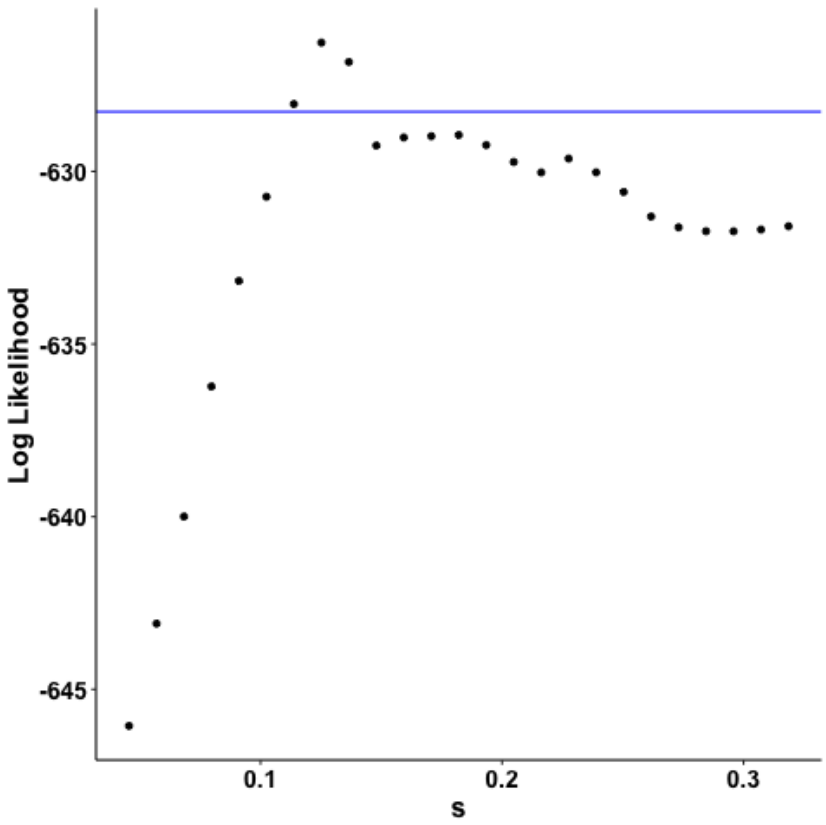

Figure 2.13: **Monte Carlo profile of the probability that an individual (who does not have COVID-19) who shows up to the emergency department with respiratory symptoms is severe enough to merit testing for COVID-19 (s).** Each point represents the parameter combination from the Monte Carlo profile for the scaling parameter s with the highest log-likelihood (with respect to observed cases) for a given value of s. All points above the blue line have likelihoods within 2-log likelihood units of the profile MLE).

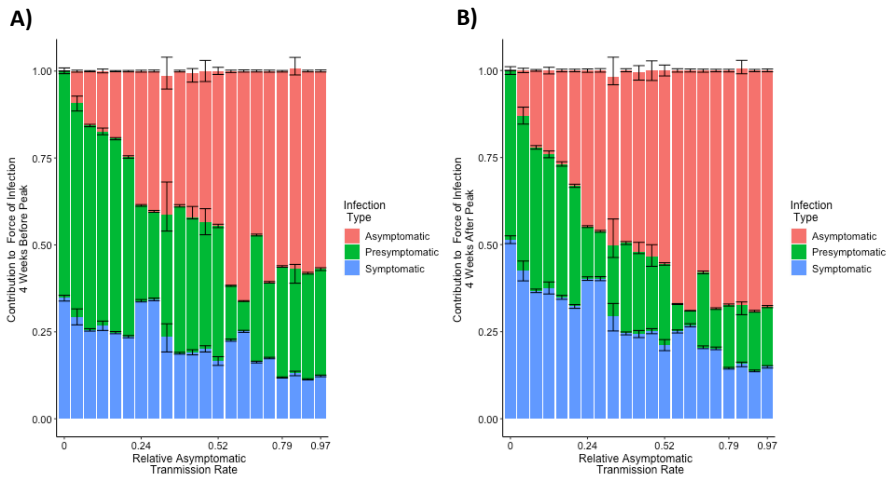

Figure 2.14: **Contributions from pre-symptomatic, symptomatic, and asymptomatic infections to the overall force of infection 4 weeks before (A) and after the peak in reported cases (B).** The calculation of the contribution to the overall force of infection from simulated trajectories of parameter combinations from the SEPIAR model supported by case and serology data as described in Figure 4 of the main manuscript was replicated using time points 4 weeks before and after the peak of reported cases on April 14, 2020 instead of at the time of the peak. As in Figure, two parameter combinations with rates of pre-symptomatic transmission b_p below 0.02 were excluded from the plot.

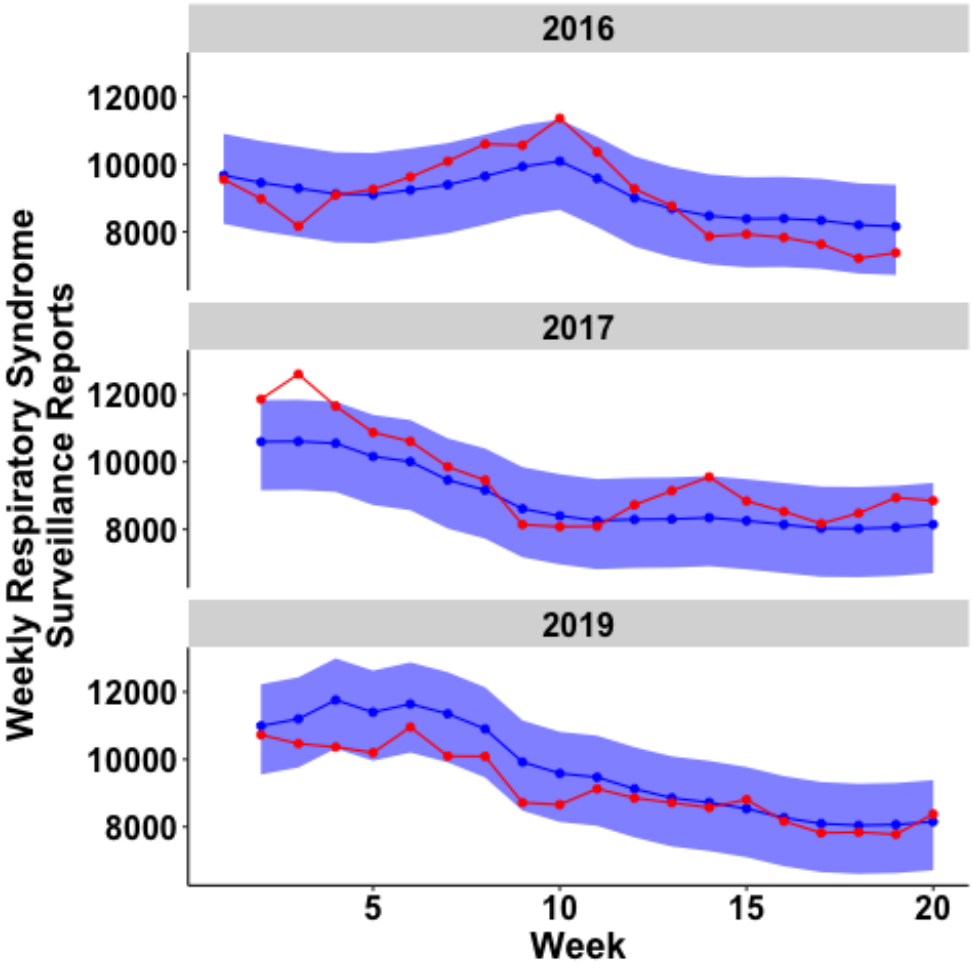

Figure 2.15: **Plot of observed respiratory syndrome surveillance reports compared to simulations from fitted statistical model** The red line corresponds to weekly respiratory infections from syndrome surveillance reports in NYC hospitals in 2016, 2017 and 2019 that were used to fit the statistical model in Section 2.6.5. The blue line represents the median estimate for the number of expected syndrome surveillance reports ($G(w, y)$ for that week and year from 100 simulations from the fitted statistical model. The shaded light blue region represents the 2.5% and 97.5% quantiles from those 100 simulations.

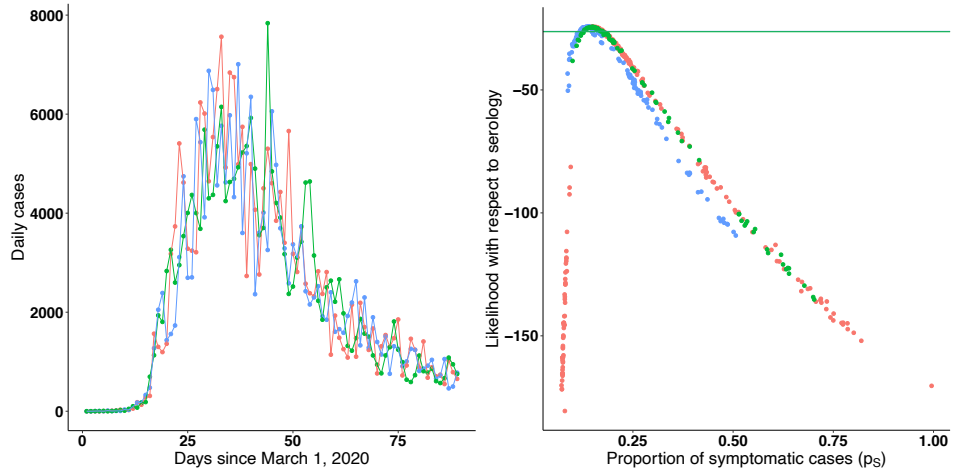

Figure 2.16: **Validation analysis results from fitting model to simulated data. We are particularly interested here in verifying the ability of the inference pipeline to estimate the value of the probability of symptomatic infection p_S used in the simulations. A) Observed data and simulated trajectories used for fitting**. The green points denote observed daily reported case counts in New York City. The red points denote daily reported cases from a representative simulated trajectory from a "low" b_a parameter combination ($b_a = 0.07$, $R_0 = 6.10$, $b_q = 0.23$, $b_p = 0.94$, $p_S = 0.15$, $p_H = 0.16$, $\gamma = 6.33$,$E_0 = 63566.34$ $z_0 = 13443$, and the overall reproductive number $R_{0_{NGM}} = 2.27$), while the blue points denote daily reported cases for a representative trajectory from a "high" b_a parameter combination ($b_a = 0.97$, $R_0 = 3.08$, $b_q = 0.16$, $b_p = 0.99$, $p_S = 0.15$, $p_H = 0.17$, $\gamma = 11.73$, $E_0 = 54806$, $z_0 = 11625$, and $R_{0_{NGM}} = 3.50$). **B) Supported parameter ranges for the proportion of cases that are symptomatic (p_S for fits to simulated and observed data**. Red and blue dots represent respectively parameter combinations supported by the case data when the model was fit to the low b_a (red) or high b_a (blue) trajectories. For comparison, the green dots represent parameter combinations supported by the case when the model was fit to observed case data (as shown in panel B of Figure 2). All parameter combinations above the green line have likelihoods within 2-log-likelihood units of the MLE defined with respect to serology. Our approach recovers the value of p_S used in the simulations, and does so accurately.

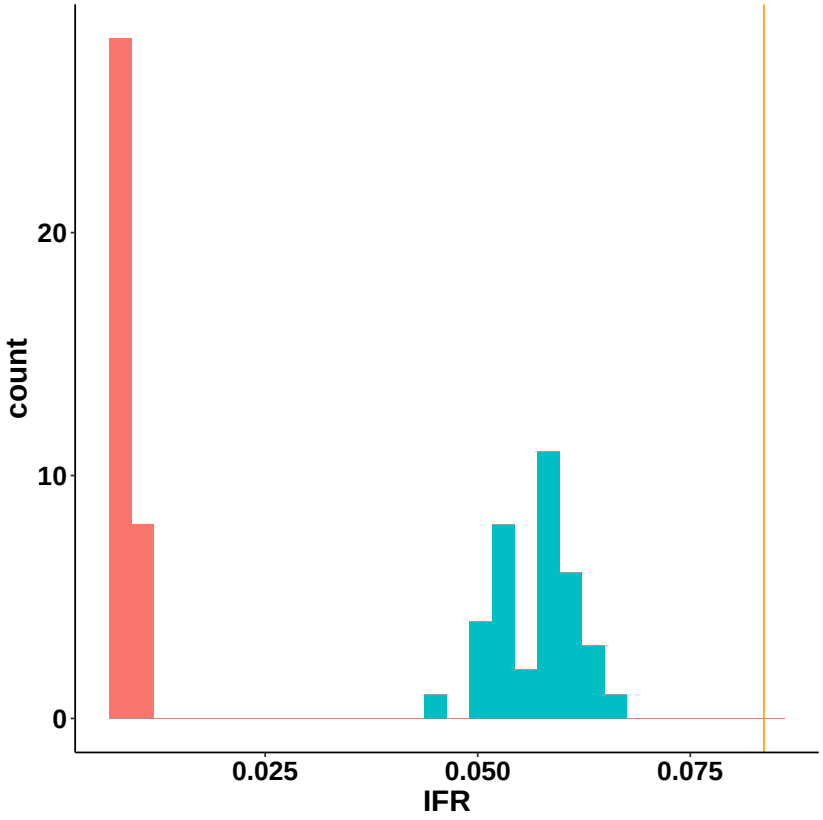

Figure 2.17: **Comparison of infection fatality ratios (IFR) estimated from fitted model parameters under different testing strategies**. The red shaded region denotes a histogram of the infection fatality ratio calculated with respect to all cases, both symptomatic and asymptomatic. The IFR was calculated for each parameter combination from the SEPIAR model that was supported by both case and serology data. The proportion of hospitalized cases that result in deaths was estimated from observed confirmed COVID-19 hospitalisations and deaths in New York City during time period of the study. The red histogram shows the range of IFR values expected under the SEPIAR model if all cases (symptomatic and asymptomatic) are observed. Each count in the histogram represents the expected IFR for one parameter combination that is supported by the case and serology data. The blue histogram shows the expected IFR if all symptomatic cases are observed. The higher IFR obtained in the blue histogram compared to the red histogram demonstrates how different testing strategies can alter the IFR. The orange line denotes the observed IFR calculated by dividing the total number of confirmed COVID-19 deaths in NYC during the study period by the total number of confirmed cases. The gap between the orange line and the blue histogram illustrates how limited testing capacity can affect the IFR that is estimated, since not all symptomatic cases were tested due to limited testing capacity early in the outbreak.

CHAPTER 3

PREDICTING RE-EMERGENCE TIMES OF DENGUE EPIDEMICS AT LOW REPRODUCTIVE NUMBERS: DENV1 IN RIO DE JANEIRO, 1986-1990.

[Originally published as: Subramanian, R., V.R. Aznar, E. Ionides, C.T. Codeço, M. Pascual. 2020. Predicting re-emergence times of dengue epidemics at low reproductive numbers: DENV1 in Rio de Janeiro, 1986-1990. Journal of the Royal Society Interface 17: 20200273.]

[Note: My colleague V.R. Aznar designed and performed several of the analyses described in this chapter, specifically the derivation of the skip threshold and the fast dynamics vector model analysis, and wrote the corresponding sections of the published manuscript. This approximately corresponds to Figure 3.1, and panel B of Figure 3.2, and panel A of Figure 3.4 (provided skip expression used in figure), in the Results section, Figures 3.20 and 3.25 (provided skip expression), and Figures 3.27 thru 3.29, as well as sections 3.8.1 and 3.8.4 in the Supporting Information. I include these results and text in this chapter for completeness since they were included in the published manuscript, while acknowledging her contribution.]

3.1 Introduction

Epidemics of arboviruses such as dengue [136], Zika [137, 138], and chikungunya [139] result in substantial global morbidity. Over the past decade, invasions of several arboviruses have triggered large outbreaks in the Western Hemisphere. In Brazil, these invasions include dengue serotype DENV4 in 2012 [140] as well as Zika [137, 141] and chikungunya [142] between 2014-2016. Predicting and understanding the re-emergence of arboviruses after these invasions has important consequences for epidemic preparedness, particularly in regions where climate factors limit mosquito transmission in the off-season. These regions typically exhibit highly intermittent seasonal epidemics, lasting one to three years with long

periods of no, or low, reported cases in between, and low mean reproductive numbers (the number of secondary cases arising from each primary case in a completely susceptible population, R_0) [140, 5, 6, 19]. Several proposed explanations include the depletion of susceptible individuals following initial epidemics [143] and the time required for their replenishment via population growth [144], inter-annual variation in climate [145, 146, 147, 148, 149], and antigenic interactions between strains of different serotypes [46, 32, 150, 151]. These temporal patterns contrast with the recurrent seasonal outbreaks observed in childhood diseases with high reproductive numbers, whose extensive study has provided the basis for our theoretical understanding of SIR (Susceptible-Infected-Recovered) dynamics in infections that confer lifelong or lasting immune protection [15, 11, 12, 152, 153, 14, 154, 8]. Statistical models of dengue transmission that take into account climate dependencies can be used to make short-term re-emergence forecasts on the order of 4 months [49] or 16 weeks [147]. Many epidemiological models that predict the re-emergence of arboviruses such as Zika [143, 52] on longer time-scales of a year [143] or several decades [52] rely however on compartmental formulations such as SIR-type approaches [143] or Ross-McDonald equations that explicitly incorporate vector transmission [52]. Both formulations assume transmission between any two individuals in the population ('well-mixed' conditions), typically at aggregated spatial scales. These process-based formulations, for example those recently applied to Zika, represent the acquisition of immunity in the population and its loss via demographic growth and turnover. These models do take into account seasonality of transmission and spatial heterogeneity in the intensity of transmission due to climate at coarse resolutions (at large city, state, or country-level scales). Nevertheless, the replenishment of a well-mixed susceptible population is frequently assumed to be the principal driver determining when the disease will re-emerge given a particular seasonal pattern for R_0 at a particular location [52]. Stochasticity can also play an important role in long-term models of re-emergence [52]. Variation in reporting rates of arboviruses between locations [155] can add further complexity. Although childhood diseases with high reproductive numbers display different dynamics from

emergent arboviruses [15, 11, 12, 152, 153], their compartmental models share a basic SIR structure given the acquisition of long-term immunity after infection. The resulting depletion and replenishment of the susceptible population is known to clearly drive inter-annual variability and re-emergence in the former [152, 14, 154]. In particular, recent theory [8] has derived analytical expressions for the number of "skip" years for a measles-like disease in the pre-vaccine era, where "skips" are defined as seasons when transmission occurs but does not cause susceptible depletion. In other words, although the number of infections increases in such seasons, it is not large enough to offset the growth in the susceptible population due to demography. The resulting expressions specifically provide a threshold condition for the number of skips expected following an initial invasion as a function of R_0. Their derivation did not include under-reporting and assumed a closed-population SIR model with 'school-term' seasonality, alternating two different rates for low and high transmission. We examine in this work whether replenishment of susceptible individuals under the typical 'well-mixed' assumption explains dengue (DENV1) re-emergence at the whole-city aggregated level. We specifically address the uncertainty inherent in such predictions at the low reproductive numbers characteristic of arboviruses, not previously considered in applications of the analytical approach. To this end, we first extend the threshold derivation to take into account population growth, continuous (sinusoidal) seasonality, and under-reporting of cases. We then fit a stochastic SIR model to observed monthly dengue case counts from the DENV1 invasion in Rio de Janeiro, Brazil from 1986-1988 [5, 19, 7] and numerically predict expected times to re-emergence. We describe high uncertainty in re-emergence times for these seasonal, low transmission regions, and show the insufficiency of susceptible replenishment in a simple SIR model to explain the short periods observed in DENV1 re-emergence. We discuss possible explanations and the need for model formulations that would scale to coarse spatial resolutions.

3.2 Results

We start with the analytical approach for a seasonally forced SIR system with intermittent outbreaks and population turnover, to consider general features of re-emergence at low R_0. In such a system, the onset of the off-season can bring an end to an initial outbreak, and the replenishment of susceptible individuals due to births and population turnover can be a major determinant of recurrence times. Let S represent the number of susceptible individuals in a population and s_0, the fraction of the population still susceptible at the end of an initial epidemic, t_0, when a prediction for the time to the next outbreak will be made. If there are enough susceptible individuals left in the population (i.e. if s0 is large), another outbreak will occur in the following year once the on-season resumes. However, if the initial outbreak was very large, s_0 may be too small, and the outbreak may "skip" one or more years. A skip year is defined as a year in which the susceptible population does not decrease, whether or not infections increase. The smaller the fraction of the susceptible population at the time of prediction (s_0), the longer it will take for the susceptible population to replenish, and the larger the number of skips that will occur. Previous theory [8] allows prediction of the number of skips that will occur given s_0. Specifically, it demonstrated that s_0 must fall below some threshold $s_c(k)$ for k skips to occur. An analytical expression was provided for $s_c(k)$ in terms of the reproductive number and population turnover rate for a closed-population SIR model with school-term seasonality [8]. The derivation of the threshold presented in [8]) requires the assumption that the transmission rate or reproductive number of the disease is high and that the fraction of the population susceptible at the time of prediction (s_0) is small. We extend this approach to take into account population growth and sinusoidal seasonality (which describes the transmission rate of dengue more accurately than a discrete high-low representation). Our derivation does not require assuming that the transmission rate or reproductive number are high or that the fraction of the population susceptible at the time of prediction is small. We follow the criteria developed in [8] (see details in [156]), which

essentially consider the sign of the logarithm of the ratio between the respective number of infections at two times, t_0 and $t_n > t_0$. A positive value indicates that an outbreak will still occur at tn; conversely a negative value indicates no outbreak at that time. By setting the logarithm of this ratio to zero, the threshold s_c is obtained (See Section 1 of the Supporting Information for details). The resulting expression for $s_c(n)$, the critical fraction of susceptible individuals required at the time of prediction for n or more skip seasons to occur, is

$$s_c(n) = 1 + (\pi(2n+1)(1 - 1/R_0) - 2\delta)/(\omega f(\omega, \delta, r, n)) \tag{3.1}$$

where $f(\omega, \delta, r, n) = (1 + e^{-r(\frac{\pi}{\omega})(2n+1)})\frac{\omega\delta}{(\omega^2+r^2)} - \frac{1-e^{-r(\frac{\pi}{\omega})(2n+1)}}{r}$,

R_0 is the annual mean of the reproductive number, δ the amplitude of seasonal transmission (as infectious contacts per person per day), ω, the transmission frequency (in days^{-1} and r, the population growth rate (also in days^{-1}). The full expression for the seasonal transmission rate is given by $\beta(t) = \beta_0(1 + \delta sin(\omega t + \phi)$, where ϕ corresponds to the phase (in radians) and β_0, to the mean seasonal transmission rate (infectious contacts per person per day). The quantity β_0 is related to the annual mean reproductive number R_0 via the expression $R_0 = \frac{\beta_0}{\gamma}$, where γ is the recovery rate (in days^{-1}).

Figure 3.1 illustrates the implications of this formula. As before, t_0 corresponds to the time of prediction, in practice usually after a large initial epidemic or invasion. Likewise, s_0 represents the fraction of the population susceptible at the time of prediction. Intuitively, the smaller the fraction of the population susceptible at the time of prediction (s_0), the longer it will take for the susceptible population to replenish, and the larger the number of skips that will occur. In practice, as we will illustrate below, values of s_0 can be computed from surveillance data provided one has an estimate of the reporting rate. For n skips to occur, the fraction of the population susceptible at the time of prediction (s_0) must fall below the susceptibility threshold $s_c(n)$. Figure 3.1A shows that the larger the number of skips

n one is considering, the smaller the threshold $s_c(n)$ that s_0 must fall below for at least n skips to occur. Let n_c denote the critical skip number corresponding to the number of skips expected at the time of prediction (t_0). We use the fraction of the population susceptible at the time of prediction (s_0) and identify the maximum value of n for which s_0 is smaller than $s_c(n)$. In the example shown in Fig. 3.1A, this fraction $s_0 = 0.7$ is smaller than $s_c(n = 6)$ and bigger than $s_c(n = 7)$, which means $n_c = 6$. We therefore expect six years of skips followed by re-emergence in the seventh year. Formally, for a given value of s_0 at the end of the transmission season, we define the critical skip number n_c as the value of n for which $s_c(n_c) > s_0 > s_c(n_c + 1)$.

With this general approach at hand, we explored the effects of the reproductive number R_0, amplitude of seasonal transmission δ and fraction of the population susceptible at time of prediction s_0, on the critical number of skips n_c (Figure 1 Panels B and C). Consideration of both the variation of the reproductive number R_0 and fraction of the population susceptible at time of prediction s0 is relevant here. Different combinations of transmission rate (β_0) and duration of the infection ($\frac{1}{\gamma}$) can yield the same R_0 but different fractions of the population susceptible at the time of prediction (Fig. 3.21). Importantly, Fig. 3.1 panels B and C show that the time to re-emergence is very sensitive to R_0. A singularity is observed as R_0 approaches 1 where the expected number of skips goes to infinity. The approach to that singularity can be very steep, meaning that small changes in R_0 can result in large increases in the expected re-emergence time. The obtained values of n_c are not as sensitive to the amplitude of seasonal transmission (Fig. 3.1 Panel B) but are sensitive to the fraction of the population susceptible at the time of prediction (Fig. 3.1 Panel C). The shift of the curve in Fig 1 Panel C for small values of the fraction of the population susceptible at time of prediction s_0 means that, for a given R_0, more time is required to replenish the susceptible population and therefore to observe a re-emergence.

We next apply this approach to the surveillance data from the 1986 invasion of DENV1 in Rio de Janeiro (Figure 3.2). The initial DENV1 invasion in Rio de Janeiro is an ideal initial

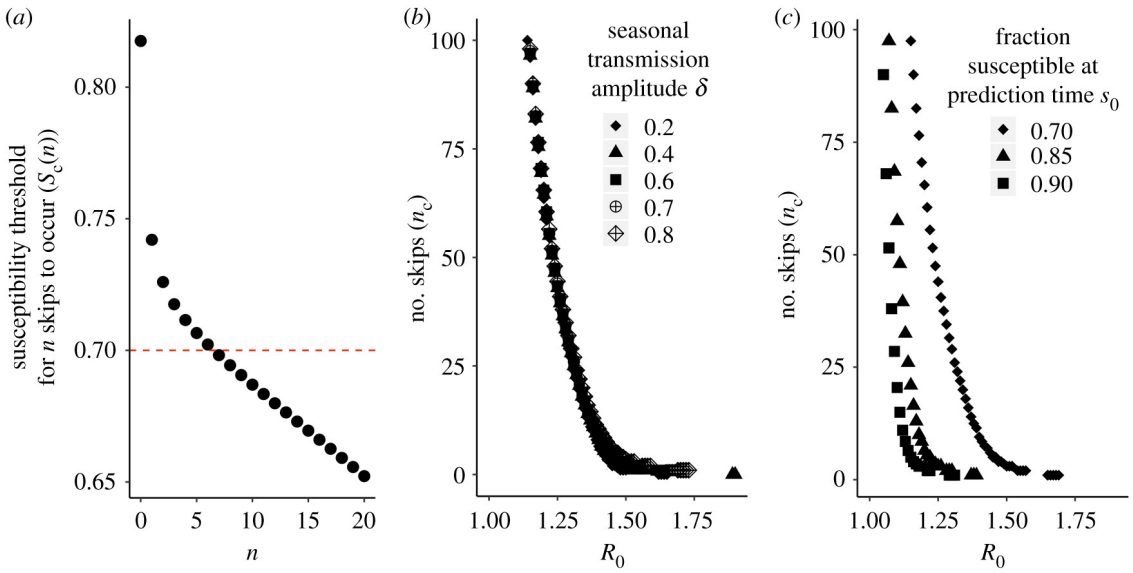

Figure 3.1: **A) Graphical illustration of how the expected number of skips (n_c) is calculated.** The black dots represent the threshold fraction of the population susceptible at the time of prediction required for n skips to occur ($s_c(n)$). The plot shows ($s_c(n)$) as a function of n (the number of skips) obtained from Equation 1 with seasonality amplitude $\delta = 0.2$ (contacts per person per day) and reproductive number $R_0 = 1.4$. In this example, the red line represents the fraction of the population susceptible at the time of prediction (s_0). If s_0 is smaller than $s_c(n)$, at least n skips will occur. To find the expected number of skips (n_c), we identify the largest number of skips n such that s_0 is smaller than the susceptibility threshold required for those skips $s_c(n)$. In this example, the red line intersects the $s_c(n)$ curve between $s_c(n = 6)$ and $s_c(n = 7)$. Therefore, a critical skip number of $n_c = 6$ is obtained. **B) and C) The critical skip value n_c as a function of** R_0 for (B) different values of the amplitude of seasonal transmission δ with $s_0 = 0.7$ and (C) different values of the fraction of the population susceptible at the time of prediction (s_0) with $\delta = 0.70$. In all three panels, the frequency of transmission ω the population turnover rate μ and population growth rate r are fixed at respective values $\omega = \frac{2\pi}{365}\text{days}^{-1}$ corresponding to an annual periodicity, $\omega = 1/(74.46 * 365))\text{days}^{-1}$ corresponding to an average lifespan of 75 years, and $r = 1.55\mu\text{days}^{-1}$ consistent with the growth of the city of Rio de Janeiro. These values were chosen for the purpose of illustration, based on the inverse of the average life expectancy in Brazil in 2012 according to the 2010 census [2] , and the interpolation of population estimates for the resident population of the municipality of Rio de Janeiro from the 1991 [3] and 2000 [4] censuses assuming exponential growth.

test case for this technique given the lack of widespread prior immunity from to prior dengue epidemics, vaccination campaigns, cross-immunity from other disease outbreaks. Specifically, the 1986 invasion occurred prior to the development of dengue vaccines. The outbreak was

the first dengue invasion in the area since the initial eradication of the Aedes aegypti mosquito in Brazil in the 1950s [157, 158, 159, 160] following a sustained intervention program that began in the 1930s and 1940s in Rio de Janeiro and other cities [158]. Cross-immunity from yellow fever vaccination appears to be very limited [161]. Given the young age distribution of the population in 1986 [162], most individuals were not alive during the period when mass yellow fever vaccination or prior dengue epidemics occurred.

We let our time of prediction t0 be equal to September 1, 1987, corresponding to the end of the initial DENV1 invasion (see panel A of Figure 3.2). In panel B of Figure 3.2, we evaluate the number of expected skips expected in Rio de Janeiro, n_c, on the basis of a range of R_0 values from 1.18 to 2.02 from the literature [9, 163]. The critical susceptibility threshold for n skips to occur ($s_c(n)$) is calculated using Equation 1 with an annual seasonality, a population growth rate interpolated from the census (see Methods section), and $\delta = 0.7$ [9]. The fraction of the population susceptible at the time of the prediction (s_0) is estimated as the difference between the total population N_0 (total population N at (t_0 =Sep. 1987)) and the total number of people infected between the start of the invasion and the time of prediction (September 1, 1987). The total number of infected people during the outbreak is computed by summing the ratio between the observed monthly cases and the reporting rate for DENV1 in the city. Literature estimates from serology during the DENV1 invasion in Rio de Janeiro indicate a reporting rate of around 3% (33) which we use and fix for this analysis. For comparison purposes, we also include the number of skips expected under a higher reporting rate of 10%. These curves show that the expected re-emergence could be very sensitive to small variation in R_0 and ρ, two quantities that are difficult to estimate with precision in the absence of serology. In particular, assuming a reporting rate of 3%, a reproductive number of 1.2 with 20% uncertainty can yield large changes in the expected re-emergence time. We highlight the potential sensitivity of the expected number of skips to the reporting rate as well to illustrate the importance of uncertainty in this parameter in cities or epidemics where its value is unknown.

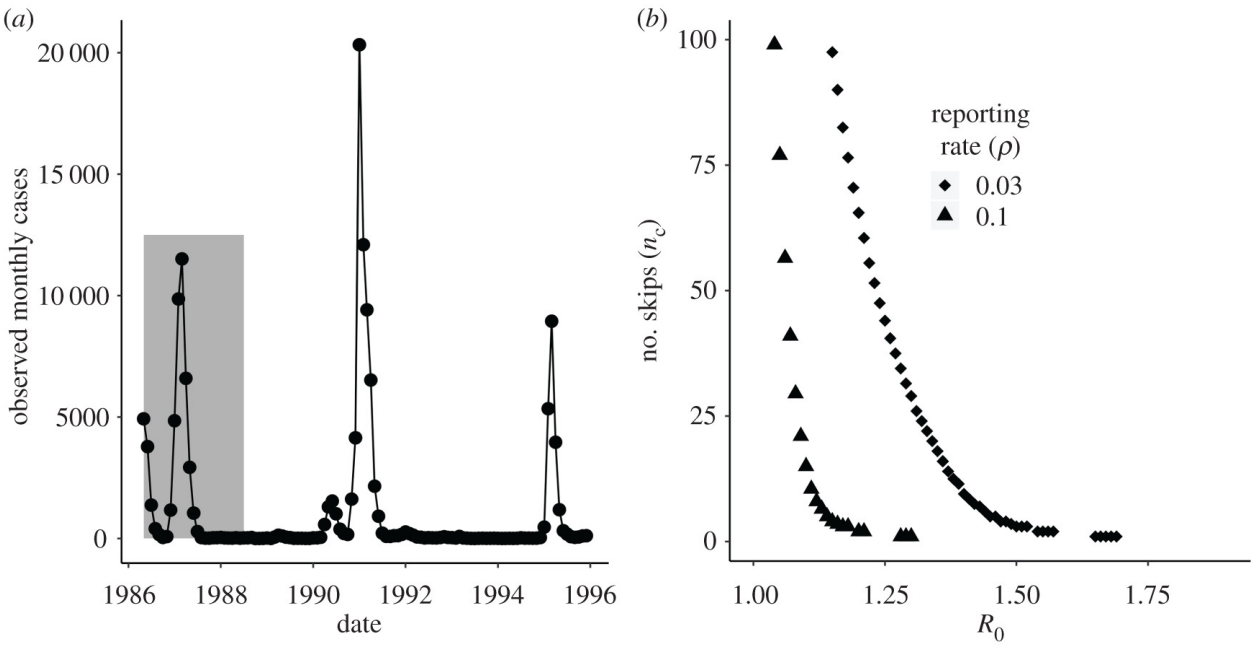

Figure 3.2: **(A) Observed dengue case data.** Monthly reported dengue cases in the city of Rio de Janeiro, Brazil from April 1986-1995. The grey shaded region denotes observations that were included in the fitting of the stochastic model from May 1, 1986 to July 1, 1988 inclusive. Serotype DENV1 re-emerged in 1990. DENV2 was first detected in the state of Rio de Janeiro in 1990 but did not become dominant until 1991 [5, 6]. Both co-circulated afterwards. We focus on the invasion of DENV1 from 1986-1987 and its initial re-emergence in DENV1 in 1990 using a single serotype transmission model. This allows us to evaluate this transmission model in a region where only one serotype was circulating, where cross-immunity could not easily be invoked to explain the absence or reduction of dengue in a given year. **(B) Deterministic critical number of DENV1 skips n_c for Rio de Janeiro from September 1988.** Expected number of skips n_c with amplitude of seasonal transmission $\delta = 0.7$ and the fraction of the population susceptible after the first DENV1 invasion as of September 1, 1987 (s_0) calculated from the data (A). We use a reporting rate ρ of 3% when calculating s_0 , consistent with serological estimates from the literature [7]. For comparison purposes, we also include the expected number of skips n_c assuming a reporting rate of 10%.

3.2.1 Replenishment of Susceptible Individuals is Insufficient to Explain Re-Emergence

To obtain more precise bounds for the reporting rate and R_0 and to determine if the depletion and replenishment of susceptible individuals could explain the rapid re-emergence

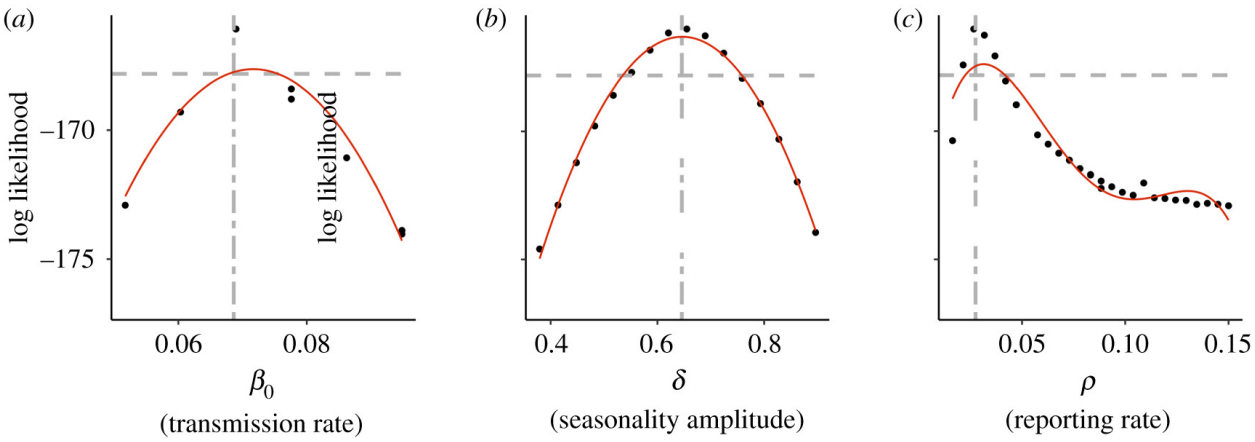

Figure 3.3: **A-C) Selected parameter profiles for the stochastic model.** Profiles of the mean annual transmission rate β_0 (A), seasonal transmission amplitude δ (B), and reporting rate ρ (C). The red curve is a polynomial fit to the subset of the profile points shown on the figure. The single dashed grey horizontal line represents the likelihood value 2 log likelihood units below the maximum likelihood estimate. This line provides an estimate of confidence intervals for the given parameter. The grey vertical line denotes the parameter value of the maximum likelihood estimate. The maximum likelihood estimate for the reporting rate in panel C is very close to the literature value obtained from serology (approximately 3 percent). [7]

of dengue in Rio de Janeiro, we fit a stochastic aggregate SIR model to case data from the DENV1 invasion from 1986-1988. The stochastic SIR model assumes that the underlying deterministic transmission rate varies seasonally as a sinusoidal function with annual mean β_0, seasonal transmission amplitude δ, frequency ω (equal to $\frac{2\pi}{365}$), and phase ϕ. The model takes into account demographic stochasticity, environmental stochasticity in the transmission rate, and measurement error due to under-reporting and variation in reporting of cases (See Materials and Methods and the Supporting Information). Panels A,B, and C of Figure 3.3 show the likelihood profile of the annual mean transmission rate, β_0 the amplitude of seasonal transmission δ and the reporting rate ρ respectively. In particular, our estimate of the reporting rate matches that from serology in the literature (Panel C).

Overall, the model is able to capture key dynamics of the DENV1 invasion including the two peaks of incidence in 1986 and 1987 and the subsequent reduction of transmission in

1988. This is shown by comparing the trajectories for an ensemble of simulations with the fitted model to the observed values of cases (Fig. 3.4). Estimated values for the transmission rate indicate a low value for R_0 (Figure 3.4 Panel C). Both of these conclusions generally hold even if one takes into account uncertainty in parameter estimates by examining all parameter combinations with log likelihoods within 2 log likelihood units of the maximum likelihood estimate (the grey region in Fig. 3.4 Panel C as well as Fig. 3.6), although some parameter combinations (not the maximum likelihood estimate) have substantial process noise (Fig. 3.6).

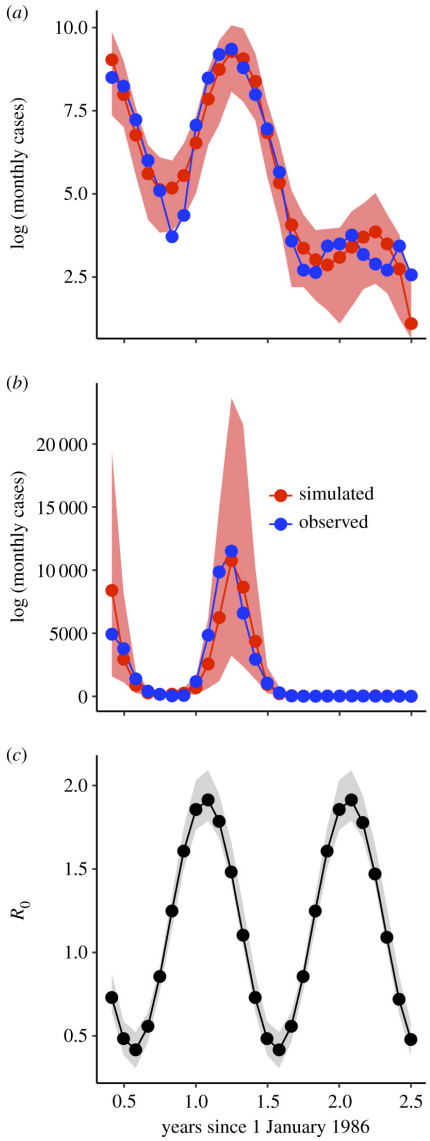

Figure 3.4: **A-B) Comparison of simulated values with the fitted model and observed data on a log (A) and regular (B) scale.** Observed monthly cases from April 1986 to June 1988 are shown in blue. Median values from 100 simulations with the maximum likelihood parameter combination are shown in red. The shaded red region denotes the 2.5% and 97.5%th quantile boundaries from those simulations. **C) Estimates for $\mathbf{R_0(t)}$.** The black line denotes the trajectory of R_0(t) for the maximum likelihood estimate. The shaded grey region represents the 2.5% and 97.5%th quantile boundaries for trajectories from all parameter combinations within 2 log likelihood units of the maximum likelihood estimate. Each parameter combination has only one seasonal trajectory for $R_0(t)$ since $R_0(t)$ is a deterministic quantity. $R_0(t)$ for all parameter estimates ranges from 1.79-2.09 in the on season to 0.31-0.52 in the off-season.

We now apply the obtained parameter estimates from the fitted model to address the expected re-emergence time on the basis of, first, the analytical expression for the skip calculation (Equation 1), and then the stochastic simulations of the fitted model. The parameter estimates used here are those for the reporting rate ρ the reproductive number R_0, and the amplitude of seasonal transmission δ from all combinations within 2 log likelihood units of the MLE. The expected number of skips following the DENV1 invasion in 1986-1988 is considerably higher than the observed 2 years. Depending on the parameter combination used, we obtain anywhere from 27 to 100 skips (Panel A of Figure 3.5). Even the fastest estimated return from the skip analysis (27 years) is much slower than the observed re-emergence time.

Forward simulation of the stochastic model likewise does not predict the rapid re-emergence of DENV1 (Panel B of Figure 3.5). Under a pulse of 20 infected individuals arriving per day, there was a low probability of re-emergence for parameter combinations with low process noise (Panel B of Figure 3.5). Only parameter combinations with high amounts of process noise (which have limited predictive value) had a non-zero emergence probability. We consider alternate pulse rates in Supplemental Figure S14. Re-emergence probabilities under forward simulation of the stochastic model thus corroborated the deterministic skip findings. The depletion of susceptible individuals from 1986-1988 and their replenishment via population growth from 1989-1990 under an aggregate SIR model was unable to explain the rapid re-emergence of DENV1 in 1990.

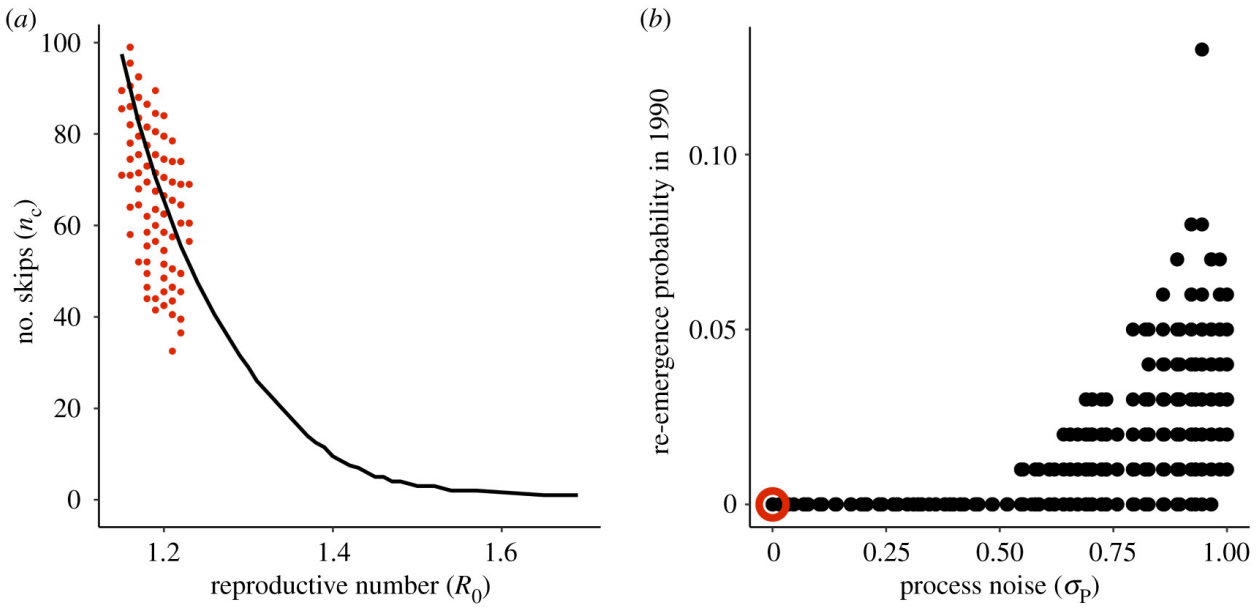

Figure 3.5: **A) Expected number of skips (n_c) calculated using parameters obtained from the fitted stochastic model.** The open circles show the expected number of skips n_c from Equation 1 using parameters and the fraction of the population susceptible after the initial DENV1 invasion (s_0) estimated from the fitted stochastic model. Each circle corresponds to one parameter combination, and we included here all parameter combinations for the fitted model with a seasonal transmission amplitude (δ of 0.7 (contacts per person per day) and a likelihood value within two log-likelihood units of the maximum likelihood estimate (MLE). See Fig. 3.20 for expected skips from parameter combinations with different values of δ and Figure S10 for parameter combinations from the profile of the recovery rate, γ. For comparison purposes, the black line shows the expected number of skips for the deterministic skip calculation from panel B of Fig. 3.2 with the reporting rate ρ fixed at the literature value of 3%. **B) Probability of epidemic in 1990 under forward stochastic simulation of fitted model.** The fitted stochastic model was simulated forward in time from 1986-1990 with population growth. A pulse of 20 infected individuals were assumed to arrive each day in January 1990. Each parameter combination within 2 log likelihood units of the maximum likelihood estimate was simulated 100 times. The re-emergence probability was calculated by determining the number of simulations in which the susceptible population decreased in 1990. The plot shows re-emergence probability as a function of the process noise intensity σ_P. Each point represents a single parameter combination. The maximum likelihood estimate parameter combination is circled in red.

Sensitivity Analysis:

To examine the robustness of our findings to adding an incubation period or altering the form of seasonality, we conducted a sensitivity analysis by considering both SIR and SEIR models with spline seasonality. The results are presented and discussed in the Supporting Information and show that our conclusions remain unchanged. (See the Supporting Information including Fig. 3.7-3.12). **Comparison with Vector Model and literature** R_0 The fitted stochastic SIR model uses a cosine function as a simplification to represent the seasonal forcing that would be created by climate variation (temperature [45]) via the changes in infected mosquitoes. To evaluate whether this simplification is realistic, we take two approaches. The first one compares the mean seasonal R_0 resulting from our model to values of this reproductive number directly estimated from time series data in the literature for DENV1 and DENV4 in Rio de Janeiro from 2010-2016. There is a close match between these very different ways to estimate R_0, and in particular the shape of the seasonality produced by our model is realistic (Fig. 3.23).

The second approach considers a simple temperature-driven vector model. To this end, we initially show that the seasonal variation in temperature in Rio de Janeiro can be approximated via a cosine function (Panel A of Fig. 3.24 and use this approximation to drive a transmission rate that includes the vector explicitly. To obtain an expression for the seasonal transmission rate we consider an explicit mosquito model with compartments for infectious and susceptible mosquitoes in which a number of parameters depend on temperature (T) (see Section 4 of the Supporting Information). By assuming fast dynamics of the mosquito (so that levels of infection in the mosquito population quickly equilibrate to the dynamics of infection in the human population), we derive the following expression for the effective transmission rate in the mosquito-human model in terms of the biting rate a(T), probability of human infection given an infectious bite $b(T$, probability of mosquito infection given biting of an infectious human $pMI(T)$, adult mosquito mortality rate μ, carrying capacity K of the

mosquito population, human population size N, and mosquito demographic function $g(T)$:

$$\beta_e ff = (\frac{a(T)^2 b(T)(pMI(T))}{\mu_M}) \frac{K}{/} N(1 - \mu_M / g(T)) \tag{3.2}$$

The function $g(T)$ is the product of the eggs laid per female mosquito per gonotrophic cycle, the mosquito egg-to-adult survival probability, and the mosquito egg-adult development rate divided by the adult mosquito mortality rate μ_M The temperature-dependence of these components was borrowed from the literature [164, 165] (see Supporting Information Section 4 for details). Under the fast dynamics assumption, this effective transmission rate β_{eff} is an implicit representation of the force of infection inflicted on humans by the vectors of the coupled human-vector model. When re-scaled between 0 and 1, β_{eff} corresponds closely with β_{MLE} the transmission rate from the fitted stochastic SIR cosine model (Panel B of Supplemental Figure S19). This close correspondence indicates that the SIR cosine model is able to capture the shape of the seasonality of DENV1 in Rio de Janeiro.

3.3 Discussion

We developed two lines of evidence regarding the uncertainty and predictability of the time to re-emergence for diseases with low reproductive numbers, on the basis of a seasonally forced SIR model under the 'well-mixed' assumption at aggregated, city-wide, scales. We showed with an analytical approach that the time to re-emergence (expressed as the number of "skip" years) was highly sensitive to small changes in R_0 and the fraction of the population still susceptible s_0 at the time of prediction (e.g. at the end of the initial outbreak). This sensitivity applies to dengue in Rio de Janeiro where re-emergence times can vary on the order of decades based on literature parameters. This uncertainty contrasts with previous applications of this analytical approach to SIR dynamics in childhood diseases such as measles with much higher R_0 values where accurate predictions of much shorter skip times have been made [8] . We also showed with a stochastic SIR model with seasonal

transmission fit to DENV1 observed case data for Rio de Janeiro from 1986-1988 that susceptible depletion and replenishment are insufficient to explain dengue re-emergence. The fitted model failed to predict by far the re-emergence of DENV1 in 1990 in terms of either the number of skips expected or the outbreak probability under forward simulation. Transmission parameters like R_0 are generally defined with respect to a particular model. Given that we aggregated cases at the city level and used a short time series, care should be taken in interpreting parameter values. Nevertheless, fitted transmission parameters correspond well with literature values and exhibit well-defined confidence intervals. Estimates of the reporting rate in particular closely match the 3% value [5] obtained via a serological study conducted during the 1986 invasion [5, 7]. Reporting rates during the onset of an epidemic may be much lower in regions that have not recently experienced transmission [7, 166] than in those with re-occurring outbreaks and an established surveillance network. This may explain why serological studies of the 1986 invasion [5, 7] and our results, estimate a lower reporting rate for dengue than studies conducted in later years in Brazil [167]. Even though different combinations of the transmission rate and duration of infection can yield the same reproductive number, the parameter estimates that compose R_0 across all models considered in the sensitivity analysis (which take into account those different combinations) are relatively well-defined. These values are also consistent with the effective reproductive number estimated for local dengue epidemics from 2012-2016 [9] and 1996-2014 [163] taking into account differences in serotype circulation and population size during those periods. More complex model structures are possible and often used for arboviruses that include an explicit representation of the vector. We expect our results to hold as this vector component should largely affect the phase and shape of seasonality in the transmission between human hosts, which we have modeled phenomenologically as a cosine wave. With two typical successive epidemic years from an emergent virus, parameter inference from such short observation period is unlikely to justify a more complex model. Nevertheless, to examine transmission seasonality further, we compared the seasonal R_0 resulting from the fitted

model to the seasonal R_0 directly estimated from time series of cases in the literature [9]. We also considered the transmission rate experienced by humans in a simple vector-human model forced by the typical seasonality of temperature in Rio de Janeiro. The shape and timing of the vector-human model's transmission rate was comparable to that of the cosine transmission rate we employed. More complex models that do not assume fast dynamics of infection in the vector relative to epidemic spread would likely exhibit a difference relative to our transmission rate, especially a delayed phase, whose consequences should be examined in future work. We posit that this difference would not influence our results on the predictability and uncertainty of re-emergence, since the values of other parameters (such as the length of infection in humans) can compensate for it. Factors that could explain the observed rapid re-emergence include inter-annual climate anomalies, antigenic evolution, or micro-scale spatial heterogeneity in transmission intensity and associated susceptible depletion. Larvae washout following flooding coupled with temperature-driven seasonality in transmission could have temporarily halted the invasion in 1988 and delayed the epidemic in 1989. Widespread flooding was reported in February 1988 [168]. Large amounts of rainfall washed away mosquito larvae in lab and field studies [26]. High rainfall negatively affected dengue transmission in Singapore [27, 28] and India [29]. The impact could be compounded in Rio de Janeiro if the high rainfall occurs during the transmission season. If the larvae population has not fully recovered before the start of the off-season, the impact of the rainfall anomaly could extend to the subsequent season. The large amount of process noise observed in the aggregate model would be consistent with this effect, given that the process noise parameter σ_{P} represents random variation in the transmission rate due to environmental factors. However, the model's inherent structure limits its ability to take into account flooding events via σ_{P}, since the magnitude of the process noise does not change between years. Incorporating an inter-annual climate driver could provide more accurate re-emergence predictions. The response to rainfall would be nonlinear: positive at low to moderate levels and negative at higher ones. Intra-serotype antigenic evolution from 1986-1990 could also

facilitate faster re-emergence. Many models focus on inter-serotype variation and assume long-lasting homosubtypic immunity [46, 32, 151]. However, the antigenic variation within and across dengue serotypes is comparable [43], and antigenic differences between strains of the same serotype influence overall dengue evolution [169]. Sequences associated with case data were unavailable, making direct analysis challenging. We cannot rule out the possibility that genetic differences between the circulating strains enabled re-infection. A future SIRS-type model (Susceptible-Infected-Recovered-Susceptible) could incorporate this intra-serotype antigenic evolution. Micro-scale spatial heterogeneity in transmission intensity and the effects of human movement between neighborhoods could also explain the rapid re-emergence. Small-scale differences in socioeconomic status and population density between neighborhoods in a large city can result in different relationships between mosquito and human population sizes, resulting in widespread heterogeneity in R_0 across neighborhoods [38]. Previous studies of mosquito trap data in the city have demonstrated that neighborhoods with differing socioeconomic characteristics have different vector population patterns [45]. In fact, schoolchildren from neighborhoods with divergent socioeconomic characteristics had varying levels of seroconversion during the 1986 invasion [7]. Human movement between neighborhoods may also influence transmission within [39] and between [170] those neighborhoods, potentially resulting in non-uniform depletion of susceptible populations between highly connected and isolated areas of a city. Whether arising through the effects of spatial heterogeneity in transmission or intra-city movement, non-uniform levels of herd immunity could enable faster re-emergence. Our findings reveal the uncertainty of re-emergence predictions with the simplest SIR models, those that would be most useful at times of emergent public health threats. Consideration of the above factors in transmission models whose goal is to inform public health over large regions, and to do so soon after, if not during, an emergent outbreak, is clearly a challenge. For example, coarse resolutions are typically used because of the scales at which the observed cases are reported, the scales at which the climate covariates are available, and the difficulties inherent in incorporating microscale variation in-

cluding connectivity. Our results should motivate further research into the central question of how we can scale microscale heterogeneity to formulate aggregated models that include it implicitly. It should also motivate the related further understanding of how such microscale heterogeneity influences susceptible depletion and replenishment in particular case studies. From such efforts, we should be able to evaluate whether the increasing availability of high-resolution data makes it feasible to parameterize transmission models at higher resolutions, or to inform new model formulations at coarser resolutions. The inability of susceptible depletion and replenishment in a simple seasonal SIR formulation at a large, city-wide scale, to explain DENV1 re-emergence has potential implications for other arboviruses. Recent long-term Zika forecasts [52] assume that susceptible depletion and replenishment brought an end to the 2015-2017 epidemics and will determine when re-emergence occurs. DENV1 and Zika share the same vector and invaded a completely susceptible population (not accounting for pre-existing cross-immunity from dengue). If factors absent from the basic model were key drivers of DENV1 inter-annual variability, it would not be unreasonable to infer that similar types of factors could have played a major role in the Zika dynamics observed from 2015-2017. Zika re-emergence could similarly occur much earlier than expected. With changing temperature patterns due to climate change, cities in Asia, Europe, and the western hemisphere that currently do not have recurrent local transmission may transition in the near future to the kinds of dynamics studied here. Our results suggest that estimates should be interpreted in the context of this sensitivity to small changes in the reporting rate and reproductive number. Factors like variation in reporting rates, micro-scale transmission heterogeneity and inter-annual climate drivers that are often ignored in long-term forecasts may thus become critical in determining re-emergence times. Overall, the large uncertainty in re-emergence times may be unavoidable for these regions. Improved models are needed together with richer data than currently used, to address the question of the relevant spatial scales of susceptible depletion.

3.4 Methods

The derivation of the expression for the number of skip years (Equation 1) is included in Section 1 of the Supporting Information. We fitted a stochastic version of the SIR model to observed monthly case counts in Rio de Janeiro from 1986-1988 to estimate parameters needed to apply this expression, and also to separately predict in parallel the time to re-emergence via numerical simulation Expected re-emergence times were then compared for the two approaches.

3.4.1 Data Description

We used monthly dengue case estimates in the city of Rio de Janeiro, Brazil from 1986-1990. Cases were reported to the local public health surveillance system [6, 19] . The case counts did not contain serotype information, but prior studies indicated that the dengue serotype DENV1 invaded the city of Rio de Janeiro in 1986 [19] and was the dominant serotype in circulation in the state of Rio de Janeiro from 1986-1990 [5] prior to the arrival of DENV2 in 1990. DENV2 did not become dominant until 1991 [5].

3.4.2 Basic Model Formulation

Because dengue infection confers full immunity to the same serotype, we considered an SIR (Susceptible-Infected-Recovered) model. The deterministic model for the number of individuals in the Susceptible (S), Infected (I), or Recovered (R) class is given by the following system of ordinary differential equations:

$$\frac{dS}{dt} = rN - \lambda(t)S - \mu_{\mathrm{H}}S \tag{3.3}$$

$$\frac{dI}{dt} = \lambda(t)S - \gamma I - \mu_{\mathrm{H}} I \tag{3.4}$$

$$\frac{dR}{dt} = \gamma I - \mu_{\mathrm{H}} R \tag{3.5}$$

$$\lambda(t) = \beta(t)\frac{I}{N} \tag{3.6}$$

$$\beta(t) = \beta_0(1 + \delta sin(\omega t + \phi)) \tag{3.7}$$

Deaths occur at rate (μ_{H}) given by the inverse of the life expectancy of Brazil in 2012 (74.49 years [2]). All individuals are born susceptible. The term r represents population growth. The human population growth rate was estimated from census resident population estimates in 1991 [3] and 2000 [4] assuming exponential growth. This rate was used to interpolate the estimated population in 1986 (See Supporting Information Section 2.1.1 for details). The per capita rate at which susceptible individuals become infected was given by the force of infection $\lambda(t)$ (Equation 5). Individuals recovered at per-capita rate γ whose inverse is the duration of infection. Estimates of the duration of infection in dengue vary. One analysis estimated that symptoms of dengue infection last 2-7 days following an incubation period of 4-10 days [171, 172]. For our analysis, we fixed the recovery rate γ to be 1/17, assuming an exponentially distributed duration of infection with mean of 17 days encapsulating the maximum extent of the combined incubation and symptomatic period in humans.

We take into account the possibility that duration of infection could vary by profiling over the duration of infection in the sensitivity analysis. The short duration of the available time series meant that fitting a formal vector model could prove difficult and could require additional assumptions in terms of which parameters could be fitted or fixed from existing formulations in the literature. We therefore used an SIR framework in which the infected stage served as a proxy for the exposed and infected human and vector compartments in a vector model of dengue transmission. A duration of infection was thus chosen that corresponds to the upper bound of the estimated pre-infectious period (4-10 days) and infectious period (2-7 days) in humans [171, 172]. We profiled over the duration of infection in the sensitivity analysis to verify that this parameterization is reasonable. This transmission rate $\beta(t)$ was represented as a cosine function with mean β_0 (units of contacts per person per day) and seasonal oscillations of amplitude δ(same units as β_0) and frequency ω, which was assumed to be annual ($\omega = \frac{2\pi}{365}$) days^{-1}. The annual mean R_0 was thus given by:

$$R_0 = \frac{\beta_0}{\gamma + \mu} \tag{3.8}$$

The observed dengue data in Rio de Janeiro consisted of monthly case counts. Serological studies of the DENV1 invasion in Rio de Janeiro also indicated substantial under-reporting [5, 7]. Let C represent the true number of monthly cases that would be obtained by summing the number of individuals entering the infected class (I) over the course of a month. For the purposes of the skip analysis, we assume that a fixed fraction ρ of the true cases C are observed, where ρ is the reporting rate.

The stochastic model is an approximation of the deterministic one used for the skip analysis. For simplicity and given the short time interval, we assumed that there was no population growth over the two and half years of the DENV1 invasion ($r = \mu_{\mathrm{H}}$) and that births and deaths occurred at rate $\mu_{\mathrm{H}} = (1/(74.9 * 365))$, which is equal to the inverse of

the average life expectancy in Brazil from the 2010 census [2]. However, population growth is taken into account when simulating forward in time from the fitted stochastic model. We also assumed that there were no recovered individuals at the start of the epidemic, so all other individuals in the population not initially infected were susceptible. We considered time in units of days and used a time step Δt of 1 day.

The stochastic model is a discrete-time model with fixed time step Δt and a discrete state space (i.e. the number of people in each compartment S, I, R, and C, at any point in time must be integers). The number of individuals who moved from one compartment to another over the course of each day was calculated via Euler simulation from the deterministic equations (See Supporting Information). Demographic stochasticity was then incorporated into the Euler approximations to obtain integer state variable values after each time step. We specifically assumed that the number of individuals making each state transition was drawn from a binomial distribution with exponentially decaying probability (See Supporting Information). Environmental noise (variation in the transmission rate $\beta(t)$ due to random environmental variation) was captured via multiplicative gamma white noise in the transmission rate as described by [173, 174]. On time step size Δt, we multiplied the transmission rate by $\frac{\Delta\Gamma}{\Delta t}$ where $\frac{\Delta\Gamma}{\Delta t}$ was drawn from a Gamma distribution with mean 1 and variance $\frac{\sigma_{\mathrm{P}}^2}{\Delta t}$.

The measurement model assumed that the observed number of monthly dengue cases ($Y(t)$) at time t were drawn from a negative binomial distribution with mean equal to the true number of monthly cases C multiplied by a reporting rate ρ, with dispersion parameter σ_{M}. More details of the measurement model can be found in Section 2.4 of the Supporting Information.

3.4.3 Fitting the stochastic model

We fitted the transmission parameters (β_0 and δ), reporting rate (ρ, process noise parameter (σ_{P}), measurement noise parameter (σ_{M}, and the number of infected individuals at the start of the outbreak in May 1986 (I_0). While the first cases of DENV1 were reported in April

1986, we started the model fitting in May 1986 to avoid complications from changes in the reporting rate as the surveillance system was established during the start of the DENV1 invasion. We used in an interpolated initial population size of 5,281,842 for Rio de Janeiro in May 1986.The model was fit using the mif2 method in the R-package pomp. The model fitting method is described further in the Supporting Information and in [175]. Calculating expected skips using parameter estimates from stochastic model Following the completion of the Monte Carlo Profiles, a maximum likelihood estimate (MLE) parameter combination was obtained from the Monte Carlo Profiles of the fitted model by selecting the parameter combination with the highest likelihood across all profiles. The table of MLE parameter values is shown in Supplemental Table ST1. All sets of parameter combinations within 2 log-likelihood units of the maximum likelihood estimate (from all profiles) were used for the expected skip calculation. The reporting rate (ρ), β_0, and δ value of each parameter combination within 2 log likelihood units of the maximum likelihood estimate were applied to a finer gridded version of the deterministic skip calculation described earlier. A distribution for the number of skips expected in Rio de Janeiro following the DENV1 invasion from 1986-1988 was obtained.

3.4.4 Stochastic Simulation

We then simulated re-emergence probabilities under the stochastic model. Each parameter combination within 2 log likelihood units of the MLE estimate from the stochastic fit was simulated again without any immigration from 1986 until 1990 but with population growth. During January 1990, "sparks" of infectious individuals were assumed to have arrived in the city at some fixed rate. There were low but non-zero levels of DENV1 incidence from 1988-1989. We chose to wait until January 1990 before introducing new DENV1 infections to be conservative, as this is when an uptick in DENV1 incidence was first observed. Had we introduced sparks earlier in 1988-1989, we would likely have observed even earlier re-emergence times. We explored rates from 5 to 100 infected individuals per day. This process

was repeated 100 times, and the probability of an epidemic occurring in 1990 was calculated. An epidemic occurrence in this situation was defined as a net decrease in the susceptible population over the course of the year (after taking into account population growth), to best match the definition of an epidemic used in the skip analysis.

3.4.5 Sensitivity Analysis

We assessed how parameter estimates of R_0 and ρ may depend on the model formulation by fitting several more complex SIR-type models to the same data using the fitting procedure described in the Methods section: an SIR Spline Model and SEIR Spline Model. As an additional sensitivity analysis, we profiled over the recovery rate for the SIR Cosine Model (Fig. 3.14). For details, see the Supporting Information.

3.4.6 Comparison with Vector Model and literature R_0

For a full description of the explicit coupled human-mosquito model with compartments for infectious and susceptible mosquitoes and comparison of transmission rates between this model and the simpler seasonally forced SIR , see the Supporting Information.

3.5 Author Contributions

R.S., V.R.-A. and M.P. designed the experiments. R.S. and V.R.-A. conducted the experiments. R.S. and E.I. developed the stochastic model fitting pipeline. C.T.C. provided the data and expertise on dengue in Rio de Janeiro. R.S., V.R.-A. and M.P. drafted the manuscript. All authors contributed to the writing of the manuscript.

3.6 Funding

R.S. was supported by a National Science Foundation Research Traineeship (no. 1735359: NRT-INFEWS: Computational data science to advance research at the energy environment

nexus). M.P. and E.I. were supported by a collaborative grant from the National Science Foundation's Division of Mathematical Sciences and the National Institutes of Health (no. 1761612: Collaborative Research: Urban Vector-Borne Disease Transmission Demands Advances in Spatiotemporal Statistical Inference). V.R.-A. was jointly supported by this grant and by the Mansueto Institute for Urban Innovation through a Mansueto Institute Postdoctoral Fellowship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the National Science Foundation.

3.7 Acknowledgments

The authors would like to thank Aaron King for his advice, and two referees for their insightful comments. This work was completed with resources and support provided by the University of Chicago's Research Computing Center.

3.8 Supporting information

3.8.1 Supplemental Methods: Analytical Skip Derivation

Background

This skip analysis is based on [8]). The goal is, given the number of susceptible after an outbreak S_0, predict how long we should expect for a new epidemic, or equivalently, the number of skipping years that will occur. In Stone et al [8]), a skipping event is when both the number of infectious and susceptible individuals increase in time. However, if the number of infectious individuals increases while the population of susceptible decreases, the event is defined as an outbreak.

A year is taken to have only two seasons and it starts in the Low season at times $t_{L\,n} = nT$

with $n \in \mathbb{N}_0$ and $T = 1\,year$. This is followed by the High season at times $t_{H\,n} = \frac{T}{2}(2n+1)$. Then, assuming $\beta = \beta_0(1 \pm \delta)$ (with $\beta = \beta_0(1-\delta)$ for the Low season and $\beta = \beta_0(1+\delta)$ for the High season), and no human population growth ($N = N_0 = const$, since the mortality and birth rates are equal, $\mu = \lambda$), they obtain for the critical threshold for the susceptible population S_0

$$\frac{S_c(k)}{N} = s_c(k) = \frac{\gamma + \mu}{\beta_0} - \frac{(k+1)\mu}{2} \tag{3.9}$$

Finally, if $S_c(k+1) < S_0 < S_c(k)$, the disease will skip k years and an outbreak will emerge on the *k+1-th* year. For details on how they obtained Eq. (3.9) see the Supplementary Information of Stone et al [8]).

The dynamical system

The SIR model Equation (3.10 shows the SIR infectious model, where λ and μ are the birth and mortality rates, β the transmission rate and γ the recovery rate.

$$\begin{aligned} \frac{dS}{dt} &= \lambda N - \beta(t)\frac{SI}{N} - \mu S \\ \frac{dI}{dt} &= \beta(t)\frac{SI}{N} - \gamma I - \mu I \\ \frac{dR}{dt} &= \gamma I - \mu R \end{aligned} \tag{3.10}$$

From Eq. (3.9) the nullcline curve of S is defined as:

$$dS/dt = 0 = \beta S\left(\frac{\lambda N - \mu S}{\beta S} - \frac{I}{N}\right) \Rightarrow \frac{I*}{N} = \frac{\lambda - \mu S/N}{\beta S/N} \tag{3.11}$$

After an outbreak, $I << I*$ and the trajectory is below the S nullcline. Under this slow build dynamics of the system the susceptible differential equation on Eq. (3.10) may be approximated by

$$\frac{dS}{dt} = \beta\frac{SI*}{N} = \lambda N - \mu S \tag{3.12}$$

Calculation of the critical value, S_C

The assumptions To obtain the critical value, $S_C(k)$ (Eq. (3.9)), Stone et al [8] consider no demographic growth, that is, $\lambda = \mu \Longrightarrow N(t) = N_0\, \forall t$. With this assumption the nullcline curve (Eq. (3.11)) is given by $i* = \mu(1-s)/(\beta s)$, where $i = I/N$ and $s = S/N$. Then they approximated $i*$ by

$$i* = \frac{\mu(1-s)}{\beta s} \simeq \frac{\mu}{\beta s} \Longrightarrow \frac{ds}{dt} \backsimeq \mu \tag{3.13}$$

In this paper we proceed as in Stone et al. [8] but considering:

1. A sinusoidal transmission rate, $\beta = \beta_0(1 - \delta sin(\omega t))$ (see Fig. (3.22)).
2. Population demography, $N = N(t)$, $\lambda \neq \mu$.
3. And a $I*$ without approximation. The approximation on Eq. (3.13) is not always applicable and in particular gives some contradictory results when $r_0 = \beta_0/(\gamma+\mu) < 1$. Specifically the threshold s_c shown on Eq.(3.9) allows outbreaks even when $r_0 < 1$.

The calculation When $I << I*$, that is, in the period of the slow build dynamics, the set of differential equations of the system is

$$\begin{aligned} \frac{dS}{dt} &= \lambda N - \mu S \\ \frac{dI}{dt} &= \beta(t)\frac{SI}{N} - \gamma I - \mu I \\ \frac{dN}{dt} &= (\lambda - \mu)N \end{aligned} \tag{3.14}$$

where, because $N = S + I + R$, we have replaced the equation of the number of individuals recovered by the equation of the total population. By defining, $Z = \frac{\beta_0}{\gamma+\mu}S = r_0 S$ and

$X = \ln(I)/(\gamma + \mu)$, and solving the equation of $\dot{N}$ with $N_0 = N(t = t_0)$, we obtain the following system, equivalent to Eq. (3.14)

$$\begin{aligned} \dot{Z} + \mu Z &= r_0 \lambda N = r_0 \lambda N_0 e^{(\lambda-\mu)(t-t_0)} \\ \dot{X} &= (1 - \delta \sin(\omega t)) \, \tfrac{Z}{N_0} e^{-(\lambda-\mu)(t-t_0)} - 1 \end{aligned} \tag{3.15}$$

The variable Z in Eq. (3.15) is described by a first order differential equation which can be solved with standard techniques, and Eq. (3.16) shows the result

$$Z(t) = (Z_0 - r_0 N_0) e^{-\mu(t-t_0)} + r_0 N(t) \tag{3.16}$$

where $Z_0 = Z(t = t_0)$. By replacing Eq. (3.16) on Eq. (3.15), defining $z_0 = Z_0/N_0$, and taking $t_0 = 0$, we obtain

$$\dot{X} = (r_0 - z_0) e^{-\lambda t} \delta \sin(\omega t) - \delta r_0 \sin(\omega t) - (r_0 - z_0) e^{-\lambda t} + r_0 - 1 \tag{3.17}$$

Then, the strategy is to evaluate the sign of the integral between t_0 (which has been taken as 0 without lost of generality) and $t_{H\,n} = \frac{T}{2}(2n + 1)$ (the high season starting time) of Eq. (3.17). In particular, by zeroing this integral, the following expression for S_C is obtained,

$$s_C = \frac{S_c}{N_0} = 1 + \frac{\pi(2n + 1)(1 - 1/r_0) - 2\delta}{\omega f(\omega, \delta, \lambda, n)} \tag{3.18}$$

where $f(\omega, \delta, \lambda, n) = \left(1 + e^{-\lambda t_{H\,n}}\right) \frac{\omega\delta}{\omega^2 + \lambda^2} - \left(1 - e^{-\lambda t_{H\,n}}\right) \frac{1}{\lambda}$.

3.8.2 Supplemental Methods: Stochastic Model Description

Parameterization

The stochastic model is an approximation of the deterministic model used for the skip analysis. For simplicity and given the short time interval, we assume that there is no population growth over the two and half years of the DENV1 invasion ($r = \mu_{\mathrm{H}}$) and that births and deaths occur at rate $\mu_{\mathrm{H}} = \frac{1}{74.49*365}\mathrm{day}^{-1}$ (the inverse of the average life expectancy in Brazil according to the 2010 census [2] expressed in units of days). Note that population growth is re-introduced in numerical simulations below that consider a longer time period. We assume that there are no recovered individuals at the start of the epidemic, so all other individuals in the population not initially infected are susceptible. The recovery rate γ is fixed at $\frac{1}{17}\mathrm{day}^{-1}$, assuming an exponentially distributed duration of infection with mean 17 days. We then profile over the recovery rate to verify that this fixed value of γ is plausible (see Sensitivity Analysis). We set $\omega = \frac{2\pi}{365}\mathrm{day}^{-1}$. The numerical solution to the stochastic model is a discrete-time model with fixed time step $\Delta t = 1$ day, and a discrete state space (i.e. the number of people in each compartment S, I, and R at any point in time must be integers). During simulation, the number of individuals who will move from one compartment to another over the course of each day is calculated via Euler simulation.

We first describe the Euler approximation for an ordinary differential equation. Then, we explain how we add demographic stochasticity (the consequence of a process acting probabilistically on a collection of discrete individuals) and environmental stochasticity (fluctuations in process rates depending on random, or unmodeled, effects that act at a population level). Finally, we specify how we model measurement noise (error in measuring the true number of dengue cases each month taking into account both a fixed mean rate of under-reporting and additional variation).

Population Growth Recall that the term r represents the per capita rate at which new individuals enter the population, while $\mu_{\mathrm{H}}$is the death rate. Let h represent the per capita

growth rate of the population (i.e. $h = r - \mu_{\mathrm{H}}$). We assume that the net population growth rate is exponential:

$$\frac{dN}{dt} = hN \tag{3.19}$$

Census population estimates from 1991 [3] and 2000 [4] were used to calculate h assuming exponential growth. This rate was used to interpolate the population in 1986. In the implementation of the stochastic model, the population growth over each time step was fed in as a covariate using the R package pomp.

Euler Simulation with Demographic Stochasticity

We simulate the underlying SIR process model with fixed time steps of constant size Δt to obtain estimates of compartment sizes at time t ($\tilde{S}(t)$, $\tilde{I}(t)$, and $\tilde{R}(t)$). The model incorporates demographic stochasticity into the Euler approximations to obtain integer state variable values after each time step. For example, suppose that we want to calculate the number of individuals who move from the susceptible to the infected class over the course of a time step. Let $\mu_{\mathrm{SI}}(t)$ represent the instantaneous rate at which individuals move from the susceptible to the infected compartment. Under the deterministic model, we have:

$$\mu_{\mathrm{SI}}(t) = \beta(t)(\frac{I(t)}{N}) \tag{3.20}$$

Let $\Delta\tilde{N}_{\mathrm{SI}}$ represent the Euler approximation for the number of individuals who move from the susceptible to infected compartment between time step t and time step $t + \Delta t$. We assume that $\Delta\tilde{N}_{\mathrm{SI}}$ is drawn from a binomial distribution with exponentially decaying probability:

$$\Delta\tilde{N}_{\mathrm{SI}} \sim Binomial(\tilde{S}(t), 1 - e^{-\mu_{\mathrm{SI}}(t)\Delta t}) \tag{3.21}$$

All of the instantaneous deterministic transition rates are shown below. The symbol $\bullet$ denotes transitions entering and leaving the study population (i.e. $\bullet S$ denotes births and $S\bullet$ deaths):

$$\mu_{\mathrm{SI}}(t) = \beta(t)(\frac{\tilde{I}(t)}{N}) \tag{3.22}$$

$$\mu_{\mathrm{IR}}(t) = \gamma \tag{3.23}$$

$$\mu_{\bullet\mathrm{S}}(t) = r \tag{3.24}$$

$$\mu_{\mathrm{S}\bullet}(t) = \mu_{\mathrm{I}\bullet}(t) = \mu_{\mathrm{R}\bullet}(t) = \mu_{\mathrm{H}} \tag{3.25}$$

Euler approximations for all compartment transitions are shown below:

$$\Delta\tilde{N}_{\mathrm{SI}} \sim Binomial(\tilde{S}(t), 1 - e^{-\mu_{\mathrm{SI}}(t)\Delta t}) \tag{3.26}$$

$$\Delta\tilde{N}_{\mathrm{IR}} \sim Binomial(\tilde{I}(t), 1 - e^{-\mu_{\mathrm{IR}}(t)\Delta t}) \tag{3.27}$$

$$\Delta\tilde{N}_{\bullet\mathrm{S}} \sim Binomial(\tilde{N}(t), 1 - e^{-\mu_{\bullet\mathrm{S}}(t)\Delta t}) \tag{3.28}$$

$$\Delta\tilde{N}_{\mathrm{S}\bullet} \sim Binomial(\tilde{S}(t), 1 - e^{-\mu_{\mathrm{S}\bullet}(t)\Delta t}) \tag{3.29}$$

$$\Delta\tilde{N}_{\mathrm{I}\bullet} \sim Binomial(\tilde{I}(t), 1 - e^{-\mu_{\mathrm{I}\bullet}(t)\Delta t}) \tag{3.30}$$

$$\Delta\tilde{N}_{\mathrm{R}\bullet} \sim Binomial(\tilde{R}(t), 1 - e^{-\mu_{\mathrm{R}\bullet}(t)\Delta t}) \tag{3.31}$$

Using these expressions, we can write an expression for the number of people in each compartment at the end of each time step (time $t + \Delta t$)

$$\tilde{S}(t + \Delta t) = \tilde{S}(t) - \Delta\tilde{N}_{\mathrm{SI}} + \Delta\tilde{N}_{\bullet\mathrm{S}} - \Delta\tilde{N}_{\mathrm{S}\bullet} \tag{3.32}$$

$$\tilde{I}(t + \Delta t) = \tilde{I}(t) + \Delta\tilde{N}_{\mathrm{SI}} - \Delta\tilde{N}_{\mathrm{IR}} - \Delta\tilde{N}_{\mathrm{I}\bullet} \tag{3.33}$$

$$\tilde{R}(t+\Delta t) = \tilde{R}(t) + \Delta\tilde{N}_{\mathrm{IR}} - \Delta\tilde{N}_{\mathrm{R}\bullet} \tag{3.34}$$

Environmental Stochasticity

As stated in the Materials and Methods section, environmental noise (variation in the transmission rate $\beta(t)$ due to random environmental variation) was captured via multiplicative gamma white noise in the transmission rate as described by [174]). On time step size Δt we multiplied the transmission rate by $\frac{\Delta\Gamma}{\Delta t}$ where $\frac{\Delta\Gamma}{\Delta t}$ was drawn from a Gamma distribution with mean 1 and variance $\frac{{\sigma_P}^2}{\Delta t}$.

The gamma white noise method used in this model is further described in [174]).

Measurement Error

Notation and Observation Times for Observed Data in Stochastic Fitting The observed data $Y_{1:\mathrm{M}}$ represent monthly observed case counts in the city of Rio de Janeiro taken at times $t_{\mathrm{OBS}_1}, t_{\mathrm{OBS}_2} ... t_{\mathrm{OBS}_M}$, where $M = 26$ observations. Observation Y_m in that sequence was taken at time t_{OBS_m} and is the number of observed infections reported between time t_{OBS_m} and time $t_{\mathrm{OBS}_{m-1}}$. Time t_{OBS_0} corresponds to the start time for the simulation of the initial epidemic (May 1, 1986). In the stochastic model simulation, the time variable t is defined as the number of days since January 1, 1986. Therefore, $t_{\mathrm{OBS}_0} = 120$. The first observation used in the fitting is then observation Y_1 measured at time $t_{\mathrm{OBS}_1} = 151$ (June 1, 1986). Note that in the raw data, the date for Y_1 is listed as "May 1986", corresponding to the total number of cases observed by the local health system between May 1, 1986 and May 30, 1986. The last observation used is $t_{\mathrm{OBS}_M} = t_{\mathrm{OBS}_{26}} = 912$ (July 1, 1988).

Monthly Accumulation of Expected Reported Cases Let C_m represent the true number of reported cumulative cases for each month at observation time t_{OBS_m}.

Recall that during the Euler Simulation of the stochastic model, the number of individuals who move from from Susceptible S to Infected I is calculated for every time step of size Δt.

This is the quantity $\Delta\tilde{N}_{\mathrm{SI}}$ from Section 2.3.

For ease of notation, let us use U_t to denote the value of $\Delta\tilde{N}_{\mathrm{SI}}$ calculated for the time step between time t and time $t + \Delta t$.

We can write an expression for C_m as the summation of the $\Delta\tilde{N}_{\mathrm{SI}}$ values over all of those time steps in month m. Formally:

$$C_m = \sum_{t=t_{\mathrm{OBS}_{m-1}}}^{t_{\mathrm{OBS}_m}} U_t \tag{3.35}$$

In the pomp implementation of the model, this issue is dealt with by designating the variable C as an accumulator variable via the accumvars() argument to the pomp constructor. The true number of reported cases are then accumulated for each (monthly) observation.

Measurement Model The measurement model relates how the true number of cumulative cases for each month C_m relates to the observed number of monthly cases Y_m.

The formulation of the model takes into account under-reporting via the reporting rate parameter ρ. However, the true reporting rate may not necessarily be constant in time, for example, as the epidemic spreads to new neighborhoods within the city of Rio or as the surveillance system is constructed and activated. To take into account additional over-dispersion in the observed case data, we incorporate a measurement model in which we assume that the observed number of monthly dengue cases (Y_m) (the observed data) at time t corresponding to the end of a reporting month are drawn from a negative binomial distribution with mean equal to the true number of cases multiplied by the reporting rate ρC_m and dispersion parameter σ_{M}. The mean of the distribution is $\mu = \rho C_m$ while the variance can be written as $\mu + \sigma_{\mathrm{M}}^2\mu^2$. Thus, when the dispersion parameter is close to zero ($\sigma_{\mathrm{M}} \to 0$), the measurement model reduces to a Poisson distribution.

Model Fitting

The model was fit using the mif2 method in the R-package pomp. The mif2 method implements the iterated filtering algorithm known as IF2 and described in [175]). For each parameter being profiled, the profileDesign function in the pomp package was used to generate a set of starting points at 30 different evenly spaced values within prescribed ranges. The function created 40 different initial sampling points drawing from a box given by the boundaries of the original parameter range. For example, for the I_0 profile, a set of 30 starting points evenly spaced between 1 and 10,000 was generated. For each of those 30 starting points, the profileDesign function created 40 different initial sampling points with the same value of I_0 but different values for the other parameters being fitted (β_0, δ, γ, ϕ, σ_{M}, and σ_{P}) where the different values were uniformly drawn from the boundaries for those parameters in the original box. This yielded a total of 1200 starting points for each parameter profile. The mif2 search from each starting point was repeated five independent times.

Sensitivity Analysis

In the sensitivity analysis, we consider two alternate models to the SIR Cosine model: the SIR Spline Model, and the SEIR Spline Model. The SIR Spline Model was an SIR model like the main stochastic model but assumed that the transmission rate was a function of three periodic cubic splines instead of a cosine function (SIR Spline Model). The spline coefficients b_1, b_2, and b_3 were fit instead of the cosine-function parameters. The second alternate model used splines but had an additional exposed class (the SEIR Spline Model) and an additional fitted initial value parameter (E_0). The duration of the incubation period was fixed at 10 days and the duration of infection at 7 days so that both periods sum to the value of the infection period fixed for the SIR Spline and Cosine Models. During the fitting of all four models, the duration of infection was fixed (at 17 days in the SIR Models and 7 days in the SEIR Models). The second part of the sensitivity analysis was a profile of the recovery rate using the SIR Cosine Model. We used a range of recovery rates corresponding to durations

of infection between 2-22 days.

3.8.3 Supplemental Results: Sensitivity Analysis

Adding an incubation period or altering the form of seasonality in the model did not alter parameter estimates for R_0, reporting rate ρ and amplitude of seasonal transmission δ better explain observed dynamics. All three models (SIR with Cosine Seasonality, SIR with Spline Seasonality, and SEIR with Spline Seasonality) gave similar parameter estimates and likelihoods with respect to 2.5 years of observed cases from 1986-1988. The SEIR model had a higher AIC score but a narrower profile for the environmental process noise standard deviation σ_{P} (Supplemental Figure 3.10, Supplemental Tables ST2 and ST3) compared to the flat σ_{P} profiles of other models. Selected profiles for all models are included in Supplemental Figures 3.9, 3.10, 3.11, and 3.12. An additional profile for the transmission phase parameter ϕ for the SIR Cosine Model is included in Supplemental Figure 3.13.

The flatness of the environmental process noise standard deviation (σ_{P} profile and high σ_{P} values for all parameter combinations with non-zero re-emergence probabilities (Figure 5) may both be additional indicators of model misspecification. Analysis of filter trajectories showed that the maximum likelihood estimate parameter combination in the SIR Cosine Model, which did not have high σ_{P} could not easily explain the rapid decrease in dengue transmission in the beginning of the transmission season in 1988 (Supplemental Figure 3.17). Some parameter combinations with high σ_{P}outperformed the maximum likelihood estimate in capturing this decrease (Supplemental Figure 3.17), but under-performed the maximum likelihood estimate in capturing the peak during the second year of the epidemic (Supplemental Figure3.18). More generally, the increased variance of the large process noise parameter combination may impact predictive utility. Varying the duration of infection did not substantially reduce the expected time to re-emergence. During the principal analysis fitting, the duration of infection was fixed at 17 days in the SIR Models and 7 days in the SEIR Models. As an additional sensitivity analysis, we profiled over the recovery rate γ for the SIR

Cosine Model (Supplemental Figure 3.14) with a range of recovery rates corresponding to durations of infection of between 2-22 days. Based on the range of parameter combinations in the profile within 2 log likelihood units of the profile peak, durations of infection between 5 and 20 days for the SIR Cosine model were supported by the data. The expected number of skips was re-calculated using all parameter combinations within 2 log likelihood units of the γ profile maximum and ranged from 25 to over 100 years (Supplemental Figure 3.15). Parameter combinations with re-emergence times longer than 100 years had values of R_0¡1 (Supplemental Figure 3.15), high reporting rates (up to 30%), and substantial process noise (Supplemental Figures 3.15 and 3.16). The support for these deterministically implausible parameterizations may be an artifact of the short length of the time series and the large magnitude of the process noise.

3.8.4 Supplemental Results: Vector Model Considerations

Epidemic model

The mosquito-human coupled dengue model considered follows the equations of [25]) but without the 'exposed' class compartment. Our purpose is to illustrate that explicit consideration of the vector is consistent with the seasonal transmission rate we have adopted in our SIR model, in particular in terms of its basic shape. For this purpose, we rely on a simple representation of the mosquito component, with two classes as in many modeling studies ([176, 177, 178, 179, 180]) . Given the temperature variation in Rio de Janeiro, the duration of the omitted exposed class would be short, and the typical mosquito lifespan would not preclude transmission.

In the model, the mosquitoes M can be susceptible W or infectious Z:

$$\frac{dW}{dt} = g(T)\, M(1 - \frac{M}{K}) - a(T)\, pMI(T)\, W\frac{I}{N} - \mu_M W \tag{3.36}$$

$$\frac{dZ}{dt} = a(T)\, pMI(T)\, W\frac{I}{N} - \mu_M Z \tag{3.37}$$

where a is the biting rate, μ_M is the adult mortality rate and pMI is the probability of transmission. The function $g(T) = EFD(T)\, pEA(T)\, MDR(T)\, \mu_M^{-1}$ depends on mosquitoes parameters: EFD corresponds to the eggs laid per female per gonotrophic cycle (number/female), pEA represents the mosquito egg-to-adult survival probability, and MDR is the mosquito egg-to-adult development rate (1/days). The human population is described by the following equation:

$$\begin{aligned} \frac{dS}{dt} &= -a(T)\, b(T)\, Z \frac{S}{N} \\ \frac{dI}{dt} &= a(T)\, b(T)\, Z \frac{S}{N} - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned} \tag{3.38}$$

We considered a mosquito mortality rate independent from temperature. This choice follows from the known difficulty in mapping laboratory thermal curves for this parameter to those from the field [178]) . In particular, field observations exhibit constancy at intermediate values of temperature (including the range of variation in Rio de Janeiro) rather than the typical bell-shaped curves from the laboratory.

By considering a quasi-static approximation to Eq. (3.37) and $\hat{W} \simeq \hat{M}$ [181], we obtain the following expression for $\hat{Z}$,

$$\hat{Z} \simeq \frac{a(T)\, pMI(T)}{\mu_M} \frac{\hat{M}\, I}{N} = \frac{a(T)\, pMI(T)}{\mu_M} \frac{I}{N} K(1 - \frac{\mu_M}{g(T)}) \tag{3.39}$$

, where $\hat{M} = K(1 - \frac{\mu_M}{g(T)})$ is obtained by zeroing the sum of Eq.(3.36) and Eq.(3.37) since $M = W + Z$. Then, we incorporate Eq.(3.39) in Eq.(3.38) and we obtain the following effective transmission rate β_{eff}

$$\beta_{eff} \simeq a(T)^2\, b(T)\, pMI(T)\, \mu_M^{-1} \frac{K}{N} (1 - \frac{\mu_M}{g(T)}) \tag{3.40}$$

Comparison of the transmission rates

To compare the behavior of the expression shown in Eq. (3.40) with the transmission rate obtained from the fitted SIR cosine model, we use a cosine function to describe the temperature of Rio de Janeiro. The dots in Fig. 3.24A show the weekly temperature reported in [9] and the solid line correspond to $TempSim(t) = 25 + 5\,cos(2\pi t/365 - 0.5)$.

The transmission rate obtained from the cosine temperature and the Eq. (3.40) (values of the parameters are taken from [164] and [165]) is shown on Fig. 3.24B, as well as the transmission rate of the fitted SIR cosine model. The values have been rescaled between 0 and 1 for a better comparison of the curves.

3.9 Supplemental Figures

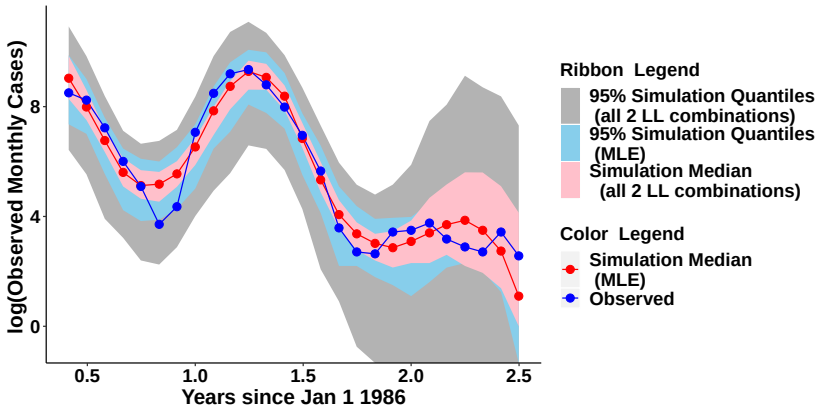

Figure 3.6: **Observed vs simulated cases from parameter combination for the stochastic SIR Cosine Model, fit to 2.5 years of DENV1 case data (fixed recovery rate)** Log of observed monthly cases from April 1986 to June 1988 are shown in blue. Simulated cases were estimated from 100 simulations for each parameter combination within 2 log-likelihood units of the highest likelihood parameter combination (the MLE). Median values from 100 simulations from the MLE the are shown in red. The range of simulation medians across all parameter combinations within 2 log-likelihood units is shaded pink. The shaded blue region denotes the 95% quantile boundaries across all 100 simulations from the MLE. The shaded grey region denotes 95% quantile boundaries from all simulations (across all parameter combinations within 2 log-likelihood units of the MLE).

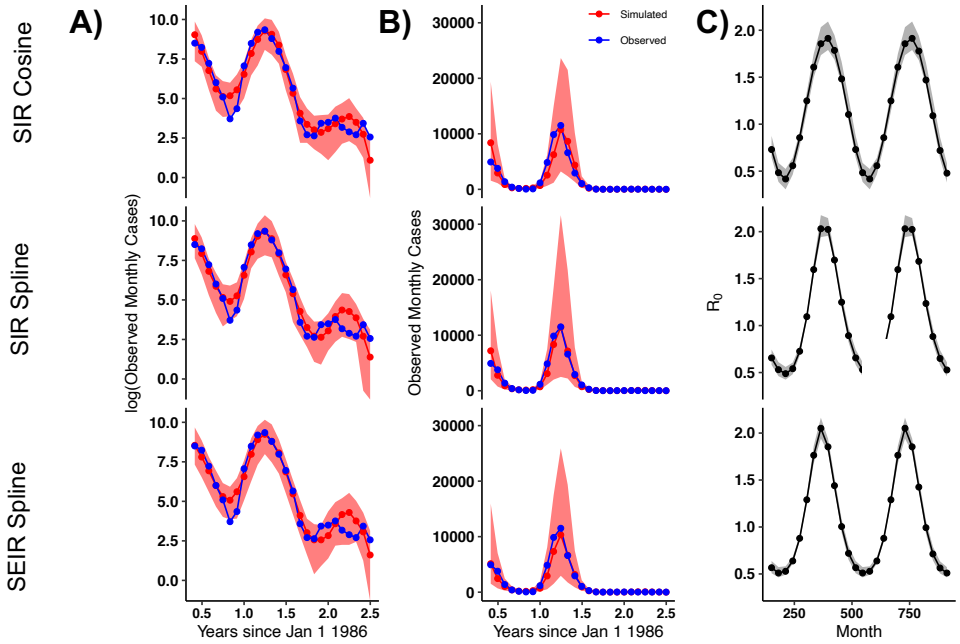

Figure 3.7: **A-B Comparison of simulated values with the fitted model and observed data on a log (A) and regular (B) scale.)** Observed monthly cases from April 1986 to June 1988 are shown in blue. Median values from 100 simulations with the maximum likelihood parameter combination are shown in red. The shaded red region denotes the 2.5% and 97.5%th quantile boundaries from those simulations. **C) Estimates for R_0(t).** The black line denotes the trajectory of R_0(t) for the maximum likelihood estimate. The shaded grey region represents the 2.5% and 97.5%th quantile boundaries for trajectories from all parameter combinations within 2 log likelihood units of the maximum likelihood estimate. Each parameter combination has only one seasonal trajectory for R_0(t) since R_0(t) is a deterministic quantity. Results for three models are shown: the SIR Cosine Model (described in the main manuscript) as well as an SIR Spline and SEIR Spline Model.

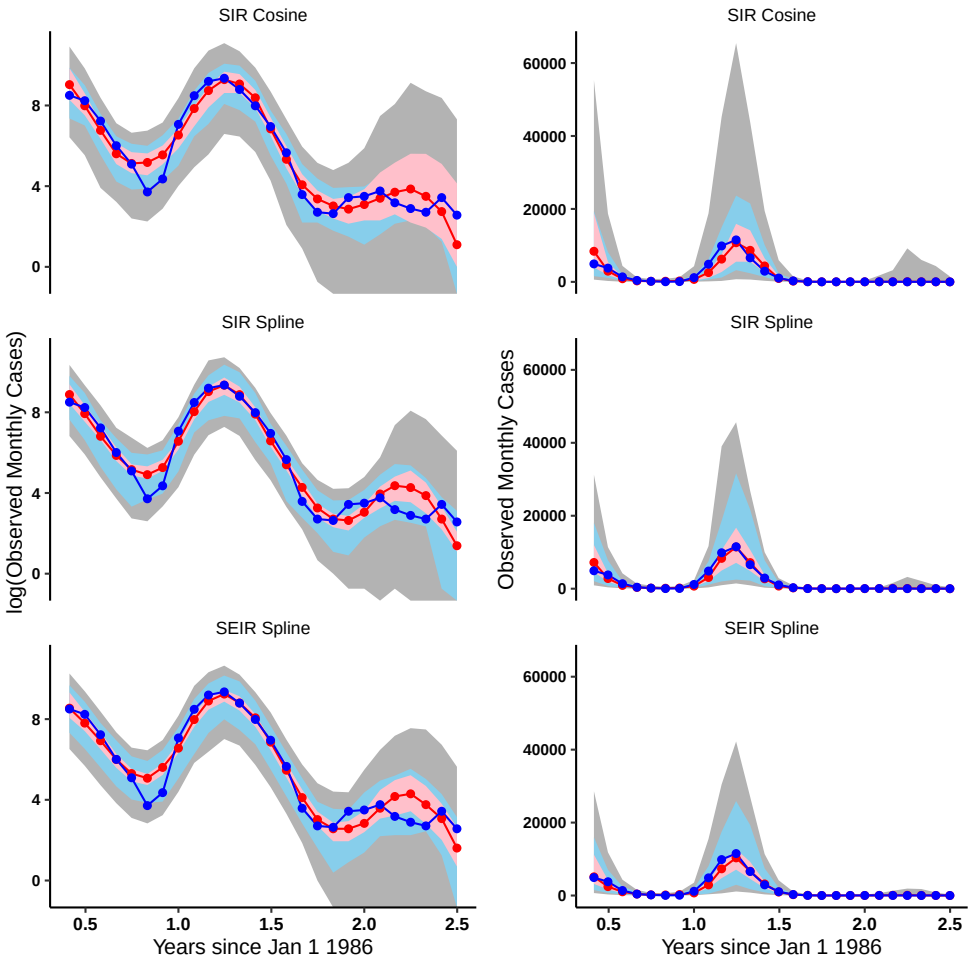

Figure 3.8: **Comparison of all examined parameter combinations within 2 log likelihood units of the maximum likelihood estimate for each fitted model with observed dengue case counts for each of 3 models.** For each parameter combination, 100 independent stochastic simulations were conducted. The right hand panel shows cases on a standard scale, while the left hand panel is on a log scale. Log of observed monthly cases from April 1986 to June 1988 are shown in blue. Simulated cases were estimated from 100 simulations for each parameter combination within 2 log-likelihood units of the highest likelihood parameter combination (the MLE). Median values from 100 simulations from the MLE the are shown in red. The range of simulation medians across all parameter combinations within 2 log-likelihood units is shaded pink. The shaded blue region denotes the 95% quantile boundaries across all 100 simulations from the MLE. The shaded grey region denotes 95% quantile boundaries from all simulations (across all parameter combinations within 2 log-likelihood units of the MLE).

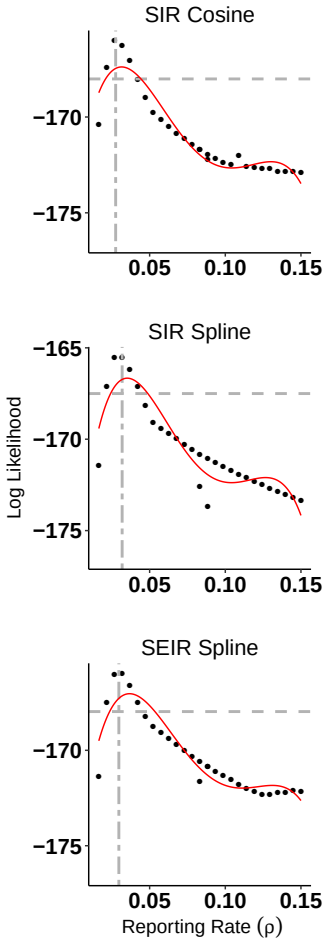

Figure 3.9: **Profiles of reporting rate (ρ) for all three models.** The red curve is a polynomial fit to the subset of the profile points shown on the figure. The single dashed grey horizontal line represents the likelihood value 2 log likelihood units below the peak of the profile. This line provides an estimate of confidence intervals for the given parameter. The grey vertical line denotes the parameter value of the maximum likelihood estimate across all parameter profiles for that model.

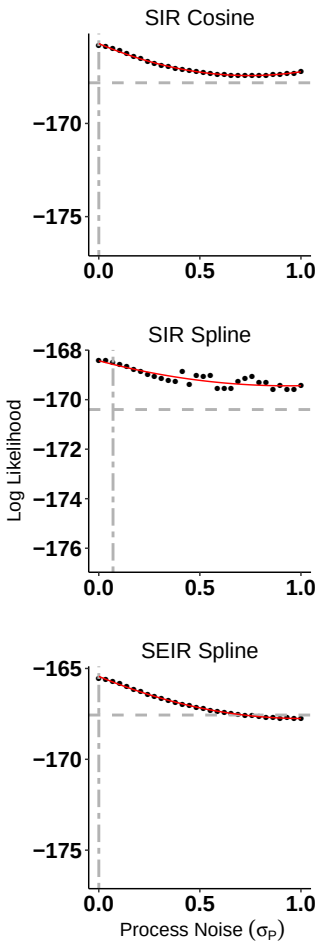

Figure 3.10: **Profiles of the environmental process noise magnitude parameter ($\sigma_{\mathbf{P}}$) for all three models.** The red curve is a polynomial fit to the subset of the profile points shown on the figure. The single dashed grey horizontal line represents the likelihood value 2 log likelihood units below the peak of the profile. This line provides an estimate of confidence intervals for the given parameter. The grey vertical line denotes the parameter value of the maximum likelihood estimate across all parameter profiles for that model.

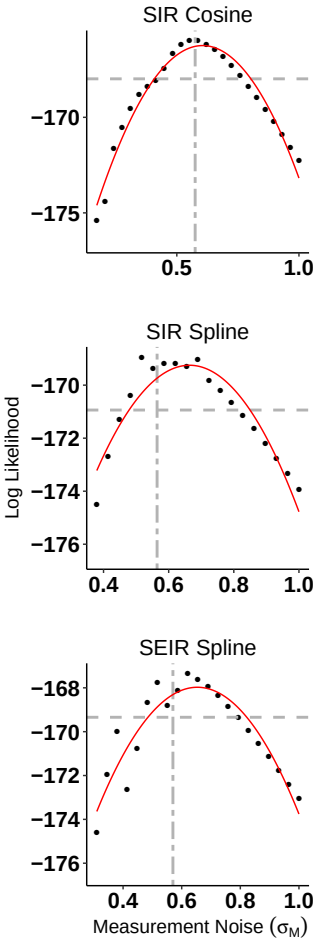

Figure 3.11: **Profiles of the measurement noise magnitude parameter ($\sigma_{\mathbf{M}}$) for all three models.** The red curve is a polynomial fit to the subset of the profile points shown on the figure. The single dashed grey horizontal line represents the likelihood value 2 log likelihood units below the peak of the profile. This line provides an estimate of confidence intervals for the given parameter. The grey vertical line denotes the parameter value of the maximum likelihood estimate across all parameter profiles for that model.

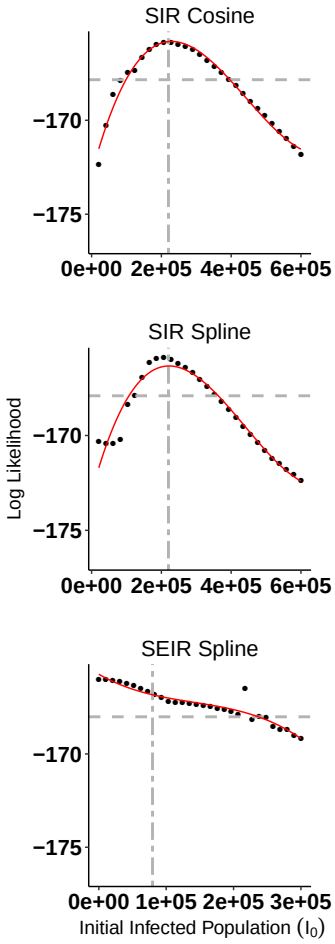

Figure 3.12: **Profiles of the initial number of infected people at the start of the simulation (I_0) for all three models.** The red curve is a polynomial fit to the subset of the profile points shown on the figure. The single dashed grey horizontal line represents the likelihood value 2 log likelihood units below the peak of the profile. This line provides an estimate of confidence intervals for the given parameter. The grey vertical line denotes the parameter value of the maximum likelihood estimate across all parameter profiles for that model.

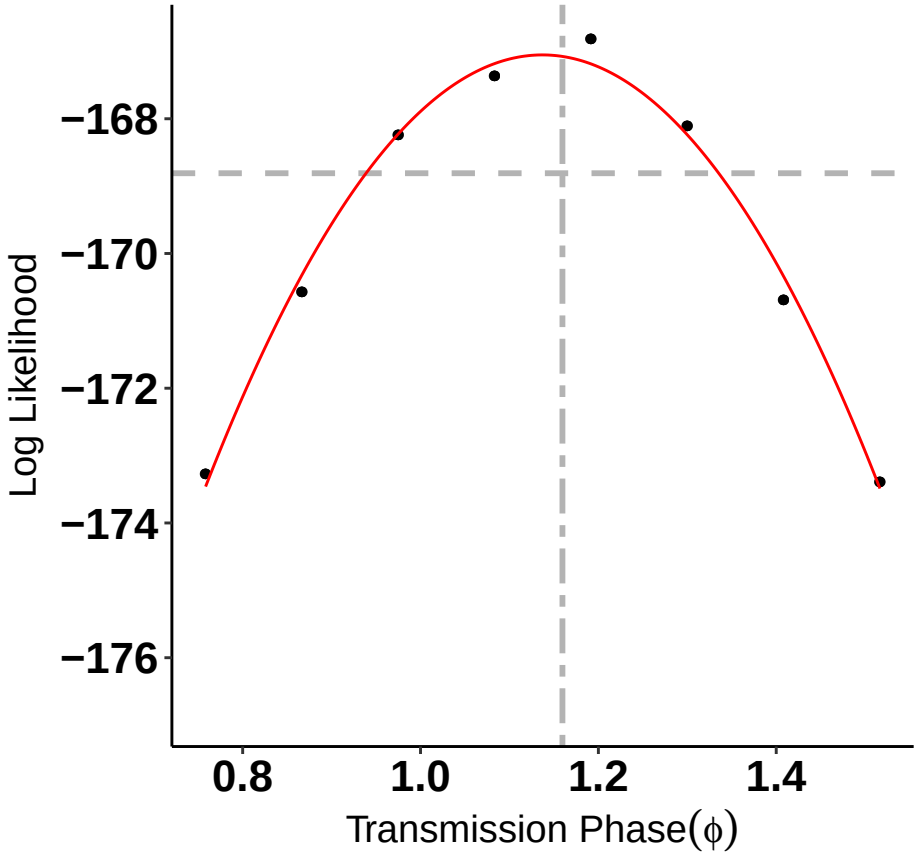

Figure 3.13: **Profiles of the phase parameter (ϕ) for the SIR Cosine Model.** The red curve is a polynomial fit to the subset of the profile points shown on the figure. The single dashed grey horizontal line represents the likelihood value 2 log likelihood units below the peak of the profile. This line provides an estimate of confidence intervals for the given parameter. The grey vertical line denotes the parameter value of the maximum likelihood estimate across all parameter profiles for that model.

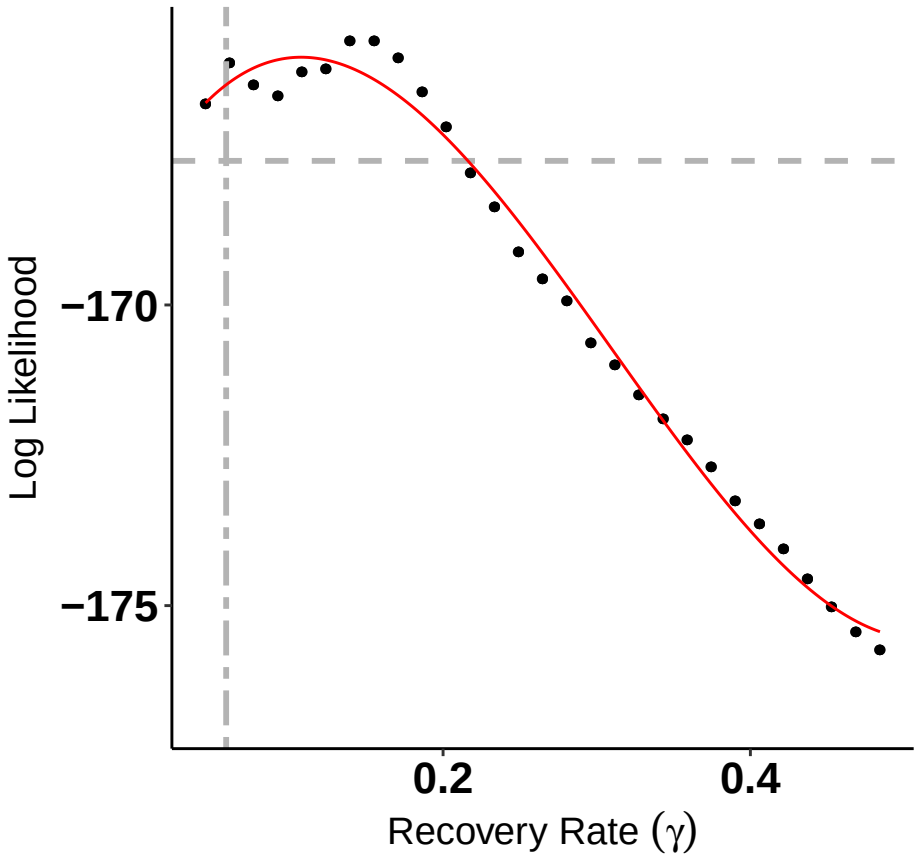

Figure 3.14: **Profiles of the recovery rate (γ) for the SIR Cosine Model.** The red curve is a polynomial fit to the subset of the profile points shown on the figure. The single dashed grey horizontal line represents the likelihood value 2 log likelihood units below the peak of the profile. This line provides an estimate of confidence intervals for the given parameter. The grey vertical line denotes the value of gamma that was used for the original fit of the SIR Cosine Model (where the recovery rate was fixed) .

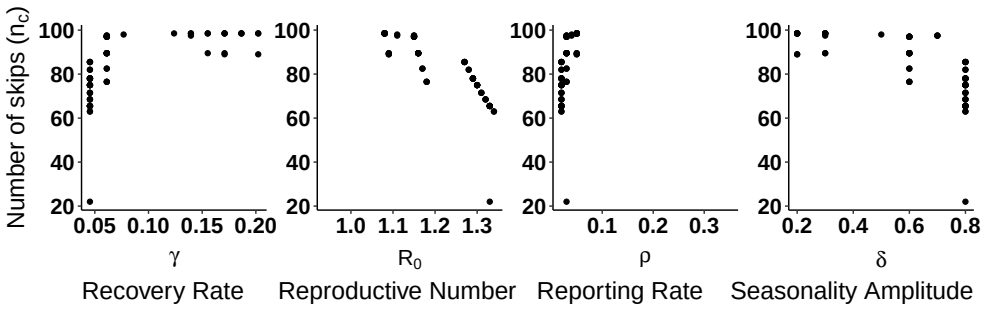

Figure 3.15: **Expected number of skips from deterministic calculation given all combinations of R_0, seasonal transmission amplitude (δ), and reporting rate (ρ) from each parameter combination of the recovery rate (γ) profile within 2 log likelihood units of the parameter combination with the highest likelihood in the original model.** In Figure 5 panel A, expected numbers of deterministic skips were calculated for all parameter combinations within 2 log-likelihood units from the maximum likelihood estimate of the SIR Cosine Model. Parameter values from the profile of the recovery rate (γ) in the sensitivity analysis shown in Supplemental Figure 3.15 were not included in the skip calculations for Figure 5 Panel A. Here, we replicate those skip calculations for all parameter combinations from the γ sensitivity analysis within 2 log-likelihood units of the original maximum likelihood estimate.

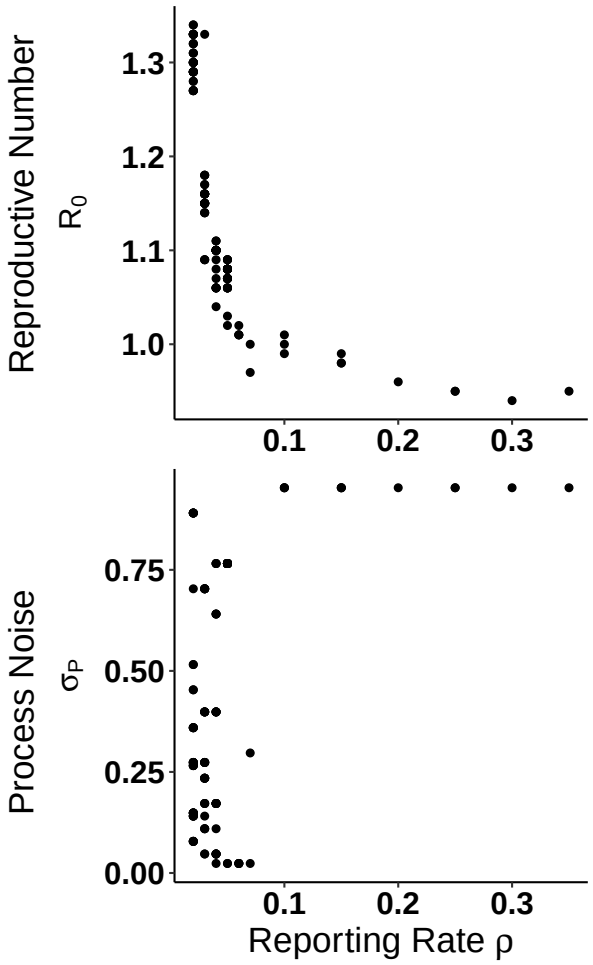

Figure 3.16: **Comparison of reporting rate (*rho*) vs. reproductive number (R_0) and environmental process noise magnitude $\sigma_{\mathbf{P}}$) for each parameter combination of the γ profile of the SIR Cosine Model within 2 log-likelihood units of the parameter combination with the highest likelihood in the original model.** This figure compares the values of R_0 and reporting rate (ρ) and process noise magntiude (σ_{P}) for all parameter combinations from the gamma profile used for the deterministic skip calculations shown in Supplemental Figure 3.15.

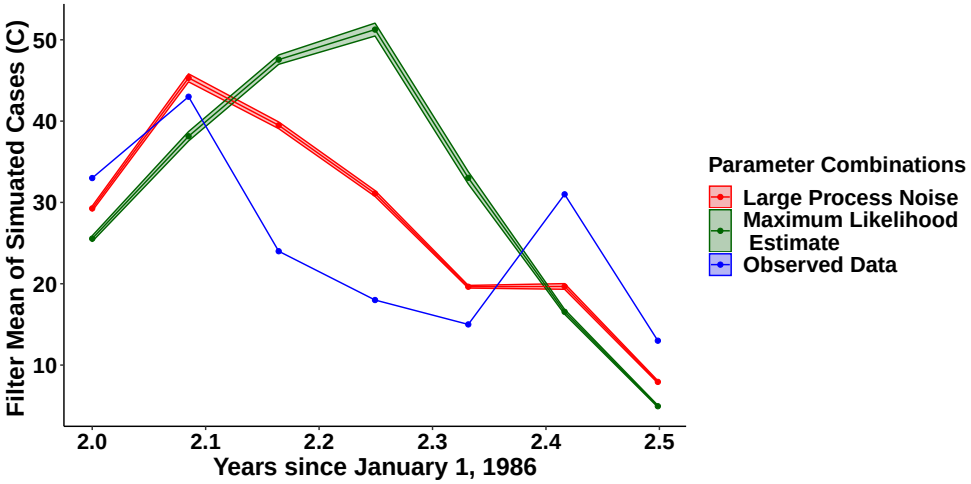

Figure 3.17: **Comparison of filter means between maximum likelihood parameter combination and parameter combination with large process noise over third year of simulation.** Average of filter means for the number of monthly cases (C) at each observed data point from 10 runs of the Sequential Monte Carlo algorithm pfilter run at MLE parameter combination (in green) and at the parameter combination with the highest likelihood out of all parameter combination with the highest permissible amount of process noise ($\sigma_{\mathrm{P}} = 1$) shown in red. The observed cases are shown in blue for comparison. Only the third year of the fit (corresponding to the period from January 1988 through July 1988) is shown. Shaded ribbons show the average of the filter mean +/- 2 times the standard deviation of the filter means (across all 10 runs).

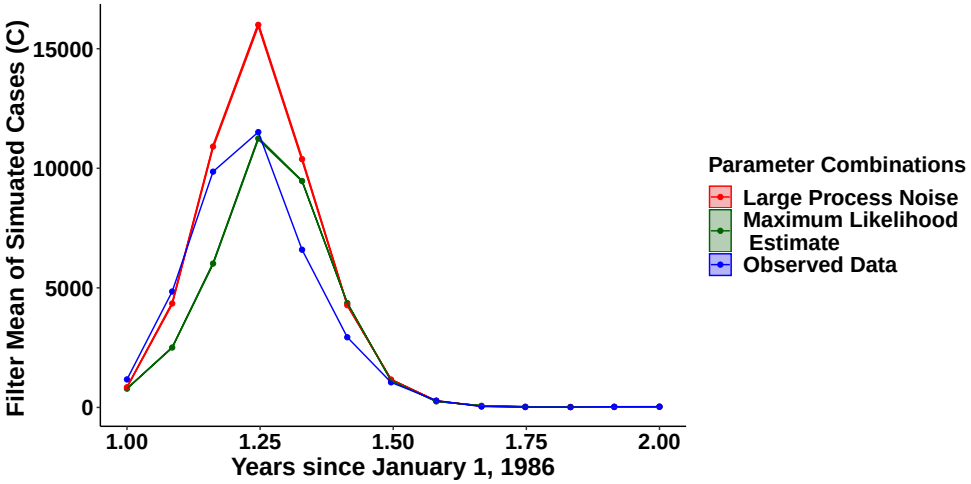

Figure 3.18: **Comparison of filter means between maximum likelihood parameter combination and parameter combination with large process noise over second year of simulation.** Average of filter means for the number of monthly cases (C) at each observed data point from 10 runs of the Sequential Monte Carlo algorithm pfilter run at MLE parameter combination (in green) and at the parameter combination with the highest likelihood out of all parameter combination with the highest permissible amount of process noise ($\sigma_{\mathrm{P}} = 1$) shown in red. The observed cases are shown in blue for comparison. Only the second year of the fit (corresponding to the period from January 1987 through December 1987) is shown. Shaded ribbons show the average of the filter mean +/- 2 times the standard deviation of the filter means (across all 10 runs).

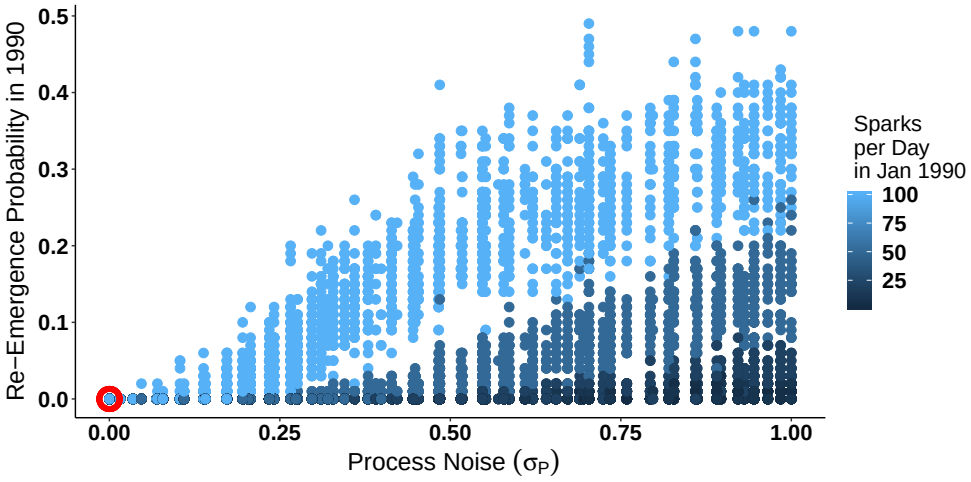

Figure 3.19: **Probability of stochastic epidemic in 1990 vs Process Noise Intensity ($\sigma_{\mathbf{P}}$) under simulation from top 2LL parameter combinations of stochastic SIR Cosine Model.** The fitted stochastic model was simulated forward in time from 1986-1990 with population growth. Daily pulse rates of 2,5,10,20, 50, and 100 infected individuals per day in January 1990 were used. Each parameter combination within 2 log-likelihood units of the maximum likelihood estimate was simulated 100 times. The re-emergence probability was calculated by determining the number of simulations in which the susceptible population decreased in 1990. The plot shows re-emergence probability as a function of the process noise intensity σ_{P}. Each point represents a single parameter combination at a particular pulse rate. Points are colored by pulse rate. MLE parameter combination points are circled in red.

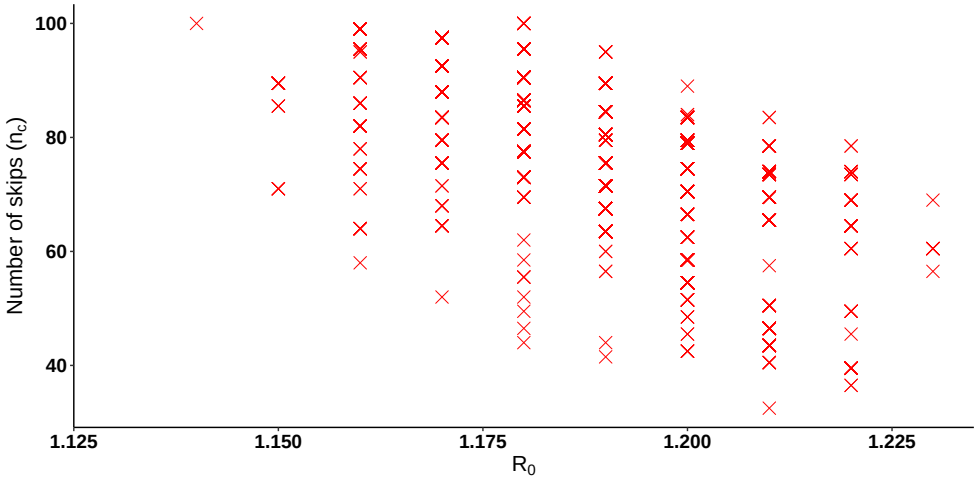

Figure 3.20: **Expected number of skips from deterministic calculation using parameter estimates from the fitted stochastic model.** The red crosses show the expected number of skips nc from Equation 1 of the main text using parameters and the fraction of the population susceptible after the initial DENV1 invasion (s_0) estimated from the fitted stochastic model. Each circle corresponds to one parameter combination, and we included here all parameter combinations for the fitted SIR Cosine model with different values of the reproductive number R_0, seasonal transmission amplitude δ, and reporting rate ρ. See Supplemental Figure 3.15 for the expected number of skips for all parameter combinations obtained from the profile of the recovery rate (γ).

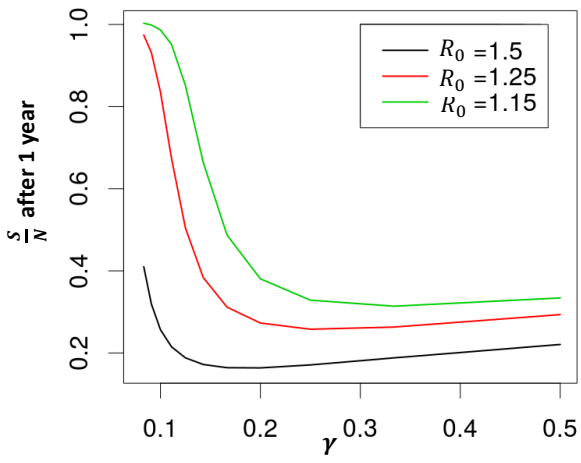

Figure 3.21: **Different combinations of mean transmission rate β_0 and recovery rate γ that yield the same reproductive number R_0 value have different values of the fraction of the population susceptible ($\frac{S}{N}$) after 1 year.** For example, suppose that we have the parameter combinations $\delta = 0.5$, $\beta_0 = 0.3$, $\gamma = 0.2$ and $\delta = 0.5$, $\beta_0 = 0.15$, $\gamma = 0.1$. While both parameter combinations give an R_0 of 1.5, the first yield an $\frac{S}{N}$ after one outbreak of around 20%, while the second gives an $\frac{S}{N}$ of approximately 40%. The plot shows values of $\frac{S}{N}$ as a function of γ with β_0 modified to give different values of R_0. For all points in the plot, the amplitude of seasonal transmission $\delta = 0.5$, the initial number of infected individuals $I(t = 0) = 1$, and the frequency of the seasonality of transmission $\omega = \frac{2\pi}{365}$, corresponding to an annual periodicity.

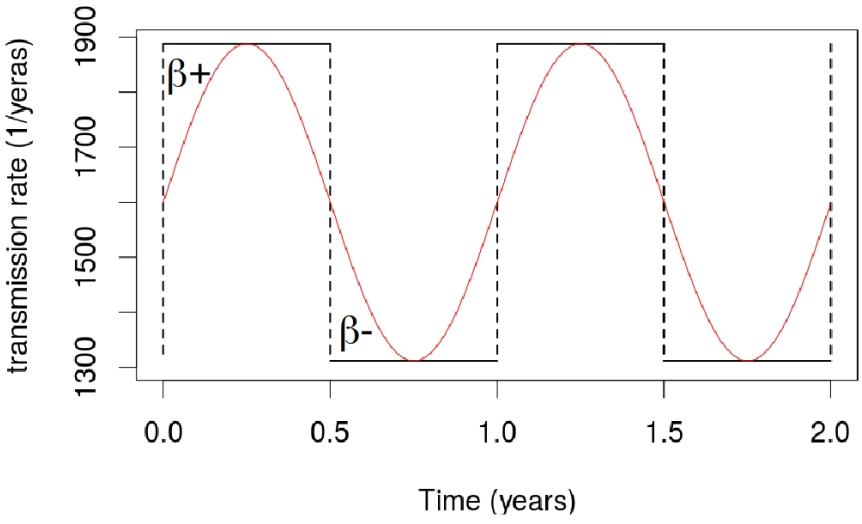

Figure 3.22: Transmission rate considered by Stone et al [8] in (black), and in this work (red).

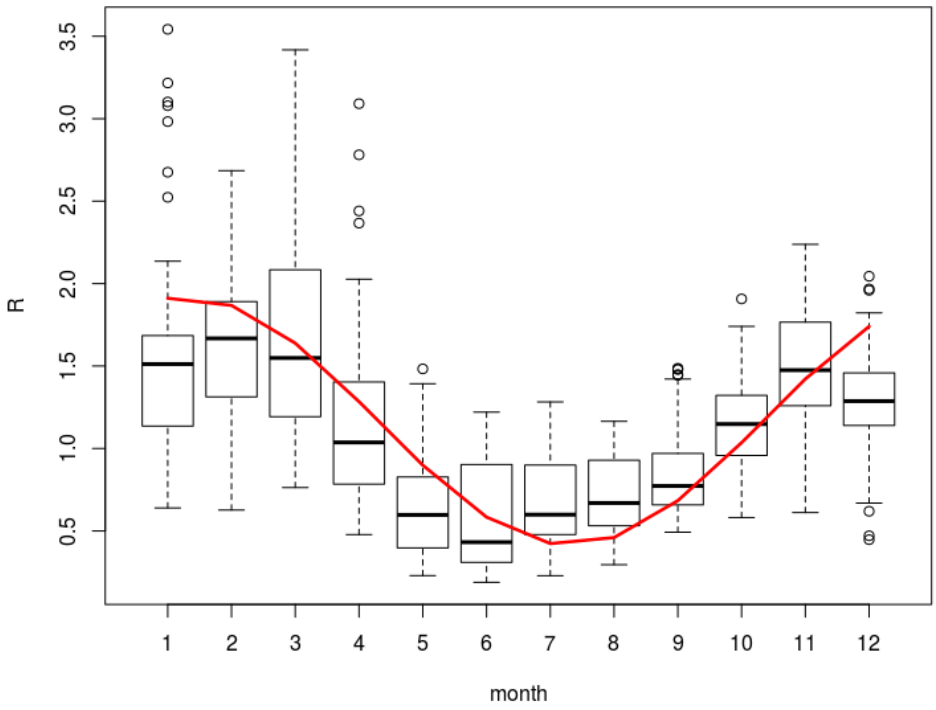

Figure 3.23: **A boxplot of R_0 for each month in Rio de Janeiro from 2010-2016 with monthly estimates from [9] .** The superimposed solid red line indicates the mean monthly R_0 obtained from our stochastic SIR cosine model with the parameter combination with the highest likelihood.

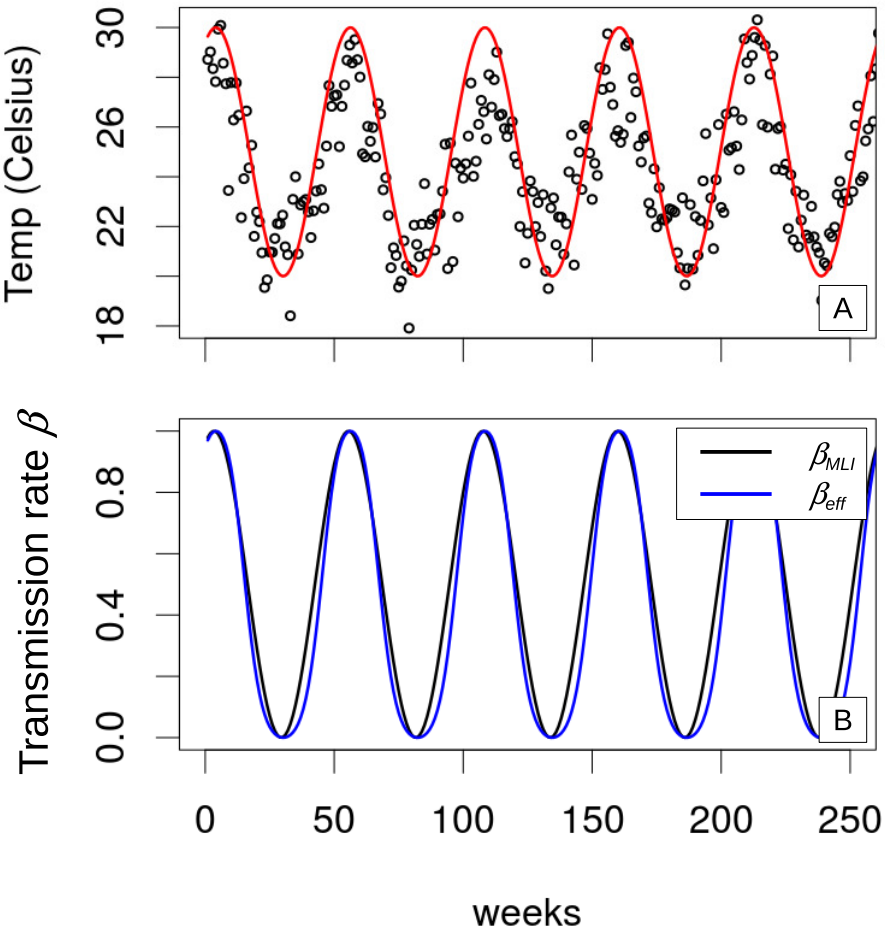

Figure 3.24: **A) Weekly temperature of the city of Rio de Janeiro.** The data is represented by black dots and cosine function by the solid red line. **B) Transmission rate re-scaled between 0 and 1.** The black line is the fitted sinusoidal transmission rate and the blue line the is the effective transmission rate shown on Eq.(3.40).

CHAPTER 4

CONNECTIVITY STRUCTURE AND POPULATION SIZE SHAPE THE SPREAD OF A NEW DENGUE SEROTYPE ACROSS A METROPOLITAN AREA

4.1 Introduction

Arboviruses such as dengue [136], Zika [137], and chikungunya [139] pose a substantial public health burden, especially in large metropolitan areas with heterogeneous climates, population densities, socio-economic conditions, and connectivity and seasonality in transmission. This burden is expected to grow in future decades as the range of suitable habitat for *Aedes aegptyi* mosquitoes increases due to both climate change [149] and urbanization. Understanding the spatiotemporal dynamics of dengue is crucial when preparing for future epidemics, but obtaining such an understanding is challenging due to the multiple factors involved in the spread of the disease. These factors include climate variables such as temperature [24] and rainfall [26, 27, 28, 29, 37], socioeconomic variables such as population density [38, 54] access to water sources [49], and human movement [39, 40, 41].

Human movement in particular is a key driver in the spatial propagation of the disease [39, 40, 41], and previous waves of dengue outbreaks due to new serotype invasions in Brazil have spread outward from large cities to mid-size cities and surrounding areas [50]. Statistical models for the spatiotemporal patterns of dengue are well-suited to tackle large regions such as the whole of Brazil, and have done so at typically coarse spatial scales at a micro-region level, identifying a combination of climate and socioeconomic factors and testing predictability of the dynamics [49, 37, 50]. Nevertheless, incorporating the structure of human movement into these statistical models used to make large-scale dengue forecasts remains an open problem [51]. Movement networks may be too complex to include fluxes from all other municipalities as covariates in a statistical model without over-fitting the

model. One class of statistical models uses conditional auto-regressive models as spatially structured random effect terms to account for neighborhood effects [49, 37]. Municipalities that are geographically neighboring, however, may have limited movement flows between them. Modifying these random effect terms to consider movement fluxes can be difficult, since calculating the determinant of the matrix used to represent this spatial structure can be computationally expensive [182]. To bypass these limitations requires a deeper understanding of the structure of human movement to identify key aspects that are most important in the spread of dengue. We use a mechanistic model in a panel setting, in an ensemble of locations for which disease surveillance has occurred in parallel, to investigate the structure of commuter movement within a large metropolitan area and identify features with epidemiological consequences. We define a "panel" model as a model in which each city's dynamics are independent of the dynamics in other cities that are being fitted, conditional on any shared or city-specific covariates.

Large metropolitan areas represent increasingly important landscapes for the overall burden of arboviruses[55, 50, 41] given the urban niche of their mosquito vectors and the concentration of human hosts. Within these areas, connectivity via commuter movement becomes a potentially meaningful driver of dengue epidemics at intermediate spatial scales, smaller than those used for country-wide analyses of climate yet larger than the neighborhood-level used for household transmission studies. Spatial coupling to the most populated core of a metropolis is expected to play a role, but is it sufficient to explain why some cities had a larger epidemic of DENV4 in 2013 than others? And if not, what other flows should be specified? To address these questions, we focus on 20 municipalities surrounding the megacity of Rio de Janeiro, Brazil. We use a process-based model to test whether incorporating movement fluxes to and from each surrounding municipality to the city of Rio de Janeiro is sufficient to explain observed dengue dynamics in the surrounding cities, especially if we account for altitudinal variation in temperature . Variation in epidemic intensity reflects endemic patterns in the most populated areas vs. explosive but less persistent transmission

in other locations. We show that including only fluxes to and from the city of Rio de Janeiro is sufficient to explain dengue dynamics in many of the outlying municipalities, with the exception of those with substantial fluxes to peri-urban areas.

By focusing on fluxes to and from large municipalities, our modeling approach provides a framework for incorporating the essential features of movement patterns. Our results show that connectivity to a large urban municipality plays an important role in the spread of dengue throughout the metropolitan area. Furthermore, large peri-urban municipalities can serve as additional hubs of commuter movement relevant to epidemiology, providing a stepping stone from a large municipality like Rio de Janeiro to more outlying areas. Both connectivity to a large central city and commuter movements between suburban municipalities may be epidemiologically important for the spread of dengue and should be incorporated into statistical models in the future.

4.2 Results

We fit a panel of mechanistic models to case data from 20 municipalities surrounding the city of Rio de Janeiro to identify how the structure of commuter movement within that region impacts the spread of new dengue serotype, and to determine the extent to which this movement structure can be approximated by only considering movement fluxes to and from its largest city. We selected the 20 cities by starting with a set of cities that spanned a latitudinal gradient from the city of Rio to the mountainous areas in the north of the state, and then trimmed that region to remove municipalities which had large fluxes to or from municipalities located outside of the region. The resulting region is highly self-contained, with at least 80% of all outbound fluxes from any city in the region staying within the region (Figure 4.2 Panel B).

There is substantial variation in the spread of DENV4 during the 2012 and 2013 epidemic seasons in these 20 municipalities. In general, cities with larger population sizes experienced higher peaks during the epidemic season and less pronounced troughs during the off-season,

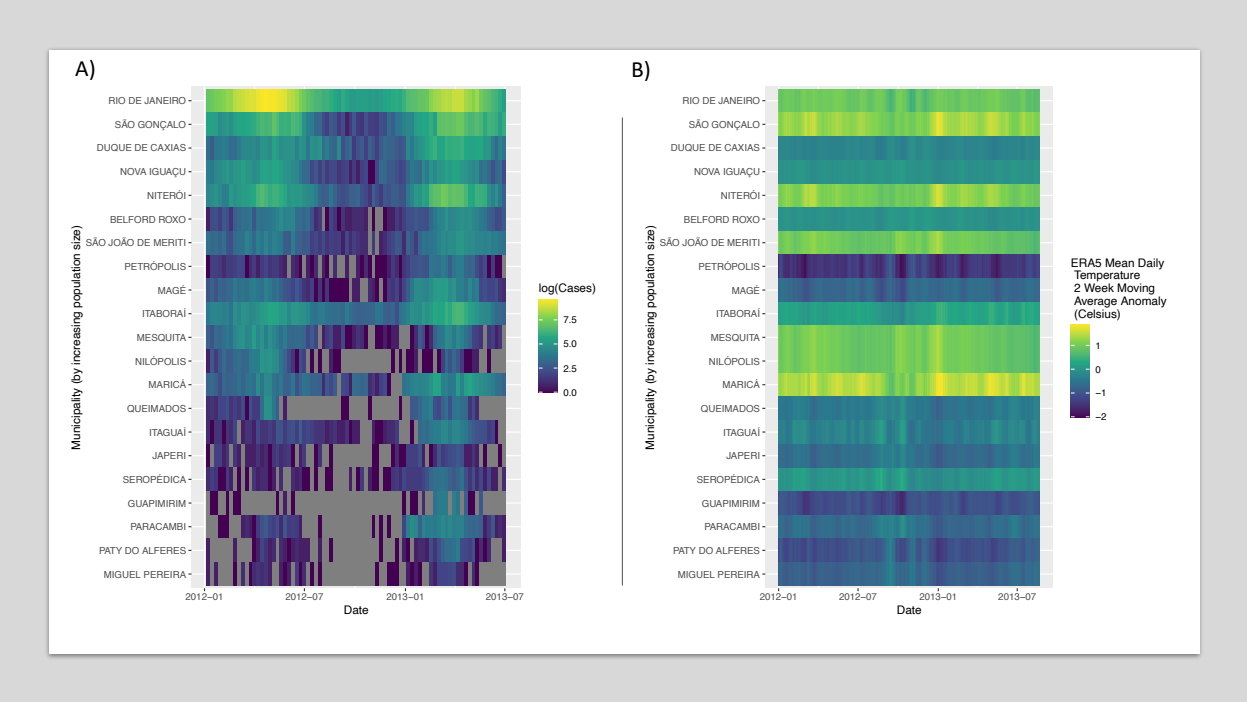

Figure 4.1: **Heatmaps of observed dengue cases and temperature anomalies for 20 municipalities surrounding the city of Rio. A) Heatmap of weekly observed dengue cases on a log scale from January 2012 thru June 2013.** Grey shaded values denote dates with zero observed cases. Municipalities are arranged in increasing order of population. **B) Heatmap of temperature anomaly for each municipality and date. We obtained mean daily temperature estimates for each municipality using ERA5 reanalysis data. Temperature measurements were smoothed using a 2-week moving average. For each date, we calculated the mean temperature across all municipalities for the smoothed temperature time series. The daily anomalies are shown here.**

while many municipalities with smaller population sizes had highly intermittent dengue outbreaks with many weeks of zero reported cases (Fig. 4.1 Panel A). These smaller cities also tended to have more observed cases during the 2013 epidemic season compared to 2012, had lower connectivity to the city of Rio de Janeiro (Fig. 4.2 Panel A), and cooler temperatures compared to the rest of the region (Fig. 4.1 Panel B). There is also considerable variation in epidemic intensity. Some outlying cities in the far north of the region exhibit intense epidemic dynamics with more explosive boom and bust patterns of incidence. Cities to the east of Rio de Janeiro, on the other hand, had higher overall prevalence but more endemic dynamics(Fig. 4.6, Panel B).

We quantified some of these differences by calculating the ratio of the total cases during the epidemic season in 2013 compared to 2012. This ratio increases with the total flux into Rio (Fig. 4.2, Panel A), suggesting that movement to and from Rio may explain why some municipalities had more cases in the first year of the outbreak than others. For cities with similarly large fluxes to Rio, municipalities with higher densities had lower peak ratios. This result is in line with studies showing a strong role for population density in shaping dengue dynamics at very small scales within the city of Rio [54]. Overall, our preliminary data analysis indicates that population size, flux to Rio, temperature, and potentially population density may all play a role in the rate of spread of DENV4 in 2012-2013.

We fit a panel of process-based models to case data from the 20 cities to disentangle the variables at work. We parameterized the models with temperature-dependent transmission rates and coupled only to the city of Rio with movement fluxes to from Rio obtained from census data. Instead of trying to simulate the dynamics within the city of Rio de Janeiro explicitly, we used the observed cases in the city as a covariate that we fed into the model. Our model captures many of the key aspects of the dynamics in most cities, including the seasonality of transmission (Figure 4.3, Supporting Figure 4.8). However, the model appears to more closely match municipalities with trajectories that are more similar to Rio de Janeiro in that they have more cases in 2012 compared to 2013 (Figure 4.3, Supporting Figure 4.8).

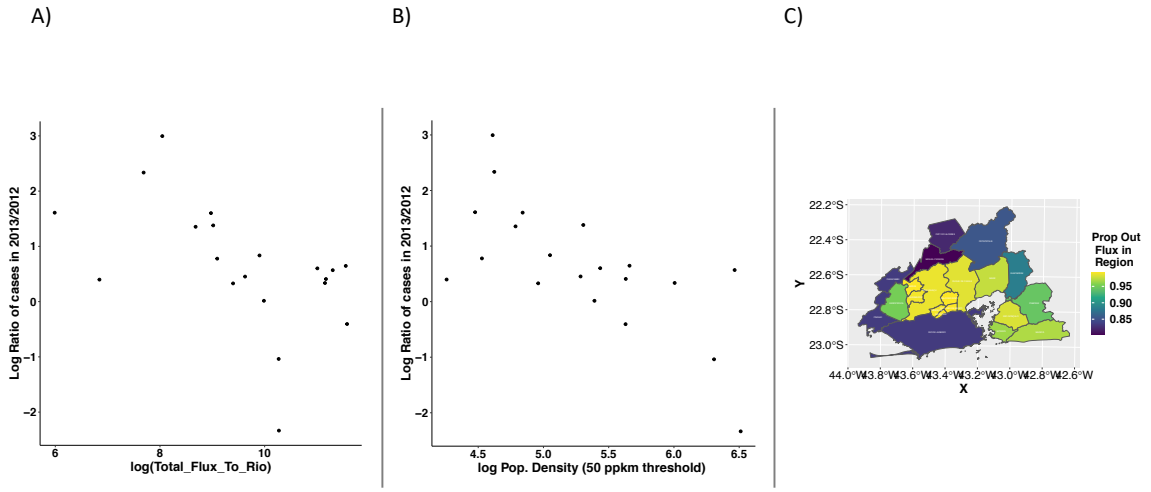

Figure 4.2: **Plot of the log ratio of total cases during the peak epidemic seasons (January-June) in 2013 vs 2012 for each of the 20 municipalities as a function of their A) log total daily flux to the city of Rio de Janeiro and B) log average population density excluding areas with less than 50 people per square kilometer.** We can see here that cities which have larger flues to the city of Rio tend to have smaller peak ratios (more cases in the first year of the epidemic compared to the second year. Points are colored by their For cities with similar total flux to Rio (such as Queimados, Mesquita, and Nilopolis) , higher population density appears to be associated with a lower peak ratio. **B) Map of the 20 cities surrounding Rio used to fit the panel model.** Each municipality is shaded by the proportion of outbound flux from that municipality to all other municipalities in the state of Rio that has a destination within the 20-city region. The proportion of outbound flux is at least 80% in all municipalities, indicating that the selected region is relatively self-contained.

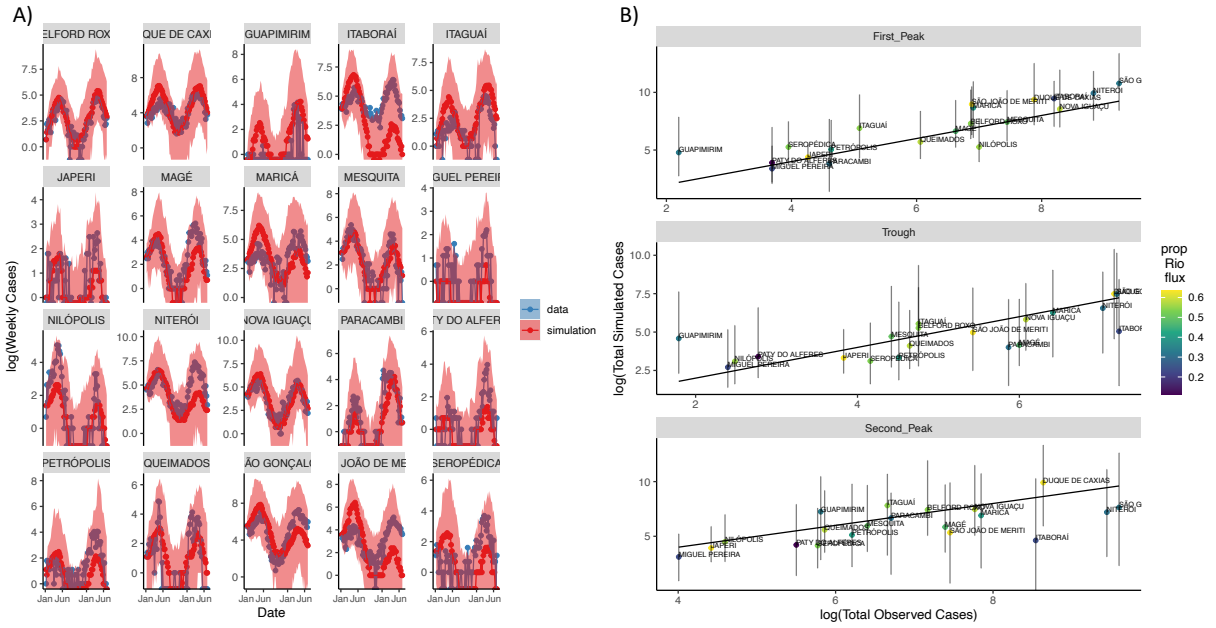

Figure 4.3: **Plots of observed and simulated cases. A) Plot of observed and simulated time series on a log scale.** The blue time series denotes the observed weekly cases for each city. The dark red line represents the median value for 100 simulations from the parameter combination with the highest likelihood from the grid search (the MLE). Red shading denotes the bounds for the 2.5% and 97.5% quantiles from the simulations. **B) Plot of total observed cases vs total simulated cases on a log scale from 100 simulations of the MLE.** We aggregate cases in each simulation trajectory across days 1-200 (first peak), 200-400 (trough) and 400-600 (second peak). The filled circles represent the median value of the total simulated cases across all 100 trajectories within each epidemic stage and municipality, while the error bars denote the 2.5% and 97.5% quantiles. The black line is the 1:1 line. The points are colored according to the proportion of flux in each municipality to or from Rio. This quantity is obtained by adding the total flux between that municipality and Rio (in both directions) and dividing by the total inbound and outbound flux from that municipality to other municipalities within the region (including Rio). For a version of this plot using the parameter combination with the second-highest log-likelihood, which was the only other parameter combination from the grid search within 2 log-likelihood units of the MLE, see Fig. 4.8.

In other words, in several municipalities the model tends to underestimate the second peak; to some degree, this under-estimation of the second peak seems to relate to how well the off-season cases during the intervening trough are captured.

In many cities, infections due to commuter movement to Rio de Janeiro make an important contribution to the force of infection (Fig. 4.4, Table 4.1, Fig. 4.13). This contribution is quite substantial in cities such as Nilopolis which are close to Rio de Janeiro and have a large proportion of their commuter movement fluxes terminating in Rio. Commuter traffic makes a smaller contribution to the force of infection in cities such as Itaborai which have a smaller proportion of traffic to Rio. Overall, the panel model with movement from Rio outperformed a version of the model without any movement fluxes (Fig. 4.4 Panel B), indicating that connectivity to Rio is an important driver of dengue dynamics in the region.

Model	**Log-Likelihood**
Panel with movement	-5117.251**
Panel without movement	-5156.804**

Table 4.1: Comparison of Likelihoods from versions of panel model with and without movement from Rio de Janeiro. The likelihood for the version with movement is obtained from the MLE of the grid search, while the likelihood for the version with no movement from Rio is obtained from the profile of the coupling paramaeter κ, specifically the parameter combination with the highest likelihood at which $\kappa = 0$. The version without movement is fully nested within the original version with movement.

The model captures some of the broad qualitative aspects of the case ratio pattern observed in the preliminary data analysis, namely that municipalities with increasing flux to Rio tend to have lower peak ratios, with more cases during the first year of the invasion (Fig. 4.5). However, the panel is unable to capture the underlying dynamics of susceptible depletion with sufficient precision because the magnitude of variation in the peak ratio within different trajectories for the same parameter combination is much larger than the magnitude of the differences between municipalities.

The model has difficulty capturing dengue dynamics in municipalities with a large proportion of movement traffic that does not originate or terminate in the city of Rio de Janeiro

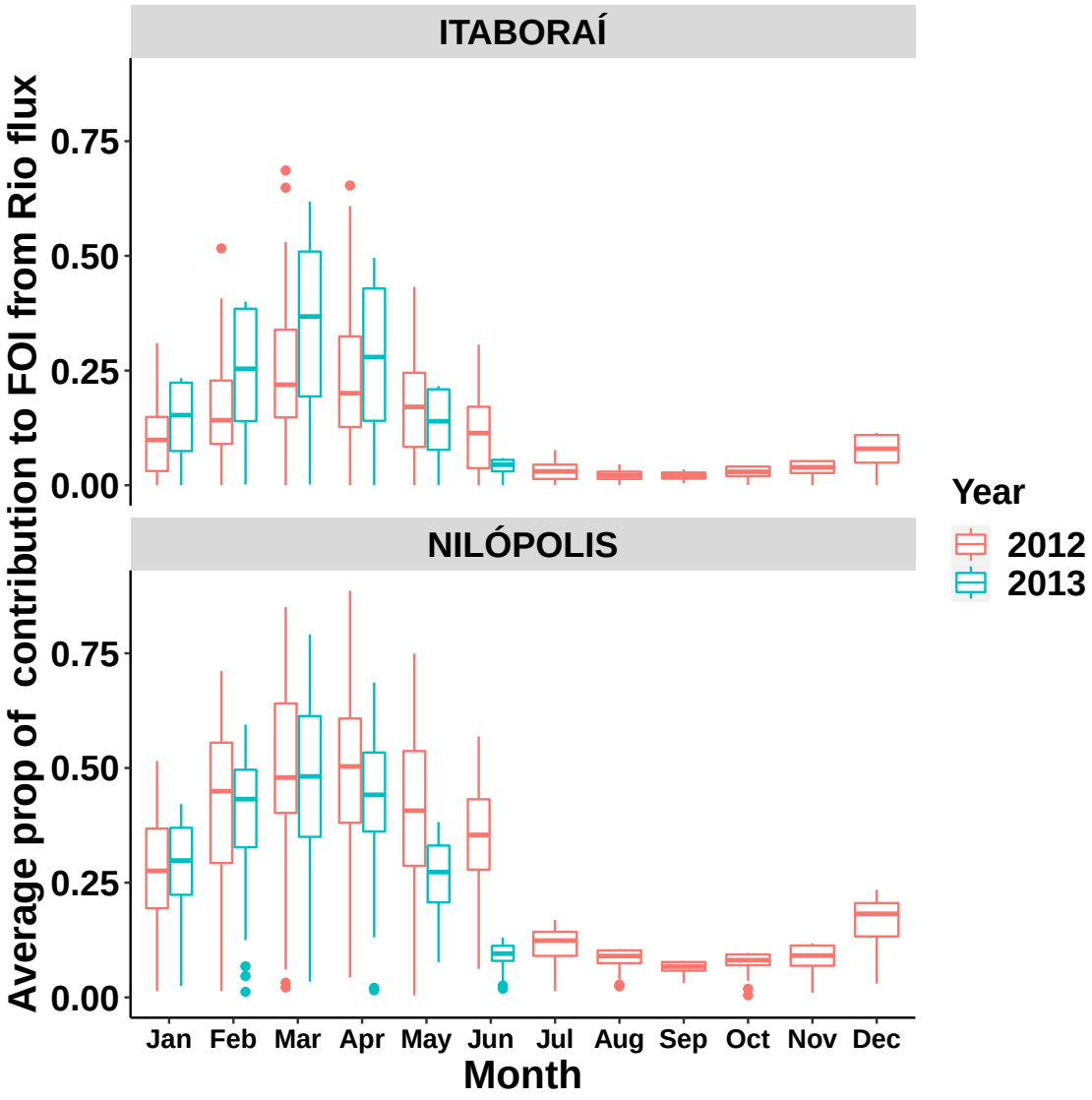

Figure 4.4: **Coupling to Rio de Janeiro has a substantial impact on dengue dynamics. Contribution to the force of infection in Itaborai and Nilopolis from commuters who work in Rio de Janeiro during the day.** The force of infection was calculated for 100 trajectories from the MLE for both municipalities, and contributions from Rio were averaged across all trajectories for each month and municipality. Movement from Rio makes a substantial contribution to the force of infection in Nilopolis, which is located just north of Rio de Janeiro and has a large proportion of flux to Rio. Movement from Rio makes a smaller contribution to the force of infection in Itaborai, which is located at the eastern edge of the region, and has a low proportion of flux to Rio. Itaborai does have substantial commuter traffic to some of the suburbs of Rio such as Niteroi which is not captured in the panel model.

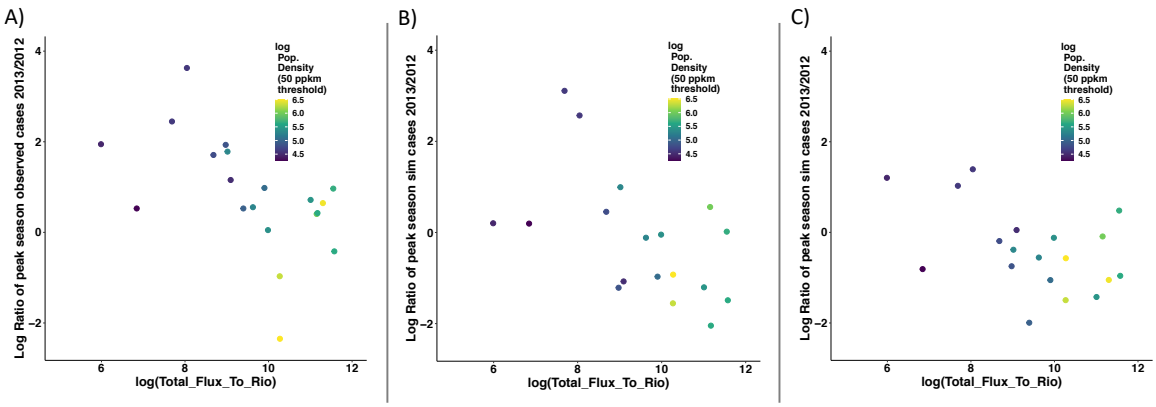

Figure 4.5: **Peak ratio plots as a function of total flux from each municipality to Rio de Janeiro colored by population density. Peak ratios were calculated using A) observed cases B) 100 simulations from the MLE and C) 100 simulations from the only other parameter combination with 2 log-likelihood units of the MLE.** For simulated peak ratios, filled points represent the median peak ratio across all trajectories. The median values of the simulations can somewhat capture the general trend of cities with higher flux having lower peak ratios (i.e. more cases in the first year). However, the simulations are generally unable to capture the differences in peak ratios between cities with the same commuter traffic to Rio but different population densities.

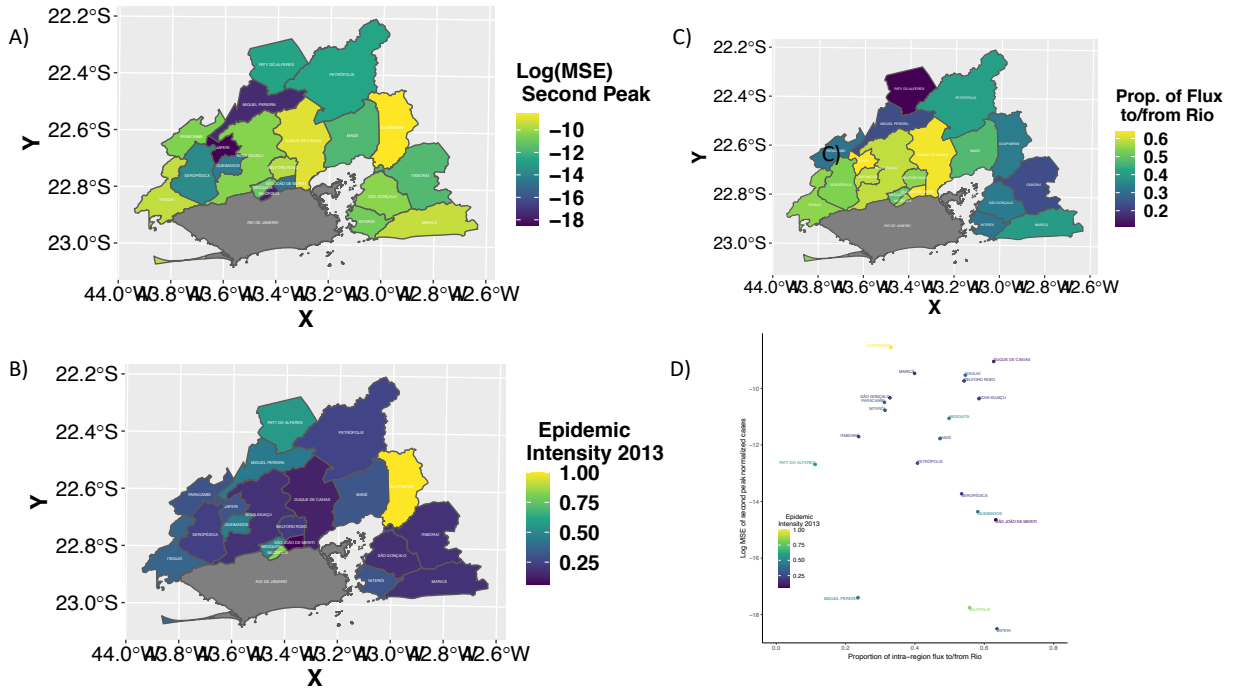

Figure 4.6: **Error Analysis for panel model. A) Map of log mean-squared error for normalized cases during second peak.** For each of 100 simulation trajectories from the MLE, we divided the simulated cases by the population of each municipality and calculated the squared error for this normalized value with respect to the observed cases divided by the population at the same time and city. We then calculated the average mean squared error across all 100 trajectories and time points within the interval of the second peak (days 400 thru 600 of the simulation). Note that many of the municipalities east of the city of Rio have higher MSE during the second peak. **B) Map of epidemic intensity in 2013.** We calculate the Shannon entropy of epidemics in each municipality in 2013 using observed case data, following the approach of Dalziel et al [10], excluding observations with zero cases. Cities with higher Shannon entropy values in 2013 had more intense epidemics, with a larger proportion of the total observed cases in 2013 concentrated within a short time interval. **C) Proportion of flux to and from Rio.** See Figure 3 for an explanation of how this quantity is calculated. Note that the municipalities east of Rio have a substantial proportion of intra-peri-urban fluxes that do not originate or end in Rio. **D) Plot of log MSE in second peak as a function of the proportion of flux to Rio**. Municipalities are colored by their epidemic intensity in 2013. Note that the municipalities east of Rio tend to have high MSE and a low proportion of flux to Rio, while municipalities with a high proportion of Rio flux can have a low or high MSE.

(Fig. 4.6 Panel D, Fig. 4.9, Fig. 4.11). Many of these municipalities are clustered around Niteroi to the east of Rio (Fig. 4.6 Panel A, Fig. 4.9). While there are some parameter combinations which have additional error in municipalities which do have a large proportion of flux to Rio (Fig. 4.6 Panel D, Fig. 4.9), this overall trend becomes quite apparent if one examines best performing parameter combination with low environmental process noise (Fig. 4.11).

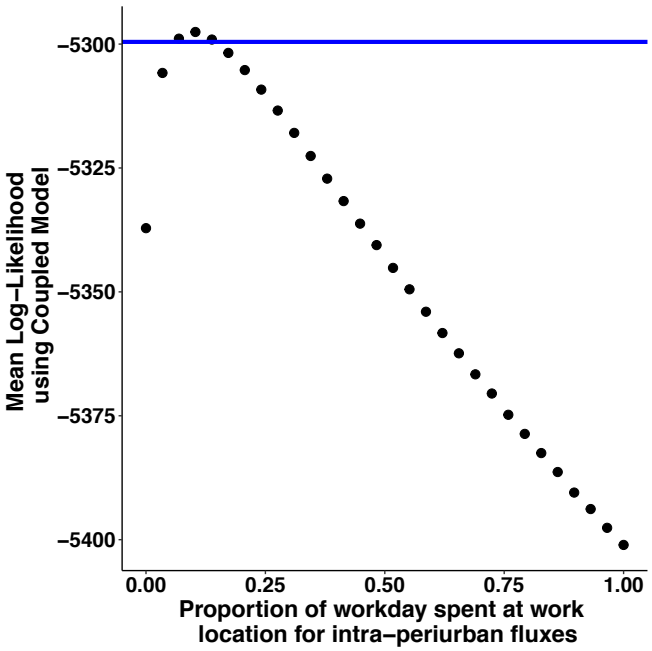

Figure 4.7: **Slice of intra-peri-urban flux parameter**. Each black dot represents the log of the mean likelihood from 10 repetitions of the block particle filter for the fully coupled model parameterized using the same values as the MLE from the gird search of the panel model with the addition of the parameter κ_{suburb}, which represents the proportion of the workday spent at the work location for movement fluxes that do not start or end in the city of Rio. The size of this parameter is a proxy for the importance of non-Rio fluxes. The blue line is equal to two log-likelihood units below the peak of the slice. The non-zero location of the slice peak may indicate that non-Rio fluxes may play a role in the dynamics. However, since the panel and coupled models are not fully nested and this slice analysis is not a true profile, we cannot make this assumption. For a version of the slice that uses the parameter combination with the second highest log-likelihood from the panel model grid search instead of the MLE, see Fig. 4.10.

To address the role of movement fluxes between cities other than Rio in the region, we took an approach which allows us to take advantage of recent methodological developments to estimate likelihood in spatiotemporal systems, without formally inferring parameters for the fully coupled spatiotemporal model (Methods). We specifically verify whether adding intra-suburban movement can improve the performance of the model by evaluating the likelihoods of two "slices" in parameter space using a fully coupled model. This fully coupled model contains an additional coupling parameter, κ_{suburb} which represents the proportion of the day that a host who is commuting to/from cities other than Rio de Janeiro spends in their work location. In practice, this parameter serves as a proxy for the strength of coupling fluxes that do not involve the city of Rio. When κ_{suburb} is zero, the coupled model becomes similar to (but not fully equivalent to) the panel model that only considers fluxes to and from the city of Rio de Janeiro. For each slice, we take the top two best performing parameter combinations from the panel model and evaluate their likelihood assuming varying values of κ_{suburb}. For both slices, parameter combinations in which κ_{suburb} had a non-zero value had a higher likelihood. This result suggests that intra-peri-urban movement fluxes may play an important role in the dynamics of the DENV4 invasion, but more formal analysis is required, including a full profile of the κ_{suburb} parameter as well as ideally fitting the full coupled model and developing a new version of the panel model that can be fully nested within it.

Overall, these results demonstrate that commuter movement plays an important role in the spread of dengue from a large city to surrounding areas within a metropolitan region, and that a panel of mechanistic models can potentially be used to capture crucial aspects of movement structure that impact this spread. However, these approaches may require further augmentation when applied to municipalities that have a large proportion of flux that is unaccounted for in the panel model.

4.3 Discussion

4.3.1 Summary of Results

We investigated how human movement within the metropolitan area of Rio de Janeiro can facilitate the spread of a new dengue serotype in an environment with substantial heterogeneity in temperature and population size. We identified particular aspects of the movement structure that are epidemiologically relevant. With a panel of mechanistic models we show that connectivity to the city of Rio de Janeiro via commuter movement plays an important role in the spread of DENV4. However, we also find that fluxes to and from the city of Rio de Janeiro are insufficient to explain the dynamics of the second year of the dengue invasion in a subgroup of municipalities, especially in the suburbs east of the city. We provide preliminary evidence that including these peri-urban fluxes may improve model performance. Overall, our results indicate that both primary movement hubs such as the city of Rio de Janeiro as well as secondary movement hubs adjacent to this central city, such as the suburb of Niteroi, play a role in the spread DENV4. Fluxes from both types of hubs should be included as epidemiologically essential features of the movement network in statistical models of dengue.

An illustrative example of the role of secondary hubs is the relationship between Rio de Janeiro, Niteroi, and Itaborai. Niteroi is a large city directly east of Rio de Janeiro. Although Niteroi is theoretically a suburb of Rio, it is itself a large city with a population of several hundred thousand people. Niteroi has substantial connectivity to Rio as well as to cities further east, such as Itaborai. Itaborai does not have itself much connectivity to Rio de Janeiro but is quite connected to Niteroi. Many of the larger municipalities to the east of Rio de Janeiro have more endemic dynamics with less deep troughs in the off-season (Fig. 4.8) which the panel model has difficulty capturing (Fig. 4.9). The panel model is unable to capture the influx of commuters from surrounding municipalities who work in Niteroi, since the panel model assumes that Niteroi is only connected to Rio de Janeiro. This influx of commuters may help explain the endemicity of cities like Niteroi. Furthermore, commuters

who work in Niteroi but live in Itaborai may play an important role in spreading dengue from Niteroi to Itaborai. The existence of these hubs may have important consequences for public health intervention strategies. If large suburbs like Niteroi serve as intermediate transition points between large cities like Rio de Janeiro and more outlying cities such as Itaborai, targeted intervention in large suburbs may help prevent the spread of dengue to outlying areas. Furthermore, if approximations in movement structure are developed for use in statistical models, our results would indicate that these approximations should also consider fluxes from secondary movement hubs from large suburbs in addition to those from primary ones form the largest cities.

4.3.2 Relationship to Previous Work

Our result that connectivity to the city of Rio de Janeiro drives the spread of a new dengue serotype between different cities in a metropolitan area builds on previous studies emphasizing the importance of human movement in the spread of dengue at the household level[39], national level[40],and between districts of a large city [41]. We use a transmission model formulation that incorporates temperature-dependent mosquito development rates [25] in line with recent mosquito models that incorporate temperature-dependence[25, 149, 164]. By assuming that the infected mosquito population does not substantially change over the time interval in which hosts are infected (the quasi-static approximation, see Chapter 3 Supplemental Results:Vector Model considerations and [181]), we obtain an expression for a temperature-dependent transmission rate based on mosquito demographic functions without an explicit representation of the susceptible and infected mosquito populations.

Our results that cities with a higher population density and connectivity to Rio de Janeiro have more cases in the first wave of the epidemic compared to the second one are in line with previous studies showing the impact of population density at fine spatial scales. Ratios of peak cases in differing seasons have been used to investigate the spread of DENV4 at much finer spatial resolutions of about a block ($250m^2$) within the city of Rio de Janeiro

[54]. At small scales, this quantity can be influenced by the rate at which new cases arrive as well as the rate at which susceptibles are depleted. Both of these aspects may depend on population density, with the peak ratio decreasing as population density increases and then decreasing again at very high population densities. In our analysis, we only see a part of this pattern corresponding to the decrease of the peak ratio with population density (and flux to Rio). If the underlying dynamics of susceptible depletion and arrival at fine scales are indeed determined by the population density of a particular location, the average population density of a municipality may be much less dense than the population density of a 250m-by-250m grid cell within the city of Rio, even if individual patches within that municipality are just as dense. We are thus unlikely to reach regimes with extremely high population density in which the case ratio increases as population density increases. Moreover, at the regional level, there is considerable heterogeneity in movement, which may play just as important a role in influencing the arrival time as the population size or density of a municipality.

Heterogeneity in connectivity may also explain observed differences in epidemic intensity, although heterogeneity in population density may also play an important role. The panel model is able to capture the explosive epidemic dynamics we observe in the small municipalities to the far northwest of the city of Rio de Janeiro, but has difficulty capturing the more endemic dynamics in the large suburbs to the east of the city such as Niteroi. These suburbs have large intra-suburban movements that are not accounted for in the panel model. The model also has difficulty capturing the dynamics in two large suburbs directly to the north of Rio de Janeiro, Nilopolis and Duque du Caxias, despite both being highly connected to Rio, Niteroi has a very high population density, which the model does not take into account. At very high population densities, the number of humans that a mosquito encounters may be very large, resulting in a high reproductive number and potentially a more explosive epidemic. The model thus under-estimates the explosive epidemic in the city during the first year of the invasion. While the average population density in Duque de Caxias is lower than that of Nilopolis, there is considerable heterogeneity in population density within the munic-

ipality that is not captured at a city-level scale. Duque de Caxias has a history of mining and polluted water sources that serve as breeding sites for mosquitoes. Dengue transmission in the highly dense areas of Duque de Caxias may explain the more intense dengue epidemics.

On a larger national scale, larger, more dense cities in the United States have been shown to have more diffuse and less intense influenza epidemics compared to smaller cities [10]. However, our results indicate that at the scale of a metropolitan region, the intensity of dengue epidemics in southeastern Brazil may have a slightly different relationship with population size, density, and connectivity. In general, larger cities and suburbs do have less explosive epidemics than small towns. However, municipalities that either have a high overall population density or contain sub-divisions with a high population density may experience more intense outbreaks compared to cities with similar population sizes but lower densities.

4.3.3 Model Caveats

There are several caveats in our analysis that need to be addressed in future versions of the model . We note that there is a large amount of stochasticity in the maximum likelihood estimate. This is not surprising, considering that we are fitting dengue time series from multiple small cities using only one shared environmental process noise parameter as well as demographic stochasticity. Dengue epidemics in these small cities exhibit highly stochastic behavior, and this is one reason why many statistical models of dengue in southeastern Brazil aggregate cases at the micro-region level rather than at the municipality level. For our purposes, this would not be effective, since the micro-region scale would be too coarse to analyze the movement structure within metropolitan area. The high amount of environmental stochasticity is also not surprising. When fitting a stochastic model with an environmental process noise term, large estimates for this term can be an indication of model mis-specification, as the inference algorithm attempts to account for discrepancies between model trajectories and observed data by increasing the magnitude of environmental noise. Some of the fluctuations in the observed data may be due to changes in other climate vari-

ables such as rainfall and humidity, which cannot be captured by our model explicitly. We added an additional term, ϵ, to represent the effect of a small amount of additional constant immigration of infected individuals which directly augments the force of infection in each city. This term helps buffer the system from some of these fluctuations. Another key way in which we know that the panel model is already mis-specified is it only contains fluxes to and from each city to Rio de Janeiro and ignores intra-peri-urban fluxes. If we look at the top parameter combination from the grid search with a small amount of environmental process noise (see Fig. 4.12), we can see that the model has a much higher log mean squared error during the second peak in cities with a lower proportion of flux to Rio. This pattern is a much sharper version of Fig. 4.4 Panel D in the results. One way in which the inference algorithm can explain this discrepancy in the second between observed and simulated data is by invoking a large amount of environmental noise. A second caveat concerns the high reporting rates in outlying municipalities observed in the MLE for the panel model. Some of these reporting rates are higher than the 33% reporting rate we assumed for Rio de Janeiro based on rates of viremia in blood samples [183], which seems unlikely. As a result, there is little depletion of susceptible hosts in many outlying municipalities under this scenario. One potential explanation may be that the reporting rate in Rio de Janeiro is higher than 33%. The confidence interval from [183] was very large, ranging from 25%-72%. If the true reporting rate in Rio is higher than 33%, then the model may be over-estimating the number of infected cases arriving in the outlying cities from Rio de Janeiro. The reporting rate was fixed at 33% since the reporting rate and the carrying capacity for the city when taken together are non-identifiable. One long term solution is to increase the length of the time series used for fitting to include the rest of 2013 and 2014, when few dengue cases were observed in the region. This will provide the model with the information that an outbreak did not occur during the third year, which may help in identifying reporting rates that are low enough to result in susceptible depletion. In our current strategy of fitting only cases in 2012 and 2013, we avoid making any assumptions about whether susceptible depletion

brought an end to the outbreak in 2014. An alternative explanation to susceptible depletion for the drop in dengue cases in 2014 could be environmental forcing due to anomalous climate conditions, such as the drought which occurred in southeastern Brazil between 2014 and 2017[37]. If a drought had indeed brought an end to the DENV4 invasion in 2014, we would expect to observe however a re-emergence of DENV4 as soon as climate conditions improved in subsequent years. While there have been outbreaks of other serotypes in recent years in Rio de Janeiro [34], there have not been any large outbreaks of DENV4 since 2014 [34, 184, 35]. An additional complication when fitting case data from 2014 is that DENV4 was not the dominant serotype in 2014[34, 184], so one cannot assume that observed cases in the state of Rio are in fact DENV4 cases. One potential solution to this challenge is to modify the measurement model for observations that occur during 2014 to place an upper bound equal to the observed number of cases.

By construction, the panel and coupled models used in this analysis are not completely nested, although they become quite similar when the parameter governing inter-peri-urban fluxes (κ_{suburb}) is set to 0. The fundamental difference between the two models is how they treat people who live in the city of Rio but work in a different city. The panel model assumes that all people who live in the city of Rio de Janeiro either work in Rio de Janeiro during the day or in city u, where u is the panel city being fitted. The coupled model, on the other hand, assumes that all individuals who live in the city of Rio stay in the city during the workday . This lack of nestedness may explain why the likelihood of the coupled model in the κ_{suburb} slice was about 200 log-likelihood units lower than the panel MLE value. One solution to this problem is to reformat the panel model so that it only includes each outlying city's commuter traffic to Rio, rather than traffic in the other direction. This new panel version can then be integrated into the fully coupled model.

4.3.4 Implications

Overall, our results provide an illustrative example of how a panel of mechanistic models can be used to quantify the impact of connectivity on the spread of a new dengue serotype in a metropolitan area. This approach obviates some of the technical challenges associated with fitting a fully spatiotemporal model to case data. The high dimensional state space of such a model can make inference challenging.

The fully coupled model we formulated could be used to further quantify the contribution of peri-urban movement hubs such as Niteroi. Particle filtering algorithms that are frequently used to estimate the likelihood of stochastic epidemiological models are often unable to estimate the likelihood of spatiotemporal models with a large number of units, since the number of particles required to explore the high-dimensional state space of the model scales exponentially with the size of the model [48]. We provided a demonstration of a recently developed technique, the block particle filter [57] implemented in the R package spatPomp [58], that avoids this problem of particle depletion and can be easily applied to evaluate the likelihood of a coupled dengue model with a specified parameterization. Iterated versions of this algorithm that are currently under development as well as several other recently developed techniques could in theory be used to fit the coupled model to case data. Many of these algorithms have been implemented in in the R package spatPomp [58], including the iterated guided intermediate re-sampling filter [185] and the iterated un-adapted bag filter [186] as well existing techniques such as the iterated ensemble Kalman Filer [187]. The block particle filter in particular is fast, computationally affordable, and reasonably accurate, and can be used to infer noise parameters. For these reasons, the iterated block particle filter once it is developed may be ideally suited for inference with fully spatiotemporal dengue models.

Epidemiologically relevant aspects of the commuter movement structure within a metropolis can be identified using a panel of mechanistic models and then incorporated into statistical models. For example, on a national scale, statistical models can be augmented

with information from the most connected provinces [51] . At the scale of a metropolitan region, one could first use a network analysis of commuter movement data to identify both key central nodes such as Rio de Janeiro as well peri-urban movement hubs such as Niteroi. Once those hubs have been identified, information from those hubs could be incorporated as co-variates in a statistical model. This approach could also be integrated with distributed lag-nonlinear models [37] that can account for lagged effects of climate variables such as rainfall. An integrated framework that can capture inter-annual variation in climate, socio-economic factors, and the key aspects of human commuter movement would be extremely useful when forecasting the spread of new dengue serotypes in large metropolitan areas.

4.4 Methods

4.4.1 Data

Weekly dengue case data from January 1, 2012 thru June December 28th, 2014 were obtained from the Info Dengue surveillance system [188]. The model simulations are initiated on January 1, 2012 approximately corresponding to the time point at which DENV4 became dominant in the city of Rio [30]. The first observation used for fitting is on January 8, 2012, while the last observation is on June 30, 2013. Daily mean temperature ERA5 re-analysis data was obtained from the Google Earth Engine database [189]. The temperature data were averaged over all locations within each municipality and were smoothed using a two-week moving average. Fine-scale population density information for each municipality was obtained from the Gridded Population of the World Dataset, accessed via Google Earth Engine [190]. For the population density measurements in the preliminary analysis, we obtained the number of individuals per square kilometer for gridded spatial cells at a 30 arc-second resolution (approximately 1km) . This resolution for the dataset is designed to be compatible with other remote sensing datasets. We then obtained the mean population density in each municipality by averaging the population density units that fell within the

boundaries of the municipality, excluding all spatial units that had fewer than 50 people per square kilometer. Although these sparsely populated areas may play a role in inhibiting dengue spread at a very fine scale, we assume that human movement at a larger scale (between municipalities) plays a larger role in driving the outbreak, and that dengue cases in these areas do not substantially contribute to overall cases within the municipality. If we did not make this adjustment, we would under-estimate the population density experienced by most individuals in that municipality.

Commuter movement fluxes were obtained from the 2010 census [2]. The fluxes describe the average raw number of daily commuters from city i to city j where i and j consist of all municipalities within the state of Rio de Janeiro for which movement fluxes are available.

4.4.2 Model Description

We use an SEIR model with Susceptible, Exposed, Infected, and Recovered sub-compartments.

Susceptible individuals in city u move to the exposed compartment at rate $\mu_{SE}(t)$. Individuals in the exposed compartment become infected at rate μ_{EI} and recover at rate γ.

The rate at which individuals move to the exposed compartment ($\mu_{SE}(t)$) in location u is the force of infection in that location $\lambda_u(t)$.

This force of infection in each location λ_u is a function of the transmission rate in unit u (β_u), the Infected population in unit u (I_u), the total population in unit u, the infected population in all other units $\tilde{u}$ within the region ($I_{\tilde{u}}$), as well the daily commuter fluxes from individuals who live in unit u and commute to unit $\tilde{u}$ ($m_{u\tilde{u}}$) and vice versa ($m_{\tilde{u}u}$).

We assume that the seasonality in transmission is driven by changes in the daily mean temperature. Let $T(u,t)$ represent the daily mean temperature at location u at time t. When estimating $T(u,t)$, we use the daily mean temperature within municipality u from the ERA5 daily mean temperature reanalysis data-set. To obtain our estimate for $T(u,t)$, we average over all grid points within municipality u and over the 14 days preceding day t. The transmis-

sion rate in each location β_u is the product of a function $\beta_{\text{fast}}(T)$of the temperature$T(u,t)$ at time t in location u and a term $\frac{d\Gamma}{dt}$ representing multiplicative white noise in the transmission rate with intensity σ_{P}. This white noise term captures environmental noise (variation in the transmission rate due to random environmental fluctuations).

We also take into account demographic stochasticity in the transition between infection classes, implemented assuming Euler-multinomially distributed transitions.

The measurement model assumes that a fraction ρ of all cases are reported, with additional negative-binomially distributed measurement noise with dispersion parameter σ_{M}. The reporting rate for the city of Rio de Janeiro is given by the parameter ρ_{rio}.

On the basis of [191], we assume that mosquitoes in each city have a different surivival probability.

We assume that the differences between the mortality rates in different municipalities are larger than the variation within the same municipality across wet and dry seasons (this is somewhat supported by the results obtained in that paper). We use the survival rate for Favela do Amorim in the dry season as our starting estimate for the lifespan in all cities, with each city fitted separately.

4.4.3 ODE Equations

SEIR Compartment Equations

$$\frac{dS_u}{dt} = \mu_H N_u(t) - \mu_H S_u(t) - \lambda_u(t) S_u(t) \tag{4.1}$$

$$\frac{dE_u}{dt} = \lambda_u(t) S_u(t) - \mu_H E_u(t) - \mu_{EI} E_u(t) \tag{4.2}$$

$$\frac{dI_u}{dt} = \mu_{EI} E_u(t) - \mu_H I_u(t) - \gamma I_u(t) \tag{4.3}$$

$$\frac{dR_u}{dt} = \gamma I_u(t) - \mu_H R_u(t) \tag{4.4}$$

SEIR Compartment Equations

Briere functions:

$$\alpha(temp_u(t)) = c_\alpha temp_u(t)(temp_u(t) - temp_{\min_\alpha})\sqrt{temp_{\max_\alpha} - temp_u(t)} \tag{4.5}$$

$$\text{EFD}(temp_u(t)) = c_{\text{EFD}} temp_u(t)(temp_u(t) - temp_{\min_{\text{EFD}}})\sqrt{temp_{\max_{\text{EFD}}} - temp_u(t)} \tag{4.6}$$

$$\text{MDR}(temp_u(t)) = c_{\text{MDR}} temp_u(t)(temp_u(t) - temp_{\min_{\text{MDR}}})\sqrt{temp_{\max_{\text{MDR}}} - temp_u(t)} \tag{4.7}$$

$$\text{b}(temp_u(t)) = c_{\text{b}} temp_u(t)(temp_u(t) - temp_{\min_{\text{b}}})\sqrt{temp_{\max_{\text{b}}} - temp_u(t)} \tag{4.8}$$

$$\text{pMI}(temp_u(t)) = c_{\text{pMI}} temp_u(t)(temp_u(t) - temp_{\min_{\text{pMI}}})\sqrt{temp_{\max_{\text{pMI}}} - temp_u(t)} \tag{4.9}$$

Quadratic functions:

$$\text{PDR}(temp_u(t)) = c_{\text{PDR}} temp_u(t)(temp_u(t) - temp_{\min_{\text{PDR}}})\sqrt{temp_{\max_{\text{PDR}}} - temp_u(t)} \quad (4.10)$$

$$\text{pEA}(temp_u(t)) = c_{\text{pEA}}(temp_u(t) - temp_{\max_{\text{pEA}}})(temp_u(t) - temp_{\min_{\text{pEA}}}) \quad (4.11)$$

$$\text{lf}(temp_u(t)) = c_{\text{lf}}(temp_u(t) - temp_{\max_{\text{lf}}})(temp_u(t) - temp_{\min_{\text{lf}}}) \quad (4.12)$$

Mosquito equations for the city of Rio

Briere functions:

$$\alpha(temp_{\text{Rio}}(t)) = c_\alpha temp_{\text{Rio}}(t)(temp_{\text{Rio}}(t) - temp_{\min_\alpha})\sqrt{temp_{\max_\alpha} - temp_{\text{Rio}}(t)} \quad (4.13)$$

$$\text{EFD}(temp_{\text{Rio}}(t)) = c_{\text{EFD}} temp_{\text{Rio}}(t)(temp_{\text{Rio}}(t) - temp_{\min_{\text{EFD}}})\sqrt{temp_{\max_{\text{EFD}}} - temp_{\text{Rio}}(t)} \quad (4.14)$$

$$\text{MDR}(temp_{\text{Rio}}(t)) = c_{\text{MDR}} temp_{\text{Rio}}(t)(temp_{\text{Rio}}(t) - temp_{\min_{\text{MDR}}})\sqrt{temp_{\max_{\text{MDR}}} - temp_{\text{Rio}}(t)} \quad (4.15)$$

$$\mathrm{b}(temp_{\mathrm{Rio}}(t)) = c_{\mathrm{b}} temp_{\mathrm{Rio}}(t)(temp_{\mathrm{Rio}}(t) - temp_{\min_{\mathrm{b}}})\sqrt{temp_{\max_{\mathrm{b}}} - temp_{\mathrm{Rio}}(t)} \quad (4.16)$$

$$\mathrm{pMI}(temp_{\mathrm{Rio}}(t)) = c_{\mathrm{pMI}} temp_{\mathrm{Rio}}(t)(temp_{\mathrm{Rio}}(t) - temp_{\min_{\mathrm{pMI}}})\sqrt{temp_{\max_{\mathrm{pMI}}} - temp_{\mathrm{Rio}}(t)} \quad (4.17)$$

Quadratic functions:

$$\mathrm{PDR}(temp_{\mathrm{Rio}}(t)) = c_{\mathrm{PDR}} temp_u(t)(temp_{\mathrm{Rio}}(t) - temp_{\min_{\mathrm{PDR}}})\sqrt{temp_{\max_{\mathrm{PDR}}} - temp_{\mathrm{Rio}}(t)} \quad (4.18)$$

$$\mathrm{pEA}(temp_{\mathrm{Rio}}(t)) = c_{\mathrm{pEA}}(temp_{\mathrm{Rio}}(t) - temp_{\max_{\mathrm{pEA}}})(temp_{\mathrm{Rio}}(t) - temp_{\min_{\mathrm{pEA}}}) \quad (4.19)$$

Fast Dynamics transmission rate

$$\beta_{\mathrm{Fast}_u}(t) = \alpha(temp_u(t))^2 b_u pMI(temp_u(t)) lf(temp_u(t)\frac{K_u(t)}{N_u(t)}(1 - \frac{\mu_M}{g(temp_u(t))}) \quad (4.20)$$

where

$$g(temp_u(t)) = EFD(temp_u(t)) pEA(temp_u(t)) MDR(temp_u(t)) lf(temp_u(t)) \quad (4.21)$$

In the previous equation, $K_u(t)$ represents the carrying capacity of the mosquito population in city u at time t. We assume that this scales linearly with the human population of city u at time t, denoted by $N_u(t)$.

We define the mosquito-human population ratio k as follows:

$$k_u = \frac{K_u(t)}{N_u(t)} \tag{4.22}$$

We fit the parameter k_u as a unit specific parameter. Plugging in the expression for k_u into our expression for $\beta_{\text{Fast}_u}(t)$ we obtain:

$$\beta_{\text{Fast}_u}(t) = \alpha(temp_u(t))^2 b_u pMI(temp_u(t)) lf(temp_u(t) k_u (1 - \frac{\mu_M}{g(temp_u(t))}) \tag{4.23}$$

$$\beta_u(t) = \beta_{\text{Fast}_u}(t) \frac{d\Gamma}{dt} \tag{4.24}$$

For the city of Rio, we define the transmission rate $\beta_{\text{Rio}}(t)$, where:

$$\beta_{\text{Rio}}(t) = \alpha(temp_{\text{Rio}}(t))^2 b_{\text{Rio}}) pMI(temp_{\text{Rio}}(t)) lf(temp_{\text{Rio}}(t) k_{\text{Rio}} (1 - \frac{\mu_M}{g(temp_{\text{Rio}}(t))}) \tag{4.25}$$

and:

$$g(temp_{\text{Rio}}(t)) = EFD(temp_{\text{Rio}}(t)) pEA(temp_{\text{Rio}}(t)) MDR(temp_{\text{Rio}}(t)) lf(temp_{\text{Rio}}(t)) \tag{4.26}$$

4.4.4 Panel Model Movement Equations

To approximate the infected cases in the city of Rio at time t, we divide the reported number of cases at time t (which is really the total monthly reported cases for the month in which t falls) by the reporting rate for the city of Rio:

$$I_{textrio}(t) = \frac{\text{riomunicases}(t)}{\rho_{\text{rio}}} \tag{4.27}$$

We assume that the city of Rio de Janeiro is the only external driver in terms of outside cases. Furthermore, we assume that the dynamics of the city of Rio are not affected at all by the dynamics in any of the surrounding cities u (i.e. Rio is strictly a source).

We also assume that the city of Rio is the only commuting destination for individuals who live in city u.

Infections from Rio de Janeiro contribute in two ways to transmission in city u.

The first is when people who live in the city of Rio de Janeiro commute to city u for work. The second is when people who live in city u commute to the city of Rio for work, at which point they experience the force of infection present in the city of Rio.

The probability that someone living in Rio de Janerio commutes to work in city u is given by the movement co-variate m_{ru}. The proportion of the workday that they spend in the city they are working in given by κ, so the total probability that that someone who lives in Rio is currently in city u during the workday is given by κm_{ru}.

We denote the quantity $m_{\text{work}_{ru}}$ to account for this adjusted movement probability that someone who lives in Rio is currently in city u during the workday:

$$m_{\text{work}_{ru}} = \kappa m_{ru} \tag{4.28}$$

We denote the corresponding quantity $m_{\text{work}_{ur}}$ to account for this adjusted movement probability that someone who lives in city u is currently in the city of Rio during the workday:

$$m_{\text{work}_{ur}} = \kappa m_{ur} \tag{4.29}$$

In this model, we assume that any time that a person who lives in city u does not spend in the city of Rio is spent in city u, and that all people who do not commute to the city of Rio spend all of their time in city u. Therefore, we can come up with an expression for $m_{\text{work}_{uu}}$, the probability that someone who lives in city u is currently in city u during the workday:

$$m_{\text{work}_{uu}} = 1 - m_{\text{work}_{ur}} \tag{4.30}$$

$$m_{\text{work}_{uu}} = 1 - \kappa m_{ur} \tag{4.31}$$

Likewise:

$$m_{\text{work}_{rr}} = 1 - m_{\text{work}_{ru}} \tag{4.32}$$

$$m_{\text{work}_{rr}} = 1 - \kappa m_{ru} \tag{4.33}$$

The force of infection experienced by someone who stays in city u during the workday ($\lambda_{\text{resident}}$) will come from both infected people from the city of Rio who are in city u during that time of the workday as well as infected people from city u who stayed in city u during the workday. Both groups will experience the transmission rate of city u at that time:

$$\lambda_{\text{resident}}(t) = \beta_u(t)\frac{m_{\text{work}_{uu}} I_u(t) + m_{\text{work}_{ru}} I_{\text{rio}}(t)}{m_{\text{work}_{uu}} N_u(t) + m_{\text{work}_{ru}} N_{\text{rio}}(t)} \tag{4.34}$$

We add in extra immigration, denoted by ϵ, which represents the average number of people who arrive in any city multiplied by the probability of transmission from those individuals. We assume that this extra immigration term is the same for all cities.

The modified form of the previous equation taking into account extra immigration via ϵ is then:

$$\lambda_{\text{resident}}(t) = \frac{[\beta_u(t)(m_{\text{work}_{uu}} I_u(t) + m_{\text{work}_{ru}} I_{\text{rio}}(t))] + \epsilon}{m_{\text{work}_{uu}} N_u(t) + m_{\text{work}_{ru}} N_{\text{rio}}(t)} \tag{4.35}$$

Fully specify by plugging in movement terms:

$$\lambda_{\text{resident}}(t) = \frac{[\beta_u(t)([1 - \kappa m_{ur}] I_u(t) + \kappa m_{ru} I_{\text{rio}}(t))] + \epsilon}{[1 - \kappa m_{ur}] N_u(t) + \kappa m_{ru} N_{\text{rio}}(t)} \tag{4.36}$$

The force of infection experienced by someone who stays in the city of Rio during the workday ($\lambda_{\text{commuter}}$) will only depend on the current infected population and total population of the city of Rio de Janeiro. This population will experience the transmission rate of the city of Rio at the time t. We assume that the additional influx of commuters from other cities does not affect either the total population present in the city of Rio at the time $N(t)$ or the infections I. Thus, we can say that:

A more complete expression may have been:

$$\lambda_{\text{commuter}}(t) = \beta_{\text{Rio}}(t)\frac{m_{\text{work}_{ur}} I_u(t) + m_{\text{work}_{rr}} I_{\text{rio}}(t)}{m_{\text{work}_{ur}} N_u(t) + [1 - m_{\text{work}_{ru}}] N_{\text{rio}}(t)} \tag{4.37}$$

However, because of these assumptions, we ignore the contribution to the FOI in Rio for people from city u

Therefore:

$$\lambda_{\text{commuter}}(t) = \beta_{\text{Rio}}(t)\frac{m_{\text{work}_{rr}} I_{\text{rio}}(t)}{m_{\text{work}_{rr}} N_{\text{rio}}(t)} \tag{4.38}$$

The modified form of the previous equation taking into account extra immigration via ϵ is then:

$$\lambda_{\text{commuter}}(t) = \frac{(\beta_{\text{Rio}}(t) m_{\text{work}_{rr}} I_{\text{rio}}(t)) + \epsilon}{m_{\text{work}_{rr}} N_{\text{rio}}(t)} \tag{4.39}$$

Fully specify by plugging in movement terms:

$$\lambda_{\text{commuter}}(t) = \frac{(\beta_{\text{Rio}}(t)[1 - \kappa m_{ru}] I_{\text{rio}}(t)) + \epsilon}{[1 - \kappa m_{ru}] N_{\text{rio}}(t)} \tag{4.40}$$

Now to calculate the total force of infection for city u, we take a weighted average of the force of infection experienced by residents of city u who stay in city u and those who commute to the city of Rio, weighted by the probability that they were in city u or Rio at that point of the workday:

$$\lambda_u = m_{\text{work}_{ur}} \lambda_{\text{commuter}}(t) + m_{\text{work}_{uu}} \lambda_{\text{resident}}(t) \tag{4.41}$$

Plugging in movement terms:

$$\lambda_u = \kappa m_{ur} \lambda_{\text{commuter}}(t) + [1 - \kappa m_{ur}] \lambda_{\text{resident}}(t) \tag{4.42}$$

Plugging in resident and commuter components:

$$\begin{aligned} \lambda_u = \kappa m_{ur} [& \frac{(\beta_{\text{Rio}}(t)[1 - \kappa m_{ru}] I_{\text{rio}}(t)) + \epsilon}{[1 - \kappa m_{ru}] N_{\text{rio}}(t)}] \\ & + [1 - \kappa m_{ur}] [\frac{[\beta_u(t)([1 - \kappa m_{ur}] I_u(t) + \kappa m_{ru} I_{\text{rio}}(t))] + \epsilon}{[1 - \kappa m_{ur}] N_u(t) + \kappa m_{ru} N_{\text{rio}}(t)}] \end{aligned} \tag{4.43}$$

Equations for model with demographic stochasticity

Rates in continuous time:

$$\mu_{S_u E_u}(t) = \lambda_u(t) \tag{4.44}$$

$$\mu_{\cdot S} = \mu_H \tag{4.45}$$

$$\mu_{S\cdot} = \mu_{E\cdot} = \mu_{I\cdot} = \mu_{R\cdot} = \mu_H \tag{4.46}$$

$$\mu_{EI} = \mu_{EI} \tag{4.47}$$

$$\mu_{IR} = \gamma \tag{4.48}$$

4.4.5 Discretizations

Eulermultinomial discretization of compartment flows from time t to time $t + \Delta t$

Several compartments have more than one exit at each time step. Instead of a binomial transition probability, we thus use the Euler multinomial distribution to model transition events using the function reulermultinom in the R package pomp as described by He and Ionides [17]).

We use the implementation of eulermultinomial transitions in: https://raw.githubusercontent.com/kingaa/sbied/master/measles/measles.R as a guide.

We summarize briefly that function here. Given a population X, a time interval Δt, and a set of rates $r_1...r_k$, the number of individuals remaining in that class or moving to other classes is multinomially distributed:

$$(X - \sum_{i=1}^{k}(dx_i), dx_1, \ldots, dx_k) \sim \text{Multinomial}(X; p_0, p_1, \ldots, p_k) \tag{4.49}$$

Each probability p_j where $j = 1...k$ is calculated as follows:

$$p_j = (1 - exp(-\sum_{i=1}^{k}(r_i \Delta t))) \frac{r_j}{(\sum_{i=1}^{k}(r_i))} \tag{4.50}$$

For our model:

$$(\tilde{S}_u(t) - \Delta\tilde{N}_{S_u\cdot} - \Delta\tilde{N}_{S_uE_u}, \Delta\tilde{N}_{S_u\cdot}, \Delta\tilde{N}_{S_uE_u}) \sim reulermultinomial(\tilde{S}_u(t), \mu_{S\cdot}, \mu_{S_uE_u}, \Delta t) \tag{4.51}$$

$$(\tilde{E}_u(t) - \Delta\tilde{N}_{E_u\cdot} - \Delta\tilde{N}_{E_uI_u}, \Delta\tilde{N}_{E_u\cdot}, \Delta\tilde{N}_{E_uI_u}) \sim reulermultinomial(\tilde{E}_u(t), \mu_{E\cdot}, \mu_{EI}, \Delta t) \tag{4.52}$$

$$(\tilde{I}_u(t) - \Delta\tilde{N}_{I_u\cdot} - \Delta\tilde{N}_{I_uR_u}, \Delta\tilde{N}_{I_u\cdot}, \Delta\tilde{N}_{I_uR_u}) \sim reulermultinomial(\tilde{I}_u(t), \mu_{I\cdot}, \mu_{IR}, \Delta t) \tag{4.53}$$

$$(\tilde{R}_u(t) - \Delta\tilde{N}_{R_u\cdot}, \Delta\tilde{N}_{R_u\cdot}) \sim reulermultinomial(\tilde{R}_u(t), \mu_{R\cdot}, \Delta t) \tag{4.54}$$

$$(\tilde{N}_u(t) - \Delta\tilde{N}_{\cdot S_u}, \Delta\tilde{N}_{\cdot S}) \sim reulermultinomial(\tilde{N}_u(t), \mu_{\cdot S_u}, \Delta t) \tag{4.55}$$

4.4.6 Compartment Transitions

$$\Delta\tilde{S}_u = \Delta\tilde{N}_{\cdot S_u} - \Delta\tilde{N}_{S_u\cdot} - \Delta\tilde{N}_{S_uE_u} \tag{4.56}$$

$$\Delta\tilde{E}_u = \Delta\tilde{N}_{S_uE_u} - \Delta\tilde{N}_{E_u\cdot} - \Delta\tilde{N}_{E_uI_u} \tag{4.57}$$

$$\Delta\tilde{I}_u = \Delta\tilde{N}_{E_uI_u} - \Delta\tilde{N}_{I_u\cdot} - \Delta\tilde{N}_{I_uR_u} \tag{4.58}$$

$$\Delta\tilde{R}_u = \Delta\tilde{N}_{I_uR_u} - \Delta\tilde{N}_{R_u\cdot} \tag{4.59}$$

$$\Delta\tilde{N}_u = \Delta\tilde{N}_{\cdot S_u} - \Delta\tilde{N}_{S_u\cdot} - \Delta\tilde{N}_{E_u\cdot} - \Delta\tilde{N}_{I_u\cdot} - \Delta\tilde{N}_{R_u\cdot} \tag{4.60}$$

$$\Delta\tilde{C}_u = \Delta\tilde{N}_{E_uI_u} \tag{4.61}$$

4.4.7 Measurement Model Equations

$$Y_u \sim NegativeBinomial(\mu = \rho\tilde{C}_u, size = \sigma_{\mathrm{M}}) \tag{4.62}$$

4.4.8 Model Fitting Strategy

We fit the panel model using the panel iterated filtering algorithm PIF implemented in the R package panelPOMP[173]. We fix the probability of mosquito infectiousness ($b(T)$ and the probability of mosquito infection ($pMI(T)$) at 0.41, which is the mean value for $b(T)$ across all temperatures experienced in the city of Rio de Janeiro during the time period being fitted. For other mosquito parameters, we use the fitted values of [25].

Some of the parameters in the model are shared, while others are city-specific. Shared parameters that we fit include the recovery rate γ, rate of leaving the exposed class μ_{EI}, the proportion of the day that commuters spend at their workplace location κ, the mosquito/human

population ratio in the city of Rio de Janeiro (k_{Rio}), the environmental noise parameter σ_{P} and the measurement noise parameter ψ . We fix the reporting rate in the city of Rio de Janeiro to be 33%, based on [183].

Specific parameters that we fit include the initial number of infected ($I_0(u)$) and exposed ($E_0(u)$) cases in each municipality u (excluding the city of Rio), as well as the reporting rate ρ_u and the ratio of mosquitoes to humans in each city (k_u).

Grid Search

For all fitted parameters except I_0 and E_0, we generate a grid of 1,000 different parameter combinations using Latin Hyper-cube Sampling via the lhs() function of the R package tcp. We then generate initial estimates of I_0 and E_0, based on the values of the recovery rate γ, rate of leaving the exposed class $\mu_E I$, unit-specific reporting rate ρ_u, and the number of cases reported during the first week of 2012 in each city.

First, we calculate the duration of infection and the duration of the internal incubation period by taking the inverse of $\gamma and \mu_E I$. To estimate I_0 initially, we start by taking the total cases reported in the first week and dividing by the reporting rate ρ. This corrects for under-reporting. We then need to correct for the difference in time scales. The quantity we have is the corrected number of new infections over the course of a whole week. However, the time step of our simulation (Δt) is 1 day. We correct for this by dividing the corrected number of new infections by 7, assuming that new infections are evenly distributed across all days of the week. We now have an expression for the expected number of new infected cases each day. One final correction we need to make is that the people in I_0 also includes those who were infected on a previous day and have not yet recovered. To account for this, we multiply the quantity we have calculated by the duration of infection (in days).

To calculate E_0, we use our estimate for I_0, and then multiply by the ratio of time that people spend in the exposed class (the duration of the intrinsic incubation period) to the time that people spend in the infected class (the duration of infection).

Because the parameter space is high-dimensional, we fit the model in four stages. In the first stage, we fix the specific parameter values at their initial estimates and only fit the shared parameters, and perform two successive runs of the panel iterated filtering algorithm pif. During the second stage, we fix the shared parameters at the ending values from the first stage, and instead fit only the unit-specific parameters and again perform two successive PIF runs. During the third stage, we again fix the unit-specific parameters at their end points from the second stage and only fit the shared parameters, although this time we only perform one run of PIF. Finally, in the fourth stage, we fix the shared parameters at their end points from the third stage and perform one PIF run fitting the unit-specific parameters.

All pif runs consisted of 50 iterations, 2,000 particles, and a random walk standard deviation of 0.005 for all fitted parameters. The likelihood was measured and the start and end of each PIF run using the panel sequential Monte Carlo algorithm pfilter in the R package panel pomp [173] with 10,000 particles and 10 replicates.

Movement flux metrics

Two metrics that we calculate for each municipality within the region include the proportion of flux to Rio and the proportion of flux that stays with in the region. For the first metric, we sum the total inbound and outbound flux from a given municipality within the region that originates or terminates in the city of Rio de Janeiro by the total inbound or outbound flux that originates or terminates within the region. For the second metric, we sum the total inbound and outbound fluxes for a given municipality within our region that stay within the region, and divide this quantity by the total total fluxes that originate or terminate from this municipality, including fluxes that originate or terminate outside of the region but within the state of Rio de Janeiro.

Mean Squared Error Calculation

We obtain 2 parameter combinations within 2-log-likelihood units of the maximum-likelihood estimate(MLE) for the grid search, including the MLE parameter combination itself. For both parameter combinations, we calculate the mean-squared normalized error by simulating from 100 trajectories of the parameter combination, dividing the number of simulated and observed cases by the population in each city, and then calculating the mean squared error for this normalized value. We calculate this metric separately for each city and time period (the first peak, trough, and second peak).

Profile

As an additional sensitivity analysis, using the maximum and minimum values of all parameter combinations from the grid search within 20-log-likelihood units of the grid search MLE as the bounds of the sampling interval for other parameters, we generate a profile of the proportion of time during the workday that commuters spend in their work destination κ. For each of 30 evenly spaced values of κ between 0 and 1 inclusive, we generate 40 parameter combinations from the bounds of all grid search parameter combinations within 20-log-likelihood units of the maximum likelihood estimate (MLE).

For each parameter combination, we perform three successive pif runs using the same tuning parameters as the grid search with one difference. Instead of fitting unit-specific and shared parameters in separate stages, we instead utilize a newly developed block feature of the algorithm which enables both shared and specific parameters to be fit concurrently in the same stage.

Slice

For both the MLE and the parameter combination with the second-highest log-likelihood (which was also within 2LL of the MLE), we calculate the likelihood for each parameter combination using the fully coupled, sampling 30 different evenly spaced values of the sub-

urban flux parameter but keeping all other parameters at the best fit value from the panel model. We use the block particle filter algorithm [57] in the R package spatPomp [58] with 200,000 particles and a block size of 1.

Selected movement equations for the fully coupled model are included in the supplement.

4.5 Funding

R.S. was supported by a National Science Foundation Research Traineeship (no. 1735359: NRT-INFEWS: Computational data science to advance research at the energy environment nexus). M.P., K.A., and E.I. were supported by a collaborative grant from the National Science Foundation's Division of Mathematical Sciences and the National Institutes of Health (no. 1761612: Collaborative Research: Urban Vector-Borne Disease Transmission Demands Advances in Spatiotemporal Statistical Inference). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the National Science Foundation.

4.6 Acknowledgments

R.S., M.P., E.I., and K.A. designed the experiments. R.S. and M.P. designed the temperature models, while R.S., K.A., and M.P. designed the movement model. R.S., K.A., E.I., and M.P designed the coding/inference pipepline and fitting strategies. R.S. performed the experiments, with assistance from K.A, E.I., and M.P. C.T.C., F.C., and M.G. provided guidance and assistance obtaining dengue case and movement data and expertise interpreting movement patterns and dengue circulation dynamics in Rio de Janeiro. R.S. and M.P wrote the manuscript. The authors acknowledge Aaron King and Victoria Romeo-Aznar for their discussions and insight, as well as Raquel M. Lana and Lais P. Freitas for providing information regarding serotype prevalence in the state of Rio de Janeiro. This work was

completed with resources and support provided by the University of Chicago's Research Computing Center.

4.7 Supporting information

4.7.1 Supplemental Figures

4.7.2 Fully Coupled Model Equations

4.8 Details for Coupled Model Equations

$$m_{\text{work}_{uv}} \tag{4.63}$$

$$m_{\text{work}_{uv}} = \begin{cases} \kappa m_{uv}, & \text{if } v = U+1 \text{ and } u \neq v \\ \kappa m_{uv}, & \text{if } u = U+1 \text{ and } u \neq v \\ 1 - \kappa m_{u(U+1)} - \sum_{v=1}^{U+1} \kappa_{\text{suburb}} m_{uv}, & \text{if } u = v \text{ and } u \neq U+1 \\ 1 - \sum_{j=1}^{U} \kappa m_{(U+1)j}, & \text{if } u = v \text{ and } u \neq U+1 \\ \kappa_{\text{suburb}} m_{uv}, & \text{otherwise} \end{cases}$$

Let $r = U + 1$, which represents the index for the city of Rio de Janeiro. We can then re-write this as:

$$m_{\text{work}_{uv}} = \begin{cases} \kappa m_{uv}, & \text{if } v = r \text{ and } u \neq v \\ \kappa m_{uv}, & \text{if } u = r \text{ and } u \neq v \\ 1 - \kappa m_{ur} - \sum_{v=1}^{U} \kappa_{\text{suburb}} m_{uv}, & \text{if } u = v \text{ and } u \neq r \\ 1 - \sum_{j=1}^{U} \kappa m_{rj}, & \text{if } u = v \text{ and } u = r \\ \kappa_{\text{suburb}} m_{uv}, & \text{otherwise} \end{cases}$$

Furthermore, define

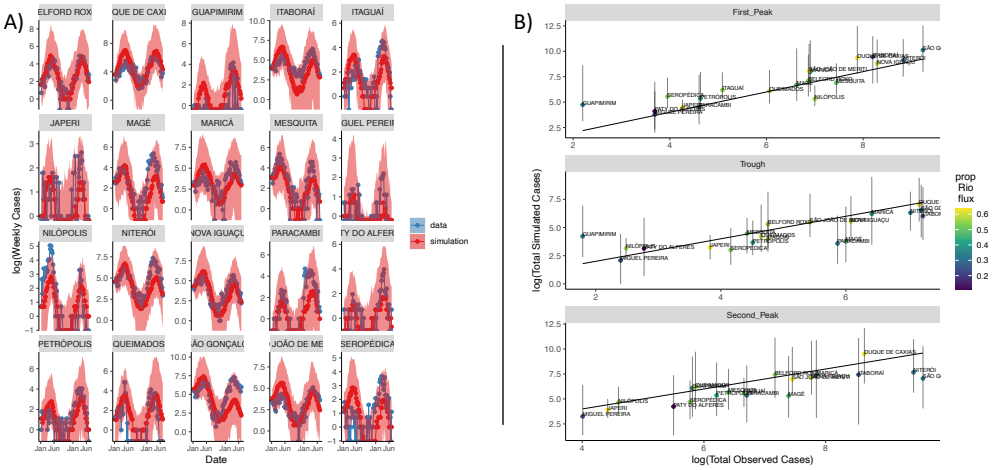

Figure 4.8: **Plots of observed and simulated cases. A) Plot of observed and simulated time series on a log scale.** The blue time series denotes the observed weekly cases for each city. The dark red line represents the median value for 100 simulations from the parameter combination with the second-highest likelihood from the grid search. Red shading denotes the bounds for the 2.5% and 97.5% quantiles from those simulations. **B) Plot of total observed cases vs total simulated cases on a log scale from 100 simulations of the parameter combination with the second highest log-likelihood.** We aggregate cases in each simulation trajectory across days 1-200 (first peak), 200-400 (trough) and 400-600 (second peak). The filled circle represent the median value of the total simulated cases across all 100 trajectories within each epidemic stage and municipality, while the error bars denote the 2.5% and 97.5% quantiles. The black line is a reference line along which total simulated cases are equal to the total observed cases. The points are colored according to the proportion of flux in each municipality to or from Rio. This quantity is obtained by adding the total flux between that municipality and Rio (in both directions) and dividing by the total inbound and outbound flux from that municipality to other municipalities within the region (including Rio).

$$I_{\text{rio}_{\text{act}}} = \frac{\text{riomunicases}}{\rho_{\text{rio}}} \tag{4.64}$$

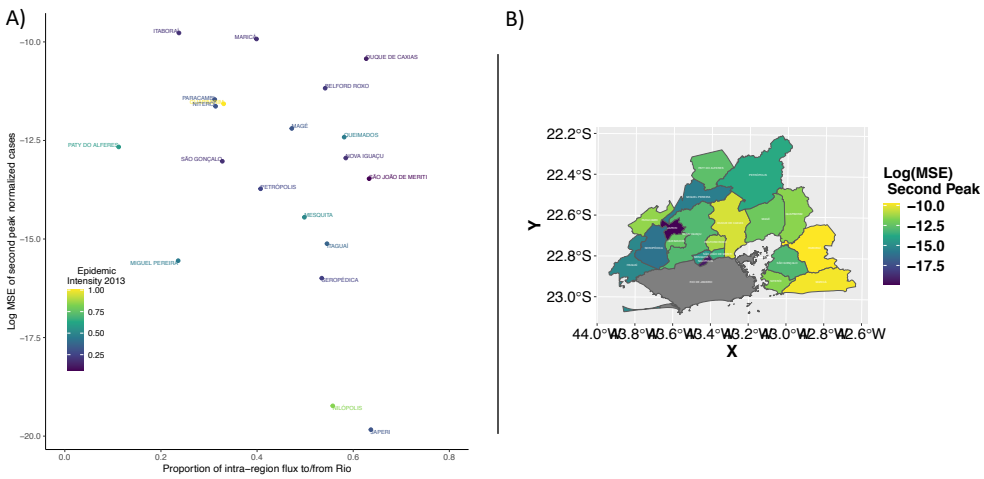

Figure 4.9: **A) Map of log mean-squared error for normalized cases during second peak.** Here we use the parameter combination with the second-highest log-likelihood instead of the MLE. **B) Plot of log MSE in second peak from second-highest log-likelihood parameter combination as a function of the proportion of flux to Rio**.

Increment the total infections in the working population in city v (J_v) by all infections in people who live in city w but work in city v.

Initially, assume that we are not calculating this quantity for Rio (so $v \neq r$):

$$J_v = m_{\text{work}_{rv}} I_{\text{rio}_{\text{act}}} + \sum_{w=1}^{U} m_{\text{work}_{wv}} I_w \tag{4.65}$$

$$J_v = m_{\text{work}_{rv}} I_{\text{rio}_{\text{act}}} + m_{\text{work}_{vv}} I_v + \sum_{w=1, w\neq v}^{U} m_{\text{work}_{wv}} I_w \tag{4.66}$$

Plugging in we have:

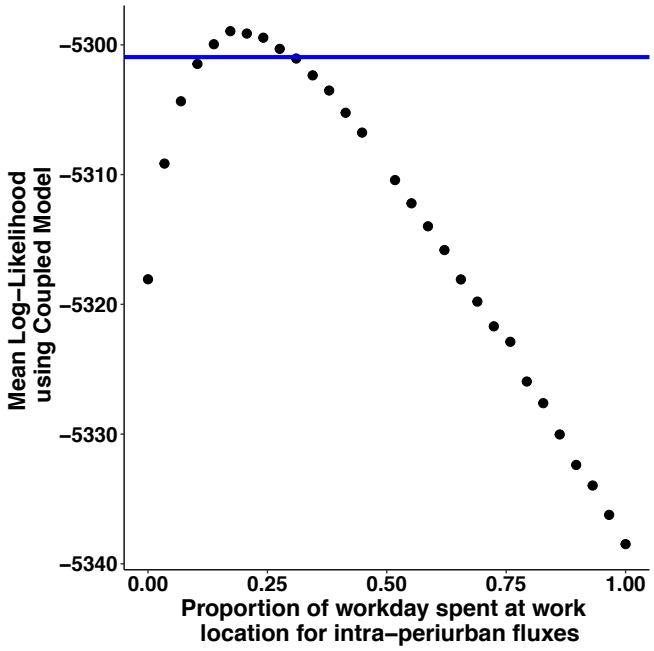

Figure 4.10: **Slice of intra-peri-urban flux parameter**. Each black dot represents the mean likelihood from 10 repetitions of the block particle filter for the fully coupled model parameterized using the same values as the parameter combination with the second highest log-likelihood from the gird search of the panel model with the addition of the parameter κ_{suburb}, which represents the proportion of the workday spent at the work location for movement fluxes that do not start or end in the city of Rio. The size of this parameter is a proxy for the importance of non-Rio fluxes. The blue line is equal to two log-likelihood units below the peak of the slice. The fact that the slice peaks at a non-zero value indicates that non-Rio fluxes may play a role in the dynamics. However, since the panel and coupled models are not fully nested and this slice analysis is not a true profile, we cannot make this assumption.

$$J_v = \kappa m_{rv} I_{\text{rio}_{\text{act}}} + [1 - \kappa m_{vr} - \sum_{j=1, j\neq v}^{U} \kappa_{\text{suburb}} m_{vj}] I_v + \sum_{w=1, w\neq v}^{U} \kappa_{\text{suburb}} m_{wv} I_w \quad (4.67)$$

Increment the total working population in city v (Q_v) by the number of people who live in city w but work in city v

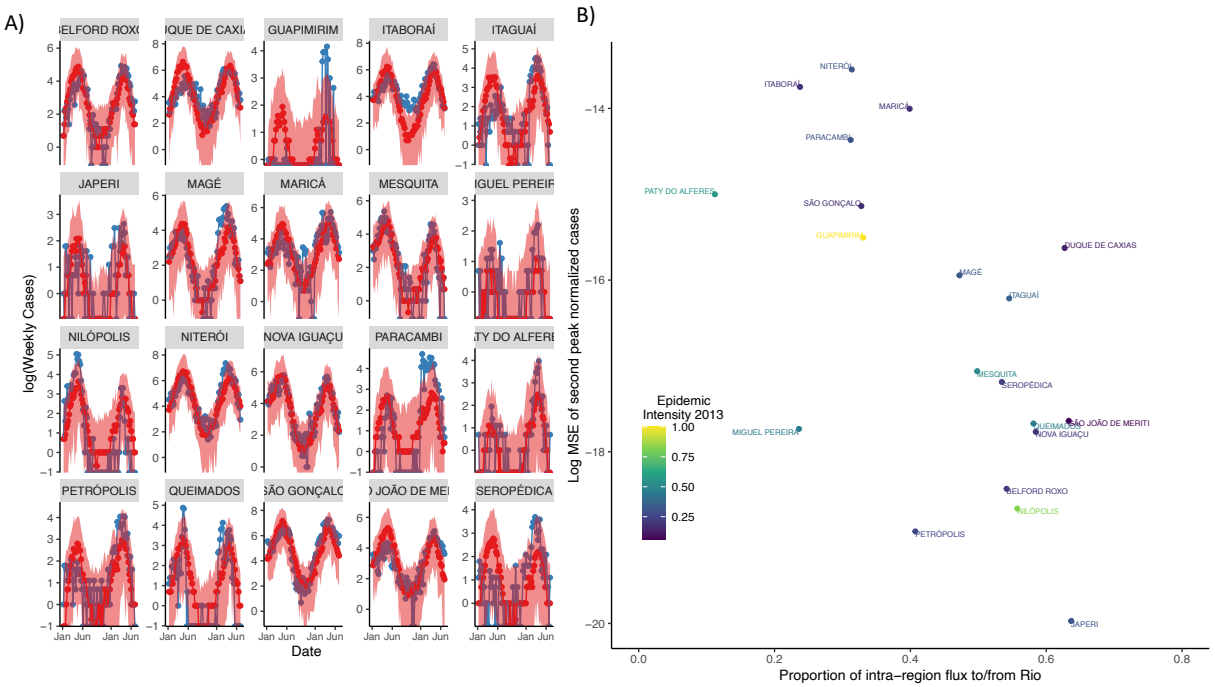

Figure 4.11: **Small sigma P MLE**. Plot of observed vs simulated cases and MSE vs proportion of Rio flux for parameter combination that had the highest log-likelihood out of all parameter combinations with low environmental noise. This log-likelihood is several hundreds units lower than the overall MLE.

$$Q_v = m_{\text{work}_{rv}} N_{\text{rio}} \ + \sum_{w=1}^{U} m_{\text{work}_{wv}} N_w \tag{4.68}$$

Break apart the summation terms:

$$Q_v = m_{\text{work}_{rv}} N_{\text{rio}} \ + m_{\text{work}_{vv}} N_v + \sum_{w=1, w\neq v}^{U} m_{\text{work}_{wv}} N_w \tag{4.69}$$

Plugging in:

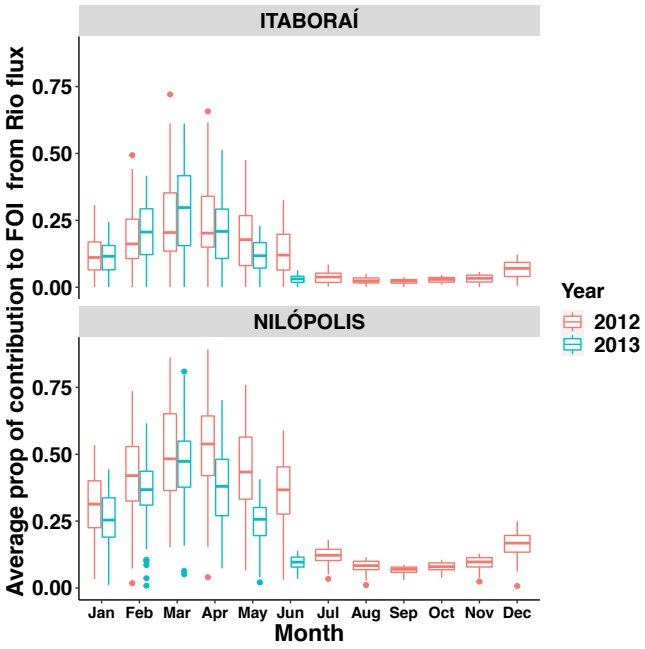

Figure 4.12: **Contribution to the force of infection in Itaborai and Nilopolis from commuters who work in Rio de Janeiro during the day.** The force of infection was calculated for 100 trajectories from the MLE for both municipalities, and contributions from Rio were averaged across all trajectories for each month and municipality. Movement from Rio makes a substantial contribution to the force of infection in Nilopolis, which is located just north of Rio de Janeiro and has a large proportion of flux to Rio. Movement from Rio makes a smaller contribution to the force of infection in Itaborai, which is located at the eastern edge of the region and has a low proportion of flux to Rio. Itaborai does have substantial commuter traffic to some of the suburbs of Rio such as Niteroi which is not captured in the panel model.

$$Q_v = \kappa m_{rv} N_{\text{rio}} + [1 - \kappa m_{vr} - \sum_{j=1, j\neq v}^{U} \kappa_{\text{suburb}} m_{vj}] N_v + \sum_{w=1, w\neq v}^{U} \kappa_{\text{suburb}} m_{wv} N_w \quad (4.70)$$

Calculate the total force of infection F experienced by an individual working in city v:

$$F_v = \frac{\beta_v J_v + \epsilon}{Q_v} dW \quad (4.71)$$

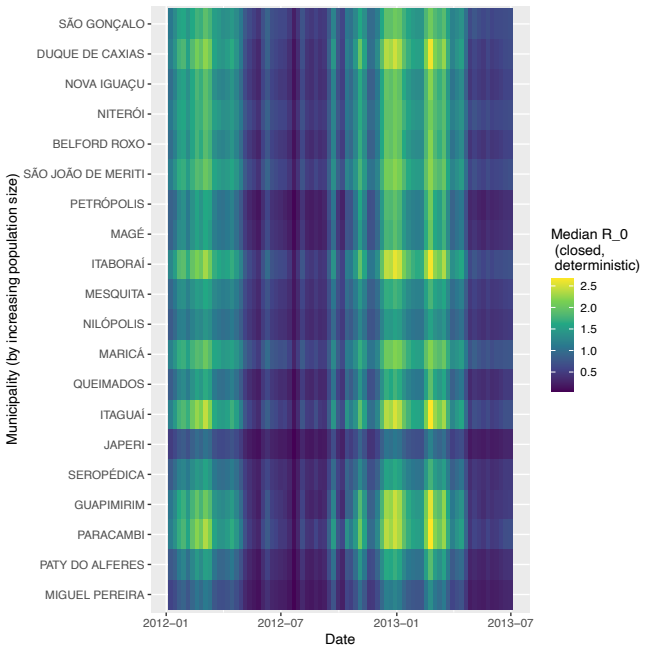

Figure 4.13: **Heatmap of the deterministic reproductive number R_0 from the MLE for each date and municipality assuming a fully closed SEIR model with no movement from Rio.** To estimate this quantity, we use the temperature-dependent transmission rate from the panel model at the MLE parameter values at each observation date for each location, ignoring environmental stochasticity.

Plugging in we obtain:

$$F_v = \frac{\beta_v[\kappa m_{rv} I_{\text{rio}_{\text{act}}} + [1 - \kappa m_{vr} - \sum_{j=1, j\neq v}^{U} \kappa_{\text{suburb}} m_{vj}] I_v + \sum_{w=1, w\neq v}^{U} \kappa_{\text{suburb}} m_{wv} I_w] + \epsilon}{\kappa m_{rv} N_{\text{rio}} + [1 - \kappa m_{vr} - \sum_{j=1, j\neq v}^{U} \kappa_{\text{suburb}} m_{vj}] N_v + \sum_{w=1, w\neq v}^{U} \kappa_{\text{suburb}} m_{wv} N_w} dW \tag{4.72}$$

Update force of infection for people living in city u with the force of infection that someone working in city v would experience. Weight by the probability that someone living in city u would work in city v:

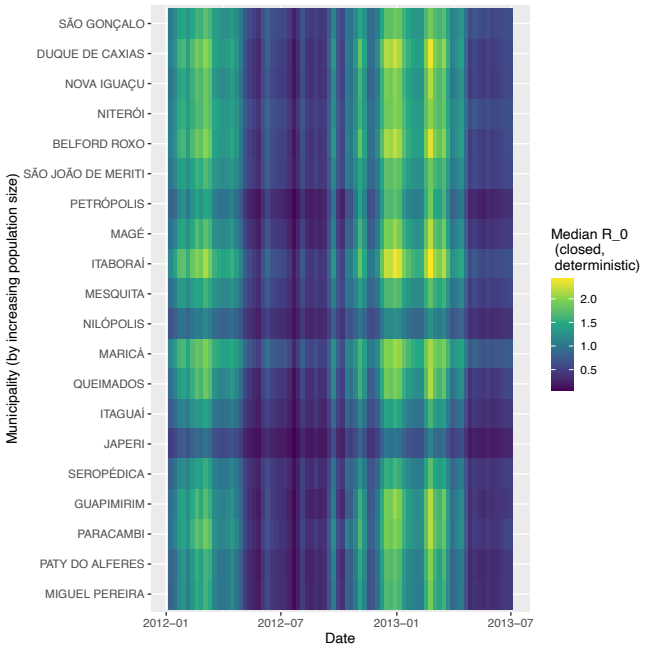

Figure 4.14: **Heatmap of the deterministic reproductive number R_0 from the parameter combination with the second-highest log-likelihood from the grid search for each date and municipality assuming a fully closed SEIR model with no movement from Rio.** To estimate this quantity, we use the temperature-dependent transmission rate from the panel model at the second-highest loglikelihood parameter values at each observation date for each location, ignoring environmental stochasticity.

$$\lambda_u = \sum_{w=1}^{U+1} m_{\text{work}_{uv}} F_v; \tag{4.73}$$

We experimented with several simplified versions of this fully coupled model assuming that only people who live in Rio de Janeiro contribute to the force of infection in Rio de Janeiro. Those equations are omitted here for brevity.

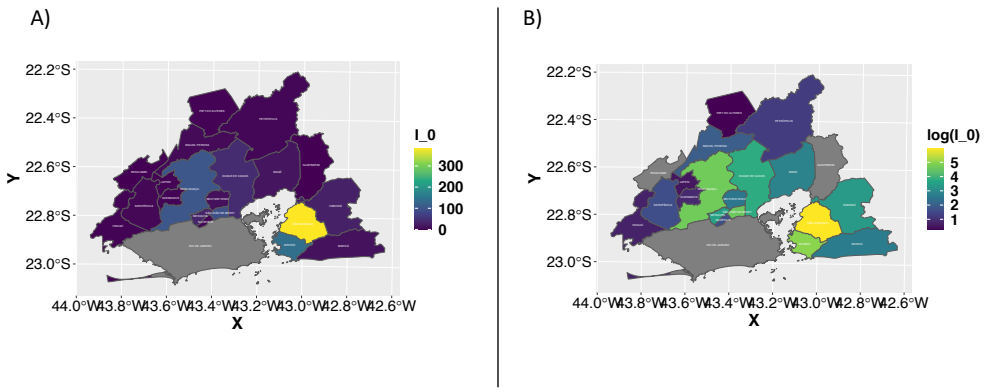

Figure 4.15: **Initial value parameters (the initial infected population I_0) for the parameter combination with the highest log-likelihood from the panel grid search (the MLE) on a A) regular and B) log scale.**

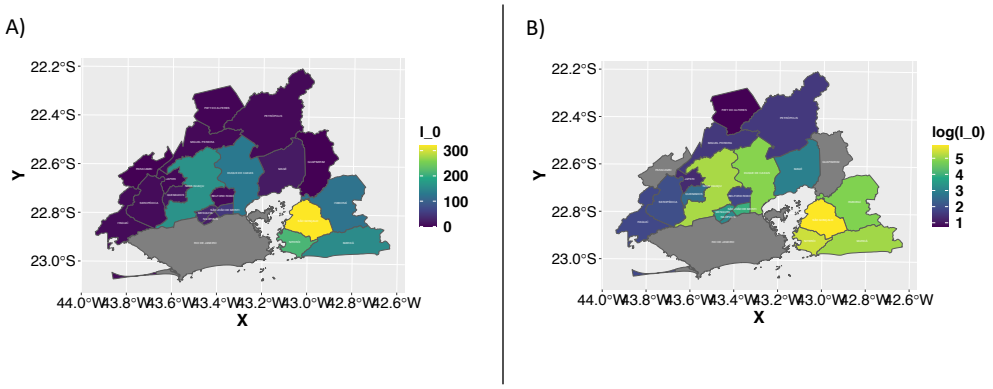

Figure 4.16: **Initial value parameters (the initial infected population I_0) for the parameter combination with the second-highest log-likelihood from the panel grid search on a A) regular and B) log scale.**

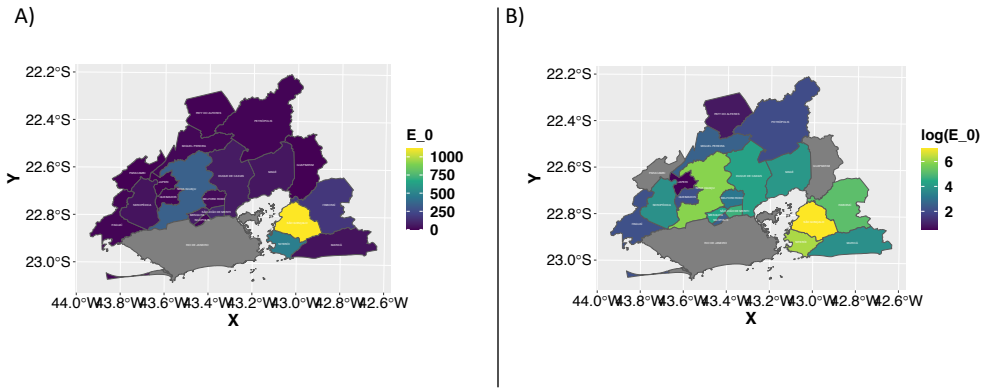

Figure 4.17: **Initial value parameters (the initial exposed population E_0) for the parameter combination with highest log-likelihood from the panel grid search (the MLE) on a A) regular and B) log scale.**

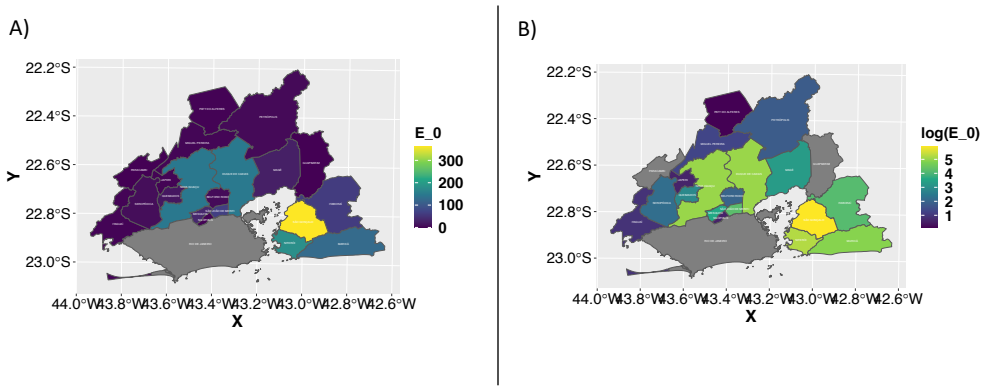

Figure 4.18: **Initial value parameters (the initial exposed population E_0) for the parameter combination with the second-highest log-likelihood from the panel grid search on a A) regular and B) log scale.**

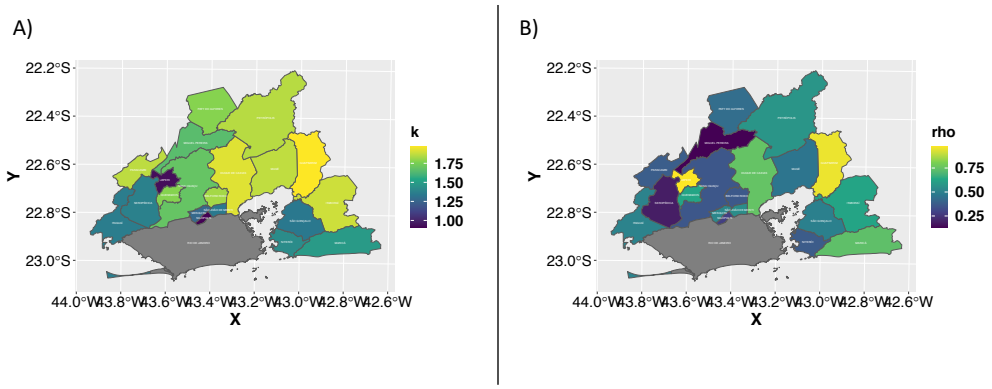

Figure 4.19: **Maps of unit-specific parameters (reporting rate ρ and mosquito to human population ratio k) from the MLE parameter combination.** Note that the value of k shown here is the parameter value multiplied by 5, as is the case within the process model. This re-scaling in conjunction with a logit parameter transformation ensures that the mosquito to human population ratio is bounded between 0 and 5.

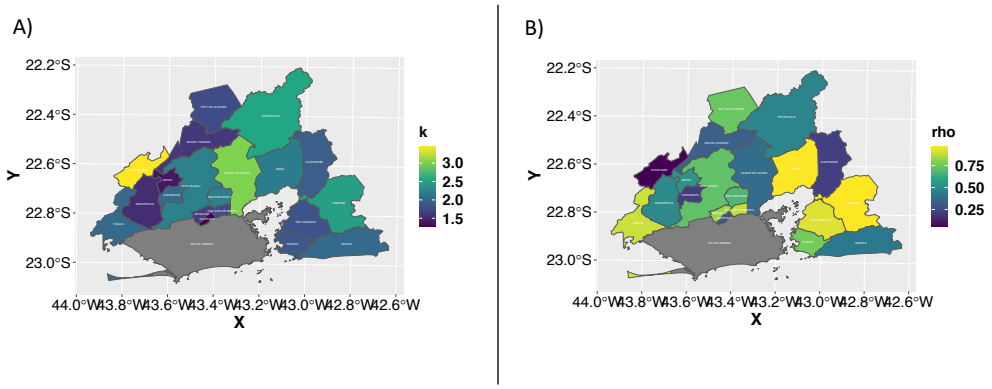

Figure 4.20: **Maps of unit-specific parameters (reporting rate ρ and mosquito to human population ratio k) from the panel grid search parameter combination with the second-highest log-likelihood.** Note that the value of k shown here is the parameter value multiplied by 5, as is the case within the process model. This re-scaling in conjunction with a logit parameter transformation ensures that the mosquito to human population ratio is bounded between 0 and 5.

CHAPTER 5

CONCLUSION

5.1 Concluding remarks

Large cities and their surrounding metropolitan areas play an important role in the transmission of emerging viruses such as COVID-19 and dengue. This thesis uses newly developed inference approaches to characterize the extent and consequences of heterogeneity in both symptom severity and drivers of transmission for these pathogens in these locations.

Quantifying heterogeneity in symptom severity and factors driving transmission in large cities can be challenging using traditional mathematical models designed for more endemic diseases such as measles [8] or seasonal influenza [10]. For newly emergent pathogens such as COVID-19, disease reporting rates may increases over time as testing capacity increases, confounding estimates of the proportion of cases that are symptomatic. The first chapter uses a model that incorporates daily changes in testing capacity to precisely estimate the fraction of COVID-19 cases that are symptomatic and shows that non-symptomatic cases contribute substantially to community transmission in New York City in Spring 2020.

The transmission of arboviruses such as dengue, Zika, and chikungunya can be influenced by numerous factors, including immunity, temperature, rainfall, population density, socio-economic status, and human movement, which can act at different scales. Understanding which scales are relevant is especially crucial for large cities, and can be viewed at a variety of scales. At a very large scale, a city may appear to be a single well-mixed population. However, at smaller scales, a large city can be viewed as a meta-population of numerous highly connected neighborhoods with considerable heterogeneity in population density and socio-economic status. This large city may itself be part of a meta-population consisting of other cities within the metropolitan area that are highly connected via commuter movement.

The second chapter shows that susceptible depletion at a city-level assuming a well-mixed population with seasonality is insufficient to explain the rapid re-emergence of dengue

serotype DENV1 following its initial invasion in 1986 in Rio de Janeiro, Brazil. This result suggests that other drivers of transmission such as inter-annual climate variation, small-scale heterogeneity in population density, and human movement between municipalities may play an important role in the spread and re-emergence of dengue in large cities.

Spatiotemporal statistical models that represent inter-annual climate variation and socioeconomic variables provide one potential tool for understanding how dengue serotypes spread in large cities [37, 51]. However, these models have difficulty incorporating human movement patterns [51]. Specific features of a movement network can be added to a statistical model, but this requires understanding which specific types of movement fluxes within a metropolitan area are epidemiologically meaningful. In the third chapter, we use a panel of mechanistic models to show that human movement to and from the city of Rio de Janeiro played an important role in the spread of dengue serotype DENV4 from 2012-2013 and suggests that additional commuter movement to and from large suburbs of Rio may also be important in facilitating the spread of the virus to more remote areas. Both types of fluxes could be incorporated into statistical models [37, 51] to improve our understanding of how new dengue serotypes spread in metropolitan areas.

Overall, this thesis uses a variety of spatial scales and model frameworks to characterize the extent and impact of heterogeneity in infection status and transmission for emerging viral pathogens.

5.2 Future directions

There are several potential lines of inquiry that could be pursued in future. Two of those lines of inquiry are described below. The results from the second chapter of the thesis suggest that treating the city as a well-mixed population may not be sufficient to capture the dynamics of susceptible depletion. These results are consistent with intriguing empirical patterns for reported cases at a very fine scale within the city of Rio de Janeiro [54]. In particular, patterns for the ratio of successive epidemic waves indicate an important role of

population density at the extremely fine scales of census tract (about a block or two). This ratio is of central interest because it reflects the interplay of herd immunity build-up and transmission seasonality. The patterns suggest a that the population density may impact the arrival time of dengue cases in a unit. Furthermore, the rate at which new "sparks" of infected cases arrive in units in the city during the 2012 DENV4 invasion can be written as a function of the population density of that unit and the total prevalence at the level of the whole city [54]. The patterns further suggest a key role of population density through arrival time. The rate at which new "sparks" of infected cases arrive in units in the city during the 2012 DENV4 invasion is both a function of the population density of that unit as well as the level of total prevalence of the whole city [54]. This suggests a simpler metapopulation model than the typical formulation requiring the full coupling among all units, in which local population density is finely resolved but coupling is global [54].

One could potentially verify whether this spark model can explain dynamics observed at a city-wide scale by simulating at a fine resolution but subsequently aggregating the simulated cases and fitting at a higher resolution. Specifically, instead of aggregating cases from geographically contiguous units, one could aggregate simulated cases from units with similar population densities. The composite time series from each population density "grouping" could then be fit to the composite time series of observed cases from the same units. If the observed total number of cases in the city and the spark arrival function are treated as known co-variates, then each grouping can be considered to be independent and the panel iterated filtering algorithm PIF [192] can be used for inference. If the depletion of susceptible hosts at a local scale using the "spark" model can explain observed patterns of dengue dynamics at a city-wide scale, this would be a strong argument for the importance of heterogeneity in population density in driving the spread of new dengue serotypes in large metropolitan areas.

Likewise, the results of the third chapter could also be further extended by fitting the fully coupled model of DENV4 transmission in the Rio metro area to case data to quantify

the role played by suburban movement hubs. Several algorithms that can cope with the curse of dimensionality can potentially be used to evaluate the likelihood of the coupled model are currently implemented in the R package spatPomp [58], including the Guided-Intermediate Resampling Filter or GIRF [185], the Un-adapted Bag Filter or UBF [186], the Ensemble Kalman Filter (EnKF) [187], and the Block Particle Filter [57]. Iterated versions of GIRF, and UBF, and EnKF that can be used for inference have also been implemented thus far in the package [58]. However, experimentation with several of these techniques using the coupled model revealed that the I-GIRF algorithm has a long run-time which may make rapid inference and model iteration difficult, while the I-UBF is extremely fast and parallelizable but computationally expensive. The I-ENKF algorithm could theoretically be used to fit the fully coupled model, although it should not be used to infer the environmental process noise and measurement noise parameters in the model. Recall that in the dengue model, the environmental stochasticity term represents the effects of climate variables such as rainfall which are known to be important but cannot be easily implemented in a mechanistic dynamic model. Since the value of this parameter is expected to be large and important, this parameter may need to be profiled when using the iterated ensemble Kalman filter for inference. This inference step would be substantially facilitated by the development of an iterated version of the block particle filter, since this algorithm is fast, computationally affordable, and reasonably accurate when evaluating the likelihood for spatiotemporal models and can be used to infer noise parameters. Additionally, to control for the complexity of the movement network, one could use a version of the coupled model for fitting with the same number of edges as the panel model, except with a different topology. For example, one could use the twenty largest commuter movement fluxes across the whole region, regardless of whether those fluxes include the city of Rio de Janeiro. Quantifying the role played by secondary movement hubs in the spread of DENV4 to outlying municipalities would provide additional information on essential movement fluxes that could be incorporated into future larger-scale statistical models of dengue and contribute to a major-scale understanding of

the drivers of dengue epidemics.

REFERENCES

[1] Daniel Stadlbauer, Jessica Tan, Kaijun Jiang, Matthew M Hernandez, Shelcie Fabre, Fatima Amanat, Catherine Teo, Guha Asthagiri Arunkumar, Meagan McMahon, Christina Capuano, et al. Repeated cross-sectional sero-monitoring of sars-cov-2 in new york city. *Nature*, 590(7844):146–150, 2021.

[2] Brazilian Institute of Geography (IBGE) and Statistics. Brazil demographic census 2010. *Rio de Janeiro, Brazil: Brazilian Institute of Geography and Statistics (IBGE)*, 2012. URL `https://censo2010.ibge.gov.br/en/noticias-censo.html?busca=1\&id=1\&idnoticia=2528\&t=life-expectancy-at-birth-was-74-6-years-in-2012\&view=noticia`.

[3] Brazilian Institute of Geography (IBGE) and Statistics. "censo demographico- 1991-rio de janeiro". *Rio de Janeiro, Brazil: Brazilian Institute of Geography and Statistics (IBGE)*, "Censo demográfico : 1991 : resultados do universo relativos as características da população e dos domicílios"(Table 1.4: "População residente, por grupos de idade, segundo ti lolesorregiàes, as Microrregiões, os Municípios,os Distritos e o sexo"): 32–41, 1991. URL `https://biblioteca.ibge.gov.br/biblioteca-catalogo?id=782\&view=detalhes`.

[4] Brazilian Institute of Geography (IBGE) and Statistics. Censo demográfico. *Rio de Janeiro, Brazil: Brazilian Institute of Geography and Statistics (IBGE)*, 7 (Tabela 7-"População residente, crescimento absoluto, participação relativa, e taxa média, geométrica de crescimento anual nos municípios mais populosos -1991/2000"), 2000. URL `https://biblioteca.ibge.gov.br/visualizacao/periodicos/308/cd_2000_v7.pdf`.

[5] Rita Maria R. Nogueira, Marize P. Miagostovich, Hermann G. Schatzmayr, Flávia B. dos Santos, Eliane S. M. de Araújo, Ana Maria B. de Filippis, Rogério V. de Souza, Sonia Maris O. Zagne, Cecília Nicolai, Mary Baran, and Gualberto Teixeira Filho. Dengue in the state of rio de janeiro, brazil, 1986-1998. *Memórias do Instituto Oswaldo Cruz*, 94:297–304, 1999. ISSN 0074-0276.

[6] R. M. R. Nogueira, M. P. Miagostovich, E. Lampe, R. W. Souza, S. M. O. Zagne, and H. G. Schatzmayr. Dengue epidemic in the state of rio de janeiro, brazil, 1990-1: Co-circulation of dengue 1 and dengue 2 serotypes. *Epidemiology and Infection*, 111 (1):163–170, 1993. ISSN 09502688, 14694409. URL `http://www.jstor.org/stable/3863761`.

[7] Luiz Tadeu Moraes Figueiredo, Silvia Maria Baeta Cavalcante, and Marcos Costa Simoes. Dengue serologic survey of schoolchildren in rio de janeiro, brazil, in 1986 and 1987. *Bull Pan Am Health Organ*, 1990.

[8] Lewi Stone, Ronen Olinky, and Amit Huppert. Seasonal dynamics of recurrent epidemics. *Nature*, 446(7135):533, 2007. ISSN 1476-4687.

[9] Claudia T Codeço, Daniel AM Villela, and Flavio C Coelho. Estimating the effective reproduction number of dengue considering temperature-dependent generation intervals. *Epidemics*, 25:101–111, 2018. ISSN 1755-4365.

[10] Benjamin D Dalziel, Stephen Kissler, Julia R Gog, Cecile Viboud, Ottar N Bjørnstad, C Jessica E Metcalf, and Bryan T Grenfell. Urbanization and humidity shape the intensity of influenza epidemics in us cities. *Science*, 362(6410):75–79, 2018.

[11] David J. D. Earn, Pejman Rohani, Benjamin M. Bolker, and Bryan T. Grenfell. A simple model for complex dynamical transitions in epidemics. *Science*, 287(5453):667–670, 2000. doi: 10.1126/science.287.5453.667. URL https://science.sciencemag.org/content/sci/287/5453/667.full.pdf.

[12] B. Finkenstädt and B. Grenfell. Empirical determinants of measles metapopulation dynamics in england and wales. *Proceedings. Biological sciences*, 265(1392):211–220, 1998. ISSN 0962-8452 1471-2954. doi: 10.1098/rspb.1998.0284. URL https://www.ncbi.nlm.nih.gov/pubmed/9493407https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1688869/.

[13] Bryan T Grenfell, Ottar N Bjørnstad, and Jens Kappey. Travelling waves and spatial hierarchies in measles epidemics. *Nature*, 414(6865):716–723, 2001.

[14] Matthieu Domenech de Cellès, Felicia M. G. Magpantay, Aaron A. King, and Pejman Rohani. The impact of past vaccination coverage and immunity on pertussis resurgence. *Science Translational Medicine*, 10(434):eaaj1748, 2018. doi: 10.1126/scitranslmed.aaj1748. URL https://stm.sciencemag.org/content/scitransmed/10/434/eaaj1748.full.pdf.

[15] Ottar N Bjørnstad, Bärbel F Finkenstädt, and Bryan T Grenfell. Dynamics of measles epidemics: estimating scaling of transmission rates using a time series sir model. *Ecological Monographs*, 72(2):169–184, 2002. ISSN 1557-7015.

[16] Alexander D Becker and Bryan T Grenfell. tsir: An r package for time-series susceptible-infected-recovered models of epidemics. *PloS one*, 12(9):e0185528, 2017.

[17] Daihai He, Edward L. Ionides, and Aaron A. King. Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. *Journal of The Royal Society Interface*, 7(43):271–283, 2010. doi: doi:10.1098/rsif.2009.0151. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2009.0151.

[18] Manoela Heringer, Thiara Manuele A Souza, Q Lima Monique da Rocha, Priscila Conrado G Nunes, Nieli Rodrigues da C Faria, Fernanda de Bruycker-Nogueira, Thaís Chouin-Carneiro, Rita Maria R Nogueira, and Flavia Barreto Dos Santos. Dengue type 4 in rio de janeiro, brazil: case characterization following its introduction in an endemic region. *BMC infectious diseases*, 17(1):1–9, 2017.

[19] Marize P Miagostovich, Rita MR Nogueira, Silvia Cavalcanti, Keyla BF Marzochi, and Hermann G Schatzmayr. Dengue epidemic in the state of rio de janeiro, brazil:

virological and epidemiological aspects. *Revista do Instituto de Medicina Tropical de São Paulo*, 35(2):149–154, 1993. ISSN 0036-4665.

[20] Ruiyun Li, Sen Pei, Bin Chen, Yimeng Song, Tao Zhang, Wan Yang, and Jeffrey Shaman. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov-2). *Science*, 368(6490):489–493, 2020. doi: 10.1126/science.abb3221. URL https://science.sciencemag.org/content/sci/368/6490/489.full.pdf.

[21] Jingbo LIANG, Hsiang-Yu Yuan, Lindsey Wu, and Dirk Udo Pfeiffer. Estimating effects of intervention measures on covid-19 outbreak in wuhan taking account of improving diagnostic capabilities using a modelling approach, 2020. URL https://www.medrxiv.org/content/medrxiv/early/2020/06/11/2020.03.31.20049387.full.pdf.

[22] New York State Department of Health. New york state statewide covid-19 testing, 2020. URL https://health.data.ny.gov/Health/New-York-State-Statewide-COVID-19-Testing/xdss-u53e.

[23] Cdc Covid-Response Team, Michelle A. Jorden, Sarah L. Rudman, Elsa Villarino, Stacey Hoferka, Megan T. Patel, Kelley Bemis, Cristal R. Simmons, Megan Jespersen, Jenna Iberg Johnson, Elizabeth Mytty, Katherine D. Arends, Justin J. Henderson, Robert W. Mathes, Charlene X. Weng, Jeffrey Duchin, Jennifer Lenahan, Natasha Close, Trevor Bedford, Michael Boeckh, Helen Y. Chu, Janet A. Englund, Michael Famulare, Deborah A. Nickerson, Mark J. Rieder, Jay Shendure, and Lea M. Starita. Evidence for limited early spread of covid-19 within the united states, january-february 2020. *MMWR. Morbidity and mortality weekly report*, 69(22):680–684, 2020. ISSN 1545-861X 0149-2195. doi: 10.15585/mmwr.mm6922e1. URL https://pubmed.ncbi.nlm.nih.gov/32497028https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7315848/.

[24] Louis Lambrechts, Krijn P. Paaijmans, Thanyalak Fansiri, Lauren B. Carrington, Laura D. Kramer, Matthew B. Thomas, and Thomas W. Scott. Impact of daily temperature fluctuations on dengue virus transmission by aedes aegypti. *Proceedings of the National Academy of Sciences*, 108(18):7460–7465, 2011. ISSN 0027-8424. doi: 10.1073/pnas.1101377108. URL https://www.pnas.org/content/108/18/7460.

[25] John H Huber, Marissa L Childs, Jamie M Caldwell, and Erin A Mordecai. Seasonal temperature variation influences climate suitability for dengue, chikungunya, and zika transmission. *PLoS neglected tropical diseases*, 12(5):e0006451, 2018.

[26] Florence Fouque, Romuald Carinci, Pascal Gaborit, Jean Issaly, Dominique J Bicout, and Philippe Sabatier. Aedes aegypti survival and dengue transmission patterns in french guiana. *Journal of Vector Ecology*, 31(2):390–400, 2006. ISSN 1081-1710.

[27] Corey M. Benedum, Osama M. E. Seidahmed, Elfatih A. B. Eltahir, and Natasha Markuzon. Statistical modeling of the effect of rainfall flushing on dengue transmission

in singapore. *PLOS Neglected Tropical Diseases*, 12(12):e0006935, 2018. doi: 10.1371/journal.pntd.0006935. URL https://doi.org/10.1371/journal.pntd.0006935.

[28] Osama M. E. Seidahmed and Elfatih A. B. Eltahir. A sequence of flushing and drying of breeding habitats of aedes aegypti (l.) prior to the low dengue season in singapore. *PLOS Neglected Tropical Diseases*, 10(7):e0004842, 2016. doi: 10.1371/journal.pntd.0004842. URL https://doi.org/10.1371/journal.pntd.0004842.

[29] Satya Ganesh Kakarla, Cyril Caminade, Srinivasa Rao Mutheneni, Andrew P. Morse, Suryanaryana Murty Upadhyayula, Madhusudhan Rao Kadiri, and Sriram Kumaraswamy. Lag effect of climatic variables on dengue burden in india. *Epidemiology and Infection*, 147:e170, 2019. ISSN 0950-2688. doi: 10.1017/S0950268819000608. URL https://www.cambridge.org/core/article/lag-effect-of-climatic-variables-on-dengue-burden-in-india/D9AC458913A43934395A64AC23BA3233.

[30] Manoela Heringer, Rita Maria R. Nogueira, Ana Maria B. de Filippis, Monique R. Q. Lima, Nieli R. C. Faria, Priscila C. G. Nunes, Fernanda B. Nogueira, and Flávia B. dos Santos. Impact of the emergence and re-emergence of different dengue viruses' serotypes in Rio de Janeiro, Brazil, 2010 to 2012. *Transactions of The Royal Society of Tropical Medicine and Hygiene*, 109(4):268–274, 01 2015. ISSN 0035-9203. doi: 10.1093/trstmh/trv006. URL https://doi.org/10.1093/trstmh/trv006.

[31] Ministério da Saúde. Secretaria de Vigilância em Saúde. Dengue: situação epidemiológica (de janeiro a abril de 2012). *Boletim epidemiológico*, 43(1), 2012. URL http://portalarquivos2.saude.gov.br/images/pdf/2014/julho/23/BE-2012-43--1--pag-11-a-15-Dengue.pdf.

[32] Nicholas G. Reich, Sourya Shrestha, Aaron A. King, Pejman Rohani, Justin Lessler, Siripen Kalayanarooj, In-Kyu Yoon, Robert V. Gibbons, Donald S. Burke, and Derek A. T. Cummings. Interactions between serotypes of dengue highlight epidemiological impact of cross-immunity. *Journal of The Royal Society Interface*, 10(86), 2013. URL http://rsif.royalsocietypublishing.org/content/10/86/20130414.abstract.

[33] Rita MR Nogueira and Ana LF Eppinghaus. Dengue virus type 4 arrives in the state of rio de janeiro: a challenge for epidemiological surveillance and control. *Memórias do Instituto Oswaldo Cruz*, 106(3):255–256, 2011. URL https://dx.doi.org/10.1590/S0074-02762011000300001.

[34] María Celeste Torres, Fernanda de Bruycker Nogueira, Carlos Augusto Fernandes, Guilherme Louzada Silva Meira, Shirlei Ferreira de Aguiar, Alexandre Otávio Chieppe, and Ana María Bispo de Filippis. Re-introduction of dengue virus serotype 2 in the state of rio de janeiro after almost a decade of epidemiological silence. *PLoS One*, 14 (12):e0225879, 2019.

[35] Ministério da Saúde. Secretaria de Vigilância em Saúde. Uma análise da situação de saúde e da epidemia pelo vírus zika e por outras doenças transmitidas pelo aedes aegypti.). *Saúde Brasil 2015/2016: uma análise da situação de saúde e da epidemia*

pelo vírus Zika e por outras doenças transmitidas pelo Aedes aegypti, pages 253–295, 2015/2016. URL http://portalarquivos2.saude.gov.br/images/pdf/2017/maio/12/2017-0135-vers-eletronica-final.pdf.

[36] Mario Recker, Konstantin B Blyuss, Cameron P Simmons, Tran Tinh Hien, Bridget Wills, Jeremy Farrar, and Sunetra Gupta. Immunological serotype interactions and their effect on the epidemiological pattern of dengue. *Proceedings of the Royal Society B: Biological Sciences*, 276(1667):2541–2548, 2009.

[37] Rachel Lowe, Sophie A Lee, Kathleen M O'Reilly, Oliver J Brady, Leonardo Bastos, Gabriel Carrasco-Escobar, Rafael de Castro Catão, Felipe J Colón-González, Christovam Barcellos, Marilia Sá Carvalho, et al. Combined effects of hydrometeorological hazards and urbanisation on dengue risk in brazil: a spatiotemporal modelling study. *The Lancet Planetary Health*, 5(4):e209–e219, 2021.

[38] Victoria Romeo-Aznar, Richard Paul, Olivier Telle, and Mercedes Pascual. Mosquito-borne transmission in urban landscapes: the missing link between vector abundance and human density. *Proceedings of the Royal Society B: Biological Sciences*, 285(1884):20180826, 2018. doi: doi:10.1098/rspb.2018.0826. URL https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2018.0826.

[39] Steven T. Stoddard, Brett M. Forshey, Amy C. Morrison, Valerie A. Paz-Soldan, Gonzalo M. Vazquez-Prokopec, Helvio Astete, Robert C. Reiner, Stalin Vilcarromero, John P. Elder, Eric S. Halsey, Tadeusz J. Kochel, Uriel Kitron, and Thomas W. Scott. House-to-house human movement drives dengue virus transmission. *Proceedings of the National Academy of Sciences*, 110(3):994–999, 2013. doi: 10.1073/pnas.1213349110. URL https://www.pnas.org/content/pnas/110/3/994.full.pdf.

[40] Amy Wesolowski, Taimur Qureshi, Maciej F Boni, Pål Roe Sundsøy, Michael A Johansson, Syed Basit Rasheed, Kenth Engø-Monsen, and Caroline O Buckee. Impact of human mobility on the emergence of dengue epidemics in pakistan. *Proceedings of the National Academy of Sciences*, 112(38):11887–11892, 2015.

[41] Guanghu Zhu, Jiming Liu, Qi Tan, and Benyun Shi. Inferring the spatio-temporal patterns of dengue transmission from surveillance data in guangzhou, china. *PLoS neglected tropical diseases*, 10(4):e0004633, 2016.

[42] Ashley L St John and Abhay PS Rathore. Adaptive immune responses to primary and secondary dengue virus infections. *Nature Reviews Immunology*, 19(4):218–230, 2019.

[43] Leah C. Katzelnick, Judith M. Fonville, Gregory D. Gromowski, Jose Bustos Arriaga, Angela Green, Sarah L. James, Louis Lau, Magelda Montoya, Chunling Wang, Laura A. VanBlargan, Colin A. Russell, Hlaing Myat Thu, Theodore C. Pierson, Philippe Buchy, John G. Aaskov, Jorge L. Muñoz-Jordán, Nikos Vasilakis, Robert V. Gibbons, Robert B. Tesh, Albert D. M. E. Osterhaus, Ron A. M. Fouchier, Anna Durbin, Cameron P. Simmons, Edward C. Holmes, Eva Harris, Stephen S. Whitehead, and Derek J. Smith. Dengue viruses cluster antigenically but not as discrete serotypes.

Science, 349(6254):1338, 2015. URL http://science.sciencemag.org/content/349/6254/1338.abstract.

[44] Leah C Katzelnick, César Narvaez, Sonia Arguello, Brenda Lopez Mercado, Damaris Collado, Oscarlett Ampie, Douglas Elizondo, Tatiana Miranda, Fausto Bustos Carillo, Juan Carlos Mercado, et al. Zika virus infection enhances future risk of severe dengue disease. *Science*, 369(6507):1123–1128, 2020.

[45] N. A. Honório, C. T. Codeço, F. C. Alves, M. A. F. M. Magalhães, and R. Lourenço-de Oliveira. Temporal distribution of aedes aegypti in different districts of rio de janeiro, brazil, measured by two types of traps. *Journal of Medical Entomology*, 46(5):1001–1014, 2009. ISSN 0022-2585. doi: 10.1603/033.046.0505. URL https://doi.org/10.1603/033.046.0505.

[46] Mario Recker, B. Blyuss Konstantin, P. Simmons Cameron, Tinh Hien Tran, Bridget Wills, Jeremy Farrar, and Sunetra Gupta. Immunological serotype interactions and their effect on the epidemiological pattern of dengue. *Proceedings of the Royal Society B: Biological Sciences*, 276(1667):2541–2548, 2009. doi: 10.1098/rspb.2009.0331. URL https://doi.org/10.1098/rspb.2009.0331.

[47] Aaron A. King, Dao Nguyen, and Edward L. Ionides. Statistical inference for partially observed markov processes via the r package pomp. *Journal of Statistical Software*, 69(12):43, 2016. ISSN 1548-7660. doi: 10.18637/jss.v069.i12. URL https://www.jstatsoft.org/v069/i12.

[48] Thomas Bengtsson, Peter Bickel, and Bo Li. Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. In *Probability and statistics: Essays in honor of David A. Freedman*, pages 316–334. Institute of Mathematical Statistics, 2008.

[49] Rachel Lowe, Trevor C. Bailey, David B. Stephenson, Tim E. Jupp, Richard J. Graham, Christovam Barcellos, and Marilia Sá Carvalho. The development of an early warning system for climate-sensitive disease risk with a focus on dengue epidemics in southeast brazil. *Statistics in Medicine*, 32(5):864–883, 2012. ISSN 0277-6715. doi: 10.1002/sim.5549. URL https://doi.org/10.1002/sim.5549.

[50] Christovam Barcellos and Rachel Lowe. Expansion of the dengue transmission area in brazil: the role of climate and cities. *Tropical medicine & international health*, 19(2):159–168, November 2013. URL https://researchonline.lshtm.ac.uk/id/eprint/3515677/.

[51] Mathew V Kiang, Mauricio Santillana, Jarvis T Chen, Jukka-Pekka Onnela, Nancy Krieger, Kenth Engø-Monsen, Nattwut Ekapirat, Darin Areechokchai, Preecha Prempree, Richard J Maude, et al. Incorporating human mobility data improves forecasts of dengue fever in thailand. *Scientific reports*, 11(1):1–12, 2021.

[52] Neil M. Ferguson, Zulma M. Cucunubá, Ilaria Dorigatti, Gemma L. Nedjati-Gilani, Christl A. Donnelly, Maria-Gloria Basáñez, Pierre Nouvellet, and Justin Lessler.

Countering the zika epidemic in latin america. *Science*, 353(6297):353–354, 2016. doi: 10.1126/science.aag0219. URL https://science.sciencemag.org/content/sci/353/6297/353.full.pdf.

[53] Ángel G. Muñoz, Madeleine C. Thomson, Anna M. Stewart-Ibarra, Gabriel A. Vecchi, Xandre Chourio, Patricia Nájera, Zelda Moran, and Xiaosong Yang. Could the recent zika epidemic have been predicted? *Frontiers in Microbiology*, 8:1291, 2017. ISSN 1664-302X. doi: 10.3389/fmicb.2017.01291. URL https://www.frontiersin.org/article/10.3389/fmicb.2017.01291.

[54] Victoria Romeo-Aznar, Laís Picinini Freitas, Oswaldo Gonçalves Cruz, Aaron A King, and Mercedes Pascual. Fine-scale heterogeneity in population density predicts wave dynamics in dengue epidemics. *medRxiv*, 2021. doi: 10.1101/2021.05.24.21257404. URL https://www.medrxiv.org/content/early/2021/05/24/2021.05.24.21257404.

[55] Derek AT Cummings, Rafael A Irizarry, Norden E Huang, Timothy P Endy, Ananda Nisalak, Kumnuan Ungchusak, and Donald S Burke. Travelling waves in the occurrence of dengue haemorrhagic fever in thailand. *Nature*, 427(6972):344–347, 2004.

[56] City of New York, Office of the Mayor. Emergency executive order no. 100, 2020. URL https://www1.nyc.gov/assets/home/downloads/pdf/executive-orders/2020/eeo-100.pdf.

[57] Patrick Rebeschini and Ramon Van Handel. Can local particle filters beat the curse of dimensionality? *The Annals of Applied Probability*, 25(5):2809–2866, 2015.

[58] Kidus Asfaw, Joonha Park, Allister Ho, Aaron A King, and Edward Ionides. Statistical inference for spatiotemporal partially observed markov processes via the r package spatpomp. *arXiv preprint arXiv:2101.01157*, 2021.

[59] Catharine I. Paules, Hilary D. Marston, and Anthony S. Fauci. Coronavirus infections—more than just the common cold. *JAMA*, 323(8):707–708, 2020. ISSN 0098-7484. doi: 10.1001/jama.2020.0757. URL https://doi.org/10.1001/jama.2020.0757.

[60] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020. ISSN 1473-3099.

[61] Centers for Disease Control and Prevention. Discontinuation of isolation for persons with covid-19 not in healthcare settings, 2020. URL https://www.cdc.gov/coronavirus/2019-ncov/hcp/disposition-in-home-patients.html\#previous-updates.

[62] Maimuna Majumder and Kenneth D Mandl. Early transmissibility assessment of a novel coronavirus in wuhan, china (january 26, 2020), 2020. URL https://ssrn.com/abstract=3524675.

[63] Steven Sanche, Yen Ting Lin, Chonggang Xu, Ethan Romero-Severson, Nick Hengartner, and Ruian Ke. High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. *Emerging Infectious Disease journal*, 26(7):1470, 2020. ISSN 1080-6059. doi: 10.3201/eid2607.200282. URL https://wwwnc.cdc.gov/eid/article/26/7/20-0282_article.

[64] Seyed M. Moghadas, Meagan C. Fitzpatrick, Pratha Sah, Abhishek Pandey, Affan Shoukat, Burton H. Singer, and Alison P. Galvani. The implications of silent transmission for the control of covid-19 outbreaks. *Proceedings of the National Academy of Sciences*, 117(30):17513–17515, 2020. doi: 10.1073/pnas.2008373117. URL https://www.pnas.org/content/pnas/117/30/17513.full.pdf.

[65] Jose Lourenco, Robert Paton, Mahan Ghafari, Moritz Kraemer, Craig Thompson, Peter Simmonds, Paul Klenerman, and Sunetra Gupta. Fundamental principles of epidemic spread highlight the immediate need for large-scale serological surveys to assess the stage of the sars-cov-2 epidemic, 2020. URL https://www.medrxiv.org/content/medrxiv/early/2020/03/26/2020.03.24.20042291.full.pdf.

[66] Ashish Goyal, Daniel B Reeves, E. Fabian Cardozo-Ojeda, Joshua T Schiffer, and Bryan T. Mayer. Wrong person, place and time: viral load and contact network structure predict sars-cov-2 transmission and super-spreading events, 2020. URL https://www.medrxiv.org/content/medrxiv/early/2020/08/07/2020.08.07.20169920.1.full.pdf.

[67] Juanjuan Zhang, Maria Litvinova, Wei Wang, Yan Wang, Xiaowei Deng, Xinghui Chen, Mei Li, Wen Zheng, Lan Yi, Xinhua Chen, Qianhui Wu, Yuxia Liang, Xiling Wang, Juan Yang, Kaiyuan Sun, Jr. Longini, Ira M., M. Elizabeth Halloran, Peng Wu, Benjamin J. Cowling, Stefano Merler, Cecile Viboud, Alessandro Vespignani, Marco Ajelli, and Hongjie Yu. Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside hubei province, china: a descriptive and modelling study. *The Lancet Infectious Diseases*, 20(7):793–802, 2020. ISSN 1473-3099. doi: 10.1016/S1473-3099(20)30230-9. URL https://doi.org/10.1016/S1473-3099(20)30230-9.

[68] Ying Liu, Albert A Gayle, Annelies Wilder-Smith, and Joacim Rocklöv. The reproductive number of covid-19 is higher compared to sars coronavirus. *Journal of Travel Medicine*, 27(2), 2020. ISSN 1708-8305. doi: 10.1093/jtm/taaa021. URL https://doi.org/10.1093/jtm/taaa021.

[69] Benjamin J. Cowling, Sheikh Taslim Ali, Tiffany W. Y. Ng, Tim K. Tsang, Julian C. M. Li, Min Whui Fong, Qiuyan Liao, Mike Y. W. Kwan, So Lun Lee, Susan S. Chiu, Joseph T. Wu, Peng Wu, and Gabriel M. Leung. Impact assessment of non-pharmaceutical interventions against coronavirus disease 2019 and influenza in hong kong: an observational study. *The Lancet Public Health*, 5(5):e279–e288, 2020. ISSN 2468-2667. doi: 10.1016/S2468-2667(20)30090-6. URL https://doi.org/10.1016/S2468-2667(20)30090-6.

[70] Tapiwa Ganyani, Cécile Kremer, Dongxuan Chen, Andrea Torneri, Christel Faes, Jacco Wallinga, and Niel Hens. Estimating the generation interval for coronavirus disease (covid-19) based on symptom onset data, march 2020. *Eurosurveillance*, 25(17):2000257, 2020. doi: doi:https://doi.org/10.2807/1560-7917.ES.2020.25.17.2000257. URL https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2020.25.17.2000257.

[71] Centers for Disease Control and Prevention. Covid-19 pandemic planning scenarios, 2020. URL https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html.

[72] Piero Poletti, Marcello Tirani, Danilo Cereda, Filippo Trentini, Giorgio Guzzetta, Giuliana Sabatino, Valentina Marziano, Ambra Castrofino, Francesca Grosso, Gabriele Del Castillo, Raffaella Piccarreta, ATS Lombardy COVID-19 Task Force, Aida Andreassi, Alessia Melegaro, Maria Gramegna, Marco Ajelli, and Stefano Merler. Probability of symptoms and critical disease after sars-cov-2 infection, 2020.

[73] Oyungerel Byambasuren, Magnolia Cardona, Katy Bell, Justin Clark, Mary-Louise McLaws, and Paul Glasziou. Estimating the extent of asymptomatic covid-19 and its potential for community transmission: systematic review and meta-analysis, 2020. URL https://www.medrxiv.org/content/early/2020/06/04/2020.05.10.20097543.

[74] Kenji Mizumoto, Katsushi Kagaya, Alexander Zarebski, and Gerardo Chowell. Estimating the asymptomatic proportion of coronavirus disease 2019 (covid-19) cases on board the diamond princess cruise ship, yokohama, japan, 2020. *Eurosurveillance*, 25(10):2000180, 2020. doi: doi:https://doi.org/10.2807/1560-7917.ES.2020.25.10.2000180. URL https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2020.25.10.2000180.

[75] Hiroshi Nishiura, Tetsuro Kobayashi, Takeshi Miyama, Ayako Suzuki, Sung-Mok Jung, Katsuma Hayashi, Ryo Kinoshita, Yichi Yang, Baoyin Yuan, Andrei R. Akhmetzhanov, and Natalie M. Linton. Estimation of the asymptomatic ratio of novel coronavirus infections (covid-19). *International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases*, 94:154–155, 2020. ISSN 1878-3511 1201-9712. doi: 10.1016/j.ijid.2020.03.020. URL https://pubmed.ncbi.nlm.nih.gov/32179137https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7270890/.

[76] Matt Feaster and Ying-Ying Goh. High proportion of asymptomatic sars-cov-2 infections in 9 long-term care facilities, pasadena, california, usa, april 2020. *Emerging Infectious Diseases*, 26(10), 2020.

[77] Qing Xie, Jing Wang, Jianling You, Shida Zhu, Rui Zhou, Zhijian Tian, Hao Wu, Yang Lin, Wei Chen, Lan Xiao, Xin Jin, Jianjuan Li, Jie Dong, Honglong Wu, Wei Zhang, Jing Li, Xun Xu, Ye Yin, Feng Mu, Weijun Chen, and Jian Wang. Effect of large-scale testing platform in prevention and control of the covid-19 pandemic: an empirical study with a novel numerical model, 2020. URL https://www.medrxiv.org/content/medrxiv/early/2020/03/20/2020.03.15.20036624.full.pdf.

[78] New York City Department of Health Hygiene and Mental. Covid-19: Data, 2020.

[79] Hiroshi Nishiura, Natalie M. Linton, and Andrei R. Akhmetzhanov. Serial interval of novel coronavirus (covid-19) infections. *International Journal of Infectious Diseases*, 93:284–286, 2020. ISSN 1201-9712. doi: https://doi.org/10.1016/j.ijid.2020.02.060. URL http://www.sciencedirect.com/science/article/pii/S1201971220301193.

[80] Marino Gatto, Enrico Bertuzzo, Lorenzo Mari, Stefano Miccoli, Luca Carraro, Renato Casagrandi, and Andrea Rinaldo. Spread and dynamics of the covid-19 epidemic in italy: Effects of emergency containment measures. *Proceedings of the National Academy of Sciences*, 117(19):10484–10491, 2020. doi: 10.1073/pnas.2004978117. URL https://www.pnas.org/content/pnas/117/19/10484.full.pdf.

[81] Jessica T Davis, Matteo Chinazzi, Nicola Perra, Kunpeng Mu, Ana Pastore y Piontti, Marco Ajelli, Natalie E Dean, Corrado Gioannini, Maria Litvinova, Stefano Merler, Luca Rossi, Kaiyuan Sun, Xinyue Xiong, M. Elizabeth Halloran, Ira M Longini, Cécile Viboud, and Alessandro Vespignani. Estimating the establishment of local transmission and the cryptic phase of the covid-19 pandemic in the usa, 2020. URL https://www.medrxiv.org/content/medrxiv/early/2020/07/07/2020.07.06.20140285.full.pdf.

[82] Hiroshi Nishiura, Tetsuro Kobayashi, Yichi Yang, Katsuma Hayashi, Takeshi Miyama, Ryo Kinoshita, Natalie M. Linton, Sung-mok Jung, Baoyin Yuan, Ayako Suzuki, and Andrei R. Akhmetzhanov. The rate of underascertainment of novel coronavirus (2019-ncov) infection: Estimation using japanese passengers data on evacuation flights. *Journal of Clinical Medicine*, 9(2):419, 2020. ISSN 2077-0383. URL https://www.mdpi.com/2077-0383/9/2/419.

[83] Reuters. Three japanese evacuees from wuhan test positive for virus, two had no symptoms, 2020. URL https://www.cnbc.com/2020/01/30/japanese-evacuees-from-wuhan-test-positive-for-virus.html.

[84] Neil Ferguson, Daniel Laydon, Gemma Nedjati Gilani, Natsuko Imai, Kylie Ainslie, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, ZULMA Cucunuba Perez, and Gina Cuomo-Dannenburg. Report 9: Impact of non-pharmaceutical interventions (npis) to reduce covid19 mortality and healthcare demand, 2020.

[85] Giacomo Grasselli, Antonio Pesenti, and Maurizio Cecconi. Critical care utilization for the covid-19 outbreak in lombardy, italy: Early experience and forecast during an emergency response. *JAMA*, 323(16):1545–1546, 2020. ISSN 0098-7484. doi: 10.1001/jama.2020.4031. URL https://doi.org/10.1001/jama.2020.4031.

[86] Jean-Louis Vincent and Fabio S. Taccone. Understanding pathways to death in patients with covid-19. *The Lancet Respiratory Medicine*, 8(5):430–432, 2020. ISSN 2213-2600. doi: 10.1016/S2213-2600(20)30165-X. URL https://doi.org/10.1016/S2213-2600(20)30165-X.

[87] Eboni G. Price-Haywood, Jeffrey Burton, Daniel Fort, and Leonardo Seoane. Hospitalization and mortality among black patients and white patients with covid-19. *New England Journal of Medicine*, 382(26):2534–2543, 2020. doi: 10.1056/NEJMsa2011686. URL https://www.nejm.org/doi/full/10.1056/NEJMsa2011686.

[88] Richardson Safiya, Hirsch Jamie S., Mangala Narasimhan, Crawford James M., McGinn Thomas, Karina W. Davidson, Consortium, and the Northwell COVID-19 Research. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with covid-19 in the new york city area. *JAMA*, 323(20): 2052–2059, 2020. ISSN 0098-7484. doi: 10.1001/jama.2020.6775. URL https://doi.org/10.1001/jama.2020.6775.

[89] Katelyn M Gostic, Lauren McGough, Edward Baskerville, Sam Abbott, Keya Joshi, Christine Tedijanto, Rebecca Kahn, Rene Niehus, James A Hay, Pablo M. De Salazar, Joel Hellewell, Sophie Meakin, James Munday, Nikos Bosse, Katharine Sherratt, Robin M Thompson, Laura F White, Jana Huisman, Jérémie Scire, Sebastian Bonhoeffer, Tanja Stadler, Jacco Wallinga, Sebastian Funk, Marc Lipsitch, and Sarah Cobey. Practical considerations for measuring the effective reproductive number, rt, 2020. URL https://www.medrxiv.org/content/medrxiv/early/2020/06/23/2020.06.18.20134858.full.pdf.

[90] Spencer J Fox, Remy Pasco, Mauricio Tec, Zhanwei Du, Michael Lachmann, James Scott, and Lauren Ancel Meyers. The impact of asymptomatic covid-19 infections on future pandemic waves, 2020. URL https://www.medrxiv.org/content/medrxiv/early/2020/06/23/2020.06.22.20137489.full.pdf.

[91] David Adam. A guide to r - the pandemic's misunderstood metric. *Nature*, 583(7816):346–348, 2020. ISSN 0028-0836. doi: 10.1038/d41586-020-02009-w. URL http://europepmc.org/abstract/MED/32620883https://doi.org/10.1038/d41586-020-02009-w.

[92] Ed Yong. The deceptively simple number sparking coronavirus fears, 2020. URL https://www.theatlantic.com/science/archive/2020/01/how-fast-and-far-will-new-coronavirus-spread/605632/.

[93] Erin Schumaker. What is r-naught for the covid-19 virus and why it's a key metric for re-opening plans, 2020. URL https://abcnews.go.com/Health/r0-covid-19-virus-key-metric-opening-plans/story?id=70868997.

[94] Natsuko Imai, Anne Cori, Ilaria Dorigatti, Marc Baguelin, Christl A Donnelly, Steven Riley, and Neil M Ferguson. Report 3: transmissibility of 2019-ncov, 2020.

[95] Jonathan M Read, Jessica RE Bridgen, Derek AT Cummings, Antonia Ho, and Chris P Jewell. Novel coronavirus 2019-ncov: early estimation of epidemiological parameters and epidemic predictions, 2020. URL https://www.medrxiv.org/content/medrxiv/early/2020/01/28/2020.01.23.20018549.full.pdf.

[96] Joseph T. Wu, Kathy Leung, Mary Bushman, Nishant Kishore, Rene Niehus, Pablo M. de Salazar, Benjamin J. Cowling, Marc Lipsitch, and Gabriel M. Leung. Estimating clinical severity of covid-19 from the transmission dynamics in wuhan, china. *Nature Medicine*, 26(4):506–510, 2020. ISSN 1546-170X. doi: 10.1038/s41591-020-0822-7. URL https://doi.org/10.1038/s41591-020-0822-7.

[97] An Pan, Li Liu, Chaolong Wang, Huan Guo, Xingjie Hao, Qi Wang, Jiao Huang, Na He, Hongjie Yu, Xihong Lin, Sheng Wei, and Tangchun Wu. Association of public health interventions with the epidemiology of the covid-19 outbreak in wuhan, china. *JAMA*, 323(19):1915–1923, 2020. ISSN 0098-7484. doi: 10.1001/jama.2020.6130. URL https://doi.org/10.1001/jama.2020.6130.

[98] Adam J. Kucharski, Timothy W. Russell, Charlie Diamond, Yang Liu, John Edmunds, Sebastian Funk, Rosalind M. Eggo, Fiona Sun, Mark Jit, James D. Munday, Nicholas Davies, Amy Gimma, Kevin van Zandvoort, Hamish Gibbs, Joel Hellewell, Christopher I. Jarvis, Sam Clifford, Billy J. Quilty, Nikos I. Bosse, Sam Abbott, Petra Klepac, and Stefan Flasche. Early dynamics of transmission and control of covid-19: a mathematical modelling study. *The Lancet Infectious Diseases*, 20 (5):553–558, 2020. ISSN 1473-3099. doi: 10.1016/S1473-3099(20)30144-4. URL https://doi.org/10.1016/S1473-3099(20)30144-4.

[99] Julien Riou and Christian L. Althaus. Pattern of early human-to-human transmission of wuhan 2019 novel coronavirus (2019-ncov), december 2019 to january 2020. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin*, 25(4):2000058, 2020. ISSN 1560-7917 1025-496X. doi: 10.2807/1560-7917.ES.2020.25.4.2000058. URL https://pubmed.ncbi.nlm.nih.gov/32019669https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7001239/.

[100] Seth Flaxman, Swapnil Mishra, Axel Gandy, H. Juliette T. Unwin, Thomas A. Mellan, Helen Coupland, Charles Whittaker, Harrison Zhu, Tresnia Berah, Jeffrey W. Eaton, Mélodie Monod, Pablo N. Perez-Guzman, Nora Schmit, Lucia Cilloni, Kylie E. C. Ainslie, Marc Baguelin, Adhiratha Boonyasiri, Olivia Boyd, Lorenzo Cattarino, Laura V. Cooper, Zulma Cucunubá, Gina Cuomo-Dannenburg, Amy Dighe, Bimandra Djaafara, Ilaria Dorigatti, Sabine L. van Elsland, Richard G. FitzJohn, Katy A. M. Gaythorpe, Lily Geidelberg, Nicholas C. Grassly, William D. Green, Timothy Hallett, Arran Hamlet, Wes Hinsley, Ben Jeffrey, Edward Knock, Daniel J. Laydon, Gemma Nedjati-Gilani, Pierre Nouvellet, Kris V. Parag, Igor Siveroni, Hayley A. Thompson, Robert Verity, Erik Volz, Caroline E. Walters, Haowei Wang, Yuanrong Wang, Oliver J. Watson, Peter Winskill, Xiaoyue Xi, Patrick G. T. Walker, Azra C. Ghani, Christl A. Donnelly, Steven Riley, Michaela A. C. Vollmer, Neil M. Ferguson, Lucy C. Okell, Samir Bhatt, and Covid-Response Team Imperial College. Estimating the effects of non-pharmaceutical interventions on covid-19 in europe. *Nature*, 584(7820):257–261, 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2405-7. URL https://doi.org/10.1038/s41586-020-2405-7.

[101] Yunjun Zhang, Yuying Li, Lu Wang, Mingyuan Li, and Xiaohua Zhou. Evaluating transmission heterogeneity and super-spreading event of covid-19 in a metropo-

lis of china. *International journal of environmental research and public health*, 17(10):3705, 2020. ISSN 1660-4601 1661-7827. doi: 10.3390/ijerph17103705. URL https://pubmed.ncbi.nlm.nih.gov/32456346https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7277812/.

[102] Clustering and superspreading potential of severe acute respiratory syndrome coronavirus 2 (sars-cov-2) infections in hong kong, 2020 2020. URL http://europepmc.org/abstract/PPR/PPR165671https://doi.org/10.21203/rs.3.rs-29548/v1.

[103] Alison P. Galvani and Robert M. May. Dimensions of superspreading. *Nature*, 438 (7066):293–295, 2005. ISSN 1476-4687. doi: 10.1038/438293a. URL https://doi.org/10.1038/438293a.

[104] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz. Superspreading and the effect of individual variation on disease emergence. *Nature*, 438(7066):355–359, 2005. ISSN 1476-4687. doi: 10.1038/nature04153. URL https://doi.org/10.1038/nature04153.

[105] M. Gabriela M. Gomes, Rodrigo M. Corder, Jessica G. King, Kate E. Langwig, Caetano Souto-Maior, Jorge Carneiro, Guilherme Goncalves, Carlos Penha-Goncalves, Marcelo U. Ferreira, and Ricardo Aguas. Individual variation in susceptibility or exposure to sars-cov-2 lowers the herd immunity threshold, 2020. URL https://www.medrxiv.org/content/medrxiv/early/2020/05/21/2020.04.27.20081893.full.pdf.

[106] Seungjae Lee, Tark Kim, Eunjung Lee, Cheolgu Lee, Hojung Kim, Heejeong Rhee, Se Yoon Park, Hyo-Ju Son, Shinae Yu, Jung Wan Park, Eun Ju Choo, Suyeon Park, Mark Loeb, and Tae Hyong Kim. Clinical Course and Molecular Viral Shedding Among Asymptomatic and Symptomatic Patients With SARS-CoV-2 Infection in a Community Treatment Center in the Republic of Korea, aug 2020. ISSN 2168-6106. URL https://doi.org/10.1001/jamainternmed.2020.3862.

[107] Young Joon Park, Young June Choe, Ok Park, Shin Young Park, Young-Man Kim, Jieun Kim, Sanghui Kweon, Yeonhee Woo, Jin Gwack, Seong Sun Kim, Jin Lee, Junghee Hyun, Boyeong Ryu, Yoon Suk Jang, Hwami Kim, Seung Hwan Shin, Seonju Yi, Sangeun Lee, Hee Kyoung Kim, Hyeyoung Lee, Yeowon Jin, Eunmi Park, Seung Woo Choi, Miyoung Kim, Jeongsuk Song, Si Won Choi, Dongwook Kim, Byoung-Hak Jeon, Hyosoon Yoo, and Eun Kyeong Jeong. Contact tracing during coronavirus disease outbreak, south korea, 2020. *Emerging Infectious Disease journal*, 26(10), 2020. ISSN 1080-6059. doi: 10.3201/eid2610.201315. URL https://wwwnc.cdc.gov/eid/article/26/10/20-1315$_{\text{a}}$rticle.

[108] Justin D. Silverman, Nathaniel Hupert, and Alex D. Washburne. Using influenza surveillance networks to estimate state-specific prevalence of sars-cov-2 in the united states. *Science Translational Medicine*, 12(554):eabc1126, 2020. doi: 10.1126/scitranslmed.abc1126. URL https://stm.sciencemag.org/content/scitransmed/12/554/eabc1126.full.pdf.

[109] Katherine M. Hiller, Lisa Stoneking, Alice Min, and Suzanne Michelle Rhodes. Syndromic surveillance for influenza in the emergency department–a systematic review. *PLOS ONE*, 8(9):e73832, 2013. doi: 10.1371/journal.pone.0073832. URL https://doi.org/10.1371/journal.pone.0073832.

[110] Plagianos Marlena Gehret, Y. Wu Winfred, McCullough Colleen, Paladini Marc, Lurio Joseph, D. Buck Michael, Calman Neil, and Soulakis Nicholas. Syndromic surveillance during pandemic (h1n1) 2009 outbreak, new york, new york, usa. *Emerging Infectious Disease journal*, 17(9):1724, 2011. ISSN 1080-6059. doi: 10.3201/eid1709.101357. URL https://wwwnc.cdc.gov/eid/article/17/9/10-1357_article.

[111] Fatima Amanat, Daniel Stadlbauer, Shirin Strohmeier, Thi H. O. Nguyen, Veronika Chromikova, Meagan McMahon, Kaijun Jiang, Guha Asthagiri Arunkumar, Denise Jurczyszak, Jose Polanco, Maria Bermudez-Gonzalez, Giulio Kleiner, Teresa Aydillo, Lisa Miorin, Daniel S. Fierer, Luz Amarilis Lugo, Erna Milunka Kojic, Jonathan Stoever, Sean T. H. Liu, Charlotte Cunningham-Rundles, Philip L. Felgner, Thomas Moran, Adolfo García-Sastre, Daniel Caplivski, Allen C. Cheng, Katherine Kedzierska, Olli Vapalahti, Jussi M. Hepojoki, Viviana Simon, and Florian Krammer. A serological assay to detect sars-cov-2 seroconversion in humans. *Nature Medicine*, 26(7):1033–1036, 2020. ISSN 1546-170X. doi: 10.1038/s41591-020-0913-5. URL https://doi.org/10.1038/s41591-020-0913-5.

[112] Ania Wajnberg, Fatima Amanat, Adolfo Firpo, Deena R. Altman, Mark J. Bailey, Mayce Mansour, Meagan McMahon, Philip Meade, Damodara Rao Mendu, Kimberly Muellers, Daniel Stadlbauer, Kimberly Stone, Shirin Strohmeier, Viviana Simon, Judith Aberg, David L. Reich, Florian Krammer, and Carlos Cordon-Cardo. Robust neutralizing antibodies to sars-cov-2 infection persist for months, 2020.

[113] Florian Krammer and Viviana Simon. Serology assays to manage covid-19. *Science*, 368(6495):1060–1061, 2020. ISSN 0036-8075. doi: 10.1126/science.abc1227. URL https://science.sciencemag.org/content/368/6495/1060.

[114] Alba Grifoni, Daniela Weiskopf, Sydney I. Ramirez, Jose Mateus, Jennifer M. Dan, Carolyn Rydyznski Moderbacher, Stephen A. Rawlings, Aaron Sutherland, Lakshmanane Premkumar, Ramesh S. Jadi, Daniel Marrama, Aravinda M. de Silva, April Frazier, Aaron F. Carlin, Jason A. Greenbaum, Bjoern Peters, Florian Krammer, Davey M. Smith, Shane Crotty, and Alessandro Sette. Targets of t cell responses to sars-cov-2 coronavirus in humans with covid-19 disease and unexposed individuals. *Cell*, 181(7):1489 – 1501.e15, 2020. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2020.05.015. URL http://www.sciencedirect.com/science/article/pii/S0092867420306103.

[115] Qian Zhang, Paul Bastard, and Zhiyong Liu. Inborn errors of type i ifn immunity in patients with life-threatening covid-19. *Science*, 370(6515), 2020. ISSN 0036-8075. doi: 10.1126/science.abd4570. URL https://science.sciencemag.org/content/370/6515/eabd4570.

[116] Paul Bastard, Lindsey B. Rosen, Qian Zhang, Eleftherios Michailidis, and Hoffmann. Autoantibodies against type i ifns in patients with life-threatening covid-19. *Science*, 370(6515), 2020. ISSN 0036-8075. doi: 10.1126/science.abd4585. URL https://science.sciencemag.org/content/370/6515/eabd4585.

[117] Helen Ward, Graham Cooke, Christina Atchison, Matthew Whitaker, Joshua Elliott, Maya Moshe, Jonathan C Brown, Barney Flower, Anna Daunt, Kylie Ainslie, Deborah Ashby, Christl Donnelly, Steven Riley, Ara Darzi, Wendy Barclay, and Paul Elliott. Declining prevalence of antibody positivity to sars-cov-2: a community study of 365,000 adults, 2020. URL https://www.medrxiv.org/content/early/2020/10/27/2020.10.26.20219725.

[118] J Zuo, A Dowell, H Pearce, K Verma, HM Long, J Begum, F Aiano, Z Amin-Chowdhury, B Hallis, L Stapley, R Borrow, E Linley, S Ahmad, B Parker, A Horsley, G Amirthalingam, K Brown, ME Ramsay, S Ladhani, and P Moss. Robust sars-cov-2-specific t-cell immunity is maintained at 6 months following primary infection, 2020. URL https://www.biorxiv.org/content/early/2020/11/02/2020.11.01.362319.

[119] Joseph R. Fauver, Mary E. Petrone, Emma B. Hodcroft, Kayoko Shioda, Hanna Y. Ehrlich, Alexander G. Watts, Chantal B. F. Vogels, Anderson F. Brito, Tara Alpert, Anthony Muyombwe, Jafar Razeq, Randy Downing, Nagarjuna R. Cheemarla, Anne L. Wyllie, Chaney C. Kalinich, Isabel M. Ott, Joshua Quick, Nicholas J. Loman, Karla M. Neugebauer, Alexander L. Greninger, Keith R. Jerome, Pavitra Roychoudhury, Hong Xie, Lasata Shrestha, Meei-Li Huang, Virginia E. Pitzer, Akiko Iwasaki, Saad B. Omer, Kamran Khan, Isaac I. Bogoch, Richard A. Martinello, Ellen F. Foxman, Marie L. Landry, Richard A. Neher, Albert I. Ko, and Nathan D. Grubaugh. Coast-to-coast spread of sars-cov-2 during the early epidemic in the united states. *Cell*, 181(5):990–996.e5, 2020. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2020.04.021. URL http://www.sciencedirect.com/science/article/pii/S0092867420304840.

[120] Roy M Anderson, B Anderson, and Robert M May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992. ISBN 019854040X.

[121] Hsiang-Yu Yuan, Guiyuan Han, Hsiangkuo Yuan, Susanne Pfeiffer, Axiu Mao, Lindsey Wu, and Dirk Pfeiffer. The importance of the timing of quarantine measures before symptom onset to prevent covid-19 outbreaks - illustrated by hong kong's intervention model, 2020. URL https://www.medrxiv.org/content/medrxiv/early/2020/05/06/2020.05.03.20089482.full.pdf.

[122] Stephen A. Lauer, Kyra H. Grantz, Qifang Bi, Forrest K. Jones, Qulu Zheng, Hannah R. Meredith, Andrew S. Azman, Nicholas G. Reich, and Justin Lessler. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and application. *Annals of internal medicine*, 172(9):577–582, 2020. ISSN 1539-3704 0003-4819. doi: 10.7326/M20-0504. URL https://pubmed.ncbi.nlm.nih.gov/32150748https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7081172/.

[123] City of New York, Office of the Mayor. Statement from mayor de blasio on bars, restaurants, and entertainment venues, 2020. URL https://www1.nyc.gov/office-of-the-mayor/news/152-20/statement-mayor-de-blasio-bars-restaurants-entertainment-venues.

[124] Press Office, Governor of New York. Governor cuomo signs the 'new york state on pause' executive order, 2020. URL https://www.governor.ny.gov/news/governor-cuomo-signs-new-york-state-pause-executive-order.

[125] U.S.Census Bureau. Quickfacts new york city, new york, 2010. URL https://www.census.gov/quickfacts/fact/table/newyorkcitynewyork/POP010210\#POP010210.

[126] U.S. Public Health Service. Priorities for testing patients with suspected covid-19 infection, 2020. URL https://www.fsmb.org/siteassets/advocacy/pdf/hhs-covid-19-testing-guidance.pdf.

[127] New York State Department of Health. Influenza laboratory-confirmed cases by county: Beginning 2009-10 season, 2020. URL https://health.data.ny.gov/Health/Influenza-Laboratory-Confirmed-Cases-By-County-Beg/jr8b-6gh6.

[128] City of New York, Department of Health. Syndromic surveillance data, 2020. URL https://a816-health.nyc.gov/hdi/epiquery/visualizations?PageType=ps\&PopulationSource=Syndromic.

[129] U.S. Food and Drug Administration. In vitro diagnostics euas:individual euas for molecular diagnostic tests for sars-cov-2, 2020. URL https://www.fda.gov/medical-devices/coronavirus-disease-2019-covid-19-emergency-use-authorizations-medical-devices/vitro-diagnostics-euas\#individual-molecular.

[130] NPR. Why it takes so long to get most covid-19 test results, 2020. URL https://www.npr.org/sections/health-shots/2020/03/28/822869504/why-it-takes-so-long-to-get-most-covid-19-test-results.

[131] Edward L. Ionides, Dao Nguyen, Yves Atchadé, Stilian Stoev, and Aaron A. King. Inference for dynamic and latent variable models via iterated, perturbed bayes maps. *Proceedings of the National Academy of Sciences*, 112(3):719–724, 2015. doi: 10.1073/pnas.1410597112. URL https://www.pnas.org/content/pnas/112/3/719.full.pdf.

[132] E. L. Ionides, C. Breto, J. Park, R. A. Smith, and A. A. King. Monte carlo profile confidence intervals for dynamic systems. *Journal of The Royal Society Interface*, 14(132):20170126, 2017. doi: doi:10.1098/rsif.2017.0126. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2017.0126.

[133] O. Diekmann, J. A. P. Heesterbeek, and M. G. Roberts. The construction of next-generation matrices for compartmental epidemic models. *Journal of the Royal Society, Interface*, 7(47):873–885, 2010. ISSN 1742-5662 1742-5689. doi:

10.1098/rsif.2009.0386. URL https://pubmed.ncbi.nlm.nih.gov/19892718https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2871801/.

[134] Centers for Disease Control and Prevention. Summary of the 2017-2018 influenza season, 2020. URL https://www.cdc.gov/flu/about/season/flu-season-2017-2018.htm.

[135] Centers for Disease Control and Prevention. Estimated influenza illnesses, medical visits, hospitalizations, and deaths in the united states — 2018–2019 influenza season, 2020. URL https://www.cdc.gov/flu/about/burden/2018-2019.html.

[136] Samir Bhatt, Peter W Gething, Oliver J Brady, Jane P Messina, Andrew W Farlow, Catherine L Moyes, John M Drake, John S Brownstein, Anne G Hoen, and Osman Sankoh. The global distribution and burden of dengue. *Nature*, 496(7446):504, 2013. ISSN 1476-4687.

[137] David Baud, Duane J Gubler, Bruno Schaub, Marion C Lanteri, and Didier Musso. An update on zika virus infection. *The Lancet*, 390(10107):2099–2109, 2017. ISSN 0140-6736.

[138] Lyle R. Petersen, Denise J. Jamieson, Ann M. Powers, and Margaret A. Honein. Zika virus. *New England Journal of Medicine*, 374(16):1552–1563, 2016. doi: 10.1056/NEJMra1602113. URL http://www.nejm.org/doi/full/10.1056/NEJMra1602113.

[139] Roger S. Nasci. Movement of chikungunya virus into the western hemisphere. *Emerging Infectious Diseases*, 20(8):1394–1395, 2014. ISSN 1080-6040 1080-6059. doi: 10.3201/eid2008.140333. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4111178/.

[140] Rita MR Nogueira and Ana LF Eppinghaus. Dengue virus type 4 arrives in the state of rio de janeiro: a challenge for epidemiological surveillance and control. *Memórias do Instituto Oswaldo Cruz*, 106:255–256, 2011. ISSN 0074-0276. URL http://www.scielo.br/scielo.php?script=sci$_{a}$rttext\&pid=S0074-02762011000300001\&nrm=iso.

[141] Camila Zanluca, Vanessa Campos Andrade de Melo, Ana Luiza Pamplona Mosimann, Glauco Igor Viana dos Santos, Claudia Nunes Duarte dos Santos, and Kleber Luz. First report of autochthonous transmission of zika virus in brazil. *Memórias do Instituto Oswaldo Cruz*, 110(4):569–572, 2015. ISSN 0074-0276 1678-8060. doi: 10.1590/0074-02760150192. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4501423/.

[142] Marcio Roberto Teixeira Nunes, Nuno Rodrigues Faria, Janaina Mota de Vasconcelos, Nick Golding, Moritz U. G. Kraemer, Layanna Freitas de Oliveira, Raimunda do Socorro da Silva Azevedo, Daisy Elaine Andrade da Silva, Eliana Vieira Pinto da Silva, Sandro Patroca da Silva, Valéria Lima Carvalho, Giovanini Evelim Coelho, Ana Cecília Ribeiro Cruz, Sueli Guerreiro Rodrigues, Joao Lídio da Silva Gonçalves Vianez, Bruno Tardelli Diniz Nunes, Jedson Ferreira Cardoso, Robert B. Tesh, Simon I. Hay, Oliver G. Pybus, and Pedro Fernando da Costa Vasconcelos. Emergence and potential

for spread of chikungunya virus in brazil. *BMC Medicine*, 13(1):102, 2015. ISSN 1741-7015. doi: 10.1186/s12916-015-0348-x. URL https://doi.org/10.1186/s12916-015-0348-x.

[143] Kathleen M. O'Reilly, Rachel Lowe, W. John Edmunds, Philippe Mayaud, Adam Kucharski, Rosalind M. Eggo, Sebastian Funk, Deepit Bhatia, Kamran Khan, Moritz U. G. Kraemer, Annelies Wilder-Smith, Laura C. Rodrigues, Patricia Brasil, Eduardo Massad, Thomas Jaenisch, Simon Cauchemez, Oliver J. Brady, and Laith Yakob. Projecting the end of the zika virus epidemic in latin america: a modelling analysis. *BMC medicine*, 16(1):180–180, 2018. ISSN 1741-7015. doi: 10.1186/s12916-018-1158-8. URL https://www.ncbi.nlm.nih.gov/pubmed/30285863https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6169075/.

[144] Claudio Jose Struchiner, Joacim Rocklöv, Annelies Wilder-Smith, and Eduardo Massad. Increasing dengue incidence in singapore over the past 40 years: Population growth, climate and mobility. *PLOS ONE*, 10(8):e0136286, 2015. doi: 10.1371/journal.pone.0136286. URL https://doi.org/10.1371/journal.pone.0136286.

[145] Anna M. Stewart-Ibarra and Rachel Lowe. Climate and non-climate drivers of dengue epidemics in southern coastal ecuador. *The American journal of tropical medicine and hygiene*, 88(5):971–981, 2013. ISSN 1476-1645 0002-9637. doi: 10.4269/ajtmh.12-0478. URL https://www.ncbi.nlm.nih.gov/pubmed/23478584https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3752767/.

[146] M. F. Vincenti-Gonzalez, A. Tami, E. F. Lizarazo, and M. E. Grillet. Enso-driven climate variability promotes periodic major outbreaks of dengue in venezuela. *Scientific reports*, 8(1):5727–5727, 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-24003-z. URL https://www.ncbi.nlm.nih.gov/pubmed/29636483https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5893565/.

[147] Yien Ling Hii, Joacim Rocklöv, Nawi Ng, Choon Siang Tang, Fung Yin Pang, and Rainer Sauerborn. Climate variability and increase in intensity and magnitude of dengue incidence in singapore. *Global Health Action*, 2(1):2036, 2009. ISSN 1654-9716. doi: 10.3402/gha.v2i0.2036. URL https://doi.org/10.3402/gha.v2i0.2036.

[148] Bernard Cazelles, Mario Chavez, Anthony J. McMichael, and Simon Hales. Non-stationary influence of el niño on the synchronous dengue epidemics in thailand. *PLOS Medicine*, 2(4):e106, 2005. doi: 10.1371/journal.pmed.0020106. URL https://doi.org/10.1371/journal.pmed.0020106.

[149] Mallory Harris, Jamie M. Caldwell, and Erin A. Mordecai. Climate drives spatial variation in zika epidemics in latin america. *bioRxiv*, page 454454, 2019. doi: 10.1101/454454. URL https://www.biorxiv.org/content/biorxiv/early/2019/07/04/454454.full.pdf.

[150] Derek A. T. Cummings, Ira B. Schwartz, Lora Billings, Leah B. Shaw, and Donald S. Burke. Dynamic effects of antibody-dependent enhancement on the fitness of

viruses. *Proceedings of the National Academy of Sciences of the United States of America*, 102(42):15259, 2005. doi: 10.1073/pnas.0507320102. URL http://www.pnas.org/content/102/42/15259.abstract.

[151] N. Ferguson, R. Anderson, and S. Gupta. The effect of antibody-dependent enhancement on the transmission dynamics and persistence of multiple-strain pathogens. *Proceedings of the National Academy of Sciences of the United States of America*, 96(2):790–794, 1999. ISSN 0027-8424 1091-6490. doi: 10.1073/pnas.96.2.790. URL https://www.ncbi.nlm.nih.gov/pubmed/9892712https://www.ncbi.nlm.nih.gov/pmc/articles/PMC15215/.

[152] Matthew Graham, Amy K. Winter, Matthew Ferrari, Bryan Grenfell, William J. Moss, Andrew S. Azman, C. Jessica E. Metcalf, and Justin Lessler. Measles and the canonical path to elimination. *Science*, 364(6440):584–587, 2019. doi: 10.1126/science.aau6299. URL https://science.sciencemag.org/content/sci/364/6440/584.full.pdf.

[153] Bryan T. Grenfell, Ottar N. Bjørnstad, and Bärbel F. Finkenstädt. Dynamics of measles epidemics: Scaling noise, determinism, and predictability with the tsir model. *Ecological Monographs*, 72(2):185–202, 2002. ISSN 0012-9615. doi: 10.1890/0012-9615(2002)072[0185:DOMESN]2.0.CO;2. URL https://doi.org/10.1890/0012-9615(2002)072[0185:DOMESN]2.0.CO;2.

[154] Jennie S. Lavine, Aaron A. King, and Ottar N. Bjørnstad. Natural immune boosting in pertussis dynamics and the potential for long-term vaccine failure. *Proceedings of the National Academy of Sciences*, 108(17):7259–7264, 2011. doi: 10.1073/pnas.1014394108. URL https://www.pnas.org/content/pnas/108/17/7259.full.pdf.

[155] Natsuko Imai, Ilaria Dorigatti, Simon Cauchemez, and Neil M. Ferguson. Estimating dengue transmission intensity from case-notification data from multiple countries. *PLOS Neglected Tropical Diseases*, 10(7):e0004833, 2016. doi: 10.1371/journal.pntd.0004833. URL https://doi.org/10.1371/journal.pntd.0004833.

[156] Ronen Olinky, Amit Huppert, and Lewi Stone. Seasonal dynamics and thresholds governing recurrent epidemics. *Journal of Mathematical Biology*, 56(6):827–839, 2008. ISSN 1432-1416. doi: 10.1007/s00285-007-0140-4. URL https://doi.org/10.1007/s00285-007-0140-4.

[157] Olivia Brathwaite Dick, José L San Martín, Romeo H Montoya, Jorge del Diego, Betzana Zambrano, and Gustavo H Dayan. The history of dengue outbreaks in the americas. *The American journal of tropical medicine and hygiene*, 87(4):584–593, 2012. ISSN 0002-9637.

[158] Octavio Pinto Severo. Eradication of the aedes aegypti mosquito from the americas. *"Yellow fever, a symposium in commemoration of Carlos Juan Finlay, 1955."*, 1955.

[159] Ilana Löwy. Leaking containers: Success and failure in controlling the mosquito aedes aegypti in brazil. *American journal of public health*, 107(4):

517–524, 2017. ISSN 1541-0048 0090-0036. doi: 10.2105/AJPH.2017.303652. URL https://pubmed.ncbi.nlm.nih.gov/28207332https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5343710/.

[160] PAHO. The feasibility of erradicating aedes aegypti in the americas. *Pan Am J Public Health*, 1, 1997.

[161] Nathalia Caroline Santiago e Souza, Alvina Clara Félix, Anderson Vicente de Paula, José Eduardo Levi, Claudio Sérgio Pannuti, and Camila Malta Romano. Evaluation of serological cross-reactivity between yellow fever and other flaviviruses. *International Journal of Infectious Diseases*, 81:4–5, 2019. ISSN 1201-9712. doi: 10.1016/j.ijid.2019.01.023. URL https://doi.org/10.1016/j.ijid.2019.01.023.

[162] Jose Alberto Magno de Carvalho. Demographic dynamics in brazil: recent trends and perspectives. *Brazilian journal of population studies*, 1(1):5–23, 1997.

[163] Flavio Codeço Coelho and Luiz Max de Carvalho. Estimating the attack ratio of dengue epidemics under time-varying force of infection using aggregated notification data. *Scientific Reports*, 5:18455, 2015. doi: 10.1038/srep18455https://www.nature.com/articles/srep18455\#supplementary-information. URL https://doi.org/10.1038/srep18455.

[164] Erin A. Mordecai, Jeremy M. Cohen, Michelle V. Evans, Prithvi Gudapati, Leah R. Johnson, Catherine A. Lippi, Kerri Miazgowicz, Courtney C. Murdock, Jason R. Rohr, Sadie J. Ryan, Van Savage, Marta S. Shocket, Anna Stewart Ibarra, Matthew B. Thomas, and Daniel P. Weikel. Detecting the impact of temperature on transmission of zika, dengue, and chikungunya using mechanistic models. *PLOS Neglected Tropical Diseases*, 11(4):e0005568, 2017. doi: 10.1371/journal.pntd.0005568. URL https://doi.org/10.1371/journal.pntd.0005568.

[165] Marcelo Otero, Nicolás Schweigmann, and Hernán G. Solari. A stochastic spatial dynamical model for aedes aegypti. *Bulletin of Mathematical Biology*, 70(5):1297, 2008. ISSN 1522-9602. doi: 10.1007/s11538-008-9300-y. URL https://doi.org/10.1007/s11538-008-9300-y.

[166] Salihu Sabiu Musa, Shi Zhao, Hei-Shen Chan, Zhen Jin, and Daihai He. A mathematical model to study the 2014–2015 large-scale dengue epidemics in kaohsiung and tainan cities in taiwan, china. *Mathematical biosciences and engineering*, 2019. ISSN 1547-1063.

[167] Natsuko Imai, Ilaria Dorigatti, Simon Cauchemez, and Neil M. Ferguson. Estimating dengue transmission intensity from sero-prevalence surveys in multiple countries. *PLoS neglected tropical diseases*, 9(4):e0003719–e0003719, 2015. ISSN 1935-2735 1935-2727. doi: 10.1371/journal.pntd.0003719. URL https://www.ncbi.nlm.nih.gov/pubmed/25881272https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4400108/.

[168] New York Times. 127 die in floods and mud slides in brazil, 02/08/1988 Feb 08 1988. URL https://search.proquest.com/docview/426748587?accountid=14657http:

//sfx.lib.uchicago.edu/sfx_local?url_ver=Z39.88-2004\&rft_val_fmt=info:ofi/fmt:kev:mtx:journal\&genre=article\&sid=ProQ:ProQ\%3Anytimes\&atitle=127+Die+in+Floods+and+Mud+Slides+in+Brazil\&title=New+York+Times\&issn=03624331\&date=1988-02-08\&volume=\&issue=\&spage=A.4\&au=AP\&isbn=\&jtitle=New+York+Times\&btitle=\&rft_id=info:eric/\&rft_id=info:doi/.

[169] Sidney Bell, Leah Katzelnick, and Trevor Bedford. Dengue antigenic relationships predict evolutionary dynamics. *bioRxiv*, page 432054, 2019. doi: 10.1101/432054. URL http://biorxiv.org/content/early/2019/04/30/432054.abstract.

[170] Giorgio Guzzetta, Cecilia A. Marques-Toledo, Roberto Rosà, Mauro Teixeira, and Stefano Merler. Quantifying the spatial spread of dengue in a non-endemic brazilian metropolis via transmission chain reconstruction. *Nature Communications*, 9(1):2837, 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-05230-4. URL https://doi.org/10.1038/s41467-018-05230-4.

[171] World Health Organization. Dengue and severe dengue., 2017.

[172] World Health Organization. Dengue and severe dengue. Report, World Health Organization. Regional Office for the Eastern Mediterranean, 2014.

[173] Carles Bretó and Edward L Ionides. Compound markov counting processes and their applications to modeling infinitesimally over-dispersed systems. *Stochastic Processes and their Applications*, 121(11):2571–2591, 2011. ISSN 0304-4149.

[174] Daihai He, Edward L Ionides, and Aaron A King. Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. *Journal of the Royal Society Interface*, 7(43):271–283, 2009. ISSN 1742-5689.

[175] Edward L Ionides, Dao Nguyen, Yves Atchadé, Stilian Stoev, and Aaron A King. Inference for dynamic and latent variable models via iterated, perturbed bayes maps. *Proceedings of the National Academy of Sciences*, 112(3):719–724, 2015. ISSN 0027-8424.

[176] Wolfgang Bock and Yashika Jayathunga. Optimal control and basic reproduction numbers for a compartmental spatial multipatch dengue model. *Mathematical Methods in the Applied Sciences*, 41(9):3231–3245, 2018. ISSN 0170-4214.

[177] Mathieu Andraud, Niel Hens, Christiaan Marais, and Philippe Beutels. Dynamic epidemiological models for dengue transmission: a systematic review of structural approaches. *PloS one*, 7(11), 2012.

[178] Oliver J Brady, Michael A Johansson, Carlos A Guerra, Samir Bhatt, Nick Golding, David M Pigott, Hélène Delatte, Marta G Grech, Paul T Leisnham, and Rafael Maciel-de Freitas. Modelling adult aedes aegypti and aedes albopictus survival at different temperatures in laboratory and field settings. *Parasites & vectors*, 6(1):351, 2013. ISSN 1756-3305.

[179] Joseph Páez Chávez, Thomas Götz, Stefan Siegmund, and Karunia Putra Wijaya. An sir-dengue transmission model with seasonal effects and impulsive control. *Mathematical biosciences*, 289:29–39, 2017. ISSN 0025-5564.

[180] Tridip Sardar, Sourav Rana, and Joydev Chattopadhyay. A mathematical model of dengue transmission with memory. *Communications in Nonlinear Science and Numerical Simulation*, 22(1-3):511–525, 2015. ISSN 1007-5704.

[181] Janis Antonovics, Yoh Iwasa, and Michael P Hassell. A generalized model of parasitoid, venereal, and vector-based transmission processes. *The American Naturalist*, 145(5): 661–675, 1995.

[182] Mitzi Morris. Spatial models in stan: Intrinsic auto-regressive models for areal data. *Stan Case Studies*, 4, 2017.

[183] Michael P Busch, Ester C Sabino, Donald Brambilla, Maria Esther Lopes, Ligia Capuani, Dhuly Chowdhury, Christopher McClure, Jeffrey M Linnen, Harry Prince, Graham Simmons, et al. Duration of dengue viremia in blood donors and relationships between donor viremia, infection incidence and clinical case reports during a large epidemic. *The Journal of infectious diseases*, 214(1):49–54, 2016.

[184] Ministério da Saúde. Secretaria de Vigilância em Saúde. Dengue: monitoramento até a semana epidemiológica (se) 32 de 2014. *Boletim epidemiológico*, 45(19), 2014. URL `http://portalarquivos2.saude.gov.br/images/pdf/2014/setembro/01/Boletim-Dengue-SE32.pdf`.

[185] Joonha Park and Edward L Ionides. Inference on high-dimensional implicit dynamic models using a guided intermediate resampling filter. *Statistics and Computing*, 30(5): 1497–1522, 2020.

[186] Edward L. Ionides, Kidus Asfaw, Joonha Park, and Aaron A. King. Bagged filters for partially observed interacting systems. *Journal of the American Statistical Association*, 0(0):1–12, 2021. doi: 10.1080/01621459.2021.1974867. URL `https://doi.org/10.1080/01621459.2021.1974867`.

[187] Geir Evensen. *Data assimilation: the ensemble Kalman filter*. Springer Science & Business Media, 2009.

[188] C. Codeco, F. Coelho, O. Cruz, S. Oliveira, T. Castro, and L. Bastos. Infodengue: A nowcasting system for the surveillance of arboviruses in brazil. *Revue d'Épidémiologie et de Santé Publique*, 66:S386, 2018. ISSN 0398-7620. doi: https://doi.org/10.1016/j.respe.2018.05.408. URL `https://www.sciencedirect.com/science/article/pii/S0398762018311088`. European Congress of Epidemiology "Crises, epidemiological transitions and the role of epidemiologists".

[189] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter

Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: https://doi.org/10.1002/qj.3803. URL `https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803`.

[190] Center for International Earth Science Information Network CIESIN Columbia University. Gridded population of the world, version 4 (gpwv4): Basic characteristics, revision 11. palisades, ny: Nasa socioeconomic data and applications center (sedac)., 2016.

[191] Rafael Maciel-De-Freitas, Claudia Torres Codeco, and Ricardo Lourenco-De-Oliveira. Daily survival rates and dispersal of aedes aegypti females in rio de janeiro, brazil. *The American journal of tropical medicine and hygiene*, 76(4):659–665, 2007.

[192] Carles Bretó, Edward L Ionides, and Aaron A King. Panel data analysis via mechanistic models. *Journal of the American Statistical Association*, 2019.