Modelling COVID-19 in the North American region with a metapopulation network and Kalman filter

Matteo Perini[1*], Teresa K. Yamana[1], Marta Galanti[1], Jiyeon Suh[1], Roselyn Kaondera-Shava[1], Jeffrey Shaman[1,2]

[1]*Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, New York, United States of America*

[2]*Columbia Climate School, Columbia University, New York, New York, United States of America*

* *Corresponding author*

Email: matteo.perini@columbia.edu

Summary

Background

Metapopulation models provide platforms for understanding infectious disease dynamics and predicting clinical outcomes across interconnected populations, particularly for large epidemics and pandemics like COVID-19.

Methods

We developed a novel metapopulation model for simulating respiratory virus transmission in the North America region, specifically for the 96 states, provinces, and territories of Canada, Mexico and the United States. The model is informed by COVID-19 case data, which are assimilated using the Ensemble Adjustment Kalman filter (EAKF), a Bayesian inference algorithm, and commuting and mobility data, which are used to build and adjust the network and movement across locations on a daily basis.

Findings

This model-inference system provides estimates of transmission dynamics, infection rates, and ascertainment rates for each of the 96 locations from January 2020 to March 2021. The results highlight differences in disease dynamics and ascertainment among the three countries.

Interpretation

The metapopulation structure enables rapid simulation at large scale, and the data assimilation method makes the system responsive to changes in system dynamics. This model can serve as a versatile platform for modeling other infectious diseases across the North American region.

Funding

US Centers for Disease Control and Prevention Contract 75D30122C14289; US NIH Grant AI163023.

Introduction

Mathematical models have been used to simulate infectious diseases outcomes, infer transmission dynamics, and predict future disease burden. These tools can inform public health strategies by testing control methods and identifying effective interventions[1]. During the SARS-CoV-2 pandemic, unprecedented data availability enabled application of mathematical models in many locations and at different geographical scales worldwide[2–5]. Metapopulation modeling approaches provide an efficient framework for simulating and evaluating the spatiotemporal progression of infectious disease over large geographic areas. In this model form, populations are aggregated within typically fixed geographic units (e.g. provinces, cities), which allows resolution of spatial disease patterns without the computational expense and micro-behavioral assumptions required for agent-based models. At large geographical scales, agent-based models require high-performance computing (HPC) clusters. For example, Bhattacharya et al. developed a platform to enable the real-time execution of an agent-based COVID-19 model for the United States on more than ten thousand CPU cores[6].

Many metapopulation models have been developed to describe the dynamics of infectious diseases at different geographic scales[7,8]; however, a key challenge in model development lies in accurately determining the movement patterns of individuals among subpopulations. Some metapopulation models use fixed or arbitrary sized geographical areas (cells) as subpopulations, which can then be aggregated to match the resolution of available case, census and movement data[9,10]. One example of a multi-national system is the GLEAM (Global Epidemic And Mobility) platform[10–12], which estimates the flux of individuals among arbitrary subpopulations centered around major transportation hubs (usually airports) and uses commuting and air travel data[13].

In 2020, the US reported the highest number of COVID-19 cases and deaths globally, with the first case identified in Washington state on January 20th and three pandemic waves manifesting during the year[14]. Canada reported its first case in Toronto on January 25th and experienced a decline in cases during the summer followed by a resurgence in the fall. Mexico had a similar epidemiological history, but the first wave developed later during the summer[15]. The literature currently lacks a comprehensive COVID-19 model for the North American region at continental scale. Here, we present a metapopulation susceptible-exposed-infectious-recovered (SEIR) for COVID-19 for the majority of the North American region (i.e. Canada, United States and Mexico) from the beginning of the pandemic to widespread availability and distribution COVID-19 vaccines. The model is coupled with a data assimilation algorithm and is informed by daily COVID-19 cases and a commuting network that is adjusted by daily mobility trends. Unlike the GLEAM platform, in this model the flux of individuals among subpopulations of the metapopulation model is calculated using the daily work commuting patterns coupled with random movements among the geographical locations, as in previous work[4].

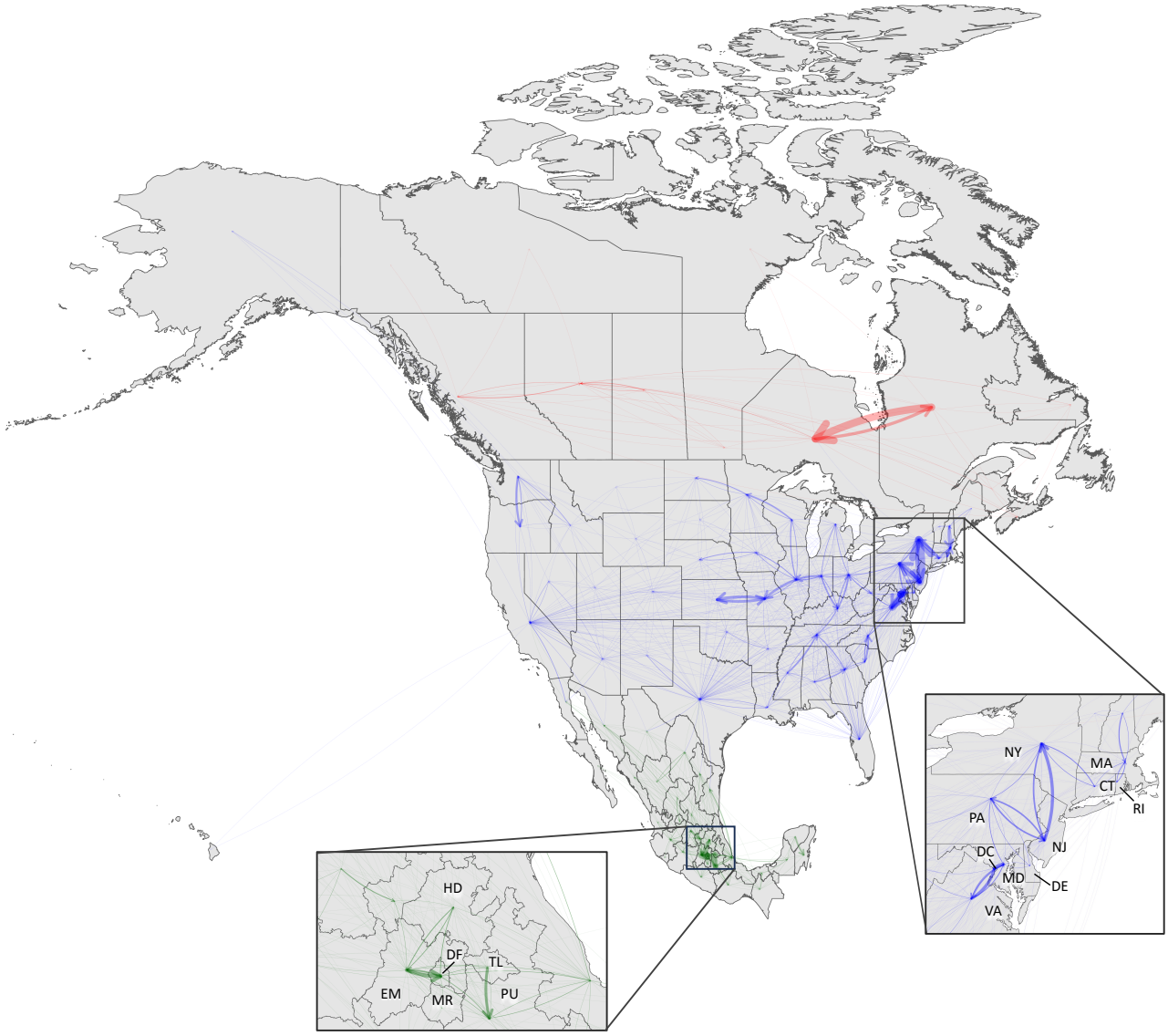
The model developed here provides temporal estimates of transmission and ascertainment rates for COVID-19 for the North American Region, filling a gap in the existing literature regarding models for this region. The flux patterns and the continental-level metapopulation structure used in this study

helps investigation of disease dynamics of COVID-19 at a larger scale, for the majority of the North American region, and can reveal dynamics that are discernible only at such a broad scale. Lastly, the work commuting matrix structure developed for this study can be paired to different compartmental models and disease surveillance data to study the dynamics of other infectious diseases such as influenza.

Methods

We developed a metapopulation susceptible-exposed-infectious-recovered (SEIR) model for the North American region. Specifically, the model represents the 96 first-level administrative divisions of Canada (10 provinces and 3 territories), United States (50 states and 1 federal district) and Mexico (31 states and 1 autonomous city) represented in **Figure 1**, with a total population of 483 million, based on census data[16–18]. Mixing is simulated as two types of movement: daily commuting and

88 random movement (i.e. all the daily movement across divisions due to reasons other than work commuting).



90 ***Figure 1 Commuting to work matrix in most of the North American region.*** *The map shows the 10 provinces and 3 territories of Canada, the 50 states and 1 federal district of United States and the 31 states*
92 *and 1 autonomous city of Mexico that have been used in the model. The arrows represent the flux of individuals commuting daily to work to another location. Arrow size represents the number of commuters,*
94 *and color represents the country of origin: red for Canada, blue for United States, green for Mexico.*

Daily commuting matrix

96 Daily commuting among locations were retrieved and derived from four national datasets: i) Canadian 2016 census (Statistics Canada) Commuting Flow from Geography of Residence to Geography of
98 Work[19]; ii) Canada Frontier Counts (Statistics Canada): Number of vehicles travelling between Canada

and the United States[20]; iii) 2011-2015 5-Year American Community Survey (ACS) Commuting Flows (United States Census Bureau)[21]; and iv) Mexican Intercensal Survey 2015 (National Institute of Statistics and Geography, INEGI)[18]. Information from these datasets were processed and combined according to the methods described in **Supplementary Note 1** to obtain the commuting work matrix represented in **Figure 1**. The matrix contains the number of people that commute daily to work in another location (states for Mexico and US, provinces or territories for Canada). To account for control measures and closures enacted during the estimation period, the commuting work matrix was scaled based on daily mobility activity, derived from the "change in workplace visitors" trends from Google Community Mobility Reports[22] (see **Supplementary Figure 1**). To account for movement between locations for purposes other than work commuting, the model includes daily random movement among locations; this movement was set to be proportional to the average number of working commuters among each location pair (see **Table 1**).

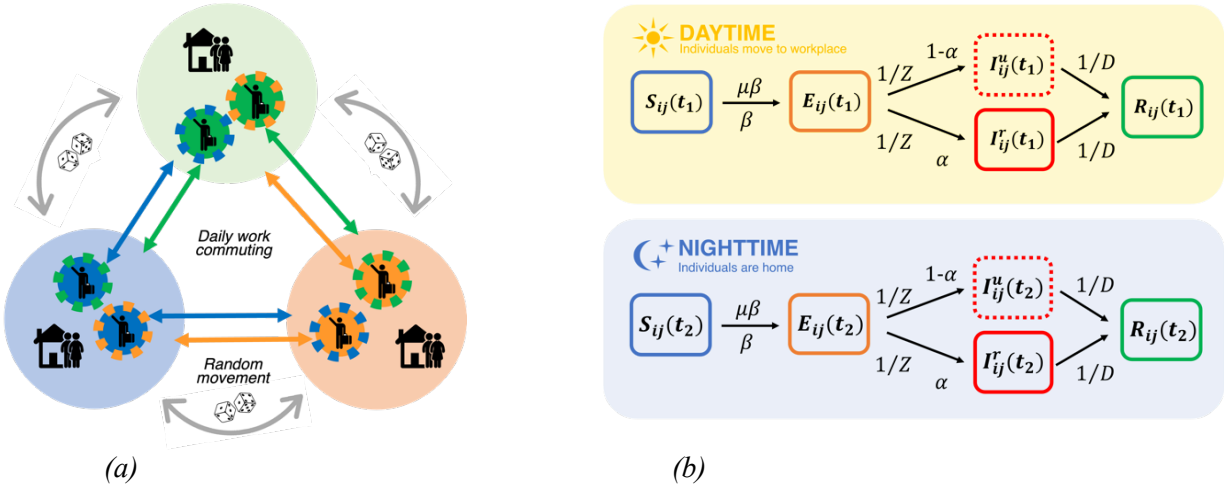
Transmission model

The metapopulation model resolves daytime and nighttime mixing differences, depicting diurnal changes in contact among subpopulations. Transmission occurs as a discrete Markov process during both day and nighttime, following the structure of previous studies in the US[4]. The transmission dynamics are described by the equations in **Supplementary Note 2** (**eqs. S4-S13**). In these equations, S_{ij}, E_{ij}, I^r_{ij}, I^u_{ij}, and N_{ij} represent the susceptible, exposed, reported infectious, unreported infectious, and total population in the subpopulation commuting from location j to location i ($i \leftarrow j$). Additionally, we assume that no individuals enter or leave the model, and that there is no loss of immunity following primary infection, given the relatively short simulation time period. For the same reason, we can compute the R_{ij} (Removed population) as $R_{ij} = \sum_1^t I^r_{ij} + \sum_1^t I^u_{ij}$ which is the cumulative sum of daily new reported and unreported cases, or the sum of the individuals that exited the two infectious compartments.

The parameters of the model are: β, the transmission rate of reported infections; μ, the relative transmissibility of unreported infections; Z, the average latency period (from infection to contagiousness); D, as the average duration of contagiousness; α, the fraction of documented infections (ascertainment rate); and θ, a multiplicative factor adjusting random movement. A distinct transmission rate, $\mu\beta$, is defined for undocumented infections: we assume that these individuals show little to no symptoms during infection and are less contagious than documented infections[5]. Each equation is integrated using a Poisson process to capture the stochastic nature of transmission dynamics. In total, the model consists of 3,268 metapopulations. To reduce the dimension of the system being estimated and to improve identifiability, we fixed the values of the parameters related to disease progression (Z, D, and μ) and the multiplicative factor for random movement (θ) based on previously published findings[4,14]. Local transmission rates β^l_t and ascertainment rates α^l_t are estimated for each of the 96 locations, l at time t. The hyperparameter R^l_t, the time-varying reproductive number, was derived using the next-generation matrix approach[4,23]:

(1) $R^l_t = \beta^l_t D[\alpha^l_t + (1 - \alpha^l_t)\mu]$

R_t^l	Time-varying reproductive number for location l at time t
D	Duration of contagiousness
β_t^l	Transmission rate of location l at time t
α_t^l	ascertainment rate of location l at time t
μ	relative transmissibility of unreported infections



(a) (b)

Figure 2. Metapopulation structure and compartmental model: (a) *Daily work commuting - during the daytime, some individuals commute from their home to their workplace in another location and mix with the populations present there. During the nighttime, those commuters return home and mix with other residents who live in the same location. Random movement - individuals may travel among locations for reasons other than work. These random visitors circulate among subpopulations following a Markov process, causing a population exchange in all locations* ***(b)*** I_{ij}^r *reported infected and* I_{ij}^u *unreported infected from location from location j to location i (i ←j);* t_1 *daytime duration;* t_2 *nighttime duration;* β *transmission rate;* α *ascertainment rate;* μ *relative transmissibility of unreported cases;* Z *latency period;* D *duration of contagiousness.*

COVID-19 cases and data assimilation

The model is informed by daily COVID-19 confirmed case data retrieved from COVID-19 Open Data — Google Health[24] starting from January 20th 2020 to March 31st 2021 for the 96 first-level administrative divisions of Canada, United States and Mexico. This time frame was chosen to capture the first three COVID-19 waves[14], before the emergence of the Delta and subsequent variants and more widespread uptake of COVID-19 vaccines. A 7-day moving average was applied to the daily cases data to smooth out daily fluctuations in reporting. The smoothing also mitigates the impact of reporting delays or inconsistencies in case data, providing a more reliable indicator of overall disease dynamics.

The estimation of state variables and parameters is carried out by data assimilation implemented using the Ensemble Adjustment Kalman Filter (EAKF) algorithm[25], as in previous studies[26,27]. Kalman filters use Bayes' rule to update state variables and parameters. Normality is assumed for the prior distribution and the likelihood so that the posterior distribution can be characterized by the mean and

the covariance. In this model, an ensemble of 300 simulations is integrated to generate a prior distribution of parameters and state variables, including estimation of the observed state variable. At each observation time point, the model is halted, and the ensemble and observation are used to calculate the Kalman gain, which is used to update the observed state variable. The unobserved state variables and parameters are then updated in proportion to the same Kalman gain. Finally, the posterior estimates are used as priors and model integration through time continues to the next observation. The formulas used to calculate the Kalman gain for the observed and unobserved variables and parameters are presented in the **Supplementary Note 3** (**eqs, S14**-**S15**). To properly balance the influence of the observational data in the assimilation process, it is crucial to estimate its error, which is typically unknown. Here, we estimated the Observational Error Variance of the observational data as shown in **eq. (2)**, similarly to prior works[5,28,29]:

(2) $OEV_t^l = \max\left(5, \frac{(O_t^l)^2}{100}\right)$

OEV_t^l Observational Error Variance of location l at time t

O_t^l average cases in location l in the week before time t

Repeated filter adjustments tend to decrease model ensemble variance reducing the impact of new observations in subsequent estimates. This may lead to divergence, in which the filter ceases adjusting the model state[25]. To avoid divergence, at each timestep we applied a multiplicative factor (1.01) to inflate the prior ensemble of the observed variable (daily reported infected I^r) and the estimated parameters (ascertainment rate α and transmission rate β). Additionally, we reinitialized the values of the estimated parameters for a fraction (2%) of the ensemble members every 7 days. This reinitialization enables the system to periodically readjust estimations whenever the ensemble variance begins to shrink, thus preventing divergence.

Further details on the SEIR-EAKF assimilation process are available in the **Supplementary Note 3**.

Model initialization

Each ensemble member was initialized with a set of parameter and state variable estimates to resemble the epidemiological conditions at the beginning of the SARS-CoV-2 pandemic in 2020. The parameter values related to disease progression (Z, D and μ) were drawn from distributions with ranges reported in **Table 1**, in accordance with Li et al[5]. These parameters are assumed to remain fixed over time and to have consistent values across all locations, representing inherent biological characteristics of the disease. The random movement factor θ is also constant over time, with a random value drawn from a uniform distribution between 0 and 0.2 assigned to each location (see **Table 1**). This value represents the relative volume of random movement compared to commuting, where $\theta = 0.15$ indicates that the number of random visitors is 15% of the average number of commuters between two locations. Exposed (E), reported infectious (I^r) and unreported infectious (I^u) individuals were initialized with random draws from a uniform distribution that ranged from 1 to 9 in each subpopulation.

The three countries exhibited substantive differences in healthcare and testing capacity during the pandemic. Mexico had particularly low testing rates[30,31], suggesting that only severe cases were assayed due to limited availability of test kits. During 2020, the United States experienced the highest number of cases, as well as relatively high nation-wide ascertainment rates, as shown in prior inference studies[4,14]. For Canada, the estimate of national testing rates or the ascertainment rate is lacking; however, Ontario, the most populous province in Canada, accounting for ~37% of the population, initially faced challenges in ramping up its testing capacity, but its centralized resource strategy was able to increase test capacity during 2020[32]. Given these differences among the three countries, we assigned country-specific initial ranges for the ascertainment rate α. For the US we used SARS-CoV-2 seroprevalence estimates and cumulative reported cases to estimate the value of the ascertainment rate. Specifically we used infection-induced seroprevalence from blood samples collected in US during July 2020 (3.5%)[33] to derive an initial estimate of α in the US, this formula is shown in **eq. (3)**:

$$\textbf{(3)}\quad \alpha_{US}^{0} = median\left(\frac{\sum cases_j}{IIS_{US} \times pop_j}\right)$$

α_{US}^{0} Initial ascertainment rate for the US

$\sum cases_j$ Cumulative reported cases of state *j* up to July 2020

IIS_{US} Infection-induced seroprevalence in the United States during July 2020 (3.5%)[33]

pop_j Population of state *j*

Based on this formula, we set the initial prior distribution mean to 0.25 for α in US.

Infection-induced seroprevalence was not available for Canada and Mexico during the time frame of interest. Instead, considering the very low testing rate for Mexico[30,31] we set the initial prior distribution mean to 0.08, corresponding to a ~70% decrease from the US. Canada faced early issues establishing testing facilities[32], so we set the initial prior distribution mean to 0.12 for α, between the values of US and Mexico (~50% decrease compared to the US).

The initial values of β were drawn for each location from the normal distribution with mean $\mu = 1.93$ and standard deviation $\sigma = 0.75$ similarly to Li et al.[5](see **Table 1**).

The SEIR-EAKF model-inference system can potentially estimate parameters and state variables values that violate physicality. This could happen, for example, if a state variable or estimated parameter is adjusted by the filter to a value below zero. To avoid this issue, we applied constraints to both the state variables and parameters, subsequently re-assigning values that fell outside specified bounds. Specifically, ensemble members associated with any state variable that exhibited values less than or equal to zero were assigned values from the previous day. The Susceptible population S_{ij} is not directly adjusted by the EAKF. Instead, it is computed using the total population count and the EAKF-adjusted state variables as $S_{ij} = N - E_{ij} - R_{ij}$. This guarantees population mass balance and is feasible under the assumption of non-reinfection among individuals. Additionally, it prevents the

variable S_{ij} from exceeding the total population, which aligns with the assumption of a constant population.

The lower bound of α was set to 0.025 corresponding to 5 reported cases every 200 infections. For Canada and US, this bound increased linearly by 0.5% at each day, or: $\alpha_{t+1}^{min} = \alpha_t^{min} * 1.005$. This increase was imposed to reflect the efforts of the two countries to increase the detection capabilities of local and national health systems[32,34–36]. Conversely, Mexico implemented a sentinel surveillance system[37] in which only the hospitalized cases and 10% of the mild cases were tested[38]. Focusing testing efforts on symptomatic individuals is a cost-effective strategy, but it also greatly increases the proportion of unreported cases, resulting in lower ascertainment rates. For this reason, the lower α bound of Mexico did not increase over time in our model. The upper bound for α was fixed for all locations at 0.6 (i.e. 60 cases per 100 infections). We allowed the transmission rate β to span a broad range of values, with the lower bound set to 0.2 and the upper bound set to 4. The descriptions, prior ranges and bounds for all the model parameters are shown in **Table 1**.

	Description	Initial prior range distribution	bounds	Parameter type
Z	Latency period	$U(2,5)\ days$	Fixed	Fixed
D	Duration of contagiousness	$U(2,5)\ days$		
μ	Relative transmissibility of unreported infections	$U(0.2,0.45)$		
θ	Random movement factor	$U(0.0,0.2)$		
β	Transmission rate	$N(1.93,0.75)$	$[0.2,4]$	Estimated
α	Ascertainment rate - Canada	$U(0.08,0.17)$	$[0.025,0.6]$*	
	Ascertainment rate – United States	$U(0.13,0.37)$		
	Ascertainment rate - Mexico	$U(0.04,0.12)$		

* Mexico lower bound is fixed to 0.025 over time; Canada and US lower bound increases by 0.5% daily: $\alpha_{t+1}^{min} = \alpha_t^{min} * 1.005$

***Table 1**: parameters description, initial prior distribution, ranges, and parameter type*

System identifiability

To assess the identifiability of the model, we first tested our framework using synthetically generated datasets. Specifically, we generated a suite of synthetic outbreaks using the model (**eqs. S4-S13**) in free simulation, each with arbitrarily assigned values of the epidemiological parameters and initial conditions for state variables. We then ran the full model-inference system 100 times assimilating the daily new reported cases time series generated by each of these free simulations to test the system's ability to accurately estimate state variable and parameter values. As in practice with actual data, we fixed the parameters related to disease progression (Z, D, and μ) and the multiplicative factor for random movement (θ), while the ascertainment rate α and transmission rate β were estimated for each location. The initial prior distribution and range of the parameters were as reported in **Table 1.**

Results

Daily commuting matrix

The daily commuting matrix depicted in **Figure 1** was obtained by combining and processing the national datasets reported in the Method section (full description in **Supplementary Note 1**); it is summarized in **Table 2**. Out of a combined population of 483 million, the Canadian population accounts for the 7.3% (35.1 million), the US for the 67.2% (324.3 million) and Mexico for the 25.6% (123.3 million). Around 8.9 million people, or 1.84% of the represented North American population, commutes daily to another state/province/territory or country to work. Inter-country commuting accounts for just the 1.12% (99,369) of the 8.9 million commuters. Finally, the percentage of internal commuters for each country is similar to the population percentage relative to the total population of the three countries.

Residence	Workplace	Commuters	% over Total Commuters
Canada	Canada	527,300	5.93
Canada	USA	1,962	0.02
USA	Canada	4,605	0.05
USA	USA	5,334,072	60.02
USA	Mexico	12,892	0.15
Mexico	USA	79,910	0.90
Mexico	Mexico	2,925,820	32.92
		Total Commuters	**% over total Population**
		8,890,144	1.84

Table 2: *Number of commuters by country of residence and workplace. Individuals with residence and workplace in the same country are commuting between states/provinces/territories within that country. The last column shows the percentage over the total commuters. The last row shows the total number of cross-border (state, province, territory, or country) commuters and its percentage over the total population of the model.*

System Identifiability

To verify the convergence of the estimated parameters (ascertainment rate α and transmission rate β) to the synthetic truth values created in free simulation, we plotted the model-generated data points over the boxplot distributions of the 3000 estimated parameter values (300 ensemble members x 100 ensemble simulations) at the end of each outbreak among the locations (see **Supplementary Figure 2**). Most of the true values fall in the interquartile of the distribution of the estimated values (68% for α and 84% for β) while all the true values falls in the 95% CI of the distribution. This demonstrates the ability of the system to estimate local time-varying α and β values.

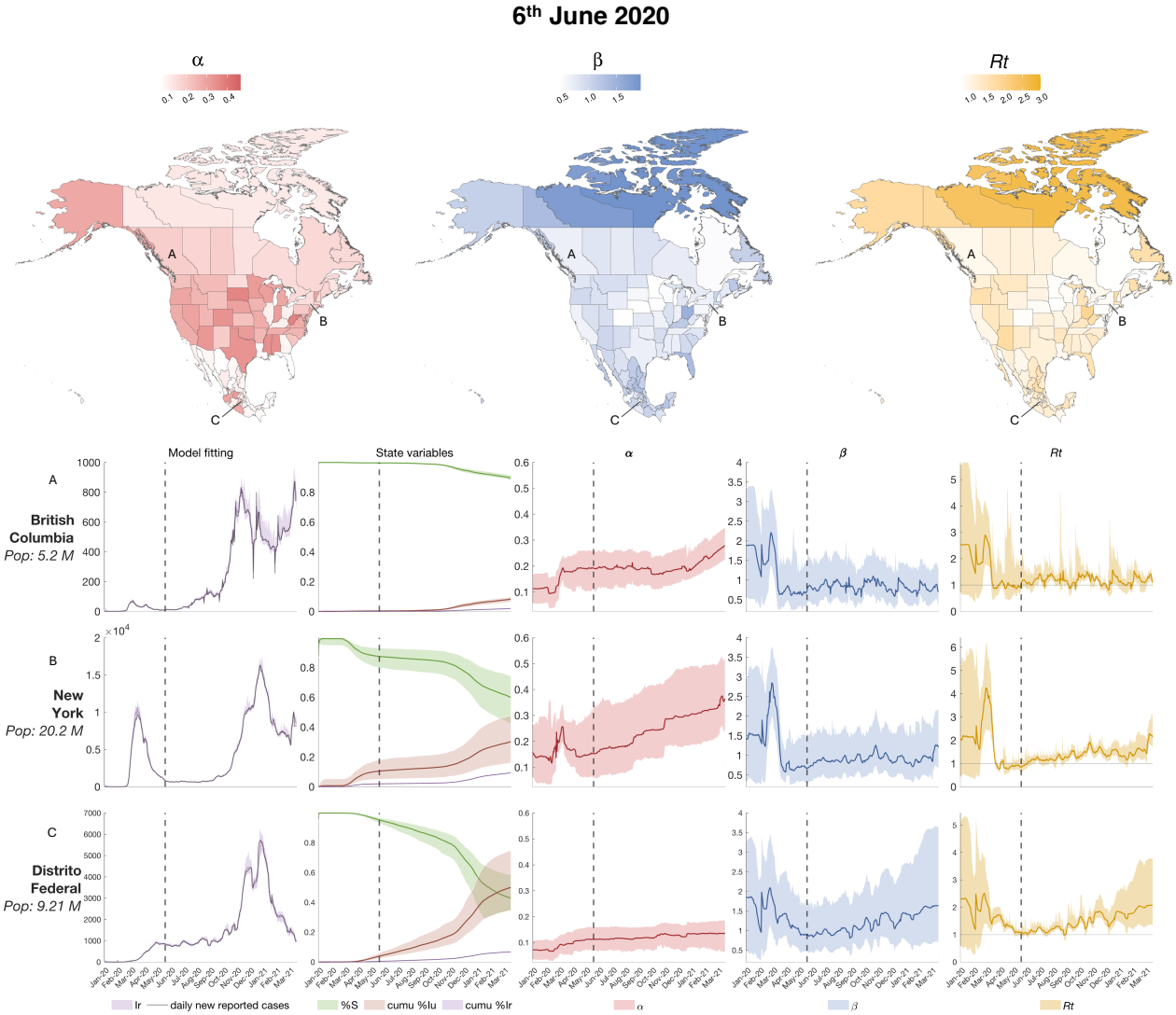
Simulation with Case Data

The model-inference system was run with the real case data of Canada, United States and Mexico starting from January 20th 2020 to March 31st 2021. During this time frame, the three countries

experienced asynchronous outbreaks as depicted in **Supplementary Figure 1.** To explore disease dynamics over time and location, we selected three timepoints when cases were declining in most of the locations of the North American region, roughly corresponding to the end of the three pandemic waves experienced in United States during 2020[14]. The first wave began in January 2020 and lasted through the spring; we selected June 6, 2020 as the first timepoint. **Figure 3** shows the estimated values of the parameters (ascertainment rate α and transmission rate β) and hyperparameters (basic reproductive number R_t) in all 96 locations of the study region at this time point. In addition, the lower panels of **Figure 3-5** show the time progression of model fitting (modeled daily new reported cases and observed daily cases), select state variables (S, cumulative I^r,cumulative I^u) and parameter and hyperparameter estimates (α, β, R_t) for a total of nine selected locations. These locations were selected among the 96 to represent the epidemiological progression in different geographical areas of the North American region, focusing on some of the most populous and epidemiologically relevant locations of the three countries. In **Figure 3** the three selected locations indicated in the maps are British Columbia (Canada), New York (United States) and Distrito Federal (Mexico City, Mexico). The sub-national parameter values for each of the three countries were aggregated to the national level using a population-weighted average. On June 6, 2020, the population-weighted average ascertainment rate α of the mean estimates in US states (0.26) was 1.5 times higher than in Canada (0.17) and 2.7 times higher than in Mexico (0.10). The population-weighted average transmission rate β of the mean estimates for Mexican states was 1.03, which is 1.4 times higher than in Canada (0.72) and the US (0.72). The population-weighted average of the time-varying reproductive number R_t of the mean estimates was 0.99 for Canadian provinces and territories, 1.11 for the United States, and 1.22 for Mexican states. These values are also reported in **Table 3**.

Location	Jun. 6th 2020	Sep. 7th 2020	Mar. 15th 2021
	Ascertainment rate α		
Canada	0.17	0.18	0.26
United States	0.26	0.31	0.36
Mexico	0.10	0.11	0.13
	Transmission rate β		
Canada	0.72	0.95	0.84
United States	0.72	0.64	0.85
Mexico	1.03	0.86	0.85
	Time-varying basic reproduction number R_t		
Canada	0.99	1.32	1.31
United States	1.11	1.04	1.49
Mexico	1.22	1.04	1.07

***Table 3**: Population-weighted average of the mean estimates values of the parameters (ascertainment rate α and transmission rate β) and hyperparameters (time-varying basic reproductive number R_t) for the three countries at three selected timepoints. The intensity of cell colors in the table corresponds to their values: higher values are represented by more intense colors.*



334 ***Figure 3. Estimated state variables and parameters on June 6, 2020:*** *the three maps on the top panel show*
the value of the of the parameters (ascertainment rate α and transmission rate β) and the hyperparameter
336 *(basic reproductive number R_t) for all 96 locations on June 6, 2020. The bottom panels show the model*
fitting (i.e. the estimated observed variable over the 7-day smoothed daily new reported cases), three state
338 *variables illustrating disease progression (the susceptible population, cumulative reported infectious and*
cumulative unreported infectious), and the parameters α, β and R_t for 3 selected locations: British
340 *Columbia (Canada), New York (US) and Distrito Federal (Mexico City, Mexico). The color shaded areas*
represent the 95% credible interval from the 300-member ensemble. The dotted vertical lines indicate the
342 *timepoint of reference for the maps.*

344 During summer 2020, some locations, particularly in the US, experienced a second wave consisting of
a resurgence of cases, with a decline at the beginning of fall. **Figure 4** shows the values of the state
346 variables and parameters on September 7, 2020. The maps show the parameter estimates across all

locations, while the temporal dynamics of state variable and parameter estimates are shown for three of the nine selected locations: Ontario (Canada), Florida (US) and Estado de México (Mexico). On September 7, 2020, the population-weighted average ascertainment rate α of the mean estimates in the United States was 0.31, which is 1.74 times higher than in Canada (0.17) and 2.8 times higher than in Mexico (0.11). The population-weighted average transmission rate β of the mean estimates in Canadian provinces and territories was 0.95, which is 1.5 times higher than in the US (0.64) and 1.1 times higher than in Mexican (0.86). The population-weighted average of the time-varying reproductive number R_t of the mean estimates was 1.32 for Canadian provinces and territories, 1.26 times higher than in the United States (1.04) and Mexican states (1.04). These values are also reported in **Table 3**.

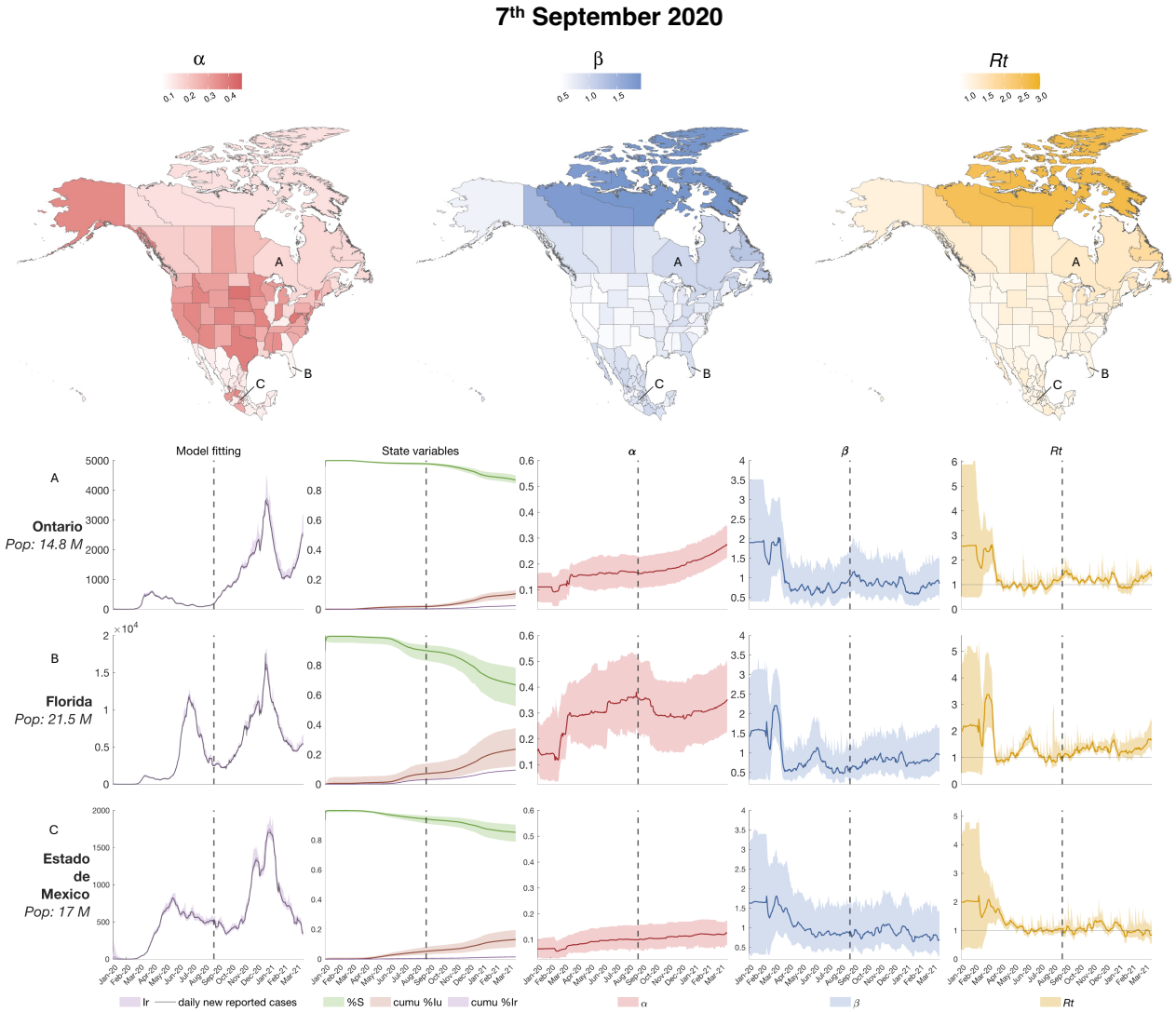


Figure 4. Estimated state variables and parameters on September 7, 2020: *the three maps on the top panel show the value of the of the parameters (ascertainment rate α and transmission rate β) and the hyperparameter (basic reproductive number R_t) for all 96 locations on September 7, 2020. The bottom*

panels show the model fitting (i.e. the estimated observed variable over the 7-day smoothed daily new reported cases), three state variables illustrating disease progression (the susceptible population, cumulative reported infectious and cumulative unreported infectious), and the parameters α, β and R_t for 3 selected locations: Ontario (Canada), Florida (US) and Estado de México (Mexico). The color shaded areas represent the 95% credible interval from the 300-member ensemble. The dotted vertical lines indicate the timepoint of reference for the maps.

Most locations experienced a more severe outbreak during the autumn-winter wave of 2020/2021, before the widespread availability of the COVID-19 vaccines. We plot the estimated parameters and hyperparameters on March 15, 2021 in **Figure 5**. The state variable and parameter estimate time series are also shown for three of the nine selected locations: Quebec (Canada), California (United States), and Jalisco (Mexico). On March 15, 2021, the population-weighted average ascertainment rate α of the mean estimates in the United States was 0.36, 1.4 times higher than Canada (0.26) and 2.7 times more than Mexico (0.13). The population-weighted average transmission rate β of the mean estimates was 0.83 in Canada and 0.85 in US and Mexico. The population-weighted average of the time-varying reproductive number R_t was 1.49 for the United States, 1.14 times higher than Canadian provinces and territories (1.31) and 1.39 times higher than Mexican states (1.04). These values are also reported in **Table 3**. All the estimated mean values of the parameters (ascertainment rate α and transmission rate β) and hyperparameters (basic reproductive number R_t) for all the location on these three timepoints are reported in **Supplementary Table 3.**

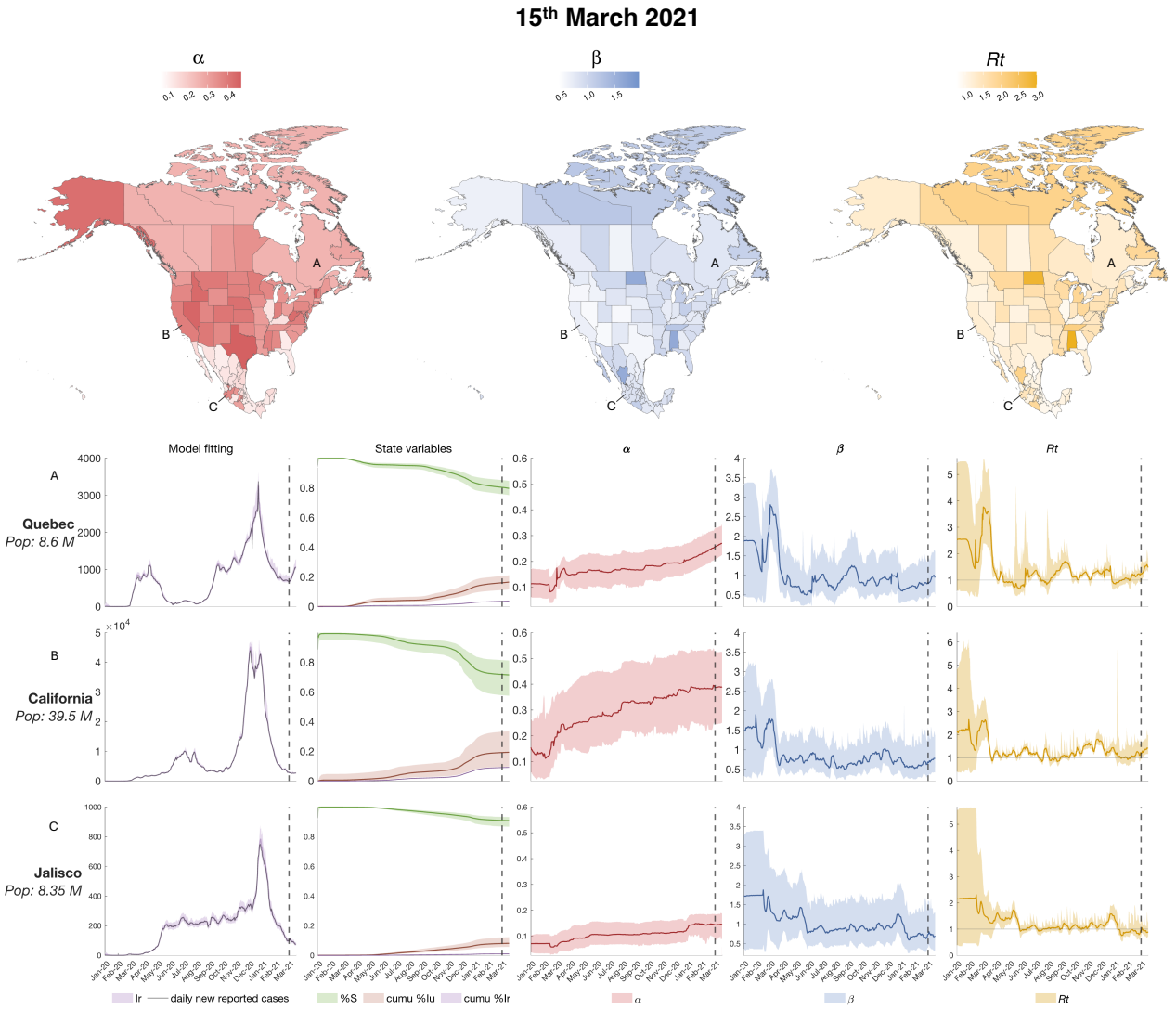


Figure 5. Estimated state variables and parameters on March 15, 2021: *the three maps on the top panel show the value of the of the parameters (ascertainment rate α and transmission rate β) and the hyperparameter (basic reproductive number R_t) for all 96 locations on March 15, 2021. The bottom panels show the model fitting (i.e. the estimated observed variable over the 7-day smoothed daily new reported cases), three state variables illustrating disease progression (the susceptible population, cumulative reported infectious and cumulative unreported infectious), and the parameters α, β and R_t for 3 selected locations: Quebec (Canada), California (United States), and Jalisco (Mexico). The color shaded areas represent the 95% credible interval from the 300-member ensemble. The dotted vertical lines indicate the timepoint of reference for the maps.*

Discussion

In this study we developed a SEIR metapopulation model structure and combined it with a data assimilation algorithm (EAKF) to reproduce COVID-19 outbreak dynamics and estimate important

epidemiological parameters across most of the North American region during the first three pandemic waves[14]. The model has demonstrated identifiability in estimating system state variables, as well as the ascertainment rates, α, and transmission rates, β, for the 96 first-level administrative divisions of Canada, United States, and Mexico. The metapopulation structure provided computational efficiency in comparison to agent based models that require HPC (high-performance computing) and cluster computing approaches to run at large scale[6]. The transmission module of the model structure relies on the daily work commuting patterns across states, provinces, and territories for disease spreading, with the capability to adjust the daily commuting matrix using Google Mobility Report trends data[22] to account for travel restrictions implemented during the pandemic.

As depicted for selected locations in **Figure 3-5** and reported for all the locations in **Supplementary Table 3**, the SEIR-EAKF system estimated large disparities in the ascertainment rates α among the three countries[4,14,30–32]. Moreover, it estimated a gradual increase in the ascertainment rates α throughout the three major COVID-19 outbreaks during 2020 and at the beginning of 2021 in each country. Specifically, the US showed substantially higher α values in almost all states across the three waves: at the last time point (March 15, 2021) the majority of states had reached an ascertainment rate of ~40%. Canada and Mexico showed smaller ascertainment rates than the US at the end of the estimation: most of the Canadian provinces and territories had an ascertainment rate of ~26%, while Mexican states had less than 20%. Note that in Canada and the United States, the increase in α is partially influenced by the increasing lower imposed during inference, whereas Mexico has a fixed α lower bound over time.

Transmission rates β remained generally low during the estimation period for the majority of locations. In the United States, the estimated β values in the majority of locations appeared to exhibit minimal variation throughout the estimation period, while in Mexico transmission rates generally decreased gradually over time. A notable exception was Distrito Federal (Mexico City), the very densely populated capital of Mexico. The relatively high β value at the end of the estimation in March 2021 (1.6) was not unexpected in this location. High population density allows for more contact opportunities, increasing the rate of transmission and the basic reproduction number as show in previous analysis[4]. A few additional exceptions to the national trends are discussed in the **Supplementary Note 4.**

The hyperparameter R_t (time-varying basic reproductive number) is proportional to β and α, as reported in **eq. (1)**. The majority of the mean estimated R_t values were above the epidemic threshold of $R_t = 1$ at all the three selected time points and all locations had mean $R_t > 1$ at least once among the three selected time point. Moreover, the lower estimated mean values for R_t never dropped below 0.75, or below 0.86 at the last time point. These values suggest a sustained epidemic in the North American region.

Overall, the three countries demonstrated distinct and quite isolated epidemiological histories. This is supported by the results presented in **Figures 3-5** and **Supplementary Table 3** where the trends estimated for the parameter and hyperparameter values can be clustered by country, highlighting the national and interconnected epidemiological developments they have undergone. This phenomenon arises from the commuting network structure depicted in **Figure 1**, where few arrows traverse national borders, illustrating three major networks (the countries) with limited interconnection.

The model developed for this work uses a new commuting matrix, based on national census data and surveys, to model worker flows in the American region (Canada, US, and Mexico). It assumes most movements are daily work commutes and accounts for additional movements with a random factor proportional to the commuting flux. Moreover, the system couples the compartmental model (SEIR) with the Ensemble Adjustment Kalman Filter (EAKF), adjusting the system state variables and parameters daily on the basis of the case data. This approach differs from other models used to estimate infectious diseases on a multi-national and/or continental scale, such as the GLEAM platfom[10–12]. Those platforms estimate the movement of individuals among various arbitrary subpopulations located around major transportation hubs and leverages commuting and air travel data. In contrast, our approach utilizes state, province, and territory boundaries to define locations and is not informed by air travel data. Instead of adjusting the estimation, the GLEAM platform uses case data and other datasets to conduct a calibration run aimed at identifying the optimal set of parameters that best fit the real data in its estimations[39].

The role of asymptomatic individuals has been shown to be central for the spread of viruses like SARS-CoV-2 and influenza, with most transmissions occurring with exposure times exceeding an hour[40]. Therefore, it is reasonable to center a model system around work-related commuting; infectious and asymptomatic individuals share the same space with coworkers for several hours, increasing their probability of infection; subsequently, each worker returns home to their resident locations, further increasing the probability of spreading disease to their families. Conversely, the GLEAM platform focuses on global disease spread, showing that the global spatiotemporal patterns of disease spreading are mainly determined by the airline network[41]. Additionally, the inclusion of air travel data enables GLEAM to capture global phenomena such as the external introduction or reintroduction of the virus during the estimation process. In contrast, the SEIR-EAKF model developed in this work is more sensitive to local changes in trends, as it leverages case data to adjust estimations. The importance of modeling each introduction event is particularly relevant at the beginning of an epidemic. This significance decreases once the epidemic in a region is primarily driven by internal transmission dynamics. Our model accommodates the external introduction of new infections through stochastic integration instead of relying on international flight data. However, the inclusion of global air travel data can improve system estimation, particularly in the initial weeks following the introduction of virus and for epidemics less prevalent than COVID-19. These differences make the two modeling approaches particularly effective at capturing different aspects and phases of epidemics. Therefore, combining these and other approaches into a single model or multi-model ensemble could improve the capacity to estimate and predict disease parameters.

The development of multinational dynamical models entails more time and effort compared to localized models because it requires the reconciliation of heterogeneous data sources. For example, the three countries in this study conducted independent census surveys that needed to be carefully interpreted and meticulously merged to ensure the resulting multinational contact network was homogeneous. The great advantage of using realistic contact networks that encompass multiple countries is the ability to infer simultaneously across broader regions (e.g. North America), rather than comparing results from separate local inference systems, thus highlighting the geographic spread of the disease across borders. Furthermore, the inference system implemented in this study can serve as a platform for modeling other respiratory infectious diseases, such as influenza, by pairing the North American commuting network developed here with other mathematical models.

Robust models for estimating disease parameters not only help to understand disease dynamics and assess responses at different locations, but they also can be adapted for use during epidemics to generate forecasts that can help health authorities in developing and implementing more informed policies. Using spatially resolved dynamical models applied at scale, it is possible to compare how various factors such as public health policies, population density, and mobility patterns affect disease spread and control in different locations. This comparative inference helps identify the most effective strategies and conditions for controlling epidemics, providing valuable insights for tailoring public health interventions to specific regions or populations. Additionally, this model can be implemented to support monitoring systems and counterfactual simulation, enhancing public health preparedness and response by enabling data-driven and location-specific strategies that can directly improve epidemic control of the region.

Contributors

Data curation: MP. Formal Analysis: MP, TKY, MG. Investigation: MP. Software: MP, TKY, MG, JS RKS. Methodology: MP, JS, RKS. Writing – original draft: MP, JS, RK. Validation: TKY, MG. Supervision: TKY, MG, JS. Writing – review & editing: TKY, MG, JS. Funding acquisition: JS. Conceptualization: JS. Project administration: JS.

Data sharing statement

The model scripts and the input data are publicly available on GitHub https://github.com/MatteoPS/NA_SEIR-EAKF

Declaration of interest

JS and Columbia University disclose partial ownership of SK Analytics. JS discloses consulting for BNI. All other authors declare no competing interests.

References

1. Pagel C, Yates CA. Role of mathematical modelling in future pandemic response policy. BMJ [Internet]. 2022 Sep 15 [cited 2024 Mar 18];e070615. Available from: https://www.bmj.com/lookup/doi/10.1136/bmj-2022-070615

2. Zhou X, Ma X, Gao S, Ma Y, Gao J, Jiang H, et al. Measuring the worldwide spread of COVID-19 using a comprehensive modeling method. BMC Med Inform Decis Mak [Internet]. 2023 Sep 15 [cited 2024 Mar 18];21(S9):384. Available from: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-023-02213-4

3. Zhang W, Xie R, Dong X, Li J, Peng P, Dr Santibanez Gonzalez E. SEIR-FMi: A coronavirus disease epidemiological model based on intra-city movement, inter-city movement and medical resource investment. Comput Biol Med [Internet]. 2022 Oct [cited 2024 Mar 18];149:106046. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0010482522007636

4. Sen Pei, Kandula S, Shaman J. Differential effects of intervention timing on COVID-19 spread in the United States. Sci Adv [Internet]. 2020 Dec 4 [cited 2023 Dec 22];6(49):eabd6370. Available from: https://www.science.org/doi/10.1126/sciadv.abd6370

5. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science [Internet]. 2020 May [cited 2023 Dec 22];368(6490):489–93. Available from: https://www.science.org/doi/10.1126/science.abb3221

6. Bhattacharya P, Machi D, Chen J, Hoops S, Lewis B, Mortveit H, et al. Novel multi-cluster workflow system to support real-time HPC-enabled epidemic science: Investigating the impact of vaccine acceptance on COVID-19 spread. J Parallel Distrib Comput [Internet]. 2024 Sep 1 [cited 2024 Jun 3];191:104899. Available from: https://www.sciencedirect.com/science/article/pii/S0743731524000637

7. Kong L, Duan M, Shi J, Hong J, Chang Z, Zhang Z. Compartmental structures used in modeling COVID-19: a scoping review. Infect Dis Poverty [Internet]. 2022 Jun 21 [cited 2024 Jan 23];11(1):72. Available from: https://idpjournal.biomedcentral.com/articles/10.1186/s40249-022-01001-y

8. Wang P, Zheng X, Liu H. Simulation and forecasting models of COVID-19 taking into account spatio-temporal dynamic characteristics: A review. Front Public Health [Internet]. 2022 Oct 18 [cited 2024 Jan 23];10:1033432. Available from: https://www.frontiersin.org/articles/10.3389/fpubh.2022.1033432/full

9. Venkatramanan S, Sadilek A, Fadikar A, Barrett CL, Biggerstaff M, Chen J, et al. Forecasting influenza activity using machine-learned mobility map. Nat Commun [Internet]. 2021 Feb 9 [cited 2024 Jun 4];12(1):726. Available from: https://www.nature.com/articles/s41467-021-21018-5

10. Balcan D, Gonçalves B, Hu H, Ramasco JJ, Colizza V, Vespignani A. Modeling the spatial spread of infectious diseases: The GLobal Epidemic and Mobility computational model. J

Comput Sci [Internet]. 2010 Aug [cited 2024 Jan 22];1(3):132–45. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1877750310000438

11. Pastore-Piontti A, Zhang Q, Gomes MFC, Rossi L, Poletto C, Colizza V, et al. Real-Time Assessment of the International Spreading Risk Associated with the 2014 West African Ebola Outbreak. In: Chowell G, Hyman JM, editors. Mathematical and Statistical Modeling for Emerging and Re-emerging Infectious Diseases [Internet]. Cham: Springer International Publishing; 2016 [cited 2024 Jan 23]. p. 39–56. Available from: http://link.springer.com/10.1007/978-3-319-40413-4_4

12. Zhang Q, Sun K, Chinazzi M, Pastore Y Piontti A, Dean NE, Rojas DP, et al. Spread of Zika virus in the Americas. Proc Natl Acad Sci [Internet]. 2017 May 30 [cited 2024 Jan 23];114(22). Available from: https://pnas.org/doi/full/10.1073/pnas.1620161114

13. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science [Internet]. 2020 Apr 24 [cited 2024 Jan 22];368(6489):395–400. Available from: https://www.science.org/doi/10.1126/science.aba9757

14. Pei S, Yamana TK, Kandula S, Galanti M, Shaman J. Burden and characteristics of COVID-19 in the United States during 2020. Nature [Internet]. 2021 Oct 14 [cited 2024 Jan 19];598(7880):338–41. Available from: https://www.nature.com/articles/s41586-021-03914-4

15. CSSEGISandData. CSSEGISandData/COVID-19 [Internet]. COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. 2024 [cited 2024 Jan 19]. Available from: https://github.com/CSSEGISandData/COVID-19

16. Statistics Canada. Population and Dwelling Count Highlight Tables, 2016 Census – Canada, provinces and territories [Internet]. [cited 2024 May 13]. Available from: https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Table.cfm?Lang=Eng&T=101&S=50&O=A

17. Bureau UC. Population [Internet]. Census.gov. [cited 2024 May 13]. Available from: https://www.census.gov/topics/population.html

18. INEGI. Intercensal Survey 2015 [Internet]. [cited 2023 Dec 22]. Available from: https://en.www.inegi.org.mx/programas/intercensal/2015/#Microdatos

19. Government of Canada SC. Commuting Flow from Geography of Residence to Geography of Work - Census Metropolitan Areas and Census Agglomerations: Main Mode of Commuting (10), Commuting Duration (6) and Sex (3) for the Employed Labour Force Aged 15 Years and Over Having a Usual Place of Work, in Private Households, 2016 Census - 25% Sample Data [Internet]. 2017 [cited 2024 Apr 3]. Available from: https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/dt-td/Rp-eng.cfm?TABID=4&LANG=E&A=R&APATH=3&DETAIL=0&DIM=0&FL=A&FREE=0&GC=0&GL=-1&GID=1354564&GK=0&GRP=1&O=D&PID=111333&PRID=10&PTYPE=109445&S=0&SHOWALL=0&SUB=0&Temporal=2017&THEME=125&VID=0&VNAMEE=&VNAMEF=%20(2017)&D1=0&D2=0&D3=0&D4=0&D5=0&D6=0

20. Statistics Canada. Number of vehicles travelling between Canada and the United States [Internet]. Government of Canada; [cited 2023 Dec 22]. Available from: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2410000201

21. US Census Bureau. 2011-2015 5-Year ACS Commuting Flows [Internet]. [cited 2023 Dec 22]. Available from: https://www.census.gov/data/tables/2015/demo/metro-micro/commuting-flows-2015.html

22. COVID-19 Community Mobility Reports [Internet]. [cited 2023 Dec 22]. Available from: https://www.google.com/covid19/mobility/

23. Diekmann O, Heesterbeek JAP, Roberts MG. The construction of next-generation matrices for compartmental epidemic models. J R Soc Interface [Internet]. 2010 Jun 6 [cited 2024 Jan 23];7(47):873–85. Available from: https://royalsocietypublishing.org/doi/10.1098/rsif.2009.0386

24. COVID-19 Open Data — Google Health [Internet]. [cited 2023 Dec 22]. Available from: https://health.google.com/covid-19/open-data/data-sources

25. Anderson JL. An Ensemble Adjustment Kalman Filter for Data Assimilation. Mon Weather Rev [Internet]. 2001 Dec [cited 2023 Dec 22];129(12):2884–903. Available from: http://journals.ametsoc.org/doi/10.1175/1520-0493(2001)129<2884:AEAKFF>2.0.CO;2

26. Shaman J, Karspeck A, Yang W, Tamerius J, Lipsitch M. Real-time influenza forecasts during the 2012–2013 season. Nat Commun [Internet]. 2013 Dec 3 [cited 2024 Jan 9];4(1):2837. Available from: https://www.nature.com/articles/ncomms3837

27. DeFelice NB, Little E, Campbell SR, Shaman J. Ensemble forecast of human West Nile virus cases and mosquito infection rates. Nat Commun [Internet]. 2017 Feb 24 [cited 2024 Jan 9];8(1):14592. Available from: https://www.nature.com/articles/ncomms14592

28. Yang W, Karspeck A, Shaman J. Comparison of Filtering Methods for the Modeling and Retrospective Forecasting of Influenza Epidemics. Ferguson NM, editor. PLoS Comput Biol [Internet]. 2014 Apr 24 [cited 2024 Jan 23];10(4):e1003583. Available from: https://dx.plos.org/10.1371/journal.pcbi.1003583

29. Pei S, Kandula S, Yang W, Shaman J. Forecasting the spatial transmission of influenza in the United States. Proc Natl Acad Sci [Internet]. 2018 Mar 13 [cited 2024 Jan 23];115(11):2752–7. Available from: https://pnas.org/doi/full/10.1073/pnas.1708856115

30. Dahal S, Banda JM, Bento AI, Mizumoto K, Chowell G. Characterizing all-cause excess mortality patterns during COVID-19 pandemic in Mexico. BMC Infect Dis [Internet]. 2021 Dec [cited 2024 Mar 20];21(1):432. Available from: https://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-021-06122-7

31. Hasell J, Mathieu E, Beltekian D, Macdonald B, Giattino C, Ortiz-Ospina E, et al. A cross-country database of COVID-19 testing. Sci Data [Internet]. 2020 Oct 8 [cited 2024 Mar 20];7(1):345. Available from: https://www.nature.com/articles/s41597-020-00688-8

32. Lee MHY, Xu G, Cheng F, Khalid AF. Testing surge capacity—A Canadian COVID-19 experience, Ontario's surge capacity for the first wave. Health Policy [Internet]. 2021 Oct [cited 2024 Mar 26];125(10):1291–6. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0168851021002025

33. Jones JM, Stone M, Sulaeman H, Fink RV, Dave H, Levy ME, et al. Estimated US Infection- and Vaccine-Induced SARS-CoV-2 Seroprevalence Based on Blood Donations, July 2020-May 2021. JAMA [Internet]. 2021 Oct 12 [cited 2024 Jan 12];326(14):1400. Available from: https://jamanetwork.com/journals/jama/fullarticle/2784013

34. The COVID Tracking Project [Internet]. The COVID Tracking Project. [cited 2024 Feb 26]. Available from: https://covidtracking.com/data/national

35. Canada PHA of. COVID-19 daily epidemiology update: Testing and variants [Internet]. aem. 2020 [cited 2024 Feb 26]. Available from: https://www.canada.ca/en.html

36. Mathieu E, Ritchie H, Rodés-Guirao L, Appel C, Giattino C, Hasell J, et al. Coronavirus Pandemic (COVID-19). Our World Data [Internet]. 2020 Mar 5 [cited 2024 Feb 26]; Available from: https://ourworldindata.org/coronavirus-testing

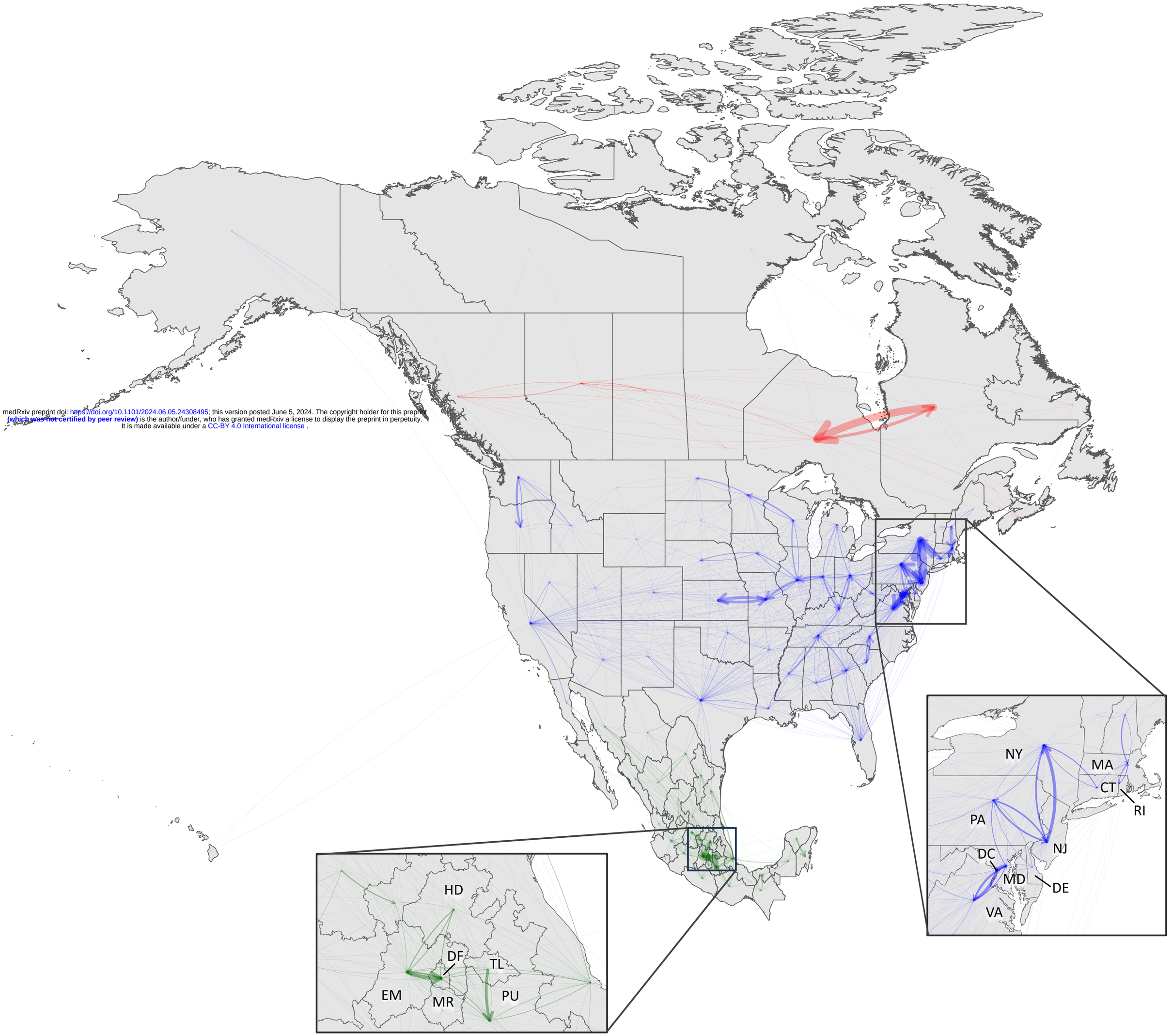
37. Dirección General de Epidemiología. Lineamiento estandarizado para la vigilancia epidemiológica y por laboratorio de enfermedad por 2019-nCoV. 2020 [Internet]. Available from: https://www.gob.mx/salud/documentos/lineamiento-estandarizado-para-la-vigilancia-epidemiologica-y-por-laboratorio-de-la-enfermedad-respiratoria-viral

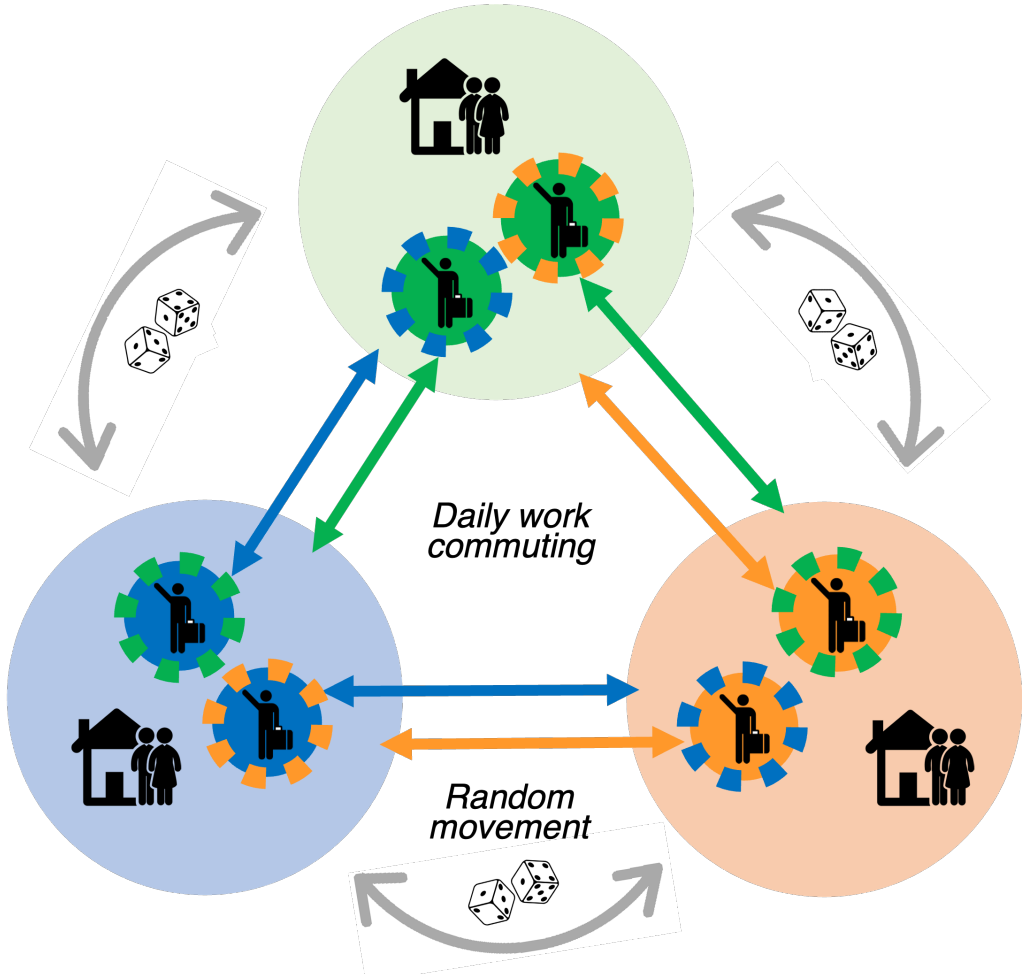
38. De La Cruz-Hernández SI. Another Vision of the Situation of the COVID-19 Pandemic in Mexico During 2020. Disaster Med Public Health Prep [Internet]. 2022 Dec [cited 2024 Apr 23];16(6):2296–8. Available from: https://www.cambridge.org/core/product/identifier/S1935789321003402/type/journal_article

39. Davis JT, Chinazzi M, Perra N, Mu K, Pastore y Piontti A, Ajelli M, et al. Cryptic transmission of SARS-CoV-2 and the first COVID-19 wave. Nature [Internet]. 2021 Dec [cited 2024 May 1];600(7887):127–32. Available from: https://www.nature.com/articles/s41586-021-04130-w

40. Ferretti L, Wymant C, Petrie J, Tsallis D, Kendall M, Ledda A, et al. Digital measurement of SARS-CoV-2 transmission risk from 7 million contacts. Nature [Internet]. 2024 Feb [cited 2024 Mar 29];626(7997):145–50. Available from: https://www.nature.com/articles/s41586-023-06952-2

41. Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco JJ, Vespignani A. Multiscale mobility networks and the spatial spreading of infectious diseases. Proc Natl Acad Sci [Internet]. 2009 Dec 22 [cited 2024 Mar 29];106(51):21484–9. Available from: https://www.pnas.org/doi/10.1073/pnas.0906910106





DAYTIME
Individuals move to workplace

$S_{ij}(t_1)$ → $E_{ij}(t_1)$ (rate $\mu\beta$, β)

$E_{ij}(t_1)$ → $I^u_{ij}(t_1)$ (rate $1/Z$, $1-\alpha$)

$E_{ij}(t_1)$ → $I^r_{ij}(t_1)$ (rate $1/Z$, α)

$I^u_{ij}(t_1)$ → $R_{ij}(t_1)$ (rate $1/D$)

$I^r_{ij}(t_1)$ → $R_{ij}(t_1)$ (rate $1/D$)

NIGHTTIME
Individuals are home

$S_{ij}(t_2)$ → $E_{ij}(t_2)$ (rate $\mu\beta$, β)

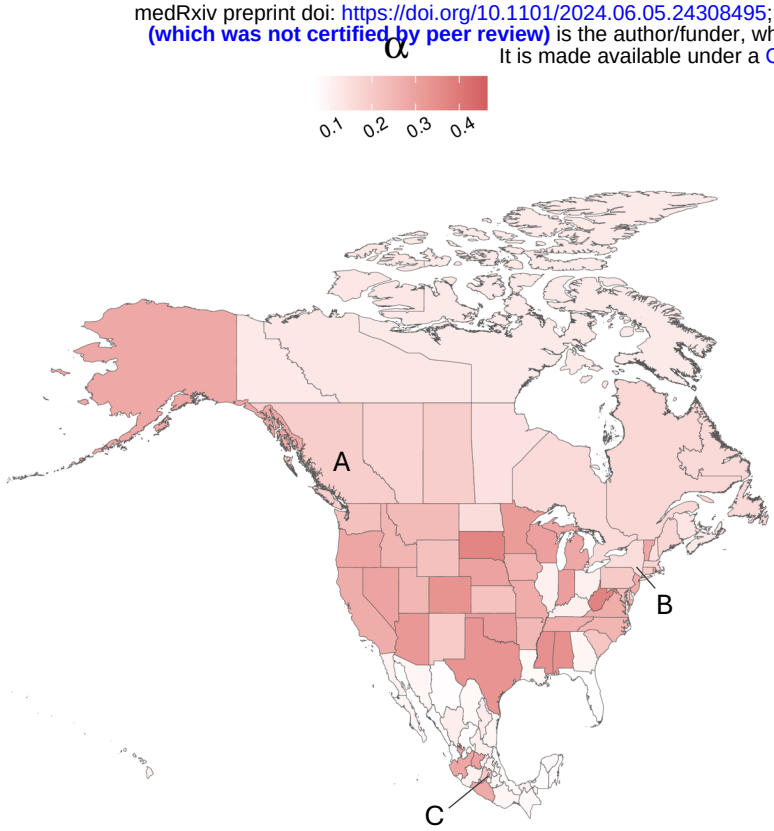
$E_{ij}(t_2)$ → $I^u_{ij}(t_2)$ (rate $1/Z$, $1-\alpha$)

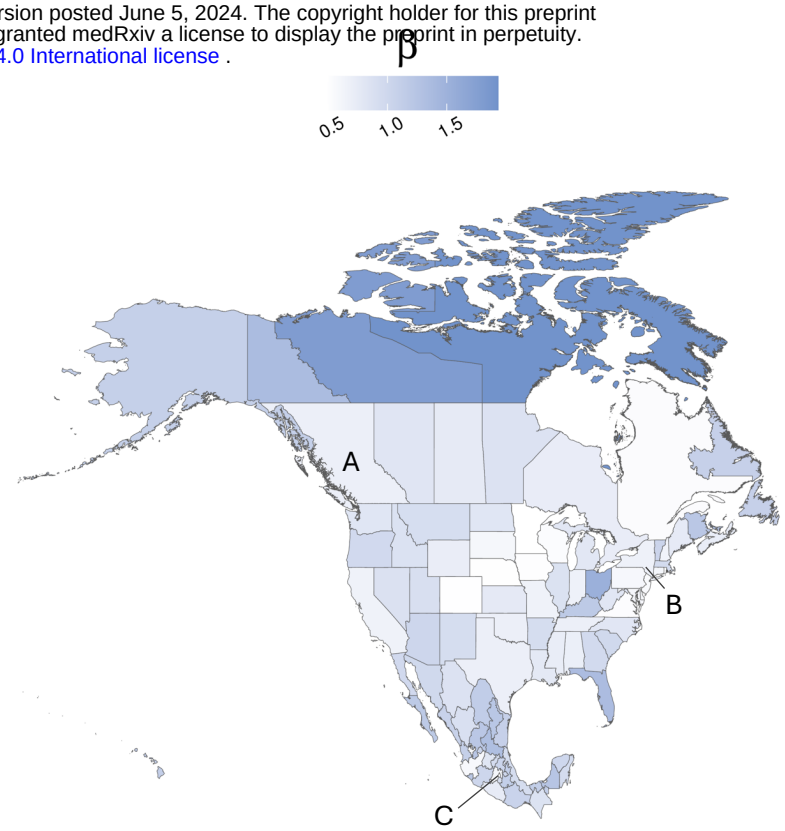
$E_{ij}(t_2)$ → $I^r_{ij}(t_2)$ (rate $1/Z$, α)

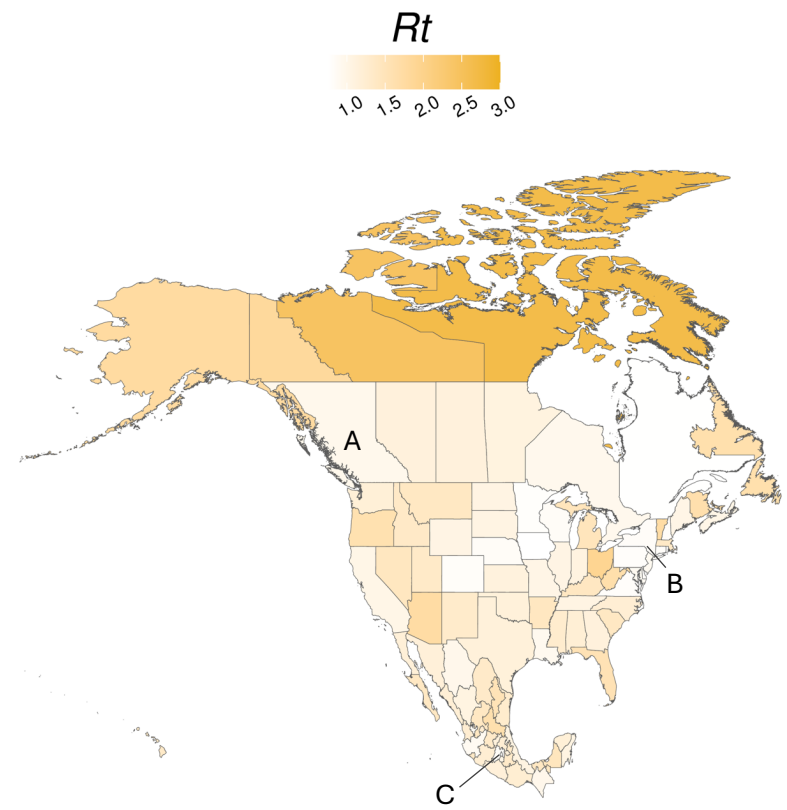
$I^u_{ij}(t_2)$ → $R_{ij}(t_2)$ (rate $1/D$)

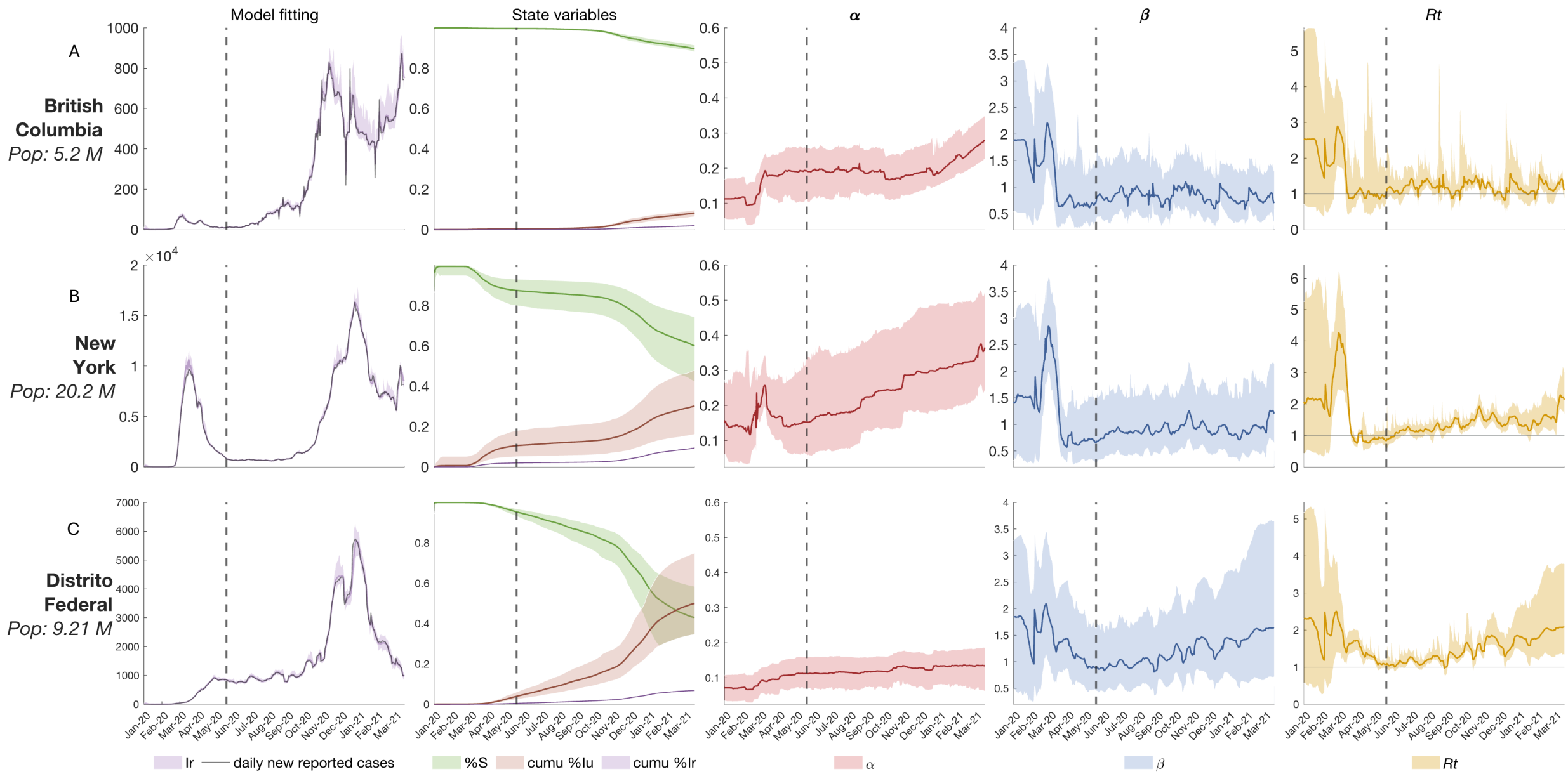
$I^r_{ij}(t_2)$ → $R_{ij}(t_2)$ (rate $1/D$)

6th June 2020

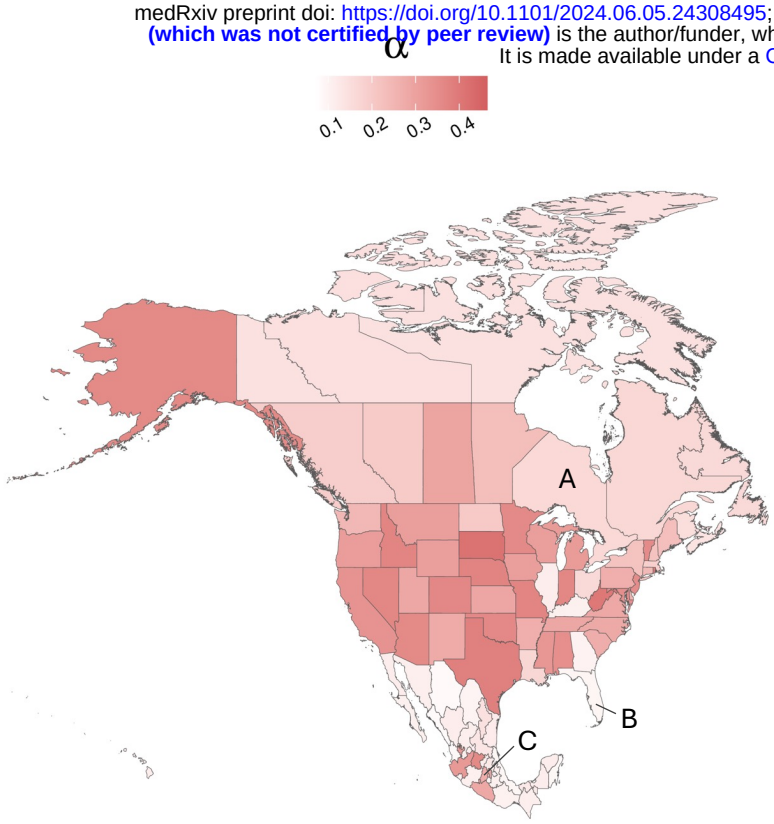


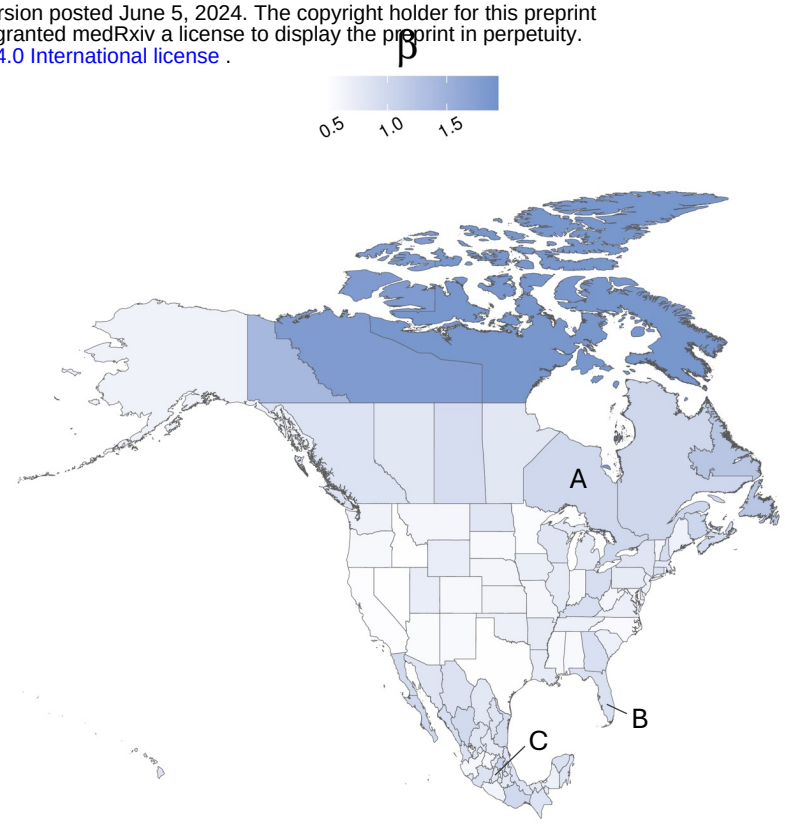


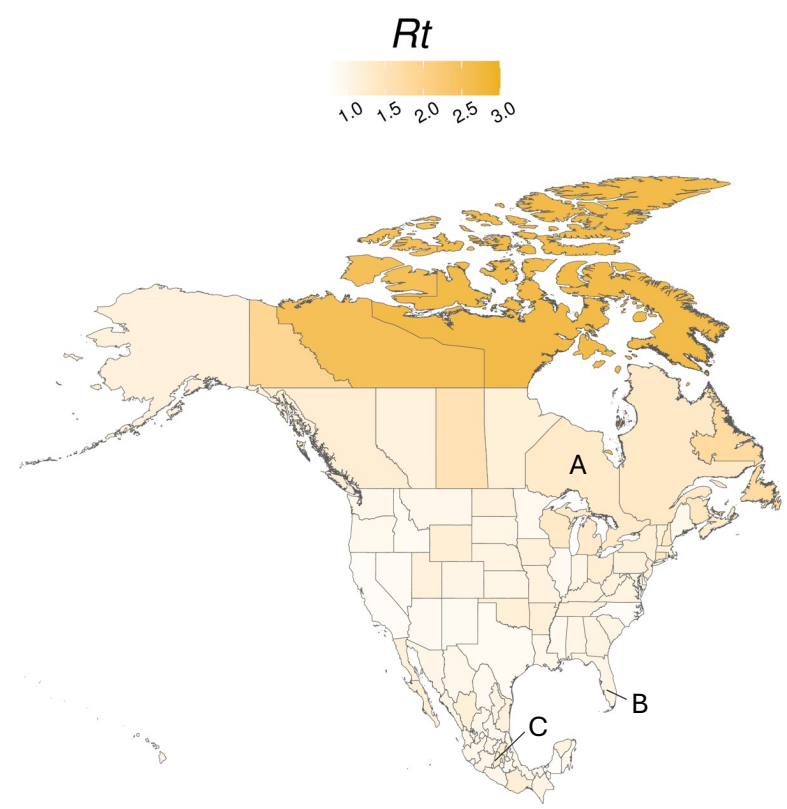


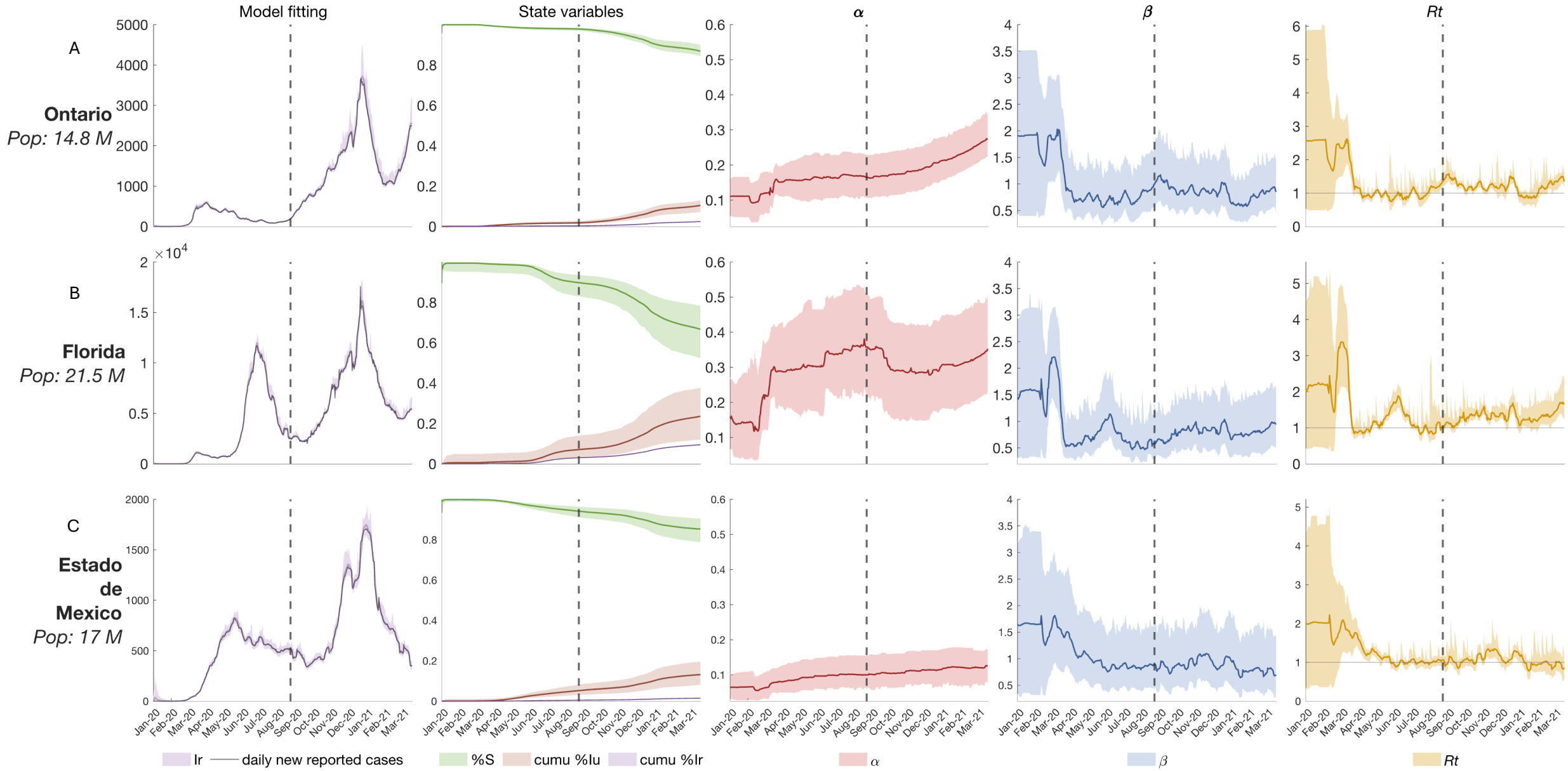


7th September 2020









15th March 2021