Machine Learning for Mechanistic Models of Metapopulation Dynamics

Jifan Li,[1] Edward L. Ionides,[2*] Aaron A. King[3], Mercedes Pascual[4], Ning Ning[1*]

[1]Department of Statistics, Texas A&M University
[2]Department of Statistics, University of Michigan
[3]Department of Ecology & Evolutionary Biology, University of Michigan
[4]Department of Environmental Studies, New York University
[*]To whom correspondence should be addressed; E-mail: ionides@umich.edu, patning@tamu.edu

Abstract

Mathematical models in ecology and epidemiology must be consistent with observed data in order to generate reliable knowledge and evidence-based policy. Metapopulation systems, which consist of a collection of sub-populations at various locations, pose technical challenges in statistical inference due to nonlinear, stochastic interactions. Difficulties encountered in these methodological tasks can obstruct the core scientific questions concerning the link between the mathematical models and the data. Progress in statistically efficient simulation-based inference for partially observed stochastic dynamic systems has enabled the development of statistically rigorous approaches to the analysis of nonlinear but low-dimensional systems. Recently, an algorithm has been developed which enables comparable inference for higher-dimensional models arising in metapopulation systems. The COVID-19 pandemic provides a situation where mathematical models and their policy implications were widely visible, and we revisit an influential metapopulation model used to inform basic epidemiological understanding early in the pandemic. Our methods support self-critical data analysis, enabling us to identify and address model limitations, and leading to a new model with substantially improved statistical fit and parameter identifiability. Our results suggest that the lockdown initiated on January 23, 2020 in China was more effective than previously thought. We proceed to recommend statistical analysis standards for future metapopulation system modeling.

Introduction

Biological populations may be structured into a collection of densely-populated communities separated by sparsely populated regions. The communities, which may be cities in a human context, comprise a metapopulation. Motivation for metapopulation modeling arises when some essential feature of the population dynamics cannot be understood from looking at a single location. Dynamics of persistence through local extinctions and reintroductions have been extensively studied in ecology [1, 2]. In epidemiology, metapopulation dynamics can be a barrier to the regional elimination and eventual eradication of a pathogen, and may determine the successful invasion of a new pathogen or a new strain of an existing pathogen [3]. In other situations, spatiotemporal dynamics may be an unavoidable component of the system under study without being the focus of the investigation [4, 5].

A recent growth in the study of metapopulation dynamics has been driven partly by the COVID-19 pandemic [6–14] and in part by methodological advances facilitating the fitting of metapopulation models to

spatiotemporal data. Until the start of this millenium, developing dynamic models with both statistical and scientific justification was a longstanding open problem for even a single community [15]. Over the past two decades, new algorithms [16–19] and software [20–22], together with ever-increasing computational resources, have enabled routine inference for low-dimensional nonlinear partially observed stochastic dynamic systems. However, fundamental algorithmic scalability issues known as the "curse of dimensionality" lead to difficulties with the high-dimensional systems arising in metapopulation inference. These issues are clearest for Monte Carlo techniques based on importance sampling [23] but are also evident in the need for variational approximations for large Monte Carlo Markov Chain (MCMC) calculations [24]. The quest for statistical methods suitable for metapopulation models of scientific interest has motivated the development of various methodological approaches. Nevertheless, limitations have persisted in the capability to carry out flexible model-based inference and rigorous model criticism. Thus, data analysis for metapopulation models has lagged behind low-dimensional time series analysis of biological systems. Recent developments enable this gap to be closed, as we demonstrate via a reanalysis of COVID-19, viewed from the context of the ability to drawm evidence-based scientific conclusions about the dynamics of the emerging pandemic in January and February 2020.

Review of metapopulation models and inference methods

Biological systems are characterized by nonlinear stochastic dynamics together with incomplete and noisy measurements [15]. We therefore focus on the class of partially observed Markov process (POMP) models [25], acknowledging that deterministic models can be conceptually useful but are problematic as statistical explanations of noisy systems [5, 26]. The Markov property asserts that the dynamic process has no memory conditional on its current state, which is algorithmically convenient while being scientifically nonrestrictive since we can choose what to include in the state. Metapopulation models consider a multivariate system state at each location and so we require methods tailored for high-dimensional POMP models. Simplifications arise if models and data are limited to binary presence-absence, or a small discrete set of values at each location [2], but we focus on situations where abundance data are available, such as case reports for infectious diseases.

A natural place to look for statistical methodology applicable to metapopulation models is among the techniques developed for inferring population dynamics at a single location, reviewed by [27, 28]. Commonly implemented approaches for POMP models can be categorized as (i) variants of MCMC; (ii) matching summary statistics between data and simulations; (iii) linearization; (iv) particle filters (i.e., sequential Monte Carlo).

In principle, MCMC techniques enable Bayesian inference or maximum likelihood via expectation-maximization algorithms [29]. In practice, successful MCMC for metapopulation models requires careful model-specific algorithm development [30].

Matching summary statistics of simulations to the corresponding data statistic is, in principle, a readily applicable inference approach for a wide class of models including metapopulation models. In the context of Bayesian inference this is called Approximate Bayesian Computing [31]. However, informative, low-dimensional summary statistics can be hard to construct even for low-dimensional nonlinear systems. This can make summary statistic methods statistically inefficient [32].

Population dynamics may be approximately linear on a log scale, and this has been used to develop linearization methods for epidemiological time series analysis [33] that have been extended to metapopulation analysis [34]. This provides a numerically convenient set of tools, but requires scientists to work within a limited class of models.

The ensemble Kalman filter (EnKF) developed in the context of massive geophysical model, combines an ensemble representation with a computationally efficient update rule inspired by the scalable linear Kalman filter, providing an approach with excellent scalability [35, 36]. For biological systems, EnKF was first demonstrated as a computationally convenient tool for compartment models at a single location [37, 38]. Subsequently, it has been applied for epidemiological metapopulation inference [6, 39]. However, the linearization in the EnKF filter update rule can be problematic for highly nonlinear systems [36]. Further, a linear update rule is not appropriate for small, discrete populations unless EnKF is embedded within a MCMC algorithm [40].

For low-dimensional systems, particle filter methods are applicable to a flexible class of models: they permit consideration of arbitrary nonlinear dynamics and require the model to be specified only via a simulator [20, 25]. Particle filters enable statistically efficient use of data, since they provide an evaluation of the likelihood function required for Bayesian or likelihood-based inference, with approximation resulting only from finite Monte Carlo effort. For high-dimensional systems, scalability considerations demand further approximations since particle filters suffer acutely from the "curse of dimensionality" [23]. In contrast, block particle filter (BPF) methods achieve scalability by updating particles through localized resampling, rather than employing the linear update of EnKF [41]. It is an empirical question which of these approximations is more successful on metapopulation models, with prior evidence favoring BPF [42].

Fitting complex models to large datasets using computationally intensive methods is the domain of machine learning. Some machine learning models, such as artificial neural networks, lack a mechanistic interpretation. Here, we focus instead on machine learning for mechanistic models. We not only compare available machine learning methods for fitting metapopulation models but, importantly, we also consider issues of model criticism. Specifically, we introduce a model diagnosis procedure to mitigate the risk of selecting an inappropriate model, an essential step in informing public policy decisions.

Metapopulation analysis of COVID-19 spread in China

Models are an irreplaceable tool to inform public policy, despite delicate issues in their implementation and interpretation [43, 44]. Some of the difficulties are operational, others are conceptual. Operationally, we seek to fit complex models using statistically valid, reproducible and transparent methods. Conceptual difficulties arise when drawing causal conclusions from fitting models to observational data, giving rise to opportunities for incorrect conclusions due to missing variables or other forms of model misspecification. A strength of metapopulation modeling is that the model can be build upon established scientific knowledge, which may give some protection against gross forms of model misspecification (for example, asserting the existence of latent dynamic variables which simply do not exist). A model assimilated to data guarantees that assumptions have been framed in a way consistent with certain facts, and evidence for predictive skill can support the value of the model construction. Indeed, it can be practically impossible to make sense of the nonlinear stochastic interactions driving biological dynamics without representing them via a model [43, 45]. We therefore seek to harness the power of models while assessing, acknowledging and minimizing their weaknesses.

As an example, we reconsider the influential analysis of COVID-19 from early in the pandemic by Li et al. [6]. This analysis provided estimates of transmission parameters and the effect of the lockdown in China using the limited data available at the time. Other teams have fitted models to address similar questions [13, 46, 47] but the study by [6] is distinctive for fitting a stochastic mechanistic metapopulation model to extensive spatiotemporal data. The results were published in May, 2020, based on reported cases from January 10 to February 8 of that year. The state-of-the-art spatiotemporal analysis was possible on

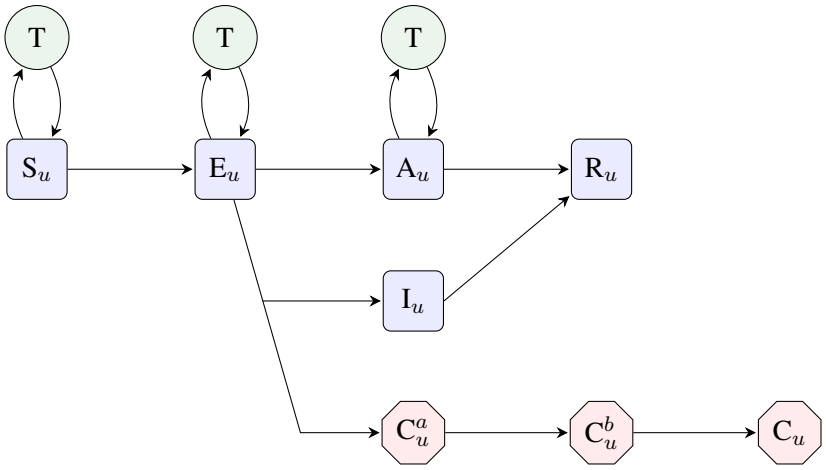

Figure 1: A flow diagram for the SEAIR metapopulation model. Each individual in city u is a member of exactly one of the square blue compartments. Individuals entering the reportable infectious compartment, I_u for city u, are simultaneously included in the delayed reporting process compartment, $\mathrm{C}_{\mathrm{u}}^{\mathrm{a}}$. Upon arrival at the final reporting compartment, C_u, the individual is included in the case report for city u. Individuals in A_u are not reportable and transmit at a reduced rate. Movement of individuals between cities occurs by transport to and from a transport compartment, T. The number of individuals moving between each pair of cities is based on 2018 data from Tencent. Movement is modeled only for susceptible, exposed, and undetected infections.

an urgent timescale because the team of researchers had developed their methodology in a sequence of previous situations [37, 38, 48, 49]. The paper is written with attention to reproducibility, and the main results are strengthened by various supporting analyses in an extensive supplement. While examining the points mentioned above, we have identified various limitations that could have been mitigated by adhering to the aforementioned recommendations. Our goal is not to criticize any specific paper, but rather to build on the timely analysis of [6] to demonstrate how recently developed techniques provide possibilities to carry out improved data analysis in future.

For our metapopulation system, the sub-populations are 373 provincial cities in China (meaning cities with administrative responsibility for an entire region) and the data are daily reported COVID-19 cases. COVID-19 dynamics are represented by an Susceptible-Exposed-Asymptomatic-Infectious-Recovered (SEAIR) epidemic model. Questions of urgent interest early in the pandemic include the relative transmissibility of reported to unreported cases, the fraction of unreported cases, and the effect on transmission of movement restrictions imposed on and around January 23 [6]. The model structure is illustrated by the flow diagram in Fig. 1. We consider different model implementations within this structure, with full model specifications provided by the equations and parameter values in Supplementary Sec. S1.

Our starting point is model M_1 which is based on the model of [6] and is fully described in Supplementary Sec. S1.3. We consider the full dataset, from January 10 to February 8, with transmission parameters re-estimated following the lockdown on January 23; these correspond to the periods 1 (January 10 to January 22) and 3 (January 24 to February 8) of [6]. Some minor differences between M_1 and [6] were introduced to enable us to place their model within the general metapopulation framework of spatiotemporal partially observed continuous-time Markov processs described by [42]. Despite these modifications, simulations from M_1, using the parameters of [6], closely match simulations from the code provided by [6] (Supplementary Sec. S1.3). However, inspection of the mobility data reveals that some small cities have no recorded in-

coming travelers, and therefore no possibility of a SARS-CoV2 introduction within M_1 (or the model of [6]) (Supplementary Sec. S1.2). This minor limitation formally results in a likelihood of zero for M_1 (i.e., it is impossible for the simulation model to reproduce the observed spatiotemporal dataset), and hence a log-likelihood of $-\infty$.

Model	loglik	df	description
M_1	$-\infty$	10	SEAIR model using the parameter values and mobility data of [6]
M_2	-14240.5	10	Adjusted mobility and measurement in M_1
M_3	-11257.9	374	Independent identically distributed negative binomial
M_4	-10825.3	375	Autoregressive negative binomial
M_5	-9088.2	12	Adding overdispersed dynamics to M_2 and refitting
M_6	-9116.5	10	Latent and infectious durations unchanged by lockdown in M_5.

Table 1: Model comparisons by log-likelihood, evaluated by a block particle filter. The degrees of freedom (df) is the number of estimated parameters.

We addressed the problematic mobility data in M_1 by adding some additional transportation based on a gravity movement model, as described in Supplementary Sec. S1.2, giving rise to model M_2. We implemented an additional adjustment between models M_1 and M_2 to align the measurement model with the ensemble Kalman filter (EnKF) inference method presented by [6]. That EnKF implementation involved specifying a quantity called the observation error variance, defined as a function of the observed cases, to quantify the uncertainty in the measurements. Within the POMP specification, the measurement variance can depend on the latent state but not directly on the observed data. To interpret the choice of EnKF observation variance within the POMP framework, we specified the measurement model for M_2 to have equivalent scaling to the choice of [6], but with dependence on the reported cases replaced by dependence on the modeled, but unobserved, exact case count.

Based on a comparison of various nonlinear spatiotemporal filters (Supplemtary Fig. S9) we evaluated the log-likelihood for M_2 using a block particle filter (Table 1). To account for model overfitting, the number of estimated parameters can be subtracted from the log-likelihood to obtain a comparison equivalent to Akaike's Information Criterion (AIC) [50]. When the difference in log-likelihood is large compared to the difference in degrees of freedom, the ordering of statistical goodness-of-fit is clear without presenting formal statistical hypothesis tests.

To find out whether this log-likelihood value suggests that the model is satisfactory, we compare it with two simple statistical models: M_3 simply models the daily case report for each city as an independent identically distributed (IID) negative binomial random variable; M_4 adds an autoregressive component to M_3 (see Supplemetary Sec. S2). We see from Table 1 that both M_3 and M_4 outperform M_2 by many units of log-likelihood. Likelihood can properly be compared between different models for the same data, with statistical uncertainty in log-likelihood differences arising on the unit scale [51]. When the fit of a mechanistic model is inferior to a simple statistical model, we learn that the mechanistic model has room for improvement as a description of the data, but we do not immediately learn what the deficiency is. The development of methods for formal statistical fitting of mechanistic models has led to increased understanding of the importance of appropriate modeling of over-dispersed variation in the stochastic dynamics [30, 52, 53]. We therefore hypothesized that the fit of M_2 could be improved by permitting additional dynamic noise.

A standard way to convert a deterministic model, constructed as a system of ordinary differential equations, into a stochastic model is to reinterpret the rates of the deterministic system as rates of a Markov chain [54]. This places limits on the mean-variance relationship of the resulting stochastic model [55]. Models allowing greater variability that permitted by this construction are said to be over-dispersed. We added mul-

tiplicative white noise to the transmission rate, following the approach of [25, 52], giving rise to model M_5. We fitted the model using an iterated block particle filter to maximize the likelihood [56, 57]. The block filter approximation has also proven helpful for spatiotemporal inference when using alternatives to particle filtering and alternatives to maximization by iterated filtering [30]. In the current context, the block particle filter was found to be more effective for likelihood evaluation than a test suite of alternative filters including the ensemble Kalman filter (Supplementary Fig. S9). The iterated block particle filter maximizes the block particle filter likelihood using an iterated filtering algorithm [19] adapted to the structure of a block particle filter.

Table 1 shows that model M_5 outperforms simple statistical benchmarks, obtaining a competitive likelihood with relatively few parameters. From a statistical perspective, M_5 is therefore an adequate statistical description of the data. However, some parameters of M_5 were weakly identified by the data, especially in the pre-lockdown time interval within which there were relatively few reported cases (Supplementary Sec. S5). When the evidence about the model parameters in the data is weak, the ambiguity may be resolved by other, unmodeled and poorly understood, aspects of the data. This risks leading to undesirable situations where substantial conclusions about questions of interest could be driven by the weaknesses of the model rather than its strengths. In Supplementary Sec. S5, we show how the flexibility of M_5 can be used to obtain a high likelihood via an unplausibly long estimated duration of infection during the pre-lockdown period, with the estimate suddenly reducing after lockdown. We resolved this issue by constraining the latent and infectious periods to be the same before and after lockdown, leading to model M_6. The additional constraints of M_6 lead to a small loss of likelihood compared to M_5, but the fit remains competitive compared to the benchmark models, and the stronger identifiability facilitates the interpretation of estimated parameters. Calculating the log-likelihood for each model in Table 1 requires extensive computation to produce a single number which contains essentially all the information about the statistical fit of the model. However, additional work is required to understand what characteristics of the models and data causes the differences in these numbers, and the practical consequences of the numerical results.

Parameter values for models M_1, M_5 and M_6 are reported in Supplementary Table S1. Here, we discuss the estimated basic reproductive number (i.e., the expected number of secondary infections from one index case in a fully susceptible population), denoted by $\mathcal{R}_0^{\mathrm{be}}$ and $\mathcal{R}_0^{\mathrm{af}}$ before and after the January 23rd lockdown. The mathematical formula used to calculate $\mathcal{R}_0^{\mathrm{be}}$ and $\mathcal{R}_0^{\mathrm{af}}$, in terms of the model parameters, is provided in Supplementary Table S1. Our estimates for model M_6, are $\mathcal{R}_0^{\mathrm{be}} = 3.51$ with confidence interval (CI) (3.31,3.72) and $\mathcal{R}_0^{\mathrm{af}} = 0.70$ with CI (0.65,0.77), where the estimates and their associated 95% CIs obtained by profile likelihood (Supplementary Sec. S6). This implies that the Chinese government non-pharmaceutical interventions instituted on and around January 23 reduced $\mathcal{R}_0$ by a factor of 5.0. By contrast, the estimates of [6] are $\widetilde{\mathcal{R}}_0^{\mathrm{be}} = 2.38$ with CI $(2.03, 2.77)$ and $\widetilde{\mathcal{R}}_0^{\mathrm{af}} = 0.98$ with CI $(0.83, 1.16)$, implying reduction by a factor of 2.4. For comparison, interventions implemented across a panel of 41 countries (34 European) were estimated to reduce $\mathcal{R}_0$ by a factor of 4.3 with CI $(2.9, 6.7)$ [58]. Our estimate for $\mathcal{R}_0$ before lockdown is toward the high end of previous estimates based on data up to February 2020, reviewed by [59]. An alternative metaopopulation analysis of the pre-lockdown China data, with a deterministic transmission model, obtained an $\mathcal{R}_0$ estimate of 3.11 with CI (2.39,4.13) [60]. Our $\mathcal{R}_0$ estimate is consistent with pre-lockdown estimates from other locations, such as New York city, for models that include asymptomatics [61].

Our model inherits the property of [6] that infections arising during the pre-lockdown period will generally be reported during the lockdown, due to the reporting delay modeled as a distributed delay with a mean of 9 days pre-lockdown and 6 days post-lockdown. Thus, the model is permitted to explain the data by inferring rapid, unreported spread prior to January 23. Despite this shared constraint on the form of the

model, conclusions of our analysis differ from [6]. Beyond the estimates of $\mathcal{R}_0$, a notable difference is that we find the estimated transmissibility of observed cases is close to that of unobserved cases, especially before lockdown (Supplementary Table S1).

Not all models are equal, and we have demonstrated an approach which evaluates the extent to which the postulated models statistically explain the observed data. Our analysis cannot disprove the possibility of an alternative model which explains the data even better via an alternative model specification, perhaps leading to alternative conclusions. Indeed, our methods are designed to facilitate others to develop and demostrate superiority to our own analysis when such advances are available.

If a mechanistic model has likelihood competitive with statistical benchmarks, it is anticipated to have simulations that are qualitatively comparable to the data. Since the model specification is inevitably imperfect, and is accounted for in the model fitting by noise processes, we expect simulations from the fitted model to have somewhat more stochastic variation than the data. By contrast, models which are structurally unable to provide sufficient variability to explain the data must give rise to simulations with too little stochasticity. Comparing model M_2 with M_5 demonstrates this (Supplementary Fig. S2). Models that have simulations with implausibly little variability give rise to claims of excessive confidence about the uncertainty surrounding estimated parameters. This phenomenon may be clearest when CIs are calculated using parametric bootstrap approach, involving re-estimation of parameters using artificial datasets simulated from a fitted model. However, it also applies for classical CI and Bayesian credible interval constructions. Thus, CIs from mechanistic models that outperform statistical benchmarks are anticipated to be conservative, whereas CIs from models with insufficient variability to explain the data are generally anti-conservative. Requiring model likelihoods to be comparable to statistical benchmarks therefore improves the credibility of uncertainty intervals as well as improving the accuracy of point estimates.

Discussion

Advances in statistical methodology will drive an increasing trend in the number of spatiotemporal models fitted to epidemiological data. The challenge of fitting intricate nonlinear models to extensive datasets makes it difficult for researchers to evaluate the limitations of their models and methods. Readers also can struggle to determine whether the proposed model has been adequately tested. It is therefore advisable to incorporate benchmarks for evaluating model performance in comparison to relatively simple statistical models [52]. This approach helps determine whether complex models provide a satisfactory level of explanatory power. In the first instance, these benchmarks can be applied to the entire dataset; subsequent analysis can focus on dissecting the contributions from various subsets of the data to gain a comprehensive understanding of which parts of the data drive the overall assessment.

Likelihood-based inference via particle filters has been considered inaccessible for metapopulation models due to the "curse of dimensionality" [23]. However, block particle filter methods can be effective on metapopulation models, as demonstrated in this paper and previously [5, 42, 57]. All high-dimensional nonlinear filters entail numerical approximation, and these can be assessed by comparing predictive skill (i.e., the estimated log-likelihood) between different filters. The ensemble Kalman filter provides a suitable point of comparison, since it has excellent scalability properties, modest capability to handle nonlinearities, and has been demonstrated on various epidemiological systems [10, 13, 37–39, 48, 49].

Software environments are critical for providing data analysis that is not only reproducible but also readily extendable. Scientists developing a data analysis should build an environment that empowers them to explore their own models and data, and then they should share this environment as part of the publication process. In practice, this involves encapsulating data analysis within a software package that immerses the

user in a documented environment where the models, methods and data used for the article can be readily be experimented with. Trustworthy data analysis should be supported by unit testing and documentation, and the quality of this support should be one of the considerations in evaluation of the data analysis. In other words, the article presenting the research should be part of a compendium [62]. The compendium for this article is comprised of the article source code, at https://github.com/jifanli/metapop_article, together with the software environment for the data analysis, provided by the R package at https://github.com/jifanli/metapoppkg.

Our research demonstrates that techniques proven effective in low-dimensional systems, such as population dynamics at one or two locations, can be extended to address larger metapopulation systems. This extension allows us to leverage well-established best practices from time series analysis, leading to a statistically principled approach. This approach enables us to identify and rectify model limitations that might otherwise remain undetected. Failure to address these weaknesses can lead to issues of irreproducibility and the provision of suboptimal policy recommendations when developing models for complex dynamic systems [44, 63]. Principles of good data analysis for population dynamics are presumably similar to general principles of data science [64] but require some adaptation to the specific situation. Here, we build on [44, 63, 64], by demonstrating the feasibility and desirability of metapopulation analysis meeting the specific set of criteria outlined below:

1. **Likelihood-based statistical inference**. A model, in conjunction with data, defines a likelihood function that quantifies the goodness of fit of the model and the data for each parameter value. For mechanistic models, it is usually impossible to write down the likelihood explicitly, but it still exists implicitly. Modern methods for implicit dynamic models permit evaluation and maximization of an implicit likelihood for metapopulation models [30, 56, 57]. Such methods extract all available information in the data about model parameters [51]. Log-likelihood is also a proper scoring rule for comparing probabilistic forecasts [65] and therefore provides a sensitive objective tool for model selection and identification of model misspecification. Whereas cross-validation and out-of-sample fit are standard benchmarks in machine learning settings [64], likelihood is better suited to situations with relatively small, spatiotemporally structured datasets.

2. **Statistical benchmarks**. The likelihood of a mechanistic model should be compared with that of a non-mechanistic statistical model, known as a benchmark [52]. A mechanistic model that statistically fits the data substantially worse than a non-mechanistic model is evidently unable to explain some aspect of the data. At the very least, this discrepancy should be identified and discussed.

3. **Residual analysis**. Introductory statistics classes, when covering linear regression, emphasize that a careful and complete data analysis involves examining deviations from the fitted model [66]. This is typically achieved by plotting residuals, a suitably rescaled measure of disparities between each observation and its corresponding fitted value. A relevant measure of residual in the current context is the *log-likelihood anomaly*, defined as the discrepancy between the mechanistic fit and a benchmark for components of the likelihood at each observation. Supplementary Sec. S7 describes how these tools were used for developing and evaluating model M_6.

4. **Uncertainty**. Reliable conclusions should be robust to plausible variations in data, models, and algorithms [64]. Standard statistical methods provide measures of uncertainty, and the validity of these measures depends critically on the statistical validity of the model. Appropriate modeling of overdispersion can be critical to accurate assessment of uncertainty for dynamic models [25, 30, 52, 53].

5. **Reproducibility and extendability**. Observational studies are not generally reproducible in an experimental sense. However, the numerical conclusions should be readily reproducible from the observations. A substantial part of the value of a computational model is that it permits *in silico* experimentation of the modeled system. The authors should build and share a computational environment that not only reproduces published numbers but also facilitates future *in silico* experimentation. Subsequent research should readily be able to challenge the assumptions of the model in light of subsequent data. In practice, this requires provision of free, open-source software environment within which the published analysis can readily be replicated, modified and extended [5, 62].

6. **Appropriate conclusions from observational data**. In the absence of a randomized controlled experiment, the care required to move from a fitted model parameter to a causal claim is well known in linear regression analysis [66]. The same principles apply to nonlinear dynamic metapopulation models: the model structure may be informed by prior scientific knowledge, and the model may statistically explain population-level data, yet observational data cannot readily rule out the possibility of alternative explanations. A model may be called hypothetically causal when it is consistent with scientifically plausible causal mechanisms, but the model fitting process does not itself validate these assumptions—this is a common situation for metapopulation modeling.

In conclusion, the study of metapopulation dynamics will continue to benefit from advances in algorithms, software, and data analysis methodologies. The models should undergo critical scrutiny to delineate their strengths and weaknesses, following evaluation procedures such as we have described in this paper. With due care, these models can unearth limitations in existing knowledge, investigate hypotheses that may extend our knowledge, and furnish us with valuable predictive tools.

Methods

Data. COVID-19 case reports, city population counts, and the time-varying matrix of movement between cities, were taken from [6]. Some erroneous numbers, revealed by our data analysis, were subsequently modified as described in Supplementary Sec. S1.

Model. Our model is a spatiotemporal partially observed Markov process (SpatPOMP) [42]. A partially observed Markov process (POMP) [20, 25] is a stochastic dynamic model comprised of: (i) a latent process having the Markov property that the future is conditionally independent of the past given the present; (ii) a measurement process, describing how the data are modeled as noisy and incomplete observations of the latent process. The SpatPOMP framework adds an extra assumption that the latent process is comprised of a collection of units, each of which has its own latent process and observation process. The latent processes for each unit can be interdependent, but the observations for a given unit are required to depend only on the latent process for that unit. In our context, each city is a unit. General notation for SpatPOMP models is set up in Supplementary Sec. S1, and this notation is subsequently used to define mathematically the specific SpatPOMP models studied in this paper (M_1, M_2, M_5 and M_6). All the models under consideration have an SEAIR structure, as described in Figure 1.

Likelihood evaluation and maximization. The log-likelihood for the SpatPOMP models was calculated using a block particle filter [41, 67]. This log-likelihood was then maximized using an iterated block particle filter [56, 57] Further discussion of these algorithms is in Supplementary Sec. S3. Using this maximization procedure, we constructed confidence intervals by profile likelihood, employing Monte Carlo adjusted profiles [68, 69] to correct for Monte Carlo variability.

Model criticism. A negative binomial autoregressive model was used to provide a non-mechanistic benchmark log-likelihood, as described in Supplementary Sec. S2. This model was also used to construct benchmark conditional log-likelihoods for each separate observation. These, differenced from the corresponding SEAIR log-likelihoods, were used to define anomalies. The anomalies were explored to identify data points which were poorly explained by the model (Supplementary Sec. S7). In preliminary data analysis, these anomalies helped to identify some errors in the data which were subsequently corrected (Supplementary Sec. S1.3).

Software environment. Numerical work was carried out in R [70]. Models and data analysis methodology were developed in an R package, `metapoppkg`, which is additionally designed to assist reproducibility and extendability of our results. Models in `metapoppkg` are implemented using `spatPomp` [67] which provides a general representation of SpatPOMP models extending the POMP model representation in `pomp` [20].

Acknowledgements

This work was supported by National Science Foundation grants DMS-1761603 and DMS-1761612. Portions of this research were conducted with Texas A&M High Performance Research Computing and University of Michigan Advanced Research Computing. We acknowledge discussions with Ethan Romero-Severson.

References

[1] I. Hanski, "Metapopulation dynamics," *Nature*, vol. 396, no. 6706, pp. 41–49, 1998.

[2] D. I. MacKenzie, J. D. Nichols, M. E. Seamans, and R. Gutiérrez, "Modeling species occurrence dynamics with multiple states and imperfect detection," *Ecology*, vol. 90, no. 3, pp. 823–835, 2009.

[3] C. J. E. Metcalf, S. F. Andriamandimby, R. E. Baker, E. E. Glennon, K. Hampson, T. D. Hollingsworth, P. Klepac, and A. Wesolowski, "Challenges in evaluating risks and policy options around endemic establishment or elimination of novel pathogens," *Epidemics*, vol. 37, p. 100507, 2021.

[4] B. Zhang, W. Huang, S. Pei, J. Zeng, W. Shen, D. Wang, G. Wang, T. Chen, L. Yang, P. Cheng, *et al.*, "Mechanisms for the circulation of influenza A (H3N2) in China: A spatiotemporal modelling study," *PLoS Pathogens*, vol. 18, no. 12, p. e1011046, 2022.

[5] J. Wheeler, A. L. Rosengart, Z. Jiang, K. Tan, N. Treutle, and E. L. Ionides, "Informing policy via dynamic models: Cholera in Haiti," *arXiv:2301.08979*, 2023.

[6] R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman, "Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2)," *Science*, vol. 368, no. 6490, pp. 489–493, 2020.

[7] J. T. Wu, K. Leung, and G. M. Leung, "Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study," *The Lancet*, vol. 395, no. 10225, pp. 689–697, 2020.

[8] P. Wang, X. Zheng, and H. Liu, "Simulation and forecasting models of COVID-19 taking into account spatio-temporal dynamic characteristics: A review," *Frontiers in Public Health*, vol. 10, 2022.

[9] K. Prieto, M. V. Chávez-Hernández, and J. P. Romero-Leiton, "On mobility trends analysis of COVID-19 dissemination in Mexico City," *Plos One*, vol. 17, no. 2, p. e0263367, 2022.

[10] J. Cascante-Vega, J. M. Cordovez, and M. Santos-Vega, "Estimating and forecasting the burden and spread of Colombia's SARS-CoV2 first wave," *Scientific Reports*, vol. 12, no. 1, pp. 1–12, 2022.

[11] C. Pizzuti, A. Socievole, B. Prasse, and P. Van Mieghem, "Network-based prediction of COVID-19 epidemic spreading in Italy," *Applied Network Science*, vol. 5, pp. 1–22, 2020.

[12] T. W. Alleman, J. Vergeynst, L. De Visscher, M. Rollier, E. Torfs, I. Nopens, J. M. Baetens, *et al.*, "Assessing the effects of non-pharmaceutical interventions on SARS-CoV-2 transmission in Belgium by means of an extended SEIQRD model and public mobility data," *Epidemics*, vol. 37, p. 100505, 2021.

[13] W. Yang, S. Kandula, M. Huynh, S. K. Greene, G. Van Wye, W. Li, H. T. Chan, E. McGibbon, A. Yeung, D. Olson, A. Fine, and J. Shaman, "Estimating the infection-fatality risk of SARS-CoV-2 in New York City during the spring 2020 pandemic wave: A model-based analysis," *The Lancet Infectious Diseases*, vol. 21, no. 2, pp. 203–212, 2021.

[14] S. Engebretsen, A. Diz-Lois Palomares, G. Rø, A. B. Kristoffersen, J. C. Lindstrøm, K. Engø-Monsen, M. Kamineni, L. Y. Hin Chan, Ø. Dale, J. E. Midtbø, K. L. Stenerud, F. Di Ruscio, R. White, A. Frigessi, and B. F. de Blasio, "A real-time regional model for COVID-19: Probabilistic situational awareness and forecasting," *PLoS Computational Biology*, vol. 19, no. 1, p. e1010860, 2023.

[15] O. N. Bjørnstad and B. T. Grenfell, "Noisy clockwork: Time series analysis of population fluctuations in animals," *Science*, vol. 293, pp. 638–643, 2001.

[16] E. L. Ionides, C. Bretó, and A. A. King, "Inference for nonlinear dynamical systems," *Proceedings of the National Academy of Sciences of the USA*, vol. 103, pp. 18438–18443, 2006.

[17] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. Stumpf, "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems," *Journal of the Royal Society Interface*, vol. 6, pp. 187–202, 2009.

[18] C. Andrieu, A. Doucet, and R. Holenstein, "Particle Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, vol. 72, pp. 269–342, 2010.

[19] E. L. Ionides, D. Nguyen, Y. Atchadé, S. Stoev, and A. A. King, "Inference for dynamic and latent variable models via iterated, perturbed Bayes maps," *Proceedings of the National Academy of Sciences of the USA*, vol. 112, no. 3, pp. 719—-724, 2015.

[20] A. A. King, D. Nguyen, and E. L. Ionides, "Statistical inference for partially observed Markov processes via the R package pomp," *Journal of Statistical Software*, vol. 69, pp. 1–43, 2016.

[21] K. Kristensen, A. Nielsen, C. W. Berg, H. Skaug, and B. M. Bell, "TMB: Automatic differentiation and Laplace approximation," *Journal of Statistical Software*, vol. 70, no. 5, 2016.

[22] P. de Valpine, D. Turek, C. J. Paciorek, C. Anderson-Bergman, D. Temple Lang, and R. Bodik, "Programming with models: Writing statistical algorithms for general model structures with NIMBLE," *Journal of Computational and Graphical Statistics*, vol. 26, no. 2, pp. 403–413, 2017.

[23] T. Bengtsson, P. Bickel, and B. 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* (T. Speed and D. Nolan, eds.), pp. 316–334, Beachwood, OH: Institute of Mathematical Statistics, 2008.

[24] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[25] C. Bretó, D. He, E. L. Ionides, and A. A. King, "Time series analysis via mechanistic models," *Annals of Applied Statistics*, vol. 3, pp. 319–348, 2009.

[26] A. A. King, M. Domenech de Celle, F. M. G. Magpantay, and P. Rohani, "Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola," *Proceedings of the Royal Society of London, Series B*, vol. 282, p. 20150347, 2015.

[27] S. Funk and A. A. King, "Choices and trade-offs in inference with infectious disease models," *Epidemics*, vol. 30, p. 100383, 2020.

[28] M. Auger-Méthé, K. Newman, D. Cole, F. Empacher, R. Gryba, A. A. King, V. Leos-Barajas, J. Mills Flemming, A. Nielsen, G. Petris, and L. Thomas, "A guide to state-space modeling of ecological time series," *Ecological Monographs*, vol. 91, no. 4, p. e01470, 2021.

[29] O. Cappé, E. Moulines, and T. Rydén, *Inference in Hidden Markov Models*. New York: Springer, 2005.

[30] M. Whitehouse, N. Whiteley, and L. Rimella, "Consistent and fast inference in compartmental models of epidemics using Poisson Approximate Likelihoods," *Journal of the Royal Statistical Society, Series B*, vol. To appear, 2023.

[31] A. J. Conlan, T. J. McKinley, K. Karolemeas, E. B. Pollock, A. V. Goodchild, A. P. Mitchell, C. P. Birch, R. S. Clifton-Hadley, and J. L. Wood, "Estimating the hidden burden of bovine tuberculosis in Great Britain," *PLoS Computational Biology*, vol. 8, p. e1002730, 2012.

[32] M. Fasiolo, N. Pya, and S. N. Wood, "A comparison of inferential methods for highly nonlinear state space models in ecology and epidemiology," *Statistical Science*, vol. 31, no. 1, pp. 96–118, 2016.

[33] O. N. Bjørnstad, B. F. Finkenstädt, and B. T. Grenfell, "Dynamics of measles epidemics: Estimating scaling of transmission rates using a time series SIR model," *Ecological Monographs*, vol. 72, no. 2, pp. 169–184, 2002.

[34] Y. Xia, O. N. Bjørnstad, and B. T. Grenfell, "Measles metapopulation dynamics: A gravity model for epidemiological coupling and dynamics," *American Naturalist*, vol. 164, no. 2, pp. 267–281, 2004.

[35] G. Evensen, *Data assimilation: The ensemble Kalman filter*. Springer Science & Business Media, 2009.

[36] G. Evensen, F. C. Vossepoel, and P. J. van Leeuwen, *Data assimilation fundamentals: A unified formulation of the state and parameter estimation problem*. Springer Nature, 2022.

[37] J. Shaman and A. Karspeck, "Forecasting seasonal outbreaks of influenza," *Proceedings of the National Academy of Sciences of the USA*, vol. 109, pp. 20425–20430, 2012.

[38] W. Yang, A. Karspeck, and J. Shaman, "Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics," *PLoS Computational Biology*, vol. 10, p. e1003583, 2014.

[39] S. C. Kramer, S. Pei, and J. Shaman, "Forecasting influenza in Europe using a metapopulation model incorporating cross-border commuting and air travel," *PLoS Computational Biology*, vol. 16, no. 10, p. e1008233, 2020.

[40] M. Katzfuss, J. R. Stroud, and C. K. Wikle, "Ensemble Kalman methods for high-dimensional hierarchical dynamic space-time models," *Journal of the American Statistical Association*, vol. 115, no. 530, pp. 866–885, 2020.

[41] P. Rebeschini and R. van Handel, "Can local particle filters beat the curse of dimensionality?," *The Annals of Applied Probability*, vol. 25, no. 5, pp. 2809–2866, 2015.

[42] E. L. Ionides, K. Asfaw, J. Park, and A. A. King, "Bagged filters for partially observed interacting systems," *Journal of the American Statistical Association*, vol. 118, no. 542, pp. 1078–1089, 2023.

[43] R. McCabe and C. A. Donnelly, "Disease transmission and control modelling at the science–policy interface," *Interface Focus*, vol. 11, no. 6, p. 20210013, 2021.

[44] A. Saltelli, G. Bammer, I. Bruno, E. Charters, M. Di Fiore, E. Didier, W. Nelson Espeland, J. Kay, S. Lo Piano, D. Mayo, R. Pielke, T. Portaluri, T. M. Porter, A. Puy, I. Rafols, J. R. Ravetz, E. Reinert, D. Sarewitz, P. B. Stark, A. Stirling, J. van der Sluijs, and P. Vineis, "Five ways to ensure that models serve society: a manifesto," *Nature*, vol. 582, pp. 428–484, 2020.

[45] E. T. Lofgren, M. E. Halloran, C. M. Rivers, J. M. Drake, T. C. Porco, B. Lewis, W. Yang, A. Vespignani, J. Shaman, J. N. Eisenberg, M. C. Eisenbergg, S. V. Marathea, Madhav andScarpinoi, K. A. Alexander, R. Meza, J. M. Ferrari, Matthew J. andHyman, L. A. Meyers, and S. Eubank, "Mathematical models: A key tool for outbreak response," *Proceedings of the National Academy of Sciences of the USA*, vol. 111, no. 51, pp. 18095–18096, 2014.

[46] M. U. Kraemer, C.-H. Yang, B. Gutierrez, C.-H. Wu, B. Klein, D. M. Pigott, O. C.-. D. W. Group, L. Du Plessis, N. R. Faria, R. Li, *et al.*, "The effect of human mobility and control measures on the covid-19 epidemic in china," *Science*, vol. 368, no. 6490, pp. 493–497, 2020.

[47] T. S. Brett, S. Bansal, and P. Rohani, "Charting the spatial dynamics of early SARS-CoV-2 transmission in Washington state," *PLoS Computational Biology*, vol. 19, no. 6, p. e1011263, 2023.

[48] W. Yang, M. Lipsitch, and J. Shaman, "Inference of seasonal and pandemic influenza transmission dynamics," *Proceedings of the National Academy of Sciences of the USA*, vol. 112, no. 9, pp. 2723–2728, 2015.

[49] S. Pei, S. Kandula, W. Yang, and J. Shaman, "Forecasting the spatial transmission of influenza in the United States," *Proceedings of the National Academy of Sciences*, vol. 115, no. 11, pp. 2752–2757, 2018.

[50] K. P. Burnham and D. R. Anderson, *Model Selection and Inference: A Practical Information-theoretic Approach*. New York: Springer-Verlag, 2nd ed., 2002.

[51] Y. Pawitan, *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press, 2001.

[52] D. He, E. L. Ionides, and A. A. King, "Plug-and-play inference for disease dynamics: Measles in large and small towns as a case study," *Journal of the Royal Society Interface*, vol. 7, pp. 271–283, 2010.

[53] T. Stocks, T. Britton, and M. Höhle, "Model selection and parameter estimation for dynamic epidemic models via iterated filtering: Application to rotavirus in Germany," *Biostatistics*, vol. 21, no. 3, pp. 400–416, 2020.

[54] M. Keeling and P. Rohani, *Modeling Infectious Diseases in Humans and Animals*. Princeton, NJ: Princeton University Press, 2009.

[55] C. Bretó and E. L. Ionides, "Compound Markov counting processes and their applications to modeling infinitesimally over-dispersed systems," *Stochastic Processes and their Applications*, vol. 121, pp. 2571–2591, 2011.

[56] E. L. Ionides, N. Ning, and J. Wheeler, "An iterated block particle filter for inference on coupled dynamic systems with shared and unit-specific parameters," *Statistica Sinica*, pp. pre–published online, 2022.

[57] N. Ning and E. L. Ionides, "Iterated block particle filter for high-dimensional parameter learning: Beating the curse of dimensionality," *Journal of Machine Learning Research*, vol. 24, pp. 1–76, 2023.

[58] J. M. Brauner, S. Mindermann, M. Sharma, D. Johnston, J. Salvatier, T. Gavenčiak, A. B. Stephenson, G. Leech, G. Altman, V. Mikulik, *et al.*, "Inferring the effectiveness of government interventions against COVID-19," *Science*, vol. 371, no. 6531, p. eabd9338, 2021.

[59] M. Park, A. R. Cook, J. T. Lim, Y. Sun, and B. L. Dickens, "A systematic review of COVID-19 epidemiology based on current evidence," *Journal of Clinical Medicine*, vol. 9, no. 4, p. 967, 2020.

[60] J. M. Read, J. R. Bridgen, D. A. Cummings, A. Ho, and C. P. Jewell, "Novel coronavirus 2019-nCoV (COVID-19): Early estimation of epidemiological parameters and epidemic size estimates," *Philosophical Transactions of the Royal Society B*, vol. 376, no. 1829, p. 20200265, 2021.

[61] R. Subramanian, Q. He, and M. Pascual, "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 of the USA*, vol. 118, no. 9, 2021.

[62] R. Gentleman and D. Temple Lang, "Statistical analyses and reproducible research," *Journal of Computational and Graphical Statistics*, vol. 16, no. 1, pp. 1–23, 2007.

[63] J. P. Ioannidis, S. Cripps, and M. A. Tanner, "Forecasting for COVID-19 has failed," *International journal of forecasting*, vol. 38, no. 2, pp. 423–438, 2022.

[64] B. Yu and K. Kumbier, "Veridical data science," *Proceedings of the National Academy of Sciences of the USA*, vol. 117, pp. 3920–3929, 2020.

[65] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.

[66] J. J. Faraway, *Linear models with R*. CRC press, 2014.

[67] K. Asfaw, J. Park, A. A. King, and E. L. Ionides, "Statistical inference for spatiotemporal partially observed Markov processes via the R package spatPomp," *arXiv:2101.01157v3*, 2023.

[68] 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*, vol. 14, pp. 1–10, 2017.

[69] N. Ning, E. L. Ionides, and Y. Ritov, "Scalable Monte Carlo inference and rescaled local asymptotic normality," *Bernoulli*, vol. 27, pp. 2532–2555, 2021.

[70] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.