Submitted to the Annals of Applied Statistics

INFORMING POLICY VIA DYNAMIC MODELS: ELIMINATING CHOLERA IN HAITI

By Jesse Wheeler[1,a], AnnaElaine Rosengart[1,b] Zhuoxun Jiang[1,c], Kevin Hao En Tan[1,d], Noah Treutle[1,e] and Edward Ionides[1,f]

[1]*Statistics Department, University of Michigan,* [a]*jeswheel@umich.edu;* [b]*aelr@umich.edu;* [c]*zhuoxunj@umich.edu;* [d]*kevtan@umich.edu;* [e]*ntreutle@umich.edu;* [f]*ionides@umich.edu*

Public health decisions must be made about when and how to implement interventions to control an infectious disease epidemic. These decisions should be informed by data on the epidemic as well as current understanding about the transmission dynamics. Such decisions can be posed as statistical questions about scientifically motivated dynamic models. Thus, we encounter the methodological task of building credible, data-informed decisions based on stochastic, partially observed, nonlinear dynamic models. This necessitates addressing the tradeoff between biological fidelity and model simplicity, and the reality of misspecification for models at all levels of complexity. As a case study, we consider a cholera epidemic in Haiti. The 2010 introduction of cholera to Haiti led to an extensive outbreak and sustained transmission until it was eliminated in 2019. We study three models developed by expert teams to advise on vaccination policies. We assess methods used for fitting and evaluating these models, leading to recommendations for future studies. Diagnosis of model misspecification and development of alternative models can lead to improved statistical fit, but caution is nevertheless required in drawing policy conclusions based on causal interpretations of the models.

1. Introduction. Regulation of biological populations is a fundamental topic in epidemiology, ecology, fisheries and agriculture. Population dynamics may be nonlinear and stochastic, with the resulting complexities compounded by incomplete understanding of the underlying biological mechanisms and by partial observability of the system variables. Quantitative models for these dynamic systems offer potential for designing effective control measures. Developing and testing such models, and assessing their fitness for guiding policy, is a challenging statistical task. Questions of interest include: What indications should we look for in the data to assess whether the model-based inferences are trustworthy? What diagnostic tests and model variations can and should be considered in the course of the data analysis? What are the possible trade-offs of increasing model complexity, such as the inclusion of interactions across spatial units?

This case study investigates the use of dynamic models and spatiotemporal data to inform a policy decision in the context of the cholera outbreak in Haiti, which started in 2010. We build on a multi-group modeling exercise by Lee et al. (2020a) in which four expert modeling teams developed models to the same dataset with the goal of comparing conclusions on the feasibility of eliminating cholera by a vaccination campaign. Model 1 is stochastic and describes cholera at the national level; Model 2 is deterministic with spatial structure, and includes transmission via contaminated water; Model 3 is stochastic with spatial structure, and accounts for measured rainfall. Model 4 has an agent-based construction, featuring considerable mechanistic detail but limited

Keywords and phrases: Partially observed Markov process, Hidden Markov model, infectious disease, cholera, sequential Monte Carlo.

ability to calibrate these details to data. The strengths and weaknesses of the agent-based modeling approach (Tracy, Cerdá and Keyes, 2018) are outside the scope of this article, and we focus on Models 1–3.

The four independent teams were given the task of estimating the potential effect of prospective oral cholera vaccine (OCV) programs. While OCV is accepted as a safe and effective tool for controlling the spread of cholera, the global stockpile of OCV doses remains limited (Pezzoli, 2020). Advances in OCV technology and vaccine availability, however, raised the possibility of planning a national vaccination program (Lee et al., 2020a). In the study, certain data were shared between the groups, including demography and vaccination history; vaccine efficacy was also fixed at a shared value between groups. Beyond this, the groups made autonomous decisions on what to include and exclude from their models; this autonomy reduced the possible effect that assumptions about the dynamic system may have on the final conclusion of the study. Despite this autonomy, and largely adhering to existing guidelines on creating models to inform policy (Behrend et al., 2020; Saltelli et al., 2020), the consensus across the four models was that an extensive nationwide vaccination campaign would be necessary to eliminate cholera from Haiti. This conclusion is inconsistent with the fact that there have been no confirmed cases since February, 2019 (Trevisin et al., 2022), despite the lack of a concentrated vaccination effort.

The discrepancy between the model based conclusions of Lee et al. (2020a) and the subsequent elimination of cholera cholera has been debated (Francois, 2020; Rebaudet, Gaudart and Piarroux, 2020; Henrys et al., 2020; Lee et al., 2020b). Suggested origins of this discrepancy include the use of unrealistic models (Rebaudet, Gaudart and Piarroux, 2020) and having unrealistic criteria for cholera elimination (Henrys et al., 2020). We find a more nuanced conclusion: attention to methodological details in model fitting, diagnosis and forecasting can improve each of the proposed model's ability to quantitatively describe observed data. These improvements results in forecasts that are more consistent with the observed outcome, without requiring major changes to the model structures. Based on this retrospective analysis, we offer suggestions on fitting mechanistic models to dynamic systems for future studies.

While this work is focused on the models proposed in Lee et al. (2020a), we believe that our suggestions are relevant to a larger class of problems. In order to determine the extent to which this work generalizes, we preformed a literature review by searching PubMed with keywords: Haiti, cholera, and model. The search resulted in 66 papers, of which 32 used dynamic models to describe the cholera epidemic in Haiti. This literature review confirmed that close attention to details in fitting, diagnosing, and forecasting is of great importance when modeling an infectious disease outbreak. The wide relevance of our findings in these papers—supplemented with studies of other epidemiological and dynamic systems—suggests that this work may also be relevant to a much broader class of problems.

We proceed by introducing Models 1–3 in Sec. 2; in Sec. 3, we present a methodological approach to examining and refining these models, and then use improved model fits to project cholera incidence in Haiti under various vaccination scenarios. This is followed by a consideration of the robustness of model based policy recommendations in Sec. 4, and a discussion in Sec. 5.

2. Mechanistic models for cholera in Haiti. Models that focus on learning relationships between variables in a dataset are called *associative*, whereas models that incorporate a known scientific property of the system are called *causal* or *mechanistic*. The danger in using forecasting techniques which rely on associative models to predict

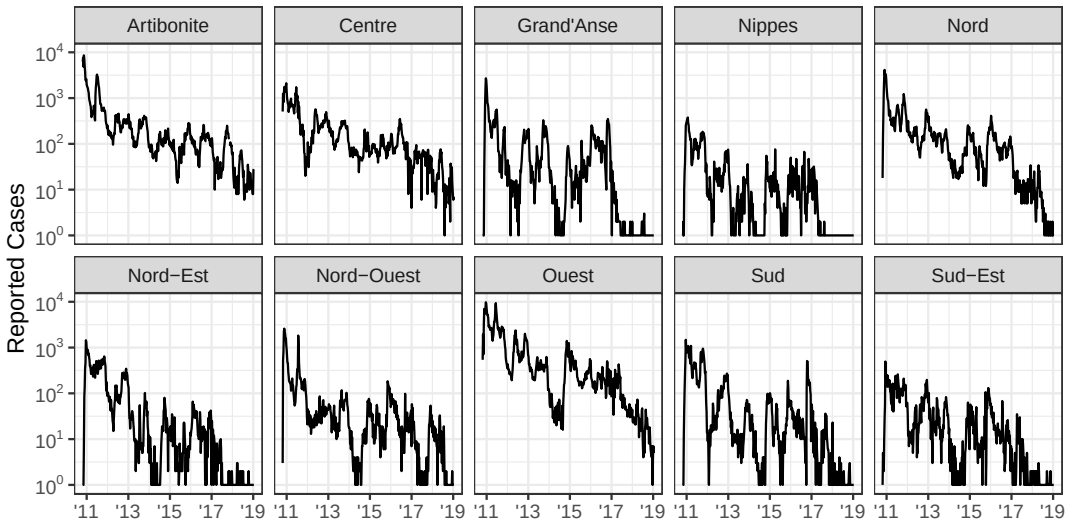

Fig 1. Reported Cholera cases in the outbreak in Haiti from 2010-2019.

the consequence of interventions is called the Lucas critique in an econometric context. Lucas et al. (1976) pointed out that it is naive to predict the effects of an intervention on a given system based entirely on historical associations. To successfully predict the effect of an intervention, a model should therefore both provide a quantitative explanation of existing data and should have a causal interpretation: a manipulation of the system should correspond quantitatively with the corresponding change to the model. This motivates the development of mechanistic statistical models, which provides a statistical fit to the available data while also supporting a causal intepretation.

The deliberate limitation of coordination between the groups of Lee et al. (2020a) allows us to treat the models as fairly independently developed expert approaches to understanding cholera transmission. However, it led to differences in notation, and in subsets of the data chosen for analysis, that hinder direct comparison. Here, we have put all three models into a common notational framework. Translations back to the original notation of Lee et al. (2020a) are given in Table S-1.

Each model describes the cholera dynamics using a latent state vector $\boldsymbol{X}^{(m)}(t)$ for each continuous time-point $t \in \mathcal{T}$, where $\mathcal{T}$ is the set of all time-points and $m \in \{1, 2, 3\}$ indexes the model. The observation at time t_n is modeled by a response vector $\boldsymbol{Y}_n^{(m)}$. The latent state vector $\boldsymbol{X}^{(m)}(t)$ consists of individuals labeled as susceptible (S), infected (I), asymptomatically infected (A), vaccinated (V), and recovered (R), with various sub-divisions sometimes considered in each model m. Models 2 and 3 have metapopulation structure, meaning that each individual is a member of a spatial unit, denoted by a subscript $u \in 1{:}U$. Here, the spatial units are the $U = 10$ Haitian administrative départements (henceforth anglicized as departments).

In the following subsections, complete descriptions of Models 1–3 are provided. While the model description is scientifically critical, as well as necessary for transparency and reproducibility, the model details are not essential to our methodological discussions of how to diagnose and address model misspecification with the purpose of informing policy. A first-time reader may choose to skim through the rest of this section, and return later.

2.1. *Model 1.* $\boldsymbol{X}^{(1)}(t) = \big(S_z(t), E_z(t), I_z(t), A_z(t), R_z(t), z \in 0:Z\big)$ describes susceptible, latent (exposed), infected (and symptomatic), asymptomatic, and recovered individuals in vaccine cohort z. Here, $z = 0$ corresponds to unvaccinated individuals, and $z \in 1{:}Z$ describes hypothetical vaccination programs. The force of infection is

$$\lambda(t) = \Big(\sum_{z=0}^{Z} I_z(t) + \epsilon \sum_{z=0}^{Z} A_z(t)\Big)^{\nu} \frac{d\Gamma(t)}{dt} \beta(t)/N \tag{1}$$

where $\beta(t)$ is a periodic cubic spline representation of seasonality, given in terms of a B-spline basis $\{s_j(t), j \in 1{:}6\}$ and parameters $\beta_{1:6}$ as

$$\log \beta(t) = \sum_{j=1}^{6} \beta_j s_j(t). \tag{2}$$

The process noise $d\Gamma(t)/dt$ is multiplicative Gamma-distributed white noise, with infinitesimal variance parameter σ^2_{proc}. Lee et al. (2020a) included process noise in Model 3 but not in Model 1, i.e., they fixed $\sigma^2_{\text{proc}} = 0$. Gamma white noise in the transmission rate gives rise to an over-dispersed latent Markov process (Bretó and Ionides, 2011) which has been found to improve the statistical fit of disease transmission models (Stocks, Britton and Höhle, 2020; He, Ionides and King, 2010).

Per-capita transition rates are given in Equations 3-10:

$$\mu_{S_z E_z} = \lambda(t) \tag{3}$$

$$\mu_{E_z I_z} = \mu_{EI}\big(1 - f_z(t)\big) \tag{4}$$

$$\mu_{E_z A_z} = \mu_{EI}\, f_z(t) \tag{5}$$

$$\mu_{I_z R_z} = \mu_{A_z R_z} = \mu_{IR} \tag{6}$$

$$\mu_{R_z S_z} = \mu_{RS} \tag{7}$$

$$\mu_{S_0 S_z} = \mu_{E_0 E_z} = \mu_{I_0 I_z} = \mu_{A_0 A_z} = \mu_{R_0 R_z} = \eta_z(t) \tag{8}$$

$$\mu_{S_z D} = \mu_{E_z D} = \mu_{I_z D} = \mu_{A_z D} = \mu_{R_z D} = \delta \tag{9}$$

$$\mu_{D S_0} = \mu_S \tag{10}$$

where $z \in 0{:}Z$. Here, μ_{AB} is a transition rate from compartment A to B. We have an additional demographic source and sink compartment D modeling entry into the study population due to birth or immigration, and exit from the study population due to death or immigration. Thus, μ_{AD} is a rate of exiting the study population from compartment A and μ_{DB} is a rate of entering the study population into compartment B.

In Model 1, the advantage afforded to vaccinated individuals is an increased probability that an infection is asymptomatic. Conditional on infection status, vaccinated individuals are also less infectious than their non-vaccinated counterparts by a rate of $\epsilon = 0.05$ in Eq. (1). In (5) and (4) the asymptomatic ratio for non-vaccinated individuals is set $f_0(t) = 0$, so that the asymptomatic route is reserved for vaccinated individuals. For $z \in 1{:}Z$, the vaccination cohort z is assigned a time τ_z, and we take $f_z(t) = c\,\theta^*(t - \tau_z)$ where $\theta^*(t)$ is efficacy at time t since vaccination for adults, taken from Lee et al. (2020a), Table S4, and $c = \big(1 - (1 - 0.4688) \times 0.11\big)$ is a correction to allow for reduced efficacy in the 11% of the population aged under 5 years. Single and double vaccine doses were modeled by changing the waning of protection; protection was assumed to be equal between single and double dose until 52 weeks after vaccination, at which point the single dose becomes ineffective.

2.2. *Model 2.* Susceptible individuals are in compartments $S_{uz}(t)$, where $u \in 1{:}U$ corresponds to the $U = 10$ departments, and $z \in 0{:}4$ describes vaccination status:

$z = 0$: Unvaccinated or waned vaccination protection.
$z = 1$: One dose at age under five years.
$z = 2$: Two doses at age under five years.
$z = 3$: One dose at age over five years.
$z = 4$: Two doses at age over five years.

Individuals can progress to a latent infection E_{uz} followed by symptomatic infection I_{uz} with recovery to R_{uz} or asymptomatic infection A_{uz} with recovery to R^A_{uz}. The force of infection depends on both direct transmission and an aquatic reservoir, $W_u(t)$, and is given by

$$\lambda_u(t) = 0.5\big(1 + a\cos(2\pi t + \phi)\big)\frac{\beta_W W_u(t)}{W_{\text{sat}} + W_u(t)} + \beta\left\{\sum_{z=0}^{4} I_{uz}(t) + \epsilon\sum_{z=0}^{4} A_{uz}(t)\right\} \tag{11}$$

The latent state is therefore described by the vector $\boldsymbol{X}^{(2)}(t) = \big(S_{uz}(t), E_{uz}(t), I_{uz}(t), A_{uz}(t), R_{uz}(t), R^A_{uz}(t), W_u, u \in 1{:}U, z \in 0{:}4\big)$. The cosine term in Eq. (11) accounts for annual seasonality, with a phase parameter ϕ. The implementation of Model 2 in Lee et al. (2020a) fixes $\phi = 0$.

Individuals move from department u to v at rate T_{uv}, and aquatic cholera moves at rate T^W_{uv}. The nonzero transition rates are

$$\mu_{S_{uz}E_{uz}} = \theta_z\,\lambda \tag{12}$$

$$\mu_{E_{uz}I_{uz}} = f\mu_{EI}, \quad \mu_{E_{uz}A_{uz}} = (1-f)\mu_{EI} \tag{13}$$

$$\mu_{I_{uz}R_{uz}} = \mu_{A_{uz}R^A_{uz}} = \mu_{IR} \tag{14}$$

$$\mu_{R_{uz}S_{uz}} = \mu_{R^A_{uz}S_{uz}} = \mu_{RS} \tag{15}$$

$$\mu_{S_{uz}S_{vz}} = \mu_{E_{uz}E_{vz}} = \mu_{I_{uz}I_{vz}} = \mu_{A_{uz}A_{vz}} = \mu_{R_{uz}R_{vz}} = \mu_{R^A_{uz}R^A_{vz}} = T_{uv} \tag{16}$$

$$\mu_{S_{u1}S_{u0}} = \mu_{S_{u3}S_{u0}} = \omega_1 \tag{17}$$

$$\mu_{S_{u2}S_{u0}} = \mu_{S_{u4}S_{u0}} = \omega_2 \tag{18}$$

$$\mu_{DW_u} = \mu_W\left\{\sum_{z=0}^{4} I_{uz}(t) + \epsilon_W\sum_{z=0}^{4} A_{uz}(t)\right\} \tag{19}$$

$$\mu_{W_uD} = \delta_W \tag{20}$$

$$\mu_{W_uW_v} = w_r T^W_{uv} \tag{21}$$

In (16) the spatial coupling is specified by a gravity model,

$$T_{uv} = v_{\text{rate}} \times \frac{\text{Pop}_u\text{Pop}_v}{D^2_{uv}}, \tag{22}$$

where Pop_u is the mean population for department u, D_{uv} is a distance measure estimating average road distance between rancomly chosen members of each population, and $v_{\text{rate}} = 10^{-12}$ was treated as a fixed constant. In (21), T^W_{uv} is a measure of river flow between departments. The unit of $W_u(t)$ is cells per ml, with dose response modeled via a saturation constant of W_{sat} in (11).

2.3. *Model 3.* The latent state is described as $\boldsymbol{X}^{(3)}(t) = \big(S_{uz}(t), I_{uz}(t), A_{uz}(t), R_{uzk}(t), W_u(t), u \in 0{:}U, z \in 0{:}4, k \in 1{:}3\big)$. Here, $z = 0$ corresponds to unvaccinated, $z = 2j - 1$ corresponds to a single dose on the jth vaccination campaign in unit u and $z = 2j$ corresponds to receiving two doses on the jth vaccination campaign. $k \in 1{:}3$

models non-exponential duration in the recovered class before waning of immunity. The force of infection is

$$\lambda_u(t) = \beta_{W_u} \frac{W_u(t)}{1 + W_u(t)} + \beta_u \sum_{v \neq u} \big(I_{v0}(t) + \epsilon A_{v0}(t)\big) \tag{23}$$

$$\mu_{S_{uz}I_{uz}} = f\,\lambda_u\big(1 - \eta_{uz}(t)\big)\,d\Gamma/dt \tag{24}$$

$$\mu_{S_{uz}A_{uz}} = (1-f)\,\lambda_u\big(1 - \eta_{uz}(t)\big)\,d\Gamma/dt \tag{25}$$

$$\mu_{I_{uz}R_{uz1}} = \mu_{A_{uz}R_{uz1}} = \mu_{IR} \tag{26}$$

$$\mu_{I_{uz}S_{u0}} = \delta + \delta_C \tag{27}$$

$$\mu_{A_{uz}S_{u0}} = \delta \tag{28}$$

$$\mu_{R_{uz1}R_{uz2}} = \mu_{R_{uz2}R_{uz3}} = 3\mu_{RS} \tag{29}$$

$$\mu_{R_{uzk}S_{u0}} = \delta + 3\mu_{RS}\,\mathbf{1}_{\{k=3\}} \tag{30}$$

$$\mu_{DW_u} = \big[1 + a\big(J(t)\big)^r\big] D_i\,\mu_W\big[I_{u0}(t) + \epsilon_W A_{u0}(t)\big] \tag{31}$$

$$\mu_{W_uD} = \delta_W \tag{32}$$

As with Model 1, $d\Gamma_u(t)/dt$ is multiplicative Gamma-distributed white noise in (24) and (25). In (31), $J_u(t)$ is a dimensionless measurement of precipitation that has been standardized by dividing the observed rainfall at time t by the maximum recorded rainfall in department u during the epidemic, and D_u is the average population density. Demographic stochasticity is accounted for by modeling non-cholera related death rate δ in each compartment, along with an additional death rate δ_C in (27) to account for cholera induced deaths among infected individuals. All deaths are balanced by births into the susceptible compartment in (28) and (30), thereby maintaining constant population in each department.

3. Statistical Analysis. We consider model fitting (Sec. 3.1) followed by diagnostic investigations (Sec. 3.2), forecasting (Sec. 3.3) and interpreting model fit (Sec. 3.4).

3.1. *Model Fitting.* Proposed mechanistic structures form a family of statistical models indexed by a parameter vector θ. Different values of θ can result in qualitative differences in the predicted behavior of the system. The complex nature of biological systems necessitates a search for modeling assumptions that combine insightful simplicity with fidelity to biological reality. For example, many models commonly used in epidemiology are motivated by reasoning about a homogeneous mixing population (Bansal, Grenfell and Meyers, 2007) which is simultaneously an avenue for powerful simplification and a source of model misspecification. Other common considerations include whether the proposed model should be stochastic or deterministic; whether the model should have change points in parameter values or should otherwise make adjustments for changes through time in the dynamic system; and whether the proposed model should include any spatial heterogeneity at a scale permissible by the observed data. In addition, elements of θ can either be chosen as constants, based on scientific reasoning and previous knowledge, or calibrated to observed data. Suitable methodology for calibrating model parameters may depend on other modeling decisions.

Each of the three models considered in this study describes cholera dynamics via unobservable states that evolve dynamically with time. Despite their similarities, these models represent a diverse selection of possible modeling assumptions: namely, the use of stochastic (Models 1 and 2) or deterministic (Model 2) equations; spatially-heterogeneous meta-population (Models 2 and 3) or spatially-aggregated (Model 1)

structure; and the use of covariates (Model 3) versus mathematical equations (Models 1 and 2) to describe a seasonal mechanism. All these structures can be described in the framework of partially observed Markov process (POMP) models, with the understanding that the deterministic Model 2 is a degenerate case of a stochastic model. In the following subsections we describe our approach to fitting these mechanistic models.

3.1.1. *Model 1.* One approach to modeling a dynamic system is through probabilistic models. With such a model, the existence of a joint density $f_{\boldsymbol{X}_{0:N}^{(m)},\boldsymbol{Y}_{1:N}^{(m)}}$ is supposed, with $\boldsymbol{X}_{0:N}^{(m)}$ denoting the unobservable Markov process of model m at times $0:N = \{0,1,\ldots,N\}$, and $\boldsymbol{Y}_{1:N}^{(m)}$ denoting the observable process of the system at times $1:N$. Under this framework, the observed data $y_{1:N}^*$—along with some unobserved value $\boldsymbol{X}_{0:N}^{(m)} = x_{0:N}^*$—are assumed to be a single realization of the model $(x_{0:N}^*, y_{1:N}^*) \sim f_{\boldsymbol{X}_{0:N}^{(m)},\boldsymbol{Y}_{1:N}^{(m)}}(x_{0:N}, y_{1:N}; \theta)$, where θ is a parameter vector that indexes the model. This modeling framework results in several advantages, including the ability to account for variability present in the system. Furthermore, because each draw from the joint distribution represents a potential outcome of the dynamic system, best/worst case scenarios under the assumptions of the model can be easily obtained via simulation.

Several algorithms exist, both frequentist and Bayesian, that can be used to obtain estimates of the parameters for this class of models. In order to retain the ability to propose models that are scientifically meaningful rather than only those that are simply statistically convenient, we restrict ourselves to parameter estimation techniques that have the plug-and-play property, which is that the fitting procedure only requires the ability to simulate the latent process instead of evaluating transition densities (Bretó et al., 2009; He, Ionides and King, 2010). Plug-and-play algorithms include Bayesian approaches like ABC and PMCMC (Toni et al., 2009; Andrieu, Doucet and Holenstein, 2010), but here we use frequentist methods to maximize model likelihoods. To our knowledge, the only plug-and-play frequentist methods that can maximize the complete model likelihood are iterated filtering algorithms, which modify the well-known particle filter (Arulampalam et al., 2002) by performing a random walk for each parameter and particle. These perturbations are carried out iteratively over multiple filtering operations, using the collection of parameters from the previous filtering pass as the parameter initialization for the next iteration, and decreasing the random walk variance at each step.

The ability to maximize the likelihood allows for likelihood-based inference, like performing statistical tests for potential improvements to the model. We demonstrate this capability by proposing a linear trend ζ in transmission in Eq. (2):

$$\log \beta(t) = \sum_{j=1}^{6} \beta_s s_j(t) + \zeta \bar{t} \tag{33}$$

Where $\bar{t}$ is the linear mapping $\bar{t} : [0, N] \to [-1, 1]$ of the time t. The proposal of a linear trend in transmission is a result of observing an apparent decrease in reported cholera infections from 2012-2019 in Fig. 1. While several factors may contribute to this decrease, one explanation is that case-area targeted interventions (CATIs), which included education sessions, increased monitoring, household decontamination, soap distribution, and water chlorination in infected areas (Rebaudet et al., 2019), may have greatly reduced cholera transmission (Rebaudet et al., 2021).

We perform a statistical test to determine whether or not the data indicate the presence of a linear trend in transmissibility. To do this, we perform a profile-likelihood

search on the parameter ζ and obtain a confidence interval via a Monte Carlo Adjusted Profile (MCAP) (Ionides et al., 2017). Model 1 was implemented in Lee et al. (2020a) by fitting two distinct phases: an epidemic phase from October 2010 through March 2015, and an endemic phase from March 2015 onward. We similarly allow the re-estimation of process and measurement overdispersion parameters (σ^2_{proc} and ψ), and require that the latent Markov process $X(t)$ carry over from one phase into the next. The resulting confidence interval for ζ is $(-0.085, -0.005)$, with the full results displayed in Fig. 2. These results are suggestive that the inclusion of a trend in transmission rate improves the quantitative ability of Model 1 to describe the observed data. The reported results for Model 1 in the remainder of this article were obtained with the inclusion of the parameter ζ.

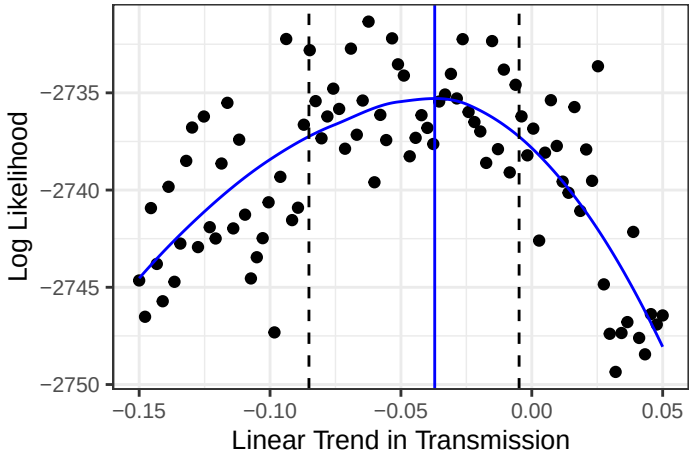

FIG 2. Monte Carlo adjusted profile of ζ. The blue curve is the profile, the blue line indicates the MLE, and the dashed lines indicate the confidence interval.

Model 1 is implemented using the `pomp` package King et al. (2009), relying heavily on the source code provided by Lee et al. (2020a), who also used the MIF2 algorithm to estimate θ. Aside from the introduction of the parameter ζ and the multiplicative Gamma-distributed white noise $d\Gamma(t)/dt$ in the process model (Eq. 1), the only change made that could increase the likelihood of the model was using a substantially increased computational effort to maximize the likelihood. While flexible, plug-and-play methods like MIF2 require a large number of computations, and the theoretic ability to maximize the likelihood depends on asymptotics in both the number of particles used and the number of MIF iterations. The large difference in likelihoods that were obtain by these two approaches (see Table 1) highlights the importance of carefully determining the necessary computational effort needed to maximize model likelihoods and acting accordingly.

3.1.2. *Model 2.* Model 2 is a deterministic model, which are degenerate probabilistic models that contain no randomness in the dynamic process or measurement. By combining a deterministic process model with a simple Gaussian measurement model, Model 2 reduces model fitting to a least squares calculation over parameters in a set of differential equations. Deterministic compartment models have a long history in the field of infectious disease epidemiology (Kermack and McKendrick, 1927; Brauer, 2017;

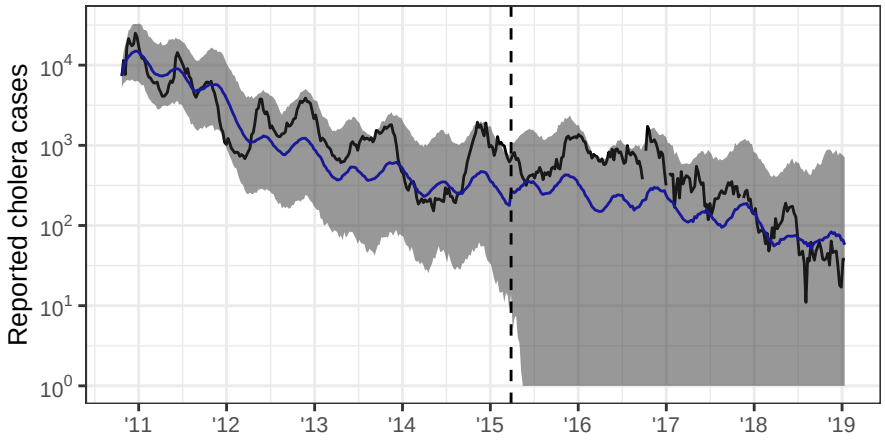

FIG 3. Simulations from Model 1 compared to reported cholera cases. The black curve is observed data, the blue curve is median of 500 simulations from the fitted model, and the vertical dashed line represents break-point when parameters are refit.

Varghese et al., 2021), and can be justified by asymptotic considerations in a large-population limit (Dadlani et al., 2020; Ndii and Supriatna, 2017).

Lee et al. (2020a) fit two versions of model 2 based on a presupposed change in cholera transmission from a epidemic phase to endemic phase that occurred in March, 2014. The inclusion of a change-point in model states and parameters increases the flexibility of the model and hence the ability to fit the observed data. The increase in model flexibility, however, results in hidden states that are inconsistent between model phases. The inclusion of a model break-point in Lee et al. (2020a) is perhaps due to a challenging feature of fitting a deterministic model via least squares: discrepancies between model trajectories and observed case counts in highly infectious periods of a disease outbreak will result in greater penalty than the discrepancies between model trajectories and observed case counts in times of relatively low infectiousness, resulting in a bias towards accurately describing periods of high infectiousness. This issue is particularly troublesome for this case study: the inability to accurately fit times of low infectiousness may result in poor model forecasts, as few cases of cholera were observed the last few years of the epidemic.

To combat this issue, we fit the model to log-transformed case counts, where the difference between periods of high and low infectiousness is less drastic. Other solutions include changing the measurement model to include overdispersion, as was done in Models 1 and 3. Here we chose to fit the model to transformed case counts rather than adding overdispersion to the measurement model with the goal of making the fewest changes to the model in Lee et al. (2020a) as possible. In practice, however, the inclusion of overdispersion in the measurement process may be preferred.

We implemented this model using the `spatPomp` `R` package (Asfaw, Ionides and King, 2021). The model was then fit using the subplex algorithm, implemented in the `subplex` package (King and Rowan, 2020). A comparison of the trajectory of the fitted model to the data is given in Fig 6.

3.1.3. *Model 3.* Both the latent and observable processes in Model 3 can be be factored into department specific processes which interact with each other. The decision to model a dynamic system via metapopulation models versus a model aggregated to a larger spatial scale is one of great scientific interest. Evidence for the former

approach has been provided in previous studies (King et al., 2015), including the specific case of heterogeneity between Haitian departments in cholera transmission (Collins and Govinder, 2014). Note that this evidence alone does not automatically discredit conclusions drawn via nationally aggregated models, as there are also reasons to prefer a simple model over a complex one (Saltelli et al., 2020; Green and Armstrong, 2015).

Fitting scientifically flexible metapopulation models is a challenging statistical problem. In particular, parameter estimation techniques based on the particle filter become computationally intractable as the number of spatial units increase. This is a result of the approximation error of particle filters growing exponentially in the dimension of the model (Rebeschini and van Handel, 2015; Park and Ionides, 2020).

Parameters that must be fit in Model 3 are shared between each department, the exception to this being the parameters β_{W_u}, and β_u, which are unique for each department $u \in 1{:}10$. To avoid the parameter estimation issue in high-dimensional models, Lee et al. (2020a) simplified the problem by using the MIF2 algorithm to separately estimate parameters for independent department-level models; the shared parameters were calibrated using the cholera incidence data from Artibonite, and the department-specific parameters (β_{W_u} and β_u) were fit using the data from their respective department. Reducing a spatially coupled model to individual units in this fashion requires special treatment of any interactive mechanisms between spatial units, such as found in Eq. (23). In particular, when considering a model for department u, the values $I_\nu(t)$ and $A_\nu(t)$, are unknown for $u \neq v \in 1{:}10$. A first order approximation of Eq. (23) for each department $v \in 1{:}10$ can be obtained using the weekly number of observed cholera cases in each department:

$$I_v(t) + A_v(t) \approx \frac{365}{7\rho} y_v^*(t) \left(\frac{1}{\delta + \delta_C + \mu_{IR}} + \frac{1-f}{f(\delta + \mu_{IR})} \right) \tag{34}$$

This approximation leads to department-specific models that are conditionally independent given the reported number of cholera infections in the remaining departments. Here, we refer to a collection of POMP models that are independent across units as a PanelPOMP.

In the case of a PanelPOMP, an extension of the IF2 algorithm, known as Panel Iterated Filtering (PIF) (Bretó, Ionides and King, 2020a), can be used to obtain the MLE, which solves the curse of dimensionality for this class of models. A major advantage of this algorithm is that PIF can be used to fit both unit-specific parameters and shared parameters; in this way, the calibration of shared parameters involves all of the available data instead of just an arbitrarily chosen subset.

We propose and use a slight modification of the PIF algorithm, which we call the Block Panel Iterated Filter (BPIF), to estimate parameters for a PanelPOMP version of Model 3. Let $l_{\mathrm{P}}(y^*_{1:N};\theta)$ denote the log-likelihood function of the PanelPOMP version of Model 3. The output of the BPIF algorithm is the MLE $\tilde{\theta} = \arg\max_\theta l_{\mathrm{P}}(y^*_{1:N};\theta)$. The BPIF algorithm is implemented in the `panelPomp` package (Bretó, Ionides and King, 2020b) as the argument `block = TRUE` in the `mif2` function. Pseudo-code for this algorithm is provided in Algorithm S1 in the supplementary material.

Fitting Model 3 in this fashion simplifies the problem of parameter estimation, but it also introduces additional technicalities that must be addressed. One concern is that of obtaining model forecasts, which was the primary goal of Lee et al. (2020a). The simplified PanelPOMP version of Model 3 relies on the observed cholera cases as a covariate, which are unavailable for use in forecasts. Similar to what was done in Lee et al. (2020a), we use the parameter estimates obtained with the independent departmental approximation of Model 3 as an estimate for the parameters in the fully coupled

version of the model, which was implemented using the `spatPomp` package. Simulations from the fully coupled version of the model using the estimated parameter vector $\tilde{\theta}$, are displayed in Fig. 4.

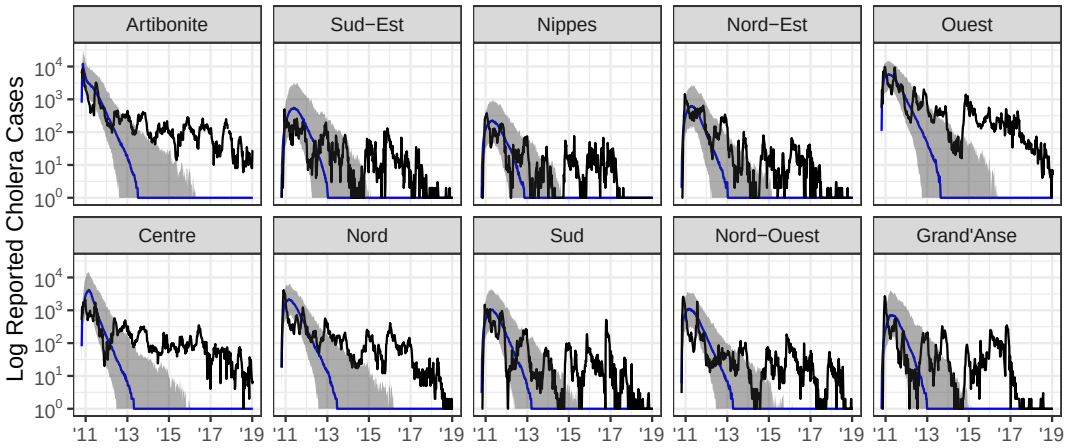

FIG 4. Simulations from initial conditions using the spatially coupled version of Model 3. The black curve represents true case count, the blue line the median of 500 simulations from the model, and the grey ribbons representing 95% confidence interval.

Let $l_{\mathrm{C}}(y^*_{1:N};\theta)$ denote the log-likelihood of the fully coupled model, and let $\hat{\theta} = \arg\max_\theta l_{\mathrm{C}}(y^*_{1:N};\theta)$. If the goal of parameter estimation is to maximize the likelihood of the fully coupled model, the approach of estimating parameters using a decoupled model is clearly sub optimal, as $l_{\mathrm{C}}(y^*_{1:N};\hat{\theta}) \geq l_{\mathrm{C}}(y^*_{1:N};\tilde{\theta})$. A natural question that arises is if the difference between $\hat{\theta}$ and $\tilde{\theta}$ is scientifically and statistically meaningful. The answer to this question requires the ability to reliably estimate $\hat{\theta}$, the answer to which is model specific and beyond the scope of this case study.

A brief analysis of the fitted parameters $\tilde{\theta}$ provides some interesting insight. Consider the large estimated exponent on rainfall: $r = 1.68 \times 10^5$ (Eq. 31). Because the value of rainfall $J(t) \in (0,1)$, the large coefficient results in $J(t)^r \approx 0$. This result may initially be surprising given the importance of rainfall as a seasonal diver of cholera infections in Haiti (see Sec. 3.4). One plausible explanation of this result is that the fitting procedure—used to obtain parameter values in the independent department model given the reported cases in other departments—has learned an over-dependence on the reported cholera cases in other departments, and thus the seasonality of cholera incidence in department u depends on the highly correlated covariate of the reported cholera cases in departments $1:U \setminus \{u\}$. Whatever the case may be, this fitting procedure results in a model that is unable to replicate the observed number of cases by simulating from initial conditions (see Fig. 4), and highlights the need for advancements in statistical methodology that permit inference on models with coupled metapopulation dynamics.

3.2. *Model Diagnostics.* Parameter calibration (whether Bayesian or frequentist) aims to find the best description of the observed data under the assumptions of the model. Obtaining the best fitting set of parameters for a given model does not, however, guarantee that the model provides an accurate representation of the system in

question. Model misspecification, which may be thought of as the omission of a mechanism in the model that is an important feature of the dynamic system, is inevitable at all levels of model complexity. To make progress, while accepting proper limitations, one must bear in mind the much-quoted observation of Box (1979) that "all models are wrong but some are useful." Beyond being good practical advice for applied statistics, this assertion is relevant for the philosophical justification of statistical inference as severe testing (Mayo, 2018, Sec. 4.8). In this section, we provide some tools and suggestions for diagnosing mechanistic models with the goal of making the subjective assessment of model "usefulness" more objective. To do this, we will rely on the quantitative and statistical ability of the model to match the observed data, which we call the model's *goodness-of-fit*, with the guiding principle that a model which cannot adequately describe observed data may not be reliable for useful purposes. Goodness-of-fit may provide evidence supporting the causal interpretation of one model versus another, but cannot by itself rule out the possibility of alternative explanations.

One common approach to assess a mechanistic model's goodness-of-fit is to compare simulations from the fitted model to the observed data. Visual inspection may indicate defects in the model, or may suggest that the observed data are a plausible realization of the fitted model. While visual comparisons can be informative, they provide only a weak and informal measure of the goodness-of-fit of a model. The study by Lee et al. (2020a) provides an example of this: their models and parameter estimates resulted in simulations that visually resembled the observed data, yet resulted in model likelihoods that were—in some cases—remarkably smaller than likelihoods that can be achieved via the likelihood based optimization techniques that were used (see Table 1). Alternative forms of model validation should therefore be used in conjunction with visual comparisons of simulations to observed data.

Another approach is to compare a quantitative measure of the model fit (such as MSE, predictive accuracy, or model likelihood) among all proposed models. These comparisons provide insight into how each model performs relative to the others. To calibrate relative measures of fit, it is useful to compare against a model that has well-understood statistical ability to fit data, and we call this model a *benchmark*. Standard statistical models, interpreted as associative models without requiring any mechanistic interpretation of their parameters, provide suitable benchmarks. Examples include linear regression, auto-regressive moving average time series models, or even independent and identically distributed measurements. The benchmarks enable us to evaluate the goodness of fit that can be expected of a suitable mechanistic model.

Goodness-of-fit alone does not guarantee that a model provides a correct causal interpretation of the model. Indeed, associative models are not constrained to have a causal interpretation, and typically are designed with the sole goal of providing a statistical fit to data. Consequently, we should not require a candidate mechanistic model to beat all benchmarks. However, a mechanistic model which falls far short against benchmarks is evidently failing to explain some substantial aspect of the data. A convenient measure of fit should have interpretable differences that help to operationalize the meaning of far short. Ideally, the measure should also have favorable theoretical properties. Consequently, we focus on log-likelihood as a measure of goodness of fit, and we adjust for the degrees of freedom of the models to be compared by using the Akaike information criterion (AIC) (Akaike, 1974).

In some cases, a possible benchmark model could be a generally accepted mechanistic model, but often no such model is available. Because of this, we use a log-linear Gaussian ARMA model as an associative benchmark, as recommended by He, Ionides and King (2010). The theory and practice of ARMA models is well developed, but the exponential

growth and decay characteristic of biological dynamics suggests applying these linear models on a log scale. Likelihoods of Models 1–3 and their respective ARMA benchmark models are provided in Table 1.

It should be universal practice to present measures of goodness of fit for published models, and mechanistic models should be compared against benchmarks. This alone would assist authors and readers to confront any major statistical limitations of the proposed mechanistic models. In addition, the published goodness of fit provides a concrete measure for subsequent research to identify and remedy limitations in the analysis, or to update the investigation based on new data or new scientific understanding. When combined with online availability of data and code, objective measures of fit provide a powerful tool to accelerate scientific progress, following the paradigm of the *common task framework* (Donoho, 2017, Sec. 6). In our literature review of the Haiti cholera epidemic [CHECK THIS] no quantitative measures of goodness of fit, and no benchmark models, were considered in any of the 24 papers which calibrated a mechanistic model to data in order to obtain scientific conclusions.

The use of benchmarks may also be beneficial when developing models with varying spatial scale, where a direct comparison between models likelihoods is meaningless. In such a case, a benchmark model could be fit to each spatial resolution being considered, and each model compared to their respective benchmark. Large advantages (shortcomings) in model likelihood relative to the benchmark for a given spatial scale that are not present in other spatial scales may provide weak evidence for (against) the spatial resolution under consideration. While these comparisons don't provide a formal statistical test of the amount of spatial resolution needed for a given model, they can be quite informative.

	Model 1	Model 2	Model 3
Log-likelihood	−2731.3 (−3132.4)[1]	−21928.0 (−29350.0)	−17883.2 (−42964.1)[2]
Number of Fit Parameters	13 (20)	16 (26)	29 (29)
AIC	5488.7 (6304.8)[1]	43888.1 (58752.1)	35824.3 (85986.2)[2]
Log-ARMA(2,1) Log-likelihood	−2802.6	−18061.9	−18061.9

TABLE 1

Log-likelihood values for each models compared to their ARMA benchmarks. Values in parenthesis are corresponding values using Lee et al. (2020a) parameter estimates. [1] *The reported likelihood is an upper bound of the likelihood of the Lee et al. (2020a) model as it is the largest likelihood obtained using their parameter calibration regime.* [2] *Lee et al. (2020a) fit Model 3 to a subset of the data (March 2014 onward, excluding data from Ouest in 2015-2016). On this subset, their model has a likelihood of* −9003.0. *On this same subset, our model has a likelihood of* −7367.5.

Similar to comparing log-likelihoods across models, an additional powerful diagnosis tool is the comparison of conditional log-likelihoods. Conditional likelihoods, defined as the density $f_{Y_k|Y_1,\ldots,Y_{k-1}}(Y_k = y_k^*|y_{1:k-1}^*)$, provide a basic description of how well the proposed model can describe each data point, given the previous observations. Comparing these results across models—including benchmark models—can help researchers identify potential model deficiencies, or errors in the observed data. Additional tools for assessing the goodness-of-fit of a model include plotting the effective sample size

of each observation (Liu, 2001), and comparing any statistic of the observed data to simulations from the model (Wood, 2010)—which is sometimes referred to as diagnostic probes (King, Nguyen and Ionides, 2016).

3.3. *Forecasts.* The central goal of a forecast is to provide an accurate estimate of the future state of a system based on currently available data. When a mechanistic model is used, forecasts may also provide estimates of the future effects of potential interventions. Forecasting models are built using available scientific understanding, but forecasting can also be a way of testing new scientific hypotheses in real time (Lewis et al., 2022). In order to provide trustworthy information, however, the reliability of the forecast should be understood. In particular, researchers should account for various forms of uncertainty present in model forecasts, and calibrate the proposed model to observed data.

The most recently available information about a dynamic system should be used to obtain reliable forecasts. Using state space model m, this means that in order to obtain a forecast up to time $N+s$, where $s \geq 0$ and N is the index for the last available data, we should simulate from the conditional density $f_{\boldsymbol{X}^{(m)}_{N:s},\boldsymbol{Y}^{m}_{N:s}}(x_{N:s}, y_{N:s} \mid \hat{\theta}, \boldsymbol{X}^{(m)}_{N} = \boldsymbol{x}^{(m)}_{N})$, where $\boldsymbol{x}^{(m)}_{N}$ is the value of the latent state at time N. Because the latent state $\boldsymbol{X}^{(m)}_{N}$ is unobservable, draws from the desired conditional density are unobtainable. Informed estimates of $\boldsymbol{X}^{(m)}_{N}$ given the observed data $\boldsymbol{Y}^{(m)}_{1:N} = y^{*}_{1:N}$ can easily be obtained using the filtering distribution. Let $\hat{\boldsymbol{X}}^{(m),i}_{N}$, $i \in 1, 2, \ldots, J$ be iid draws from the filtering distribution at time N, with density $\hat{\boldsymbol{X}}^{(m),i}_{N} \sim f_{\boldsymbol{X}^{(m)}_{N}|\boldsymbol{Y}^{(m)}_{1:N}}(x_N \mid \hat{\theta}, y^{*}_{1:N})$. A single model forecast can then be obtained by simulating from the model $f_{\boldsymbol{X}^{(m)}_{N:s},\boldsymbol{Y}^{(m)}_{N:s}}(x_{N:s}, y_{N:s} \mid \hat{\theta}, \boldsymbol{X}^{(m)}_{N} = \hat{\boldsymbol{X}}^{(m),i}_{N})$. Obtaining forecasts in this fashion allows us to obtain results that are consistent with the model assumptions and the observed data. In this particular case study, projecting the future state of the cholera epidemic starting from latent states that are draws from the filtering distribution allows the proposed models to benefit from the fact that very few cholera cases had been observed between 2018 and 2019, and that cases appear to be decreasing. This alone partially explains the unrealistic forecasts of Lee et al. (2020a), which were obtained by simulating the model from initial conditions; Table S7 of their supplement material shows that of their simulations that were consistent with observing zero cases in 2019, all forecasts resulted in the elimination of cholera.

It has been noted that the uncertainty in just a single parameter can lead to drastically different forecasts (Saltelli et al., 2020); parameter uncertainty should therefore be considered when obtaining model forecasts to influence policy. If a Bayesian technique is used for parameter estimation, the most natural way to account for parameter uncertainty is to obtain J simulations from the model where each simulation is obtained using parameters $\Theta_{1:J}$ drawn from the posterior distribution $\Theta_{1:J} \sim f_{\Theta}\big(\theta|\boldsymbol{Y}^{(m)}_{1:N} = y^{*}_{1:N}\big)$. In the frequentist case, one approach is empirical Bayes: endow to set $\Theta_K = \{\theta_1, \theta_2, \ldots, \theta_K\}$ with a discrete uniform distribution, such that $P\big(\Theta_j = \theta_k\big) = \frac{1}{K}$ for all values $k \in 1\!:\!K$. In this case, the posterior distribution of $\Theta_j|\boldsymbol{Y}^{(m)}_{1:N}$ can be expressed as: $P\big(\Theta_j = \theta_k|\boldsymbol{Y}^{(m)}_{1:N} = y^{*}_{1:N}\big) = \frac{l(y^{*}_{1:N};\theta_k)}{\sum_{l=1}^{K} l(y^{*}_{1:N};\theta_l)}$. To complete the formulation, we then choose the set Θ_K such that each $\theta_k \in \Theta_K$ is a single output of the MIF2 algorithm, which has been shown to be within a neighborhood of the MLE Ionides et al. (2015). This approach has been taken in previous studies (King et al., 2015), and the result is simply weighing parameter sets based on their corresponding

likelihoods: parameters with higher likelihoods should be used more frequently in model forecasts than those that correspond to lower likelihoods. This can be done for both deterministic and stochastic models, but requires a large number of computations, especially as the number of observations and model parameters increase. Because our direct aim is not to provide forecasts for policy recommendation, we do not include parameter uncertainty in our forecasts. Instead, we use the projections from point estimates to highlight a major deficiency of deterministic models, which is that the only variability in model projections is a result of parameter uncertainty, which leads to over-confidence in forecasts King et al. (2015). To the credit of deterministic models, of the four fitted models in Lee et al. (2020a), Model 2 provides the most apparently accurate forecasts. This perhaps demonstrates that while deterministic models describe the systems in a less realistic way, the relative ease in fitting these models potentially results in fewer modeling errors. As the data analysis becomes more refined, however, the deficiencies of deterministic models become increasingly apparent.

Forecasting can be used to simulate various interventions on a system; this feature is useful to inform policy and was the primary goal of Lee et al. (2020a). Outcomes of their study include estimates for the probability of cholera elimination and cumulative number of cholera infections under several possible vaccination scenarios. Mimicking their efforts, we define cholera elimination as having less than one infection of cholera over at least 52 consecutive weeks, and provide forecasts under the following vaccination scenarios:

$V0$: No additional vaccines are administered.
$V1$: Vaccination limited to the departments of Centre and Artibonite, deployed over a two-year period.
$V2$: Vaccination limited to three departments: Artibonite, Centre, and Ouest deployed over a two-year period.
$V3$: Countrywide vaccination implemented over a five-year period.
$V4$: Countrywide vaccination implemented over a two-year period.

Simulations from probabilistic models (Models 1 and 3) represent possible trajectories of the dynamic system under the scientific assumptions of the models. Estimates of the probability of cholera elimination can therefore be obtained as the proportion of simulations from these models that result in cholera elimination. The results of these projections are summarized in Figs. 5–8, and suggest that cholera elimination was likely even without increased vaccination efforts—consistent with observed reality (Trevisin et al., 2022).

Probability of elimination estimates of this form are not meaningful for deterministic models, as the trajectory of these models only represent the mean behavior of the system rather than individual potential outcomes. We therefore do not provide probability of elimination estimates under Model 2. Still, trajectories obtained by Model 2 are consistent with the simulation results of Models 1 and 3, and suggest that cholera was in the process of being eliminated from Haiti.

In additional to probability of elimination estimates, we provide estimates for the cumulative number of infections under each vaccination scenario from February 2019 – February 2024. Notably, the median number of cumulative cholera infections under the no-vaccination scenario using Models 1 and 3 were $9,200$ and 800, respectively. While there is remaining time during this projection period in which new cholera infections can be detected, up to this point our estimates are far more consistent with the observed number of reported cholera cases than the corresponding estimates from Lee et al. (2020a), which were approximately $400,000$ and $1,000,000$.

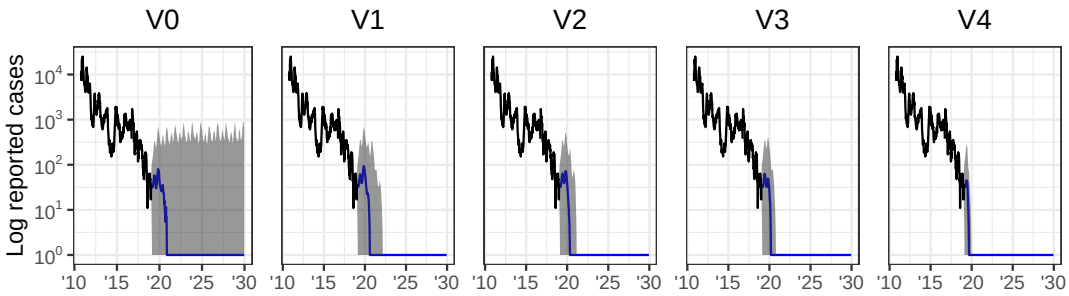

FIG 5. Simulations of Model 1 under each vaccination scenario. Blue line indicates the simulated median of reported cases, and the ribbon represents 95% of simulations.

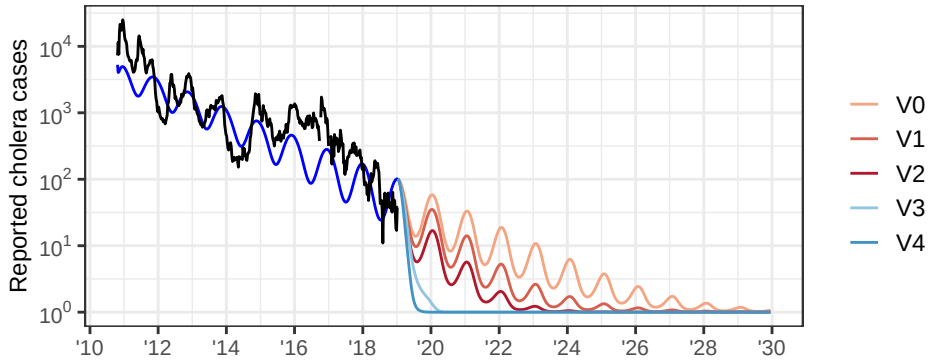

FIG 6. Simulated trajectory of Model 2 (blue curve) and projections under the various vaccination scenarios. Reported cholera incidence is shown in black.

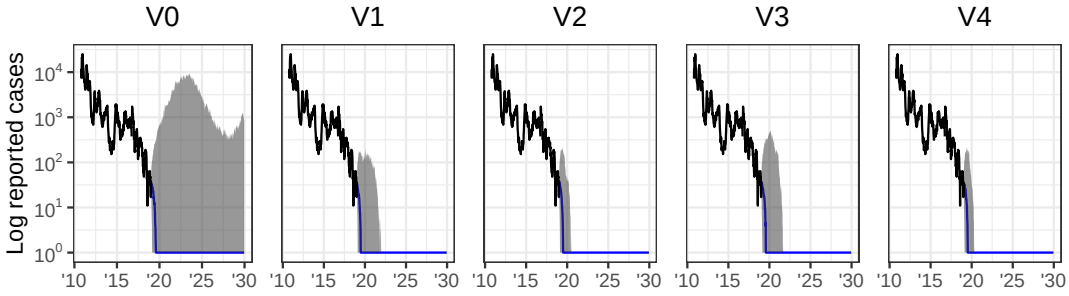

FIG 7. Simulations of Model 3 under each vaccination scenario. Blue line indicates the simulated median of reported cases, and ribbon represents 95% of simulations.

3.4. *Reckoning Scientific Understanding With Fitted Models.* The resulting mechanisms in a fitted model can be compared to current scientific knowledge about a system. Agreement between model based inference and our current understanding of a system may taken as a confirmation of both model based conclusions and our scientific understanding. On the other hand, discrepancies found in these comparisons may have the potential to spark new scientific knowledge (Ganusov, 2016). The potential for new dis-

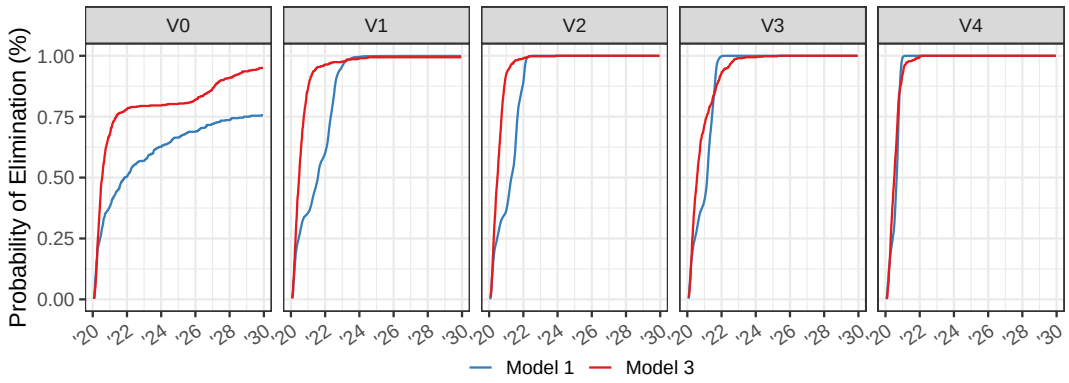

FIG 8. Probability of elimination across simulations for a 10 year period. Compare to Figure 3A of Lee et al. (2020a).

coveries based on dynamic modeling outcomes, however, is rooted in a model's ability to effectively approximate the system in question, and is therefore inexorably linked to a models goodness-of-fit. As such, conclusions based on dynamic models that poorly fit observed data should not be considered as much as those resulting from models that describe observed data well.

In our analysis, we demonstrate how one may compare model mechanisms with current scientific understanding by examining the results of fitting the flexible cubic spline term in Model 1 (Eq. (1)–(2)); the cubic splines permit flexible estimation of seasonality in the force of infection. After calibrating model parameters to observed data, we explore potential patterns in the seasonal transmission rate by plotting the average value of β in a typical year. Fig. 9 shows that the estimated seasonal transmission rate β mimics the rainfall dynamics in Haiti, despite Model 1 not having access to rainfall data. This result provides evidence that rainfall played an important role in cholera transmission in Haiti, consistent with previous studies (Lemaitre et al., 2019; Eisenberg et al., 2013). The estimated seasonality also features an increased transmission rate during the Fall, which was noticed at an earlier stage of the epidemic (Rinaldo et al., 2012).

While dynamic models may potentially result in meaningful scientific insights, one must be careful not to assume the results of a single analysis as a de-facto feature of the system. It is important instead to recognize and assess the modeling simplifications and assumptions that were used in order to arrive at the conclusions. In epidemiological studies, for example, it is unreasonable to expect that our quantitative understanding of individual-level processes will perfectly match model parameters that were fit to population-level case counts (He, Ionides and King, 2010) [ED: ARE THERE BETTER REFERENCES FOR THIS?], making direct interpretation estimated parameters difficult.

Our case study provides an example of this in the parameter estimate for the duration of natural immunity due to cholera infection μ_{RS}^{-1}. Under the framework of Model 2, the best estimate for this parameter is 1.97×10^{16}, suggesting that individuals have effectively permanent immunity to cholera once infected, and thus the pool of susceptible individuals decreases over time. The interpretation of this result is muddled due to the fact that the data ranged from 2010-2019, and therefore estimates of immunity longer than 10 years—the upper end of previous estimates of natural immunity (King et al.,

Mechanism	Model 1		Model 2		Model 3	
Infection (day)	$\mu_{IR}^{-1} = 2.0$	(6)	$\mu_{IR}^{-1} = 7.0$	(14)	$\mu_{IR}^{-1} = 2.0$	(26)
Latency (day)	$\mu_{EI}^{-1} = 1.4$	(5)	$\mu_{EI}^{-1} = 1.3$	(13)	—	
Seasonality	$\beta_{1:6} = (1.6, 1.1, 1.3, 1.1, 1.5, 0.9)$	(2)	$a = 0.4$	(11)	$a = 19.49$ $r = 1.68 \times 10^5$	(31)
Immunity (year)	$\mu_{RS}^{-1} = 8.0$	(7)	$\mu_{RS}^{-1} = 2 \times 10^{16}$ $\omega_1^{-1} = 1.0$ $\omega_2^{-1} = 5.0$	(15) (17) (17)	$\mu_{RS}^{-1} = 8.0$	(29)
Vaccine efficacy	—		$\theta_{1:4} = (0.80, 0.76$ $0.57, 0.48)$	(12)	$\eta_{ud}(t)$	
Birth/death (yr)	$\mu_S^{-1} = 44.9$ $\delta^{-1} = 134.2$	(9)	—		$\delta^{-1} = 63.0$	(27)
Symptomatic frac.	$f_z(t) = c\theta^*(t - \tau_d)$	(4-5)	$f = 0.2$	(13)	$f = 1.00$	(25)
Asymptomatic infectivity	$\epsilon = 0.05$	(1)	$\epsilon = 0.001$ $\epsilon_W = 10^{-7}$	(11) (19)	$\epsilon = 1$ $\epsilon_W = 0.010$	(23) (31)
Human to human	$\beta_{1:6}$ as above	(1)	$\beta = 4.51 \times 10^{-15}$	(11)	$\beta_{1:10} = (2.16, 1.02, 0.35, 0.22, 2.18, 0.59, 1.44, 2.09, 0.44, 0.28) \times 10^{-6}$	(23)
Water to human	—		$W_{\text{sat}} = 10^5$ $\beta_W = 1.06$	(11)	$\beta_{W_{1:10}} = (1.23, 5.21, 7.23, 8.31, 1.35, 8.28, 2.73, 0.24, 3.54, 3.91)$	(23)
Human to water	—		$\mu_W = 1.17 \times 10^4$	(19)	$\mu_W = 2.71 \times 10^{-5}$	(31)
Water survival (wk)	—		$\delta_W^{-1} = 3$	(20)	$\delta_W^{-1} = 0.36$	(32)
Mixing exponent	$\nu = 0.97$	(1)	—		—	
Process noise($\text{wk}^{1/2}$)	$\sigma_{\text{proc}} = (0.31, 0.35)$	(1)	—		$\sigma_{\text{proc}} = 0.032$	(25)
Reporting rate	$\rho = 0.898$	(S16)	$\rho = 0.20$	(S17)	$\rho = 0.89$	(S18)
Observation overdispersion	$\psi = (356.16, 59.55)$	(S16)			$\psi = 117.11$	(S18)

TABLE 2

References to the relevant equation are given in parentheses. Parameters in blue were fixed based on scientific reasoning and not fitted to the data. [N] denotes parameters added during our re-analysis, not considered by Lee et al. Translations back into the notation of Lee et al. (2020a) are given in Table S1.

2008)—effectively result in the same model dynamics. The depletion of susceptible individuals may also be attributed to confounding mechanisms that are unaccounted for in the model—such as vaccination and other non-pharmaceutical interventions that reduce cholera transmission (Trevisin et al., 2022; Rebaudet et al., 2021)—that were not accounted for in the model. Perhaps the best interpretation of the estimated parameter, then, is that under model assumptions, the model most adequately describes the observed data by having a steady decrease in the number of susceptible individuals. The poor model fit compared to a log-linear benchmark (see Table 1), and the large inconsistenty between model parameters and our understanding of the system, may be taken as evidence of model mispecificaiton.

4. Robustness of model based conclusions. [BETTER NAME FOR THIS?]

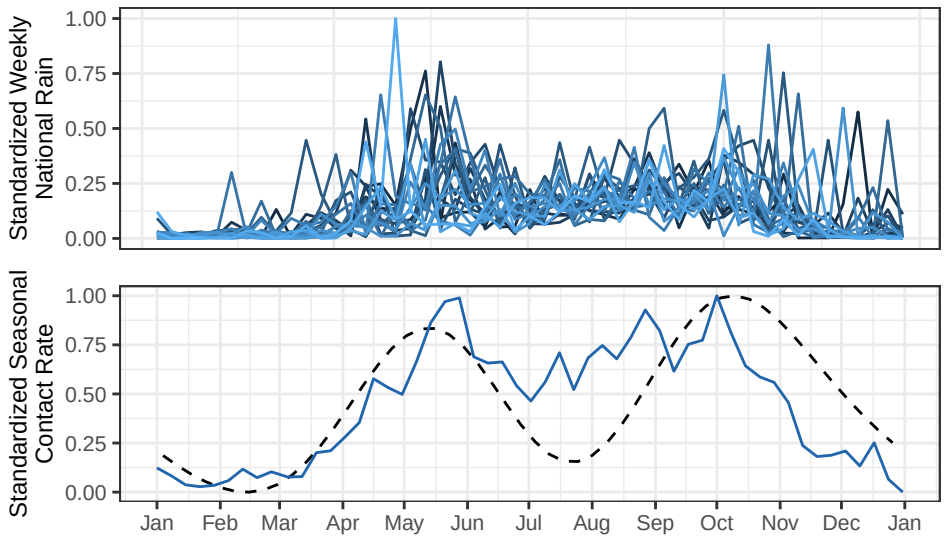

FIG 9. (Top) weekly rainfall in Haiti, lighter colors representing more recent years. (Bottom) estimated seasonality in the transmission rate (dashed line) plotted alongside mean rainfall (solid line).

A model which aspires to provide quantitative guidance for assessing interventions should provide a quantitative statistical fit for available data. However, strong statistical fit does not guarantee a correct causal structure: it does not even necessarily require the model to assert a causal explanation. A causal interpretation is strengthened by corroborative evidence. For example, reconstructed latent variables (such as numbers of susceptible and recovered individuals) should make sense in the context of alternative measurements of these variables; parameters that have been calibrated to data should make sense in the context of alternative lines of evidence about the phenomena being modeled.

If a mechanistic model including a feature (such as a representation of a mechanism, or the inclusion of a covariate) fits better than mechanistic models without that feature, and also has competitive fit compared to associative benchmarks, this may be taken as evidence supporting the scientific relevance of the feature. As for any analysis of observational data, we must be alert to the possibility of confounding. For a covariate, this shows up in a similar way to regression analysis: the covariate under investigation could be a proxy for some other unmodeled or unmeasured covariate. For a mechanism, the model feature could in principle explain the data by helping to account for some different unmodeled phenomenon. In the context of our analysis, the estimated trend in transmission rate could be explained by any trending variable (such as hygiene improvements, or changes in population behavior), resulting in confounding from colinear covariates. Alternatively, the trend could be attributed to a decreasing reporting rate rather than decreasing transmission rate, resulting in confounded mechanisms. The robust statistical conclusion is that a model which allows for change fits better than one which does not—we argue that a decreasing transmission rate is a plausible way to explain this, but the incidence data themselves do not provide enough information to pin down the mechanism.

5. Discussion. The ongoing global COVID-19 pandemic has provided a clear example on how government policy may be affected by the conclusions of scientific models.

This article demonstrates that fitting appropriate scientific models remains a challenging statistical task, and therefore great care is needed when fitting scientific models for policy recommendations. We provided a few suggestions that may aid the fitting of mechanistic models such as comparing model likelihoods to a benchmark. Improved model fits allows for meaningful statistical inference that may provide valuable insight on a dynamic system and may improve the accuracy of model forecasts. Caution is nonetheless needed when making policy based on modeling conclusions, as model misspecification may invalidate conclusions.

In this article we argue that careful attention to important statistical details could have led the models in Lee et al. (2020a) to correctly predict the imminent cholera elimination. We acknowledge the benefit of hindsight: our demonstration of a statistically principled route to obtain better-fitting models with more accurate predictions does not rule out the possibility of discovering other models that fit well yet predict poorly. We used the same data and models, and even much of the same code, as Lee et al. (2020a), and yet ended up with drastically different conclusions. At a minimum, we have shown that the conclusions are sensitive to details in how the data analysis is carried out, and that attention to statistical fit (including numerical issues such as likelihood maximization) can lead to improved policy guidance.

We acknowledge there are limitations to this study, such as the inability to fit model parameters to the fully-coupled version of Model 3. Promising theoretical and methodological developments (Ning and Ionides, 2021; Ionides, Ning and Wheeler, 2022) based on the Block Particle Filter (Rebeschini and van Handel, 2015) may potentially be used to obtain better parameter estimates for Model 3 in future work.

Inference for mechanistic time series models offers opportunities for understanding and controlling complex dynamic systems. This case study has investigated issues requiring attention when applying powerful new statistical techniques that can enable statistically efficient inference for a general class of partially observed Markov process models. Care must be taken to ensure that the computationally intensive numerical calculations are carried out adequately. Once that is accomplished, care is required to assess what causal conclusions can properly be inferred given the possibility of alternative explanations consistent with the data. Studies that combine model development with thoughtful data analysis, supported by a high standard of reproducibility, build knowledge about the system under investigation. Cautionary warnings about the difficulties inherent in understanding complex systems (Saltelli et al., 2020; Ioannidis, Cripps and Tanner, 2020; Ganusov, 2016) should motivate us to follow best practices in data analysis, rather than avoiding the challenge.

5.1. *Reproducibility and Extendability.* Lee et al. (2020a) published their code and data online, and this reproducibility facilitated our work. By design, the models were coded and analyzed independently, leading to differing implementation decisions. Robust data analysis requires not only reproducibility but also extendability: if one wishes to try new model variations, or new approaches to fitting the existing models, or plotting the results in a different way, this should be not excessively burdensome. Scientific results are only trustworthy so far as they can be critically questioned, and an extendable analysis should facilitate such examination (Gentleman and Temple Lang, 2007).

We provide a strong form of reproducibility, as well as extendability, by developing our analysis in the context of an `R` package, `haitipkg`. Using a software package mechanism supports documentation, standardization and portability that promote extendability. In the terminology of Gentleman and Temple Lang (2007), the source code for this article is a *dynamic document* combining code chunks with text. In addition to

reproducing the article, the code can be extended to examine alternative analysis to that presented. The dynamic document, together with the R packages, form a *compendium*, defined by Gentleman and Temple Lang (2007) as a distributable and executable unit which combines data, text and auxiliary software (the latter meaning code written to run in a general-purpose, portable programming environment, which in this case is R).

Funding. This work was supported by National Science Foundation grants DMS-1761603 and DMS-1646108.

SUPPLEMENTARY MATERIAL

Eliminating cholera in Haiti: Supplement
This document contains additional details for Models 1–3, as well as a translation table that facilitates comparisons between these models and those described in Lee et al. (2020a). The supplement also demonstrates our capability to faithfully replicate the results of Lee et al. (2020a).

`haitipkg`
This `R` package is contained in a public GitHub repository: jeswheel/haitipkg. The package contains all of the data and code used to create and fit the models, as well as other useful functions that were used in this article.

`jesseuwheeler/haiti`
This GitHub repository contains the `.Rnw` files that were used to create this document and the supplement material.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19** 716-723. https://doi.org/10.1109/TAC.1974.1100705

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

Arulampalam, M. S., Maskell, S., Gordon, N. and Clapp, T. (2002). A Tutorial on Particle Filters for Online Nonlinear, Non-Gaussian Bayesian Tracking. *IEEE Transactions on Signal Processing* **50** 174 – 188.

Asfaw, K., Ionides, E. L. and King, A. A. (2021). `spatPomp`: R package for Statistical Inference for Spatiotemporal Partially Observed Markov Processes. `https://github.com/kidusasfaw/spatPomp`.

Bansal, S., Grenfell, B. T. and Meyers, L. A. (2007). When individual behaviour matters: homogeneous and network models in epidemiology. *Journal of the Royal Society Interface* **4** 879–891.

Behrend, M. R., Basáñez, M.-G., Hamley, J. I. D., Porco, T. C., Stolk, W. A., Walker, M., de Vlas, S. J. and for the NTD Modelling Consortium (2020). Modelling for policy: The five principles of the Neglected Tropical Diseases Modelling Consortium. *PLOS Neglected Tropical Diseases* **14** 1-17. https://doi.org/10.1371/journal.pntd.0008033

Box, G. E. (1979). Robustness in the strategy of scientific model building. In *Robustness in statistics* 201–236. Elsevier.

Brauer, F. (2017). Mathematical epidemiology: Past, present, and future. *Infectious Disease Modelling* **2** 113-127. https://doi.org/10.1016/j.idm.2017.02.001

Bretó, C. and Ionides, E. L. (2011). Compound Markov counting processes and their applications to modeling infinitesimally over-dispersed systems. *Stochastic Processes and their Applications* **121** 2571–2591. https://doi.org/10.1016/j.spa.2011.07.005

Bretó, C., Ionides, E. L. and King, A. A. (2020a). Panel Data Analysis via Mechanistic Models. *Journal of the American Statistical Association* **115** 1178-1188. https://doi.org/10.1080/01621459.2019.1604367

Bretó, C., Ionides, E. L. and King, A. A. (2020b). panelPomp: Statistical Inference for PanelPOMPs (Panel Partially Observed Markov Processes) R package version 0.10.0.2.

Bretó, C., He, D., Ionides, E. L. and King, A. A. (2009). Time Series Analysis via Mechanistic Models. *Annals of Applied Statistics* **3** 319–348.

Collins, O. C. and Govinder, K. S. (2014). Incorporating heterogeneity into the transmission dynamics of a waterborne disease model. *Journal of Theoretical Biology* **356** 133-143. https://doi.org/10.1016/j.jtbi.2014.04.022

Dadlani, A., Afolabi, R. O., Jung, H., Sohraby, K. and Kim, K. (2020). Deterministic Models in Epidemiology: From Modeling to Implementation. https://doi.org/10.48550/ARXIV.2004.04675

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics* **26** 745–766.

Eisenberg, M. C., Kujbida, G., Tuite, A. R., Fisman, D. N. and Tien, J. H. (2013). Examining rainfall and cholera dynamics in Haiti using statistical and dynamic modeling approaches. *Epidemics* **5** 197-207. https://doi.org/10.1016/j.epidem.2013.09.004

Francois, J. (2020). Cholera remains a public health threat in Haiti. *The Lancet Global Health* **8** e984.

Ganusov, V. V. (2016). Strong inference in mathematical modeling: a method for robust science in the twenty-first century. *Frontiers in microbiology* **7** 1131.

Gentleman, R. and Temple Lang, D. (2007). Statistical analyses and reproducible research. *Journal of Computational and Graphical Statistics* **16** 1–23.

Green, K. C. and Armstrong, J. S. (2015). Simple versus complex forecasting: The evidence. *Journal of Business Research* **68** 1678-1685. Special Issue on Simple Versus Complex Forecasting. https://doi.org/10.1016/j.jbusres.2015.03.026

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

Henrys, J. H., Lerebours, G., Achille, M. A., Moise, K. and Raccurt, C. (2020). Cholera in Haiti. *The Lancet Global Health* **8** e1469.

Ioannidis, J. P., Cripps, S. and Tanner, M. A. (2020). Forecasting for COVID-19 has failed. *International Journal of Forecasting*.

Ionides, E. L., Ning, N. and Wheeler, J. (2022). An iterated block particle filter for inference on coupled dynamic systems with shared and unit-specific parameters. https://doi.org/10.48550/ARXIV.2206.03837

Ionides, E. L., Nguyen, D., Atchadé, Y., Stoev, S. and King, A. A. (2015). Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. *Proceedings of the National Academy of Sciences of the USA* **112** 719—724. https://doi.org/10.1073/pnas.1410597112

Ionides, E. L., Breto, C., Park, J., Smith, R. A. and King, A. A. (2017). Monte Carlo profile confidence intervals for dynamic systems. *Journal of the Royal Society Interface* **14** 1–10.

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character* **115** 700–721.

King, A. A., Nguyen, D. and Ionides, E. L. (2016). Statistical Inference for Partially Observed Markov Processes via the R Package pomp. *Journal of Statistical Software* **69** 1–43.

King, A. A. and Rowan, T. (2020). subplex: Unconstrained Optimization using the Subplex Algorithm R package version 1.6.

King, A. A., Ionides, E. L., Pascual, M. and Bouma, M. J. (2008). Inapparent infections and cholera dynamics. *Nature* **454** 877–880.

King, A. A., Ionides, E. L., Bretó, C. M., Ellner, S. and Kendall, B. (2009). pomp: Statistical inference for partially observed Markov processes. R package, available at `http://cran.r-project.org/web/packages/pomp`.

King, A. A., Domenech de Cellès, M., Magpantay, F. M. and Rohani, P. (2015). Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. *Proceedings of the Royal Society B: Biological Sciences* **282** 20150347.

Lee, E. C., Chao, D. L., Lemaitre, J. C., Matrajt, L., Pasetto, D., Perez-Saez, J., Finger, F., Rinaldo, A., Sugimoto, J. D., Halloran, M. E., Longini, I. M., Ternier, R., Vissieres, K., Azman, A. S., Lessler, J. and Ivers, L. C. (2020a). Achieving coordinated national immunity and cholera elimination in Haiti through vaccination: A modelling study. *The Lancet Global Health* **8** e1081–e1089.

Lee, E. C., Ternier, R., Lessler, J., Azman, A. S. and Ivers, L. C. (2020b). Cholera in Haiti–Authors' reply. *The Lancet Global Health* **8** e1470–e1471.

Lemaitre, J., Pasetto, D., Perez-Saez, J., Sciarra, C., Wamala, J. F. and Rinaldo, A. (2019). Rainfall as a driver of epidemic cholera: Comparative model assessments of the effect of intra-seasonal precipitation events. *Acta Tropica* **190** 235-243. https://doi.org/10.1016/j.actatropica.2018.11.013

Lewis, A. S. L., Rollinson, C. R., Allyn, A. J., Ashander, J., Brodie, S., Brookson, C. B., Collins, E., Dietze, M. C., Gallinat, A. S., Juvigny-Khenafou, N., Koren, G., McGlinn, D. J., Moustahfid, H., Peters, J. A., Record, N. R., Robbins, C. J., Tonkin, J. and Wardle, G. M. (2022). The power of forecasts to advance ecological theory. *Methods in Ecology and Evolution*. https://doi.org/10.1111/2041-210X.13955

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer, New York. https://doi.org/10.1007/978-0-387-76371-2

Lucas, R. E. et al. (1976). Econometric policy evaluation: A critique. In *Carnegie-Rochester conference series on public policy* **1** 19–46.

Mayo, D. G. (2018). *Statistical Inference as Severe Testing*. Cambridge: Cambridge University Press.

Ndii, M. Z. and Supriatna, A. K. (2017). Stochastic mathematical models in epidemiology. *Information* **20** 6185–6196.

Ning, N. and Ionides, E. L. (2021). Iterated Block Particle Filter for High-dimensional Parameter Learning: Beating the Curse of Dimensionality. https://doi.org/10.48550/ARXIV.2110.10745

Park, J. and Ionides, E. L. (2020). Inference on high-dimensional implicit dynamic models using a guided intermediate resampling filter. *Statistics & Computing* **30** 1497–1522.

Pezzoli, L. (2020). Global oral cholera vaccine use, 2013–2018. *Vaccine* **38** A132-A140. Cholera Control in Three Continents: Vaccines, Antibiotics and WASH. https://doi.org/10.1016/j.vaccine.2019.08.086

Rebaudet, S., Gaudart, J. and Piarroux, R. (2020). Cholera in Haiti. *The Lancet global health* **8** e1468.

Rebaudet, S., Bulit, G., Gaudart, J., Michel, E., Gazin, P., Evers, C., Beaulieu, S., Abedi, A. A., Osei, L., Barrais, R. et al. (2019). The case-area targeted rapid response strategy to control cholera in Haiti: a four-year implementation study. *PLoS neglected tropical diseases* **13** e0007263.

Rebaudet, S., Dély, P., Boncy, J., Henrys, J. H. and Piarroux, R. (2021). Toward Cholera Elimination, Haiti. *Emerging infectious diseases* **27** 2932.

Rebeschini, P. and van Handel, R. (2015). Can local particle filters beat the curse of dimensionality? *The Annals of Applied Probability* **25** 2809–2866.

Rinaldo, A., Bertuzzo, E., Mari, L., Righetto, L., Blokesch, M., Gatto, M., Casagrandi, R., Murray, M., Vesenbeckh, S. M. and Rodriguez-Iturbe, I. (2012). Reassessment of the 2010–2011 Haiti cholera outbreak and rainfall-driven multiseason projections. *Proceedings of the National Academy of Sciences* **109** 6602-6607. https://doi.org/10.1073/pnas.1203333109

Saltelli, A., Bammer, G., Bruno, I., Charters, E., Di Fiore, M., Didier, E., Nelson Espeland, W., Kay, J., Lo Piano, S., Mayo, D. et al. (2020). Five ways to ensure that models serve society: a manifesto.

Stocks, T., Britton, T. and Höhle, M. (2020). Model selection and parameter estimation for dynamic epidemic models via iterated filtering: application to rotavirus in Germany. *Biostatistics* **21** 400–416.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A. and Stumpf, M. P. H. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface* **6** 187-202.

Tracy, M., Cerdá, M. and Keyes, K. M. (2018). Agent-Based Modeling in Public Health: Current Applications and Future Directions. *Annual Review of Public Health* **39** 77-94. PMID: 29328870. https://doi.org/10.1146/annurev-publhealth-040617-014317

Trevisin, C., Lemaitre, J. C., Mari, L., Pasetto, D., Gatto, M. and Rinaldo, A. (2022). Epidemicity of cholera spread and the fate of infection control measures. *Journal of the Royal Society Interface* **19** 20210844.

Varghese, A., Kolamban, S., Sherimon, V., Lacap, E. M., Ahmed, S. S., Sreedhar, J. P., Al Harthi, H., Shuaily, A. and Salim, H. (2021). SEAMHCRD deterministic compartmental model based on clinical stages of infection for COVID-19 pandemic in Sultanate of Oman. *Scientific Reports* **11** 1–19.

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* **466** 1102–1104.