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. Quantitative models for dynamic systems offer potential for designing effective control measures. Regulation of biological populations is a fundamental topic in epidemiology, ecology, fisheries and agriculture. Quantitative models for these 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. 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

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

has an agent-based construction, featuring considerable mechanistic detail but limited 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. OCV is accepted as a safe and effective tool for controlling the spread of cholera, however the available stockpile and production of OCV doses is insufficient to meet global needs (Lee et al., 2020a; Pezzoli, 2020) [NEW CHOLERA VACCINES, NEW STUDIES OF EXISTING VACCINES, AND INCREASED PRODUCTION VOLUMES ALL AROSE DURING 2010-2019. MAYBE SOMETHING LIKE: Advances in OCV technology and vaccine availability raised the possibility of planning a national vaccination program (Lee et al., 2020a; Pezzoli, 2020)] 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 without additional vaccination efforts (Ferguson, 2022).

The failure of Lee et al. (2020a) to correctly predict the elimination of cholera has been debated (Francois, 2020; Rebaudet, Gaudart and Piarroux, 2020; Henrys et al., 2020; Lee et al., 2020b). Rebaudet, Gaudart and Piarroux (2020) suggested that the models proposed by Lee et al. (2020a) were too unrealistic. 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.

We proceed by introducing Models 1–3 in Sec. 2; in Sec. 3, we present a methodological approach to examining and refining these models. In Sec. 4, we use improved model fits to project cholera incidence in Haiti under various vaccination campaigns. We then conclude with a discussion on the use of mechanistic models to inform policy decisions 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 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.

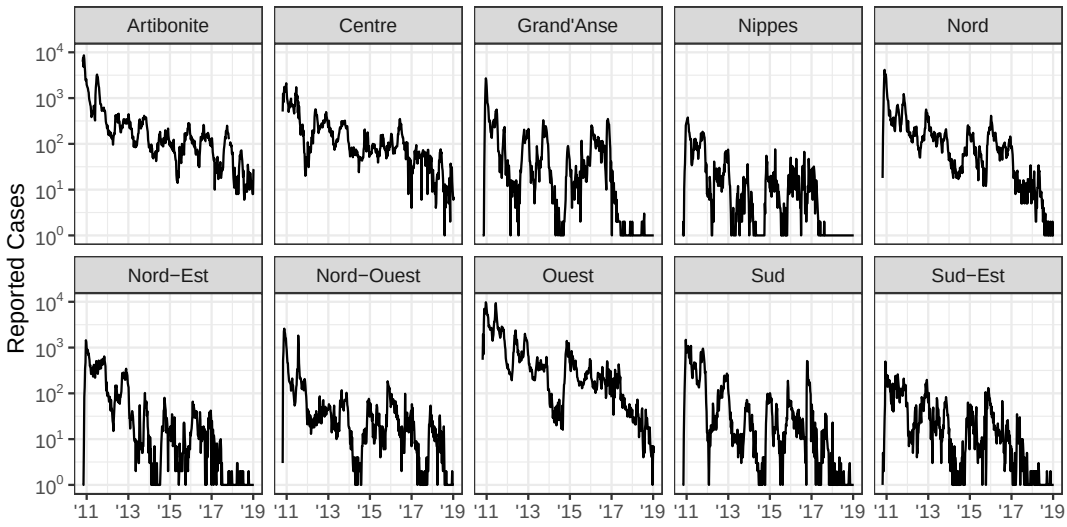

Fig 1. Reported Cholera cases in the outbreak in Haiti from 2010-2019.

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 for $m \in \{1, 2, 3\}$. 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 being 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

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

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 original implementation of Model 2 in Lee et al. (2020a) fixes the phase at $\phi = 0$.

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. We note that 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) and forecasting (Sec. 3.3).

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. While this section is focused on the elements of θ that are calibrated to data, we note that the decision of which elements to fix and which to estimate has consequences for model interpretability, as discussed in Sec. 4.

All three model 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 metapopulation (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 a probabilistic model, we suppose the existence of a joint density $f_{\boldsymbol{X}_{0:N}^{(m)},\boldsymbol{Y}_{1:N}^{(m)}}$, 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 $y_{1:n}^*$ are assumed to be a single realization of the model $y_{1:n}^* \sim f_{\boldsymbol{X}_{0:N}^{(m)},\boldsymbol{Y}_{1:N}^{(m)}}(x_{0:N}, y_{1:N}; \theta)$, where θ is a parameter vector that indexes the model. Using a probabilistic model 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.

Once a model for the system has been proposed, the parameter vector θ needs to be estimated using the observed data. There exist several algorithms, both frequentist and Bayesian, that can be used to obtain estimates of the parameters in stochastic dynamic 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). Similar to Model 2, this model was originally implemented in two parts: an epidemic phase from October 2010 through March 2015, and an endemic phase from March 2015 onward. As before, we continue to allow the re-estimation of some model parameters at the start of the endemic phase ($\rho, \sigma^2_{\text{proc}}$ and ψ) but require that the latent Markov process $X(t)$ carry over from one phase

into the next. The resulting confidence interval for ζ is $(-0.109, -0.014)$, 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.

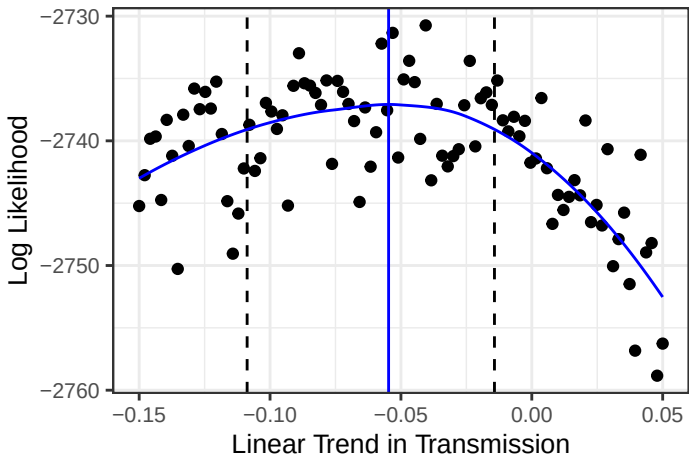

Fig 2. Monte Carlo adjusted profile of ζ. The blue curve is the profile, the blue line indicates the MLE, and the dashed lines indicate the confidence interval.

Model 1 is implemented using the `pomp` package King et al. (2009), which contains an implementation of the IF2 algorithm that was used to compute the likelihood profile. Simulations from the fitted model compared to the observed data are given in Fig 3.

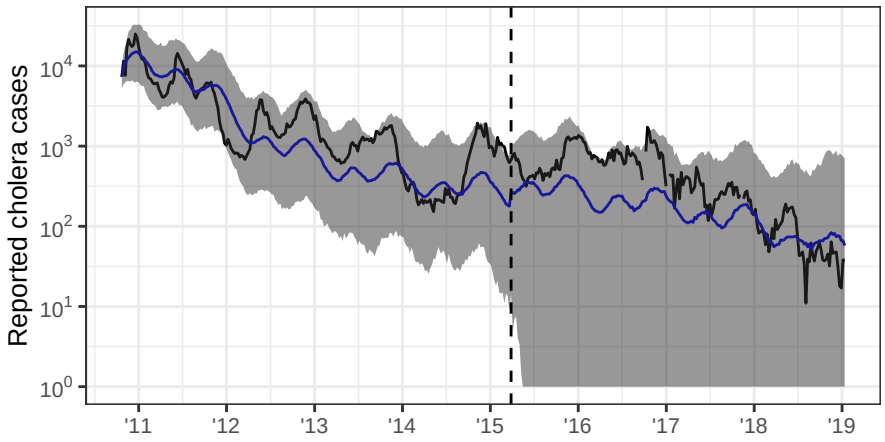

Fig 3. Simulations from Model 1 compared to reported cholera cases. The black curve is observed data, the blue curve is median of 500 simulations from the fitted model, and the vertical dashed line represents break-point when parameters are refit.

3.1.2. *Model 2.* Model 2 is a deterministic model, which can be viewed as a special case of a POMP (or SpatPOMP) with no randomness in the dynamic model. 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; 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) originally 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. We follow their decision to re-estimate model parameters at this breakpoint, but we include a requirement that the latent state $\boldsymbol{X}^{(2)}(t)$ at the start of the endemic period must be the same as the state at the end of the epidemic period. This additional constraint is sensible for a mechanistic interpretation of the latent state, though it adds difficulty to the task of obtaining a model fit that closely resembles the observed data. To combat this added difficulty, and to allow for a possible shift in the seasonality component in the force of infection, we introduce an additional phase parameter ϕ in Eq. 11.

We further increase model flexibility by fitting the parameter μ_W rather than treating it as fixed. 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 8.

3.1.3. *Model 3.* Model 3 is also of a probabilistic model. In this model, both the latent and observable processes can be factored into department specific processes which interact with each other. We denote the latent and measurement processes for Model 3 as $\boldsymbol{X}^{(3)}(t_{0:N}) = X^{(3)}_{1:U,0:N}$, and $\boldsymbol{Y}^{(3)}(t_{1:N}) = Y_{1:U,1:N}$. The decision to model the system via metapopulation models versus a model aggregated to a larger spatial scale is one of great scientific interest, and evidence for the former approach has been provided in previous studies (King et al., 2015). Note that this evidence alone does not automatically discredit conclusions drawn via nationally aggregated models, as an argument can easily be made that one should prefer a simple model over a complex one (Saltelli et al., 2020; Green and Armstrong, 2015). Still, researchers that intend to use a mechanistic model to describe a dynamic system should design the model to be as realistic as their scientific understanding of the system and their computational abilities permit. Fitting scientifically flexible metapopulation models, however, remains a challenging statistical problem; this is due to the fact that the approximation error of particle filters grows exponentially in the dimension of the model (Rebeschini and van Handel, 2015; Park and Ionides, 2020). Algorithms that are based on the particle filter therefore become computationally intractable as the number of spatial units increase.

Parameters that must be fit in Model 3 are primarily 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 fit this model, Lee et al. (2020a) simplified the parameter estimation problem by fitting independent department-level models to the data. 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-heterogeneous 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 over fitting each unit-specific model separately 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 then fit the panel version of Model 3 using a slight modification of the PIF algorithm, which we call the Block Panel Iterated Filter (BPIF). The BPIF algorithm is implemented in the `panelPomp` package (Bretó, Ionides and King, 2020b) in the R programming language, and can be used by adding the argument `block = TRUE` in the `mif2` function. Pseudo-code for this algorithm is provided in Algorithm S1 in the supplementary material. The reported parameters were obtained using the PanelPOMP version of Model 3, and simulations for the various vaccination campaigns (Sec. 4) were obtained using these same parameters in the SpatPOMP version of the model.

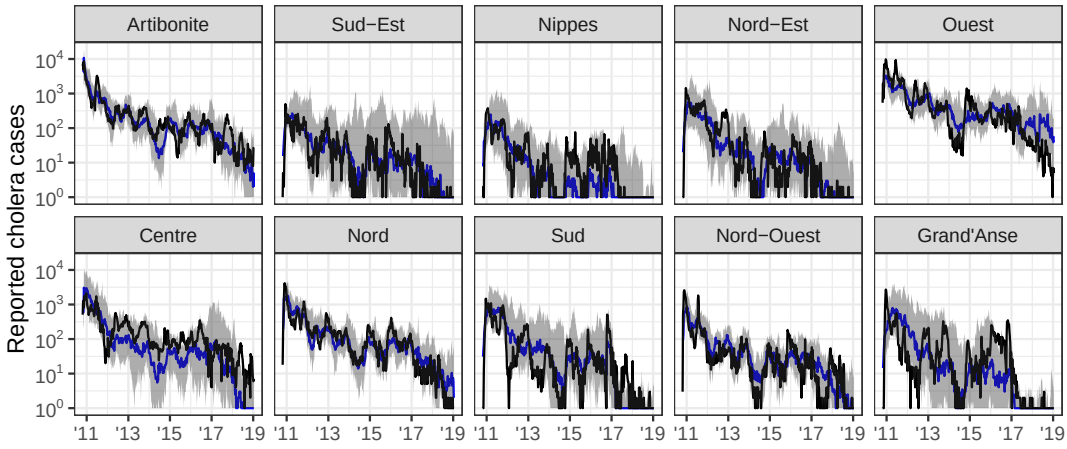

FIG 4. Simulations from initial conditions using panelPOMP version of Model 3. The black curve represents true case count, the blue curve the median of 500 simulations from the model, and the grey ribbons representing 95% confidence interval.

Fitting Model 3 as a PanelPOMP 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. To address this, we use the MLE obtained for the PanelPOMP approximation of Model 3 as an estimate for the parameters in fully coupled version of the model, which was implemented using the `spatPomp` package. Simulations from the SpatPOMP version of the model, using the parameters that were fit with the PanelPOMP version of the model, are displayed in Fig. 5.

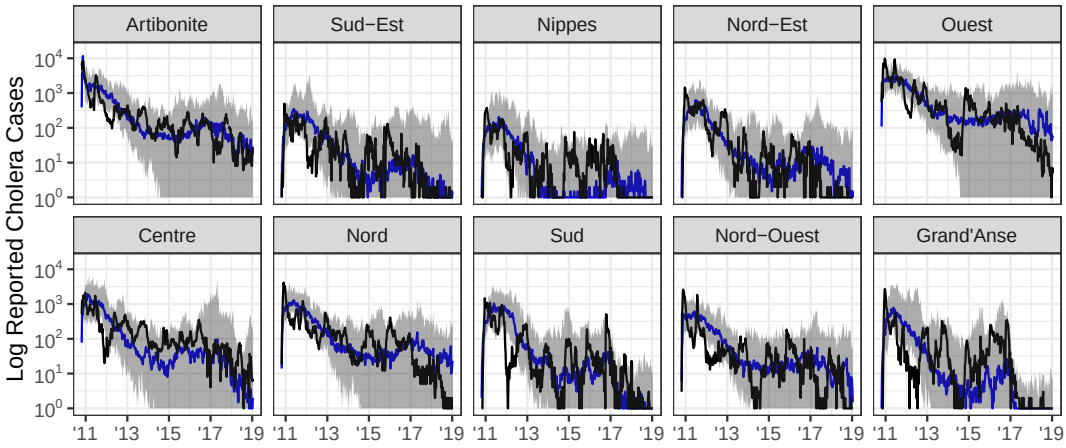

FIG 5. Simulations from initial conditions using SpatPOMP version of Model 3. The black curve represents true case count, the blue line the median of 500 simulations from the model, and the grey ribbons representing 95% *confidence interval.*

While simulations from the SpatPOMP version of Model 3 resemble the observed data, we note in Table 1 that the SpatPOMP version of Model 3 has a lower log-likelihood than the PanelPOMP version using the same model parameters. A natural question to ask is if the MLE obtained for the PanelPOMP version of Model 3—obtained using the approximation in Eq. (34)—could result in a drastically different set of parameters than those that would be obtained by estimating the MLE in the fully coupled model. Unfortunately, the answer to this question would require the ability to accurately and reliably estimate the MLE of the fully coupled model, which is beyond the scope of this article. This 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." [I WAS WONDERING WHETHER TO SAY SOMETHING LIKE, This is not just good practical advice for applied statistics, but matches broader appreciation of the limitations of all scientific knowledge (REF). DO YOU THINK THIS IS USEFUL, OR A DISTRACTION?] In this section, we discuss some tools and

suggestions for diagnosing mechanistic models with the goal of making the subjective assessment of model "usefulness" more objective by relying on the quantitative and statistical ability of the model to match the observed data, which we call the model's goodness-of-fit [GOODNESS-OF-FIT SEEMED INITIALLY TO ME A MEASURE OF "RIGHT VS WRONG" RATHER THAN "USEFUL VS USELESS". I THINK NOW I SEE YOUR POINT: GOODNESS OF FIT CAN GIVE US SOME INDICATION OF HOW WRONG A MODEL IS, WITH THE GUIDING PRINCIPLE THAT A MODEL WHICH IS TOO FAR WRONG IS PROBABLY NOT RELIABLE FOR USEFUL PURPOSES. MAYBE WE HAVE TO EXPLAIN THIS TO THE READER?] 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 *benchmarks*. 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) [REF].

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* (, Sec. 6). In our literature review of the Haiti cholera epidemic [CHECK THIS. HOW/WHERE SHOULD WE DESCRIBE THIS?] no quantitative measures of goodness of fit, and no benchmark models, were considered in any of the ?? papers which calibrated a mechanistic model to data in order to obtain scientific conclusions.

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.

The use of a common benchmark may also be beneficial when developing models with varying spatial scale, as direct comparisons between models fit to data with different levels of spatial aggregation are meaningless (e.g. comparing Model 1 to Model 3). [I DON'T QUITE UNDERSTAND THIS YET. BENCHMARKS TEND TO APPLY TO SPECIFIC DATA SETS, SO CANNOT READILY BE USED TO COMPARE DIFFERENT LEVELS OF AGGREGATION. LIKELIHOOD BENCHMARKS CAN BE USED TO STUDY THE VALUE OF EXPLICIT SPATIAL COUPLING. HOW TO BENCHMARK ACROSS AGGREGATION LEVELS IS, I THINK, SOMETHING OF AN OPEN PROBLEM.] Likelihoods of Models 1–3 and their respective ARMA benchmark models are provided in Table 1.

	Model 1	Model 2	Model 3 (panelPOMP)	Model 3 (SpatPOMP)
Log Likelihood	-2731.3 $(-3416.3)^1$	-40137.8 (-28006.9)	-17601.3^2	-19746.6 $(-60060.4)^3$
Number of Fit Parameters	13 (20)	21 (26)	39 (29)	39 (29)
AIC	5488.7 $(6872.6)^1$	80317.6 (56065.9)	35280.7^2	39571.1 $(120178.9)^3$
ARMA(2, 1) Log Likelihood	-2802.6	-18061.9	-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.

[NEW PARAGRAPH:] 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 [REF???] and comparing any statistic of the observed data to simulations from the model, which is sometimes referred to as

diagnostic probes (for example, the autocorrelation function (ACF), as was done in King et al. (2015)).

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 undersood. In particular, researchers should account for various forms of uncertainty present in model forecasts, and calibrate the proposed model to observed data.

[I THINK WE WILL HAVE TO EXPLAIN THIS CAREFULLY IF THE POINT IS TO APPEAR NONTRIVIAL. IT IS WORTH NOTING THAT LEE ET AL DID NOT PROPERLY FORECAST FROM THE BEST ESTIMATE OF THE CURRENT STATE. HOWEVER, IT WILL BE OBVIOUS TO MOST SATISTICAL READERS THAT THIS IS NOT THE BEST THING TO DO. IT MAY NOT BE SO OBVIOUS THAT THE DIFFERENCE IS IMPORTANT IF AND ONLY IF A DETERMINISTIC MODEL IS INADEQUATE. THE QUESTION OF WHETHER TO USE A DETERMINISTIC MODEL REMAINS OF CURRENT INTEREST...] As an example, we note that simulations from a well-fit mechanistic model may closely resemble the observed data $y^*_{1:N}$. In such comparisons, these simulations are random draws from the complete estimated joint distribution $f_{\boldsymbol{X}^{(m)}_{0:N},\boldsymbol{Y}^{(m)}_{1:N}}\left(x_{0:N}, y_{1:N} \mid \hat{\theta}\right)$, where m indexes the model used for simulations. It can therefore be tempting to use simulations from this model up to time $N+s$, $f_{\boldsymbol{X}^{(m)}_{0:N+s},\boldsymbol{Y}^{(m)}_{1:N+s}}\left(x_{0:N+s}, y_{1:N+s} \mid \hat{\theta}\right)$, with $s \geq 0$ to project the dynamic system up to a future time $N+s$, as it has been done in previous studies (Lee et al., 2020a; **?**). This approach, however, does not take advantage of the information about the state of the system that is contained in the observed data. Note that each of the models in question (and, more generally, all models that are described as POMPs) are Markovian, that is, the history of the process $\{\boldsymbol{X}^{(m)}_s, s \leq t\}$ for model m is uninformative about the future of the process $\{\boldsymbol{X}^{(m)}_s, s \geq t\}$, given the current state $\boldsymbol{X}^{(m)}_t$. In other words, observing the data $\boldsymbol{Y}^{(m)}_N = y^*_N$ at time N provides more information about the future state of the system than the initial conditions. In this case, draws from the conditional density $f_{\boldsymbol{X}^{(m)}_{N:s},\boldsymbol{Y}^{m}_{N:s}}\left(x_{N:s}, y_{N:s} \mid \hat{\theta}, \boldsymbol{X}^{(m)}_N = \boldsymbol{x}^{(m)}_N\right)$ should be preferred as forecasts, as these simulations account for the most recent known state of the system.

$\boldsymbol{X}^{(m)}_N$, however, is unobservable and therefore 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, however, via 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 $\hat{\boldsymbol{X}}^{(m),i}_N \sim$

[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]Parameters to department-specific models not provided by Lee et al. (2020a), as department fits were only used as an intermediary step to obtain parameter estimates of the coupled model.

[3]Model 3 was originally fit to only a subset of the data starting from March 2014 and did not include a large portion of data from Ouest in 2015-2016. On this subset, the parameters provided by Lee et al. (2020a) achieved a likelihood of -13202.1. On this same subset of data, our model achieved a likelihood of -8583.7.

$f_{\boldsymbol{X}_N^{(m)}|\boldsymbol{Y}_{1:N}^{(m)}}(x_N \mid \hat{\theta}, y^*_{1:N})$. A single model forecast can then be obtained by simulating from the model $f_{\boldsymbol{X}_{N:s}^{(m)},\boldsymbol{Y}_{N:s}^{(m)}}(x_{N:s}, y_{N:s} \mid \hat{\theta}, \boldsymbol{X}_N^{(m)} = \hat{\boldsymbol{X}}_N^{(m),i})$. Intuitively, simulating the model starting at the filtering distribution of the most recently available time point is a more appropriate way to project a stochastic dynamic system into the future, as it is not expected that each simulation from initial conditions will result in a latent state at time N consistent with the model and the observation $\boldsymbol{Y}_N^{(m)} = y^*_N$. In this particular case study, projecting the future state of the cholera epidemic in Haiti starting from the 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 (i.e., the number of susceptible individuals may be small).

Another consideration to make when obtaining model forecasts is that of parameter uncertainty. It has been noted that the uncertainty in just a single parameter can lead to drastically different projections (Saltelli et al., 2020). One possible approach to account for parameter uncertainty in model forecasts is by obtaining confidence intervals for each parameter, sampling parameters from the confidence intervals, simulating the model with the resulting parameter set, and then weighing the resulting model projections based on the likelihood of the given set of parameters, as was done in King et al. (2015). [THE "OBVIOUS" THING TO DO HERE IS A BAYESIAN APPROACH. IN SOME SENSE, KING15 IS EMPIRICAL BAYES. IT MAY TAKE CARE TO DISCUSS THIS WITHOUT OPENING UP A CAN OF WORMS - I CAN HAVE A GO AT MAKING SUGGESTIONS, NEXT TIME I'M WORKING ON THE MS.]

Note that in order to obtain projections that are consistent with the observed data, one must first be able to sample from the filtering distribution given each set of parameters. This approach 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 the focus of this study is on model fitting and evaluation, we do not provide model projections accounting for parameter uncertainty. 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 the projections. This observation is consistent with those made in King et al. (2015), and suggests that stochastic models should be preferred over deterministic models.

[Replaced Ed's comment with this:] We note—the credit of deterministic models—that 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 and useful 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.

4. Results. [SHOULD THESE FIRST TWO PARAGRAPHS BE MOVED TO THE DISCUSSION?] 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; parameter values which fit the 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 models, 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.

In a similar fashion, one can take advantage of certain mechanistic features contained in a particular model in order to make inference on a system. Examples of this are as diverse as estimating the effective reproductive number (R_0) of an infectious disease (He, Ionides and King, 2010) or investigating interactions between pedestrians and autonomous vehicles (Domeyer et al., 2022). In our analysis, we demonstrate this ability by examining the results of fitting the flexible cubic spline term in Model 1 (Eq. (1)–(2)), which allows for a flexible estimation of seasonality in the force of infection. After fitting the model, we explore potential patterns in the seasonal transmission rate by plotting the average value of β in a typical year. Fig. 6 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 is a potential driver of cholera infections in Haiti.

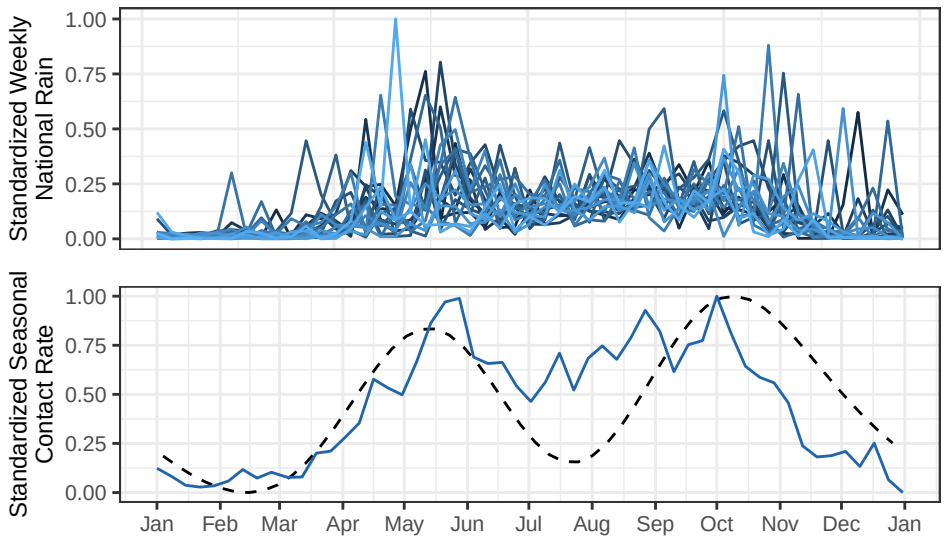

Fig 6. (Top) weekly rainfall in Haiti, lighter colors representing more recent years. (Bottom) estimated seasonality in the transmission rate (dashed line) plotted alongside mean rainfall (solid line).

While such inference is useful, 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.

An additional benefit of mechanistic modeling is ability 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 in the 10-year projection period, 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. In this case study, estimates of the probability of cholera elimination can therefore be obtained as the proportion of simulations from the fitted model that result in cholera elimination. The results of these projections are summarized in Figs. 7–10. These results suggest that cholera elimination was likely, even without increased vaccination efforts, which is consistent with observed reality (Ferguson, 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 $2,058$ and $623,538$, 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 $1,000,000$.

[PARAGRAPH REMOVED]

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 in question and may improve the accuracy of model-based

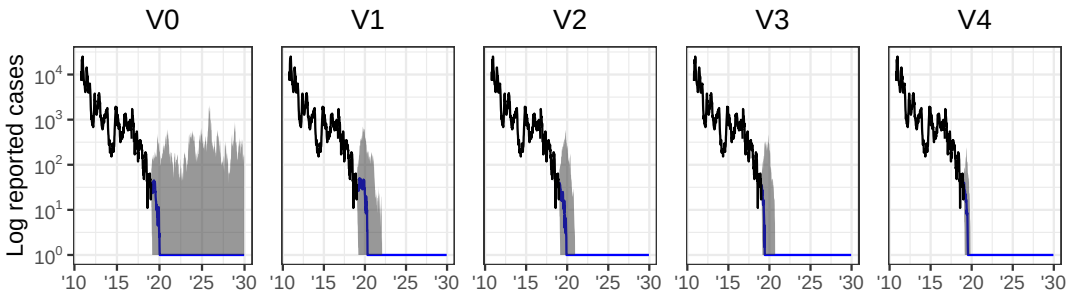

Fig 7. Simulations of Model 1 under each vaccination scenario. Blue line indicates the median of model simulations, and ribbon represents 95% of simulations. The various vaccination campaigns made no practical difference in the median scenario, but a drastic difference in the extreme cases.

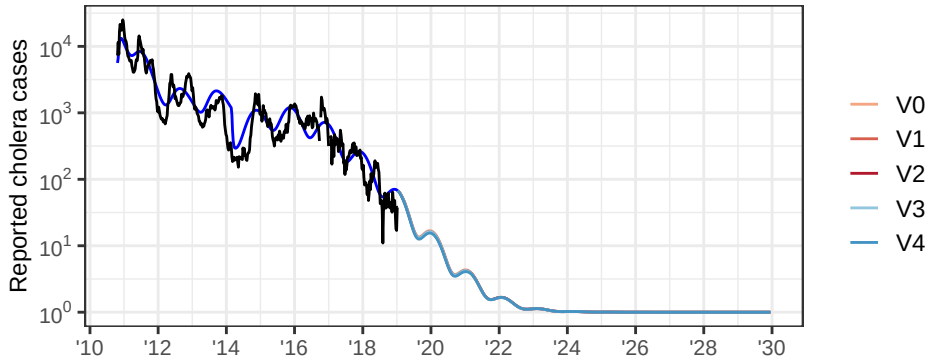

Fig 8. Simulated trajectory of Model 2 (blue curve) and projections under the various vaccination scenarios. Reported cholera incidence is shown in black.

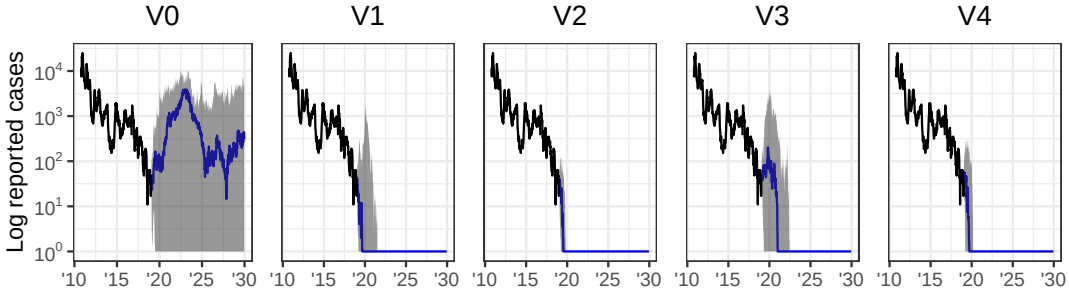

Fig 9. Simulations of Model 3 under each vaccination scenario. Blue line indicates the median of model simulations, and ribbon represents 95% of simulations. The various vaccination campaigns made no practical difference in the median scenario, but a drastic difference in the extreme cases.

projections. Caution is nonetheless needed when making policy based on modeling conclusions, as model misspecification may invalidate conclusions.

Various suggestions been made about why Lee et al. (2020a) failed to accurately predict the eventual eradication of cholera from Haiti, including model misspecification, overly difficult elimination criteria, and a potential conflict of interest (Rebaudet,

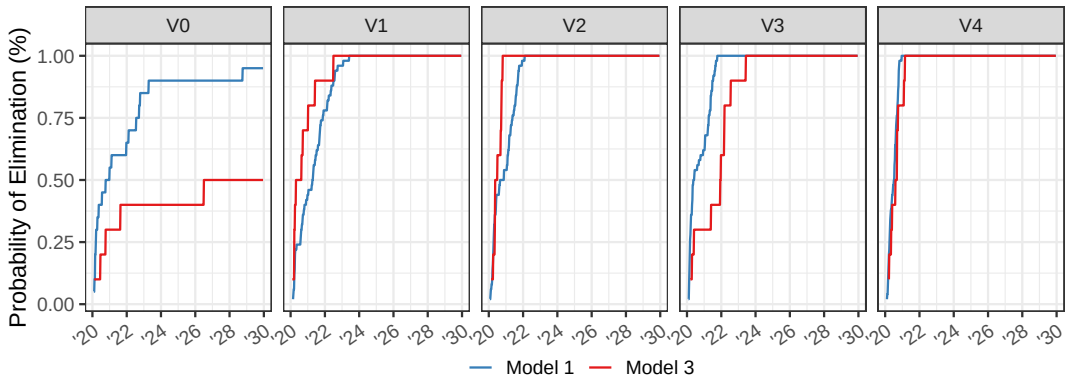

FIG 10. Probability of elimination across simulations for a 10 year period. Compare to Figure 3A of Lee et al. (2020a).

Gaudart and Piarroux, 2020; Henrys et al., 2020). Here, we instead argue that careful attention to important statistical details could have correctly resulted in the conclusion of 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; one example was the inability to fit model parameters to the fully-coupled SpatPOMP 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 fit the SpatPOMP version of 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) 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

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	$\beta_{1:6} = (1.4, 1.2, 1.2, 1.1, 1.4, 0.9)$	(2)	$a = 0.4$	(11)	$a = 19.49$ $r = 1.68 \times 10^5$	(31)
Immunity (year)	$\mu_{RS}^{-1} = 8.0$	(7)	$\mu_{RS}^{-1} = 5$ $\omega_1^{-1} = 1.0$ $\omega_2^{-1} = 0.2$	(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)	$\beta = 1.14 \times 10^{-17}$	(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	—		$W_{\text{sat}} = 10^5$ $\beta_W = 6.08$	(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	—		$\mu_W = 822$	(19)	$\mu_W = 2.71 \times 10^{-5}$	(31)
Water survival (wk)	—		$\delta_W^{-1} = 3.86 \times 10^{11}$	(20)	$\delta_W^{-1} = 0.36$	(32)
Mixing exponent	$\nu = 0.98$	(1)	—		—	
Process noise($\text{wk}^{1/2}$)	$\sigma_{\text{proc}} = (0.32, 0.35)$	(1)	—		$\sigma_{\text{proc}} = 0.032$	(25)
Reporting rate	$\rho = (0.577, 0.833)$	(S16)	$\rho = 0.20$	(S17)	$\rho = 0.89$	(S18)
Observation overdispersion	$\psi = (398.45, 65.71)$	(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.

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: zjiang2/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

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.

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

Domeyer, J. E., Lee, J. D., Toyoda, H., Mehler, B. and Reimer, B. (2022). Driver-Pedestrian Perceptual Models Demonstrate Coupling: Implications for Vehicle Automation. *IEEE Transactions on Human-Machine Systems* 1-10. https://doi.org/10.1109/THMS.2022.3158201

Ferguson, S. (2022). Haiti's winning the fight against cholera. *Forbes*.

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

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

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.

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

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

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.

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

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.