Innovations in Likelihood-Based Inference for State Space Models

by

Jesse Wheeler

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2025

Doctoral Committee:

Professor Edward L. Ionides, Chair
Professor Aaron A. King
Assistant Professor Jeffrey Regier
Professor Kerby Shedden


PUT YOUR OWN FRONTISPIECE IMAGE
HERE IF YOU WANT

Jesse Wheeler
jeswheel@umich.edu
ORCID iD: 0000-0003-3941-3884

ACKNOWLEDGEMENTS

Put your acknowledgements text here. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim

nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF PROGRAMS

PROGRAM

LIST OF APPENDICES

ABSTRACT

State space models are widely used for conducting time series analysis. Developing a state space model involves proposing mathematical equations that describe how a data-generating system evolves over time and how observations of the system are obtained. These models are particularly useful when a scientific hypothesis about system dynamics exists, as is common when modeling ecological populations or tracking infectious disease outbreaks over time. However, except for the simplest cases, state space models do not permit closed-form expressions of their likelihood functions, presenting challenges for inference. This thesis presents three projects that introduce innovations in likelihood-based inference for state space models.

The first project proposes a novel approach for performing inference on Auto Regressive Moving Average (ARMA) time series models, which are formally linearly Gaussian state space models. ARMA models are among the most frequently taught and widely used methods for time series analysis. In this project, I demonstrate that existing algorithms and software for parameter estimation often produce sub-optimal parameter estimates with surprising frequency. I introduce a novel random initialization algorithm designed to leverage the structure of the ARMA likelihood function to help overcome these optimization shortcomings. Additionally, I demonstrate that profile likelihoods offer superior confidence intervals compared to those based on the Fisher information matrix, which is the current standard practice for ARMA modeling.

The second project presents a likelihood-based analysis of the 2010-2019 cholera outbreak in Haiti. This work explores three distinct state space models for cholera incidence data and demonstrates the effectiveness of recently developed algorithms for performing inference in a high-dimensional setting. A key focus of this project is to assess the strengths and limitations of using state space models to inform public health policy decisions. Existing methodologies and workflows for this purpose are evaluated, and revised data analysis strategies that lead to better statistical fit and outcomes are presented. For example, I demonstrate a reproducible framework for diagnosing model misspecification and subsequently developing enhancements that result in better recommendations for policy decisions.

The third project proposes a simulation-based algorithm designed to perform maximum

likelihood estimation for a class of high-dimensional state space models. This algorithm, called the Marginalized Panel Iterated Filter (MPIF), significantly enhances the capability of iterated filtering algorithms to estimate parameters for large collections of independent but related state space models. Improvements in parameter estimates and empirical convergence rates are achieved by addressing the issue of particle depletion that occurs when performing iterated filtering on models that have high-dimensional parameter spaces. Theoretical support for the algorithm is provided through an analysis of iterating marginalized Bayes maps. Additionally, asymptotic theory demonstrating the convergence of general iterated filtering algorithms for panel models without the marginalization step is presented.

CHAPTER 1

Introduction

This is a dissertation template. Here is the introduction text. As with this chapter, add other chapters to the dissertation and cite them in the "main.tex" file.

1.1 Example Section

This dissertation template combines the work of our fore-academics to build a format that works for all of those who use LaTeX[**?**]. We sincerely hope you find it useful [**?**].

You can add acronyms in the text and front table! See the comments around the acronyms command for how to include them in the front table. The first time you use an acronym in the text, use the "\ac{}" command like with this **TLA!** (**TLA!**) or **SOA!** (**SOA!**). After you use them the first time, it'll just appear as the acronym (like this: **TLA!**).

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

1.2 Example Tables

You can make a table as follows [**?**]. Use can cross-reference items by doing things like Chapter 1.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec

ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

left column	**right column**
entry1	entry2
entry3	entry4

Table 1.1: An example table with things in it. See https://www.latex-tables.com/ for a relatively-easy latex table generator. Other options include https://truben.no/table/old/ and https://www.tablesgenerator.com/, but you'll find others online, too! Here, we're also showing that you can create URLs that link to external websites while still being underlined. You can use "\url{}" for shorter URLs, but they will roll over the end of the line if they're too long. In that case, it's more accessible if you use "\href{<url>}{<text>}" instead because it breaks at spaces in the text.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

1.3 Example Figure

Check out Figure 1.1. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit

Figure 1.1: The Big House, taken from Michigan Radio's article about it [**?**]

blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris.

Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

CHAPTER 2

Likelihood Based Inference for ARMA Models

2.1 Introduction

Auto-regressive moving average (ARMA) models are the most well known and frequently used approach to modeling time series data. The general ARMA model was first described by Whittle [21], and popularized by Box and Jenkins [2]. Today, ARMA models are a fundamental component of various academic curricula, leading to their widespread use in both academia and industry. ARMA models are as foundational to time series analysis as linear models are to regression analysis, and they are often used in conjunction for regression with ARMA errors. A Google Scholar search for articles from 2024 onward that include the phrase "time series" and the term "ARMA" (or variants) yields over 18,000 results. While not all these articles focus on the same models discussed here, the importance of this model class to modern science cannot be overstated. Given the ubiquity of ARMA models, even small improvements in parameter estimation constitute a significant advancement of statistical practice.

A commonly used extension of the ARMA model is the *integrated* ARMA model, which extends the class of ARMA models to include first or higher order differences. That is, an autoregressive integrated moving average (ARIMA) model is an ARMA model fit after differencing the data in order to make the data stationary. Additional extensions include the modeling of seasonal components (SARIMA), or the inclusion of external regressors (SARIMAX). Our methodology can readily be extended to these model classes as well, but here we focus on ARMA modeling for simplicity.

We demonstrate that the most commonly used methodologies and software for estimating ARMA model parameters frequently yield sub-optimal estimates. This assertion may seem surprising given the extensive application and study of ARMA models over the past five decades. A natural question arises: if these optimization issues exist, why have they not been addressed? There are three plausible explanations: the first is that the potential for

sub-optimal results has largely gone unnoticed; the second is a satisfactory solution has not yet been discovered; and the third is a general indifference to the problem. It is likely that a combination of these factors has deterred prior exploration of this issue. For instance, most practitioners may be unaware of the problem, while those who have noticed it either did not prioritize it or were unable to provide a general method to resolve it. In this article, we address all three possible explanations by demonstrating the existence of an existing shortcoming, explaining why the problem has nontrivial consequences, and proposing a readily applicable and computationally efficient solution.

Imperfect likelihood optimization has an immediate consequence of complicating standard model selection procedures. Algorithms in widespread use lead to frequent inconsistencies in which a smaller model is found to have a higher maximized likelihood than a larger model within which it is nested. This is mathematically impossible but occurs in practice when the likelihood is imperfectly maximized, and is commonly observed using contemporary methods for ARMA models. Such inconsistencies are a distraction for the interpretation of results even when they do not substantially change the conclusion of the data analysis. Removing numeric inconsistencies can, and should, increase confidence in the correctness of statistical inferences.

There are various software implementations available for the estimation of ARMA model parameters. In this article, we focus on the standard implementations in R (`stats` package) and Python (`statsmodels` module), which we selected due to their widespread usage. While both implementations offer multiple ways to estimate parameters, the default approach in both software packages is to perform likelihood maximization, assuming that the error terms are Gaussian. The challenges in parameter estimation arise in this situation because there is no closed-form expression for the likelihood function, though computational algorithms do exist for maximizing ARMA likelihoods [9].

We begin by providing essential background information on the estimation of ARMA model parameters. We then present our proposed approach for parameter estimation, which leads to parameter values with a likelihood that is never lower and sometimes higher than the standard method. This is followed by a motivating example and a discussion of the potential implications of our proposed method. We also discuss the construction of standard errors for our maximum likelihood estimate. Specifically, we show that estimates of the standard error for model parameters that are default output of R and Python can be misleading, and we provide a reliable alternative. Throughout the article, we use the `stats::arima` function from the R programming language as a baseline for comparison. The same methodology for fitting parameters is used in the `statsmodels.tsa` module in Python, and we demonstrate that our results apply to this software as well (**??**).

2.2 Maximum Likelihood for ARMA Models

Following the notation of Shumway and Stoffer [18], a time series $\{x_t; t = 0, \pm 1, \pm 2, \ldots\}$ is said to be ARMA(p, q) if it is (weakly) stationary and

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q}, \tag{2.1}$$

with $\{w_t; t = 0, \pm 1, \pm 2, \ldots\}$ denoting a mean zero white noise (WN) processes with variance $\sigma_w^2 > 0$, and $\phi_p \neq 0$, $\theta_q \neq 0$. We refer to the positive integers p and q of Eq. 2.1 as the autoregressive (AR) and moving average (MA) orders, respectively. A non-zero intercept could also be added to Eq. 2.1, but for simplicity we assume that the time series $\{x_t; t = 0, \pm 1, \pm 2, \ldots\}$ has zero mean. We denote the set of all model parameters as $\psi = \{\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma_w^2\}$. Our objective is to estimate model parameters using the observed uni-variate time series data.

Given the importance of ARMA models, numerous methods have been developed for parameter estimation. For instance, parameters can be estimated through methods such as Bayesian inference [15, 5] or neural networks [6], among many others. Specialized methods include those for integer-valued data [19], approaches robust to outliers [4], and non-Gaussian ARMA processes [14]. While each methodology has its merits, we focus on the approach most widely adopted in statistical practice: likelihood maximization, under the assumption that the WN process $\{w_t\}$ is Gaussian.

A relevant method for estimating model parameters is minimization of the conditional sum-of-squares (CSS). The CSS estimate is fast to compute, but it does not posses the statistical efficiency of the maximum likelihood estimate (MLE). However, the CSS method plays a role in likelihood maximization, so we briefly describe it here. By solving for the WN term w_t, Eq. 2.1 can be written as

$$w_t = x_t - \sum_{i=1}^{p} \phi_i x_{t-i} - \sum_{j=1}^{q} \theta_j w_{t-j}. \tag{2.2}$$

A natural estimator would involve minimizing the sum of squares $\sum_{t=1}^{n} w_t^2$. However, since only $x_1, x_2, \ldots, x_n$ are observed and w_t is recursively defined in Eq. 2.2 using values of x_{t-p} and w_{t-q}, directly minimizing this sum is intractable. The CSS method addresses this issue by conditioning on the first p values of the process, assuming $w_p = w_{p-1} = \ldots = w_{p+1-q} = 0$, and minimizing the conditioned sum $\sum_{t=p+1}^{n} w_t^2$. While the CSS method provides an attractive solution do to its relative simplicity and easiness to compute, it ignores the error terms for the first few observations. This is particularly concerning when the time series is short or when

there are missing observations. CSS minimization was previously popular because methods for likelihood maximization were considered prohibitively slow, though this is no longer the case with currently available hardware and software [17].

For likelihood maximization, Eq. 2.1 is reformulated as an equivalent state-space model. Although there are several ways this can be done, the approach of [9] is widely used. In this approach, we let $r = \max(p, q+1)$ and extend the set of parameters so that $\bar{\psi} = \{\phi_1, \ldots, \phi_r, \theta_1, \ldots, \theta_{r-1}, \sigma_w^2\}$, with some of the $\phi_i s$ or $\theta_i s$ being equal to zero unless $p = q+1$. We define a latent state vector $z_t \in \mathbb{R}^r$, along with transition matrices $T \in \mathbb{R}^{r \times r}$ and $Q \in \mathbb{R}^{r \times 1}$, enabling the recovery of the original sequence $\{x_t\}$ using Equations 2.3 and 2.4. For a detailed explanation of how to define the latent state z_t and transition matrices T and Q in order to recover the ARMA model, we refer readers to Chapter 3 of Durbin and Koopman [7]. Along with initializations for the mean and variance of z_0, these equations allow for the exact computation of the likelihood of the ARMA model via the Kalman filter [12], which can subsequently be optimized by a numeric procedure such as the BFGS algorithm [8].

$$z_t = T z_{t-1} + Q w_t, \tag{2.3}$$

$$x_t = \begin{pmatrix} 1 & 0 & \ldots & 0 \end{pmatrix} z_t. \tag{2.4}$$

Numeric black-box optimizers require an initial guess for parameter values. For ARMA models, this task is non-trivial because the valid parameter region is defined in terms of the roots of a polynomial associated with the parameters, as discussed in the next section. The default strategy in R and Python is to use the CSS estimator for initialization. This is an effective approach because the CSS estimator asymptotically converges to the MLE [18], and may therefore be close to the global maximum when there are sufficiently many observations. However, the CSS initialization is less useful with limited data, or when there are missing observations. The CSS estimate may also lie outside the valid parameter region, and in such cases, parameters are reinitialized at the origin. Both software implementations also allow for manual selection of initial parameter values, but finding suitable initializations manually can be challenging due to complex parameter inter-dependencies.

The log-likelihood function of ARMA models is often multimodal [17], and therefore this single initialization approach can result in parameter estimates corresponding to local maxima (see Fig 2.1). This is true even for a carefully chosen initialization, such as the CSS estimate. A common strategy to optimize multimodal loss functions is to perform multiple optimizations using different initial parameters. However, we have found no instances of practitioners using a multiple initialization strategy for estimating ARMA model parameters. This may be explained by a general unawareness of the possibility of converging to a local

maximum or because obtaining a suitable collection of initializations for ARMA models is nontrivial. For example, independently initializing parameters at random can place the parameter vector outside the region of interest. Furthermore, uniform random sampling generally fails to adequately cover the plausible parameter region (**??**).

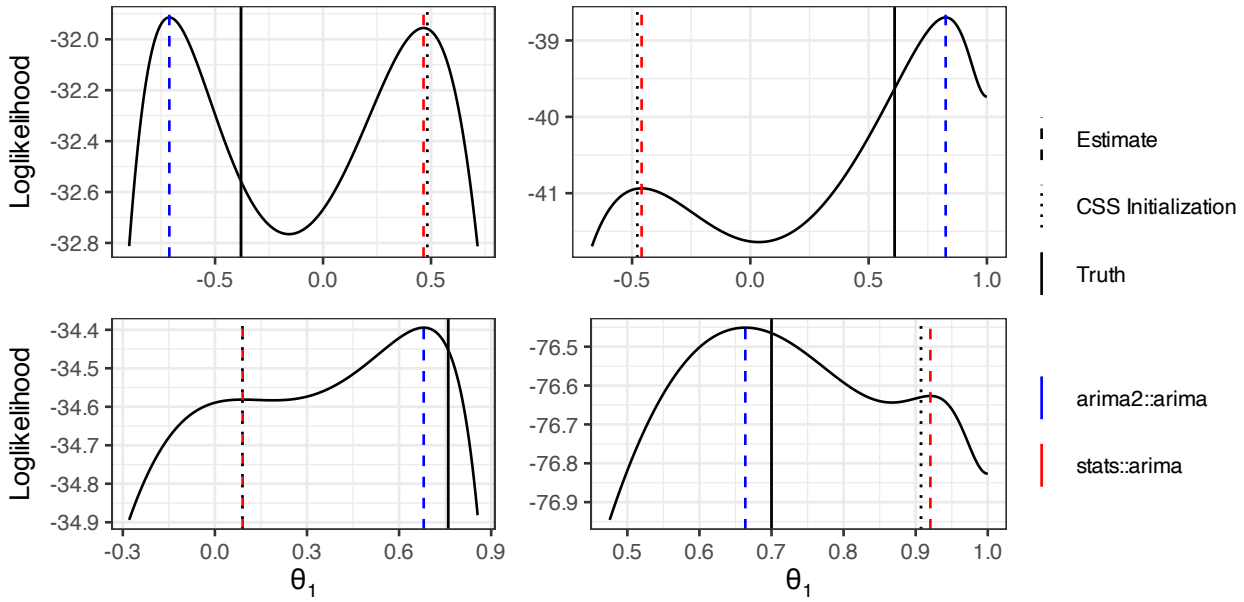

Figure 2.1: The profile log-likelihood of data simulated from four distinct MA(1) models, demonstrating a few examples of multimodal likelihood surfaces. The solid, black line indicates the true value of θ_1; the dotted line is the CSS-initialization. The dashed lines correspond to the estimate $\hat{\theta}_1$ using `stats:arima` (red) and our proposed algorithm (implemented in `arima2::arima`, blue).

2.2.1 A Novel Multi-start Algorithm

To obtain random parameter initializations, parameter sets must correspond to *causal* and *invertible* ARMA processes; definitions are in Chapter 3 of Shumway and Stoffer [18]. Let $\{\phi_i\}_{i=1}^p$ and $\{\theta_i\}_{i=1}^q$ be the coefficients of the ARMA(p, q) model (Eq. 2.1), and define $\Phi(x) = 1-\phi_1 x-\phi_2 x^2-\ldots-\phi_p x^p$ as the AR polynomial, and $\Theta(x) = 1+\theta_1 x+\theta_2 x^2+\ldots+\theta_q x^q$ as the MA polynomial. An ARMA model is *causal* and *invertible* if the roots of the AR and MA polynomials lie outside the complex unit circle. Therefore in order to obtain valid, random parameter initializations, we sample the roots of $\Phi(x)$ and $\Theta(x)$ and use these values to reconstruct parameter initializations.

It is an easier task to sample *inverted* roots, as the sufficient conditions for causality and invertibility now require that the inverted roots lie inside the complex unit circle, a region

easier to sample uniformly. The roots of the polynomials can be real or complex; complex roots must come in complex conjugate pairs in order for all of the corresponding model parameters to be real. The simplest approach would be to sample all inverted root pairs (z_1, z_2) within the complex unit circle by uniformly sampling angles and radii (lines 10-14 of Algorithm 1). However, this would imply almost surely all root pairs are complex, and some model parameters would only be sampled as positive (or negative). For instance, consider an AR(2) model. The AR polynomial is:

$$\Phi(x) = 1 - \phi_1 x - \phi_2 x^2 = (1 - z_1 x)(1 - z_2 x) = 1 - (z_1 + z_2)x - (z_1 z_2)x^2$$

In this equation, if both z_1, z_2 are complex conjugates, then $\phi_2 = z_1 z_2 > 0$. As such, the only way that $\phi_2 < 0$ is if $z_1, z_2 \in \mathbb{R}$. Similar results hold for the MA coefficients with opposite signs for the coefficients. This issue is directly addressed in lines 5-8 of Algorithm 1: root pairs are sampled as real with probability $p = \sqrt{1/2}$, and real pairs sampled with the same sign with probability p, such that the product (and sums) of each pair is positive with probability 1/2. We sample conjugate pairs within an annular disk on the complex plane to avoid trivial and approximately non-stationary cases. The radii of both the inner and outer circles defining the disk are defined using γ in lines 7, 10, and 13 of Algorithm 1.

Parameter redundancy in ARMA models occurs when the polynomials $\Phi(x)$ and $\Theta(x)$ share one or more roots, leading to an overall reduction in model order. This complicates parameter initialization, optimization, and identifiability. The ARMA model (Eq. 2.1) can be rewritten as:

$$\Phi(B)x_t = \Theta(B)w_t, \tag{2.5}$$

where B is the *backshift* operator, i.e., $Bx_t = x_{t-1}$. Using the fundamental theorem of algebra, Equation 2.5 can be factored into

$$(1 - \lambda_1 B)\dots(1 - \lambda_p B)x_t = (1 - \nu_1 B)\dots(1 - \nu_q B)w_t,$$

where $\{\lambda_i\}_{i=1}^p$ and $\{\nu_j\}_{j=1}^q$ are the inverted roots of $\Phi(B)$ and $\Theta(B)$, respectively. If $\lambda_i = \nu_j$ for any $(i, j) \in \{1, \dots, p\} \times \{1, \dots, q\}$, then the roots will cancel each other out, resulting in an ARMA model of smaller order. As an elementary example, consider the ARMA$(1, 1)$ and

Algorithm 1: MLE for ARMA Models.
Inputs (defaults):
First parameter initialization $\psi_0 = (\phi_1^0, \ldots, \phi_p^0, \theta_1^0, \ldots, \theta_q^0)$ (CSS estimate).
Minimum acceptable polynomial root distance $\alpha > 0$, $(\alpha = 0.01)$.
Probability of sampling a root pairs as real $0 \leq p \leq 1$, $(p = \sqrt{1/2})$.
Bounds on inverted polynomial roots $\gamma \in (0, 0.5)$, $(\gamma = 0.05)$.
Numeric optimization routine $f(\psi)$ [9].
Stopping Criterion (stop if last M iterations do not improve log-likelihood $\ell(\psi)$).

Get preliminary estimate: $\hat{\psi}_0 = f(\psi_0)$; set $k = 0$;
repeat Until stopping criterion met
    Set AR and MA roots $\{z^{\text{AR}}\}_{i=1}^{p} = 0_p$, $\{z^{\text{MA}}\}_{i=1}^{q} = 0_q$; increment k;
    while $\min_{i,j} |z_i^{AR} - z_j^{MA}| < \alpha$, *for both AR and MA polynomials* **do**
        Sample paired roots as real with probability p;
        for *all real pairs* **do**
            Sample root magnitudes from $U(\gamma, 1-\gamma)$;
            Sample signs with $P(\text{sign}(z_1) = \text{sign}(z_2)) = p$;
        for *all complex pairs* **do**
            sample angle: $\tau \sim U(0, \pi)$; sample radius: $r \sim U(\gamma, 1-\gamma)$;
            set $z_1 = r\cos(\tau) + ir\sin(\tau)$; set $z_2 = \bar{z}_1$;
        if *Number of roots is odd (non-paired root)* **then**
            sample τ uniformly from the set $\{0, \pi\}$; sample $r \sim U(\gamma, 1-\gamma)$;
            set $z = r\cos(\tau)$;
    Calculate coefficients $\psi_k = (\phi_1^k, \ldots, \phi_p^k, \theta_1^k, \ldots, \theta_q^k)$ using sampled roots;
    Estimate $\hat{\psi}_k = f(\psi_k)$;
until;
Set $\hat{\psi} = \arg\max_{j \in 0:k} \ell\left(\hat{\psi}_j\right)$;

ARMA$(2, 2)$ models in equations 2.6 and 2.7.

$$x_t = \frac{1}{3}x_{t-1} + w_t + \frac{2}{3}w_{t-1}, \tag{2.6}$$

$$x_t = \frac{5}{6}x_{t-1} - \frac{1}{6}x_{t-2} + w_t + \frac{1}{6}w_{t-1} - \frac{1}{3}w_{t-2}. \tag{2.7}$$

While these two models appear distinct at first glance, re-writing the models in polynomial form (Eq. 2.5) shows that these two models are actually equivalent after canceling out the common factors on each side of the equation.

In a similar fashion, it is possible that the roots are not exactly equal but are approximately equal. In this case, the ratio of factors becomes close to one, resulting in a similar effect to when the roots exactly cancel. We avoid the possibility of *nearly canceling roots* in parameter initializations by requiring the minimum Euclidean distance between inverted polynomial

roots to be greater than α. This is done in line 4 of Algorithm 1, though the condition is rarely triggered if the order of the model is of typical size ($p, q < 4$).

Our sampling scheme is combined with existing procedures for numeric optimization of model log-likelihoods (lines 1 and 16), as well as the default initialization strategy for ψ_0 used by existing software (such as the CSS initialization). Doing so guarantees that final estimates correspond to likelihood values greater than or equal to the currently accepted standards in the software environment where the algorithm is implemented. For this article, both the numeric optimization procedure $f(\cdot)$ and the parameter initialization strategy to obtain ψ_0 are those implemented in `stats::arima`. The stopping criterion was chosen so that the algorithm stops trying new initial values when no new maximum has been found using the last M parameter initializations. Alternative stopping criterion can be used (see for example, [10]), but we found that this simple heuristic works well in practice.

Algorithm 1 is implemented in the R package `arima2`, available on the Comprehensive R Archive Network (CRAN) [20]. The package features the function `arima2::arima`, which is an adaptation of the `stats::arima` function modified to incorporate the adjustments specified by Algorithm 1.

2.2.2 Simulation Studies

```
## data/sim1_results.rds
```

To investigate the extent to which the standard approach for ARMA parameter estimation results in improperly maximized likelihoods, we conduct a series of simulation studies. It is challenging to obtain precise estimates of how frequently current standards lead to sub-optimal parameter estimates due to the varied applications of ARMA models in practice, the diversity in data sizes (n) and model orders (p, q), and the differing degrees to which an ARMA model adequately describes the data-generating process. Therefore, we restrict our simulation studies to idealized scenarios where the data-generating process is Gaussian-ARMA, recognizing that likelihood maximization is easiest for this model class, thereby resulting in conservative estimates of how frequently our algorithm improves model likelihood.

In the first simulation study, we simulate time series data of lengths $n \in \{50, 100, 500, 1000\}$ from Gaussian-ARMA models with known orders $(p, q) \in \{1, 2, 3\}^2$, generating 36,000 unique models and datasets. We avoid any models that contain parameter redundancies by requiring the data generating model to have a minimum distance of 0.1 between all roots of $\Phi(x)$ and $\Theta(x)$. We further restrict model coefficients so that they do not lie near boundary conditions. Models of the same order of the generating data are fit to the data, simplifying the the

problem further by avoiding the order selection step that is necessary in most data analyses. In doing so, we attempt to answer the question of how often sub-optimal estimates may arise using existing software in the case where the parameter estimation procedure should be as easy as possible for the given combinations of (n, p, q).

Even in this extremely simplified scenario, existing software failed to properly maximize model likelihoods in at least 20.8% of the simulated datasets—evidenced by an improvement obtained using Algorithm 1. Though this improvement may appear modest, an improvement in 20.8% of the large number of published ARMA models would affect many papers—a number measured in thousands of papers since 2024 following our estimate in the introduction. Furthermore, time series analysis courses and textbooks often recommend fitting multiple ARMA models to a dataset, and here we only fit one for each algorithm. Consequently, the probability that at least one candidate model is not properly optimized increases significantly in practice. The rate of improvement obtained using our algorithm increases with model complexity and decreases with more observations (Fig. 2.2). For example, likelihoods improved in 55.1% of the simulations when $p = q = 3$ and $n = 50$; models of this size and number of observations are not uncommon in published research studies.

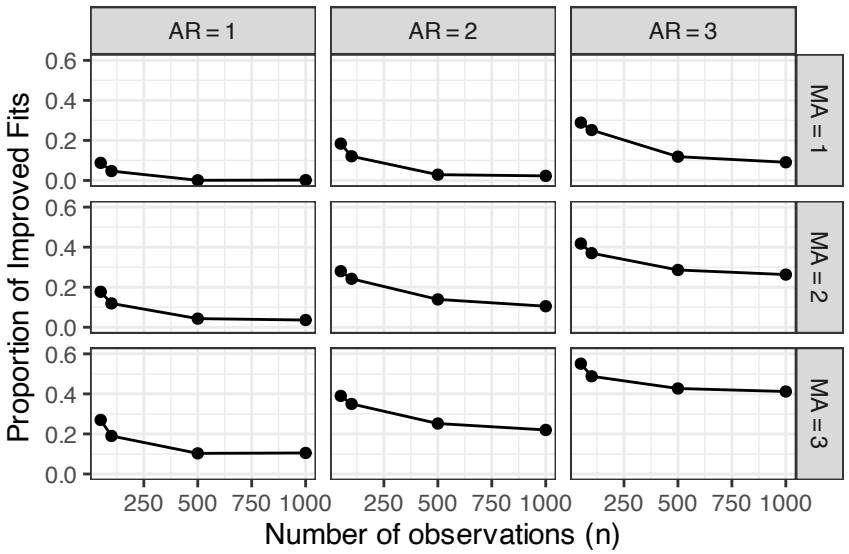

Figure 2.2: Proportion of simulated data with improved likelihood from using multiple restarts (Algorithm 1).

Importantly, we do not claim to improve likelihood for all ARMA models and datasets. In this simulation study, our likelihood maximum routine did not improve likelihoods for many of the simulated data. However, these results demonstrate that there is a very real potential for obtaining sub-optimal parameter estimates when using only a single parameter

initialization, even in the most idealistic scenarios. Rather than having to worry if a single initialization is sufficient to fit a given model, it is preferable to adopt methods that make such situations rare. The primary limitation of our algorithm is that the potential for improved fits comes at the cost of increased computation times. In our simulation study, however, the average time to estimate parameters using our approach was 0.6 seconds, a computational expense that is worth the effort in many situations.

The median log-likelihood improvement in this simulation study was 0.66, with an interquartile range of $(0.22, 1.46)$. Among the most common motivations for fitting ARMA models is to model serial correlations in a regression model; in this setting, the discovered shortcomings in log-likelihood are often enough to change the outcome of the analysis. For instance, consider modeling $y_i = \beta x_i + \epsilon_i$, where $\beta \in \mathbb{R}$, and we model the error terms $\epsilon_i \sim \text{ARMA}(p, q)$. For now, we will assume the order (p, q) is fixed. We may wish to test the hypothesis $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$. A standard approach to doing this is a likelihood ratio test, and using Wilks' theorem to get an approximate test. We denote ll_0 and ll_1 as the maximum log-likelihood of the model under H_0 and H_1, respectively. The standard approximation is to assume $2\Delta = 2(ll_1 - ll_0) \sim \chi^2_1$. Using a significance level of $\alpha = 0.05$, we would reject H_0 if $\Delta \geq 1.92$. Given that $E_{H_0}[\Delta] = 0.5$, subtracting the reported log-likelihood deficiencies (which has a median value of 0.66) of existing software to either or both ll_0, ll_1 could change the outcome of this test.

2.2.2.1 Parameter Uncertainty

Improved parameter estimation leads to modified standard error estimates, which are default outputs in R and Python. These standard errors result from the numeric optimizer's estimate of the gradient of the log-likelihood, used to approximate the Fisher information matrix. These standard errors, though not inherently of interest, are sometimes used justify the inclusion of a parameter in a model [3, Chatper 9]. We extend our simulation study to examine this approach and how our algorithm impacts the estimates. For each of the 36,000 generative models from the previous study, we generate 100 additional datasets, estimating the MLE and Bonferroni-adjusted 95% confidence intervals using both estimated standard errors and profile likelihood confidence intervals (PLCIs) from Wilks' theorem. Fig 2.3 shows PLCIs had better or equivalent nominal coverage than Fisher-based confidence intervals across all combinations of p, q, and n. We found that the interval estimation method mattered more for confidence interval performance than the specific parameter estimation algorithm. The relevance of this result is explored further in Section 2.3.1.

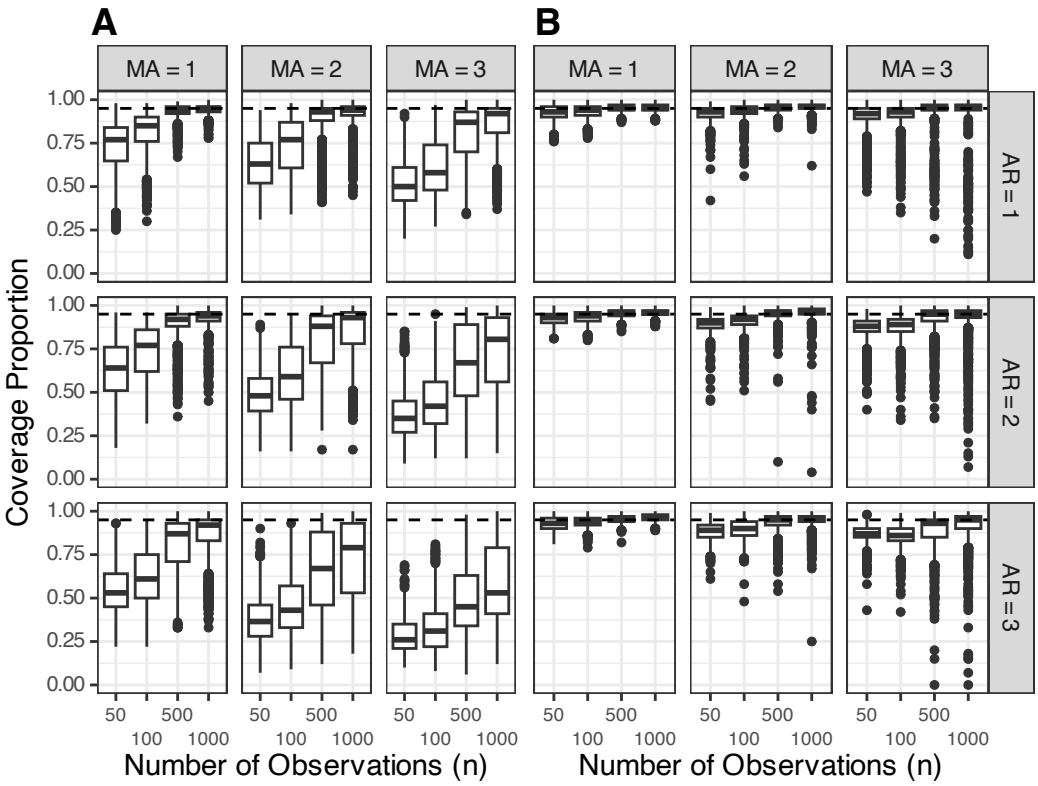

Figure 2.3: Proportion of models that achieved nominal coverage of Bonferroni adjusted 95% confidence intervals. The dashed line denotes the target coverage level. (A) Confidence intervals created using Fisher's information matrix. (B) Confidence intervals created using profile likelihoods.

2.2.2.2 AIC Table Consistency

A more realistic situation than the previous simulation study involves estimating the model order (p, q) as well as obtaining parameter estimates. Fitting multiple models raises the chance that at least one candidate model was not properly optimized. It may also necessitate fitting larger models than needed, leading to parameter redundancies that make proper optimization more challenging.

A contemporary approach involves fitting several candidate models and selecting the one that minimizes a criterion like Akaike's information criterion (AIC) [1]. This can be done by explicitly creating a table of all candidate models and their corresponding AIC values; in this case issues of improper maximization become more apparent. For instance, a table of AIC values may contain numeric inconsistencies, where a larger model may have lower estimated likelihoods than a smaller model within which it is nested (for an example, see Section 2.3). This type of result can make a careful practitioner feel uneasy, as there is

evidence that at least one candidate model was not properly optimized. Evidence of improper optimization may be less evident when relying on software that automates this process, such as the automated Hyndman-Khandakar algorithm [11], but the potential for sub-optimal estimates remains.

We conducted an additional simulation study to investigate numeric inconsistencies that may arise when fitting multiple model parameters. As before, we simulated 1000 unique models and datasets of size $n \in \{50, 100, 500, 1000\}$ from Gaussian ARMA(p, q) models for $(p, q) \in \{1, 2, 3\}^2$. To avoid models with parameter redundancies, we ensured a minimum distance of 0.1 between all roots of $\Phi(x)$ and $\Theta(x)$ and excluded models with coefficients near boundary conditions. For each dataset, AIC tables were created for model sizes $(p, q) \in \{0, 1, 2, 3\}^2$.

The single parameter initialization approach resulted in AIC table inconsistencies in 45.6% of the simulated datasets. Although our proposed algorithm significantly mitigates this issue, it does not guarantee that all model likelihoods are fully maximized. This is illustrated in Fig 2.4, where a non-zero percentage of AIC tables remain inconsistent, even as the algorithm's stopping criterion grows. The ARMA$(1, 1)$ panel in Fig 2.4 illustrates the increasing difficulty of parameter estimation when dealing with parameter redundancies. In such cases, it is often necessary to adjust additional parameters in the numeric optimization routine. For example, our R implementation of the algorithm relies on the generic BFGS optimizer in the `stats::optim` function. Modifying the optimization method or the default hyperparameters can lead to improved fits or faster convergence rates. While the default parameters of the numeric optimizer are generally adequate, increasing the maximum number of algorithmic iterations can be beneficial for fully maximizing the likelihood for challenging models and data.

The contemporary approach of using AIC—or any other information based criteria—to select model order involves fitting unnecessarily large models, leading to parameter redundancies that complicate likelihood optimization. The use of AIC for ARMA model order selection has theoretical support, particularly for forecasting, as ARMA models inspired the original AIC paper [1]. However, without proper likelihood maximization, a strategy that considers only a single parameter initialization may not truly minimize AIC. In this framework, likelihood maximization and over-parameterization are interconnected: all candidate models must be maximized for likelihood, or users risk selecting over-parameterized models that fail to minimize the intended information criterion. For the current study, the choice of AIC versus other popular information criteria such as the corrected AIC (AICC) or Bayesian information criterion (BIC) is unimportant: all of these approaches rely on proper optimization of the likelihood function, which is the problem we are addressing here.

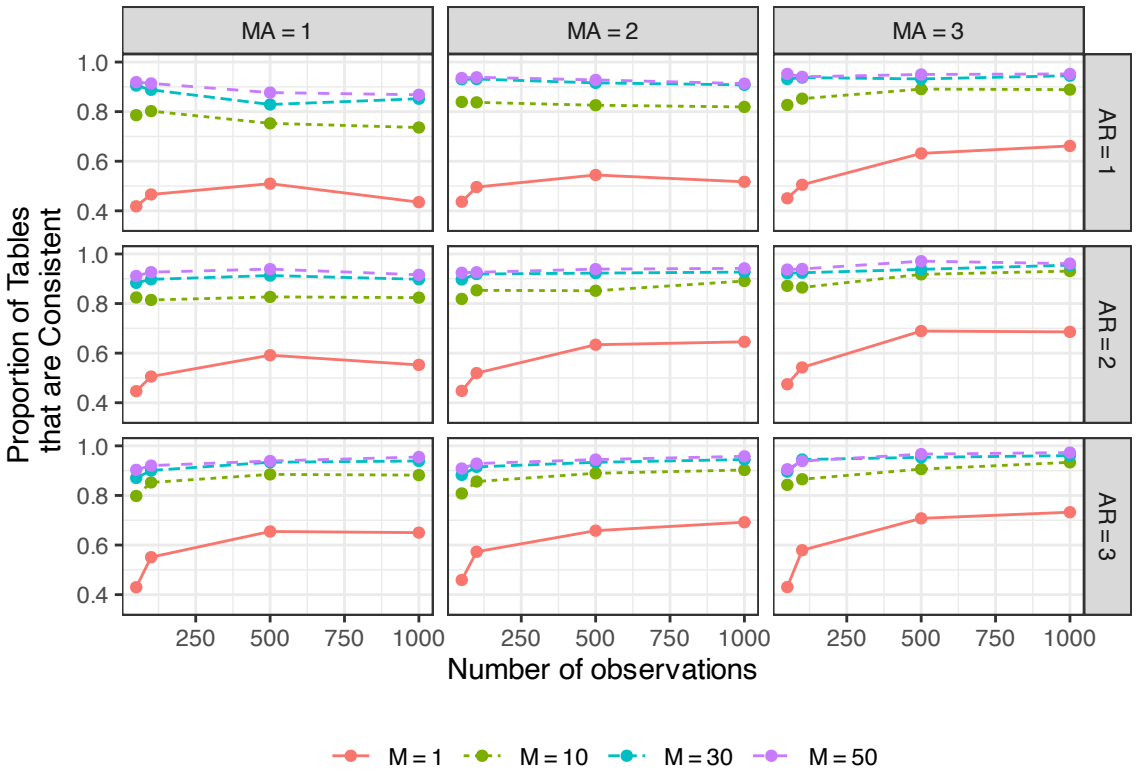

Figure 2.4: Data is generated from ARMA(p, q) models with $(p, q) \in \{1, 2, 3\}^2$, and the corresponding AIC table is created. The Y-axis shows the percentage of tables that were consistent. M is the number of times a maxima is observed before the algorithm terminates, so $M = 1$ corresponds to the standard maximization procedure.

Classical ARMA modeling addresses this by recommending diagnostic plots to determine appropriate model order and advising against simultaneously adding AR and MA components. In this approach, the additional difficulty in parameter estimation associated with fitting models containing parameter redundancies is avoided by not fitting overly complex models when possible. Despite this, shortcomings in likelihood maximization can occur even in models without parameter redundancies (Figs 2.1, 2.2, and 2.4), necessitating the exploration of multiple parameter initializations. Further, the increasing preference of using automated software to pick the model size using an information criterion suggests the importance of using software that reliably maximizes model likelihoods even in the presence of over-parameterization.

2.3 Annual Depths of Lake Michigan

In this example, we illustrate how improperly maximized likelihoods can lead to inconsistencies and uncertainty in a real data analysis scenario. Additionally, we show how the common practice of using the estimated standard error for calibrated parameters can misleadingly support the inclusion of model parameters. We consider a dataset containing annual observations on the average depth of Lake Michigan-Huron, recorded the first day of each year from 1860-2014 (Fig 2.5) [16]. We wish to develop an ARMA model for these data, which is a standard task in time series analysis [18].

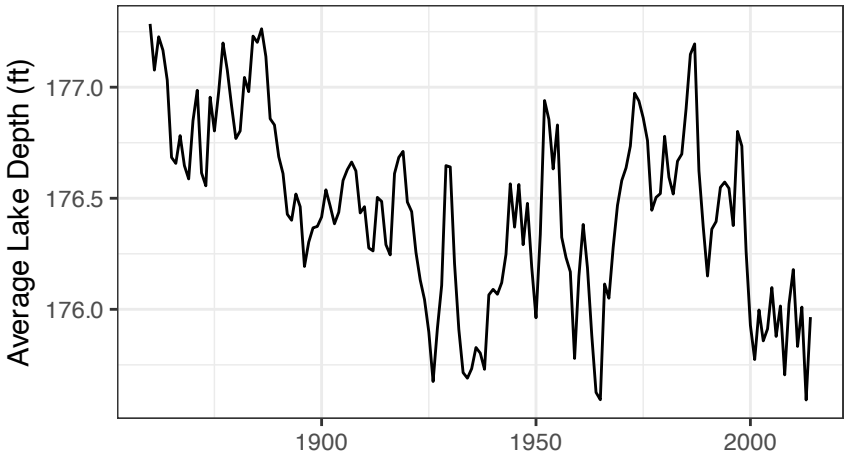

Figure 2.5: Average depth of Lake Michigan-Huron from 1860-2014.

Diagnostic tests, such as sample autocorrelation and normal quantile plots for residuals, suggest that it is reasonable to model the data in Fig 2.5 as a weakly stationary Gaussian ARMA(p, q) process for some non-negative integers p and q. While an ARIMA model may also be reasonable for these data, we first consider fitting an ARMA model because we would like to avoid the possibility of over-differencing the data. The next step is to determine appropriate values of p and q; after some initial investigation, multiple combinations of p and q seem plausible, and therefore we decide to choose the values of p and q that minimize the AIC. For simplicity, we create a table of AIC values for all possible combinations of $(p, q) \in \{0, 1, 2, 3\}^2$ (Table 2.1). Using the AIC as the model selection criterion, the selected model size is ARMA$(2, 1)$.

Recall that the AIC is defined as:

$$\text{AIC} = -2 \max_{\psi} \ell(\psi; x^*) + 2d, \tag{2.8}$$

where $\ell(\psi; x^*)$ denotes the log-likelihood of a model indexed by parameter vector $\psi \in \mathbb{R}^d$,

(a) Single parameter initialization.

	MA0	MA1	MA2	MA3
AR0	166.8	46.6	7.3	-15.0
AR1	-38.0	-37.4	-35.5	-33.8
AR2	-37.3	-38.4	-36.9	-34.9
AR3	-35.5	-35.2	-33.0	-33.3

(b) Multiple parameter initializations.

	MA0	MA1	MA2	MA3
AR0	166.8	46.6	7.3	-15.0
AR1	-38.0	-37.4	-35.5	-33.8
AR2	-37.3	-38.4	-36.9	-34.9
AR3	-35.5	-36.9	-36.4	-36.2

Table 2.1: AIC values for an ARMA(p, q) model fit to Lake Michigan-Huron depths. Table 2.1a was computing using only a single parameter initialization. Table 2.1b was computed using Algorithm 1. Highlighted cells show where the likelihood was improved (AIC reduced) using our algorithm.

$d \geq 1$, given the observed data x^*. In the case of an ARMA model with an intercept, $d = p + q + 2$, where the additional parameter corresponds to a variance estimate. If either p or q increases by one, then a corresponding increase in AIC values greater than two suggests that the *inclusion* of an additional parameter resulted in a *decrease* in the maximum of the log-likelihood, which is mathematically impossible under proper optimization. Several such cases are present in Table 2.1a, for example increasing from an ARMA$(2, 2)$ model to a ARMA$(3, 2)$ model results in a decrease of 1.0 log-likelihood units. In this case, using our multiple restart algorithm eliminates all instances of mathematical inconsistencies (Table 2.1b). We refer to tables that have log-likelihood values larger for any smaller nested model within the table as *inconsistent*.

Suppose a scientist is confronted with a mathematically implausible table of nominally maximized likelihoods (Table 2.1a). How much should they worry about this? Is it acceptable to publish scientific results that demonstrate a nominally maximized likelihood is not, in fact, maximized? Can researchers confidently trust the scientific implications of a fitted model if there is evidence of improper optimization in some of the candidate models? Given a choice, a researcher should prefer to use maximization algorithms reliable enough to make such situations rare. In the Lake Michigan example, improved estimation does not change which model is selected or the final parameter estimates, but it does remove inconsistencies that could lead to these concerns (Table 2.1b).

Minimizing the AIC (or an alternative information criterion) is not the only accepted approach to order selection. A classical perspective on model selection involves consulting sample autocorrelation plots, partial autocorrelation plots, conducting tests such as Ljung-Box over various lags, studying the polynomial roots of fitted models, and checking properties of the residuals of the fitted models [2, 3, 18]. This approach helps avoid fitting models that are possibly over-parameterized. However, additional computational power and increasing

Table 2.2: Parameter values of ARMA(p, q) model fit to Lake Michigan-Huron depth data.

	ϕ_1	ϕ_2	θ_1	Intercept
Estimate	-0.053	0.791	1.000	176.460
s.e.	0.052	0.053	0.024	0.121

volumes of data have favored automated data analysis strategies that fit many models and evaluate them using a model selection criterion. In principle, a simple model selection criterion such as AIC can address parsimony and guard against over-parameterization as well. Diagnostic inspection can be combined with these automated approaches. For example, a table of AIC values can be generated, and models with promising likelihoods can be explored further [3].

When possible, there may be general agreement that the best approach is to combine modern computational resources with careful attention to model diagnostics, considering the data and the scientific task at hand. Improved maximization facilitates this process by eliminating distractions resulting from incomplete maximization.

2.3.1 Parameter uncertainty

Default output from fitting an ARMA model in R or Python includes estimates for parameter values and their standard errors, calculated using Fisher's information matrix. If the ARMA model with the lowest AIC value is chosen to describe the Lake Michigan data, then an ARMA$(2, 1)$ model is selected. The estimated coefficients and standard errors obtained after fitting this model are reported in Table 2.2. The small standard error for $\hat{\theta}_1$ reported in this table suggests a high-level of confidence that the parameter has a value near 1. Taken at face value, these estimates seem to strongly favor the inclusion of the MA(1) term in the model.

However, our simulation studies have suggested that these confidence intervals can be misleading, and that PLCIs are more reliable alternatives. The 95% PLCI for the parameter (Fig 2.6A) is much larger than the confidence interval created using these standard errors. The steep curve in the immediate vicinity of $\hat{\theta}_1$ may explain the small standard error estimates for this parameter and the corresponding tight confidence intervals created using Fisher's identity matrix. Alternative evidence indicates the potential for nearly canceling roots (Fig 2.7), in which case the MA(1) term may not be needed in the model.

Both types of confidence intervals considered in this example rely on asymptotic justifications, but we can further investigate the finite sample properties using a simulation study. We fit both ARMA$(2, 1)$ and AR(1) models to the data, and conduct a boot-strap simulation study by simulating 1000 datasets from each of the fitted models. We then re-estimate

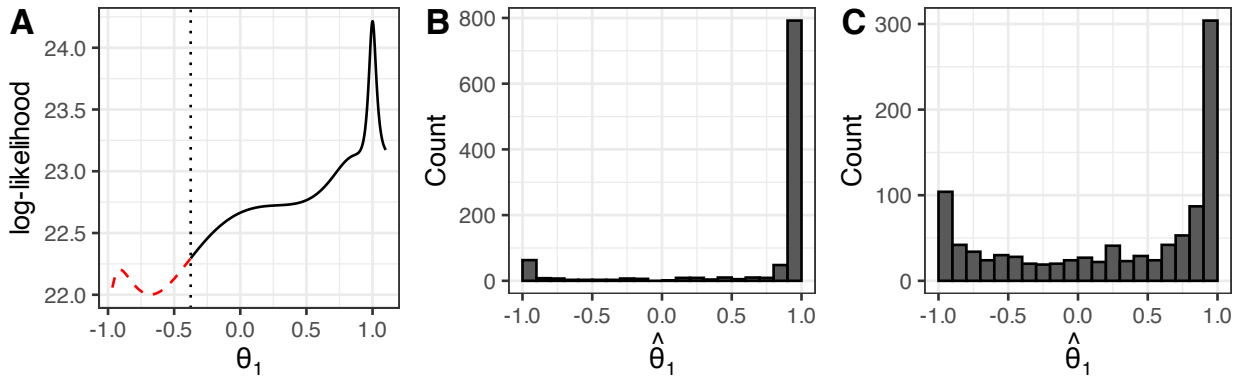

Figure 2.6: Evidence for an AR(1) model for the Lake Michigan-Huron data. (A) Profile likelihood confidence interval (PLCI) for θ_1 which includes the value $\theta_1 = 0$. The vertical dotted line represents the lower end of the approximate confidence interval; all points on the solid black line lie within the confidence interval, and points on the dashed red line are outside the interval. (B) Histogram of re-estimated θ_1 values using simulated data simulated from the ARMA$(2, 1)$ model that was calibrated to the Lake Michigan-Huron data. (C) Histogram of re-estimated θ_1 values using data simulated from the AR(1) model that was calibrated to the Lake Michigan-Huron data.

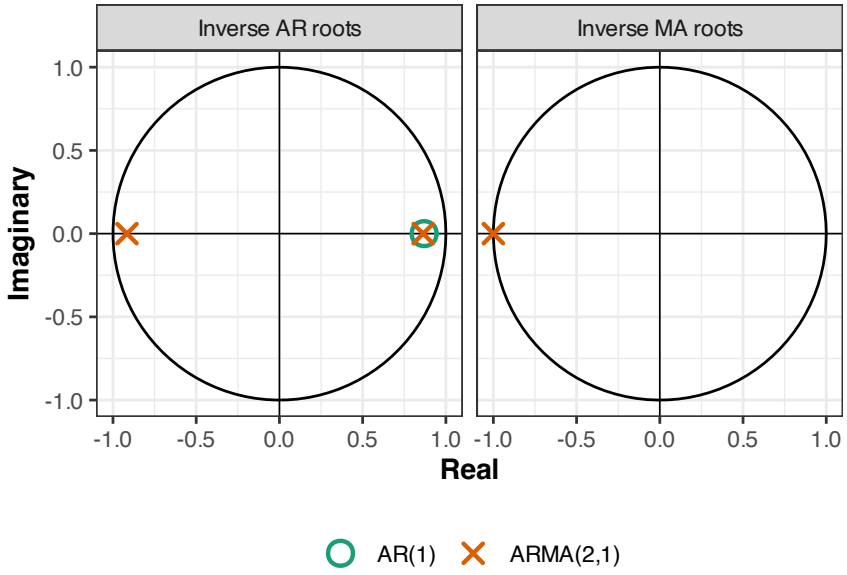

Figure 2.7: Inverted AR and MA polynomial roots to the fitted AR(1) and ARMA$(2, 1)$ models to the Lake Michigan-Huron data using a single parameter initialization.

an ARMA$(2, 1)$ model to each of these datasets and record the estimated coefficients. A histogram containing the estimated values of $\hat{\theta}_1$ when the data are generated from the ARMA$(2, 1)$ and AR(1) models fit to the Lake Huron-Michigan data are shown in Fig 2.6B

and Fig 2.6C, respectively.

The shape of this histogram in Fig 2.6B mimics that of the profile log-likelihood surface in Fig 2.6A, confirming that a large confidence interval is needed in order to obtain a 95% confidence interval. In Fig 2.6C, a large number of $\hat{\theta}_1$ coefficients are estimated near 1 when the generating model is AR(1). Combining this result with the nearly canceling roots of the ARMA$(2, 1)$ model (Fig 2.7), we cannot reject the hypothesis that the data were generated from a AR(1) model, even though the Fisher information standard errors suggest that the data should be modeled with a nonzero θ_1 coefficient.

2.4 Discussion

A significant motivation for our work is the observation that commonly used statistical software that purports to maximize ARMA model likelihoods fails to do so for a large number of examples. In addition to improving parameter estimates, proper maximization of ARMA model likelihoods is crucial because ARMA models are often used to model serial correlations in regression analyses. In this context, researchers may perform likelihood ratio hypothesis tests for regression coefficients, and the validity of these tests depends on proper likelihood optimization.

An important consequence of improved likelihood maximization is better model selection. A common approach to selecting an ARMA model involves fitting different sizes of models and choosing the one that minimizes an information criterion, such as the AIC. Fitting multiple models results in having a higher probability that at least one candidate model was not properly maximized. Since AIC assumes the parameters correspond to maximized likelihoods, enhancements in likelihood maximization can lead to different model selections. Consequently, methodology relying on existing estimation methods—like the popular `auto.arima` function in the `forecast` package in R [11], which minimizes the AIC of a group of candidate models without explicitly displaying an AIC table—will be impacted by improved estimates.

Our proposed algorithm is supported by existing theory on likelihood evaluation of linear state-space models via the Kalman Filter [12], the same as the current existing standard approach for parameter estimation. The simulation studies that we have conducted, however, demonstrate the importance of considering multiple parameter initializations in order to fully maximize model likelihoods. These simulations provide a conservative estimate of how frequently our algorithm results in improved likelihoods compared to existing standards. A common situation where our algorithm is expected to provide even larger improvements than those reported here is in the presence of missing data, a primary motivator of the likelihood maximization procedure of existing software [17]. In this situation, the well-informed CSS

initialization is not available, and the default approach is to initialize at the origin, resulting in a greater need to attempt multiple parameter initializations.

Parameter estimates corresponding to higher likelihood values are not necessarily scientifically preferrable to alternative regions of parameter space with lower likelihood values [13]. Sometimes, our improved estimates may result in models with nearly canceling roots, parameters near boundary conditions, or otherwise unfavorable statistical properties. On other occasions, our method can rescue a naive optimization attempt from a local maximum having those unfavorable properties. Practitioners should carefully evaluate fitted models to ensure they are appropriate for the data and problem at hand.

The primary limitation of our approach is that it achieves higher likelihoods at the cost of processing speed, which is more pronounced with large datasets. However, our algorithm is most necessary for small datasets ($n \ll 10000$), where default parameter initialization strategies may perform poorly. Therefore, our algorithm is most beneficial for small to moderate sample sizes, where the additional computational cost is generally negligible. The compute time of our algorithm is approximately K times slower than the default approach, where K is the number of unique parameter initializations. This is only an approximation of the actual additional cost as not all initializations require the same amount of processing time in order to converge. In particular, initializations that are already close to local maximum will generally converge much quicker than those that are further away.

Our proposed algorithm for ARMA parameter estimation significantly advances statistical practice by addressing a frequently occurring optimization deficiency. Because existing software can also be leveraged to mitigate the issue, the largest contribution of this work may be highlighting the prevalence of this optimization problem. Traditional random initialization approaches software fail to uniformly cover the entire range of possible models and often produce many initializations outside the accepted range. Our algorithm offers a computationally efficient and practically convenient solution, providing a robust approach to parameter initialization and estimation that ensures adequate coverage of all possible models. We have shown that it provides a new standard for best practice in the field of time series analysis.

CHAPTER 3

Informing policy via dynamic models: Cholera in Haiti

Abstract

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. We assess current methodological approaches to these issues via a case study of the 2010-2019 cholera epidemic in Haiti. We consider three dynamic models developed by expert teams to advise on vaccination policies. We evaluate previous methods used for fitting these models, and we demonstrate modified data analysis strategies leading to improved statistical fit. Specifically, we present approaches for diagnosing model misspecification and the consequent development of improved models. Additionally, we demonstrate the utility of recent advances in likelihood maximization for high-dimensional nonlinear dynamic models, enabling likelihood-based inference for spatiotemporal incidence data using this class of models. Our workflow is reproducible and extendable, facilitating future investigations of this disease system.

Author summary

Quantitative understanding of infectious disease transmission dynamics relies upon mathematical models informed by scientific knowledge and relevant data. The models aim to provide a

statistical description of the trajectory of an epidemic and its uncertainty, together with a representation of the underlying biological mechanisms. Evaluation of success at these goals is necessary in order for a model to provide a reliable tool for guiding evidence-based public policy interventions. In this article, we conduct a re-analysis of the 2010-2019 cholera outbreak in Haiti. We use this case study to investigate current procedures for fitting mechanistic models to time series data, while identifying limitations of these methodologies and proposing remedies. Our analysis presents methodology for diagnosing how well a model describes observed data. Using objective measures to assess model fit ensures that our evaluation is based on quantifiable criteria. Incorporating reproducibility into this assessment results in a framework that enables the validation or refinement of model based inferences when revisiting the data, facilitating scientific discovery. Our data analysis workflow is supported by recent advances in algorithms, software and hardware, which facilitate statistical fitting of nonlinear stochastic dynamic models to observed incidence data. However, inference for high-dimensional systems remains a methodological challenge. One of the models under consideration involves spatially coupled stochastic meta-populations, and we demonstrate how a recently developed algorithm permits likelihood-based inference and model diagnostics in this setting. We contend that raising the currently accepted standards of infectious disease modeling will result in a greater ability of scientists and policy makers to understand and respond to future infectious disease outbreaks.

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 models for dynamic systems, 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 public health policy in the context of the cholera outbreak in Haiti, which started in 2010. Various dynamic models were developed to study this outbreak: searching PubMed

with keywords "Haiti, cholera, model" we obtained 22 studies that utilized deterministic mechanistic dynamic models [?] and 11 studies that used stochastic models [? ? ? ? ? ? ? ? ? ? ?]. Incidence data on the outbreak are available at various spatial scales, motivating 17 studies in our literature review to consider spatially explicit dynamic models [? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?]. Here we focus on a multi-group modeling exercise by Lee et al. [?] 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 ability to calibrate these details to data. These modeling strategies were selected by Lee et a. [?] to represent the range of approaches used in the research community. We focus on Models 1–3, as the strengths and weaknesses of the agent-based modeling approach [?] are outside the scope of this article. In addition, agent-based models were less widely used, the agent based model in [?] being the only model of this class that was found in our literature review. The data that were used in [?], and that we reanalyze, are displayed in Fig. 3.1.

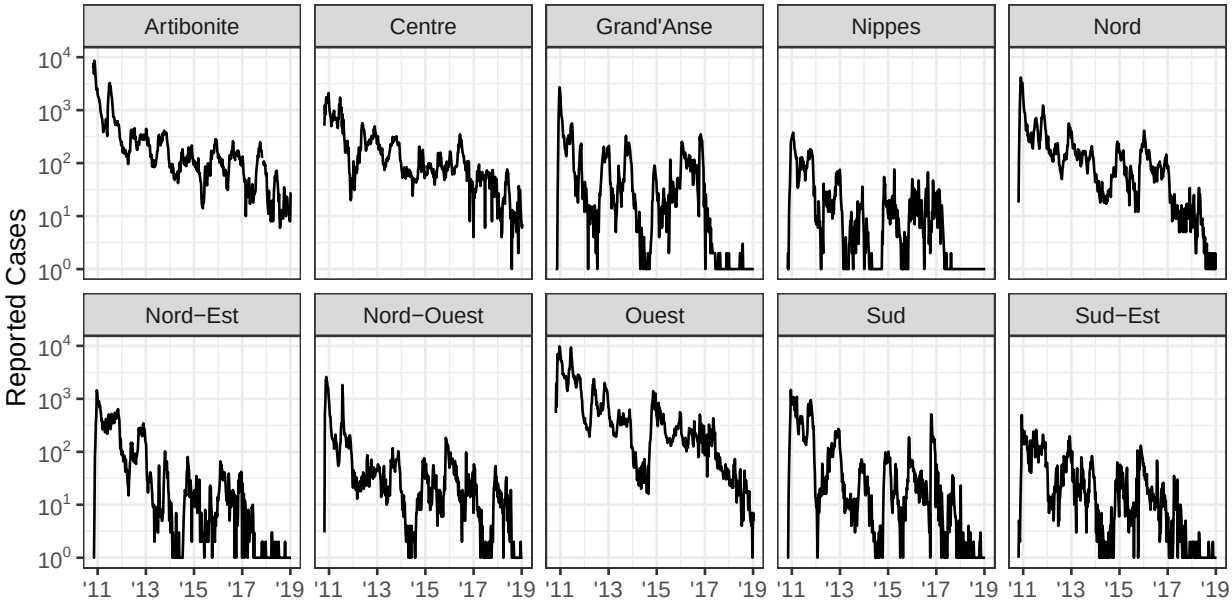

Figure 3.1: **Weekly cholera cases.** Weekly reported cholera cases in Haiti from October 2010 to January 2019 for each of the 10 administrative departments.

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 [**?**]. Advances in OCV technology and vaccine availability, however, raised the possibility of planning a national vaccination program. The possibility of controlling the Haiti cholera outbreak via OCV was considered by various research groups [**?** **?** **?** **?** **?** **?** **?** **?** **?** **?**]. In the Lee et al. [**?**] 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. Despite their autonomy, the four independent teams obtained a consensus that an extensive nationwide vaccination campaign would be necessary to eliminate cholera from Haiti, estimating that a large number of cumulative cholera cases would be observed in the absence of additional vaccination efforts (Figure 3 and 4 of [**?**]). These forecasts are inconsistent with the prolonged period with no confirmed cholera cases between February, 2019 and September, 2022 [**?**]. Though cholera has recently reemerged in Haiti [**?** **?**], the inability to accurately forecast cholera incidence from 2019-2022 prompts us to consider retrospectively what may have been done differently in order to obtain more reliable conclusions, leading to recommendations for future studies.

The discrepancy between the model-based conclusions of Lee et al. [**?**] and the prolonged absence of cholera in Haiti has been debated [**?** **?** **?** **?**]. Suggested origins of this discrepancy include the use of unrealistic models [**?**] and unrealistic criteria for cholera elimination [**?**]. 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. This improved ability may result in more accurate forecasts and facilitates the exploration of model assumptions. Based on this retrospective analysis, we offer suggestions on fitting mechanistic models to dynamic systems for future studies.

Numerous guidelines have been proposed for using mechanistic models to inform policy, reviewed in [**?**]. Behrend et al. [**?**] identify the importance of stakeholder engagement, transparency, reproducibility, uncertainty communication, and testable model outcomes. These and related principles are echoed by other influential articles [**?** **?**]. Additional literature emphasizes model calibration and evaluation techniques [**?** **?** **?**]. These guidelines often lack implementation specifics. As an example, [**?**] largely adhere to the principles of [**?**]—though assessing the extent of stake-holder engagement is challenging—yet their projections are inconsistent with actual cholera incidence data from 2019 to 2022, demonstrating the limitations of current standards. We provide methodology for rigorous statistical calibration and evaluation of dynamic models (as advocated by [**?**]), thereby expanding on

the prevailing guidance. We specifically emphasize principles that prove essential in our case study. Complementary methodological suggestions arising from a spatio-temporal analysis of COVID-19 are detailed in [**?**].

Our recommendations are presented in the context of a case study, with the goal of demonstrating how careful adherence to statistical principles may result in improved model fits. We proceed by introducing the general modeling scheme employed by Models 1–3 and provide details of each individual model; we then describe how each model is calibrated to data, and present a systematic approach to examining and refining these models. Specifically, we focus on how to develop and test variations of the proposed models, as well as diagnosing the models once they have been assimilated to incidence reports. This includes a comprehensive tutorial on performing inference with Model 3 (**??**), a highly non-linear, spatially explicit stochastic model, a challenging task that is possible due to recent methodological advancements. We then use the improved model fits to project cholera incidence in Haiti under various vaccination scenarios considered by Lee et al. [**?**]. Finally, we conclude with a discussion of the results, in which we relate our general recommendations for model based inference of biological systems to the case study of the Haiti cholera outbreak.

Materials and methods

Mechanistic models for disease modeling

Mechanistic models representing biological phenomena are valuable for epidemiology and consequently for public health policy [**?** **?**]. More broadly, they have useful roles throughout biology, especially when combined with statistical methods that properly account for stochasticity and nonlinearity [**?**]. In some situations, modern machine learning methods can outperform mechanistic models on epidemiological forecasting tasks [**?** **?**]. The predictive skill of non-mechanistic models can reveal limitations in mechanistic models, but cannot readily replace the scientific understanding obtained by describing the biological dynamics of the system in a mathematical model [**?** **?**].

In this article, we refer to models that focus on learning relationships between variables in a dataset as *associative*, whereas models that incorporate a known scientific property of the system we call *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. [**?**] 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 models, which provides a statistical fit to the available data while also supporting a causal interpretation. Despite their limited ability to project the effect of interventions on a system, associative models can be effectively used to make inference on an certain features of a system. In our literature review, there were 14 studies that used associative models to describe various aspects of the cholera epidemic [**? ? ? ? ? ? ? ? ? ? ? ? ? ? ?**].

The four mechanistic models of Lee et al. [**?**] were deliberately developed with limited coordination. This 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 created a common notational framework that facilitates model comparison, and put all comparable model parameters—including parameters that were estimated or held constant—into Table 3.1. Translations back to the original notation of Lee et al. [**?**] are given in the supplement (**??**).

Each model describes the cholera dynamics as a partially observed Markov process (POMP) with a latent state vector $\boldsymbol{X}(t)$ for each continuous time point t. N observations on the system are collected at time points $t_1, \ldots, t_N$, written as $t_{1:N}$. The observation at time t_n is modeled by the random vector $\boldsymbol{Y}_n$. While the latent process exists between observation times, the value of the latent state at observations times is of particular interest. We therefore write $\boldsymbol{X}_n = \boldsymbol{X}(t_n)$ to denote the value of the latent process at the nth observation time, and $\boldsymbol{X}_{1:N}$ is the collection of latent state values for all observed time points. The observable random variables $\boldsymbol{Y}_{1:N}$ are assumed to be conditionally independent given $\boldsymbol{X}_{0:N}$. Together, with the density for the initial value of the latent state $\boldsymbol{X}_0 = \boldsymbol{X}(t_0)$, each model defines a joint density $f_{\boldsymbol{X}_{0:N},\boldsymbol{Y}_{1:N}}\Big(\boldsymbol{x}_{0:N}, \boldsymbol{y}_{1:N}; \theta\Big)$, where θ is a parameter vector that indexes the model. The observed data $\boldsymbol{y}^*_{1:N}$, along with the unobserved true value of the latent state, are modeled as a realization of this joint distribution.

Because of the probabilistic nature of both the unobserved latent state and the observable random variables, it is possible to consider various marginal and conditional densities of these two jointly random vectors. An important example is the marginal density of the observed random vector $\boldsymbol{Y}_{1:N}$, evaluated at the observed data $\boldsymbol{y}^*_{1:N}$, as shown in Eq. (3.1):

$$f_{\boldsymbol{Y}_{1:N}}\Big(\boldsymbol{y}^*_{1:N}; \theta\Big) = \int f_{\boldsymbol{X}_{0:N},\boldsymbol{Y}_{1:N}}\Big(\boldsymbol{x}_{0:N}, \boldsymbol{y}^*_{1:N}; \theta\Big) d\boldsymbol{x}_{0:N}. \tag{3.1}$$

Table 3.1: **Model parameters.**

nism	Model 1	Model 2	Model 3
on (day)	$\mu_{IR}^{-1} = 2.0^{\dagger}$ (3.8)	$\mu_{IR}^{-1} = 7.0^{\dagger}$ (3.16)	$\mu_{IR}^{-1} = 5.0^{\dagger}$ (3.28)
y (day)	$\mu_{EI}^{-1} = 1.4^{\dagger}$ (3.7)	$\mu_{EI}^{-1} = 1.3^{\dagger}$ (3.15)	
ality	$\beta_{1:6} = (1.4, 1.2, 1.1, 1.1, 1.4, 1.0)$ (3.4) $\zeta = -0.04^{*}$ (3.34)	$a = 0.4^{\dagger}$ (3.13) $\phi = 0.97^{*}$ (3.13)	$a = 1.00$ (3.32) $r = 0.78$ (3.32)
ity (yr)	$\mu_{RS}^{-1} = 8.0^{\dagger}$ (3.9)	$\mu_{RS}^{-1} = 1.4 \times 10^{11}$ (3.17) $\omega_1^{-1} = 1.0^{\dagger}$ (3.19) $\omega_2^{-1} = 5.0^{\dagger}$ (3.20)	$\mu_{RS}^{-1} = 8.0^{\dagger}$ (3.30)
death (yr^{-1})	$\mu_S = 10^{-2} \times 2.23^{\dagger}$ (3.11) $\delta = 10^{-3} \times 7.5^{\dagger}$ (3.11)		$\delta = 10^{-2} \times 1.59^{\dagger}$ (3.29) $\delta_C = 1.46^{\dagger}$
frac.	$f_z(t) = c\vartheta^{*}(t - \tau_d)^{\dagger}$ (3.6-3.7)	$f = 0.2^{\dagger}$ (3.15)	$f = 0.25^{\dagger}$ (3.27)
t. ity	$\epsilon = 0.05^{\dagger}$ (3.3)	$\epsilon = 0.001^{\dagger}$ (3.13) $\epsilon_W = 10^{-7\dagger}$ (3.21)	$\epsilon = 1^{\dagger}$ (3.25) $\epsilon_W = 0.008$ (3.32)
to human	$\beta_{1:6}$ as above (3.3)	$\beta = 5.97 \times 10^{-15}$ yr^{-1} (3.13)	$\beta_{1:10} = (0.82, 0.02, 0.38, 0.21, 0.51, 0.51, 0.35, 0.12, 0.26, 0.10) \times 10^{-6}$ yr^{-1} (3.25)
to human		$W_{\text{sat}} = 10^{5\dagger}$ (3.13) $\beta_W = 1.1$ yr^{-1} (3.13)	$\beta_{W_{1:10}} = (4.70, 21.00, 24.97, 27.14, 5.28, 30.70, 10.17, 0.99, 11.89, 12.82)$ yr^{-1} (3.25)
to water survival		$\mu_W = 179$ wk^{-1} (3.21) $\delta_W^{-1} = 3^{\dagger}$ (3.22)	$\mu_W = 9.77 \times 10^{-7}$ $\frac{km^2}{wk}$ (3.32) $\delta_W^{-1} = 0.11$ (3.33)
exponent	$\nu = 0.98$ (3.3)		
s noise	$\sigma_{\text{proc}} = (0.09, 0.12)^{*}$ (3.3)		$\sigma_{\text{proc}} = 0.218$ (3.27)
ing rate	$\rho = 0.679$ (??)	$\rho = 0.20^{\dagger}$ (??)	$\rho = 0.98$ (??)
ation e	$\psi = (279.15, 78.33)$ (??)	$\psi = 1.319$ (??)	$\psi = 88.58$ (??)
Values	$I_{0,0} = 7298$ $E_{0,0} = 350$		$I_{0,0}^{3,4} = (21, 6)^{*}$ (??)
ane eters			$\beta_{W3,9}^{hm} = (36.88, 31.64)^{*}$ (3.25) $h_{3,9}^{hm} = (98.98, 58.43)^{*}$ (3.25)

relevant equation are given in parentheses. Parameters that were fixed and not calibrated using the data
†; all fixed parameters values were chosen to match the fixed parameter values of [**?**]. Parameters that
g our re-analysis and were not considered by Lee et al. are indicated with *. Confidence intervals for model
ven in the supplement (**??**). Translations back into the notation of [**?**] are given in **??**.

When treated as a function of the parameter vector θ, this marginal density is called the *likelihood function*, which is the basis of likelihood based statistical inference.

Using the conditional independence of $\boldsymbol{Y}_{1:N}$ given $\boldsymbol{X}_{0:N}$ and the Markov property of $\boldsymbol{X}_{0:N}$, the joint density can be re-factored into the useful form given in Eq. (3.2):

$$f_{\boldsymbol{X}_{0:N},\boldsymbol{Y}_{1:N}}\Big(\boldsymbol{x}_{0:N},\boldsymbol{y}_{1:N};\theta\Big) = f_{\boldsymbol{X}_0}\Big(\boldsymbol{x}_0;\theta\Big)\prod_{n=1}^{N} f_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1}}\Big(\boldsymbol{x}_n|\boldsymbol{x}_{n-1};\theta\Big) f_{\boldsymbol{Y}_n|\boldsymbol{X}_n}\Big(\boldsymbol{y}_n|\boldsymbol{x}_n\Big). \tag{3.2}$$

This factorization is useful because it demonstrates that POMP models may be completely described using three parts: the *initialization model* for the latent states $f_{\boldsymbol{X}_0}\Big(\boldsymbol{x}_0;\theta\Big)$; the *one-step transition density*, or *the process model* $f_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1}}\Big(\boldsymbol{x}_n|\boldsymbol{x}_{n-1};\theta\Big)$; and the *measurement model* $f_{\boldsymbol{Y}_n|\boldsymbol{X}_n}\Big(\boldsymbol{y}_n|\boldsymbol{x}_n\Big)$. In the following subsections, we describe Models 1–3 in terms of these three components. The latent state vector $\boldsymbol{X}(t)$ for each model consists of individuals labeled as susceptible (S), infected (I), asymptomatically infected (A), vaccinated (V), and recovered (R), with various sub-divisions sometimes considered. The observable random vector $\boldsymbol{Y}_{1:N}$ represents the random vector of cholera incidence data for each model; Models 2 and 3 have metapopulation structure, meaning that each individual is a member of a spatial unit, denoted by a subscript $u \in 1:U$, in which case we denote the observed data for each unit using $\boldsymbol{y}^*_{1:N} = y^*_{1:N,1:U}$. Here, the spatial units are the $U = 10$ Haitian administrative départements (henceforth anglicized as departments).

While the complete 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. Additional details about the numeric implementation of these models are provided in a supplemental text (**??**). While each of the dynamic models considered in this manuscript can be fully described using the mathematical equations provided in the following section, diagrams of dynamic systems can be helpful to understand the equations. For this reason, we provide flow chart diagrams for Models 1–3 in supplement figures (**??**, **??** and **??**).

Model 1

The latent state vector $\boldsymbol{X}(t) = \Big(S_z(t), E_z(t), I_z(t), A_z(t), R_z(t), z \in 0:Z\Big)$ describes susceptible, latent (exposed), infected (and symptomatic), asymptomatic, and recovered individuals in vaccine cohort z at time t. Here, $z = 0$ corresponds to unvaccinated individuals, and $z \in 1:Z$ describes hypothetical vaccination programs. Each program z indexes differences in both the number of doses administered (one versus two doses per individual) and the round

of vaccine administration, separating individuals into compartments with distinct dynamics based on vaccination status. The force of infection is

$$\lambda(t) = \Big(\sum_{z=0}^{Z} I_z(t) + \epsilon \sum_{z=0}^{Z} A_z(t) \Big)^{\nu} \frac{d\Gamma(t)}{dt} \beta(t)/N, \tag{3.3}$$

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

$$\beta(t) = \bar{\beta} \exp\Big(\sum_{j=1}^{6} \beta_j s_j(t) \Big), \tag{3.4}$$

where $\bar{\beta} = 1$ (wk)$^{-1}$ is a dimensionality constant. The process noise $d\Gamma(t)/dt$ is multiplicative Gamma-distributed white noise, with infinitesimal variance parameter σ^2_{proc}. Lee et al. [**?**] 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 [**?**] which has been found to improve the statistical fit of disease transmission models [**? ?**].

For any time point in $t_{1:N}$, the process model $f_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1}}\big(\boldsymbol{x}_n|\boldsymbol{x}_{n-1};\theta\big)$ is defined by describing how individuals move from one latent state compartment to another. Per-capita transition rates are given in Eqs. (3.5)-(3.12):

$$\begin{aligned}
\mu_{S_z E_z} &= \lambda(t), && (3.5)\\
\mu_{E_z I_z} &= \mu_{EI}\big(1 - f_z(t)\big), && (3.6)\\
\mu_{E_z A_z} &= \mu_{EI}\, f_z(t), && (3.7)\\
\mu_{I_z R_z} &= \mu_{A_z R_z} = \mu_{IR}, && (3.8)\\
\mu_{R_z S_z} &= \mu_{RS}, && (3.9)\\
\mu_{S_0 S_z} &= \mu_{E_0 E_z} = \mu_{I_0 I_z} = \mu_{A_0 A_z} = \mu_{R_0 R_z} = \eta_z(t), && (3.10)\\
\mu_{S_z \bullet} &= \mu_{E_z \bullet} = \mu_{I_z \bullet} = \mu_{A_z \bullet} = \mu_{R_z \bullet} = \delta, && (3.11)\\
\mu_{\bullet S_0} &= \mu_S, && (3.12)
\end{aligned}$$

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 $\bullet$ modeling entry into the study population due to birth or immigration, and exit from the study population due to death or immigration. Thus, $\mu_{A\bullet}$ is a rate of exiting the study population from compartment A and $\mu_{\bullet B}$ 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. (3.3). In Eqs. (3.7) and (3.6) 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 $f_z(t) = c\,\vartheta^*(t - \tau_z)$ where $\vartheta^*(t)$ is efficacy at time t since vaccination for adults, a step-function represented in Table S4 of [**?**], 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 modeled as equal between single and double dose until 52 weeks after vaccination, at which point the single dose becomes ineffective.

The latent state vector $\boldsymbol{X}(t)$ is initialized by setting the counts for each compartment and vaccination scenario $z \neq 0$ as zero, and introducing initial-value parameters $I_{0,0}$ and $E_{0,0}$ such that $R_0(0) = 0$, $I_0(0) = \text{Pop} \times I_{0,0}$, $E_0(0) = \text{Pop} \times E_{0,0}$ and $S_0(0) = \text{Pop} \times (1 - I_{0,0} - E_{0,0})$, where Pop is the total population of Haiti. The measurement model describes reported cholera cases at time point n come from a negative binomial distribution, where only a fraction (ρ) of new weekly cases are reported. More details about the initialization model $f_{\boldsymbol{X}_0}\Big(\boldsymbol{x}_0;\theta\Big)$ and the measurement model $f_{\boldsymbol{Y}_n|\boldsymbol{X}_n}\Big(\boldsymbol{y}_n|\boldsymbol{x}_n\Big)$ for Models 1–3 are provided a supplement text (**??** and **??**).

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.

Like Model 1, the process model $f_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1}}\Big(\boldsymbol{x}_n|\boldsymbol{x}_{n-1};\theta\Big)$ is primarily defined via the description of movement of individuals between compartments, however Model 2 also includes a dynamic description of a latent bacterial compartment as well. 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

$$\lambda_u(t) = 0.5\Big(1 + a\cos(2\pi t + \phi)\Big)\frac{\beta_W\, W_u(t)}{W_{\rm sat} + W_u(t)} + \beta\left\{\sum_{z=0}^{4} I_{uz}(t) + \epsilon\sum_{z=0}^{4} A_{uz}(t)\right\}. \tag{3.13}$$

The latent state is therefore described by the vector $\boldsymbol{X}(t) = \Big(S_{uz}(t), E_{uz}(t), I_{uz}(t), A_{uz}(t), R_{uz}(t), R^A_{uz}(t), W_u, u \in 1\!:\!U, z \in 0\!:\!4\Big)$. The cosine term in Eq. (3.13) accounts for annual seasonality, with a phase parameter ϕ. The Lee et al. [**?**] implementation of Model 2 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

$$\begin{aligned}
\mu_{S_{uz}E_{uz}} &= (1-\vartheta_z)\,\lambda_u(t), && (3.14)\\
\mu_{E_{uz}I_{uz}} &= f\mu_{EI}, \quad \mu_{E_{uz}A_{uz}} = (1-f)\mu_{EI}, && (3.15)\\
\mu_{I_{uz}R_{uz}} &= \mu_{A_{uz}R^A_{uz}} = \mu_{IR}, && (3.16)\\
\mu_{R_{uz}S_{uz}} &= \mu_{R^A_{uz}S_{uz}} = \mu_{RS}, && (3.17)\\
\mu_{S_{uz}S_{vz}} &= \mu_{E_{uz}E_{vz}} = \mu_{I_{uz}I_{vz}} = \mu_{A_{uz}A_{vz}} = \mu_{R_{uz}R_{vz}} = \mu_{R^A_{uz}R^A_{vz}} = T_{uv}, && (3.18)\\
\mu_{S_{u1}S_{u0}} &= \mu_{S_{u3}S_{u0}} = \omega_1, && (3.19)\\
\mu_{S_{u2}S_{u0}} &= \mu_{S_{u4}S_{u0}} = \omega_2, && (3.20)\\
\mu_{\bullet W_u} &= \mu_W\left\{\sum_{z=0}^{4} I_{uz}(t) + \epsilon_W\sum_{z=0}^{4} A_{uz}(t)\right\}, && (3.21)\\
\mu_{W_u\bullet} &= \delta_W, && (3.22)\\
\mu_{W_uW_v} &= w_r T^W_{uv}. && (3.23)
\end{aligned}$$

In Eq. (3.18) the spatial coupling is specified by a gravity model,

$$T_{uv} = v_{\rm rate} \times \frac{{\rm Pop}_u{\rm Pop}_v}{D^2_{uv}}, \tag{3.24}$$

where ${\rm Pop}_u$ is the mean population for department u, D_{uv} is a distance measure estimating average road distance between randomly chosen members of each population, and $v_{\rm rate} = 10^{-12}\,{\rm km^2 yr^{-1}}$ was fixed at the value used in [**?**]. In Eq. (3.23), 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 $W_{\rm sat}$ in Eq. (3.13). In Eq. (3.14), ϑ_z denotes the vaccine efficacy for each vaccination campaign $z \in Z$, with $\vartheta_0 = 0$, $\vartheta_1 = 0.429q$, $\vartheta_2 = 0.519q$, $\vartheta_3 = 0.429$, and $\vartheta_4 = 0.519$ Here, $q = 0.4688$ represents the reduced efficacy of the vaccination for children under the age of five years, and the values 0.429 and 0.519 are the median effectiveness of one

and two doses over their effective period respectively, according to Table S4 in the supplement material of Lee et al. [**?**]. Because vaccine efficacy remains constant, individuals in this model transition from a vaccinated compartment to the susceptible compartment at the end of the vaccine coverage period.

The starting value for each element of the latent state vector $\boldsymbol{X}(0)$ are set to zero except for $I_{u0}(0) = y^*_u(0)/\rho$ and $R_{u0}(0) = \text{Pop}_u - I_{u0}(0)$, where $y^*_u(0)$ is the reported number of cholera cases in department u at time $t = 0$. Reported cases are described using a log-normal distribution, with the log-scale mean equal to the reporting rate ρ times the number of newly infected individuals. See the supplement material on model initializations for more details (**??**).

Model 3

The latent state is described as $\boldsymbol{X}(t) = \Big(S_{uz}(t), I_u(t), A_u(t), R_{uzk}(t), W_u(t), u \in 0\!:\!U, z \in 0\!:\!4, k \in 1\!:\!3\Big)$. 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 processes model $f_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1}}\Big(\boldsymbol{x}_n|\boldsymbol{x}_{n-1};\theta\Big)$ describes the movement of individuals between latent compartments, as well as the birth and death process of local, unobserved bacterial compartments $W_u(t)$. The force of infection is

$$\lambda_u(t) = \Big(\beta_{W_u} + 1_{(t\geq t_{hm})}\beta^{hm}_{W_u}e^{-h^{hm}_u(t-t_{hm})}\Big)\,\frac{W_u(t)}{1+W_u(t)} + \beta_u\sum_{v\neq u}\Big(I_v(t) + \epsilon A_v(t)\Big), \tag{3.25}$$

where t_{hm} is the time Hurricane Matthew struck Haiti [**?**], and $1_{(A)}$ is the indicator function for event A. In [**?**], $\beta^{hm}_{W_u}$ and h^{hm}_u were set to zero for all u; the need to account for the effect Hurricane Matthew had on cholera transmission for this model is explored in Sec. S5 of the supplement.

Per-capita transition rates are used for both compartments representing human counts

and the aquatic reservoir of bacteria; these rates are given in Eqs. (3.26)–(3.33).

$$\mu_{S_{uz}I_u} = f\,\lambda_u\Big(1-\vartheta_{uz}(t)\Big)\,d\Gamma/dt, \tag{3.26}$$

$$\mu_{S_{uz}A_u} = (1-f)\,\lambda_u\Big(1-\vartheta_{uz}(t)\Big)\,d\Gamma/dt, \tag{3.27}$$

$$\mu_{I_uR_{uz1}} = \mu_{A_uR_{uz1}} = \mu_{IR}, \tag{3.28}$$

$$\mu_{I_uS_{u0}} = \delta+\delta_C,\ \ \mu_{A_uS_{u0}} = \delta \tag{3.29}$$

$$\mu_{R_{uz1}R_{uz2}} = \mu_{R_{uz2}R_{uz3}} = 3\mu_{RS}, \tag{3.30}$$

$$\mu_{R_{uzk}S_{u0}} = \delta+3\mu_{RS}\,\mathbf{1}_{\{k=3\}}, \tag{3.31}$$

$$\mu_{\bullet W_u} = \Big[1+a\Big(J_u(t)\Big)^r\Big]\mathrm{Den}_u\,\mu_W\Big[I_u(t)+\epsilon_W A_u(t)\Big], \tag{3.32}$$

$$\mu_{W_u\bullet} = \delta_W. \tag{3.33}$$

As with Model 1, $d\Gamma_u(t)/dt$ is multiplicative Gamma-distributed white noise in Eqs. (3.26) and (3.27). In Eq. (3.32), $J_u(t)$ 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 Den_u is the 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 Eq. (3.29) to account for cholera induced deaths among infected individuals. All deaths are balanced by births into the susceptible compartment in Eqs. (3.29) and (3.31), thereby maintaining constant population in each department.

Similar to Model 1, there are no distinct compartments for individuals under five years of age, and the vaccination efficacy is taken as a age adjusted weighted average of the efficacy for individuals both over and under five years of age: $\vartheta_{uz}(t) = c\vartheta^*(t-\tau_{uz})$, where τ_{uz} is the vaccination time for unit u and vaccination campaign z. The value c and the function ϑ^* are equivalent to those described in the Model 1 description.

Latent states are initialized using an approximation of the instantaneous number of infected, asymptomatic, and recovered individuals at time t_0 by using the first week of cholera incidence data. Specifically, we set $I_{u0}(0) = \frac{y^*_{1u}}{\rho(\delta+\delta_C+\mu_{IR})}$, $A_{u0}(0) = \frac{1-f}{f}I_{u0}(0)$, $R_{u0k} = y^*_{1u} - I_{u0}(0) - A_{u0}(0)$, and we initialize $W_u(0)$ by enforcing the rainfall dynamics supposed by the one step transition model; all other compartments that represent population counts are set to zero at time t_0. For each unit u with zero case counts at time t_1, this initialization scheme results in having zero individuals in the Infected and Asymptomatic compartments, as well as having no bacteria in the aquatic reservoir. In reality, it is plausible that some bacteria or infected individuals were present in unit u but went unreported. Therefore, for departments with zero case counts in week 1, we estimate the number of infected individuals

rather then treating this value as a constant (**??**). Finally, reported cholera cases are modeled using a negative binomial distribution with mean equal to a fraction (ρ) of individuals in each unit who develop symptoms and seek healthcare, and with over-dispersion parameter ψ (**??**).

Model Fitting

Each of the three models considered in this study describes cholera dynamics as a partially observed Markov process (POMP) [**?**], with the understanding that the deterministic Model 2 is a special case of a Markov processes solving a stochastic differential equation in the limit as the noise parameter goes to zero. Each model is indexed by a parameter vector, θ, and different values of θ can result in qualitative differences in the predicted behavior of the system. Therefore, the choice of θ used to make inference about the system can greatly affect model-based conclusions [**?**]. Elements of θ can be fixed at a constant value based on scientific understanding of the system, but parameters can also be calibrated to data by maximizing a measure of congruency between the observed data and the model's mechanistic structure. Calibrating model parameters to observed data does not guarantee that the resulting model successfully approximates real-world mechanisms, since the model description of the dynamic system may be incorrect and does not change as the model is calibrated to data. However, the congruency between the model and observed data serves as a proxy for the congruency between the model and the true underlying dynamic system. As such, it is desirable to obtain the best possible fit of the proposed mechanistic structure to the observed data.

In this article we follow [**?**] by calibrating the parameters of each of our models using maximum likelihood, as described in Eq. (3.1). The likelihood for each of the fitted models—and the corresponding AIC values for model comparisons that include an adjustment for the number of calibrated parameters—is provided in Table 3.2. In the following subsections we describe in detail our approach to calibrating the three proposed mechanistic models to observed cholera incidence data. The main alternative to maximum likelihood estimation is Bayesian inference via Markov chain Monte Carlo, used to analyze the Haiti cholera epidemic by [**? ? ? ? ? ? ? ? ? ? ?**].

Calibrating Model 1 Parameters

Model 1 is a highly nonlinear over-dispersed stochastic dynamic model, favoring a scientifically plausible description of cholera dynamics rather than one that is statistically convenient [**?**]. This results in the inability to obtain a closed form expression of the joint model density—described in Eq. (3.2). Therefore in order to perform likelihood based inference on this model, we are restricted to use parameter estimation techniques that have the *plug-and-play*

Table 3.2: **AIC values for Models 1–3 and their benchmarks.**

	Model 1	Model 2	Model 3
Log-likelihood	−2728.1 (−3030.9)[1]	−21957.3 (−29367.4)	−17332.9 (−33832.6)[2]
Number of Fit Parameters	15 (20)	6 (6)	34 (29)
AIC	5486.3 (6101.8)[1]	43926.5 (58746.9)	34733.9 (67723.2)[2]
Benchmark AIC	5585.3	36961.0	35945.2

Values in parentheses are corresponding values obtained using the models of [**?**]. [1]The reported likelihood is an upper bound of the likelihood of the model in [**?**]. [2]In [**?**], Model 3 was fit 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 −9721.2. On this same subset, our model has a likelihood of −7219.5. Details of estimating the likelihood of the models used in [**?**] are provided in the supplement (**??**).

property, which is that the fitting procedure only requires the ability to simulate the latent process rather than evaluating transition densities [**?** **?**]; in the context of the notation and definitions employed in this article, this means that we only require the ability to simulate from $f_{\boldsymbol{X}_0}\big(\boldsymbol{x}_0;\theta\big)$ and $f_{\boldsymbol{X}_n|\boldsymbol{X}_{n-1}}\big(\boldsymbol{x}_n|\boldsymbol{x}_{n-1};\theta\big)$ rather than needing to evaluate these densities. Plug-and-play algorithms include Bayesian approaches like ABC and PMCMC [**?** **?**], but here we use algorithms that enable maximum likelihood estimation. To our knowledge, the only plug-and-play methods that have been effectively used to maximize the likelihood for arbitrary nonlinear POMP models are iterated filtering algorithms [**?**], which modify the well-known *particle filter* [**?**].

The particle filter, also referred to as sequential Monte Carlo, is a simulation based method that is frequently used in Bayesian inference to approximate the posterior distribution of latent states. This algorithm can also be used to accurately approximate the log-likelihood of a POMP model, defined as the integral in Eq. (3.1). Iterated filtering algorithms, such as IF2 [**?**], extend the particle filter 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. With a sufficient number of iterations, the resulting parameter values converge to a region of the parameter space that maximizes the model likelihood.

The ability to maximize the likelihood allows for likelihood-based inference, such as performing statistical tests for potential model improvements. We demonstrate this capability by proposing a log-linear trend ζ in transmission in Eq. (3.4):

$$\beta(t) = \bar{\beta} \exp\Big(\sum_{j=1}^{6} \beta_s s_j(t) + \zeta \bar{t} \Big), \tag{3.34}$$

where $\bar{t} = \frac{t-(t_N+t_0)/2}{t_N-(t_N+t_0)/2}$, so that $\bar{t} \in [-1, 1]$. The proposal of a trend in transmission is a result of observing an apparent decrease in reported cholera infections from 2012-2019 in Fig. 3.1. 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 [**?**], may have substantially reduced cholera transmission over time [**?**].

We perform a statistical test to determine whether or not the data indicate the presence of a trend in transmissibility. To do this, we perform a profile-likelihood search on the parameter ζ and obtain a 95% confidence interval via a Monte Carlo Adjusted Profile (MCAP) [**?**]. Lee et al. [**?**] implemented Model 1 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 95% confidence interval for ζ is $(-0.098, -0.009)$, with the full results displayed in Fig. 3.2. These results are suggestive that the inclusion of a trend in the transmission rate improves the quantitative ability of Model 1 to describe the observed data. The maximum likelihood estimate for ζ corresponds to a 7.3% reduction to the transmission rate over the course of the outbreak, with a 95% confidence interval of (1.8%, 17.9%) for the overall reduction in transmission. The reported results for Model 1 in the remainder of this article were obtained with the inclusion of the parameter ζ. The inclusion of a trend in transmission rate demonstrates a class of model variation that can be highly beneficial to consider: the model variation has a plausible scientific justification, and is easily testable using likelihood based methods.

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 phenomenon or unmeasured covariate.

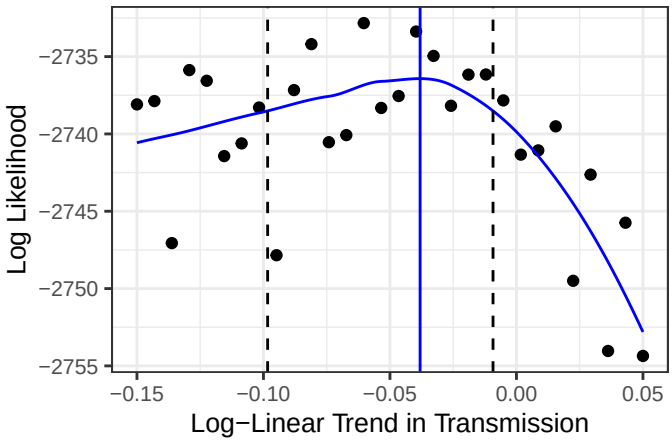

Figure 3.2: **Confidence interval for the log-linear trend in transmission.** Monte Carlo adjusted profile (MCAP) of ζ for Model 1. The blue curve is the MCAP, the vertical blue line indicates the MLE, and the vertical dashed lines indicate the 95% confidence interval.

The statistical evidence of a trend in transmission rate in this model could be explained by any trending variable (such as hygiene improvements, or changes in population behavior), resulting in confounding from collinear covariates. Alternatively, it is possible that the negative trend observed in the incidence data could be attributed to a decreasing reporting rate rather than decreasing transmission rate. This could be formally tested by comparing models with either trend specification. We did not do this because evidence suggests that reporting rate was maintained or increased (Figure 1 of [**?**]). We instead argue that a decreasing transmission rate is a plausible way to explain the decrease in cases over time, as there is alternative evidence that supports this model [**?** **?** **?**]. It is not practical to test all remotely plausible model variations, yet a strongly supported conclusion should avoid ruling out untested hypotheses. The robust statistical conclusion for our analysis is that a model which allows for change fits better than one which does not, and a trend in transmission is a plausible way to do this.

We implemented Model 1 using the `pomp` package [**?**], relying heavily on the source code provided by Lee et al. [**?**]. Both analyses used the `mif2` implementation of the IF2 algorithm to estimate θ by maximum likelihood. One change we made in the statistical analysis that led to larger model likelihoods was increasing the computational effort in the numerical maximization. While IF2 enables parameter estimation for a large class of models, the theoretic ability to maximize the likelihood depends on asymptotics in both the number of particles and the number of filtering iterations. Many Monte Carlo replications are then required to quantify and further reduce the error. The large increase in the log-likelihood for Model 1 (Table 3.2) can primarily be attributed to increasing the computational effort used

to calibrate the model. This result highlights the importance of carefully determining the necessary computational effort needed to maximize model likelihoods and acting accordingly. In this case study, this was done by performing standard diagnostics for the IF2 and particle filter algorithms[**? ? ? ?**]. Given the considerable computational costs of simulation-based algorithms, we find it useful to perform an initial assessment using hyperparameter values—such as the number of particles, filtering iterations, and replicates based on different parameter initializations—that enable relatively quick calculations. The insights obtained from this preliminary analysis help in accurately determining the amount of computation that is required to achieve reliable outcomes. Simulations from the initial conditions of our fitted model are plotted against the observed incidence data in Fig. 3.3.

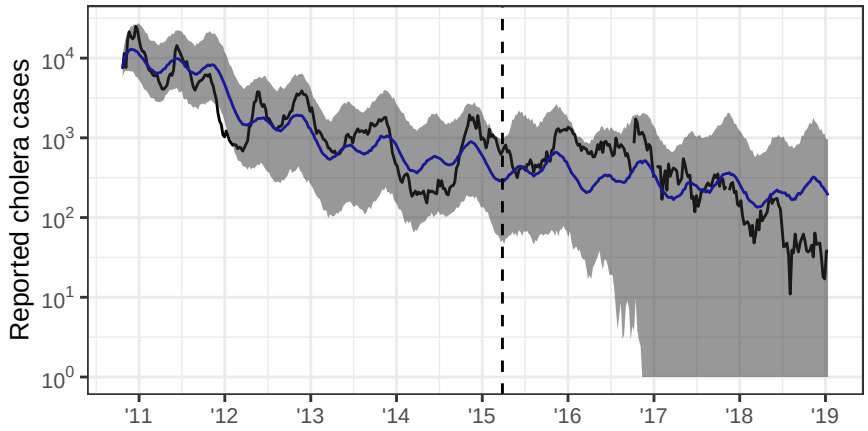

Figure 3.3: **Simulations from Model 1 compared to reported cholera cases.** The black curve is observed data, the blue curve is median of 500 simulations from initial conditions using estimated parameters, and the vertical dashed line represents break-point when parameters are refit.

Calibrating Model 2 Parameters

Model 2 is a deterministic compartmental model defined by a set of coupled differential equations. The use of deterministic compartment models have a long history in the field of infectious disease epidemiology [**? ? ?**], and can be justified by asymptotic considerations in a large-population limit [**? ?**]. Because the process model of Model 2 is deterministic, maximum likelihood estimation reduces to a least squares calculation when combined with a Gaussian measurement model (**??**). Lee et al. [**?**] 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 increased the flexibility of the model and hence the ability to fit the observed data. The

increase in model flexibility, however, resulted in hidden states that were inconsistent between model phases. The inclusion of a model break-point by Lee et al. [**?**] 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. This results in a bias towards accurately describing periods of high infectiousness. This bias is particularly troublesome for modeling cholera dynamics in Haiti: the inability to accurately fit times of low infectiousness may result in poor model forecasts, as few cases of cholera were observed in the last few years of the epidemic.

To combat this issue, we fit the model to log-transformed case counts, since the log scale stabilizes the variation during periods of high and low incidence. An alternative solution is to change the measurement model to include overdispersion, as was done in Models 1 and 3. This permits the consideration of demographic stochasticity, which is dominant for small infected populations, together with log scale stochasticity (also called multiplicative, or environmental, or extra-demographic) which is dominant at high population counts. Here we chose to fit the model to transformed case counts rather than adding overdispersion to the measurement model with the goal of minimizing the changes to the model proposed by Lee et al. [**?**].

We implemented this model using the `spatPomp` `R` package [**?**]. The model was then fit using the subplex algorithm [**?**]. A comparison of the trajectory of the fitted model to the data is given in Fig. 3.4.

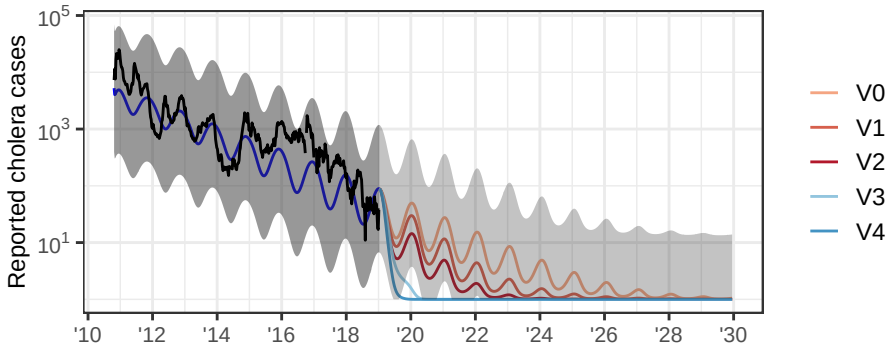

Figure 3.4: **Simulated trajectory of Model 2.** The black line shows the nationally aggregated weekly cholera incidence data. The blue curve from 2012-2019 is the trajectory of the calibrated version of Model 2. Projections under the various vaccination scenarios, which are discussed in detail in the **Forecasts** subsection are also included. The gray ribbons represent a 95% interval obtained from the log-normal measurement model. To avoid over-plotting, measurement variance is only plotted for the V0 vaccination scenario.

Calibrating Model 3 Parameters

Model 3 describes cholera dynamics in Haiti using a metapopulation model, where the hidden states in each administrative department has an effect on the dynamics in other departments. The decision to address metapopulation dynamics using a spatially explicit model, rather than to aggregate over space, is double-edged. Evidence for the former approach has been provided in previous studies [**?**], including the specific case of heterogeneity between Haitian departments in cholera transmission [**?**]. However, a legitimate preference for simplicity can support a decision to consider nationally aggregated models [**? ?**].

In our literature review, 17 articles considered dynamic models that incorporate spatial heterogeneity [**? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?**]. All but four [**? ? ? ?**] of these studies used deterministic dynamic models: this greatly simplifies the process of calibrating model parameters to incidence data, though deterministic models can struggle to describe complex stochastic dynamics. The model in [**?**] was fit using an Ensemble Kalman Filter (EnKF) [**?**]; though EnKF scales favorably with the number of spatial units, it relies on linearization of latent states which can be problematic for highly nonlinear systems [**? ?**]. Alternative approaches used to fit stochastic models included making additional simplifying assumptions to aid in the fitting process [**?**], and using MCMC algorithms [**? ?**] which require specific structures in the latent dynamics, making these algorithms non plug-and-play. In this subsection, we present how the recently developed iterated block particle filter (IBPF) algorithm [**? ?**] can be used to fit a spatially explicit stochastic dynamic model to incidence data.

One issue that arises when fitting spatially explicit models is that parameter estimation techniques based on the particle filter become computationally intractable as the number of spatial units increases. This is a result of the approximation error of particle filters growing exponentially in the dimension of the model [**? ?**]. To avoid the approximation error present in high-dimensional models, Lee et al. [**?**] simplified the problem of estimating the parameters of Model 3 by creating an approximate version of the model where the units are independent given the observed data. 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. (3.25). Because the simplified, spatially-decoupled version of Model 3 implemented in [**?**] relies on the observed cholera cases, the calibrated model cannot readily be used to obtain forecasts. Therefore, in order to obtain model forecasts, Lee et al. [**?**] used the parameters estimates from the spatially-decoupled approximation of Model 3 to obtain forecasts using the fully coupled version of the model. This approach of model calibration and forecasting avoids the issue of particle depletion, but may also be problematic. One concern is that cholera dynamics in department u are highly related to the

dynamics in the remaining departments; calibrating model parameters while conditioning on the observed cases in other departments may therefore lead to an over-dependence on observed cholera cases. Another concern is that the two versions of the model are not the same, resulting in sub-optimal parameter estimates for the spatially coupled model, as parameters that maximize the likelihood of the decoupled model almost certainly do not maximize the likelihood of the fully coupled model. These two concerns may explain the unrealistic forecasts and low likelihood of Model 3 in [**?**] (Table 3.2).

At the time Lee et al. [**?**] conducted their study, there was no known algorithm that could readily be used to maximize the likelihood of an arbitrary meta-population POMP model with coupled spatial dynamics, which justifies the spatial decoupling approximation that was used to calibrate model parameters. For our analysis, we calibrate the parameters of the spatially coupled version of Model 3 using the IBPF algorithm [**?**]. This algorithm extends the work of Ning and Ionides [**?**], who provided theoretic justification for the version of the algorithm that only estimates unit-specific parameters. The IBPF algorithm enables us to directly estimate the parameters of models describing high-dimensional partially-observed nonlinear dynamic systems via likelihood maximization. The ability to directly estimate parameters of Model 3 is responsible for the large increase in model likelihoods reported in Table 3.2. Simulations from the fitted model are displayed in Fig. 3.5.

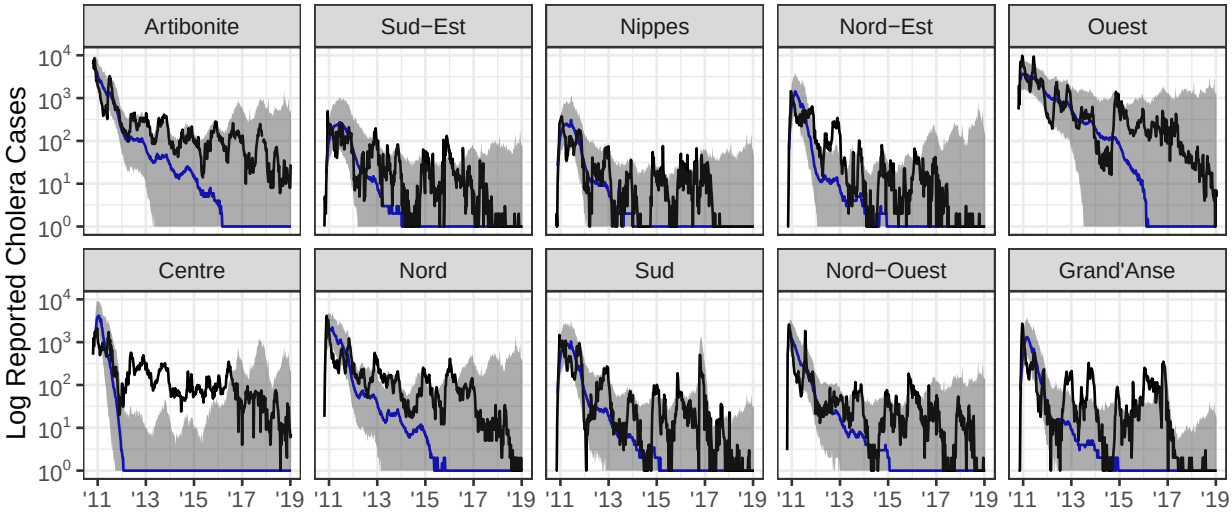

Figure 3.5: **Simulations from Model 3 compared to reported cholera cases.** Simulations from initial conditions using the spatially coupled version of Model 3. The black curve represents true case count, the blue line the median of 500 simulations from the model, and the gray ribbons representing 95% confidence interval.

Model Diagnostics

The goal of parameter calibration—whether done using Bayesian or frequentist methods—is to find the best description of the observed data in the context 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 under investigation. 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 George Box [**?**] 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 [**?**]. In this section, we discuss some tools 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 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. [**?**] 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 considerably smaller than likelihoods that can be achieved (see Table 3.2). 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, which provide insight into how each model performs relative to the others, are quite common [**? ?**]. 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 (ARMA) time series models, or even independent

and identically distributed measurements. Benchmarks enable us to evaluate the goodness of fit that can be expected of a suitable mechanistic model.

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. Therefore, 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) [1].

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 simple negative binomial model with an auto regressive mean as our associative benchmark; this model is described in (3.35).

$$Y_n|Y_{n-1} \sim \mathrm{NB}\Big(\alpha + \beta Y_{n-1}, \varphi\Big), \tag{3.35}$$

where $\mathrm{E}\Big(Y_n|Y_{n-1}\Big) = \alpha+\beta Y_{n-1}$, and $\mathrm{Var}\Big(Y_n|Y_{n-1}\Big) = \mathrm{E}\Big(Y_n|Y_{n-1}\Big)+\mathrm{E}\Big(Y_n|Y_{n-1}\Big)^2/\varphi$. To obtain a benchmark for models with a meta-population structure, we fit independent auto-regressive negative binomial models to each spatial unit. Under the assumption of independence, the log-likelihood of the benchmark on the entire collection of data can be obtained by summing up the log-likelihood for each independent model. In general, a spatially explicit model may not have well-defined individual log-likelihoods, and, in this case, comparisons to benchmarks must be made at the level of the joint model.

In the case where the case counts are large, an alternative benchmark recommended by He et al. [**?**] is a log-linear Gaussian ARMA model; the theory and practice of ARMA models is well developed, and these linear models are appropriate on a log scale due to the exponential growth and decay characteristic of biological dynamics. We use the auto regressive negative binomial model, however, because the large number of weeks with zero recorded cholera cases in department level data makes a benchmark based on a continuous distribution problematic. Log-likelihoods and AIC values of Models 1–3 and of their respective benchmark models are provided in Table 3.2. Models that are fit to the same datasets can be directly compared using AIC values, making it a useful tool to compare to benchmark models. Though Models 2 and 3 are both fit to department level incidence reports, their AIC values are not directly comparable due to the way Model 3 initializes latent states (**??**).

It should be universal practice to present measures of goodness of fit for published models, and mechanistic models should be compared against benchmarks. In our literature review of the Haiti cholera epidemic, no non-mechanistic benchmark models were considered in any of the 32 papers that used dynamic models to describe cholera in order to obtain scientific conclusions. Including benchmarks would help authors and readers to detect and confront any major statistical limitations of the proposed mechanistic models. In addition, the published goodness of fit provides a concrete point of comparison for subsequent scientific investigations. 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* [**?**].

The use of benchmarks may also be beneficial when developing models at differing spatial scales, where a direct comparison between model likelihoods is meaningless. In such a case, a benchmark model can be fit to each spatial resolution being considered, and each model compared to their respective benchmark. Large advantages (or 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 (or against) the statistical fit of models across a range of spatial resolutions.

Comparing model log-likelihoods to a suitable benchmark may not be sufficient to identify all the strengths and weaknesses of a given model. Additional techniques include the inspection of conditional log-likelihoods of each observation given the previous observations in order to understand how well the model describes each data point (**??**). Other tools include plotting the effective sample size of each observation [**?**]; plotting the values of the hidden states from simulations (**??**); and comparing summary statistics of the observed data to simulations from the model [**?** **?**].

Corroborating Fitted Models with Scientific Knowledge

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 be taken as a confirmation of both model-based conclusions and our scientific understanding. On the other hand, comparisons may generate unexpected results that have the potential to spark new scientific knowledge [**?**].

In the context of our case study, we demonstrate how the fit of Model 1 corroborates other evidence concerning the role of rainfall in cholera epidemics. Specifically, we examine the results of fitting the flexible cubic spline term in Model 1 (Eqs. (3.3)–(3.4)). The cubic splines permit flexible estimation of seasonality in the force of infection, $\beta(t)$. Fig. 3.6 shows that the

estimated seasonal transmission rate β mimics the rainfall dynamics in Haiti, despite Model 1 not having access to rainfall data. This is consistent with previous studies that incorporated rainfall as an important part of their mechanistic model or otherwise argue that rainfall is an important driver of cholera dynamics in Haiti [**?** **?** **?** **?** **?** **?** **?** **?** **?**]. The estimated seasonality also features an increased transmission rate during the fall, which was noticed at an earlier stage of the epidemic [**?**]. The high transmission rate in the fall may be a result of the increase transmission that occurred in the fall of 2016, when hurricane Matthew struck Haiti [**?**].

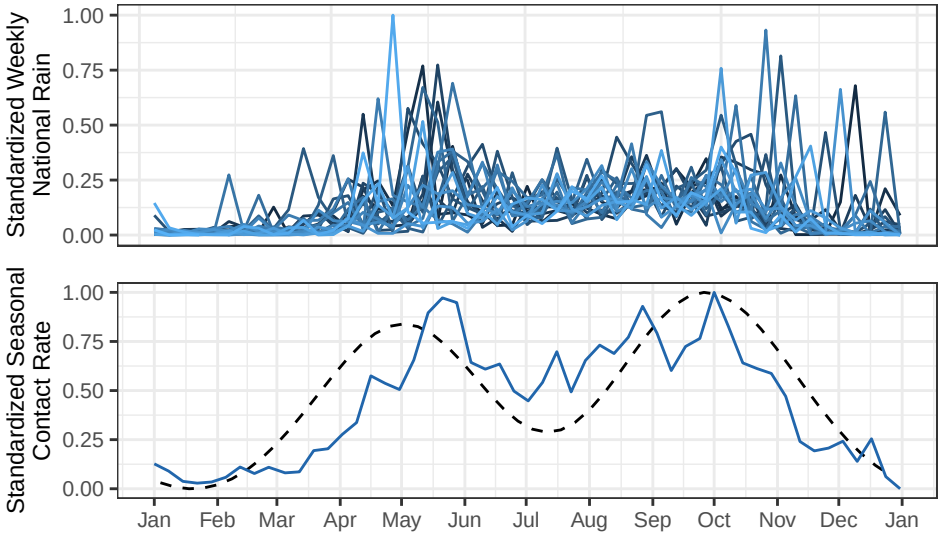

Figure 3.6: **Seasonality of Model 1 transmission compared to rainfall data.** (Top) weekly rainfall in Haiti, lighter colors representing more recent years. (Bottom) estimated seasonality in the transmission rate (dashed line) plotted alongside mean rainfall (solid line). The outsized effect of rainfall in the fall may be due to Hurricane Matthew, which struck Haiti in October of 2016 and resulted in an increase of cholera cases in the nation.

For any model-based inference, it is important to recognize and assess the modeling simplifications and assumptions that were used in order to arrive at the conclusions. In epidemiological studies, for example, quantitative understanding of individual-level processes may not perfectly match model parameters that were fit to population-level case counts, even when the model provides a strong statistical fit [**?**]. This makes direct interpretation of estimated parameters delicate.

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 1.4×10^{11} yr, suggesting that individuals have permanent immunity to cholera once infected. Rather than interpreting this as scientific evidence that

individuals have permanent immunity from cholera, this result suggests that Model 2 favors a regime where reinfection events are a negligible part of the dynamics. The depletion of susceptible individuals may be attributed to confounding mechanisms—such as localized vaccination programs and non-pharmaceutical interventions that reduce cholera transmission [**?** **?**]—that were not accounted for in the model. Perhaps the best interpretation of the estimated parameter, then, is that under the modeling framework that was used, the model most adequately describes the observed data by having a steady decrease in the number of susceptible individuals. The weak statistical fit of Model 2 compared to a log-linear benchmark (see Table 3.2) cautions us against drawing quantitative conclusions from this model. A model that has a poor statistical fit may nevertheless provide a useful conceptual framework for thinking about the system under investigation. However, a claim that the model has been validated against data should be reserved for situations where the model provides a statistical fit that is competitive against alternative explanations.

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 [**?**]. Similarly, parameters that have been calibrated to data should make sense in the context of alternative lines of evidence about the phenomena being modeled, while making allowance for the possibility that the interpretations of parameters may vary when modeling across differing spatial scales.

In the supplement material (**??**), we explore in more detail the process of model fitting and diagnostics for Model 3. Here we demonstrate that the model outperforms its benchmark model on the aggregate scale. However, when focusing on the spatial units with the highest incidence of cholera, Model 3 performs roughly the same as a simple benchmark. By comparing simulations from the fitted model to the filtering distribution, we see that the reconstructed latent states of the model favor higher levels of cholera transmission than what is typically observed in the incidence data. These results hint at the possibility of model mispecification, and warrant a degree of caution in interpreting the model's outputs.

Results

Forecasts

Forecasts are an attempt to provide an accurate estimate of the future state of a system based on currently available data, together with an assessment of uncertainty. Forecasts from mechanistic models that are compatible with current scientific understanding may also provide estimates of the future effects of potential interventions. Further, they may enable real-time testing of new scientific hypotheses [**?**].

Forecasts of a dynamic system should should be consistent with the available data. It is particularly important that forecasts are consistent with the most recent information available, as recent data is likely to be more relevant than older data. While this assertion may seem self-evident, it is not the case for deterministic models, for which the initial conditions together with the parameters are sufficient for forecasting, and so recent data may not be consistent with model trajectories. Epidemiological forecasts based on deterministic models are not uncommon in practice, despite their limitations [**?**]. Lee et al. [**?**] chose to obtain forecasts from all of their models by simulating forward from initial conditions, rather than conditioning forecasts based on the available data. This decision is possibly as a result of using a deterministic model, as forecasts from different models may only be considered comparable if they are obtained in the same way, which is most easily done be by simulating from initial conditions because Model 2 is deterministic.

In contrast, for non-deterministic Models 1 and 3, we obtain forecasts by simulating future values using latent states that are harmonious with the most recent data. This is done by simulating forward from latent states drawn at the last observation time (t_N) from the filtering distribution $f_{\boldsymbol{X}_N|\boldsymbol{Y}_{1:N}}(\boldsymbol{x}_N|\boldsymbol{y}^*_{1:N};\hat{\theta})$. The decision to obtain model forecasts from initial conditions partially explains the unsuccessful forecasts of Lee et al. [**?**]. Table S7 in their supplement material, which contains results that were not discussed in their main article, shows that the subset of their simulations with zero cholera cases from 2019-2020 also correspond with its disappearance until 2022. These results support our argument that forecasts should be made by ensuring the starting point for the forecast is consistent with available data.

Uncertainty in just a single parameter can lead to drastically different forecasts [**?**]. Therefore, parameter uncertainty should also be considered when obtaining model forecasts to influence policy. If a Bayesian technique is used for parameter estimation, a natural way to account for parameter uncertainty is to obtain simulations from the model where each simulation is obtained using parameters drawn from the estimated posterior distribution. For frequentist inference, one possible approach is obtaining model forecasts from various

values of θ, where the values of θ are sampled proportionally according to their corresponding likelihoods [**?**] (**??**). Both of these approaches share the similarity that parameters are chosen for the forecast approximately in proportion to their corresponding value of the likelihood function, $f_{\boldsymbol{Y}_{1:N}}(\boldsymbol{y}^*_{1:N};\theta)$. In this analysis, we do not construct forecasts accounting for parameter uncertainty as our focus is on the estimation and diagnosis of mechanistic models, rather than providing forecasts intended to influence policy. Furthermore, we use the projections from a single point estimate to highlight the deficiency of deterministic models that the only variability in model projections is a result of parameter and measurement uncertainty, which can lead to over-confidence in forecasts [**?**].

The primary forecasting goal of Lee et al. [**?**] was to investigate the potential consequences of vaccination interventions on a system to inform policy. One outcome of their study include estimates for the probability of cholera elimination under several possible vaccination scenarios. Mimicking their approach, we define cholera elimination as an absence of cholera infections for at least 52 consecutive weeks, and we 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. Because Model 1 only accounts for national level disease dynamics, the pre-determined department-specific vaccination campaigns are carried out by assuming the vaccines are administered in one week to the same number of individuals that would have obtained vaccines if explicitly administered to the specific departments. We refer readers to [**?**] and the accompanying supplement material for more details. 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. 3.7.

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

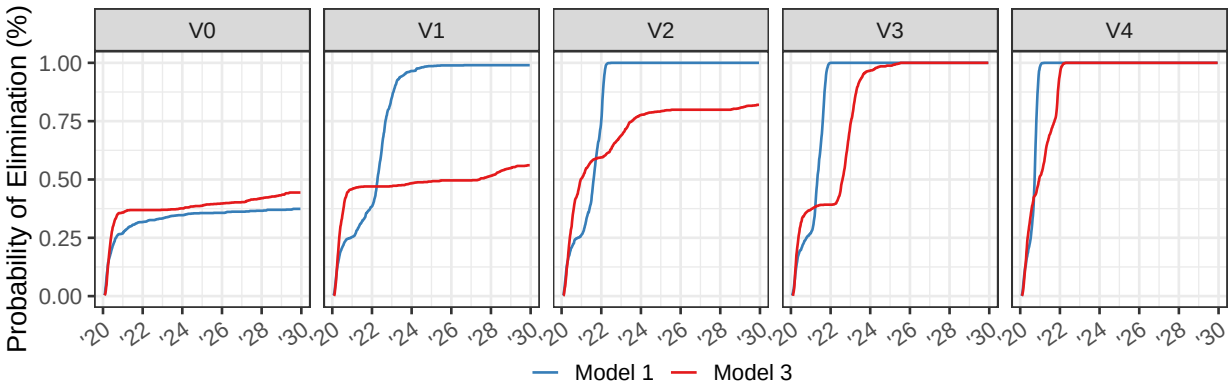

Figure 3.7: **Simulated probability of elimination using Models 1 and 3.** Probability of cholera elimination, defined as having zero cholera infectious for at least 52 consecutive weeks, based on 10 year simulations from calibrated versions of Models 1 and 3. Compare to Fig. 3A of [**?**].

rather than individual potential outcomes. We therefore do not provide probability of elimination estimates under Model 2, but show trajectories under the various vaccination scenarios using this model (Fig. 3.4).

Discussion

The ongoing global COVID-19 pandemic has demonstrated how government policy may be affected by the inferences drawn from mathematical modeling [**?**]. However, the development of credible models—which are supported by data and can provide quantitative insights into a dynamic system—remains a challenging task. In this article, we demonstrated opportunities available for raising the current standards of statistical inference for mathematical models of biological systems.

We presented methodology consistent with existing guidelines [**?**] but going beyond standard practice. In particular, we showed the value of comparing the likelihood of fitted mechanistic models versus non-mechanistic benchmarks, a practice that has been previously advocated for [**?**] but was not done by any of the studies in our literature review. These comparisons, along with other likelihood based diagnostics, help identify specific limitations of proposed models. Diagnostic tools include likelihood profile methods, which help to assess parameter identifiability and enable the construction of confidence intervals for parameter estimates [**?** **?**]. When reaching conclusions, it is important to consider potential consequences of confounded variables and model misspecification.

Model diagnostics are a key tool for exposing unresolved model limitations and improving

model fit. In our case study, we compared the three models from Lee et al. [**?**] to statistical benchmarks, revealing areas for improvement. For example, comparisons of Model 3 to a benchmark revealed its inadequacy in accounting for the post-hurricane increase in transmission, leading to a beneficial model refinement. When a mechanistic model is competitive with statistical benchmarks, we have a license to begin critical evaluation of its causal implications. If a model falls far behind simple benchmarks, there is likely to be substantial limitation in the data analysis that should be identified and remedied. In our case study, the re-calibrated version of Model 1 outperformed its benchmark, so we proceeded to examine causal implications. When doing so, we found that the fitted model provides a causal description of the dynamic system that is consistent with known features of the system, such as the importance of rainfall as a driver of cholera infection. The congruency between causal implications of the model and our belief about the dynamic system, coupled with a strong quantitative description of observed data relative to a benchmark, provides support for viewing the model as a plausible quantitative representation of the system under investigation.

When fitting a mechanistic model to a dynamic system, the complexity of the model warrants consideration. Mathematical models provide simplified representations of complex systems, with the simplicity serving both to facilitate scientific understanding and to enable statistical inference on unknown parameters. In our case study, employing deterministic dynamics in Model 2 was found to be an over-simplification by comparing model fit with benchmarks. Model 3 is distinct in that it is both stochastic and has a meta-population structure, making it challenging to draw likelihood-based inferences. In this paper, we demonstrated how this model class can be calibrated to incidence data using the innovative IBPF algorithm. One of only a few examples of fitting a nonlinear non-Gaussian meta-population model via maximum likelihood [**? ?**], this case study exemplifies the algorithm's potential benefits and provides an example for future researchers on a possible approach to fitting a high-dimensional non-linear model.

Likelihood-based methods aid in determining an appropriate level of model complexity. Models fit to the same data can be compared using a criteria such as AIC. Nested model variations are particularly useful as they enable formal statistical testing of the nested features via likelihood ratio tests. Our case study demonstrated the examination of nested model features for all three models. Model 1 investigated a time-varying transmission rate; Model 2 assessed a phase-shift parameter in seasonal cholera peaks; Model 3 incorporated hurricane-related parameters.

Unmodeled features of a dynamic system can lead to spurious or misleading parameter estimates if the features substantially impact observed data. In deterministic models, features

that cannot be explained by measurement error must be accounted for by the choice of parameters. For our case study, some of the parameter estimates for the deterministic Model 2 are implausible, such as the infinite immunity discussed above, and this may be explained by compensation for model misspecification. Incorporating demographic and environmental stochasticity into models can mitigate the impact of unmodeled features. Stochastic phenomena are not only arguably present in biological systems, but their inclusion in a model also allows observed data variations to be attributed to inherent uncertainty rather than to distorted parameter values. Models 1 and 3 suggest the presence of extra-demographic stochasticity [**?** **?** **?**], as evidenced by the confidence intervals for the corresponding parameter σ_{proc} (**??**).

If forecasts are an important component of a modeling task, the forecasts should be consistent with the available data, particularly at the most recently available time points. In our case study, we did this by simulating forward from the filtering distribution, as this procedure conditions latent variables on the available data. This type of forecasting, however, is not directly available using a deterministic model, where future dynamics are fully determined by initial conditions and parameter values. This can result in over-confident model forecasts [**?**]. Despite their limitations, deterministic models can offer valuable insights into dynamic systems [**?**]. In [**?**], the forecasts from the deterministic Model 2 were qualitatively more consistent with the observed disappearance of cholera than the stochastic models. In our case study, we found improvements to Models 1 and 3 that resulted in improved forecasts for these models.

In our case study, we found that additional attention to statistical details could have resulted in an enhanced statistical fit to the observed incidence data. This would have improved the accuracy of the policy guidance resulting from the study. We used the same data, models, and much of the same code used by Lee et al. [**?**], but we arrived at drastically different conclusions. Specifically, each of the re-calibrated models predicted with moderate probability that cholera would disappear from Haiti. Although there have been new cases of cholera in Haiti, this conclusion aligns more with the prolonged absence of cholera cases from 2019-2022. We acknowledge the benefit of hindsight: our demonstration of a statistically principled route to obtain better-fitting models resulting in more robust insights does not rule out the possibility of discovering other models that fit well yet predict poorly.

Mechanistic models offer 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. Researchers should ensure that intensive numerical calculations are adequately executed. Using benchmarks and alternative model specifications

to assess statistical goodness-of-fit should also should be common practice. Once a model has been adequately calibrated to data, 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 [**?** **?** **?**] should motivate us to follow best practices in data analysis, rather than avoiding the challenge.

Reproducibility and Extendability

Lee et al. [**?**] published their code and data online, and this reproducibility facilitated our work. 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 not be excessively burdensome. Scientific results are only trustworthy so far as they can be critically questioned, and an extendable analysis should facilitate such examination [**?**].

We provide a strong form of reproducibility, as well as extendability, by developing our analysis in the context of a software package, `haitipkg`, written in the R language [**?**]. Using a software package mechanism supports documentation, standardization and portability that promote extendability. In the terminology of Gentleman and Temple Lang [**?**], 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 [**?**] 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).

CHAPTER 4

Conclusion

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

APPENDIX A

Example Appendix 01

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

A.1 Sample appendix section

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

A.1.1 Sample appendix subsection

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in

sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

A.2 Another sample appendix section

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

APPENDIX B

Example Appendix 02

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

BIBLIOGRAPHY

[1] H. Akaike. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6):716–723, 1974.

[2] George Box and Gwilym Jenkins. Time Series Analysis: Forecasting and Control. San Francisco, Holdan-Day, 1970.

[3] Peter J. Brockwell and Richard A. Davis. Time Series: Theory and Methods. Springer Series in Statistics. Springer, New York, NY, 1991.

[4] Yacine Chakhchoukh. A new robust estimation method for arma models. IEEE Transactions on Signal Processing, 58(7):3512–3522, 2010.

[5] Siddhartha Chib and Edward Greenberg. Bayes inference in regression models with arma (p, q) errors. Journal of Econometrics, 64(1):183–206, 1994.

[6] K.H. Chon and R.J. Cohen. Linear and nonlinear arma model parameter estimation using an artificial neural network. IEEE Transactions on Biomedical Engineering, 44(3):168–174, 1997.

[7] James Durbin and Siem Jan Koopman. Time Series Analysis by State Space Methods, volume 38. OUP Oxford, 2012.

[8] Roger Fletcher. Practical Methods of Optimization. John Wiley & Sons, 2000.

[9] G. Gardner, A. C. Harvey, and G. D. A. Phillips. Algorithm AS 154: An algorithm for exact maximum likelihood estimation of autoregressive-moving average models by means of Kalman filtering. Journal of the Royal Statistical Society. Series C (Applied Statistics), 29(3):311–322, 1980.

[10] William E. Hart. Sequential stopping rules for random optimization methods with applications to multistart local search. SIAM Journal on Optimization, 9(1):270–290, 1998.

[11] Rob J Hyndman and Yeasmin Khandakar. Automatic time series forecasting: The forecast package for R. Journal of Statistical Software, 26(3):1–22, 2008.

[12] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. Journal of Basic Engineering, 82(1):35–45, 03 1960.

[13] L. Le Cam. Maximum likelihood: An introduction. International Statistical Review / Revue Internationale de Statistique, 58(2):153–171, 1990.

[14] K.-S. Lii. Identification and estimation of non-gaussian arma processes. IEEE Transactions on Acoustics, Speech, and Signal Processing, 38(7):1266–1276, 1990.

[15] John F. Monahan. Fully bayesian analysis of arma time series models. Journal of Econometrics, 21(3):307–331, 1983.

[16] NOAA. Monthly average master gauge water levels (1860-present): Lake michigan-huron, 2016. Accessed: Jan 24, 2016.

[17] Brian D Ripley. Time series in R 1.5.0. The Newsletter of the R Project Volume, 2:2, June 2002.

[18] Robert H. Shumway and David S. Stoffer. Time Series Analysis and Its Applications: With R Examples. Springer Texts in Statistics. Springer International Publishing, Cham, 2017.

[19] Christian H Weiß and Fukang Zhu. Mean-preserving rounding integer-valued arma models. Journal of Time Series Analysis, 46:530–551, 2024.

[20] Jesse Wheeler, Noel McAllister, and Sylvertooth. arima2. https://cran.r-project.org/web/packages/arima2/index.html, 2023.

[21] Peter Whittle. Hypothesis Testing in Time Series Analysis, volume 4. Almqvist & Wiksells boktr., 1951.