5-2018

Modeling and Optimization of Dynamical Systems in Epidemiology using Sparse Grid Interpolation

Aditya P. Sai
Purdue University

MODELING AND OPTIMIZATION OF DYNAMICAL SYSTEMS IN EPIDEMIOLOGY USING SPARSE GRID INTERPOLATION

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Aditya P. Sai

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2018

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Nan Kong, Chair

Weldon School of Biomedical Engineering

Dr. Gregery T. Buzzard

Department of Mathematics

Dr. Taeyoon Kim

Weldon School of Biomedical Engineering

Dr. David M. Umulis

Weldon School of Biomedical Engineering

Approved by:

Dr. George R. Wodicka

Head of the Graduate Program

It won't be soon before long.

ACKNOWLEDGMENTS

This work is the culmination of years of effort, across multiple, overlapping fields of research. I would first like to thank my advisor Dr. Nan Kong. His passion, patience, and keen sense for engaging academic research challenged me from our very first meeting to conceptualize and execute meaningful, impactful research. My productivity in the latter half of my graduate career can be attributed entirely to the creative liberties intrinsic to his research philosophy. My late advisor, Dr. Ann Rundell, provided me with a solid foundation as a graduate student that enabled future success. There are no words to describe the enormity of the opportunity she gave me. I also owe a debt of gratitude to Drs. Buzzard, Kim, and Umulis, who I have had the great honor of interacting with at various times in my graduate tenure.

I have also had the tremendous fortune of collaborating with peers and colleagues who have been more knowledgeable than me. Their experience and respective skillsets were an invaluable resource to my development. Nevertheless, I would be remiss if I did not mention them here: Nimisha Bajaj, Ankush Chakrabarty, MD Shariar Karim, Wei-an Lin, Zhouyang Lou, Thembi Mdluli, Jeff Perley, Joyatee Sarker, Carolina Vivas-Valencia. Much of the undying gratitude and respect I developed for my peers stems from the indispensable minutiae they imparted to me about everyday life in an immersive, highly rewarding environment.

Finally, to my family. I would like to thank my father, for having convinced me to start this journey, and my mother, for encouraging me to complete it. My brother, who despite being my junior, has set an enviable standard for success that I hope to surpass in the future. The countless members of my extended family, who relentlessly cheered me on through every peak and valley, deserve many thanks. I will be forever changed for having undertaken this intellectual endeavor, in no small part thanks to all of these wonderful people.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABBREVIATIONS

abbr	abbreviation
DBSCAN	Density-based Spatial Clustering of Applications with Noise
FIM	Fisher Information Matrix
GA	Genetic Algorithm
GMM	Gaussian Mixture Model
ISS	Infection-based Sensor Selection
LHS	Latin Hypercube Sampling
MPC	Model Predictive Control
NSS	No Sensor Selection
ODE	Ordinary Differential Equation
PSO	Particle Swarm Optimization
RSS	Random Sensor Selection
SDE	Stochastic Differential Equation
SIR	Susceptible-Infectious-Removed
SIVR	Susceptible-Infectious-Vaccinated-Removed
SRN	Stochastic Reaction Network
TSS	Targeted Sensor Selection

ABSTRACT

Sai, Aditya P. Ph.D., Purdue University, May 2018. MODELING AND OPTIMIZATION OF DYNAMICAL SYSTEMS IN EPIDEMIOLOGY USING SPARSE GRID INTERPOLATION. Major Professor: Nan Kong.

Infectious diseases pose a perpetual threat across the globe, devastating communities, and straining public health resources to their limit. The ease and speed of modern communications and transportation networks means policy makers are often playing catch-up to nascent epidemics, formulating critical, yet hasty, responses with insufficient, possibly inaccurate, information. In light of these difficulties, it is crucial to first understand the causes of a disease, then to predict its course, and finally to develop ways of controlling it. Mathematical modeling provides a methodical, *in silico* solution to all of these challenges, as we explore in this work. We accomplish these tasks with the aid of a surrogate modeling technique known as sparse grid interpolation, which approximates dynamical systems using a compact polynomial representation.

Our contributions to the disease modeling community are encapsulated in the following endeavors. We first explore transmission and recovery mechanisms for disease eradication, identifying a relationship between the reproductive potential of a disease and the maximum allowable disease burden. We then conduct a comparative computational study to improve simulation fits to existing case data by exploiting the approximation properties of sparse grid interpolants both on the global and local levels. Finally, we solve a joint optimization problem of periodically selecting field sensors and deploying public health interventions to progressively enhance the understanding of a metapopulation-based infectious disease system using a robust model predictive control scheme.

1. INTRODUCTION

1.1 Objectives

Throughout history, mathematical modeling has empowered the public health domain to effectively confront and eliminate threats, ranging from smallpox to malaria [1]. Models based on mathematically formulated principles are necessary to elucidate the observed epidemiological phenomena arising from the complexity of disease interactions on numerous spatiotemporal scales. They are employed to address the following:

1. predict the future course of an epidemic through analysis of its transmission mechanisms,
2. align these model forecasts to available data to restrict the number of viable model hypotheses, thereby improving the current state of knowledge, and finally,
3. determine the optimal control strategy to halt and eventually stop the spread of disease, while operating within existing constraints.

In furtherance of these objectives, we present a surrogate modeling framework to rapidly identify and assess the model structures and epidemiological processes responsible for shaping the profile of an infectious disease. The framework makes use of sparse grid interpolation, a polynomial interpolation technique that produces a parsimonious, high-fidelity approximation model that can be examined repeatedly without reference to the original model, avoiding prohibitively expensive simulations.

1.2 Background

Infectious diseases are a continual threat to societies worldwide. They can wreak havoc on unsuspecting populations, strain health care infrastructures, and restrict the movement of peoples, goods and services. The number and variety of outbreaks traced to these infectious diseases have been steadily increasing for decades [2]. Infectious diseases, such as lower respiratory infections, diarrhoeal diseases, HIV/AIDS, and tuberculosis, currently constitute 4 of the top 10 leading causes of death worldwide [3]. Furthermore, 44% of childhood deaths under five years are attributed to infectious diseases like pneumonia, diarrhoeal diseases, malaria, HIV/AIDS, and measles [4]. While these diseases may no longer pose the imminent threat that they did in the past, there are still regions of the world coping with infectious disease outbreaks. One of the tools now increasingly available at our disposal is mathematical modeling. With mathematical models, researchers in the field of epidemiology can characterize ongoing outbreaks, make comparisons with historical data, and even project future scenarios of the evolving disease with and without medical interventions, all using a simplified mechanistic description of an infectious disease.

Mathematical models can predict the dynamics of an epidemic to provide insight on how to prevent undesirable outcomes [5]. While model predictions may sacrifice quantitative exactness for qualitative correctness, their underlying assumptions render them invaluable approximations of reality [6]. We can extrapolate from current information the number of infected individuals, the duration of the epidemic, the peak incidence, the final size, and ultimately, the entire epidemic curve, providing us with the expected number of cases at each point in time. With this information, we can forecast the occurrence of developing a disease with its respective risk factors [7]. When models fail to predict accurately, this failure can provide opportunities for further epidemiological and experimental studies to discriminate among the competing transmission mechanisms. The deficiencies in our current understanding of the disease of interest can contribute to the design and analysis of epidemiological surveys,

suggest optimal data collection strategies, identify prevailing trends, and quantify the uncertainty in current forecasts [8, 9].

In order to use mathematical models effectively, there must also be confidence that the values used for the various parameters in the model correspond to reality. These parameters encode various, possibly credible epidemiological hypotheses. Although certain parameters can be determined on the basis of prior knowledge, other parameters are often heterogeneous or unobservable in nature. These include the transmission parameters that characterize the unique spreading network of the underlying disease, which must be estimated by fitting the model to the available data. However, available epidemiological data is often incomplete, oversimplified, and subject to measurement and underreporting errors. Nevertheless, models built on such imperfect data can be used as platforms to test hypotheses that may be experimentally difficult or expensive. Fitting epidemiological models to real data can become a key issue during the first phase of an outbreak, where potential interventions have more effect. Models can forecast disease progression and help health officials plan for the latter portion of an outbreak by calculating the parameters from data collected at the start of an epidemic. The diverse set of transmission mechanisms which contribute to the proliferation of each disease can be clarified when equipped with available epidemiological data. Discerning these transmission mechanisms requires quantitative enumeration of the relevant disease components, i.e., the mathematical model.

In epidemiology, it is often impossible to conduct clinical trials or experiments to compare different interventions, due to practical (e.g., expensive, time-consuming) or ethical (e.g., subjecting individuals to lethal pathogens, withholding treatments in control group) constraints. In these cases, mathematical models can evaluate and optimize multiple (often competing) interventions in an attempt towards prudent, efficient decision-making. Accurate modeling and prediction of disease occurrence are critical prerequisites to informative development of intervention strategies. Understanding how diseases begin and spread can ultimately shed light into how they can

be curtailed. As they edge closer to reality, these models can even highlight weak links on the transmission chain, where control efforts should be focused, to prevent, control and eventually eradicate diseases [10, 11]. Furthermore, control frameworks built around mathematical models can respect economic constraints imposed by limited resources when analyzing potential control strategies, eventually informing public policy [12]. Policy makers need to be able to easily interrogate prospective models for relevant intervention outcomes during critical public health situations.

Models restrict their scope of analysis to a particular demographic unit, whether it be a single individual or an entire population. Individual-level models can explicitly incorporate causal factors in disease transmission related to individual behavior and movement, adding a higher level of heterogeneity [13]. Examples of individual-level models include agent-based models and contact networks [13–15]. Agent-based models imbue each individual, or agent, with attributes and directives that enable them to act asynchronously and autonomously, leading to complex, emergent epidemiological phenomena at the population level [16–20]. This bottom-up approach enables the explicit description of both individual nuances in behavior, and global trends in disease spread. Agents operate at discrete time steps during which they move through the simulation environment and perform pre-programmed actions. Consequently, their risk of infection is inevitably linked to their individual behavior. On the other hand, contact networks compromise between the depth of agent-based models and the mathematical simplicity of population-level models by projecting a population's heterogeneous contact patterns onto a graph-theoretic structure, labeling nodes as individuals, and edges as possible contacts [21–24]. Each disease is characterized by the degree distribution of the underlying contact network. Disease propagation in contact networks is explained by a theory known as bond percolation, whereby the size of the infected subgraph can be reliably predicted based on the network's connectivity [21]. Contact networks can also be configured to evolve with respect to time by coupling changes in connectivity to an ordinary differential equation model; the resulting dynamic contact networks evolve according to a form of neighbor exchange,

where individuals have constant degree but swap contacts over time. Interventions can be intuitively applied by manipulating this network structure [21, 25]. In spite of the gains in detail provided by these models, the degree of individuation comes at a price of increased computational burden. Furthermore, the absence of individualized data for model validation and the preference for feasible, population-level interventions in the public health domain limit the applicability of individual-level models to planning and forecasting of outbreaks.

Population-based modeling, on the other hand, is suited to modeling large-scale epidemics and pandemics over broad homogeneous areas. Compartmental models are the mainstay of population-based mathematical modeling in epidemiology. The target population is segmented into distinct units, or compartments, based on each individual's epidemiological status. A hallmark of compartmental models is the susceptible-infected-removed, or SIR, model [26]. The susceptible class can incur the disease but are not yet infected. The infectious class are currently infected and can transmit the disease to others. The removed class are removed from the infection process entirely. A common representation for deterministic epidemic models is ordinary differential equations (ODEs), where the threat of infectious agents invading the population is assumed to change with time [27]. Dynamics emanating from compartmental models exist within a coarse-grained continuum [15]. These ODEs can be fairly complex, depending on the degree of nonlinear interactions involved, requiring the use of numerical methods. Of course, deterministic modeling has its drawbacks. The assumptions held of homogeneously mixing populations and disease persistence, where the infection never completely ceases but can regenerate from small pockets of residual infection, are often criticized as unrealistic [28, 29].

Stochastic models overcome the deficits of their deterministic counterparts by incorporating the many random components involved in propagating infections, like transmission and migration processes. After all, diseases, like all biological phenomena, are stochastic in nature [30]; all natural populations experience some degree of stochasticity. It is important to realize that a given historical record of an epidemic

is but only one possible realization of the underlying process, of which there are infinitely many. Probability distributions govern the outcomes generated from stochastic models. The most important difference between deterministic and stochastic epidemic models is asymptotic dynamics. Eventually stochastic solutions converge to the disease-free state even though the corresponding deterministic solution converges to an epidemic equilibrium [28]. Stochastic models are preferable when studying small communities, where they tend to predominate, and can encapsulate the variability inherent in transmission, recovery, birth and death processes.

Every infectious disease has a unique spatio-temporal "fingerprint", a characteristic of the particular environment and pathogen, which is reflected in its spreading pattern across the population [11]. Any accurate representation of the underlying contact networks (i.e., mathematical model) must account for these epidemiological patterns, in addition to the resulting nonlinearities within the model [30]. One of the principal challenges in epidemiological modeling is realistically estimating transmission rates in spatially structured, heterogeneous host populations, in which hosts differ in susceptibility [31]. Population heterogeneity can endow systems with a complex range of dynamics, where multiple transmission rates determine the spatio-temporal evolution of epidemics in ways that are quite different from homogeneous transmission [31, 32]. On the other hand, classical deterministic epidemic models implicitly assume that space is homogeneous and excludes spatial variation. However, there are instances where spatial homogeneity does not adequately account for the observed behavior of disease transmission. Metapopulation models reflect the spatial heterogeneity in disease transmission that occurs in loosely coupled subpopulations, acting as a "population of populations" where every subpopulation, or patch, contains a local population of individuals [33, 34]. Controlling disease transmission at the metapopulation level is more practical from the viewpoint of policy makers, presenting a manageable level of analysis for potential interventions. It reconciles the countervailing currents of aggregation, meant to operationalize decision making, and disaggregation, meant to provide situational realism [35]. Each subpopulation, or

patch, describes movement of individuals between discrete spatial patches that can be groups, households, villages, cities, provinces, countries, etc. These patches also account for differences in infection risk as the infectious agent moves among them. Factors such as spatial connectivity, environmental conditions, and mobility models can also affect the likelihood that a disease will persist in a given patch.

1.3 Organization of Thesis

Our overall contribution in this thesis work is a set of studies that couples mathematical models of infectious diseases with computational techniques for navigating the space of potential epidemiological scenarios. The objective of these studies is to uncover the necessary public health quantities, like the number of cases and the basic reproductive number, to address different public health challenges:

- Produce model forecasts reflecting the number of cases over time, that are generated from a multi-dimensional parameter space consisting of relevant, sensitive epidemiological parameters, without available data
- Make efficient use of limited, incomplete data to estimate heterogeneous, unobservable parameters to tailor specific interventions for the particular situational context
- Devise efficient, informative field deployment strategies of pathogen sensors in order to collect pathogen information optimally for designing effective intervention strategies

The remainder of this thesis is organized as follows. Chapter 2 focuses on predicting the future course of an epidemic that lacks data and active controls. We explore a stochastic differential equation-based model of a susceptible-infected-vaccinated-removed model and utilize sparse grid interpolation to investigate relevant parameter values that would lead to reduction or complete elimination in the expected cumulative number of cases. Furthermore, we examine how the presence of noise af-

fects accomplishment of this objective with a comparison between deterministic and stochastic models. Other examples in this chapter, outside the field of infectious disease epidemiology, include a breast cancer cell population model and a biochemical network model of the JAK-STAT signaling pathway.

Chapter 3 proposes a parameter estimation approach involving two disease models by exploiting successive sub-grids of the parameter space to identify parameter values consistent with available case data. We conduct a comparative study of various established algorithms, in the domains of cluster analysis and metaheuristics, to both select ranges for local sparse grid interpolants and sample them comprehensively for improved simulations that reflect available outbreak data. Among the models chosen for this endeavor is a stochastic reaction network depicting a SIR process of influenza.

Chapter 4 applies an optimal control strategy with prospective public health interventions to minimize the number of infected individuals within a metapopulation model of cholera with limited information derived from sensor estimates and case data. The underlying algorithm implements an adaptive, multiscenario model predictive control scheme to optimize potential interventions in light of repeating data assimilation cycles that incorporate incoming sensor observations to reconstruct missing state measurements. Sensors for observation in each time interval are chosen according to a predictive optimization criterion that emphasizes minimizing uncertainty in future sensor observations, while simultaneously prioritizing present needs. We present results comparing the usage of different criteria to acquire sensor observations in order to minimize the societal impact of cholera on multiple, interacting populations.

Finally, Chapter 5 concludes the thesis with a discussion on the topics covered and future extensions.

2. SPARSE GRID INTERPOLATION OF ITÔ STOCHASTIC MODELS IN EPIDEMIOLOGY AND SYSTEMS BIOLOGY

2.1 Preface

The research described in this chapter has been published in IAENG International Journal of Applied Mathematics [36].

2.2 Abstract

Certain dynamical models may be unwieldy to simulate repetitively, especially if the models contain uncertainty. This is evident in both epidemiology and systems biology, where inherent biological variability and a spectrum of plausible model hypotheses exist. Surrogate modeling using sparse grid interpolation can alleviate the burden associated with increasing dimension of the parameter space. By leveraging multivariate tensor products across a predefined set of points, sparse grid interpolants are able to provide a promising surrogate model to answer pressing domain-related questions. Specifically, we explore Itô stochastic differential equation-based models, with examples of a susceptible-infectious-vaccinated-removed epidemiological model, a breast cancer tumor population model, and a biochemical network model of the JAK-STAT signal cascade presented. Surrogate modeling is performed to satisfy model-based objectives that implicitly incorporate the presence of noise. Overall, sparse grid interpolation is an effective computational modeling tool, enabling researchers in the epidemiology and systems biology communities to interrogate models of interest for key insight into biological phenomena.

2.3 Introduction

Biological phenomena are inherently complex. This complexity can be simplified for human understanding with mathematical models. Mathematical models condense key biological assumptions and knowledge into a unified representation [37]. Two biological domains that have benefited from mathematical modeling are epidemiology and systems biology. Epidemiology aims to characterize the dynamics of disease spread throughout a population [8]. Systems biology is concerned with the systems-level representation of biological functions and mechanisms underpinning cellular networks [38]. Examples in both domains are commonly represented as mechanistic and semi-mechanistic mathematical models using ordinary differential equations (ODEs), which often have to be solved numerically using discretized approximations of the true solution. However, randomness and heterogeneity can also influence biological systems, requiring the use of stochastic processes [39, 40].

Consider Itô stochastic differential equations (SDEs):

$$d\mathbf{X}(t) = f(\mathbf{X}, t, \boldsymbol{\theta})dt + g(\mathbf{X}, t, \boldsymbol{\theta})d\mathbf{B}(t). \quad \mathbf{X}(0) = X_0. \tag{2.1}$$

where $\mathbf{X} \in \mathbb{R}^N$ is a continuous time stochastic process, $\mathbf{B} \in \mathbb{R}^M$ is a Brownian motion process, $t \in [0, T]$ is time, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^P$ is a vector of model parameters, $f(\cdot) : \mathbb{R}^N \times [0, T] \times \boldsymbol{\Theta} \to \mathbb{R}^N$ is the drift term (deterministic component), X_0 are the initial conditions, and $g(\cdot) : \mathbb{R}^N \times [0, T] \times \boldsymbol{\Theta} \to \mathbb{R}^{N \times M}$ is the diffusion term (stochastic component). Examples of SDE-based models in epidemiology and systems biology include the human nervous system [41–43], cancer tumors [44], predator-prey systems [45, 46], and a glucose regulatory system for diabetes patients [47].

Complex system dynamics can be difficult to simulate when a large number of model parameters have to be considered [48–50]. Furthermore, local searches of these parameters may be insufficient to characterize the wide range of possible behaviors. Sparse grids allow for global, computationally efficient exploration of the parameter space $\boldsymbol{\Theta}$ using tensor-product quadrature [51–53]. These approximations of the underlying model mitigate the curse of dimensionality associated with the increasing

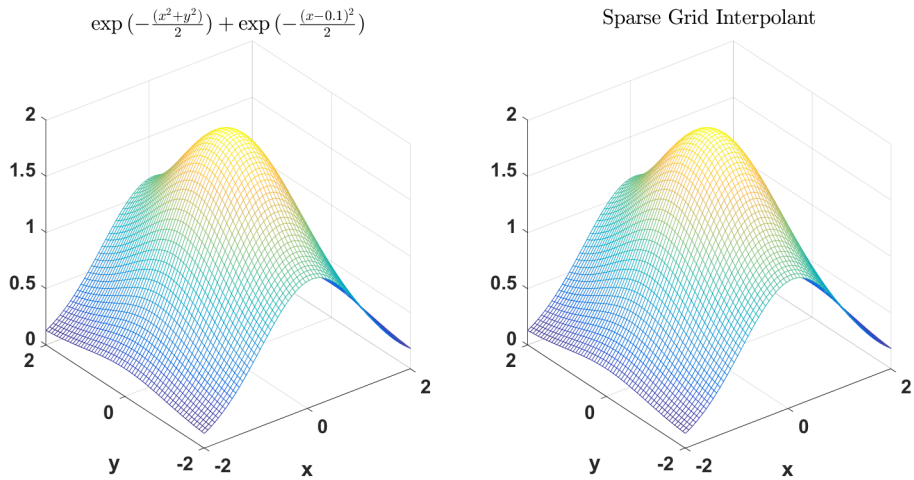


Fig. 2.1. Exponential function evaluated on a grid $[-2, 2] \times [-2, 2]$. Both the original function (*left*) and the sparse grid interpolant (*right*) are shown. The interpolant was produced with a relative error of 0.021%, absolute error 0.00039, and 129 support nodes.

dimension of $\boldsymbol{\Theta}$ by selecting the grid points, or support nodes, in a hierarchical fashion [52–54]. This is done so that nodes from a previous level of refinement can be reused in higher levels of refinement. Once the original model has been evaluated at these support nodes and the interpolant has been constructed, the resulting surrogate model can be used in model-based optimization without having to directly integrate the underlying model, which is often computationally prohibitive. The concept of sparse grid interpolation, and surrogate modeling in general, is not unlike that of compressive sensing, where a compressible signal is recovered from a limited number of measurements [55]. Fig. 2.1 demonstrates the application of sparse grid interpolation to a simple 3-dimensional exponential function. Sparse grids have been applied to other stochastic models, such as stochastic partial differential equations with random inputs [56–61], backwards stochastic differential equations with random inputs [62], and differential algebraic equations with random parameters [63].

We demonstrate the application of sparse grid interpolation to approximating the dynamics of Itô SDE-based models in different biological contexts. In Section 2.4,

we discuss the concept of sparse grids, and the necessary numerical techniques for effective interpolation. Then, in Section 2.5, we present examples of sparse grid interpolation through targeted computational experiments that approach domain-related problems. Specifically, we examine the role that noise plays in perturbing normal biological function, or whether there is any discernible influence of noise at all. Finally, in Section 2.6, we summarize the significance of our work and propose future avenues of research.

2.4 Methodology

2.4.1 Sparse Grid Interpolation

In sparse grid interpolation, the support nodes are selected in a predefined manner; a nested, hierarchical sampling scheme [52, 54, 64] recycles nodes from lower levels of resolution to use in higher levels.

A mathematical formulation of sparse grids now follows from [51, 52, 64–67]. Consider a function $f : [0, 1]^d \to \mathbb{R}$ that is to be interpolated on a finite number of support nodes. Dimensions that are not of unit length can be rescaled. Here, f represents the sample average of multiple SDE trajectories sampled at discrete time points. For a given f, a univariate interpolation function can be constructed:

$$\mathcal{U}^i(f) = \sum_{j=1}^{m_i} a_j^i \cdot f(x_j^i), \tag{2.2}$$

where $i \in \mathbb{N}$, $a_j^i \in C([0, 1])$, $a_j^i(x_l^i) = \delta_{jl}$, $l \in \mathbb{N}$ are the univariate basis functions, and $x_j^i \in X^i = \{x_1^i, \ldots, x_{m_i}^i\}$, $x_j^i \in [0, 1]$, $1 \leq j \leq m_i$, are the support nodes.

Extending this interpolation function to multi-dimensional cases (i.e. $d \geq 1$), the corresponding multivariate formula, using the full tensor product formulation, is as follows:

$$(\mathcal{U}^{i_1} \otimes \cdots \otimes \mathcal{U}^{i_d})(f) = \sum_{j_1=1}^{m_{i_1}} \cdots \sum_{j_d=1}^{m_{i_d}} (a_{j_1}^{i_1} \otimes \cdots \otimes a_{j_d}^{i_d}) f(x_{j_1}^{i_1}, \ldots, x_{j_d}^{i_d}). \tag{2.3}$$

The number of support nodes required for the full tensor product representation is $\prod_{j=1}^{d} m_{i_j}$, which is computationally intractable for high dimensions d.

The Smolyak construction aims to substantially decrease the number of support nodes used while preserving the interpolation properties observed in the 1-dimensional case. Define the difference function $\Delta^i = \mathcal{U}^i - \mathcal{U}^{i-1}$, $\mathcal{U}^0 = 0$ and multi-index $\mathbf{i} \in \mathbb{N}^d$, $|\mathbf{i}| = i_{i_1} + \cdots + i_{i_d}$. Now, define the Smolyak interpolant as:

$$A_{n+d,d}(f) = \sum_{k=0}^{n} \sum_{|\mathbf{i}|=k+d} (\Delta^{i_1} \otimes \cdots \otimes \Delta^{i_d})(f). \tag{2.4}$$

The inner sum can be further expressed as

$$\sum_{|\mathbf{i}|=k+d} \sum_{\mathbf{j}} (a_{j_1}^{i_1} \otimes \cdots \otimes a_{j_d}^{i_d})(f(\mathbf{x}_{\mathbf{j}}^{\mathbf{i}}) - A_{k+d-1,d}(f(\mathbf{x}_{\mathbf{j}}^{\mathbf{i}}))), \tag{2.5}$$

where $\mathbf{j}$ is the multi-index $(j_1, \ldots, j_d)$, $j_l = 1, \ldots, m_{i_l}^{\Delta}$, $l = 1, \ldots, d$, and the points $\mathbf{x}_{\mathbf{j}}^{\mathbf{i}} = (x_{j_1}^{i_1}, \ldots, x_{j_d}^{i_d})$, $x_{j_l}^{i_l}$ is the j_l^{th} element of $X_{\Delta}^{i_1} = X^{i_l} \backslash X^{i_l - 1}$, $X^0 = \emptyset$, and $m_{i_l}^{\Delta} = |X_{\Delta}^{i_l}|$. The support nodes can be chosen in an hierarchical manner such that $X^i \subset X^{i+1}$, $i \in \{i_1, \ldots, i_d\}$.

It is also useful to compute the absolute (E_{abs}^n) and relative (E_{rel}^n) errors of the Smolyak interpolant using correction terms known as hierarchical surpluses ($w_{\mathbf{j}}^{k,\mathbf{i}}$):

$$w_{\mathbf{j}}^{k,\mathbf{i}} = f(\mathbf{x}_{\mathbf{j}}^{\mathbf{i}}) - A_{k+d-1,d}(f(\mathbf{x}_{\mathbf{j}}^{\mathbf{i}})), \tag{2.6}$$

$$E_{abs}^n = \max_{\mathbf{i},\mathbf{j}} w_{\mathbf{j}}^{n,\mathbf{i}}, \tag{2.7}$$

$$E_{rel}^n = \frac{\max_{\mathbf{i},\mathbf{j}} w_{\mathbf{j}}^{n,\mathbf{i}}}{\max_{\mathbf{i},\mathbf{j}} f(\mathbf{x}_{\mathbf{j}}^{\mathbf{i}}) - \min_{\mathbf{i},\mathbf{j}} f(\mathbf{x}_{\mathbf{j}}^{\mathbf{i}})}. \tag{2.8}$$

The conventional sparse grid fails to consider the impact errors can have on the quality of the interpolant produced. Adaptive sparse grids [51] build on the conventional formulation by using generalized error indicators that consider the influence of the error in comparison to the necessary computational work:

$$g_{\mathbf{j}} = \max \left\{ w \frac{|\Delta^{\mathbf{j}} f|}{|\Delta^{\mathbf{1}} f|}, (1-w) \frac{n_{\mathbf{1}}}{n_{\mathbf{j}}} \right\}, \tag{2.9}$$

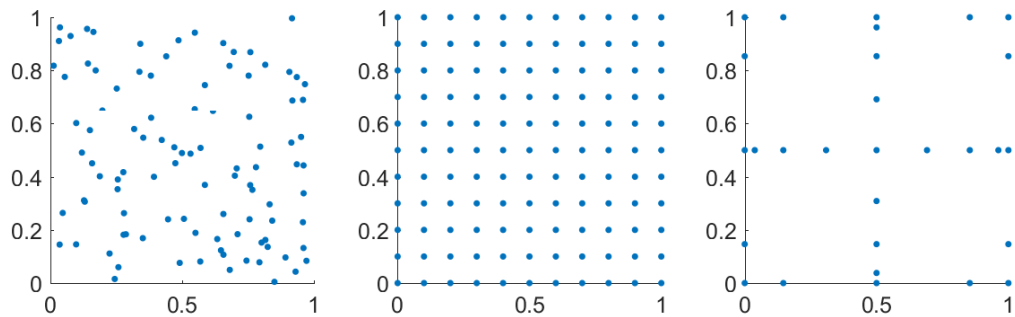


Fig. 2.2. Compared to randomly (*left*), and uniformly (*center*) sampled grids, sparsely sampled grids, like the Chebyshev-Gauss-Lobatto grid (*right*), strategically sample the parameter space to produce error controlled surrogate models that use fewer samples.

where $w \in [0, 1]$ is a weight for the error indicator $g_{\mathbf{j}}$, $n_{\mathbf{k}}$ is the number of function evaluations for an index set $\mathbf{k}$. Conventional sparse grids are formed when $w = 0$, and only the number of function evaluations are considered. When $w = 1$, the error indicators will decay with increasing indices. Intermediate values of w compromise between excessive work and high error.

Grid Type

The approximation properties of the sparse grid rely on basis functions to select the required support nodes. Chebyshev-based node distributions can be used for higher-order polynomial interpolation, where the function to be interpolated is smooth and higher accuracy is required [68]. In this work, we use Chebyshev-Gauss-Lobatto nodes [66], which are defined as follows:

$$m_i = \begin{cases} 1, & i = 1 \\ 2^{i-1} + 1, & i > 1 \end{cases} \tag{2.10}$$

$$x_j^i = \begin{cases} -\cos\frac{\pi\cdot(j-1)}{m_i-1}, & m_i > 1 \\ 0, & m_i = 1, \end{cases} \tag{2.11}$$

where m_i is the number of support nodes for level i, and x_j^i is the position of the jth node at level i, $j = 1, \ldots, m_i$.

Time Domain Interpolation

In addition to inteprolation across the parameter space, there is also the issue of time domain interpolation. Choosing nodes in the time domain to accurately represent a trajectory may influence the accuracy of the resulting sparse grid interpolant. Time intervals can be either uniform or non-uniform. With non-uniform time points, a possibility is to utilize the extrema of the Chebyshev polynomials as was done in [49, 50] for ODE models:

$$T_s^\ell = T_{min}^\ell + \left(1 - \cos\left(\frac{\pi s_\ell}{d}\right)\right)\frac{T_{max}^\ell - T_{min}^\ell}{2}, \tag{2.12}$$

where $\ell \in \{1, \ldots, n\}$ is a vector of indices corresponding to model outputs, d is the degree of the interpolating Lagrange polynomial, T_s^ℓ is a vector of sampling times, T_{min}^ℓ is the minimum time, T_{max}^ℓ is the maximum time, and $s_\ell = [0, \ldots, d]$. Choosing the extrema of Chebyshev polynomials can reduce the effect of poor interpolation on the edges of an interval that occur when using equidistant nodes, a problem known as the Runge phenomenon [69].

Once the model outputs are sampled at these times, they can be evaluated at other times t, $T_{min}^\ell \leq t \leq T_{max}^\ell$:

$$\tilde{y}_\ell(\boldsymbol{\theta}, t) = L_d^\ell(t) \cdot \hat{y}_\ell(\boldsymbol{\theta}, T_s^\ell), \tag{2.13}$$

where $\tilde{y}_\ell(\boldsymbol{\theta}, t)$ is the interpolated model output with parameters $\boldsymbol{\theta}$ at time t, $\hat{y}_\ell(\boldsymbol{\theta}, T_s^i)$ is the sparse grid model output sampled at the times T_s^ℓ, L_d^ℓ is the Lagrange interpolating polynomial for the ℓth model output with degree d, defined in [70].

Simulation Conditions

Matlab was used as the simulation environment for the models discussed here. The Euler-Maruyama method, a first-order stochastic Taylor expansion, was used to integrate SDEs [71–73]:

$$\mathbf{X}(t_{k+1}) = \mathbf{X}(t_k) + f(\mathbf{X}(t_k), k\delta t, \boldsymbol{\theta})\delta t + g(\mathbf{X}(t_k), k\delta t, \boldsymbol{\theta})(B(t_k) - B(t_{k-1})), \quad (2.14)$$

where δt is the integration time step. Sparse grid interpolation was performed using the Sparse Grid Interpolation Toolbox [68].

Each model had to be tuned for compatibility with sparse grid interpolation by choosing both the simulation conditions and the number of realizations. Simulation conditions for the model, such as initial conditions, timespan of the simulation, desired model states, and parameters to include in the parameter space, were determined first. These conditions were defined in large part to conform with the scope of the examples presented in this work.

2.5 Computational Experiments

2.5.1 SIVR Model

We first examine a model describing the spread of an infectious disease, known as the susceptible-infectious-vaccinated-removed (SIVR) model [74]. This system includes a vaccination mechanism by which certain individuals may avoid infection for a limited period of time. It is described as follows:

$$dS = [\mu - \beta SI - (\mu + \phi)S]dt - \sigma SIdB(t) \quad (2.15)$$

$$dI = [\beta SI + \rho\beta VI - (\lambda + \mu)I]dt + \sigma(S + \rho V)IdB(t) \quad (2.16)$$

$$dV = [\phi S - \rho\beta VI - \mu V]dt - \rho\sigma VIdB(t) \quad (2.17)$$

$$dR = [\lambda I - \mu R]dt. \quad (2.18)$$

Susceptible individuals (S) can contract the infection, after which they are infected (I), and can infect other susceptible individuals. Vaccinated individuals (V) may

be partially resistant to infection upon vaccination, but are not completely immune. After recovering from an infection, removed individuals (R) no longer participate in the infection process. The values of each disease state are expressed as percentages by normalizing to the overall population size. The parameters of interest in this model and the predefined parameter ranges are described in Table 2.1. The stochastic perturbations in the SIVR model have been integrated into models of real-world diseases, such as HIV [75].

For demonstration purposes, we investigate those epidemiological parameter values that result in the average number of cases being less than some percentage of the total population $\mathcal{C}_{crit}$. Minimizing the average number of cases is a practical disease eradication objective that would also bound the number of deaths in a real-world context. Expressed mathematically, our goal is to obtain the set of acceptable parameters

$$\boldsymbol{\Theta}_A = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} | \mathbb{E}[C_{\boldsymbol{\theta}}(T)] < \mathcal{C}_{crit}\}, \tag{2.19}$$

where $\mathbb{E}[C_{\boldsymbol{\theta}}(T)]$ is the expected number of cases at time T with parameters $\boldsymbol{\theta} = \{\lambda, \beta, \mu, \phi, \rho, \sigma\}$. We set $T = 100$ days, with $\mathbf{X}(0) = [0.85, 0.1, 0.05, 0]^\intercal$. Additionally, we define C as follows:

$$C = I + R, \tag{2.20}$$

with $C(0) = 0.1$. This formulation of the number of cases captures the percentage of the population who have experienced the infection process. We also compare the ODE and SDE versions of the model to determine what, if any, differences exist, in trying to determine the percentage of acceptable parameters and the basic reproductive number $\mathcal{R}_0$. The ODE-based sparse grid interpolant produced had a relative error of 0.75% and an absolute error of 0.0071 with 209 support nodes, while the SDE-based interpolant had a relative error of 0.83% and an absolute error of 0.008 with 427 support nodes. The number of realizations for the SDE model at each point in the parameter space was selected to satisfy a statistical error criterion [76, 77]:

$$\epsilon_S = c_0 \frac{\mathscr{S}(C_\theta(T), K)}{\sqrt{K}} \leq TOL \tag{2.21}$$

where $c_0 \geq 1.65$, $\mathscr{S}(C_\theta(T), K)$ is the sample standard deviation of $C_\theta(T)$ with K realizations, and $TOL = 10^{-3}$ is the error tolerance. 10,000 parameter samples from the given ranges in Table 2.1 were obtained through Latin Hypercube Sampling (LHS). These ranges were determined through manual tuning to avoid negative dynamics or dynamics outside the normalized range $[0, 1]$. Then, model dynamics corresponding to these sampled parameters were interpolated using the surrogate model.

Table 2.1.
Parameters of SIVR model, with definitions and ranges used in sparse grid interpolation.

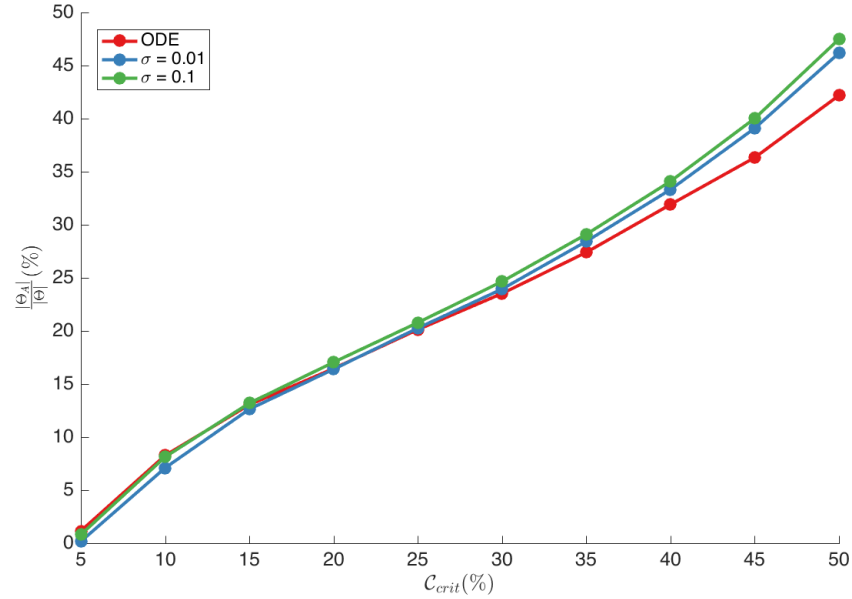
Parameter	Definition	Units	Range
λ	Recovery rate	days^{-1}	$[0, 0.01]$
μ	Birth/death rate	days^{-1}	$[0, 0.01]$
β	Transmission rate	days^{-1}	$[0, 0.4]$
ϕ	Vaccination rate	days^{-1}	$[0, 0.1]$
ρ	Vaccination efficacy	dimensionless	$[0, 0.01]$
σ	Environmental noise	days^{-1}	$[0.01, 0.1]$

Fig. 3.3(a) illustrates how the percentage of acceptable parameters increases as the case threshold is increased. For all three modeling contexts, there is a drastic increase in the number of cases as nearly half of parameters are deemed acceptable as a case threshold of 50% is allowed. There appears to be no saturation point by the 50% mark for $\mathcal{C}_{crit}$, as there is a continual ascent.

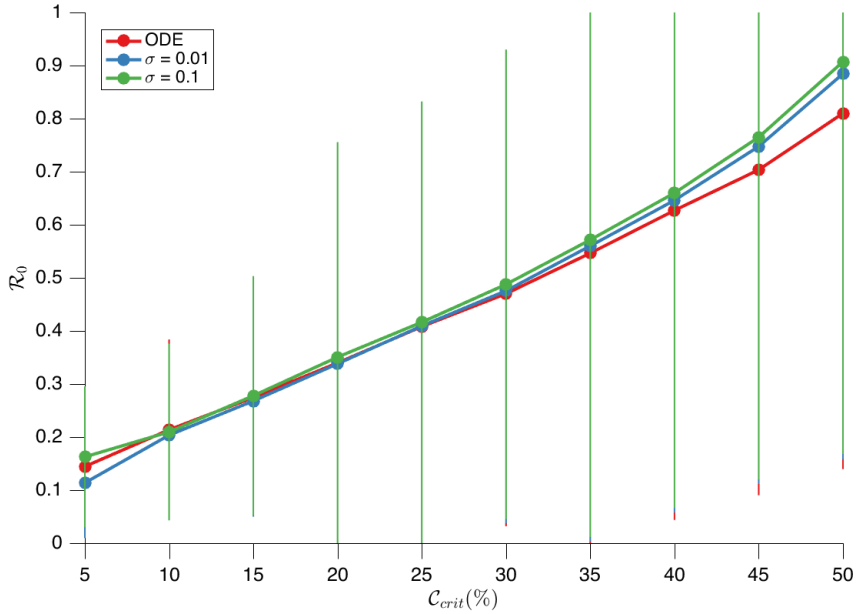
Figure 3.3(b) depicts the mean and standard deviation of $\mathcal{R}_0$ values computed for each parameter set as a function of $\mathcal{C}_{crit}$. For this model, $\mathcal{R}_0$, the basic reproductive number, is defined as [74]:

$$\mathcal{R}_0 = \frac{\beta}{\mu + \lambda} \frac{\mu + \rho\phi}{\mu + \phi}. \tag{2.22}$$

Medical professionals often refer to the $\mathcal{R}_0$ value of particular diseases to inform them of the current state of the disease. Knowledge of the maximum case loads possible to sustain a given $\mathcal{R}_0$ value gives a meaningful target in terms of available



(a) Percentage of acceptable parameters as a function of $\mathcal{C}_{crit}$.



(b) $\mathcal{R}_0$ values as a function of $\mathcal{C}_{crit}$ (Mean ± SD).

Fig. 2.3. The impact of varying $\mathcal{C}_{crit}$ on modeling and epidemiological measures.

resource allocation strategies and treatment options. Lower case loads translate to lower reproductive potentials, as the disease fails to adequately propagate for increased transmission. This is observed for the ODE and SDE models, as $\mathcal{R}_0 \leq 1$. The reproductive number for acceptable parameter-based simulations increases as the allowable case burden increases, but the average reproductive number remains below 1, suggesting that scenarios where 50% or less of the population have experienced infections are in disease contexts where the disease fails to adequately propagate. The average reproductive potential of the disease (over 10,000 simulated parameter values) indicates a non-escalation of a disease outbreak into a full-scale epidemic. Averages and standard deviations for both sets rise with $\mathcal{C}_{crit}$, indicating a process where borderline unacceptable parameters are slowly pushed to the acceptable set, raising the averages of both sets in the process. This transfer significantly alters the composition of the unacceptable set by introducing more variability in the form of a higher standard deviation. By leaving the unacceptable set and joining the acceptable set, the standard deviation of the acceptable set increases in accommodating these formerly unacceptable parameter values.

We acknowledge that in the attempt to demonstrate the link between the number of cases and the basic reproductive number, there are limitations to this study, especially when it comes to choosing a stochastic epidemic model and specifying the number of days to simulate. Our purpose in comparing ODE and SDE results was to determine if there was any difference between the two modeling approaches and what they could be attributed to. Differences were observed for both Figs. 3.3(a) and 3.3(b) as $\mathcal{C}_{crit}$ increased above 30%. The difference between the ODE and SDE results may be due to the fact that the sample mean of the diffusion term in the SDE model, is non-zero. This non-zero sample mean may propagate through the interpolation process to produce interpolated results that differ from the ODE results. Moreover, the non-zero sample mean may be a result of applying the statistical error criterion in choosing the number of realizations. In certain cases, the number of realizations cho-

sen may be fewer; these few realizations would have a larger influence on the expected value being computed, especially if they did not represent the population mean well.

The limitations in devising an objective that explores the number of cases occurring by a certain time point, is that any dynamics after that time are not accounted for. We extended the duration of simulation to mitigate this possibility, and highlight some results about our decision to choose 100 days in lieu of a longer time period like 200 days, but acknowledge that this may not cover all possible scenarios. We computed the relative error in choosing $T = 100$ days as opposed to a longer time span (e.g., $T = 200$ days) across the 10,000 sampled parameters using the ODE model. Figure 2.4 depicts a boxplot of the relative errors for the set of considered parameters. A median relative error of 3.92% and a mean relative error of 8.26% was found.

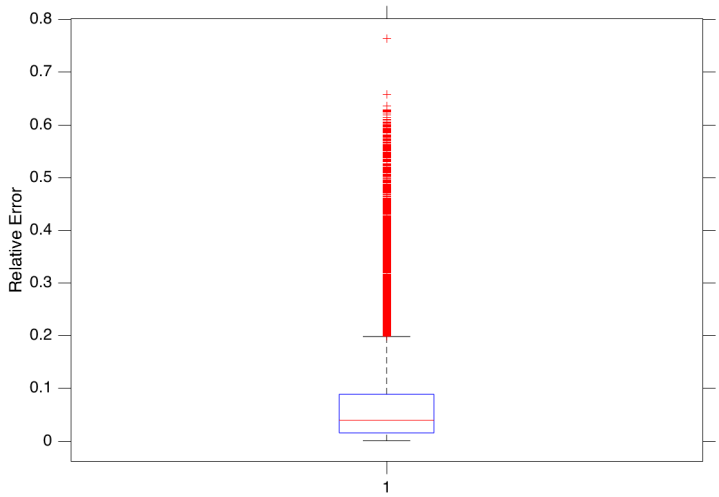


Fig. 2.4. Boxplot of relative errors of cases derived at $T = 100$ days as opposed to $T = 200$ days across 10,000 parameters sampled using LHS.

2.5.2 MCF-7 Breast Cancer Model

The MCF-7 breast cancer model was developed to predict tumor responses to radiotherapy and other therapeutic treatments [78]. To capture the deleterious and vari-

able effects of radiation on cancer cells, the model added noise terms to the cell death rates for the three cancer sub-populations being studied. These sub-populations, sorted according to radiotherapy sensitivity, represented stages of the cell cycle: the gap phase (G), the synthesis phase (S), and the mitosis phase (M). The model is described as follows:

$$dG = [-(\alpha + q_1)G + 2\gamma M]dt - \sigma G dB_1(t) \tag{2.23}$$

$$dS = [\alpha G - (\beta + q_2)S]dt - \sigma S dB_2(t) \tag{2.24}$$

$$dM = [\beta S - (\gamma + q_3)M]dt - \sigma M dB_3(t) \tag{2.25}$$

where q_i, $i = 1, 2, 3$ are the specific death rates for each sub-population, α is the transition rate from G to S, β is the transition rate from S to M, γ is the transition rate from M to G, and σ is the magnitude of the stochastic noise.

Table 2.2.
Parameters of MCF-7 model, with definitions and ranges used in sparse grid interpolation.

Parameter	Definition	Range
α	Transition rate from G to S	$[-0.0052, 0.0918]$
β	Transition rate from S to M	$[0.0315, 0.1333]$
γ	Transition rate from M to G	$[0.1744, 0.9055]$
σ	Environmental noise	$[0, 0.1]$

In addition to incorporating stochastic noise into the cancer model, [78] introduced a measure known as the tumor lifespan L, defined as the amount of time needed to eradicate the cancer:

$$L = \min\{t : G(t) + S(t) + M(t) = 0\}. \tag{2.26}$$

The tumor lifespan was introduced to evaluate cancer treatment effectiveness. Multiple treatment strategies can be ranked based on how much they reduced L. A mean

tumor lifespan of 175 hours was calculated for the nominal parameters presented in [78].

While L has been evaluated on parameters found to best fit existing data on this form of breast cancer, understanding the impact that the stochastic noise term has on L would clarify its influence on cancer proliferation. To accomplish this, we employ sparse grid interpolation to observe the tumor lifespan landscape for 200 MCF-7 cancer cells at the end of 200 hours with varying noise levels. The parameters used to form the parameter space, and their associated parameter ranges as reported in [78], are described in Table 2.2. The sparse grid interpolant produced had a relative error of 0.93% and an absolute error of 0.1719 with 249 support nodes.

Fig. 2.5 illustrates this landscape in 3-dimensional form for 10,000 uniformly sampled points in the parameter space, with varying noise levels. If there were still cancer cells present at the end of 200 hours, the tumor lifespan was set to 200 hours. The top row, where only γ is varied, shows a clear discrepancy between areas of decreased tumor lifespan and the maximum plateau of 200 hours. Specifically, for $\alpha \leq 0.01$ and $\beta \leq 0.08$, the tumor lifespan declines to as much as 110 hours. Lower transition rates tend to suspend cell viability and lifespan. On the other hand, higher transition rates retain the existing cellular machinery, promoting cell growth and division. Increasing the noise levels also did not significantly alter this landscape or the minimum lifespans. Observing the tumor lifespan landscape for α and γ, where β is held constant reveals some interesting features. The bottom row of Fig. 2.5 highlights two distinct regions of decreased tumor lifespan, where $\alpha \leq 0.005$ and $0.17 \leq \gamma \leq 0.28$, $0.55 \leq \gamma \leq 0.9$. The minimum lifespan attained in these areas are approximately 150 hours. While this area appears for all three noise levels, what differentiates each level is the prevalence of abnormal contours emblematic of noise. Noise pervades the decreased lifespan areas in the form of peaks, starting at the minima of both parameters. The quantity and size of these peaks increase as the noise level increases.

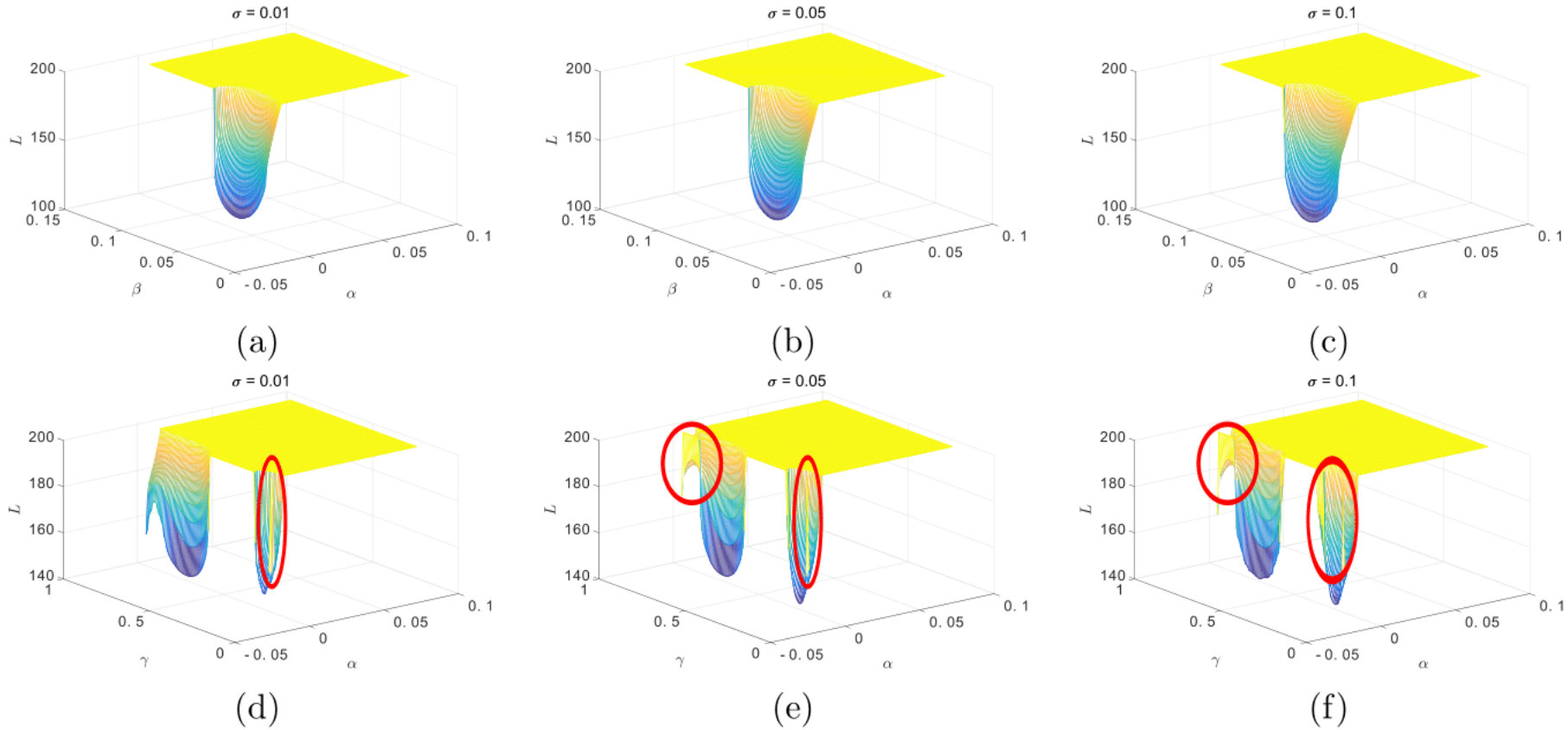


Fig. 2.5. Tumor lifespan landscape with varying noise levels. Top row varies α and β, with $\gamma = 0.3655$. Bottom row varies α and γ, with $\beta = 0.0824$. Red circles denote regions distorted by noise.

2.5.3 JAK-STAT Signaling Pathway

Parameter estimation in systems biology aims to reconstruct dynamic inter- and intracellular biochemical relationships from available data [79, 80]. The JAK-STAT signaling pathway SDE, derived from an earlier ODE model [81], is described as follows [82, 83]:

$$dx_1 = [-k_1 x_1 EpoR + 2k_4 z_1]dt + \sigma x_1 dB(t) \tag{2.27}$$

$$dx_2 = [k_1 x_1 EpoR - k_2 x_2^2]dt \tag{2.28}$$

$$dx_3 = [-k_3 x_3 + \frac{1}{2} k_2 x_2^2]dt \tag{2.29}$$

$$dx_4 = [k_3 x_3 - k_4 z_1]dt \tag{2.30}$$

$$dz_1 = \Gamma(t)[x_3 - z_1]dt \tag{2.31}$$

$$\Gamma(t) = \frac{\alpha}{1 - A^\alpha \exp(-\alpha t)}. \tag{2.32}$$

This model of the JAK-STAT signaling pathway can be described by a number of steps [81]. Erythropoietin receptor ($EpoR$) is activated by erythropoietin hormone binding, phosphorlyating cytoplasmic STAT5 (x_1). Phosphorylated STAT5 (x_2) then proceeds to dimerize (x_3), after which it is then imported into the nucleus (x_4). In the nucleus, dissociation and dephosphorylation of STAT5 occur with a time delay (z_1).

A readily measurable output of this system is the total phosphorylated STAT5 y, defined as follows:

$$y = s(x_2 + 2x_3), \tag{2.33}$$

where s is a scaling parameter.

We rely on a nonparametric simulated maximum likelihood approach using kernel density estimation for parameter estimation [84]. The approach approximates the transition densities of the maximum likelihood function by comparing all generated realizations with observed data. We note that parameter estimation approaches have been applied previously using sparse grid interpolation [85–87]. The corresponding

Table 2.3.
Parameters of JAK-STAT model, with definitions and ranges used in sparse grid interpolation.

Parameter	Definition	Range
k_1	STAT5 phosphorylation rate	$[0.015, 0.025]$
k_2	STAT5 dimerization rate	$[0.015, 0.025]$
k_3	Nuclear import rate	$[0.1, 0.15]$
k_4	Nuclear export rate	$[0.05, 0.1]$
α	Delay function parameter	$[0.05, 0.5]$
A	Delay function parameter	$[10^{-4}, 10^{-2}]$
σ	Environmental noise	$[0.05, 0.2]$

log likelihood function was then computed at the support nodes and subsequently interpolated across the parameter space described in Table 2.3. The optimal parameter estimates minimized the log likelihood function. We set the duration of the simulation at 60 minutes, and $\mathbf{X}(0) = [2.3, 0.01, 0.01, 0.01, 0]^\intercal$. Data obtained from [81] was used for parameter estimation. The sparse grid interpolant produced had a relative error of 0.52% and an absolute error of 0.38 with 481 support nodes. 10,000 LHS sampled parameters were generated from the prescribed parameter ranges, and the corresponding trajectories were estimated using the sparse grid interpolant. We plot and compare the results for three different noise levels, shown in Figure 2.6.

The log likelihood values for $\sigma = 0.05$, 0.1, and 0.2, were $6.1893{*}10^{-4}$, $4.353{*}10^{-4}$, and 4.5854, respectively. Higher noise levels resulted in a dramatic loss of fit quantitatively, although all noise levels possessed great qualitative fits. This example demonstrates the applicability of sparse grid interpolation to parameter estimation of SDEs within a maximum likelihood framework.

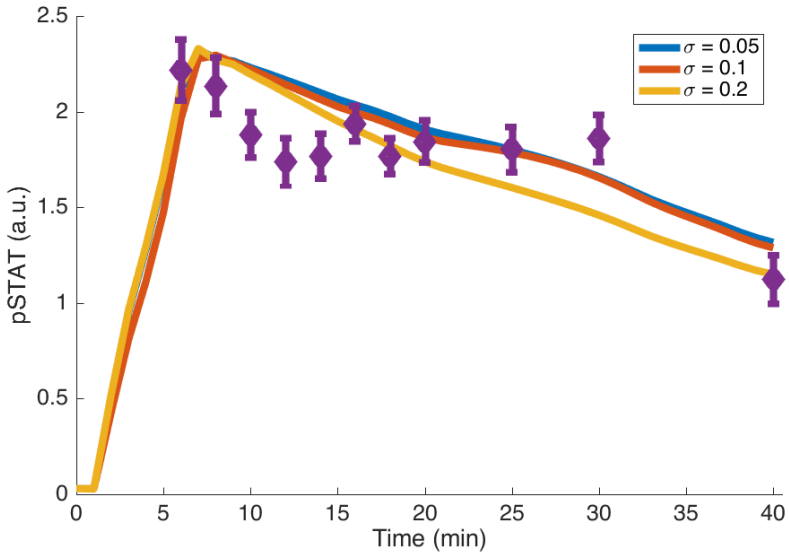


Fig. 2.6. Results of parameter estimation with JAK-STAT pathway model across three different noise levels. Dataset is in purple (mean ± SD).

2.6 Conclusion

Sparse grids produce effective interpolants without sacrificing much of the modeling accuracy and incurring the cost of unnecessary model evaluations. These unnecessary model evaluations materialize in both the parameter and uncertainty spaces, with multiple parameter values and realizations necessary for an adequate model description. The approach discussed here interpolates the solution provided by an average SDE trajectory at each support node in a parameter space of moderate dimension. The stochastic noise was also considered as a dimension of the parameter space, and played an important role in the examples presented. Our work serves as a computationally efficient surrogate modeling-based exploration of the stochastic dynamics of SDE models. We acknowledge our limitations in truly capturing the stochastic process underlying these models, especially the SIVR model. To address this in the future, we endeavor to explore more complex forms of noise and output higher statistical moments in the interpolation process.

3. PARAMETER ESTIMATION IN EPIDEMIOLOGY USING SPARSE GRID INTERPOLATION

3.1 Preface

The research described in this chapter has been submitted to the Journal of Biological Dynamics.

3.2 Abstract

We consider the problem of using time-series data to calibrate compartment-based epidemiological models. Our two-stage algorithm identifies potentially optimal regions of the parameter space and directs computational effort towards resolving the dynamics of these regions. To facilitate this endeavor, we rely on sparse grid interpolation, a popular numerical discretization technique for the treatment of high dimensional, multivariate problems, to capture the dynamics underlying both global and local spaces. By employing cluster analysis techniques and metaheuristic algorithms, we show through two case studies that definitive gains in performance can be made to produce simulated outcomes consistent with available data to infer epidemiologically relevant parameters.

3.3 Introduction

Mathematical models of biological phenomena rely on parameters to capture model behavior [88]. Parameter values must be estimated with accuracy to provide any meaningful insight into critical biological problems. Limited prior knowledge on parameter regimes often prohibits targeted or smart sampling strategies, hindering efforts at successful parameter estimation. In the domain of epidemiology, many

parameters are not easily derived from literature, nor directly observable from available data, and yet are indispensable to characterizing the force of infection within a population. Parameter estimation in epidemiology usually relies on approaches like Bayesian [89–91], likelihood-based [92–95], evolutionary computing [96], and least squares methods [97, 98]. We propose an alternative parameter estimation strategy that can operate independently of prior parameter estimates on models containing many parameters. In furtherance of this approach, we use sparse grid interpolation, a surrogate modeling technique, to estimate relevant model dynamics across a predefined parameter space. By enclosing the estimation problem within a proxy environment, sufficient samples can be taken to obtain a comprehensive assessment of parameter fitness at a fraction of the cost of directly simulating the model. Furthermore, we attack the parameter estimation problem by making use of both global and local searches of the parameter space. This approach, previously pursued in [99, 100], is at the crux of our proposed two-stage algorithm. A two-stage approach has been explored previously by [101] to infer parameters of the basic reproductive number for a discrete age-structured model using incidence data from one or multiple disease outbreaks. The first stage involved a direct estimation of the parameters to generate priors, which were then refined by a second stage of maximum likelihood estimation. When applied to influenza-like illness data, the approach obtained good estimates of the age-dependent basic reproductive number and the population's age-specific susceptibility. However, the use of maximum likelihood optimization may not entirely avoid local minima and may be inappropriate for high dimensional parameter spaces.

The purpose of this paper is to suggest an intuitive, easily implementable two-stage algorithm to inform parameter values for population-based epidemiological models equipped with available time-series data. This paper is organized as follows. We review sparse grid interpolation in Section 3.4. Section 3.5 revisits an earlier method of identifying acceptable parameters and proposes a two-stage parameter estimation algorithm, which makes use of cluster analysis and metaheuristic algorithms. Cluster analysis specifically addresses the selection of ranges for localized searches, while

metaheuristic algorithms embed within the sampling process to iteratively locate improved parameter values. In Section 3.6, we demonstrate our approach on two compartment-based infectious disease models, which depict well-mixed population flows of individuals in various epidemiological states [8]. Results show an improvement in parameter estimation with fewer model evaluations when either cluster analysis or metaheuristic algorithms are employed. Section 3.7 analyzes our findings and offers some perspective. Finally, Section 3.8 summarizes our contribution and suggests future extensions.

3.4 Sparse Grid Interpolation

Approaching a problem like parameter estimation using mathematical models entails its own challenges. A sufficiently well parameterized model may require a high-dimensional parameter space. At these higher dimensions, the model may even be computationally expensive to simulate, deeming the parameter estimation problem intractable. On the other hand, the global diversity of model behaviors desired for accurate parameter estimation may be forfeited by compromising on the simulation effort. Computationally intensive models also present a similar obstacle, where it is desired to minimize the number of direct model evaluations as much as possible.

Sparse grid interpolation presents a viable, parsimonious solution to these challenges. By sampling the parameter space strategically and selectively, sparse grid interpolants closely approximate the target model [52, 53, 65, 66]. The interpolant is constructed by combining basis functions at a set of sparsely sampled points across the parameter space. By interrogating the interpolant rather than the target model, excessive and costly model evaluations can be avoided. The concept of sparse grid interpolation can be traced back to the Russian mathematician Smolyak, who developed an efficient technique to extend tensor product formulas for numerical integration, or quadrature, to multiple dimensions [102]. Smolyak's algorithm takes the partial tensor product of univariate quadrature rules instead of the full tensor

product representation, minimizing the number of points used while maintaining an error up to a logarithmic factor [52, 53, 65, 66, 103]. Important features of sparse grid interpolation include hierarchical decomposition and dimensional adaptivity. The hierarchical property of sparse grids allows for points to be reused at higher levels of refinement, meaning only points unique to the higher level are evaluated [52, 54, 65]. Adaptive sparse grids place more points along dimensions of the parameter space that contribute most to the interpolation error to produce a smoother, more accurate interpolant [51, 104].

Sparse grids have been used to aid efforts in parameter identification before [85–87]. Adaptive sparse grid-based optimization was used to identify promising regions of the parameter space with respect to alignment with available data, with further extensions in robustness analysis [86], and multi-scenario control [87]. In particular, [86] demonstrated that the quality of an 18-dimensional sparse grid-based parameter estimation method improved when the number of model evaluations increased. Furthermore, the sparse grid approach outperformed a standard optimization method when the same number of model evaluations were considered for both. These early approaches tended to interpolate the cost function itself, but could not entirely avoid irregularities in the function that could degrade the quality of the interpolant. Later approaches [49, 50], including our work, interpolated the actual model dynamics, resulting in a far more accurate interpolant with fewer model evaluations. We make use of the Matlab-based Sparse Grid Interpolation Toolbox [68] for this work.

3.5 Two-Stage Algorithm

The two-stage algorithm searches for potentially optimal parameters on both the global and local scales. The local stage of the algorithm relies on the concept of local grids. Local grids were introduced in [49], and further explored in [50], to enables searches of local subspaces once the global search was exhausted. Interpolants constructed on local grids, when rendered sufficiently accurate, can improve upon the

results obtained from their global counterparts. However, the original concept was intended towards model-based experiment design for reducing uncertainty in model dynamics. Furthermore, local grids were originally intended to identify a sufficient number of parameters to satisfy a given criterion, not necessarily to determine which parameters best minimized the difference between simulated outcomes and observed data. Figure 3.1 illustrates the overall algorithm. In this work, we define a parameter to be either a point within, or a dimension of, the search space, depending on the context. We also define a parameter value to be a particular numerical value for a parameter. We describe each stage in the following.

3.5.1 Global Stage

Construct global interpolant

The global stage scans the entire parameter space for potentially optimal regions by constructing a global interpolant. We stipulate that the interpolant must possess a relative error less than 1%. This level of accuracy ensures that there is enough global confidence in the interpolated trajectories.

Sample global interpolant

We sample the global interpolant using Latin Hypercube Sampling (LHS) to obtain more comprehensive coverage of the global space. A simple, unweighted sum of squared errors cost function compares the interpolated trajectories and the available data. After the costs of all sampled parameters have been computed, we choose those parameters whose costs are below a model-dependent threshold to form the initial parameter set for the local stage.

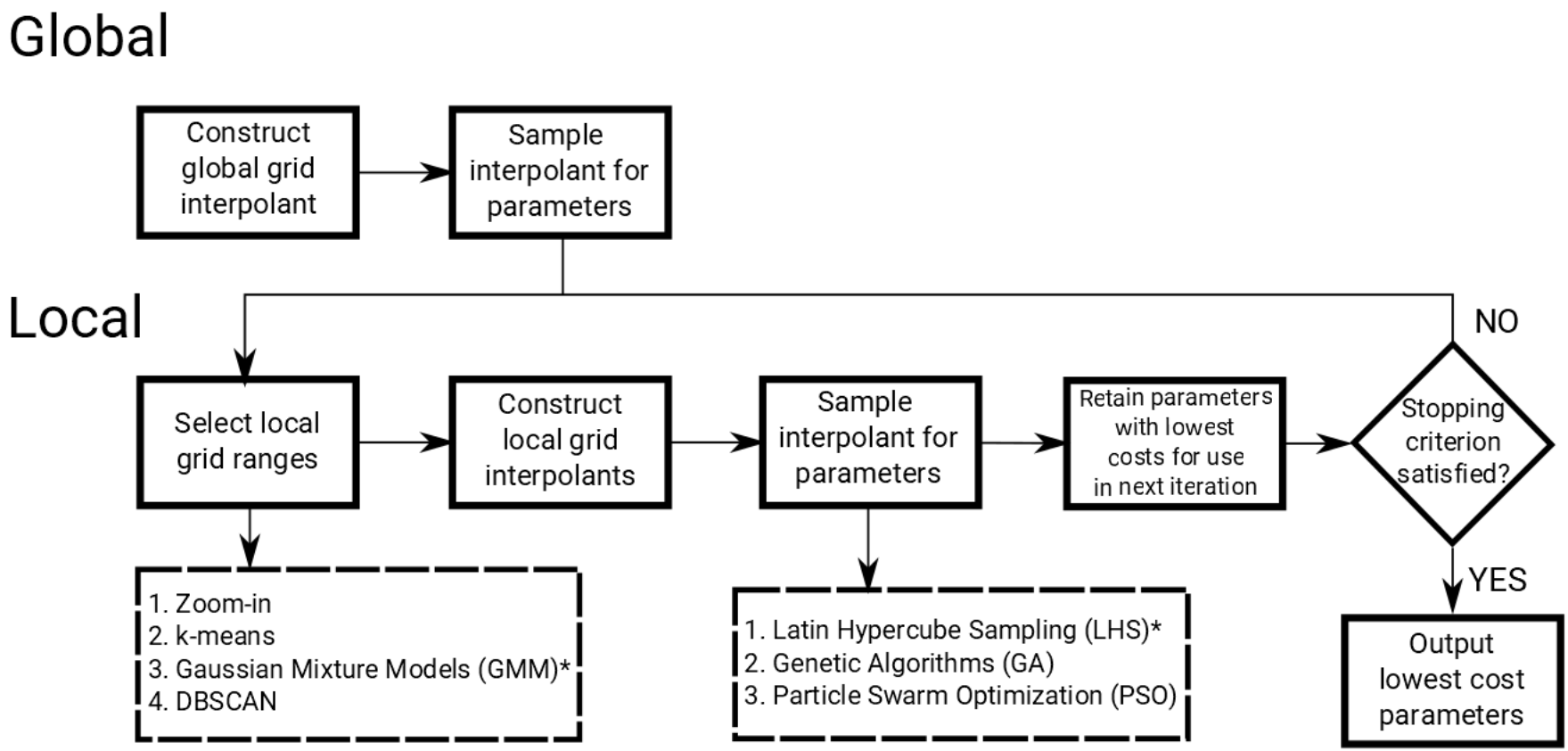


Fig. 3.1. Overall two-stage algorithm. A variety of methods are available for selecting local grid ranges and sampling the local interpolant. Asterisks indicate methods used in [49, 50].

3.5.2 Local Stage

Select local grid ranges

The local stage initiates with the incoming parameter set collected during the global stage. Here, we first decide how to construct the local grids. Specifically, we must choose how many local grids to create, and their respective ranges. In this work, we explore various methods that address these questions. The majority of these methods are based on some form of cluster analysis. Clustering methods, seen as an extension of multistart methods in the context of global optimization, can avoid the redundancy of detecting the same local minima repeatedly by isolating neighborhoods of local optima in order to conduct efficient, productive searches [105, 106].

In choosing clustering methods, we opt for methods that exhibit diversity in their clustering approach and appropriateness to the target model. Figure 3.2 displays each clustering method's approach towards arbitrary data distributions. Table 3.1 summarizes the methods chosen, along with how they select the number of clusters. k-means and Gaussian mixture models (GMMs) specify the number of clusters *a priori*, so we introduce objective functions for both methods to select the number of clusters. k-means clusters are determined by using an objective associated with the silhouette method, where clusters are well-separated and appropriately categorized. This objective function is defined as follows:

$$N_C^* = \arg\max_{N_C} \bar{\mathcal{S}}(N_C) \cdot \mathcal{S}_{min}(N_C) \tag{3.1}$$

where $\bar{\mathcal{S}}(N_C)$ is the average silhouette coefficient, and $\mathcal{S}_{min}(N_C)$ is the minimum silhouette coefficient, for all parameters in N_C clusters. The objective function penalizes negative silhouette coefficients, which suggest mis-clustering and lack of cohesion within a cluster, and seeks higher average silhouette coefficients, which indicate good separation.

In [49, 50], GMMs were used to cluster parameters, using a criterion based on minimizing the volume of overlap between the local grids, which we use here:

$$N_C^* = \arg\min_{N_C} \mathcal{V}(N_C) \tag{3.2}$$

where $\mathcal{V}(N_C)$ is the volume of overlap between N_C clusters.

DBSCAN is capable of recognizing clusters of arbitrary shape, while accounting for noise and outliers in the underlying data [107]. Moreover, it does not require the number of clusters to be specified by the user, as it relies on user-defined parameters, like the threshold distance and the minimum number of neighbors, to define clusters. For all clustering methods, we limit the maximum number of clusters that can be created to avoid creating too many local interpolants. Both the k-means and GMM clustering methods will incrementally increase N_C until this limit is reached to determine the optimal number of clusters.

We compare these clustering methods to an iterative magnification method, which we term zoom-in. Zoom-in is a greedy, divide-and-conquer approach that enlarges areas where previously optimal parameter estimates were found to locate better solutions. The method works by selecting the N_C parameters with the lowest costs, and computes a hyperrectangle around each parameter. The volume of each hyperrectangle is determined by extending the search range along each dimension by $\alpha\%$ of the parameter's value in both directions. Both N_C and α are defined by the user for the zoom-in algorithm. We explore the tuning of these parameters in Section 3.6.

Once the local grid ranges have been specified, the local interpolants are then created, with more stringent accuracy requirements than the global grid. We impose a limit on the relative and absolute errors of the local interpolants to $10^{-3}\%$ and 10, respectively. Because the goal of the algorithm is parameter estimation, attention must be paid to the overall accuracy of the grid so that outrageously unrealistic cost estimates are avoided and the interpolated dynamics serve as a reasonable approximation to the actual dynamics.

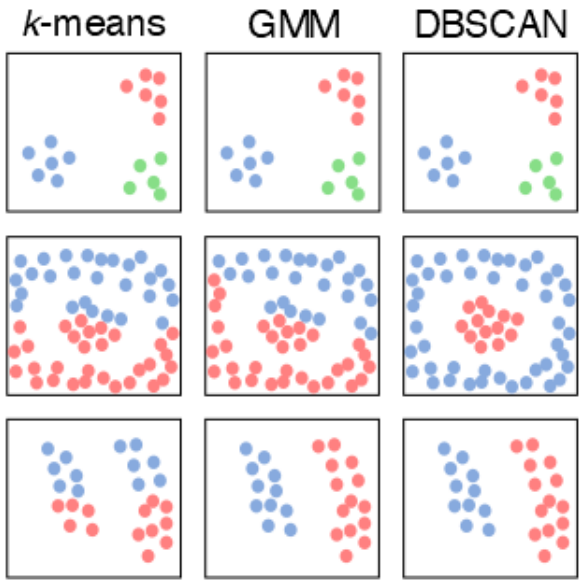


Fig. 3.2. Graphical depiction of each cluster analysis method on different data distributions.

Sample local interpolant

The two-stage algorithm then samples parameters from each local interpolant. The interpolant generates approximated trajectories for each parameter, which can then be compared with observed data to compute a cost. Those parameters with the lowest costs are retained, where they can be used in future iterations of the algorithm. Here, we apply metaheuristic algorithms, which continuously navigate the search space in order to determine near-optimal solutions in a reasonable amount of time. We choose population-based metaheuristic algorithms because the cost of computing population fitness compared to individual fitness is negligible and the entire population can be updated simultaneously. These metaheuristic algorithms are then compared to LHS. Briefly, we detail these metaheuristics:

1. **Genetic Algorithms (GAs)**

   GAs provide a stochastic heuristic solution to global optimization by relying on evolution-based concepts such as crossover and mutation to produce new and

Table 3.1.
Description of clustering methods used, and how they are deployed by two-stage algorithm.

Clustering Method	Description	Number of Clusters
k-means	Centroid-based method that partitions points based on distance to k cluster means	Maximizing separation and cohesion of clusters
Gaussian Mixture Models (GMM)	Model-based method that assigns each point soft membership to a cluster defined by a Gaussian distribution	Minimizing volume of overlap between clusters
Density-based Spatial Clustering of Applications with Noise (DBSCAN)	Density-based method that groups points according to compactness and proximity to neighboring points	Selected internally with no user input

improved candidate solutions [108, 109]. We adapt a GA for parameter estimation from [96], with a population undergoing selection, migration, crossover, and mutation. The GA is implemented as follows:

(a) A preliminary set of parameters is generated using LHS.

(b) The cost of each parameter in the population is calculated. Those parameters with the lowest costs are retained.

(c) A new group of parameters are introduced by migration, where LHS produces more random samples.

(d) Crossover is initiated, where the existing parameters contribute randomly to spawn a new set of parameters.

(e) Mutation affects a certain percentage of the parameters, replacing certain values with new ones randomly.

(f) The process either returns to step (b) if the requisite number of iterations have not been completed, or terminates.

2. **Particle Swarm Optimization (PSO)**

PSO, a swarm intelligence algorithm, tries to improve the quality of candidate solutions by deploying a population of particles to move throughout the search space [110]. Their movement is dictated by simple mathematical formulae of physical concepts like position and velocity. The velocity of each particle dictates the rate at which each particle traverses the search space. Each particle is influenced by the best positions that it (personal best) and the entire swarm (global best) have attained thus far. By integrating this knowledge iteratively, the swarm is eventually driven towards the best solution. PSO is implemented as follows:

(a) The swarm is initialized with random position and velocity vectors within the parameter space.

(b) The costs associated with the particles' positions are evaluated.

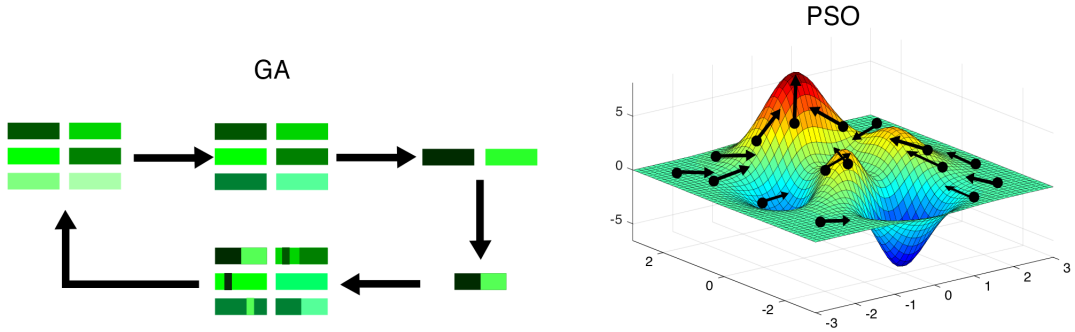
(c) The global and personal best positions of the swarm are revised. The global best positions and costs of the current iteration are retained.

(d) The position and velocity vectors are subsequently updated, with consideration for the global and personal best positions of the swarm.

(e) If the number of iterations has reached its maximum, the algorithm is terminated. Otherwise, the process restarts at step (b).

Once the local interpolants have been completely sampled, the remaining parameter set, which represents parameters with the lowest anticipated costs, is sorted and

filtered to retain a certain number of parameters for the next iteration of the local stage. The two-stage algorithm reports the lowest cost of all parameters found in the current iteration.



(a) Genetic algorithms. Prospective candidate vectors undergo selection, migration, crossover and mutation within the parameter space.

(b) Particle swarm optimization. Particles update their positions and velocities based on personal and global knowledge.

Fig. 3.3. Metaheuristic algorithms used in this work.

Stopping Criteria

Once the optimal parameters for the current iteration have been found, as detailed in Section 3.5.2, the process repeats. The current parameter set is passed to the initial step of the local stage, discussed in Section 3.5.2. We devised two stopping criteria, which upon satisfying either one, the two-stage algorithm will terminate:

1. The number of overall iterations.

2. When no change in the minimum cost was observed after consecutive iterations.

3.6 Numerical Studies

We conduct two numerical studies examining variants of the proposed two-stage algorithm. The first numerical study compares the methods described in Section 3.5.2

for selection of local grid ranges, while the second considers the sampling strategies for the local interpolant detailed in Section 3.5.2. We analyze the selection and sampling steps separately to evaluate its individual impact on parameter estimation.

3.6.1 Influenza Model: Selecting Local Grid Ranges

Our first model is a stochastic reaction network (SRN) of an influenza outbreak that occurred at an English boarding school in 1978, a well-recorded episode in the medical literature [111]. The epidemiological system, described as the classic SIR model in SRN form, is as follows:

$$S + I \xrightarrow{k_1} 2I \tag{3.3}$$

$$I \xrightarrow{k_2} R. \tag{3.4}$$

Model variables and parameters are listed in Table 3.2.

While the SIR model has commonly been simulated deterministically using ODEs, it may not be entirely valid in this case. The continuous variables within the ODEs are an ensemble average of their stochastic, discrete integer-valued counterparts over many replications. An epidemiological system comprises several discrete-valued processes, where a positive integer number of infected individuals must make contact to propagate the disease. Stochastic models are appropriate when both the population size and the number of infected individuals are small [112, 113], as is the case here. Stochastic models also permit the possibility of an epidemic-free state [28]. Therefore, we opt for the SRN representation of the SIR model.

The SRN consists of an expansive state space composed of all possible transitions between individuals in various epidemiological states, which are modeled as multivariate Markovian population processes. To efficiently compute the probability mass function of the population process, [114] simulated the SRN numerically using a novel implementation of the implicit Euler method, which relied on the degree of advancement (DA), a stochastic counting process that tracked the number of occurrences of

Table 3.2.
Model variables and parameters of influenza model along with feasible ranges. Values for state variables indicate initial conditions, described in [111].

State variables	Definition	Initial Condition
S	Susceptible individuals	762 people
I	Infectious individuals	1 person
R	Recovered individuals	0 people
Parameters varied		Range
k_1	Infection rate	$2.18 \times 10^{-4} - 2.18 \times 10^{-2}$ days^{-1}
k_2	Recovery rate	$0.044036 - 4.4036$ days^{-1}

every reaction within the system. By exploiting the DA process and its finite sample space, determining the populations of each epidemiological state is effectively distilled to recursively evaluating an implicit ODE of the probability mass function of the DA process.

A potential bottleneck for parameter estimation of this SRN from a computational efficiency standpoint is the number of computations to perform with respect to the size of the matrix to be inverted at each time step, the generator matrix. The dimension of the generator matrix reflects the number of distinct states in the sample space of the DA process and approximately scales as Q^2, where $Q = (S(0)+1)(S(0)+I(0)+1)$. The generator matrix for this model therefore contains $(763 \times 764)^2 \approx 3.40 \times 10^{11}$ elements. Simulating the SRN across time with a small time step for every possible parameter value would be time-consuming. Fortunately, the resulting trajectories are reasonably smooth to deem an interpolation approach appealing in the broader context of parameter estimation. Therefore, we embed the implicit Euler method within the sparse grid interpolation framework to evaluate the model where it needs to and interpolate trajectories where it doesn't. We interpolate the mean number of infected persons predicted by the model. While tensor-based approaches to parameter

estimation of SRNs [115], and sparse grid methods for approximating the underlying chemical master equation [116] have been studied, we believe that our approach best combines the advantages of a numerically sound solver with a proven, high-fidelity approximation model for this particular problem.

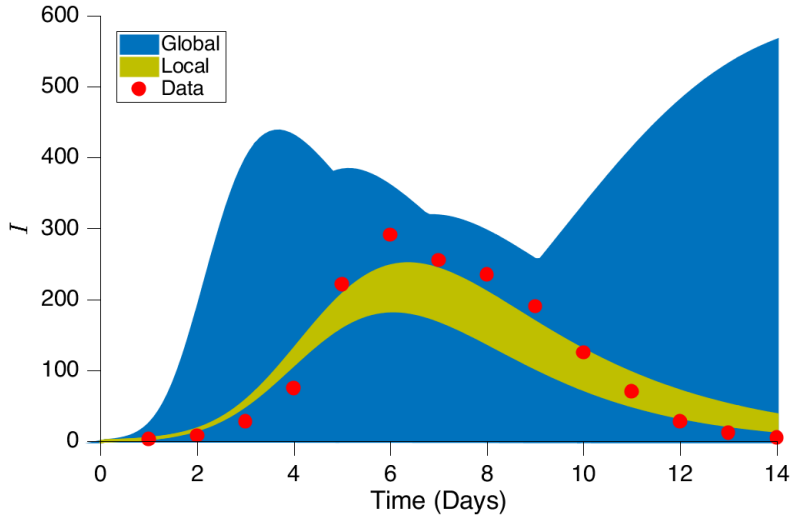


Fig. 3.4. Simulations of the number of cases for the influenza model against actual data (red dots). Blue (gold) trajectories obtained by simulating parameters obtained from the global (local) stage.

We first assess the utility of cluster analysis on selecting local grid ranges for the influenza model. The 2-dimensional global sparse grid interpolant required 2,177 model evaluations, yielding a relative error of 0.29%. The interpolant identified 70 parameters with costs less than the threshold of 3×10^5, with the corresponding dynamics illustrated in blue in Figure 3.4. It is clear that these trajectories cover a dynamically diverse range, overlapping with the actual data. The minimum cost found in the global stage was 1.15×10^5.

Figure 3.4 also shows the dynamics obtained from the local stage, highlighted in gold. Interestingly enough, not only do all clustering methods outperform the zoom-in method, but they also converge on virtually the same parameter values. Furthermore, they complete their search in fewer iterations, as they find no further improvement

after three iterations (Figure 3.5). At the end of the local stage, the zoom-in and clustering methods improve by 49% and 0.5% respectively. However, this understates the improvement of the clustering methods from the minimum cost found in the global stage. While the zoom-in method improved by 63% from the global minimum cost, the other methods outperform this cost by an astounding 90%, producing parameter values with a cost of 1.26×10^4. In general, the clustering methods saw an improvement in the minimum cost of over 70% compared to zoom-in. Moreover, the number of model evaluations needed when the clustering methods are used are at least 20% lower than that of the zoom-in method, as seen in Table 3.3. DBSCAN was the best in terms of minimizing both computational effort and the deviation of model from data. An additional comparison was made to the constrained optimization solver *fmincon*, with two different algorithms, sequential quadratic programming (SQP) and the interior-point algorithm. Both variants of *fmincon* identified the same minima as the clustering methods, but with significantly fewer SRN model evaluations when started from the best parameter obtained in the global stage. This lends more confidence to the obtained minima, and the ability of the clustering methods to identify it, albeit with more model evaluations.

In Figure 3.6, we show how the different methods perform in selecting the local grid ranges on their first iteration. While the clustering methods are able to partition a wider space into successively distinct subspaces, the zoom-in method magnifies the region around the initial optimal estimate, moving relatively little across all five iterations. This is also reflected on the final parameter values found. The clustering methods and the *fmincon* methods settle on $[0.0023, 0.3431]$, while the zoom-in method settles at $[0.0024, 0.4748]$. We also note that the basic reproductive number, the average number of infections caused by an infected individual in an entirely susceptible population [10], was 0.0051 and 0.0067, based on the parameters determined by the zoom-in and clustering methods respectively. This corresponds with the observation that the outbreak quickly surged and abated over the course of two weeks.

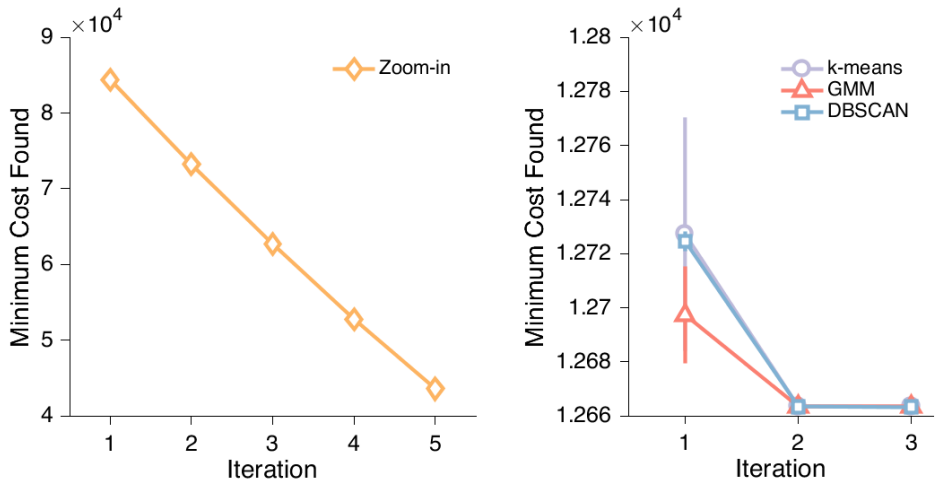


Fig. 3.5. Minimum costs found through several iterations of local stage of the two-stage algorithm. Left figure shows the zoom-in method, with $\alpha = 5\%$, $N_C = 2$.

Table 3.3.
Number of model evaluations taken for the zoom-in method and the clustering methods in Figure 3.5. * indicates results obtained from using Matlab's *fmincon* when starting from the global best parameter.

Method	Number of Model Evaluations (mean ± SD)	Improvement over zoom-in (%)	Minimum Cost Found (mean ± SD)	Improvement over zoom-in (%)
Zoom-in	1290	-	43592.5 ± 2.46	-
k-means	883.5 ± 82.84	32	12663.58 ± 0.56	71
GMM	979 ± 63.47	24	12663.41 ± 0.03	71
DBSCAN	718 ± 13.86	44	12663.4 ± 10^{-7}	71
SQP*	180	87	12663.4	71
Interior-point*	48	96	12663.4	71

3.6.2 Cholera Model: Sampling Local Interpolants

The second model describes population dynamics during a cholera outbreak. It was originally used to analyze a 2008-2009 epidemic of the water-borne disease in Zim-

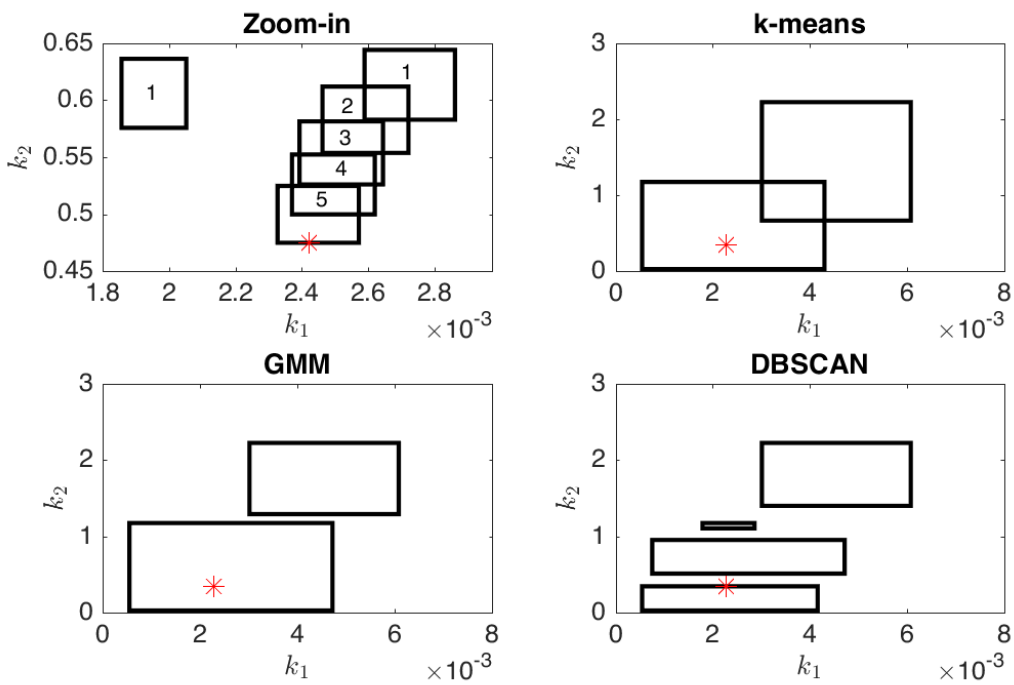


Fig. 3.6. Performance of different clustering methods in dividing the parameter space. Numbers for zoom-in method indicate iteration. All other clustering methods show clusters formed in the first iteration only.

babwe [117]. The ordinary differential equation (ODE)-based model considered both human-to-human and environment-to-human transmission pathways. The ODEs are as follows:

$$\frac{dS}{dt} = \mu N - \beta_e S \frac{B}{\kappa + B} - \beta_h SI - \mu S \tag{3.5}$$

$$\frac{dI}{dt} = \beta_e S \frac{B}{\kappa + B} + \beta_h SI - (\gamma + \mu) I \tag{3.6}$$

$$\frac{dR}{dt} = \gamma I - \mu R \tag{3.7}$$

$$\frac{dB}{dt} = \xi I - \delta B. \tag{3.8}$$

The model notation and parameter values are summarized in Table 3.4. We consider a 9-dimensional parameter space, whose ranges are a subspace of those suggested

by [118, 119]. We narrowed the initial parameter range until we found a suitable starting point for the two-stage algorithm. Data from the current outbreak in Yemen, which reports the cumulative number of cases reported in the first few months of the 2017 epidemic [3], is compared with a variable representing the cumulative number of cases, described as follows:

$$\frac{dC}{dt} = \beta_e S \frac{B}{\kappa + B} + \beta_h SI. \tag{3.9}$$

Our usage of this model is motivated by a variety of reasons. Efforts to model the Yemen cholera epidemic thus far have relied on statistical models that forecast the growth of the outbreak [120]. We rely on an established mathematical model that can characterize existing trends and attribute transmission to multiple pathways. Unlike the influenza model, the cholera model uses data from a current outbreak. There is a significant amount of parameter uncertainty surrounding new and evolving outbreaks, which translates to more uncertain parameters and larger search ranges. Our two-stage algorithm can easily accommodate these needs. Moreover, as the initial conditions for S, I, and B have not been clearly specified, we include them as parameters in the overall parameter space for estimation.

We apply the two-stage algorithm with an emphasis on testing the different sampling approaches discussed in Section 3.5.2 on the cholera model. The 9-dimensional global sparse grid interpolant required 1,919 ODE model evaluations with an estimated relative error of 0.49%. The interpolant identified 124 parameters with costs less than the specified threshold of 10^5. Dynamics corresponding to these parameters are illustrated in blue in Figure 3.7. There are varying degrees of qualitative fit to the data, with two distinct qualitative trends. One set of trajectories appears to rise slowly before plateauing, while another seems to exponentially increase. The minimum cost found in the global stage was 4.139×10^3. These 124 parameters are then passed into the local stage.

Dynamical results for the local stage are also presented in Figure 3.7 in gold. These trajectories follow the observed data fairly closely, showing a uniform qualita-

Table 3.4.
Model variables and parameters of cholera model along with feasible ranges. Value of R indicates initial condition $R(0)$. All state variables denote individuals in thousands.

State variables	Definition	Range
S	Susceptible individuals	$2.5 \times 10^4 - 2.9 \times 10^4$ people
I	Infectious individuals	0-3 people
R	Recovered individuals	0 people
B	*V. cholerae* concentration	0-10^6 cells/ml
Parameters varied		
μ	Birth/death rate	10^{-5}-10^{-4} days^{-1}
γ	Recovery rate	0-10 days
ξ	Bacterial contamination rate	0-10 cells/ml/day/person
δ	Bacterial death rate	3-41 days
β_e	Environmental contact rate	0-0.1 days^{-1}
β_h	Human contact rate	10^{-8}-10^{-7} people^{-1}days^{-1}
N	Total population	$2.5 \times 10^4 - 2.9 \times 10^4$ people

tive pattern. The individual iterations of the local stage are shown in Figure 3.8a, where $\alpha = 5\%$, and $N_C = 2$. Based on the figure, the standard LHS, GA and PSO implementations show respective decreases of 12, 10, and 13% across the five iterations. More importantly, all three methods improved on the costs obtained from the global stage by at least 40%. The LHS and GA runs appear to overlap for much of the iterations, with a slight edge for GA. In terms of computational burden, Table 3.5 indicates that the genetic algorithm implementation took fewer model evaluations on average. This was due to at least one replication of the algorithm stopping before the limit on the maximum number of iterations of the local stage was reached, thus preventing further local interpolants from being created unnecessarily. Again, we perform a comparison to *fmincon* and the results demonstrate that based on evaluations

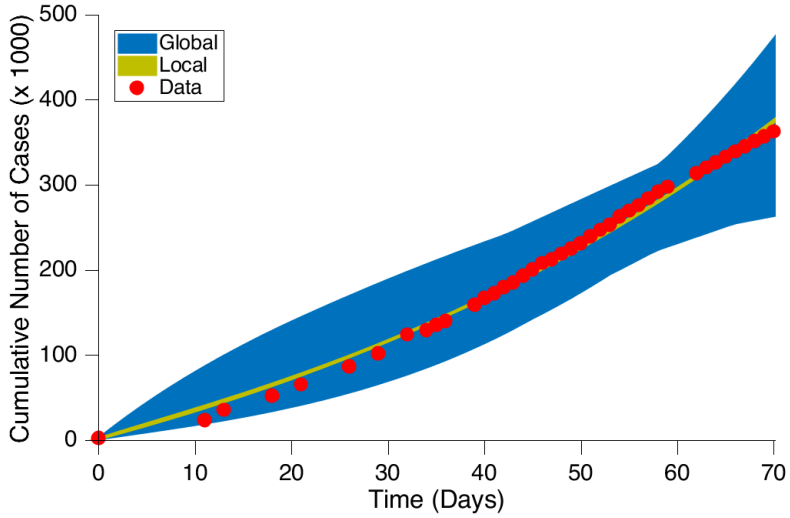


Fig. 3.7. Simulations of the number of cases for the cholera model against actual data (red dots). Blue (gold) trajectories obtained by simulating parameters obtained from the global (local) stage.

at the local stage, there is actually diminished performance by *fmincon* in terms of converging to an improved solution. *fmincon* identified different minima that did not match those of metaheuristic algorithms, and actually performed worse than LHS, but required fewer ODE model evaluations overall. *fmincon* terminated when the step size dropped below 10^{-12}. Like the influenza model, we compute the basic reproductive number for the various methods. The estimates of $\mathcal{R}_0$ range from 1.5304 to 2.4802, suggesting a continuation of infection that accurately reflects reality.

We implement the two metaheuristic algorithms assuming that the local subspaces were selected using the zoom-in method. In addition to doing a straightforward application of the metaheuristic algorithms towards sampling of the local interpolants, we also explored how tuning the parameters of the zoom-in method, namely the range of enlargement around a parameter value α, and the number of local grids to construct N_C, would alter the results. Figure 3.8 show the results of tuning the zoom-in algorithm for α and N_C, respectively. Each metaheuristic shows a different preference for these parameters. GA improves when α was increased, possibly because

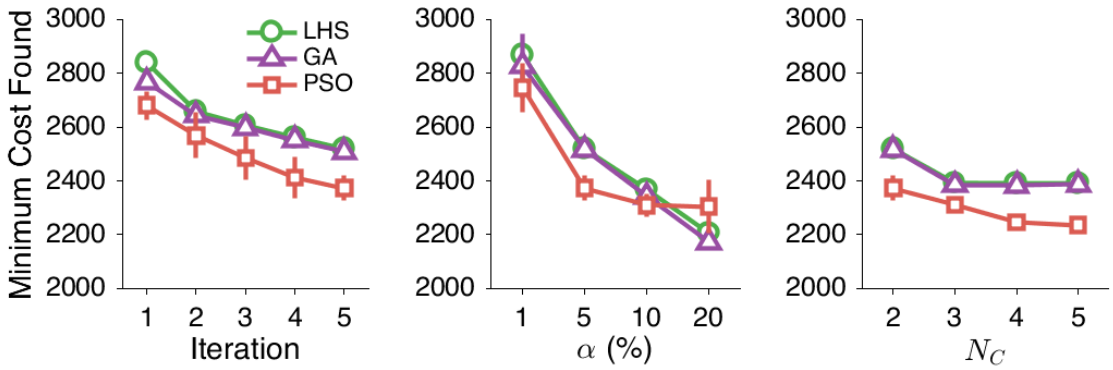


Fig. 3.8. Performance of the local stage of the two-stage algorithm for various metaheuristic algorithms on the cholera model, when tuning for user-defined parameters N_C and α.

an increase in the surrounding area of an optimal parameter estimate permitted more variation in the population, and therefore enabled the pertinent genetic operators (selection, migration, crossover) to produce better candidate solutions. In fact, the best fit between the cholera model and the data on hand is achieved when $\alpha = 20\%$, as the GA reached a minimum cost of 2.172×10^3. On the other hand, increasing N_C seemed to improve the performance of PSO, as it reached its best value at 2.235×10^3 when $N_C = 5$. This may be due to the exploratory nature of PSO, which could navigate and locate optimal values in the parameter space better than LHS and GA when given more opportunities to do so.

3.7 Discussion

This work has analyzed the various options for selecting and sampling local subspaces for two-stage parameter estimation. We examined cluster analysis and metaheuristic algorithms as suitable components of an algorithm to repetitively narrow a large search space in the hopes of determining progressively better fits to observed data. As our results demonstrate, certain gains in parameter estimation can be expected when deploying these methods.

Table 3.5.
Number of model evaluations taken for the LHS benchmark and the metaheuristic algorithms in Figure 3.8a. * indicates results obtained from using Matlab's *fmincon* when starting from the global best parameter.

Method	Number of Model Evaluations (mean ± SD)	Improvement over LHS (%)	Minimum Cost Found (mean ± SD)	Improvement over LHS (%)
LHS	10083.6 ± 8.47	-	2519.26 ± 4.87	-
GA	9683.2 ± 1259.72	4	2508.98 ± 3.96	0.4
PSO	10137 ± 102.61	-0.5	2373 ± 47.34	6
SQP*	337	97	3226.32	-28
Interior-point*	22	97	3323.89	-32

In terms of local grid selection, the remarkable contrast in performance between the clustering algorithms and the zoom-in method reflects the importance of successfully segregating and scrutinizing regions when informed with prior parameter information. What the zoom-in method lacks, and the clustering methods ultimately capitalize on, is the delicate balance between exploration and exploitation of the parameter space. By over-relying on exploitation of its current position, the zoom-in method insufficiently explores the parameter space, settling for minor gains without a broader view of the search space. In fact, the zoom-in method continually improves, albeit incrementally. Given sufficient iterations, it may eventually settle into the same minima as the clustering methods. The clustering methods switch from a rapid exploration to deep exploitation strategy, and are even able to terminate early once the largest improvement has been made in the first two iterations. The fact that all three clustering methods, while being based on completely different concepts of clustering, converged to the same neighborhood and minima, is a testament to the benefits of employing cluster analysis as a tool to local grid selection. Our cluster

selection criteria for each clustering method enable optimal coverage of the parameter space previously inhabited by parameters obtained in the global stage.

However, the use of cluster analysis has an important caveat. The feasibility of creating local interpolants comes into question when a local grid range delineated by a cluster becomes too large. The resulting local interpolant can become poor in quality, a problem exacerbated by the dimensionality of the parameter space. While we experienced no such obstacle in a low-dimensional parameter space, it is a potential hurdle that will have to be overcome by designing cluster selection criteria that is more stringent than what we proposed, and may produce more clusters with smaller ranges. However, an increase in granularity will inevitably lead to an increase in computational effort.

On the other hand, comparing sampling approaches reveals a more ambiguous picture. GA and PSO offer modest, if not negligible, improvements over a standard LHS approach. It would appear that much of the effort needed to perform effective parameter estimation rests on identifying appropriate local subspaces rather than sampling them. There are limitations to what GA and PSO can do once a particular parameter range has been established. It is also somewhat expected that both the GA and PSO approaches possessed more variability between replications. Metaheuristic algorithms are prone to uncertainty in how the solution space is explored with each run. Our results show that while both methods are superior to LHS, they pale in comparison to the clustering methods. Moreover, the randomness and approximate nature of these variants calls into question whether such results would be expected in other models. However, these results can be partially abrogated by the fact that the sampling approaches are layered on top of the zoom-in method. So, within the context of sampling local interpolants, there are alternatives to a conventional random sampling method. The GA implementation took fewer iterations than the other approaches, resulting in fewer model evaluations on average. The PSO approach conversely required more model evaluations than even the LHS approach to deliver a quantitatively better fit. Therefore, given a decision between these two alternatives

to LHS, one has the option to either prioritize fewer model evaluations (GA), or a quantitatively better fit to data (PSO).

Through modifying the zoom-in method for possible improvements in sampling local interpolants, we also observe some interesting trends. The algorithm parameters N_C and α modulate the breadth and depth capabilities of the zoom-in method, respectively. While increasing N_C may be an effective way of improving on the minimum cost, we caution that increasing α may not necessarily do so. At large values of α, the resulting local interpolant may no longer be able to accurately characterize all the potential dynamics within its multi-dimensional space. This may lead to interpolated dynamics and costs that may conflict with reality, squandering computational resources towards incorrect parameter estimates.

3.8 Conclusion

We have demonstrated the viability of epidemiological model calibration using a sparse grid-based two-stage algorithm. We have evaluated the impact of identifying local subspaces via cluster analysis and sampling them using metaheuristic algorithms, both of which can independently improve the quality of parameter estimation with less computational burden. We recommend that both options be considered when investigating models for possible fits. The clustering methods partition the search space effectively to probe disparate regions, while the metaheuristic algorithms empirically test combinations of parameter values to search intelligently through the multi-dimensional space. Based on the results presented, we suggest that cluster analysis offers more advantages to quickly determining promising regions, but the value of exploiting those regions with metaheuristic techniques are not to be discounted in certain situations.

While sparse grid interpolation permits global interpolation of the model dynamics, we cannot claim for certain that the results obtained from our parameter estimation strategy are truly optimal, except that based on our ensemble of simulations,

we found no alternative optima. The limitations in resolving all dynamics over large parameter ranges or dimensions prevents a comprehensive analysis. However, we are confident that the large gains in matching simulation outcomes with existing data exhibited by our computational experiment presents a compelling parameter estimation procedure for similar models. Issues of parameter identifiability, where parameters are unable to be estimated uniquely, are often attributed to defects in model structure or data insufficiency. If left unresolved, these inaccurate parameter estimates may prevent public health researchers from appropriately characterizing the current epidemiological situation, hindering successful intervention strategies in the process. Resolving identifiability issues may require sensitivity analyses to determine the most important model parameters on model outputs [121–123]. We also acknowledge limitations in implementing genetic algorithms and particle swarm optimization with the appropriate hyperparameters (i.e., the number of iterations and the number of candidate solutions available at each iteration). We selected these hyperparameters with the runtime of the algorithm in mind, but adaptively adjusting these hyperparameters as the algorithm runs may be worthwhile.

The algorithm presented here can easily be extended to performing parameter estimation on more complicated models, such as metapopulation models, whose parameters are more numerous yet spatially refined. Fitting parameters for larger models will provide more detailed information on the local status of a disease and facilitate computation of regional epidemiological metrics, like the basic reproduction number. Understanding the epidemic on the local level will enable researchers to craft interventions better tailored to the afflicted population, saving lives and halting the spread of devastating diseases.

4. OPTIMAL MULTI-PERIOD POINT OF CARE SENSOR SELECTION FOR CHOLERA MODELING AND CONTROL

4.1 Preface

The research described in this chapter is in preparation for submission to *Operations Research*.

4.2 Abstract

Epidemics present enormous resource allocation problems for unsuspecting populations. On top of that, information related to the state of the epidemic often arrives at erratic bursts, laden with reporting errors and time delays. Cholera, a water-borne bacterial disease, is no exception to these challenges. We consider an adaptive, multi-scenario model predictive control algorithm to regulate the infected population using a set of accessible real-world interventions, given partial information at select times. We also deploy a sensor selection scheme that embeds within the control framework to select future sensor configurations for bacterial concentration measurements based on projected model dynamics. Sparse grid interpolation is employed to produce future model dynamics as a function of data-consistent model parameters and admissible control signals. A comparative study of sensor selection criteria is conducted to highlight the societal and economic benefits of jointly monitoring infected individuals and quantifying bacterial concentration uncertainty.

4.3 Introduction

4.3.1 Cholera Modeling

Cholera is an acute water-borne, diarrheal disease caused by the bacterium *V. cholerae*. Ingesting food or water contaminated with the bacterium can cause excessive diarrhea and vomiting that, if left untreated, can lead to severe dehydration and death [124–127]. Worldwide, cholera is responsible for 1.3-4 million cases, resulting in 21-143,000 deaths, establishing endemicity in 69 countries, mainly in Africa and South Asia [128]. The ability to sense and detect the presence of cholera in particular locations confers an advantage in mounting specific public health policies. Since the bacteria cannot be sensed directly, surrogate quantities must be measured to ascertain bacterial distributions. Satellite imaging has been used to measure observable proxies for cholera outbreaks in coastal regions, including sea surface temperature and height, chlorophyll A levels, precipitation, air temperature, local climate phenomena, plankton biomass, sunlight, sea salinity, vegetation and soil content [129]. [130] demonstrated a link between cholera outbreaks and sea surface temperature and height, postulating that plankton blooms at warmer temperatures facilitate the proliferation of *V. cholerae* and emit chlorophyll, which can then be measured spatiotemporally by satellites. The growth of bacteria within coastal regions would then extend into inland waters via networked waterways as a result of increased sea surface height, multiplying the opportunities for unmitigated contact with potential hosts through water consumption. Early warning systems for cholera outbreaks that incorporate these forms of ancillary data have also been proposed to provide as much lead time as possible to warn potential target communities of impending outbreaks [131, 132]. A sensitivity analysis using remotely sensed data sets of the aforementioned environmental parameters in the Lake Kivu region of the Democratic Republic of the Congo has suggested that seasonality and local climate phenomena contribute significantly to endemic cholera incidence compared to other factors [133].

Modeling cholera incidence through mathematical modeling provides public health policy makers guidance on how to effectively target the disease with available interventions. Cholera models integrate host-pathogen dynamics through direct human-to-human and indirect environment-to-human transmission pathways. While transmission occurs almost exclusively via contaminated water or food, human-to-human transmission occurs on a relatively quicker time scale with a lower infectious dose [117]. Epidemic models of cholera assume a single bacterial strain, an entirely susceptible population, and a short time scale that omits climate and bacteriophage dynamics [119]. The probability of infection is dose-dependent, with the bacterial concentration modeled as reaching a saturation point [134]. This observation is reflected in most mechanistic cholera models, which are often expressed as differential equations representing the known disease states.

Many cholera models fail to account for any spatially explicit properties of different geographic areas dealing with the same cholera outbreak. They assume that parameters derived from accumulated, national data can be applied homogeneously to subnational regions. The implications of such an approach may fail to adequately account for the particular dynamics taking place in the disaggregated subunits [119,135]. Metapopulation models examine cholera dynamics with a spatially refined lens, often based on administrative [136], or geographic boundaries [137]. The dynamics between these spatially distinct areas, or patches, is of interest as well. With waterways connecting these patches, bacteria shed from one population can spread to others with ease [136,138–140]; hydrological connections can also be modeled to reflect downstream movement and infection by pathogens [137,141]. Asymmetric population movements between the patches may redistribute the pathogen unevenly. The overlay of human mobility and hydrological networks can add a higher degree of realism to spatiotemporal cholera models.

The primary interventions studied by cholera models include: vaccination, treatment, sanitation, hygiene, awareness, chlorination, and quarantine. Oral cholera vaccines have been effective in curtailing the spread of the disease where it is known

to have spread [127, 142]. Treatment with antibiotics is recommended for severe cases [127]. Disposing of human waste properly by modern sanitation standards and enforcing common hygiene practice in cholera can also alleviate the disease burden. Awareness in the form of educational campaigns and spreading of rumors has also been advanced by certain models [143–145]. Finally, chlorinating the water supply and administering quarantine practices are rarer options that have been explored by fewer models [144, 146]. Metapopulation models have also been the subject of control studies, in which different populations experience varying levels of intervention specific to their circumstances [144, 147–150]. However, one prominent drawback of these models and their control strategies has been that the entire control trajectory is available from the start of the simulation, with no consideration for new data that could alter the decision landscape. Furthermore, there is no consideration among the existing cholera control literature for the utility and availability of cholera pathogen sensors that may provide additional insight into pathogen dynamics and assist in developing more effective intervention strategies.

4.3.2 Model Predictive Control (MPC)

MPC is an iterative optimization-based control technique, where a stabilizing feedback control is designed to satisfy a performance criterion subject to state and control constraints [151]. It is predicated on repeated, online use of a dynamical prediction model. The prediction model mathematically approximates the true system, or *plant*, and provides forecasts of future system behavior over some time interval, known as the *prediction horizon*. At the beginning of each iteration, new, possibly infrequent, plant measurements inform the prediction model as to its current state. The controller samples possible trajectories over the prediction horizon, which originate from the current state and are generated by candidate control sequences, in order to solve a constrained optimization problem in the current time interval. The performance index, or objective function, being optimized favors control sequences

that best minimize the deviation between model outputs and desired trajectories with the least amount of effort, respecting restrictions on both outputs and controls in the process. After the optimal control sequence is selected, the elements of the sequence corresponding to the immediate future are then used to update the plant, as the prediction horizon moves to the next time interval and the procedure is repeated.

Sparse grid interpolation has been previously applied to MPC to estimate future model dynamics over the prediction horizon as a function of the current measurement state and admissible control signals [152, 153]. A sparse grid-based adaptive model predictive control method was applied to control the differentiation of a cancer cell line by formulating a multiscenario adaptive MPC approach, wherein model parameters consistent with available data were jointly considered for the optimization problem [152]. The optimal control sequence was determined by employing a constrained nonlinear optimization solver in Matlab. Sparse grids were specifically used to construct input and parameter domain interpolants that predicted future model output dynamics over the prediction horizon. The resulting interpolants were used in solving the optimization problem. [153] expanded on this approach by populating a pool of candidate models to inform the selection of a weighted, consensus-based control sequence, involving multiobjective optimization to identify Pareto-optimal solutions for control in T-cell signaling pathways. A main difference between prior sparse grid-based MPC approaches and our proposed method is the domain of application. We focus our efforts in epidemiology, a field that MPC lends itself well to, especially considering issues such as data insufficiency in times of outbreaks, and predictive forecasting of interventions.

4.3.3 Sensor Selection

Sensor selection primarily refers to the problem of determining an optimal sensor configuration to guarantee some prescribed performance related to estimating the target environment. Mathematically, sensor selection is a difficult combinatorial problem

(i.e., exactly $\binom{M}{L}$ possibilities of choosing L distinct sensors out of M available ones). Sensor systems typically operate under resource constraints that prevent concurrent resource use at all times. Sensor selection schemes dynamically select subsets of available sensors to use at each time during a measurement period in order to optimize some performance metric and minimize inevitable information loss. The performance metric quantifies a system requirement, such as information redundancy, energy efficiency, estimation accuracy or detection probability. Time is usually partitioned into a series of decision epochs with L sensors chosen in each epoch. In traditional feedback control, sensors determine the state of the plant through periodic measurements; these measurements inform the controller to execute certain control policies, which in turn influences the plant [154]. Once a sensor is selected and the corresponding measurement is obtained, information relevant to the performance metric is extracted from sensor data. This information must substantiate the merit of each potential sensor configuration in the next decision epoch, either statistically or heuristically.

Various functions of the Fisher Information matrix (FIM) have been used as objective functions for sensor selection [155–158]. However, using the FIM requires initial parameter estimates for the sensor measurement model, and the resulting computation provides only a local measure of the information value [154]. There is a close connection between sensor selection and the D-optimal experiment design problem, in that both attempt to choose a subset of possible measurements from available choices [155]. Traditionally, sensor selection schemes employ a reactive selection policy, wherein future sensor configurations are chosen based on prior sensor measurements. Predictively quantifying the information value of a particular sensing action before it is taken is difficult. In this work, we incorporate predictive sensor selection into our control implementation. Sensor selection is used to select certain locations for future pathogen sensing. Having additional data in the form of sensor observations of pathogen concentrations constrains the space of plausible epidemiological explanations so they better reflect the real-world situation. With pathogen sensing, we

achieve more clarity in a disease model, enabling identification of effective infectious disease interventions.

4.3.4 Data Assimilation

Tracking and predicting the full evolution of a new outbreak is notoriously challenging. The model may contain numerous inaccuracies, and observational data may be incomplete and irregular. Data assimilation uses the latest available observations and knowledge of error associated with observations to create new sets of forecasts and estimate the current state of the population and the epidemic. It has been applied to model and track numerous emerging epidemics [89, 159–165]. Using data assimilation methods can increase the accuracy, reliability of epidemic tracking by incorporating data as it arrives to better reflect the observed fidelity of the observations. The observations can recursively inform and train the model so that current conditions are better depicted and evolving outbreak characteristics are better matched.

A variety of data assimilation methods exist. The effectiveness of a particular method depends on model size and structure, as well as the quality of the observations. Filtering techniques, such as the Kalman filter and its offshoots, iteratively update, or adjust, model simulation estimates of the dynamic state, using real-world observations of that state, as the model is integrated through time. Because the state is intermittently and imprecisely measured, the filter balances the relative information contained in the observations with the model simulation. At the same time, the filtering process can also be used to estimate epidemiologically significant states and parameters within a model, creating uncertainty intervals for the estimates. Current approaches to disease modeling place confidence in historical data or on the ability to predict future outbreaks, ignoring the needs of public health officials to understand currently unfolding situations in ways that extract meaningful knowledge. Models are traditionally parameterized and calibrated when constructed, but the underlying parameters may be dependent on unobservable dynamic factors such as human con-

tact patterns, and hydrological flows. When executed properly, data assimilation can alert public health practitioners to epidemiological anomalies arising either through pathogen evolution or changes in the population [160]. In this work, we assume the values of various model states and parameters are only partially described, and that incoming measurements are incomplete and noisy. Therefore, state estimation using a data assimilation approach is also required. At each time step, we reconstruct our understanding of the current state by performing a parameter identification procedure to identify those model parameters whose simulated outcomes best match incoming measurements. These parameters are then used to compute estimates of the current state.

In this work, we solve a joint optimization problem involving both sensor selection and intervention optimization of a cholera epidemic. Both components of the optimization problem make use of prospective dynamics occurring over the prediction horizon. We obtain periodically incoming information sourced from a limited number of field-deployed sensors that measure pathogen concentrations. This information is then used to construct estimates of unobserved system states and to select admissible real-world interventions to apply at a number of locations in a spatially heterogeneous metapopulation model. Here, we present results from a study comparing different sensor selection criteria in order to highlight features essential to disease eradication in a metapopulation model.

4.4 Methodology

4.4.1 Mathematical Preliminaries

We apply our adaptive, multiscenario MPC scheme to a metapopulation model for cholera [137]. Each population within the overall metapopulation model contains compartments for susceptible, infected, and recovered individuals, along with a separate compartment for the infectious agent, the bacteria *V. cholerae*. Table 4.1 displays all the states and parameters in the metapopulation model. In this case, M refers to

the number of populations, or sites, within the model that experience the outbreak concurrently. We consider a situation where only $L < M$ sites can be monitored in a given time interval, or decision epoch, to obtain bacterial concentrations.

$$\frac{dx_S^i}{dt} = \mu(H_i - x_S^i) - \mathcal{F}_i(t)x_S^i + \rho x_R^i - u_i x_S^i, \tag{4.1}$$

$$\frac{dx_I^i}{dt} = \mathcal{F}_i(t)x_S^i - (\gamma + \mu + \alpha)x_I^i, \tag{4.2}$$

$$\frac{dx_R^i}{dt} = \gamma x_I^i - (\rho + \mu)x_R^i + u_i x_S^i, \tag{4.3}$$

$$\frac{dx_B^i}{dt} = -\delta x_B^i - \ell\Big(x_B^i - \sum_{j=1}^{M} P_{ji}\frac{H_j}{H_i}x_B^j\Big) + \xi\mathcal{G}_i(t), \tag{4.4}$$

$$\mathcal{F}_i(t) = \beta\Big[(1-m)\frac{x_B^i}{x_B^i + \kappa} + m\sum_{j=1}^{M} Q_{ij}\frac{x_B^j}{x_B^j + \kappa}\Big], \tag{4.5}$$

$$\mathcal{G}_i(t) = (1-m)x_I^i + m\sum_{j=1}^{M} Q_{ji}x_I^j, \tag{4.6}$$

$$Q_{ij} = \frac{H_j e^{\frac{-d_{ij}}{D}}}{\sum_{k\neq i}^{M} H_k e^{\frac{-d_{jk}}{D}}}, \tag{4.7}$$

$$i = 1, \ldots, M. \tag{4.8}$$

4.4.2 Algorithm

The proposed control algorithm is presented in Figure 4.1, and is outlined in this section. Each iteration of k represents one decision epoch.

1. *Offline Sensor Selection* Let $k = 0$. We solve the offline sensor selection problem at $t_k = t_0$, by selecting the L sites with the highest numbers of infected individuals, which will serve as the sensor sites for $[t_0, t_1]$. We label the resulting sites $\zeta(t_0)$.

2. *Parameter Identification* Let $k \leftarrow k + 1$. At time t_k, we obtain information from the target populations in the form of (limited) measurements. These measurements consist of the current number of infected individuals $y_I(t_k)$, and

Table 4.1.
Meaning of states and parameters in metapopulation model.

Variable	Meaning	Units
x_S^i	Number of susceptible individuals in population i	individuals
x_I^i	Number of infected individuals in population i	individuals
x_R^i	Number of recovered individuals in population i	individuals
x_B^i	Bacterial concentration in population i	cells/ml
H_i	Initial population size in population i	individuals
μ	Birth/death rate	days^{-1}
β	Transmission rate	days
κ	Bacterial concentration for half infection	cells/ml
ρ	Cholera immunity rate	days^{-1}
γ	Cholera recovery rate	days^{-1}
α	Cholera-induced mortality rate	days^{-1}
δ	Bacterial death rate	days^{-1}
ξ	Bacterial contamination rate	cells/ml/individual/day
ℓ	Bacterial dispersal rate	days^{-1}
P_{ij}	Probability of pathogen movement from i to j	dimensionless
Q_{ij}	Probability of human movement from i to j	dimensionless
m	Population connectivity parameter	dimensionless
d_{ij}	Distance from i to j	km
D	Distance parameter	km

the bacterial measurements at the previously selected sensors $y_B^i(t_k)$, $i \in \zeta(t_{k-1})$. We perform parameter identification using sparse grid interpolants. The sparse grid interpolant can be constructed across different spaces depending on k:

- If $k = 1$, then the interpolant is constructed over a combined parameter and initial condition space $\theta_1 \otimes \cdots \otimes \theta_{n_\theta} \otimes \mathcal{X}_1 \otimes \cdots \otimes \mathcal{X}_{4M-3}$, where n_θ is

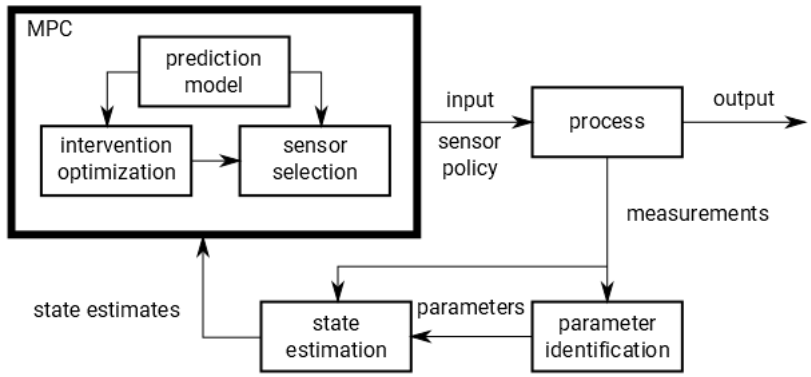


Fig. 4.1. Diagram of proposed control algorithm, which utilizes elements of model predictive control and sensor selection to derive optimal sensor policies given limited, periodic information.

the number of uncertain parameters, and $\mathcal{X}_i$ is the range of the ith initial condition. This combined space consists of those parameters and states that are unknown to us.

- If $k > 1$, then the interpolant is constructed over the parameter space $\theta_1 \otimes \cdots \otimes \theta_{n_\theta}$.

In both cases, the interpolant computes estimated trajectories $\tilde{x}_S(t_k, \theta)$, $\tilde{x}_I(t_k, \theta)$ $\tilde{x}_R(t_k, \theta)$, and $\tilde{x}_B(t_k, \theta)$. Prospective acceptable parameters are identified by fitting $\tilde{y}_I(t_k, \theta)$, $\tilde{y}_B^i(t_k, \theta)$ to $y_I(t_k)$, $y_B^i(t_k)$, $i \in \zeta(t_{k-1})$ with the following cost function:

$$C_k(\theta) = \sum_{i=1}^{M} \left(y_I^i(t_k) - \tilde{y}_I^i(t_k, \theta) \right)^2 + \sum_{j \in \zeta(t_{k-1})} \left(y_B^j(t_k) - \tilde{y}_B^j(t_k, \theta) \right)^2. \tag{4.9}$$

Prospective parameters are sampled from the grid points and through Latin Hypercube Sampling (LHS). After sorting the parameters by cost, we select the N_A parameters with the lowest costs to form the current parameter set Θ^k. We compute probabilities for each of the parameters in Θ^k:

$$p_k(\theta) = \frac{\frac{1}{C_k(\theta)}}{\sum_{\theta \in \Theta^k} \frac{1}{C_k(\Theta)}}. \tag{4.10}$$

3. *State Estimation* Due to our incomplete knowledge of the state of the system, we need to estimate its current state so we can further advance the system based on our most recent measurements. The previous step identified parameters most consistent with recent data, along with their associated probabilities. We make use of these quantities to compute state estimates of $x_S(t_k)$, $x_R(t_k)$, $x_B^i(t_k)$, $i \notin \zeta(t_{k-1})$ using parameter-based estimates $\tilde{x}_S(t_k, \theta)$, $\tilde{x}_R(t_k, \theta)$, $\tilde{x}_B^i(t_k, \theta)$, $i \notin \zeta(t_{k-1})$.

$$\bar{x}_A(t_k) = \mathbb{E}_k[\tilde{x}_A(t_k)] = \sum_{\theta \in \Theta^k} p_k(\theta)\tilde{x}_A(t_k, \theta), \quad A = \{S, R, B^i, i \notin \zeta(t_{k-1})\} \tag{4.11}$$

4. *Intervention Optimization* We now proceed to solve the intervention optimization problem at t_k. To do so, we require forecasts of dynamics over the prediction horizon H_p, which we obtain using sparse grid interpolation. The problem is stated as

$$\begin{aligned} \underset{u}{\text{minimize}} \quad & J_{\mathcal{I}}^k(u, \tilde{x}_B, \tilde{x}_I, \theta) \\ \text{subject to} \quad & \dot{\tilde{x}} = \tilde{F}(\tilde{x}, u, \theta) \\ & \tilde{x} \in \mathbb{X} \\ & u \in \mathbb{U} \\ & \theta \in \Theta^k \\ & t_k \leq t \leq t_{k+H_p} \end{aligned} \tag{4.12}$$

where the objective function is

$$J_{\mathcal{I}}^k(u, \tilde{x}_B, \tilde{x}_I, \theta) = \mathbb{E}_k\left[\sum_{j=1}^{H_p}\sum_{i=1}^{M} \mathcal{A}\alpha_\theta \tilde{y}_I^i(u, t_{k+j}|t_k) + \sum_{j=1}^{H_p}\sum_{i=1}^{M} \mathcal{C}u^i(t_{k+j}|t_k)\tilde{x}_S^i(u, t_{k+j}|t_k)\right] \tag{4.13}$$

The first part of Equation 4.13 inside the expectation refers to the cost associated with each cholera-related death. This is inferred by multiplying the number of infected individuals by the cholera-induced death rate and the cost per death. The second part refers to the cost of the applied intervention. Since the interventions primarily act on the individuals of each site who are susceptible and not yet infected, we multiply the control magnitude by the number

of predicted susceptible individuals who would receive the intervention and its benefits. The admissible control space $\mathbb{U}$ consists of limits on the intervention level at each site as well as a limit on the combined sum of interventions at all sites in each time interval.

5. *Online Sensor Selection* We solve the sensor selection problem at t_k. The sensor selection criterion is configured to select those sites that pose the greatest uncertainty in their bacterial concentrations. We quantify the uncertainty by measuring the variance of the predicted concentrations at each future intervention application time in the prediction horizon. This is a similar criterion to one used in model-based experiment design algorithms for reducing dynamical uncertainty [48–50]. Measurement selection for future experiments preferred future measurements that had the highest amount of uncertainty attached to them. We make use of the predictive interpolants generated in step 4.

$$\begin{aligned} \arg\max_{z} \quad & J^k_{\mathcal{S}}(z, \tilde{x}_B, y_I, \theta) = \sum_{i=1}^{M} z^i g(\tilde{x}^i_B(u, t_k, \theta)) + \sum_{i=1}^{M} z^i y^i_I(t_k) \\ \text{subject to} \quad & z = [z^1, \ldots, z^M] \\ & \sum_{i=1}^{M} z^i = L \\ & z^i \in \{0, 1\} \end{aligned} \tag{4.14}$$

$g(\tilde{x}^i_B(t_k, \theta))$ is a function of estimated bacterial concentrations, i.e.,

$$g(\tilde{x}^i_B(t_k, \theta)) = \max_{t_{k+1} \leq t \leq t_{k+H_p}} \mathbb{V}_k\Big[\tilde{x}^i_B(u, t|t_k)\Big] \tag{4.15}$$

$$\mathbb{V}_k\Big[\tilde{x}^i_B(t|t_k)\Big] = \sum_{\theta \in \Theta^k} p_k(\theta)\Big(\tilde{x}^i_B(u, t, \theta|t_k) - \mu^i_B(u, t, \theta|t_k)\Big)^2 \tag{4.16}$$

$$\mu^i_B(t, \theta|t_k) = \sum_{\theta \in \Theta^k} p_k(\theta)\tilde{x}^i_B(u, t, \theta|t_k) \tag{4.17}$$

Once we determine z, then we can compute $\zeta(t_k)$ as the set of indices of measurable outputs,

$$\zeta(t_k) = \{i : z^i = 1\}. \tag{4.18}$$

6. *Termination* If $k = N$, then terminate. Return controlled output $y_I(t)$, and input $u(t)$ trajectories upon termination, $t = [t_0, \ldots, t_N]$. Otherwise, return to Step 2.

Table 4.2.
Initial conditions for each site, and nominal parameter values for metapopulation model. * denotes parameters considered uncertain. Values for ℓ and m were retained from [137].

Variable	Value(s)
x_S	$[6000, 700, 9000, 4600, 900]$
x_I	$[3000, 200, 500, 200, 50]$
x_R	$[1000, 100, 500, 200, 50]$
x_B	$[100000, 50000, 10000, 50000, 50000]$
μ	4.56×10^{-5}
β*	1
κ	100000
ρ*	9.13×10^{-4}
γ*	0.01
α*	0.004
δ*	0.1
ξ*	1
ℓ	1.83
P_{ij}	$P_{12} = P_{23} = P_{34} = P_{45} = 1$
m	0.69
d_{ij}	$d_{12} = d_{23} = d_{34} = d_{45} = 50$
D	100

4.5 Results

We demonstrate the impact sensor selection criteria can have on reducing the number of infections and the bacterial reservoir responsible for perpetuating cholera, as well as minimizing predicted costs of interventions over time. Specifically, we compare four distinct sensor selection criteria:

1. **No sensor selection (NSS)** No pathogen sensors are deployed. The only information available is periodic updates on the number of infected individuals at all sites.

2. **Random sensor selection (RSS)** Pathogen sensors are deployed randomly at certain (but not all) sites, with their corresponding measurements arriving along with information on the number of infected individuals at all sites.

3. **Infection-based sensor selection (ISS)** Pathogen sensors are deployed at certain sites with the highest number of infected individuals. The number of infected individuals at all sites is also provided.

4. **Targeted sensor selection (TSS)** Pathogen sensors are deployed at certain sites according to the sensor selection problem defined in Equation 4.14, which optimizes sensor sites based on both the number of infected individuals and prospective bacterial uncertainty. The number of infected individuals at all sites is also provided.

Given a series of possible intervention application and bacterial sensor configuration times $[t_0, \ldots, t_N]$, we consider the previously described model in Section 4.4.1 with $M = 5$ populations, or sites, where only $L = 2$ sites can be measured for bacterial concentrations at each time interval. A vaccination program is implemented as an intervention at each of the sites, with $\mathcal{A}$, the cost of a cholera-related death, set to \$4500 [166], and $\mathcal{C}$, the cost of vaccinating an individual, set to \$6 [167]. Vaccination is implemented within the model by diverting currently susceptible, and potentially infected, individuals to the recovered state directly. The populations are connected

linearly, with bacterial movement being predominantly downstream, from site 1 to site 5. The simulation is conducted over the course of $t_N = 32$ days, with $N = 8$ total decision epochs, each lasting four days, allowing for adjustments in interventions and assimilation of new, incoming data. A prediction horizon of $H_P = 3$ decision epochs, or 12 days was selected. Parameter identification and state estimation are performed using $N_A = 1000$ parameters in each iteration. Ranges for initial conditions and uncertain parameter values were created by perturbing their value by 10%.

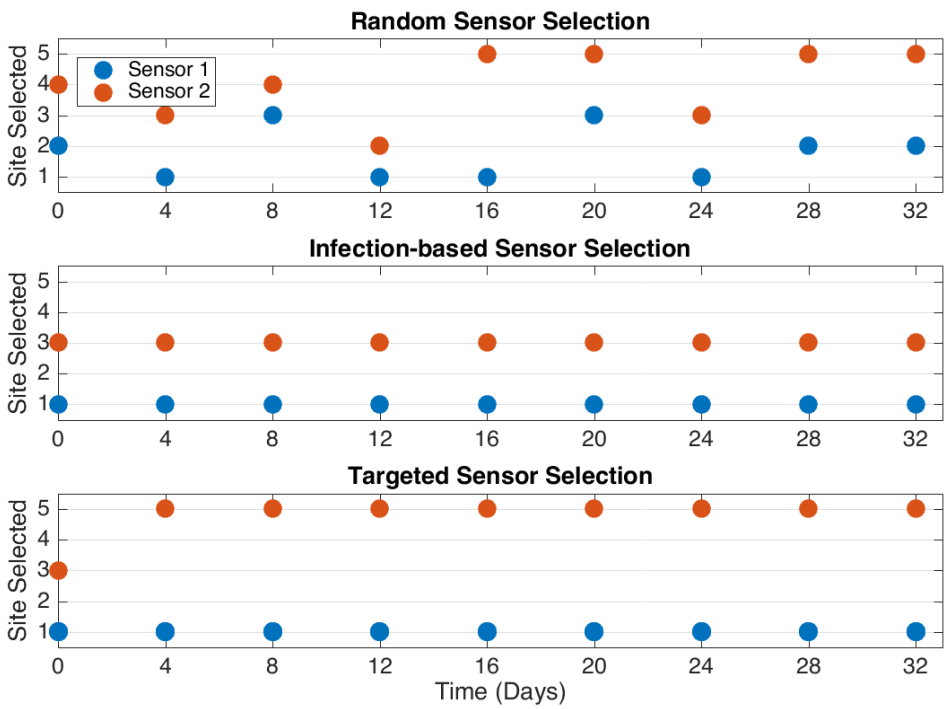


Fig. 4.2. Sites selected for sensing with different sensor selection criteria.

4.5.1 Examination of Sensor Policies

We first present the sensor policies derived from each sensor selection criterion for the duration of the simulation, shown in Figure 4.2. Of particular interest is the ISS and TSS criteria. ISS identifies sites with the highest number of infected

individuals; incidentally, sites 1 and 3 begin with the simulation with the highest numbers of infected individuals, and continue to maintain this trend throughout the simulation. Logically, placing sensors in these sites would make sense to measure the corresponding bacterial reservoirs. On the other hand, TSS selects sites 1 and 5 continuously throughout the duration of the simulation. The rationale for choosing these sites in TSS is revealed through analysis of Figs. 4.3 and 4.4.

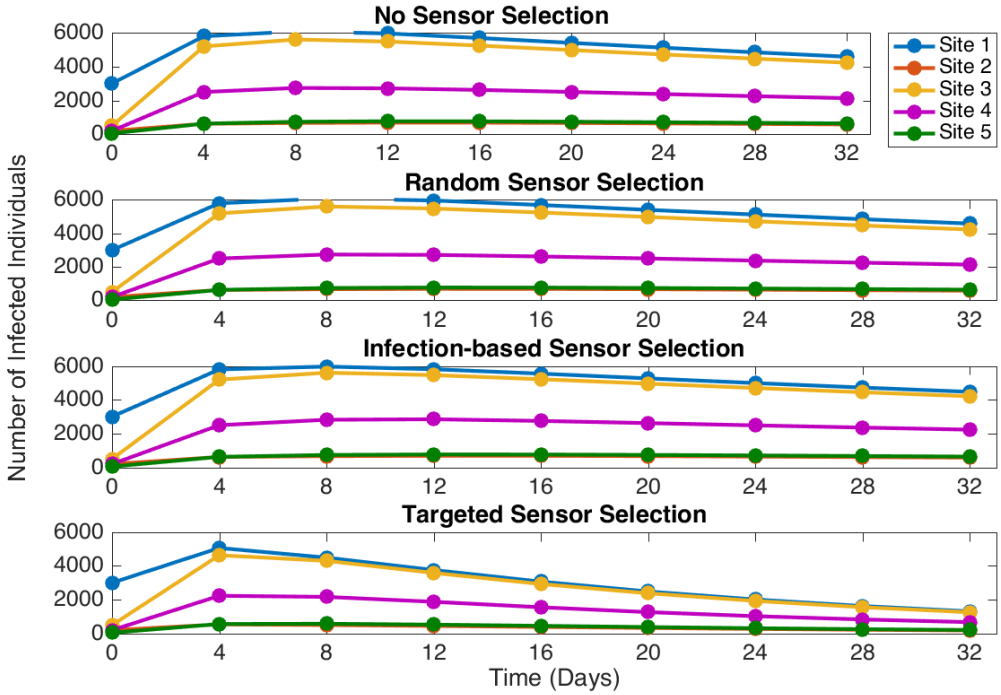


Fig. 4.3. Number of infected individuals across each site for different sensor selection criteria.

Fig. 4.3 displays the number of infected individuals at each site through time for the various sensor selection criteria studied. For all four criteria, sites 1 and 3 predominate the cholera case loads. Site 1 initially started with the highest number of infections, which peaks 4-8 days after the start of the simulation, before decreasing, like the rest of the sites. Site 3 has the same population size as, and is downstream of, site 1. The movement of pathogen and individuals from site 1 down towards site

3 explains the rise of site 3's infections to the same level as site 1. Sites 2, 4, and 5, all smaller population centers, do not accumulate as many infections due to smaller movements of people and pathogens. However, what differentiates the sensor selection criteria is how quick the descent in infections is. TSS is able to drastically lessen the number of infections over time, synchronizing the trajectories for both site 1 and 3, whereas the other criteria are unable to appreciably affect the number of infections to the same extent. Both RSS and NSS reach peaks of infection impacting nearly 60% of both site 1 and 3's total populations by day 8. ISS provides intermediate performance between the uninformed pathogen sensing duo of NSS, RSS and the predictive sensing of TSS.

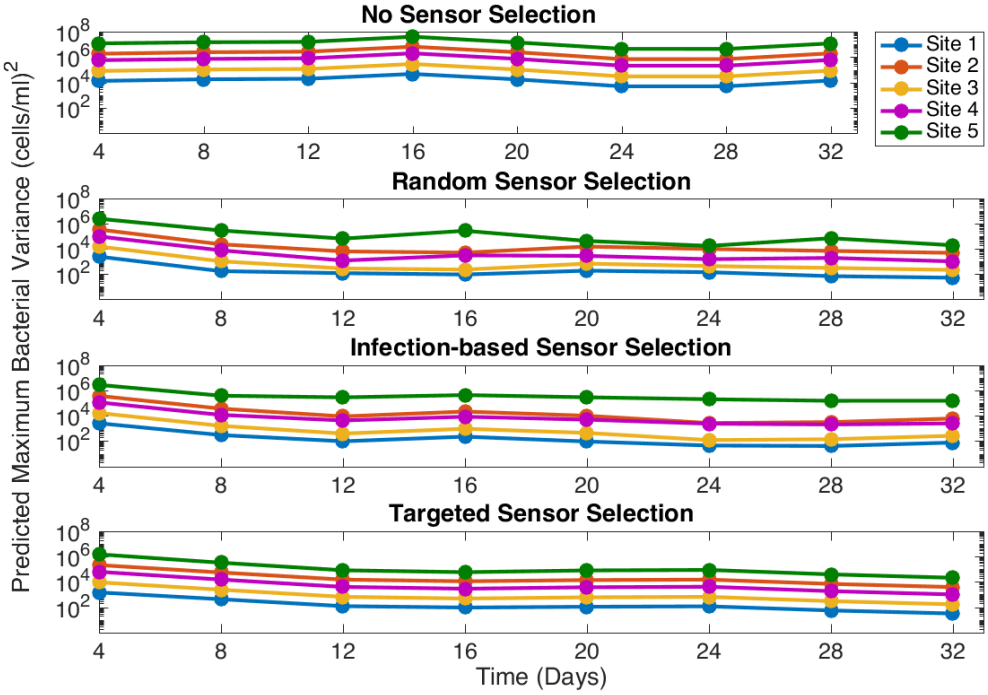


Fig. 4.4. Predicted maximum variance of bacterial concentrations across each site for different sensor selection criteria.

Fig. 4.4 reveals the state of uncertainty as it pertains to bacterial concentrations across each site for the different sensor selection criteria. These values are the max-

imum bacterial variance values $g(\tilde{x}_B^i(t_k, \theta))$, $i = 1, \ldots, M$, obtained as part of the solution to the online sensor selection problem (Equation 4.14). Because the online sensor selection problem involves prospective bacterial dynamics, these values are not actual, but predicted. It is evident across all criteria that site 5 contains the most uncertainty in its bacterial concentration when accounting for the multiple data-consistent parameters obtained in the parameter identification step. This is not surprising, considering the downstream flow of pathogens would inevitably lead to a rise in the bacterial reservoir in site 5. Furthermore, the bacterial variance of site 5 in RSS, ISS, and TSS decreases to below 10^6 cells^2/ml^2. Having no sensors available provides no recourse as the bacterial variance across all sites tends to fluctuate simultaneously. Taken together, Figs. 4.3 and 4.4 justify the selection of sites 1 and 5 for TSS; site 1 has the highest number of cholera infections, and site 5 has the highest uncertainty of bacterial concentrations.Fig. 4.3 clearly confirms the importance of sites 1 and 3 for ISS as hotbeds for infection.

4.5.2 Overall Impact of Sensor Policies

When viewed overall in terms total number of infections, TSS outperforms all other sensor selection criteria in terms of minimizing disease impact (Fig. 4.6). An 80% decrease in the number of infections is observed as a result of incorporating our sensor selection scheme compared to the alternative policies. While TSS contains the peak infections to day 4, the other sensor policies peak at day 8 before declining approximately 20% from that peak by the end of the simulation. NSS, RSS, and ISS essentially overlap in their profiles, suggesting in this aspect, these criteria are identifying similar outcomes.

The bacterial reservoir is another metric to measure the performance of the different selection criteria. While we do not directly optimize our interventions to minimize the bacterial concentrations in each site, our efforts to minimize the spread of infections indirectly limits the potential contamination of the bacterial reservoir by

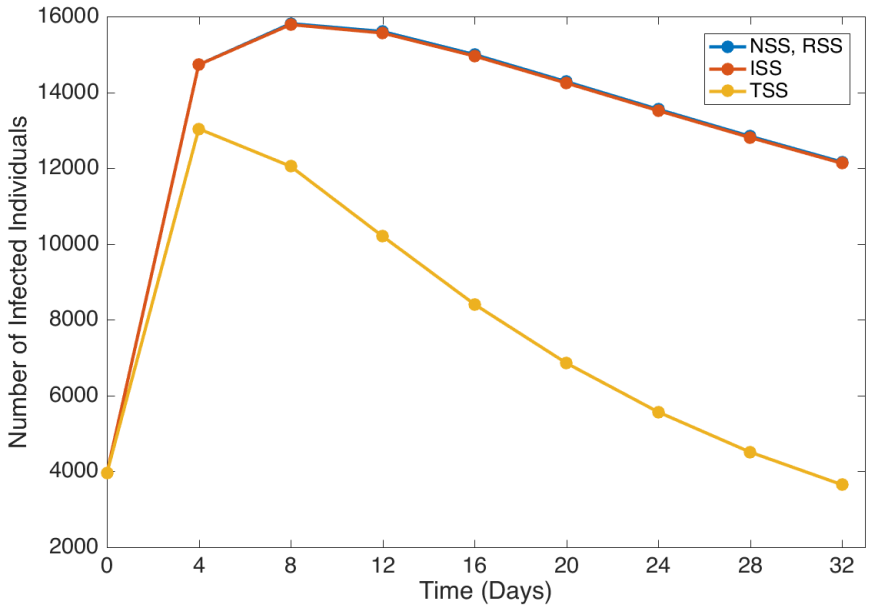


Fig. 4.5. Total number of infections throughout duration of simulation for different sensor selection criteria.

potential cholera patients. These patients, by not developing the infection, prevent further contribution to the growth of bacteria by not becoming infected themselves. The corresponding bacterial reservoir of that site fails to sustain itself in light of falling infections.

An interesting point to note in Figure 4.6 is that the bacterial concentrations for NSS, RSS, and ISS hover around 10^5 cells/ml, which is the infectious dose of bacteria needed to infect half of the population, the model variable κ. Coincidentally, approximately 12,000 infections were present by day 32 for these criteria, representing 45% of the overall population, whereas TSS ended with fewer than 4,000 infections. This improvement by TSS represents a 70% reduction over the other sensor selection criteria, which is complimented by an equivalent reduction in the bacterial reservoir.

Finally, we point out the costs in vaccinating the populations under each of the sensor selection scenarios, depicted in Fig. 4.7. These costs reflect the cumulative costs over the prediction horizon, so anticipating higher cholera-related casualties, or

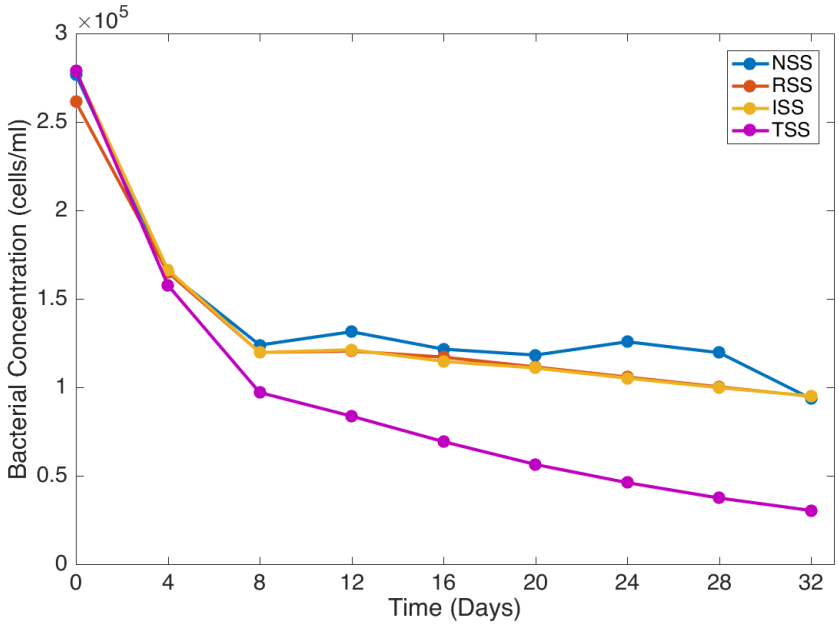


Fig. 4.6. Total bacterial concentrations throughout duration of simulation for different sensor selection criteria.

vaccines administered induces a higher cost. From the outset, TSS provides a clear advantage with reduced costs that decline quickly over time. The final cost is approximately 80% lower for using the uncertainty-based criterion over the alternatives. This is mainly due to the reduced costs associated with cholera-induced deaths and reduced administration of vaccines as a result of prior control actions.

As to what causes the difference between our proposed sensor selection criterion and the alternatives, different choices as to what data to acquire leads to drastically different outcomes in disease elimination and pathogen eradication. The targeted sensor selection criterion exploits the link between pathogen and host dynamics by prioritizing sites that have both the highest rates of cholera infection and the most uncertainty in bacterial dynamics with respect to previously determined data-consistent parameters. The function of a sensor policy is to deliver specific data related to the pathogen concentrations at selected sites. This data is then fit to parameter-derived simulations, which then attempt to reconstruct the unobserved states. The requisite

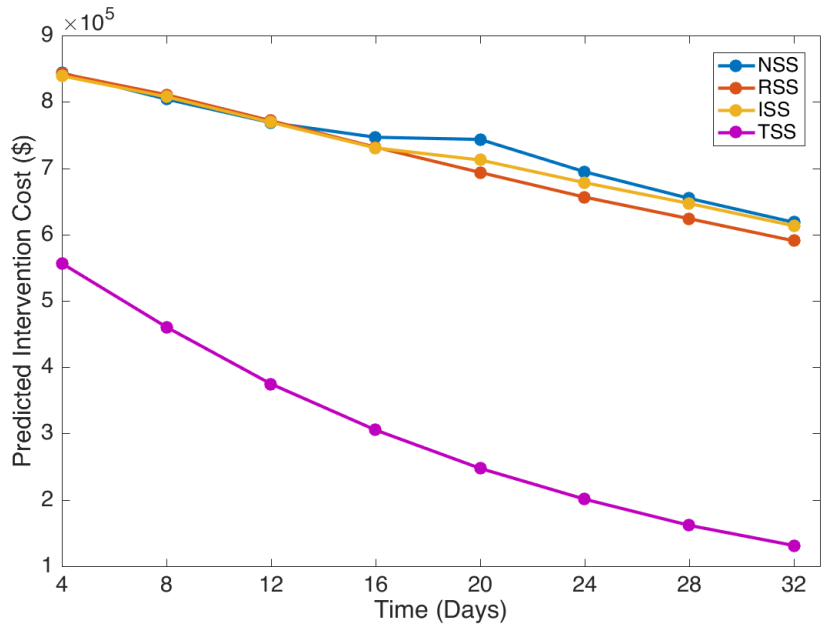


Fig. 4.7. Total predicted intervention costs throughout duration of simulation for different sensor selection criteria.

parameters and their probabilities change with respect to the observations available for fitting. It is not so obvious that the difference in the results presented here is due to the quality of the state estimates, but may be due in large part to these parameters and their probabilities, which influence the intervention optimization step. Intervention optimization includes an expectation with respect to the data-consistent parameters included in Θ^k for iteration k. These parameters can influence the decision landscape and the ultimate determination of an optimal control sequence. Each sensor selection criterion allocates its interventions differently based on the information provided, leading to different outcomes.

4.6 Conclusion

In this work, we presented an iterative control algorithm, relying on the principles of model predictive control, sensor selection, and recurring data assimilation, to mitigate the spread of cholera across a metapopulation system consisting of interconnected

populations. Using vaccination as the intervention of choice, we studied the effects of various sensor selection criteria on the overall objectives of disease eradication, both on the host and pathogen level. We address a joint optimization problem comprised of sequentially optimizing for admissible interventions and specifying future sensor policies. Our results demonstrate the efficacy and importance of incorporating the right variables, mainly the number of infected individuals and the uncertainty in bacterial concentrations, to make informed decisions on where to sense next.

The bacterial variance results in Figure 4.4 may raise the question of why we do not include a sensor selection criterion that would explicitly select sites with the highest pathogen concentrations. Site 5 had the highest predicted bacterial concentration uncertainty, but it also had the highest concentration of bacteria. With TSS, it is unclear which feature predominated. This may provide a more simplistic metric with which to select sensor sites that would not require a characterization of uncertainty. The predictive interpolants used to forecast future infections and vaccinations are a function of the parameter space and the possible controls implemented over the prediction horizon. The quality of these interpolants may alter the appeal of certain control regimes as compared to others, and is not to be overlooked. We plan to adjust the construction of these interpolants so that their results will not be relied upon in the event they are of suboptimal quality.

Another avenue of future work is to explore more complicated topologies than the linear configuration presented here, where sites have numerous inlets and outlets of pathogen and human flow that may temporarily make the site a hotspot in one decision epoch, only for such patterns to subside at the next decision epoch. Additional interventions besides the vaccination program considered here, such as hygiene and sanitation, will also be considered in this approach. Distributed [168] and decentralized [169] control may provide a real-world analog to localized medical decision-making by empowering each site to compose its own intervention strategy, with or without knowledge of other sites. The resulting intervention strategies can

be compared with those of a centralized decision-maker, demonstrated by the work presented here.

5. CONCLUSIONS AND FUTURE WORK

5.1 Conclusions

Infectious disease outbreaks are critical humanitarian episodes that require vast amounts of resources to properly contain and overcome. They also produce a wealth of opportunities that mathematical modeling can take advantage of. This thesis has presented examples of mathematical modeling that have examined the circumstances underlying future disease growth, improved parameter estimation strategies for matching observed outcomes in ongoing epidemics, and iterative predictive control approaches that can thrive in data-scarce environments where incoming measurement frequency is less than optimal.

Understanding the inherent transmission and recovery mechanisms of any disease, particularly one described by stochastic processes, would be invaluable for its management. Chapter 2 demonstrated how possible epidemiological scenarios arrange themselves based on desirable outcomes, in this case, the growth in cases below a certain threshold. The reproductive potential of certain scenarios were also assessed to determine the necessary case loads to avert a self-sustaining epidemic. Finally, we discussed the implications of selecting and modeling an appropriate representation of the epidemiological system, be it deterministic or stochastic.

Given outbreak data, the task of calibrating model parameters presented an opportunity to advance existing sparse grid-based parameter estimation approaches. In Chapter 3, we exploited the global-local hierarchy of interpolant creation and sampling to conduct a comparative study of various cluster analysis and metaheuristic methods using a two-stage algorithm to enhance the quality of fitting. We applied our algorithm to an epidemic model of cholera, which used ongoing data from the outbreak in Yemen, and a stochastic reaction network model of influenza. As the

number of iterations in the two-phase algorithm increased, further improvements in the search for data-consistent parameters were made. By examining the multitude of available clustering and metaheuristic methods available for optimization in the relevant parameter space, we were able to demonstrate the improved performance of certain combinations of methods over others.

Our final contribution in this thesis was in the area of epidemic control. In Chapter 4, we developed a model predictive control algorithm for a metapopulation model that combines sparsely sampled data with an intelligent sensor selection scheme that favors the most uncertain sites for bacterial concentration measurements. The defining feature of this chapter is the introduction of a joint optimization problem which encompasses both intervention optimization and sensor deployment. Various sensor selection schemes were compared for performance in terms of reducing the infected populations with minimal economic cost.

The impact of this work is not to be understated. There is a need for early identification and detection of emerging diseases, epidemics, and pandemics. Modeling can help prevent or mitigate the real-time threats of epidemic growth, setting quantitative intervention targets as events progress, providing real-time logistic allocation strategies and estimates. Ultimately, the success of these modeling approaches depends on the ability to predict and extrapolate the many avenues of transmission that an infectious pathogen avails itself to, in order to formulate the necessary, calculated response. Public health surveillance systems operating in real time would benefit from the exploratory modeling studies presented in this work.

5.2 Future Work

5.2.1 Disease Awareness

Social, economic, and cultural factors influence the spread of diseases. Word-of-mouth can often be the prevailing mode of information flow, either to the benefit or detriment of affected communities. People often correct for their behaviors in

the midst of an outbreak, and most models fail to capture this. Previously, the impact of behavior has been incorporated into the direct and indirect transmission rates [170], and as separate compartments of educated and uneducated susceptible individuals [145]. These non-pharmaceutical interventions enable exploration of the intangible, behavioral dynamics in play during the initial phases of the outbreak where it is essentially untouched. However, quantifying the effects of behavioral interventions is challenging from a modeling standpoint, particularly due to lack of adequate data. Testing verifiable hypotheses of a population's attitudes and actions in response to outbreaks is vital to integrating more detailed layers of behavior on top of existing epidemic processes.

5.2.2 Time Delays

Delay mathematical models provide an additional degree of realism by approximating the lags between identifying an intervention and adequately implementing it. [171, 172] studied cholera models wherein disinfectants and insecticides to sanitize bacterial reservoirs were applied after the bacterial density was measured, introducing a time lag into the ODE model. The resulting delay differential equation model explored how different combinations of intervention concentrations and time delays could effectively control the pathogen growth. Delays in the control variables have also been explored [173]. Another way of incorporating latency for either incubation periods or resource delays is to add an additional compartment to the traditional compartmental model approach. This is commonly done with the exposed compartment, E, or in the case of quarantines that separate known or suspected infectious individuals, the quarantine compartment Q.

5.2.3 Alternative Data Sources

Updating the current state of an evolving outbreak has traditionally relied on official government records, which often come late and consist of numerous reporting

errors. Today's real-time technologies, especially those emanating from the digital and social media landscape, can inform modelers relatively quickly as to qualitative changes in new outbreaks. Aggregation of multiple informal data sources, coupled with traditional information streams can improve the specificity and accuracy of localized public health risks [174, 175]. Of course, technology deserts present enormous gaps in coverage that conversely tend to also have the greatest disease burdens. On the other hand, [176] assimilated medical documentation from various sources to serve as proxies to estimate the reproductive potential of an ongoing influenza epidemic, so there are opportunities at multiple levels of the information hierarchy.

5.2.4 Improvements to Sparse Grid Interpolation

Our usage of sparse grid interpolation has typically relied on it producing ranges for the parameters, inputs, and initial conditions of interest. These ranges were derived from what is essentially a uniform distribution. Adapting the sparse grid construction process to accommodate other statistical distributions that these quantities may be derived from, such as Gaussian, Beta, and Gamma distributions, could help tailor surrogate models to the needs of public health researchers. Scanning the parameter space comprehensively with the underlying distribution in mind could validate plausible theories as to the statistical origins underlying those parameters. Additionally, adaptively incorporating realizations into the sampling process for SDE-based models within the sparse grid construction would enable further exploration and exploitation of the stochastic processes for disease growth, by identifying the minimum necessary number of realizations for adequate characterization of disease dynamics. Infectious diseases are, by nature, stochastic, nonlinear, and often chaotic [30], so faithfully approximating these dynamics will bring models one step closer to reality.

REFERENCES

REFERENCES

[1] F. Brauer, "Mathematical epidemiology is not an oxymoron," *BMC Public Health*, vol. 9, no. SUPPL. 1, pp. 1–11, 2009.

[2] K. F. Smith, M. Goldberg, S. Rosenthal, L. Carlson, J. Chen, C. Chen, and S. Ramachandran, "Global rise in human infectious disease outbreaks," *Journal of The Royal Society Interface*, vol. 11, no. 101, pp. 20 140 950–20 140 950, 2014.

[3] WHO, "The top 10 causes of death," 2017.

[4] "The Millennium Development Goals Report," United Nations, Tech. Rep., 2010.

[5] D. Mollison, V. Isham, and B. Grenfell, "Epidemics: Models and Data," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, vol. 157, no. 1, p. 115, 1994.

[6] A. Huppert and G. Katriel, "Mathematical modelling and prediction in infectious disease epidemiology," *Clinical Microbiology and Infection*, vol. 19, no. 11, pp. 999–1005, 2013.

[7] A. Krämer, M. Akmatov, and M. Kretzschmar, "Principles of Infectious Disease Epidemiology," 2009, pp. 85–99.

[8] H. W. Hethcote, "The Mathematics of Infectious Diseases," *SIAM Review*, vol. 42, no. 4, pp. 599–653, 2000.

[9] M. J. Keeling and L. Danon, "Mathematical modelling of infectious diseases," *British Medical Bulletin*, vol. 92, no. 1, pp. 33–42, 2009.

[10] E. S. Allman and J. A. Rhodes, *Mathematical Models in Biology: An Introduction*. Cambridge University Press, 2004.

[11] D. Chen, B. Moulin, and J. Wu, "Introduction to Analyzing and Modeling Spatial and Temporal Dynamics of Infectious Diseases," in *Analyzing and Modeling Spatial and Temporal Dynamics of Infectious Diseases*, 2015, pp. 1–17.

[12] J. Glasser, M. I. Meltzer, and B. Levin, "Mathematical modeling and public policy: responding to health crises." *Emerging infectious diseases*, vol. 10, no. 11, pp. 2050–2051, 2004.

[13] D. Chen, "Modeling the Spread of Infectious Diseases: A Review," in *Analyzing and Modeling Spatial and Temporal Dynamics of Infectious Diseases*, 2015, pp. 19–42.

[14] M. Kretzschmar and J. Wallinga, "Mathematical Models in Infectious Disease Epidemiology," 2009, pp. 209–221.

[15] C. I. Siettos and L. Russo, "Mathematical modeling of infectious disease dynamics," *Virulence*, vol. 4, no. 4, pp. 295–306, 2013.

[16] P. Patlolla, V. Gunupudi, A. R. Mikler, and R. T. Jacob, "Agent-based simulation tools in computational Epidemiology," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3473 LNCS, pp. 212–223, 2006.

[17] L. Perez and S. Dragicevic, "An agent-based approach for modeling dynamics of contagious disease spread," *International Journal of Health Geographics*, vol. 8, no. 1, pp. 1–17, 2009.

[18] B. Roche, J. M. Drake, and P. Rohani, "An Agent-Based Model to study the epidemiological and evolutionary dynamics of Influenza viruses," *BMC Bioinformatics*, vol. 12, 2011.

[19] F. Miksch, C. Urach, P. Einzinger, and G. Zauner, "A flexible agent-based framework for infectious disease modeling," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8407 LNCS, pp. 36–45, 2014.

[20] S. Venkatramanan, B. Lewis, J. Chen, D. Higdon, A. Vullikanti, and M. Marathe, "Using data-driven agent-based models for forecasting emerging infectious diseases," *Epidemics*, 2016.

[21] L. A. Meyers, "Contact network epidemiology: Bond percolation applied to infectious disease prediction and control," *Bulletin of the American Mathematical Society*, vol. 44, no. 1, pp. 63–86, 2007.

[22] E. Volz and L. A. Meyers, "Susceptible-infected-recovered epidemics in dynamic contact networks," *Proceedings of the Royal Society B: Biological Sciences*, vol. 274, no. 1628, pp. 2925–2934, 2007.

[23] E. Volz, "SIR dynamics in random networks with heterogeneous connectivity," *Journal of Mathematical Biology*, vol. 56, no. 3, pp. 293–310, 2008.

[24] N. B. Dimitrov and L. A. Meyers, "Mathematical Approaches to Infectious Disease Prediction and Control," *Risk and Optimization in an Uncertain World*, no. November 2017, pp. 1–25, 2010.

[25] B. Pourbohloul, L. A. Meyers, D. M. Skowronski, M. Krajden, D. M. Patrick, and R. C. Brunham, "Modeling control strategies of respiratory pathogens," *Emerging Infectious Diseases*, vol. 11, no. 8, pp. 1249–1256, 2005.

[26] W. O. Kermack and A. G. McKendrick, "A Contribution to the Mathematical Theory of Epidemics," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 115, no. 772, pp. 700–721, 1927.

[27] R. Pitman, D. N. Fisman, G. S. Zaric, M. Postma, M. Kretzschmar, W. J. Edmunds, and M. Brisson, "Dynamic Transmission Modeling: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group-5," *Medical Decision Making*, vol. 32, no. 5, pp. 712–721, 2012.

[28] L. J. S. Allen, "An Introduction to Stochastic Epidemic Models," *Mathematical Epidemiology*, vol. 1945, no. 3, pp. 81–130, 2008.

[29] T. Britton, "Stochastic epidemic models: A survey," *Mathematical Biosciences*, vol. 225, no. 1, pp. 24–35, 2010.

[30] H. Heesterbeek, R. M. Anderson, V. Andreasen, S. Bansal, D. De Angelis, C. Dye, K. T. D. Eames, W. J. Edmunds, S. D. W. Frost, S. Funk, T. D. Hollingsworth, T. House, V. Isham, P. Klepac, J. Lessler, J. O. Lloyd-Smith, C. J. E. Metcalf, D. Mollison, L. Pellis, J. R. C. Pulliam, M. G. Roberts, and C. Viboud, "Modeling infectious disease dynamics in the complex landscape of global health," *Science*, vol. 347, no. 6227, pp. aaa4339–aaa4339, 2015.

[31] A. R. Cook, W. Otten, G. Marion, G. J. Gibson, and C. a. Gilligan, "Estimation of multiple transmission rates for epidemics in heterogeneous populations." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 51, pp. 20 392–20 397, 2007.

[32] S. A. Levin and R. Durrett, "From Individuals to Epidemics," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 351, no. 1347, pp. 1615–1621, 1996.

[33] R. Levins, "Some Demographic and Genetic Consequences of Environmental Heterogeneity for Biological Control," *Bulletin of the Entomological Society of America*, vol. 15, no. 3, pp. 237–240, 1969.

[34] J. Arino and P. van den Driessche, "Disease spread in metapopulations," in *Nonlinear Dynamics and Evolution Equations*. Providence, Rhode Island: American Mathematical Society, 2006, vol. 48, pp. 1–12.

[35] G. E. Glass, J. L. Aron, J. H. Ellis, and S. S. Yoon, "Applications of GIS technology to disease control," *Papers on Population WP 93-05*, pp. vi, 39 p., 1993.

[36] A. Sai and N. Kong, "Sparse Grid Interpolation of It o Stochastic Models in Epidemiology and Systems Biology," *IAENG International Journal of Applied Mathematics*, vol. 48, no. 1, pp. 45–52, 2018.

[37] R. M. May, "Uses and Abuses of Mathematics in Biology," *Science*, vol. 303, no. 5659, pp. 790–793, 2004.

[38] H. Kitano, "Computational Systems Biology," *Nature*, vol. 420, no. 6912, pp. 206–10, 2002.

[39] T. Székely and K. Burrage, "Stochastic Simulation in Systems Biology," *Computational and Structural Biotechnology Journal*, vol. 12, no. 20-21, pp. 14–25, 2014.

[40] S. Ditlevsen and A. Samson, "Introduction to Stochastic Models in Biology," in *Springer-Verlag Berlin Heidelberg*, ser. Lecture Notes in Mathematics, M. Bachar, J. Batzel, and S. Ditlevsen, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, vol. 2058, pp. 3–35.

[41] T. Manninen, M.-L. Linne, and K. Ruohonen, "Developing Itô stochastic differential equation models for neuronal signal transduction pathways." *Computational biology and chemistry*, vol. 30, no. 4, pp. 280–291, 2006.

[42] A. Saarinen, M.-L. Linne, and O. Yli-Harja, "Modeling Single Neuron Behavior Using Stochastic Differential Equations," *Neurocomputing*, vol. 69, pp. 1091–1096, 2006.

[43] ——, "Stochastic Differential Equation Model for Cerebellar Granule Cell Excitability," *PLoS Computational Biology*, vol. 4, no. 2, p. e1000004, 2008.

[44] O. Chis and D. Opris, "Mathematical Analysis of Stochastic Models for Tumor-Immune Systems," *Romania*, pp. 1–19, 2009.

[45] C. Ji, D. Jiang, and N. Shi, "Analysis of a predator-prey model with modified Leslie-Gower and Holling-type II schemes with stochastic perturbation," *Journal of Mathematical Analysis and Applications*, vol. 359, no. 2, pp. 482–498, 2009.

[46] H. Zhou and M. Liu, "Analysis of a Stochastic Predator-Prey Model in Polluted Environments," *IAENG International Journal of Applied Mathematics*, vol. 46, no. 4, pp. 445–456, 2016.

[47] A. K. Duun-Henriksen, S. Schmidt, R. M. Røge, J. B. Møller, K. Nørgaard, J. B. Jørgensen, and H. Madsen, "Model Identification using Stochastic Differential Equation Grey-Box Models in Diabetes," *Journal of diabetes science and technology*, vol. 7, no. 2, pp. 431–40, 2013.

[48] M. M. Donahue, G. T. Buzzard, and A. E. Rundell, "Experiment design through dynamical characterisation of non-linear systems biology models utilising sparse grids." *IET systems biology*, vol. 4, no. 4, pp. 249–62, 2010.

[49] J. N. Bazil, G. T. Buzzard, and A. E. Rundell, "A Global Parallel Model based Design of Experiments Method to Minimize Model Output Uncertainty," *Bulletin of mathematical biology*, vol. 74, no. 3, pp. 688–716, 2012.

[50] T. Mdluli, G. T. Buzzard, and A. E. Rundell, "Efficient Optimization of Stimuli for Model-Based Design of Experiments to Resolve Dynamical Uncertainty," *PLoS Computational Biology*, vol. 11, no. 9, p. e1004488, 2015.

[51] T. Gerstner and M. Griebel, "Dimension-Adaptive Tensor-Product Quadrature," *Computing*, vol. 71, no. 1, pp. 65–87, 2003.

[52] H.-J. Bungartz and M. Griebel, "Sparse grids," *Acta Numerica*, vol. 13, p. 147, 2004.

[53] T. Gerstner and M. Griebel, "Sparse Grids," in *Encyclopedia of Quantitative Finance*. Chichester, UK: John Wiley & Sons, Ltd, 2010, vol. 13, pp. 147–269.

[54] A. Klimke and B. Wohlmuth, "Algorithm 847: spinterp: Piecewise Multilinear Hierarchical Sparse Grid Interpolation in MATLAB," *ACM Transactions on Mathematical Software*, vol. 31, no. 4, pp. 561–579, 2005.

[55] M. Andrecut, "Stochastic recovery of sparse signals from random measurements," *Engineering Letters*, vol. 19, no. 1, pp. 1–6, 2011.

[56] F. Nobile, R. Tempone, and C. G. Webster, "A Sparse Grid Stochastic Collocation Method for Partial Differential Equations with Random Input Data," *SIAM Journal on Numerical Analysis*, vol. 46, no. 5, pp. 2309–2345, 2008.

[57] X. Ma and N. Zabaras, "An Adaptive Hierarchical Sparse Grid Collocation Algorithm for the Solution of Stochastic Differential Equations," *Journal of Computational Physics*, vol. 228, no. 8, pp. 3084–3113, 2010.

[58] N. Agarwal and N. R. Aluru, "A Domain Adaptive Stochastic Collocation Approach for Analysis of MEMS under Uncertainties," *Journal of Computational Physics*, vol. 228, no. 20, pp. 7662–7688, 2009.

[59] ——, "Weighted Smolyak Algorithm for Solution of Stochastic Differential Equations on Non-uniform Probability Measures," *International Journal for Numerical Methods in Engineering*, vol. 85, no. 11, pp. 1365–1389, 2011.

[60] M. Liu, Z. Gao, and J. S. Hesthaven, "Adaptive Sparse Grid Algorithms with Applications to Electromagnetic Scattering under Uncertainty," *Applied Numerical Mathematics*, vol. 61, no. 1, pp. 24–37, 2011.

[61] S. Sankaran and A. L. Marsden, "A Stochastic Collocation Method for Uncertainty Quantification and Propagation in Cardiovascular Simulations." *Journal of Biomechanical Engineering*, vol. 133, no. 3, p. 31001, 2011.

[62] G. Zhang, M. Gunzburger, and W. Zhao, "A Sparse-Grid Method for Multi-Dimensional Backward Stochastic Differential Equations," *Journal of Computational Mathematics*, vol. 31, no. 3, pp. 221–248, 2013.

[63] R. Pulch, "Stochastic Collocation and Stochastic Galerkin Methods for Linear Differential Algebraic Equations," *Journal of Computational and Applied Mathematics*, vol. 262, pp. 281–291, 2014.

[64] T. Gerstner and M. Griebel, "Sparse grids (Quantitative Finance)," *Encyclopedia of Quantitative Finance*, vol. 13, p. 5, 2008.

[65] ——, "Numerical Integration using Sparse Grids," *Numerical Algorithms*, vol. 18, pp. 209–232, 1998.

[66] V. Barthelmann, E. Novak, and K. Ritter, "High Dimensional Polynomial Interpolation on Sparse Grids," *Advances in Computational Mathematics*, vol. 12, pp. 273–288, 2000.

[67] A. Klimke, K. Willner, and B. Wohlmuth, "Uncertainty Modeling using Fuzzy Arithmetic based on Sparse Grids: Applications to Dynamic Systems," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 12, no. 6, pp. 745–759, 2004.

[68] A. Klimke, "Sparse Grid Interpolation Toolbox User's Guide," *IANS report*, 2006.

[69] A. Gil, J. Segura, and N. M. Temme, *Numerical Methods for Special Functions*. Society for Industrial and Applied Mathematics, 2007.

[70] G. T. Buzzard and D. Xiu, "Variance-based Global Sensitivity Analysis via Sparse-Grid Interpolation and Cubature," *Communications in Computational Physics*, vol. 9, no. 3, pp. 542–567, 2011.

[71] D. J. Higham, "An Algorithmic Introduction to Numerical Simulation of Stochastic Differential Equations," *SIAM Review*, vol. 43, no. 3, pp. 525–546, 2001.

[72] T. Sauer, "Numerical Solution of Stochastic Differential Equations in Finance," in *Handbook of Computational Finance*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, vol. 47, no. 1, pp. 529–550.

[73] ——, "Computational Solution of Stochastic Differential Equations," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 5, no. 5, pp. 362–371, 2013.

[74] E. Tornatore, P. Vetro, and S. M. Buccellato, "SIVR Epidemic Model with Stochastic Perturbation," *Neural Computing and Applications*, vol. 24, no. 2, pp. 309–315, 2014.

[75] D. Jiang, Q. Liu, N. Shi, T. Hayat, A. Alsaedi, and P. Xia, "Dynamics of a Stochastic HIV-1 Infection Model with Logistic Growth," *Physica A: Statistical Mechanics and its Applications*, vol. 469, pp. 706–717, 2017.

[76] A. Szepessy, R. Tempone, and G. E. Zouraris, "Adaptive weak approximation of stochastic differential equations," *Communications on Pure and Applied Mathematics*, vol. 54, no. 10, pp. 1169–1214, 2001.

[77] K.-S. Moon, A. Szepessy, and G. E. Zouraris, "Stochastic Analysis and Applications Convergence Rates for Adaptive Weak Approximation of Stochastic Differential Equations Convergence Rates for Adaptive Weak Approximation of Stochastic Differential Equations," *Stochastic Analysis and Applications*, vol. 233, no. 23, pp. 511–558, 2005.

[78] A. Oroji, M. Omar, and S. Yarahmadian, "An Îto stochastic differential equations model for the dynamics of the MCF-7 breast cancer cell line treated by radiotherapy," *Journal of Theoretical Biology*, vol. 407, pp. 128–137, 2016.

[79] J. R. Banga, "Optimization in Computational Systems Biology," *BMC Systems Biology*, vol. 2, p. 47, 2008.

[80] J. R. Banga and E. Balsa-Canto, "Parameter Estimation and Optimal Experimental Design," *Essays In Biochemistry*, vol. 45, pp. 195–210, 2008.

[81] I. Swameye, T. G. Muller, J. Timmer, O. Sandra, and U. Klingmuller, "Identification of Nucleocytoplasmic Cycling as a Remote Sensor in Cellular Signaling by Databased Modeling," *Proceedings of the National Academy of Sciences*, vol. 100, no. 3, pp. 1028–1033, 2003.

[82] C. Surulescu, "On Some Stochastic Differential Equation Models with Applications to Biological Problems," ICAM, WWU Muenster, Tech. Rep., 2011.

[83] C. Surulescu and N. Surulescu, "Some Classes of Stochastic Differential Equations as an Alternative Modeling Approach to Biomedical Problems," 2013, pp. 269–307.

[84] A. S. Hurn, K. A. Lindsay, and V. L. Martin, "On the efficacy of simulated maximum likelihood for estimating the parameters of stochastic differential equations," *Journal of Time Series Analysis*, vol. 24, no. 1, pp. 45–63, 2003.

[85] M. M. Donahue and G. T. Buzzard, "Parameter identification with adaptive sparse grid-based optimization for models of cellular processes," *Methods in Bioengineering: Systems Analysis of Biological Networks*, pp. 1–32, 2009.

[86] M. M. Donahue, G. T. Buzzard, and A. E. Rundell, "Robust parameter identification with adaptive sparse grid-based optimization for nonlinear systems biology models," in *2009 American Control Conference*. IEEE, 2009, pp. 5055–5060.

[87] S. L. Noble, G. T. Buzzard, and A. E. Rundell, "Feasible parameter space characterization with adaptive sparse grids for nonlinear systems biology models," in *Proceedings of the 2011 American Control Conference*. IEEE, 2011, pp. 2909–2914.

[88] J. Sun, J. M. Garibaldi, and C. Hodgman, "Parameter estimation using metaheuristics in systems biology: a comprehensive review." *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, vol. 9, no. 1, pp. 185–202, 2012.

[89] L. M. A. Bettencourt and R. M. Ribeiro, "Real time bayesian estimation of the epidemic potential of emerging infectious diseases," *PLoS ONE*, vol. 3, no. 5, 2008.

[90] F. C. Coelho, C. T. Codeço, and M. G. M. Gomes, "A Bayesian framework for parameter estimation in dynamical models," *PLoS ONE*, vol. 6, no. 5, pp. 1–6, 2011.

[91] L. Held, "Bayesian Methods in Epidemiology," in *Handbook of Epidemiology*. New York, NY: Springer New York, 2014, pp. 1161–1193.

[92] G. Katriel, R. Yaari, A. Huppert, U. Roll, and L. Stone, "Modelling the initial phase of an epidemic using incidence and infection network data: 2009 H1N1 pandemic in Israel as a case study," *Journal of The Royal Society Interface*, vol. 8, no. 59, pp. 856–867, 2011.

[93] L. F. White, J. Wallinga, L. Finelli, C. Reed, S. Riley, M. Lipsitch, and M. Pagano, "Estimation of the reproductive number and the serial interval in early phase of the 2009 influenza A/H1N1 pandemic in the USA," *Influenza and other Respiratory Viruses*, vol. 3, no. 6, pp. 267–276, 2009.

[94] L. F. White and M. Pagano, "Transmissibility of the influenza virus in the 1918 pandemic," *PLoS ONE*, vol. 3, no. 1, 2008.

[95] R. Yaari, G. Katriel, L. Stone, E. Mendelson, M. Mandelboim, and A. Huppert, "Model-based reconstruction of an epidemic using multiple datasets: understanding influenza A/H1N1 pandemic dynamics in Israel," *Journal of The Royal Society Interface*, vol. 13, no. 116, p. 20160099, 2016.

[96] O. Akman and E. Schaefer, "An evolutionary computing approach for parameter estimation investigation of a model for cholera," *Journal of Biological Dynamics*, vol. 9, no. 1, pp. 147–158, 2015.

[97] G. Chowell, "Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts," *Infectious Disease Modelling*, vol. 2, no. 3, pp. 379–398, 2017.

[98] A. Smirnova and G. Chowell, "A primer on stable parameter estimation and forecasting in epidemiology by a problem-oriented regularized least squares algorithm," *Infectious Disease Modelling*, vol. 2, no. 2, pp. 268–275, 2017.

[99] E. Balsa-Canto, M. Peifer, J. R. Banga, J. Timmer, and C. Fleck, "Hybrid optimization method with general switching strategy for parameter estimation," *BMC Systems Biology*, vol. 2, no. 1, p. 26, 2008.

[100] M. Rodriguez-Fernandez, P. Mendes, and J. R. Banga, "A hybrid approach for efficient and robust parameter estimation in biochemical pathways," *Biosystems*, vol. 83, no. 2-3, pp. 248–265, 2006.

[101] R. Yaari, I. Dattner, and A. Huppert, "A two-stage approach for estimating the parameters of an age-group epidemic model from incidence data," *Statistical Methods in Medical Research*, pp. 1–16, 2017.

[102] S. Smolyak, "Quadrature and interpolation formulas for tensor products of certain classes of functions," *Soviet Mathematics, Doklady*, vol. 4, pp. 240–243, 1963.

[103] C. Zenger, "Sparse Grids," in *Parallel Algorithms for Partial Differential Equations*, ser. Notes on Numerical Fluid Mechanics, W. Hackbusch, Ed., vol. 31. Vieweg, 1991, pp. 241–251.

[104] H. Yserentant, "Sparse grids, adaptivity, and symmetry," *Computing (Vienna/New York)*, vol. 78, no. 3, pp. 195–209, 2006.

[105] C. G. Moles, P. Mendes, and J. R. Banga, "Parameter estimation in biochemical pathways: a comparison of global optimization methods." *Genome research*, vol. 13, no. 11, pp. 2467–74, 2003.

[106] A. Rinnooy Kan and G. Timmer, "Stochastic Global Optimization Methods Part I: Clustering Methods," *Mathematical Programming*, vol. 39, no. 1, pp. 27–56, 1987.

[107] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 226–231.

[108] J. H. Holland, "Adaptation in Natural and Artificial Systems," *Ann Arbor MI University of Michigan Press*, vol. Ann Arbor, p. 183, 1975.

[109] M. Srinivas and L. M. Patnaik, "Genetic Algorithms: A Survey," *Computer*, vol. 27, no. 6, pp. 17–26, 1994.

[110] J. Kennedy and R. Eberhart, "Particle swarm optimization," *Neural Networks, 1995. Proceedings., IEEE International Conference on*, vol. 4, pp. 1942–1948, 1995.

[111] Anonymous, "Influenza in a boarding school," *British Medical Journal*, vol. 1, p. 587, 1978.

[112] M. J. Keeling and P. Rohani, *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, 2011.

[113] M. E. Woolhouse, C. Dye, J. F. Etard, T. Smith, J. D. Charlwood, G. P. Garnett, P. Hagan, J. L. Hii, P. D. Ndhlovu, R. J. Quinnell, C. H. Watts, S. K. Chandiwana, and R. M. Anderson, "Heterogeneities in the transmission of infectious agents: implications for the design of control programs." *Proc Natl Acad Sci U S A*, vol. 94, no. 1, pp. 338–342, 1997.

[114] G. Jenkinson and J. Goutsias, "Numerical integration of the master equation in some models of stochastic epidemiology," *PLoS ONE*, vol. 7, no. 5, 2012.

[115] S. Liao, T. Vejchodský, and R. Erban, "Tensor methods for parameter estimation and bifurcation analysis of stochastic reaction networks," *Journal of The Royal Society Interface*, vol. 12, no. 108, p. 20150233, 2015.

[116] M. Hegland, A. Hellander, and P. Lötstedt, "Sparse grids and hybrid methods for the chemical master equation," *BIT Numerical Mathematics*, vol. 48, no. 2, pp. 265–283, 2008.

[117] Z. Mukandavire, S. Liao, J. Wang, H. Gaff, D. L. Smith, and J. G. Morris, "Estimating the reproductive numbers for the 2008-2009 cholera outbreaks in Zimbabwe," *Proceedings of the National Academy of Sciences*, vol. 108, no. 21, pp. 8767–8772, 2011.

[118] I. C.-H. Fung, "Cholera transmission dynamic models for public health practitioners," *Emerging Themes in Epidemiology*, vol. 11, no. 1, p. 1, 2014.

[119] Y. H. Grad, J. C. Miller, and M. Lipsitch, "Cholera Modeling," *Epidemiology*, vol. 23, no. 4, pp. 523–530, 2012.

[120] H. Nishiura, S. Tsuzuki, B. Yuan, T. Yamaguchi, and Y. Asai, "Transmission dynamics of cholera in Yemen, 2017: a real time forecasting," *Theoretical Biology and Medical Modelling*, vol. 14, no. 1, p. 14, 2017.

[121] A. Lloyd, "Sensitivity of Model-Based Epidemiological Parameter Estimation to Model Assumptions," in *Mathematical and Statistical Estimation Approaches in Epidemiology*. Dordrecht: Springer Netherlands, 2009, pp. 123–141.

[122] D. P. Moualeu-Ngangue, S. Röblitz, R. Ehrig, and P. Deuflhard, "Parameter identification in a tuberculosis model for Cameroon," *PLoS ONE*, vol. 10, no. 4, pp. 1–20, 2015.

[123] Y. Kao and M. C. Eisenberg, "Practical unidentifiability of a simple vector-borne disease model : implications for parameter estimation and intervention assessment," pp. 1–31, 2017.

[124] R. I. Glass and R. E. Black, *Cholera*, D. Barua and W. B. Greenough, Eds. Boston, MA: Springer US, 1992, no. January 1992.

[125] J. B. Harris, R. C. LaRocque, F. Qadri, E. T. Ryan, and S. B. Calderwood, "Cholera," *The Lancet*, vol. 379, no. 9835, pp. 2466–2476, 2012.

[126] S. Jahan, "Cholera Epidemiology, Prevention and Control," in *Significance, Prevention and Control of Food Related Diseases*. InTech, apr 2016, vol. 56, no. C, pp. 604–609.

[127] WHO, "Cholera Fact sheet," 2017.

[128] M. Ali, A. R. Nelson, A. L. Lopez, and D. A. Sack, "Updated global burden of cholera in endemic countries," *PLoS Neglected Tropical Diseases*, vol. 9, no. 6, pp. 1–13, 2015.

[129] T. E. Ford, R. R. Colwell, J. B. Rose, S. S. Morse, D. J. Rogers, and T. L. Yates, "Using satellite images of environmental changes to predict infectious disease outbreaks," *Emerging Infectious Diseases*, vol. 15, no. 9, pp. 1341–1346, 2009.

[130] B. Lobitz, L. Beck, A. Huq, B. Wood, G. Fuchs, A. S. G. Faruque, and R. Colwell, "Climate and infectious disease: Use of remote sensing for detection of Vibrio cholerae by indirect measurement," *Proceedings of the National Academy of Sciences*, vol. 97, no. 4, pp. 1438–1443, 2000.

[131] A. S. Jutla, A. S. Akanda, and S. Islam, "A framework for predicting endemic cholera using satellite derived environmental determinants," *Environmental Modelling and Software*, vol. 47, pp. 148–158, 2013.

[132] A. Jutla, A. S. Akanda, A. Huq, A. S. G. Faruque, R. Colwell, and S. Islam, "A water Marker monitored by satellites to predict seasonal endemic cholera," *Remote Sensing Letters*, vol. 4, no. 8, pp. 822–831, 2013.

[133] F. Finger, A. Knox, E. Bertuzzo, L. Mari, D. Bompangue, M. Gatto, I. Rodriguez-Iturbe, and A. Rinaldo, "Cholera in the Lake Kivu region (DRC): Integrating remote sensing and spatially explicit epidemiological modeling," *Water Resources Research*, vol. 50, no. 7, pp. 5624–5637, 2014.

[134] C. T. Codeço, "Endemic and epidemic dynamics of cholera: the role of the aquatic reservoir," *BMC Infectious Diseases*, vol. 1, no. 1, p. 1, 2001.

[135] Y. Grad, J. C. Miller, and M. Lipsitch, "Challenges to quantitative modelling of cholera disease transmission," *The quarterly update on epidemiology from the South African Centre for Epidemiological Modelling and Analysis (SACEMA)*, 2012.

[136] E. Bertuzzo, R. Casagrandi, M. Gatto, I. Rodriguez-Iturbe, and A. Rinaldo, "On spatially explicit models of cholera epidemics." *Journal of the Royal Society, Interface / the Royal Society*, vol. 7, no. 43, pp. 321–33, 2010.

[137] A. Rinaldo, E. Bertuzzo, L. Mari, L. Righetto, M. Blokesch, M. Gatto, R. Casagrandi, M. Murray, S. M. Vesenbeckh, and I. Rodriguez-Iturbe, "Reassessment of the 2010-2011 Haiti cholera outbreak and rainfall-driven multiseason projections," *Proceedings of the National Academy of Sciences*, vol. 109, no. 17, pp. 6602–6607, 2012.

[138] E. Bertuzzo, S. Azaele, A. Maritan, M. Gatto, I. Rodriguez-Iturbe, and A. Rinaldo, "On the space-time evolution of a cholera epidemic," *Water Resources Research*, vol. 44, no. 1, pp. 1–8, 2008.

[139] E. Bertuzzo, L. Mari, L. Righetto, M. Gatto, R. Casagrandi, M. Blokesch, I. Rodriguez-Iturbe, and A. Rinaldo, "Prediction of the spatial evolution and effects of control measures for the unfolding Haiti cholera outbreak," *Geophysical Research Letters*, vol. 38, no. 6, pp. 1–5, 2011.

[140] L. Mari, E. Bertuzzo, L. Righetto, R. Casagrandi, M. Gatto, I. Rodriguez-Iturbe, and A. Rinaldo, "Modelling cholera epidemics: the role of waterways, human mobility and sanitation," *Journal of The Royal Society Interface*, vol. 9, no. 67, pp. 376–388, 2012.

[141] D. L. Chao, M. E. Halloran, and I. M. Longini, "Vaccination strategies for epidemic cholera in Haiti with implications for the developing world," *Proceedings of the National Academy of Sciences*, vol. 108, no. 17, pp. 7081–7085, 2011.

[142] J. D. Clemens, "Vaccines in the time of cholera," *Proceedings of the National Academy of Sciences*, vol. 108, no. 21, pp. 8529–8530, 2011.

[143] R. P. Sanches, C. P. Ferreira, and R. A. Kraenkel, "The Role of Immunity and Seasonality in Cholera Epidemics," *Bulletin of Mathematical Biology*, vol. 73, no. 12, pp. 2916–2931, 2011.

[144] T. Bakhtiar, "Optimal intervention strategies for cholera outbreak by education and chlorination," *IOP Conference Series: Earth and Environmental Science*, vol. 31, no. 1, 2016.

[145] C. Yang, X. Wang, D. Gao, and J. Wang, "Impact of Awareness Programs on Cholera Dynamics: Two Modeling Approaches," *Bulletin of Mathematical Biology*, vol. 79, no. 9, pp. 2109–2131, 2017.

[146] M. Al-arydah, A. Mwasa, J. M. Tchuenche, and R. J. Smith, "Modeling Cholera Disease With Education and Chlorination," *Journal of Biological Systems*, vol. 21, no. 04, p. 1340007, 2013.

[147] A. R. Tuite, J. Tien, M. Eisenberg, D. J. D. Earn, J. Ma, and D. N. Fisman, "Cholera epidemic in Haiti, 2010: Using a transmission model to explain spatial spread of disease and identify optimal control interventions," *Annals of Internal Medicine*, vol. 154, no. 9, pp. 593–601, 2011.

[148] J. B. Njagarah and F. Nyabadza, "Modelling Optimal Control of Cholera in Communities Linked by Migration," *Computational and Mathematical Methods in Medicine*, vol. 2015, 2015.

[149] D. Posny, J. Wang, Z. Mukandavire, and C. Modnak, "Analyzing transmission dynamics of cholera with public health interventions," *Mathematical Biosciences*, vol. 264, no. 1, pp. 38–53, 2015.

[150] M. R. Kelly, J. H. Tien, M. C. Eisenberg, and S. Lenhart, "The impact of spatial arrangements on epidemic disease dynamics and intervention strategies," *Journal of Biological Dynamics*, vol. 10, no. 1, pp. 222–249, 2016.

[151] E. F. Camacho and C. Bordons, *Model Predictive control*, ser. Advanced Textbooks in Control and Signal Processing. London: Springer London, 2007.

[152] S. L. Noble, L. E. Wendel, M. M. Donahue, G. T. Buzzard, and A. E. Rundell, "Sparse-Grid-Based Adaptive Model Predictive Control of HL60 Cellular Differentiation," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 2, pp. 456–463, 2012.

[153] J. P. Perley, J. Mikolajczak, M. L. Harrison, G. T. Buzzard, and A. E. Rundell, "Multiple Model-Informed Open-Loop Control of Uncertain Intracellular Signaling Dynamics," *PLoS Computational Biology*, vol. 10, no. 4, 2014.

[154] A. O. Hero and D. Cochran, "Sensor Management: Past, Present, and Future," *IEEE Sensors Journal*, vol. 11, no. 12, pp. 3064–3075, 2011.

[155] S. Joshi and S. Boyd, "Sensor selection via convex optimization," *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 451–462, 2009.

[156] S. P. Chepuri and G. Leus, "Sensor selection for estimation, filtering, and detection," in *2014 International Conference on Signal Processing and Communications (SPCOM)*, vol. 26, no. 5. IEEE, 2014, pp. 1–5.

[157] X. Shen and P. K. Varshney, "Sensor selection based on generalized information gain for target tracking in large sensor networks," *IEEE Transactions on Signal Processing*, vol. 62, no. 2, pp. 363–375, 2014.

[158] X. Shen, S. Liu, and P. K. Varshney, "Sensor selection for nonlinear systems in large sensor networks," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 50, no. 4, pp. 2664–2678, 2014.

[159] P. Costa, J. Dunyak, and M. Mohtashemi, "Models, prediction, and estimation of outbreaks of infectious disease," *Proceedings. IEEE SoutheastCon, 2005.*, pp. 1–5, 2005.

[160] L. M. A. Bettencourt, R. M. Ribeiro, G. Chowell, T. Lant, and C. Castillo-Chavez, "Towards Real Time Epidemiology: Data Assimilation , Modeling and Anomaly Detection of Health Surveillance Data Streams," *Intelligence and Security Informatics: Biosurveillance*, pp. 79–90, 2007.

[161] J. Shaman and A. Karspeck, "Forecasting seasonal outbreaks of influenza," *Proceedings of the National Academy of Sciences*, vol. 109, no. 50, pp. 20 425–20 430, 2012.

[162] W. Yang, A. Karspeck, and J. Shaman, "Comparison of Filtering Methods for the Modeling and Retrospective Forecasting of Influenza Epidemics," *PLoS Computational Biology*, vol. 10, no. 4, 2014.

[163] L. Cobb, A. Krishnamurthy, J. Mandel, and J. D. Beezley, "Bayesian tracking of emerging epidemics using ensemble optimal statistical interpolation," *Spatial and Spatio-temporal Epidemiology*, vol. 10, pp. 39–48, 2014.

[164] W. Qian, N. D. Osgood, and K. G. Stanley, "Integrating epidemiological modeling and surveillance data feeds: A Kalman Filter based approach," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8393 LNCS, no. Mcmc, pp. 145–152, 2014.

[165] H. Gupta, K. K. Verma, and P. Sharma, "Using Data Assimilation Technique and Epidemic Model to Predict TB Epidemic," *International Journal of Computer Applications*, vol. 128, no. 9, pp. 975–8887, 2015.

[166] A. Naficy, M. R. Rao, C. Paquet, D. Antona, A. Sorkin, and J. D. Clemens, "Treatment and Vaccination Strategies to Control Cholera in Sub-Saharan Refugee Settings," *JAMA*, vol. 279, no. 7, p. 521, 1998.

[167] Global Task Force on Cholera Control, "Ending cholera - A global roadmap to 2030," World Health Organization, Tech. Rep., 2017.

[168] P. D. Christofides, R. Scattolini, D. Muñoz de la Peña, and J. Liu, "Distributed model predictive control: A tutorial review and future research directions," *Computers and Chemical Engineering*, vol. 51, pp. 21–41, 2013.

[169] A. Bemporad and D. Barcelli, "Decentralized model predictive control," *Lecture Notes in Control and Information Sciences*, vol. 406, pp. 149–178, 2010.

[170] X. Wang, D. Gao, and J. Wang, "Influence of human behavior on cholera dynamics," *Mathematical Biosciences*, vol. 267, pp. 41–52, 2015.

[171] A. K. Misra, S. N. Mishra, A. L. Pathak, P. Misra, and R. Naresh, "Modeling the effect of time delay in controlling the carrier dependent infectious disease - Cholera," *Applied Mathematics and Computation*, vol. 218, no. 23, pp. 11 547–11 557, 2012.

[172] A. K. Misra and V. Singh, "A delay mathematical model for the spread and control of water borne diseases," *Journal of Theoretical Biology*, vol. 301, pp. 49–56, 2012.

[173] M. Elhia, M. Rachik, and E. Benlahmar, "Optimal Control of an SIR Model with Delay in State and Control Variables," *ISRN Biomathematics*, vol. 2013, no. 50, pp. 1–7, 2013.

[174] K. Khan, S. J. McNabb, Z. A. Memish, R. Eckhardt, W. Hu, D. Kossowsky, J. Sears, J. Arino, A. Johansson, M. Barbeschi, B. McCloskey, B. Henry, M. Cetron, and J. S. Brownstein, "Infectious disease surveillance and modelling across geographic frontiers and scientific specialties," *The Lancet Infectious Diseases*, vol. 12, no. 3, pp. 222–230, 2012.

[175] R. Chunara, J. R. Andrews, and J. S. Brownstein, "Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak," *American Journal of Tropical Medicine and Hygiene*, vol. 86, no. 1, pp. 39–45, 2012.

[176] V. Marmara, A. Cook, and A. Kleczkowski, "Estimation of force of infection based on different epidemiological proxies: 2009/2010 Influenza epidemic in Malta," *Epidemics*, vol. 9, pp. 52–61, 2014.

APPENDICES

A. LIST OF PUBLICATIONS

Research reported in this thesis are in boldface.

Published

1. Aditya Sai, and Nan Kong, "Surrogate Modeling of Stochastic Dynamical Systems", Engineering Letters, vol. 26, no.1, pp 1-6, 2018.
2. Aditya Sai, and Nan Kong, "**Sparse Grid Interpolation of Itô Stochastic Models in Epidemiology and Systems Biology**", IAENG International Journal of Applied Mathematics, vol. 48, no.1, pp 45-52, 2018.

In revision

3. Aditya Sai, Carolina Vivas-Valencia, and Nan Kong, "A Comparative Study for Using Active Learning in Calibrating Disease Simulation Models", *In revision*, 2018.

In submission

4. Aditya Sai, and Nan Kong, "**Parameter Estimation in Epidemiology using Sparse Grid Interpolation**", *In submission*, 2018.
5. Aditya Sai, and Nan Kong, "Exploring the Information Content of Noise-induced Dynamics in Glioma Differentiation using Surrogate Modeling", *In submission*, 2018.

In preparation

6. Aditya Sai, and Nan Kong, "**Optimal Multi-period Point of Care Sensor Selection for Epidemic Modeling and Control**", *In preparation.*
7. Carolina Vivas-Valencia, Aditya Sai, and Nan Kong, "Characterization of Colorectal Cancer Progression Parameters Between Men and Women using a Discrete-Event Simulation", *In preparation.*
8. Aditya Sai, Thembi Mdluli, Gregery Buzzard, and Ann Rundell, "Nexperiment: Computational Platform for Model-based Design of Experiments to Reduce Dynamical Uncertainty", *In preparation.*

VITA

VITA

Aditya Prakash Sai was born in Staten Island, NY, and spent his childhood in Edison, NJ. He graduated from Rutgers, The State University of New Jersey, in May 2013 with a Bachelor of Science in Biomedical Engineering and Computer Science with a minor in Mathematics. He then joined the research lab of Dr. Ann Rundell in the Weldon School of Biomedical Engineering at Purdue University in 2013, serving as a research assistant in the Rundell lab from 2013 to 2016. He is currently a Ph.D. candidate in the Biomedical Analytics and Systems Optimization lab, advised by Dr. Nan Kong. His research interests include data science, computational modeling, machine learning, and algorithm development.