UNIVERSITY OF CALIFORNIA
SANTA CRUZ

EXTENSIONS OF EMPIRICAL DYNAMIC MODELING FOR PREDICTION AND MANAGEMENT IN ECOLOGICAL SYSTEMS

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

APPLIED MATHEMATICS

by

Bethany Johnson

December 2022

The Dissertation of Bethany Johnson
is approved:

Dr. Marcella Gomez, Chair

Dr. Stephan B. Munch

Dr. Hongyun Wang

Dr. George Sugihara

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Extensions of empirical dynamic modeling for prediction and management in ecological systems

by

Bethany Johnson

Humans simultaneously depend on and affect the health of natural ecosystems on a global scale, so it is important to establish ecosystem management practices that will ensure longevity and mutualism in the relationship between humans and nature. For decades, scientists have worked in a single-species paradigm to inform most management decisions in ecology. Specifically, species have traditionally been modeled and assessed individually, with limited consideration of how they interact with other species and drivers in their ecosystems. This has led to inaccurate predictions in the past, so there has been a recent push to account for more complexity in ecological models, as this would facilitate better management decisions. While one natural extension is to incorporate multiple variables into mechanistic models, this is challenging and inefficient with our current understanding of ecosystems. Alternatively, data-driven models offer a way to predict population dynamics without requiring specific inputs for all ecosystem components.

In this dissertation, we explore empirical dynamic modeling, a data-driven approach to forecasting which is derived from principles of dynamical systems theory. Empirical dynamic modeling is a promising tool that accounts for system complexity without requiring strong assumptions or full system observations. However, it cannot cope with some limitations that are common in ecological datasets, including short time series and missing samples. Thus, we develop extensions of empirical dynamic modeling to address these limitations. We then apply this approach along with optimal control methods to generate management decisions in ecological pest control scenarios. Throughout the dissertation, we demonstrate the effectiveness of our method developments on a wide range of simulated data examples in addition to empirical data from high-impact terrestrial and aquatic ecosystems.

To my family

Mom, Dad, Genevieve, and Olivia

Acknowledgments

This dissertation includes reprints of published material. The coauthors in those publications supervised the research that forms the basis for the dissertation. Specific acknowledgements for each project are contained within the corresponding chapters.

My achievements throughout graduate school would not have been possible without the support of my mentors, friends, and family.

Above all, I would like to thank my advisors who made my graduate experience so positive. Steve Munch – I appreciate your endless patience and guidance over the last five years. You pushed me to become a better researcher and encouraged me to be proud of my work. I am truly grateful for everything I've learned from you over the years. Marcella Gomez – thank you for supporting me through every step of this process and for always believing in me. You inspire me in more ways than you know, and I am thankful for everything you have done to challenge and encourage me. A big thanks to my committee members, Hongyun Wang and George Sugihara, as well, for graciously giving me their time and insightful feedback as I worked on this research.

Thank you to the current and former members of the Munch lab. A special thank you must go to Tanya Rogers for showing me what it means to be a detail-oriented and passionate researcher. You are truly inspiring, and I am grateful to have worked with you. Tara Dolan, Uttam Bhat, Chenghan Tsai, Antoine Brias, Kenneth Gee, Burleigh Charlton, Dylan Esguerra, Vadim Karatayev, Darian Sorenson, Lucas Medeiros, and Vicki Sheng – you all motivate me with your work ethic, desire to learn, and overall positive attitudes. Thank you for making our weekly lab meetings thought-provoking and enjoyable over the years.

I am overwhelmed with gratitude for the wonderful friends I made in my graduate program. Sara Nasab, Martin Rodriguez, Tenavi Nakamura-Zimmerer, Giovanny Marquez, Kenneth Caluya, Paul Wintz, Jacob Noone Wade, George Labaria, Evan Burton, Bram Rodgers, and Ian May – thank you for making our department feel like a family. I have learned so much from each of you, and I'll always cherish the memories from our classes, social hours (even the virtual ones), climbing days, and bonfire nights.

I recently had the privilege of working with a brilliant team at X, the Moonshot Factory. Thank you to all of my teammates at X for helping me improve as a scientist, communicator, and collaborator, and for always encouraging me to prioritize my Ph.D. work even when we had deadlines to meet. You all made me laugh harder than I thought was possible.

Thank you, Steven Haug, for keeping me company through my entire graduate school journey. You believed in me even when I did not believe in myself, and I could not have done this without your love and support.

Finally, thank you to my one-of-a-kind family. Mom, Dad, Genevieve, and Olivia – somehow you have helped me stay (almost) sane throughout this journey. Thank you for encouraging me to have the confidence to keep going and accomplish my goals.

Chapter 1

Introduction

Humans are highly dependent on nature. More than half of the global gross domestic product relies on food, medicine, and materials sourced from natural ecosystems (Herweijer et al., 2020). The fisheries and agricultural industries are some of the primary examples. Every year, fisheries generate over \$200 billion (Dyck and Sumaila, 2010) and provide more than 15% of animal protein consumed worldwide (Béné et al., 2015), and farming generates over \$100 billion annually and supports approximately 2.6 million jobs in the United States alone (USDA, 2021).

While maximizing economic yield or minimizing loss are often the primary objectives considered by decision makers in these industries, decision makers should also ensure that policies avoid exploiting ecosystems in unsustainable ways and avoid taking actions that could interfere with non-focal ecosystem services. The best way to achieve this is through the development of sustainable management practices for natural ecosystems.

Traditional approaches to management in ecology have leaned on a single-species framework. Specifically, decisions have historically been based on simplified models of population dynamics, which do not adequately capture various aspects of ecosystems, including species interactions, environmental drivers, and anthropogenic effects. This restrictive framework has led to costly failures in numerous fields ranging from fisheries collapse (Skern-Mauritzen et al., 2016) to poisoning of non-target organisms after pesticide applications (Beketov et al., 2013; van der Werf, 1996) to "threat cascades," which have involved the accidental suppression of beneficial species through elimination of invasive ones (Bergstrom et al., 2009; Geary et al., 2019; Zavaleta et al., 2001). In light of this, over the past several decades, some scientists have claimed that more accurate ecological forecasting could resolve most of the current problems in management (Clark, 2001; Dietze, 2017), which has created a strong push to make ecology a more predictive science (Dietze et al., 2018; Evans et al., 2012) with the development of complex ecosystem modeling techniques (Christensen et al., 1996; Pikitch, 2004).

Ecosystem models aim to take a holistic approach by incorporating many ecosystem components and processes into mathematical models. However, the natural world makes ecosystem modeling challenging (Geary et al., 2020; Sugihara et al., 1984). Nonlinear interactions between numerous species and exogenous variables occur at different time scales ranging from days to decades. While some scientific fields have validated laws to abide by when constructing mathematical models, efficient theories derived from first-principles are somewhat sparse in ecology (Marquet et al., 2014). Realistically, we have limited mechanistic understanding of how climate inputs, trophic network interactions, anthropogenic actions, genetic variability, and other factors influence each other. Furthermore, we typically only have data for a few of these variables on short time scales. This leads to strong uncertainty in model structures and parameters, so ecosystem models often involve numerous simplifying assumptions. Small changes in these assumptions can lead to large differences in model predictions and the management advice they generate (Wood and Thomas, 1999).

While there has been notable recent progress in quantitative ecosystem modeling, some scientists have pointed out that ecological theory has produced minimal practical value and many developments have yet to meet practice (Arkema et al., 2006; Mcevoy, 2017). As a result, the same issues have existed for over sixty years in fisheries management (Maunder and Piner, 2015) and responses to ecological events are often still reactive in practice, rather than carefully anticipated and controlled (Oliver and Roy, 2015). Due to these issues, ecosystem modeling and management has been classified as a "wicked problem" – a seemingly intractable problem with no generalizable solution (DeFries and Nagendra, 2017).

In light of this, there is a need to establish methods that provide suitable and practical contributions to ecological management. While there is certainly a great deal of "wickedness" in ecosystem management, it may be possible to take steps toward a more generalizable solution by developing flexible methods for ecological prediction and control that abandon the need for complete mechanistic understanding of ecosystem complexity.

1.1 Empirical dynamic modeling

As an alternative to ecosystem models, empirical dynamic modeling (EDM) (Sugihara, 1994; Sugihara and May, 1990; Ye et al., 2015a) is a framework that allows us to work efficiently and directly with empirical data without requiring strong assumptions. Each chapter in this dissertation uses EDM and contains a description of the method, but we provide a more thorough overview here for pedagogical purposes.

EDM is grounded in Takens's theorem of time-delay embedding (Takens, 1981), which we paraphrase as: *Let M be a compact manifold of dimension m. Then for generic pairs (ϕ, y), where $\phi : M \to M$ is a diffeomorphism (i.e. a dynamical system describing the flow on M) and $y : M \to \mathbb{R}$ is a differentiable observation function (i.e. time series), the map $\Phi_{(\phi,y)} : M \to \mathbb{R}^{2m+1}$ given by*

$$\Phi_{(\phi,y)}(x) := \left(y(x), y(\phi(x)), y\left(\phi^2(x)\right), \ldots, y\left(\phi^{2m}(x)\right)\right)$$

(i.e. time lagged observations) is generically an embedding of M in $\mathbb{R}^{2m+1}$.

An *embedding* is a global one-to-one, smooth transformation of the manifold M that preserves all topological properties. Thus, this theorem implies that the image of the attractor of a multi-dimensional deterministic dynamical system can be reconstructed using the time series of a single state variable, $x_0, \ldots, x_T$. From the single-variable time series, *embedding vectors* of the form $[x_t, x_{t-1}, \ldots, x_{t-E}]^\top$, $t \in (E, \ldots, T)$ can used as synthetic coordinates to reconstruct the attractor, and all of the important mathematical features of the original attractor will hold in the reconstructed attractor.

As an illustrative example, consider a tri-trophic food-chain dynamical system given by

$$\begin{aligned} \dot{x} &= rx(1-x) - \frac{axy}{1+bx} \\ \dot{y} &= \frac{axy}{1+bx} - \frac{cyz}{1+dy} - fy \\ \dot{z} &= \frac{cyz}{1+dy} - gz \end{aligned} , \tag{1.1}$$

(Hastings and Powell, 1991). In an ecological context, these equations model the change in population size through time of a producer (x), a grazer (y), and a consumer (z). Figure. 1.1(a) shows the "teacup" attractor of this system, and

Fig. 1.1(b) shows the reconstructed attractor, which was constructed by taking two lags of the x time series and plotting the evolution of the system in the lagged-coordinate space. Importantly, the reconstruction clearly preserves the global "teacup" shape of the original attractor.

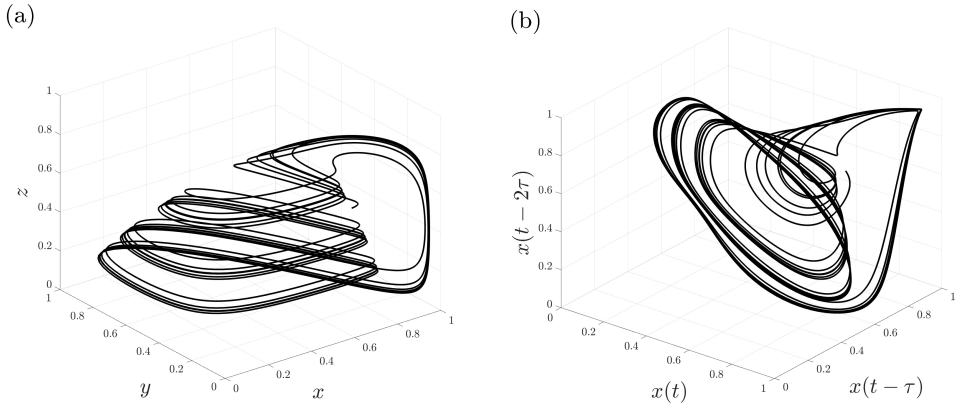

Figure 1.1: Illustration of Takens's theorem with the state space attractor of system (1.1) and its corresponding reconstructed attractor using time lags of variable x. With parameter values $r = 2$, $a = 10$, $b = 3$, $c = .2$, $d = 2$, $f = 0.8$, $g = 0.016$, we solved the system of ODEs using a fourth-order Runge-Kutta method with a time step 0.01. (a) The original trajectory of plotted in coordinates for each variable (x, y, and z) results in a "teacup" attractor. (b) The reconstructed attractor using time lags of x (lag parameter $\tau = 28$) as synthetic corrdinates. The reconstruction resembles the "teacup" shape of the original attractor.

Takens's theorem has widespread applications. Fundamentally, it allows any system analysis to be performed with only partial observations of the system. For instance, dynamical invariants (e.g. correlation dimension, Lyapunov exponents) of a system can be estimated using observations of just a single state variable (Sugihara, 1994). Furthermore, it ensures that the evolution of some local neighborhood of points in the original system maps directly to the evolution of a neighborhood of points in the reconstructed system. Thus, forecasts can be made based on observations of previous ecosystem states that were similar to the current state. This fundamentally means that forecasting with EDM is not simply an extrapolation of the recent temporal dynamics, but a guided prediction based on historical ecosystem trajectories.

More specifically, forecasting is possible because the theorem implies that the present value of a state variable is a function of its past values. That is, $x_t = f(\mathbf{x_{t-1}})$ where $\mathbf{x_{t-1}} = [x_{t-1}, \ldots, x_{t-E}]^\top$ for some unknown function f and embedding dimension E (Deyle et al., 2016; Perretti et al., 2013; Sugihara, 1994; Ye et al., 2015a). For a stochastic system, an analogous argument shows that this approach models the conditional mean $E(x_t|x_{t-1}, \ldots, x_{t-E}) = f(\mathbf{x_{t-1}})$ (Stark et al., 2003). Any non-parametric function approximation scheme can be used to approximate f. As a result, EDM allows us to predict nonlinear dynamics without requiring data from all variables or an explicit model structure. EDM is used in a variety of scientific fields and is increasingly being used in ecology to forecast population size (Deyle et al., 2013; Glaser et al., 2014a; Liu et al., 2012; Perretti et al., 2013), identify causal connections (Clark et al., 2015; Sugihara et al., 2012; Van Nes et al., 2015), quantify species interactions (Deyle et al., 2016; Rogers et al., 2020), and identify chaotic dynamics (Dixon et al., 1999; Hsieh et al., 2005; Rogers et al., 2022).

The primary benefits of empirical dynamic modeling are (1) it does not require specification of a model structure, which allows for more flexibility, as it can capture any patterns present in the historical data, (2) it allows us to work efficiently with partial observations of a system, and (3) it has been shown to outperform standard mechanistic models (Munch et al., 2018). There are, however, several important limitations of EDM. Since EDM aims to reconstruct dynamics based on historical data, the quality and quantity of data affect the accuracy of the method. Since many applications and developments of EDM have been in high quality, data-abundant fields, such as physics, there has not been much consideration for how to adapt the method when there are limitations in the data. One particularly problematic feature that is often encountered in ecological data is short time series. Another is that many ecological systems are sampled at irregular temporal intervals. This is problematic because Takens's theorem is valid only for uniformly sampled time series from a dynamical system. Both of these issues greatly limit the applicability of EDM in ecology.

1.2 Summary of contributions

In this dissertation we explore and develop extensions of the standard EDM framework to address some common issues that prevent EDM from being used in ecological applications. With these extensions, we aim to make EDM a more generalizable and universal tool for ecological forecasting. Since forecasting is only one step in ecosystem management, we also address how to make robust decisions in ecological systems by combining EDM forecasts with optimal control theory. In all of the chapters of this dissertation, our goal is to provide a proof of concept and an investigation of generalizability through simulation experiments, which represent a wide range of ecological scenarios. We use these simulation experiments to identify conditions in the underlying dynamics that might affect the accuracy of our forecasts and control recommendations. We also aim to provide additional proof of concept through several empirical examples, which include real ecological data for a range of species in high-impact industries including fisheries and agriculture.

In Chapter 2, we address the limitation of short time series in ecology in an analysis of how to leverage spatial information. Short time series are highly common, but most ecological surveys also contain data from multiple spatial locations. In the ecological literature, data from site-specific time series have been concatenated to effectively lengthen the time series and improve predictions. In the physics literature, however, it is more common to incorporate data from neighboring locations directly into the embedding vectors. We develop a physically-informed extension of this framework, which accounts for species dispersal. Using a series of simulations experiments, we evaluate the relative effectiveness of each of these approaches on short time series. We also apply the methods to empirical data for several commercially valuable fish species. Through this analysis, we establish general guidelines for coping with short time series within the EDM framework.

In Chapter 3, we address another major limitation, which is that EDM can only be applied to uniformly sampled time series data. This creates a particular challenge in ecological applications because there are many reasons for ecological data to be missing or sampled irregularly. For example, weather conditions, funding constraints, equipment malfunctions, and seasonal sampling can all contribute to irregularity in time series. With this in mind, we develop a framework that allows EDM to operate robustly on irregularly sampled time series or series with missing data. We evaluate the framework in numerous scenarios of missing data through simulations, and we also demonstrate the method on empirical data for an insect pest which damages grain and potato crops.

While the standard EDM framework allows us to efficiently cope with single-species data, in many ecological applications, species are influenced by harvesting or other anthropogenic controls. When we have historical data for these controls, incorporating them as inputs in the EDM framework can improve prediction accuracy and opens up the possibility of using these tools for quantitative ecosystem management (Boettiger et al., 2015; Brias and Munch, 2021; Giron-Nava et al., 2017; Liu et al., 2012). In Chapter 4, we explore connections between EDM forecasts and stochastic dynamic programming in an insect pest management setting. In this analysis, we quantify how much improvement, if any, is possible by using EDM forecasting with optimal control theory rather than status quo approaches in this concrete example. We explore several scenarios of control through simulations and further validate the methods on empirical data for an agricultural pest and an insect disease vector.

We summarize the proposed method developments and their applications in Chapter 5. We also address limitations of our methods and extensions of this work that we have left for future studies. Several supplemental analyses and additional supporting information about the main chapters are available in the Appendix.

We intentionally made each chapter in this dissertation independent by including all relevant background information and method descriptions within each chapter. Accordingly, there may be some repetition across chapters, but readers are welcome to approach the chapters out of order.

Chapter 2

Leveraging Spatial Information with Empirical Dynamic Modeling

This chapter is a reprint of Johnson et al. (2021) as it appears in Methods in Ecology and Evolution (Volume 12, Issue 2, October 2020). Stephan B. Munch and Marcella Gomez are coauthors of this paper. The dissertation author was the primary investigator and author of this paper.

2.1 Abstract

1. There has been a recent demand for forecasting in ecology, particularly in the field of ecosystem management. Empirical dynamic modelling (EDM), an equation free nonlinear forecasting method, is receiving growing attention, but it requires long time series to produce accurate predictions. Though

most ecological time series are short, spatial replicates are often available. Here we explore how utilizing available spatial data can improve our ability to forecast ecological dynamics.

2. There are several ways to incorporate spatial information into EDM and not all have been applied in ecology. We compare spatial EDM approaches used in ecology and physics and introduce a flexible Bayesian model that makes use of prior movement information.

3. We test these methods on simulated data generated with three population dynamics models with varying levels of complexity, time series length, spatial symmetry and heterogeneity. Adding spatial data generally improves accuracy, though the best method depends on the spatial process. We applied the methods to empirical fisheries data, highlighting the complexity of real population dynamics.

4. Leveraging spatial data is an effective way to overcome the problem of short ecological time series. Since the best forecasting method depends on the underlying dynamics, we suggest that users apply several in concert and that this may be useful in identifying spatial heterogeneity in dynamics.

2.2 Introduction

Over the past three decades there have been numerous calls to make ecology a more predictive science (Clark, 2001; Dietze, 2017; Dietze et al., 2018; Evans et al., 2012). One of the most pressing demands for ecological forecasting is in the field of ecosystem management (Christensen et al., 1996; EPAP, 1999; Pikitch, 2004), which accounts for interactions among competing ecosystem services when setting management policy. Unfortunately, in most ecosystems, we do not have complete knowledge of all the system variables and their interactions. In such partially observed systems, nonlinear interactions between numerous species and exogenous variables make mechanistic predictions difficult (Wood and Thomas, 1999).

Alternatively, equation-free approaches, collectively called empirical dynamic modelling (EDM) (Sugihara and May, 1990; Takens, 1981; Ye et al., 2015a), can be used to make predictions in partially observed systems. The state variables of a dynamical system evolve through time and tend to converge on a set of values (e.g. a fixed point, limit cycle or complex shape) referred to as the attractor for the system (Chang et al., 2017). Takens' theorem (Takens, 1981) demonstrates that time lags of a single variable can be used as synthetic coordinates to reconstruct an image of the attractor. In an ecological context, this means that it is possible to use historical data from a single species to reconstruct the dynamics even when we do not have data for other species in the community. In practice, this is done by constructing a collection of short segments of the time series, $\{x_{t-E}, x_{t-E+1}, \ldots, x_t\}$ and using this 'library' to identify similar segments. If the segments are long enough and the dynamics are smooth, similar segments imply similar future states allowing robust, model-free predictions to be made. See Movie S1 of (Ye et al., 2015a) for an explanation. EDM is increasingly used in ecology to forecast population size (Perretti et al., 2013), identify causal connections (Sugihara et al., 2012), quantify species interactions (Deyle et al., 2016; Rogers et al., 2020), and identify chaotic dynamics (Sugihara, 1994; Sugihara and May, 1990). However, EDM requires long time series at relatively short sampling intervals in order to fully resolve the attractor. This often limits the accuracy of EDM in ecology where many time series are fairly short (20-50 years).

There are several avenues to overcome the problem of short time series. For instance, libraries from several state variables (i.e. species) can be concatenated to increase the density of points on a common attractor (Banbrook et al., 1997; Hsieh et al., 2008; Sugihara, 1994). This method improves predictions, but the obvious trade-off is that it requires data for multiple species. In cases where data for only one species are available from multiple locations, it may be possible to use the available spatial information and improve forecasts.

The state of the art for incorporating spatial information into EDM varies by discipline. In ecology, it is common to concatenate data from multiple sites to fill in the attractor more completely (Clark et al., 2015; Glaser et al., 2014b). Hierarchical approaches have also been used to combine information across sites (Munch et al., 2017; Rogers and Munch, 2020). Although these methods often improve forecasts, they rely on two opposing requirements. First, the spatial replicates must be similar enough that their dynamics are mutually informative. Second, the replicates must be different enough to provide new information. For instance, highly synchronized replicates contribute very little. These methods use only lags of the local time series to construct an average attractor across multiple sites. However, nearby populations can exchange migrants or share common drivers. In this case, the dynamics at a focal site depend on the state of its neighbours, so lags from neighbouring sites may provide additional information for forecasts.

Spatial EDM (sEDM), uses both spatial neighbours and temporal lags to leverage this spatial coupling. sEDM has been used in a variety of fields since its introduction (Parlitz and Merkwirth, 2000; Ørstavik and Stark, 1998) including forecasting coastline changes (Grimes et al., 2015) and sunspots (Covas, 2017; Covas and Benetos, 2019). Although sEDM seems like a promising avenue for forecasting in ecology, there is, to our knowledge, no ecological application of it, and it has not been compared to the method of concatenating data. Here we aim to determine which of these approaches more effectively utilizes spatial information to improve ecological forecasts.

Since sEDM is typically applied outside ecology, there are three limitations to address to make it suitable for ecological forecasting. First, sEDM applications usually construct predictions for each focal location independently. That is, although sEDM inputs include lags of the focal site and neighbours, the library of data used to make predictions is based on predictor–target pairs centred at the focal site. Including predictor–target pairs centred at other locations by concatenating libraries (Clark et al., 2015; Glaser et al., 2014a) could possibly improve sEDM predictions. This would simultaneously account for dynamic influences of dispersal or gradients and also borrow information across locations on the shape of the attractor.

Second, standard sEDM implementation makes simplifying assumptions including symmetric coupling between sites and identical dynamics across sites. Since many ecological systems break these assumptions (e.g. with advection and spatial heterogeneity (Barnett et al., 2019; Kolasa et al., 1991; Largier, 2003)), these assumptions should be relaxed when evaluating these methods.

Finally, there is no consensus on how to choose which neighbours and lags to include in attractor reconstructions with sEDM. Mutual information and false nearest neighbours can be used (Abarbanel, 1997; Covas, 2017; Covas and Benetos, 2019; Kantz and Schreiber, 2004), or combinations of neighbours and lags may be selected to minimize prediction error on some subset of training data (Bialonski et al., 2015). However, these methods force the spatial and temporal lags to be equally spaced, and a more flexible lag spacing may be preferable (Judd and Mees, 1998).

Automatic relevance determination (ARD; (MacKay and Neal, 1994; Neal, 1996)), a technique used in machine learning, is useful for selecting relevant lags in Bayesian approaches to EDM (Munch et al., 2017). However, the multivariate embedding theorem (Deyle and Sugihara, 2011) asserts that, from a theoretical perspective, sEDM is inherently non-identifiable. Here, it is important to distinguish between model identifiability, a well-known statistical problem, and dynamical identifiability which arises because there are multiple ways to exactly reconstruct dynamics (Deyle and Sugihara, 2011). For instance, in a predator–prey system, we could write a model using either the current densities of both species, lags of the predator or lags of the prey. All three formulations are equally valid and produce identical dynamics. In contrast, statistical identifiability occurs whenever multiple parameter combinations produce identical values for the likelihood and can often be eliminated by reparameterization. The lack of dynamical identifiability in sEDM implies that different lag selections are equally valid and will produce nearly equivalent fits, making it highly likely that a standard implementation of ARD would settle at a local maximum. This is an issue because settling at a local maximum in-sample could reduce out-of-sample predictive accuracy. Moreover, lack of identifiability implies that ARD may select different lags for two realizations of the same dynamics, reducing mechanistic interpretability of lag selections. Ideally, lag selections should be consistent across repetitions of equivalent dynamics.

To overcome this problem, we introduce a modification of ARD intended to resolve this identifiability issue. We use the physical network topology of a spatiotemporal system to derive a Bayesian prior for the expected relevance of each lag. By using this prior, ARD is more likely to converge to a collection of lags that are biologically and physically plausible, without restricting the inference to a fixed set.

Here we examine how to leverage spatial information with EDM and address how our modifications of sEDM compare to the standard implementation of sEDM and other EDM methods that are used in ecology. Specifically, we compare the performance of several combinations incorporating local lags (EDM) versus spatial lags (sEDM) and local versus concatenated libraries. For models with spatial lags, we compare performance with and without the informative prior. We present results on simulated data generated with population dynamics models and also on empirical data for several widely distributed marine species.

2.3 Methods

2.3.1 Empirical dynamic modeling and spatial extensions

EDM is grounded in Takens' theorem of time-delay embedding (Takens, 1981), which states that an image of the attractor of a multi-dimensional deterministic dynamical system can be reconstructed using the time series of a single state variable, $x_0, \dots, x_T$. From the single-variable time series, *embedding vectors* of the form $[x_t, x_{t-1}, \dots, x_{t-E}]^\top$ for $t \in (E, \dots, T)$ are used as synthetic coordinates to reconstruct the attractor. This is equivalent to saying that the present value of a state variable can be written as a function of its past values. That is, $x_t = f(\mathbf{x_{t-1}})$ where $\mathbf{x_{t-1}} = [x_{t-1}, \dots, x_{t-E}]^\top$ for some unknown function f and embedding dimension E.

As an intuitive example of why EDM should work, consider a predator-prey system in which the dynamics fall on a limit cycle. Given a single observation of the prey abundance, it is impossible to predict the abundance in next step because the observation can either be on a increasing or a decreasing part of the cycle. This ambiguity disappears if we augment the current prey abundance with any information on which side of the cycle we currently reside. Obviously, the current predator abundance is one possibility. The current direction of change in prey abundance (e.g. increasing or decreasing) would also be sufficient. Prey abundance a short time earlier provides the same information. Thus, time lags effectively allow us to reconstruct dynamics in partially observed coupled systems. See (Deyle et al., 2016; Perretti et al., 2013; Sugihara, 1994; Ye et al., 2015a) for more details on EDM in an ecological context.

For spatial extensions of EDM, consider a single-variable time series at N locations, $x_0^1, ..., x_T^1, ..., x_0^N, ...x_T^N$. To make use of these spatial replicates in EDM, the common approach in ecology is to stitch the N time series together and construct embedding vectors from the long concatenated time series (e.g. Hsieh et al. (2008)). This approach relies on the assumption that all replicates come from the same underlying attractor (Banbrook et al., 1997), so data from any site can inform dynamics of the others.

Alternatively, the idea of spatial EDM (Parlitz and Merkwirth, 2000; Ørstavik and Stark, 1998) is to use embedding vectors that include lags of both the focal location as well as its nearest spatial neighbors. Essentially, this says that the state at time t and site i is a function of the past values of the state in sites $i-S, \ldots, i, \ldots, i+S$, given by

$$x_t^i = f(\mathbf{x_{t-1}^{i-S}}, \ldots, \mathbf{x_{t-1}^{i}}, \ldots, \mathbf{x_{t-1}^{i+S}}), \tag{2.1}$$

for some unknown function f and embedding dimensions E and S. Here $\mathbf{x_{t-1}^i} = \left[x_{t-1}^i, \ldots, x_{t-E}^i\right]^\top$.

Regardless of the way spatial data are used, our goal is to approximate the function f from data so that it can be iterated forward to make predictions. There is a vast literature on function approximation techniques that can be used in this context (Iatan, 2016; Judd, 1999; Stalph, 2014), but the most commonly used in ecological literature are piece-wise constant models (e.g. Simplex (Sugihara and May, 1990)) , local linear regression (e.g. S-map (Sugihara, 1994)), splines (e.g. (Ellner and Turchin, 1995)), and Gaussian process (GP) regression (Munch et al., 2017). We implement EDM and sEDM via GP regression to facilitate incorporating auxiliary biological and physical information through prior specification (Thorson et al., 2014). Additionally, the GP framework admits a hierarchical structure to share information across different locations (Rogers and Munch, 2020), though we do not pursue this here.

Here we provide a brief introduction to GP regression, define a naive ARD prior, and then develop a novel ARD prior for sEDM. For additional background, Rasmussen and Williams provide an overview of modelling with GP regression (Rasmussen and Williams, 2006), and Munch and colleagues provide examples of using GP regression specifically for EDM (Munch et al., 2017).

In general, we want to fit a model for

$$x^i_{t+1} = f(\mathbf{x}) + \epsilon_t \tag{2.2}$$

where the approximation error, ϵ_t, has mean 0 and variance V. Note that the input x has a different structure for EDM and sEDM as described above. To infer f, we assume a GP prior and update using the observed time series. The GP is a continuous generalization of the multivariate normal distribution, completely specified by a mean function, $m(\mathbf{x}) = E[f(\mathbf{x})]$, and a covariance function, $c(\mathbf{x}, \mathbf{x}') = E[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$ (Rasmussen and Williams, 2006). Without information on the characteristics of f a priori, we use a constant prior mean function $m(\mathbf{x}) = 0$. Consistent with previous GP applications, we use a squared-exponential covariance function.

$$c(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^\top M(\mathbf{x} - \mathbf{x}')\right) \tag{2.3}$$

with

$$M = \begin{bmatrix} \phi_1 & 0 & \dots & 0 \\ 0 & \phi_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \phi_n \end{bmatrix}.$$

Here, ϕ_i is the inverse characteristic length scale and governs how much f varies in the direction of the i^{th} input (Rasmussen and Williams, 2006). Note that Equation (2.3) has $n+1$ free parameters, $\phi_1, ..., \phi_n, \sigma_f^2$. For σ_f^2, we use a weakly informed prior distribution, Beta$(1.1, 1.1)$. we use the same prior distribution for V, the variance in approximation error, which limits variance in the predictions to be less than twice the total variance in the observed data (assuming that the data have been scaled to unit standard deviation prior to analysis).Priors for ϕ are described below. We estimate these hyperparameters by maximizing the marginal likelihood using the resilient backpropagation (Rprop) algorithm (Riedmiller and Braun, 1993). Subsequent predictions are made conditional on these maximum a posteriori estimates (MAP; Rasmussen and Williams (2006)).

Note that when $\phi_i = 0$, f is constant in the direction of the i^{th} input. To encourage sparsity in f, previous GP-EDM applications (Munch et al., 2017; Rogers and Munch, 2020) have used identical half-normal priors for each ϕ, i.e.

$$p(\phi_j) = \sqrt{\frac{2}{\pi\gamma}} \exp\left(-\frac{\phi_j^2}{2\gamma}\right), \tag{2.4}$$

for all $1 \leq j \leq n$. A value of $\gamma = \frac{\pi}{2}$ is typically chosen so f will have a single extremum over a unit interval of the input on average (Sacks and Ylvisaker, 1966), which regularizes f and avoids over-fitting (Munch et al., 2017). While this prior has been successful for GP-EDM, setting identical priors does nothing to ameliorate identifiability issues in sEDM (Fig. 2.1(a)) and we refer to this model as sEDM with a naive prior. Next, we derive an informed prior for ϕ that makes use of the physical network to promote identifiability.

2.3.2 Physically-informed prior specification

In order to facilitate identifiability, we take advantage of the fact that some combinations of sites are more likely to have coupled dynamics than others. If sites are near each other, it is likely for them to be coupled through similar environmental conditions and dispersal. This coupling decreases for sites that are far apart, so we expect the length scale parameters to decay with distance. We modify the ARD framework to account for this by setting independent but informative half-normal priors for the length scale parameters allowing the mean for each to decrease with distance to the target. For diffusive coupling or, equivalently, nearest neighbour dispersal, the prior expected value of each ϕ is given by

$$E(\phi_{\delta,\tau}) \approx \frac{a_{\delta,\tau}}{\sqrt{2\sigma^2}} \tag{2.5}$$

where $a_{\delta,\tau} = \frac{1}{\sqrt{2\pi v\tau}} e^{-\frac{\delta^2}{2v\tau}}$. Here, δ and τ are the spatial and temporal distance from the target respectively, v is the mean-square dispersal distance, measured in whatever units are used to measure distance between sites and σ^2 is the variance of the data. See Appendix A.2 for a derivation of eq. (2.5). Since the expected value of the half-normal distribution is $E(\phi) = \sqrt{\frac{2\gamma}{\pi}}$, we set $\gamma = \frac{\pi}{2}(E(\phi))^2$ in eq. (2.4).

Figure 2.1(a) demonstrates the identifiability problem in naive sEDM by displaying length scale estimates for 20 independent but identical simulations of a two-species competition model with nearest neighbour dispersal (see Appendix

A.1 for model and parameter values). In each simulation, we generated 20 time points from random initial conditions on a one-dimensional lattice with seven sites. Using data from only one species, we included four time lags from each site in embedding vectors. ARD identified which of those 28 inputs were relevant for making predictions.

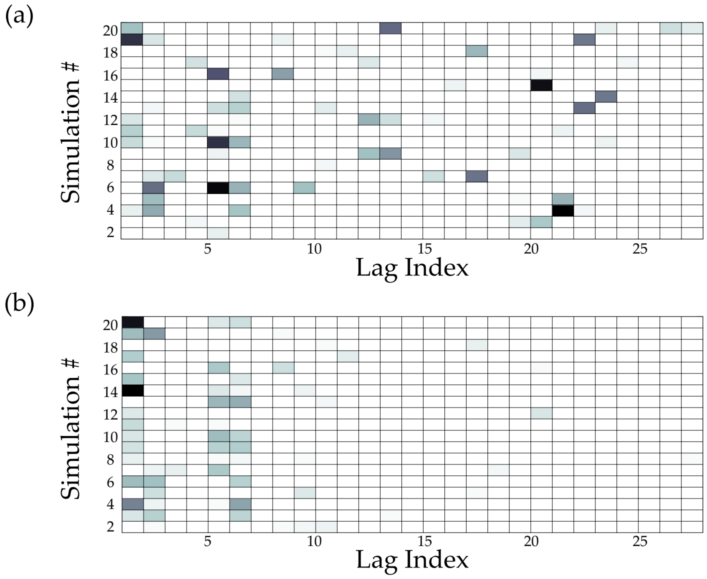

Figure 2.1: ARD outputs are shown for (a) naive sEDM and (b) informed sEDM. Each row corresponds to a single simulation of a two-species competition model with 7 spatial patches and 20 time points. sEDM framework included 28 'lags' in attractor reconstructions. Lag indices are ordered $x^i_{t-1}, \ldots, x^i_{t-4}, x^{i-1}_{t-1}, \ldots, x^{i-1}_{t-4}, x^{i+1}_{t-1}, \ldots, x^{i+1}_{t-4}, \ldots, x^{i+3}_{t-1}, \ldots, x^{i+3}_{t-4}$. Shaded lag indices were determined by ARD to be relevant for making predictions (darker colors indicate stronger relevance) and white indices were determined to be irrelevant. The strength and index of relevant lags is not consistent over the 20 simulations in (a), highlighting an identifiability issue. More agreement across simulations in (b) demonstrates the utility of the informed prior.

Since the dynamics for each simulation were generated by the same model, ARD should select the same inputs across simulations. However, consistent with the multivariate embedding theorem (Deyle and Sugihara, 2011) and earlier work (Kantz and Schreiber, 2004), the selected lags varied widely, indicating an identifiability problem. Although forecasting may or may not suffer, this lack of identifiability makes interpretation of ‘relevance’ based on the estimated length scales impossible. Figure 2.1(b) shows analogous outputs when the physically informed prior specification is used and highlights its ability to improve identifiability.

2.3.3 Local vs. spatial methods

Here we describe the distinction between the methods used. Fig. 2.2 shows the relationship between input embedding vectors and their outputs for EDM versus sEDM with and without an informed prior. We can construct multiple embedding vector/output pairs through time so that all N locations have a library of data available to train the GP. Figure 2.3 shows the different uses of these libraries.

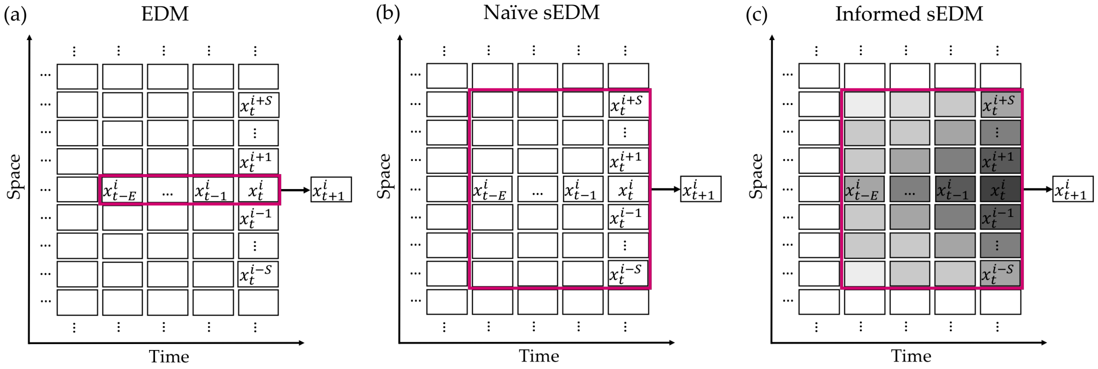

Figure 2.2: Embedding vectors that map to the one-step-ahead target at location i. The red box shows entries that are included in embedding vectors for EDM (a), spatial EDM with a naive prior (b), and spatial EDM with a physically-informed prior (c). The shading of entries in the embedding vectors represents the prior specification for length scales. Darker colors represent higher values of $\mathbb{E}(\phi)$.

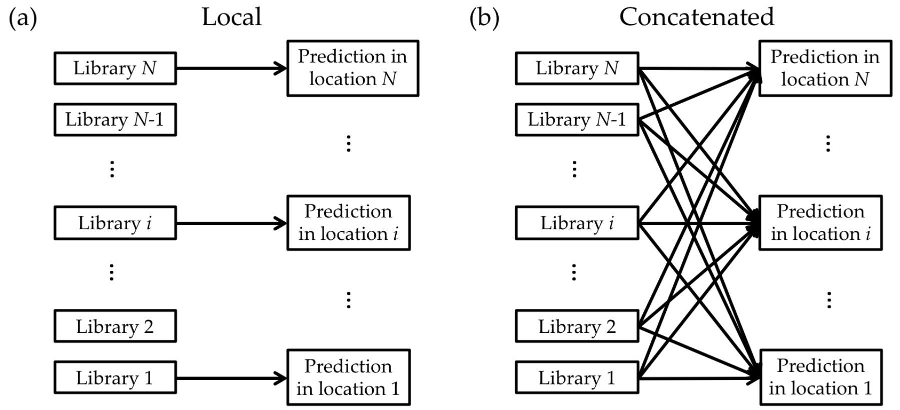

Figure 2.3: Libraries of embedding vectors used to train the GP and make predictions at location i. A local library uses only embedding vectors centered at the focal site (a). A concatenated library combines libraries of data across multiple sites to train the GP (b). Note that a concatenated library may combine data from all available locations or just a subset of locations.

2.3.4 Forecasting population dynamics

Here we compare six forecasting methods: EDM (local), EDM (concatenated), naive sEDM (local), naive sEDM (concatenated), informed sEDM (local) and informed sEDM (concatenated). We evaluate these on both simulated and empirical data.

Simulated data

Forecast performance of EDM methods are sensitive to a variety of factors including the length of time series, system dimension (e.g. number of coupled species), complexity of dynamics (e.g. periodic, chaotic, etc.) and spatial characteristics (e.g. dispersal, heterogeneity,etc.). To understand these factors, we varied them in three ecological models (Table 2.1). Importantly, these models cover a range of dimensions and ecological interactions to increase the generality of our results.

We used the Table 2.1 models to simulate spatiotemporal data on a one-dimensional lattice with periodic boundary conditions. For each simulation, we assigned random initial conditions to every variable in every site and removed the first 100 time points to avoid transients. We split the time series into 'training' data for fitting the GP hyperparameters, and 'testing' data for estimating the one-step-ahead forecast accuracy, out-of-sample. For all models, we used a lattice size of $N = 75$, test set of length $T_{test} = 20$. We used the abundance data for only one species, x, to train and test the methods and ignored the other species abundances.

The training time series length, T_{train}, varied by simulation. We repeated every model 100 times to produce summary statistics for forecast errors. GP regression has $\mathcal{O}(n^3)$ complexity stemming from a matrix inversion step that is required to compute the posterior predictive distribution. Hence, to ease the computational burden, we restricted our forecasting to a subset of locations (i.e. 10 randomly chosen sites used for both training and testing). Importantly, all methods used the same locations within each simulation, but locations varied across the repetitions.

Note that the informed sEDM prior (eq. 2.5) has one additional parameter v for the mean-square dispersal distance. This parameter is not easily estimated from data. To avoid choosing v arbitrarily, we fit the model with v fixed at either: $v = 10$ (i.e. short dispersal distance and steep decay in prior ϕ) or $v = 2^8 \cdot 10 = 2560$ (i.e. long dispersal and nearly identical prior). Note that $v = 2560$ was chosen arbitrarily but is large enough such that all lags are assigned similar priors, making the naive prior a special case of the informed prior. Importantly, users applying this method should consider the units of distance in their data and may need to adjust the candidate values of v. With both sets of hyperparameters corresponding to the different v values, we evaluated leave-one-out predictions on

Name	Model	Parameters
Ricker (1 species: x)	$f(x) = xe^{r(1-x)}$	$r = \{3, 3, 3.5\}$
Predator-Prey (2 species: x,y)	$f(x,y) = xe^{a-x-\frac{by}{(1+\alpha x)(1+\beta y)}}$ $g(x,y) = ye^{-c+\frac{dx}{(1+\alpha x)(1+\beta y)}}$	$a = \{2.4, 2.6, 2.8\}$ $b = 2$, $c = 0.1$, $d = 1.75$, $\alpha = 0.1$, $\beta = 0.1$
Host- Parasitoid- Parasitoid (3 species: x,y,z)	$f(x,y,z) = xe^{r(1-x/K)-ay^{-m+1}-bz^{-n+1}}$ $g(x,y,z) = x\left(1 - e^{-ay^{-m+1}-bz^{-n+1}}\right)$ $h(x,y,z) = x\left(1 - e^{-ay^{-m+1}-bz^{-n+1}}\right)$	$m = 0.7$, $n = 0.4$ $K = 50$, $a = 0.4$, $r = \{3.2, 3.2, 3.6\}$, $b = \{0.75, 0.84, 0.84\}$

Table 2.1: Model structure and parameter values for {periodic, chaotic (high spatial synchrony), and chaotic (low synchrony)} dynamics of a single-species Ricker model (Ricker, 1954), a two-species predator prey model (Zhang et al., 2018), and a three-species host-parasitoid-parasitoid model (Yu et al., 2009). we simulate spatiotemporal data with these models on a one-dimensional lattice with time series updates given by

$$x^i_{t+1} = \left[(1-\mu)f(\bullet^i_t) + \frac{\mu}{2}\left(f(\bullet^{i-1}_t) + f(\bullet^{i+1}_t)\right)\right]e^{\xi^i_t}$$

$$y^i_{t+1} = \left[(1-\mu)g(\bullet^i_t) + \frac{\mu}{2}\left(g(\bullet^{i-1}_t) + g(\bullet^{i+1}_t)\right)\right]e^{\xi^i_t}$$

$$z^i_{t+1} = \left[(1-\mu)h(\bullet^i_t) + \frac{\mu}{2}\left(h(\bullet^{i-1}_t) + h(\bullet^{i+1}_t)\right)\right]e^{\xi^i_t}$$

The big dot notation is given by $\bullet \equiv x$ (Model 1), $\bullet \equiv x$,y (Model 2), and $\bullet \equiv x, y, z$ (Model 3). μ is the nearest neighbor dispersal rate given by $\mu = \{0.1, 0.25, 0.25\}$ (Model 1), and $\mu = \{0.25, 0.25, 0.25\}$ (Model 2), $\mu = \{0.2, 0.3, 0.1\}$ (Model 3). For all simulations, the noise term was drawn independently from a normal distribution, i.e. $\xi^i_t \sim N(-s^2/2, s^2)$ with $s = 0.1$ (Model 1), $s = 0.05$ (Model 2), and $s = 0.1$ (Model 3).

the training data. The v associated with the lower prediction error on training data was used on testing data.

We measured forecast accuracy with the root mean squared error (RMSE) given by

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{j=1}^{T}\left(x^{i,j}_{\text{predicted}} - x^{i,j}_{\text{observed}}\right)^2}.$$

where the sum is over the testing data for each site. We chose to use the RMSE instead of Pearson's correlation coefficient—the criterion commonly used in ecological EDM studies—because the RMSE is more widely used in other disciplines and measures whether the absolute magnitude of the predictions and observations match. As a performance benchmark we calculated the RMSE obtained when using the mean of the training data to predict all points in the test data. We refer to this as the 'mean predictor'.

For both EDM and sEDM we used a temporal embedding dimension up to $E = \sqrt{T_{train}}$ (Cheng and Tong, 1992). In sEDM, we set the spatial embedding dimension to $S = 2$ (i.e. embedding vectors contain five locations) and used ARD to prune irrelevant inputs.

We used this setup in four simulation experiments. The first two addressed how dynamical complexity and time series length affect the forecast performance. These assumed symmetric dynamics (i.e. equal dispersal to nearest neighbours) and spatial homogeneity (i.e. identical parameter values for all sites). Since this may not apply for ecological dynamics (Barnett et al., 2019; Kolasa et al., 1991; Largier, 2003), two additional experiments addressed the impact of asymmetry and heterogeneity.

Effects of dynamical regime

We determined how dynamical complexity influences forecast performance by simulating each model in Table 3.1 with three parameter sets generating periodic dynamics, chaotic dynamics with high spatial synchrony and chaotic dynamics with low synchrony for $T_{train=25}$. See Table 3.1 for parameter values and Table A.1 in Appendix A.3 for synchrony measures.

Effects of time series length

To examine the effect of time series length on forecasts, we simulated each model using the parameters for chaotic (high synchrony) dynamics for time series lengths $T_{train} = 25, 50$ and 75.

Effects of asymmetric coupling

To test whether advection influences forecasts, we simulated dynamics with uni-directional dispersal. That is, updates were given by

$$x^i_{t+1} = \left[(1-\mu)f(\bullet^i_t) + \frac{\mu}{3}\left(f(\bullet^{i+1}_t) + f(\bullet^{i+2}_t) + f(\bullet^{i+3}_t)\right)\right]\epsilon^{\xi^i_t}. \tag{2.6}$$

Dynamics were simulated with $T_{train} = 25$ and chaotic (high synchrony) parameters. Figure 2.6 illustrates the difference between the two types of coupling compared.

Effects of spatial heterogeneity

To generate heterogeneity, we set a spatial gradient in the growth rate of x for each model (Model 1: r, Model 2: a, Model 3: r). Since simulations had periodic boundary conditions, the gradient was given by a sinusoidal function with period N. For example, for Model 1, we defined $r_i = H_r \sin\left(\frac{2\pi}{N}i\right) + r_0$ for each location i. The amplitude H_r scales the heterogeneity in growth rate, which we varied from 0 to 1. This measure of heterogeneity requires model-specific knowledge that would not be accessible in real applications. As a practical empirical measure, we used the spatial variance in mean abundance given by $H_m = \frac{\max \bar{x}_i - \min \bar{x}_i}{\min \bar{x}_i}$, where $\bar{x}_i$ is the average abundance of the species in site i over the time series. Models in this analysis were simulated with $T_{train} = 25$ and chaotic (high synchrony) parameters.

Empirical data

To further evaluate the forecasting utility of these methods, we compared them on data for several marine species from the northeast United States continental shelf. The NOAA Northeast Fisheries Science Center (NEFSC) fall bottom trawl survey has sampled from Cape Hatteras, North Carolina to the Gulf of Maine since 1963 (Politis et al., 2014). To avoid any effects of survey design changes, we restricted our analysis to offshore strata from 1973 to 2008. We predicted dynamics of three different species: longfin squid (*Loligo pealeii*), silver hake (*Merluccius bilinearis*) and butterfish (*Peprilus triacanthus*). These species were chosen because they are widely distributed and have short generation times.

We compared forecast performance for data aggregated at both coarse and fine spatial resolutions. The coarse resolution aggregated the data into four major regions: Mid-Atlantic Bight, Southern New England, Georges Bank and Gulf of Maine. These regions are grouped to have similar biophysical characteristics and have been used in previous studies (Lucey and Nye, 2010; Nye et al., 2009; Walsh et al., 2015). The fine resolution used the survey strata as spatial sites. Figure 2.9 shows the midpoint of all strata and their corresponding major regions. At both resolutions, we used the Euclidean distance between midpoints as δ in Equation (2.5). We selected a temporal embedding dimension between $E = 5$ and $E = 7$ depending on which minimized the average forecast errors. At the coarse resolution, lags from the nearest site to the focal site were included in the sEDM embedding vectors and libraries from all regions were used in the concatenated methods. In the fine resolution analysis, the embedding vectors used lags from the two strata closest to the focal site, and concatenated methods included libraries from sites within the same major region as the focal site.

At both resolutions, population abundance was estimated by averaging the number of individuals in the catch over all tows in the site each year. We computed sequential forecasts (i.e. forecasts that only used data earlier in the time series to predict the next point) through the entire time series and reported errors on the last 10 time series points. We predicted log population abundance, $\ln(x_{t+1})$, for all species.

2.4 Results

2.4.1 Results on simulated data

Overall, simulations show that utilizing spatial information in EDM is advantageous. However, the characteristics of the dynamics influence which method most effectively leverages spatial information.

Effects of Dynamical Regime

Figure 2.4 shows the mean and standard deviation RMSE of one-step-ahead forecasts of the Table 2.1 models with three levels of complexity. Even with a short training time series of $T_{train} = 25$, concatenating spatial replicates provides accurate predictions (75% average error reduction from the mean predictor), and these methods consistently outperform local EDM, which ignores spatial data. Alternatively, including spatial lags without concatenating libraries (local sEDM) does not improve prediction over local EDM, and these methods are more sensitive to dynamical complexity; forecast errors increase as dynamics become more complex and less synchronized. See Table A.2 for RMSE and variance explained (R^2) for greater interpretability.

Dynamical complexity plays a role in the forecasting ability of these methods but does not qualitatively change their performance compared to one another.

Complexity primarily influences how much improvement is made by using spatial data. For instance, when dynamics are simple (e.g. periodic) all methods perform well and concatenated libraries only reduce forecast errors from local libraries by about 20on average, but concatenated libraries reduce errors by much more (65%) when the dynamics are complex (i.e. chaotic and asynchronous).

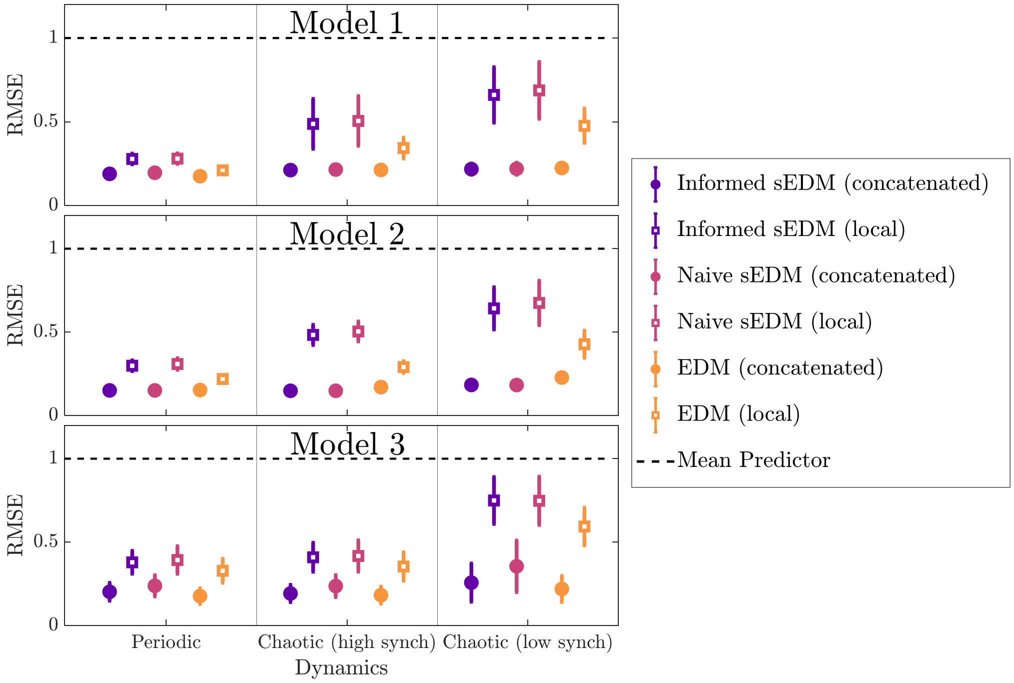

Figure 2.4: Average RMSE (points) and standard deviation (error bar) on testing data over 100 simulations of each model from Table 2.1. Results are shown for local (open boxes) and concatenated (closed circles) informed sEDM (purple), naive sEDM (pink), and EDM (yellow). Each model was simulated with 3 parameter sets generating periodic dynamics (left), highly synchronous chaotic dynamics (middle), and asynchronous chaotic dynamics (right).

Effects of Time Series Length

Longer time series yield better predictions for all methods (Figure 2.5; Table A.3). As the length of the training time series increases from 25 to 75, the RMSE decreases from about 50% to at least 70% average error reduction from mean predictor. As with dynamical complexity, the performance ranking of the methods does not change with time series length; concatenated methods are consistently best and local sEDM methods are consistently worst. For short time series ($T_{train} = 25$), concatenation reduces forecast error compared to local libraries by about 50%, but all methods produce similar accuracy for long time series ($T_{train} = 75$), so the benefit of concatenation decreases with time series length.

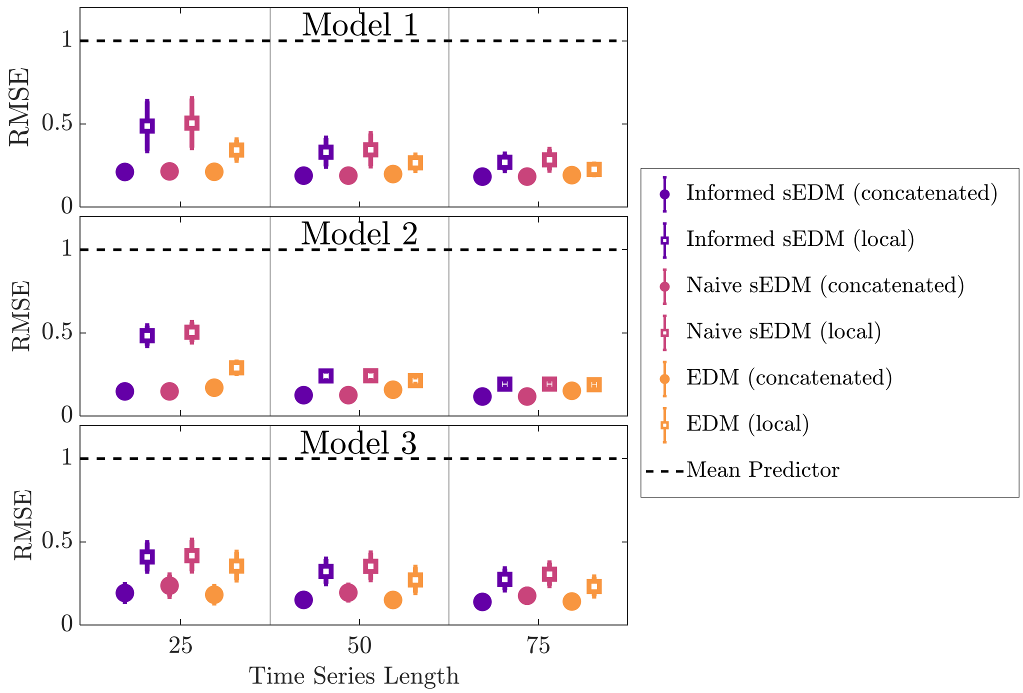

Figure 2.5: Average RMSE (points) and standard deviation (error bar) on testing data over 100 simulations of each model from Table 2.1. Results are shown for local (open boxes) and concatenated (closed circles) informed sEDM (purple), naive sEDM (pink), and EDM (yellow). Each model was simulated with 3 time series lengths: $T_{train} = 25$ (left), $T_{train} = 50$ (middle), and $T_{train} = 75$ (right).

Effects of asymmetry

Interestingly, the results are qualitatively different when dispersal is unidirectional compared to symmetric (Figure 2.7; Table A.4). When advection is present, local sEDM outperforms local EDM (10% average reduction in RMSE) even on the short time series of 25 points. Concatenating time series still produces a substantial improvement. Hence, any method of incorporating spatial data is better than implementing the standard local EDM method in this case.

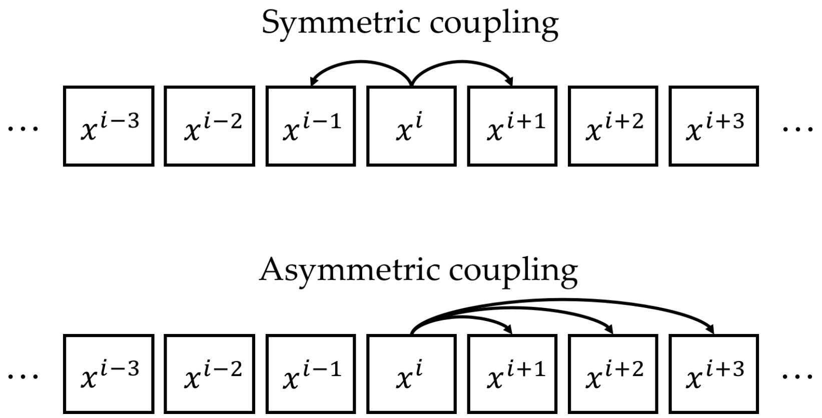
Symmetric coupling
$\cdots$ x^{i-3} x^{i-2} x^{i-1} x^{i} x^{i+1} x^{i+2} x^{i+3} $\cdots$
Asymmetric coupling
$\cdots$ x^{i-3} x^{i-2} x^{i-1} x^{i} x^{i+1} x^{i+2} x^{i+3} $\cdots$

Figure 2.6: Illustration of dispersal for symmetric and asymmetric coupling schemes.

Effects of spatial heterogeneity

As we increase spatial heterogeneity, the RMSEs for all concatenated methods increase until they become worse than their corresponding local methods (Figure 2.8). In contrast, local methods are relatively insensitive to spatial heterogeneity. This demonstrates that when spatial replicates become too different, they can no longer be concatenated.

One final result that pertains to all simulations is that sEDM with a physically informed prior provides minimal predictive advantage over the naive implementation, typically providing no more than 10% reduction in forecast error.

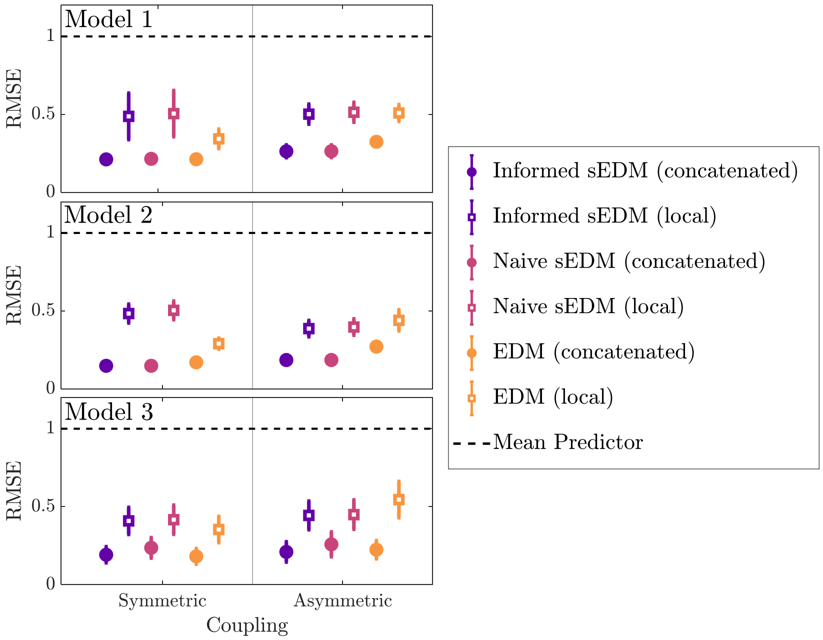

Figure 2.7: Average RMSE (points) and standard deviation (error bar) on testing data over 100 simulations of each model from Table 2.1. Results are shown for local (open boxes) and concatenated (closed circles) informed sEDM (purple), naive sEDM (pink), and EDM (yellow). Each model was simulated with symmetric nearest neighbor dispersal (left) and asymmetric unidirectional dispersal (right).

2.4.2 Results on empirical data

Results from simulations suggest that utilizing spatial information is likely to provide the strongest advantage over standard local EDM when time series are short and dynamics are complex and asynchronous. This is closely analogous to the data from NEFSC bottom trawl survey. The time series is only 36 years long, the dynamics vary irregularly and they are not highly synchronized (Table A.1 in Appendix A.3). In general, these data are ideal for evaluating EDM methods because they come from a system that is estimated to support over 5,000 species (Fautin et al., 2010) and we do not have complete understanding of all relevant species and drivers.

Concatenated methods always provide more accurate predictions than their

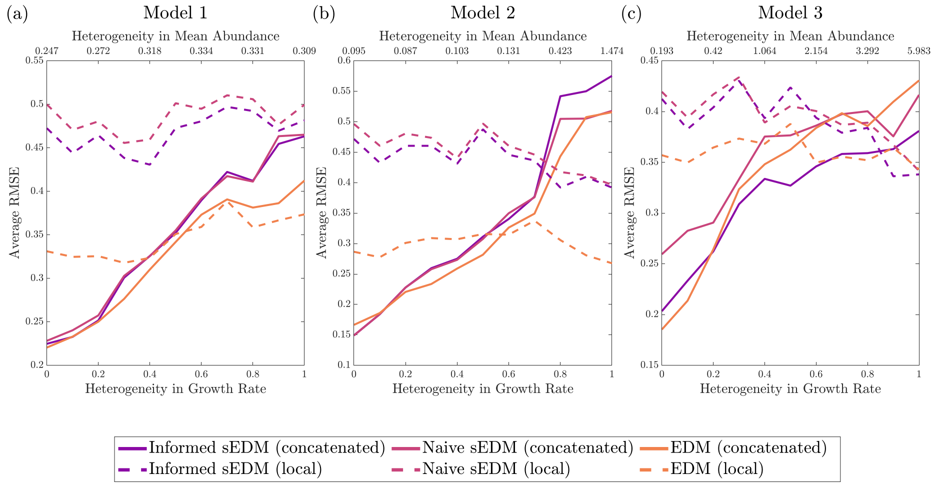

Figure 2.8: Average RMSE on testing data over 50 simulations of each model from Table 2.1 for various degrees of heterogeneity. Results are shown for local (dotted lines) and concatenated (solid lines) informed sEDM (purple), naive sEDM (pink), and EDM (yellow). The measure of heterogeneity in growth rate is given by the amplitude of the sinusoidal function used to vary the growth rate parameter. The auxiliary x-axis at the top of each figure is the empirical measure of heterogeneity in mean abundance (see Methods). Note that this empirical measure is not linear nor monotonic due to complexity and noise in dynamics.

corresponding local methods, but the improvement is more substantial at the fine spatial resolution where local methods lose accuracy (Figure 2.10). Specifically, concatenating data from all regions at the coarse resolution produces forecast errors that are approximately 12% lower than corresponding local methods on average. Concatenating data from sites within the local region at the fine resolution reduces error by 20% on average. In most cases, there are only slight differences between EDM and sEDM performance, and the best-performing method is not consistent across species. See Tables A.5 and A.6 for exact RMSE and R^2.

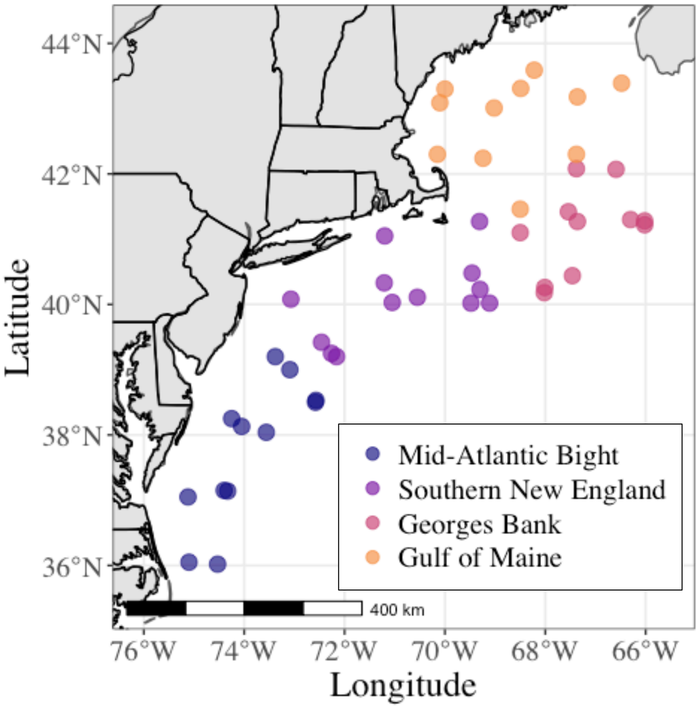

Figure 2.9: Strata locations from the NEFSC fall bottom trawl survey. Dots are midpoints of offshore strata, and colors indicate which major region each stratum is in.

2.5 Discussion

Here we explored various ways to leverage spatial data when forecasting ecological dynamics in partially observed systems. Specifically, we evaluated the forecast performance of the ecologists' approach (concatenated EDM), the physicists' approach (local naive sEDM) and our dispersal-motivated prior to combine information from multiple sites. We compared these methods to classic local EDM, which does not use spatial information. Results indicate that when dynamics are primarily homogeneous, any concatenated method produces substantially better forecasts than local EDM. If dynamics differ significantly among spatial replicates, however, concatenated methods produce poorer predictions since they rely on the assumption that spatial replicates have comparable dynamics. Applying both ap-

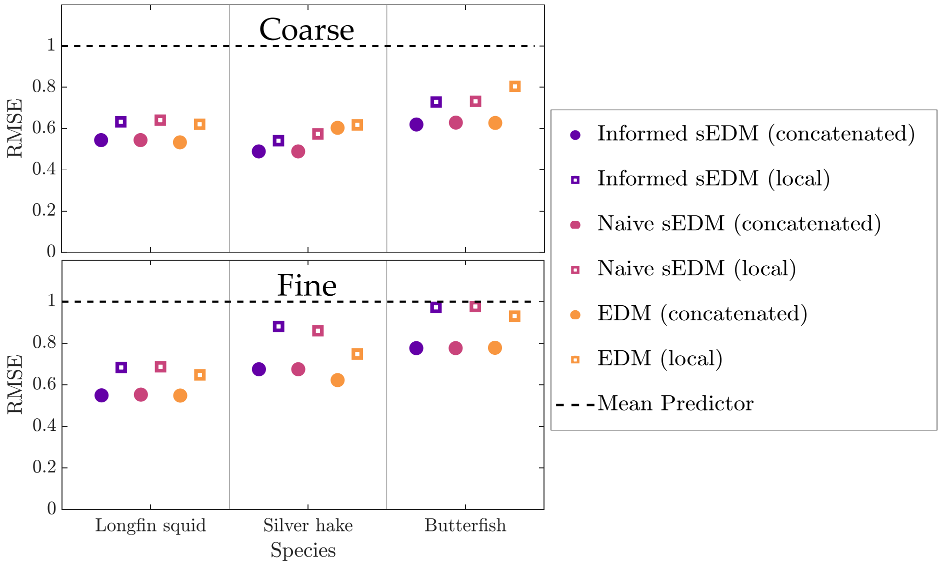

Figure 2.10: Sequential forecast RMSEs on NEFSC bottom trawl survey data for three species: longfin squid, silver hake, and butterfish. Results are shown for local (open boxes) and concatenated (closed circles) informed sEDM (purple), naive sEDM (pink), and EDM (yellow). Top (bottom) panels show results for coarse (fine) spatial resolution data separated by major biophysical region (survey strata).

proaches to a given data set may provide a direct way of identifying heterogeneous dynamics.

Dynamic complexity and time series length play a role in the forecasting ability of all methods. Generally, as complexity decreases and time series length increases, forecasts improve (Figures 2.4 and 2.5). Intuitively, when dynamics are regular (e.g. periodic), the attractor can be filled with a short time series and all methods predict dynamics well. When dynamics are chaotic, more data are required to reconstruct the attractor.

Regardless of complexity and time series length in Figures 2.4 and 2.5, local sEDM methods consistently have higher errors than local EDM, implying that it is not advantageous to incorporate spatial data through embedding vectors alone. This result is somewhat counterintuitive since EDM is a special case of sEDM. But we can understand this as a trade-off between in-sample fitting and outof- sample forecasting: sEDM has more length scale parameters and better in-sample fits on training data, but local EDM is more robust out-of-sample. ARD is insufficient to eliminate the unnecessary spatial lags when dynamics are complex and time series are short.

In the first two simulation experiments, concatenated methods have lower prediction errors than local methods. This is likely because concatenated methods fit a single GP on data from 10 locations while local methods fit a separate GP on data from each location. Therefore, concatenation uses 10 times more data than local methods. Figure A.2 shows that the forecast errors of local and concatenated methods converge as the time series length for local methods increases. Thus, under homogeneous conditions, concatenating data from 10 locations is essentially equivalent to having a time series that is 10 times longer. Therefore, concatenation is an effective way to overcome limitations of short time series provided that the spatial replicates contribute independent information, which is easily determined by measuring the improvement in forecast performance.

With symmetric dispersal, local EDM outperforms local sEDM, but the opposite occurs with unidirectional dispersal (Figure 2.7) as the length scale parameters in sEDM can accommodate asymmetry. Spatial heterogeneity is also important because it hinders the utility of concatenated methods (Figure 2.8) since they assume stationarity across replicates (i.e. all replicates must come from the same attractor; Banbrook et al. (1997)). This assumption is reasonable if the underlying dynamical equations are identical for all spatial replicates. However, if they vary across sites, information is not easily shared between them.

Given the simulation results, it is reasonable to imagine cases that could yield different qualitative outcomes. Figure A.3 shows that when strong heterogeneity and asymmetry are both present, local sEDM methods are most accurate and concatenated methods are least accurate, contradicting the results in Figures 2.4 and 2.5 in nearly every way. Importantly, we cannot identify precise thresholds of asymmetry or heterogeneity at which these changes occur. We present this simply as an important area for future work and a proof of concept that these aspects of the dynamics are instrumental in the utility of spatial extensions of EDM.

Although simulations reveal that none of these methods perform uniformly well in every scenario, the fact that their performance depends on characteristics of the dynamics offers interesting potential for diagnosing ecological conditions. For example, if we did not know anything about the dynamics of a system, comparing forecast performance across all of these methods may provide insights into whether advection and heterogeneity are influencing the dynamics. This knowledge of ecological conditions could be useful when making management decisions.

In our empirical analysis on coarse resolution data from the NEFSC survey, local sEDM outperformed local EDM on silver hake and butterfish data but did not on longfin squid data. In light of simulations, this is consistent with advective dispersal of silver hake and butterfish. Interestingly, there is empirical evidence that dynamics in this region are advection-dominated from particle tracking oceanographic models (Lynch et al., 2014), stable isotope studies (Clarke et al., 2009) and genetic data (Mach et al., 2011). Furthermore, concatenated methods produced slightly lower forecast errors than local methods at the coarse resolution. Given the complexity and asymmetry in the empirical data (Table A.1), we might expect concatenation to provide a stronger improvement (e.g. right panels of Figure 2.4). However, it is possible that a moderate level of heterogeneity hinders the utility of concatenation in these cases. There is ample spatial heterogeneity in environmental drivers (e.g. temperature, salinity, dissolved oxygen) that are often linked to differences in population growth (Hofmann and Powell, 1998), and the ecological communities (e.g. prey, predators, competitors) also differ substantially between these regions. The empirical measures of heterogeneity on these for all species at the coarse resolution further support this deduction: $H_m = 1.54$ for longfin squid, $H_m = 1.23$ for silver hake and $H_m = 1.82$ for butterfish.

In contrast, at fine resolution, concatenated methods provided greater predictive improvement. This is likely due to the geographic ranges spanned: At the coarse resolution, concatenated methods used data from along the entire coast whereas at the fine resolution, concatenation was restricted within biophysical regions. As a result, there is lower heterogeneity in concatenated data at the fine resolution ($H_m = 1.10$ for longfin squid, $H_m = 0.97$ for silver hake, and $H_m = 1.06$ for butterfish) allowing for information to be more easily shared across sites.

Note that local methods performed more accurately at the coarse spatial resolution data than they did at the fine resolution. One explanation for this is that averaging abundance measures over several strata at the coarse resolution distorts nonlinear signals that appear at the fine resolution (Glaser et al., 2014a; Sugihara et al., 1999, 1990; Ye et al., 2015a), making it easier to predict coarse dynamics accurately.

All analyses reveal that informed sEDM usually produces more accurate predictions than naive sEDM, but the improvement is typically no more than 1%–5% (Tables A.2, A.3, A.4, A.5, A.6). This suggests that although our informed prior helps resolve the lack of identifiability (Figure 2.1) and improves interpretability, it does not substantially improve predictive accuracy. This is consistent with the multivariate embedding theorem's notion that different combinations of lags can produce equivalent predictions (Deyle and Sugihara, 2011). Thus, using the informed prior is not necessarily helpful if the goal is to improve predictions, though it may facilitate inference about ecological mechanisms based on the ARD selection of relevant lags (Browne et al., 2008; Marwala, 2015; Munch et al., 2018).

Regardless of whether EDM is used for inference or prediction, it is crucial for EDM users to understand the importance of temporal and spatial scale. If the temporal scale of the dynamics is too large for EDM to capture with time lags or if samples are not taken with high frequency, none of these methods will give accurate forecasts (e.g. Appendix A.9) and a linear or other parametric modelling approach may be more suitable. Similarly, the spatial scale at which data are collected may influence the heterogeneity across sites. For example, if data are collected over a large geographic range at distant sites, dynamics likely vary across locations due to environmental gradients. In this case, concatenated methods are unlikely to perform well. Generally, given the ubiquity of spatial heterogeneity in ecological systems at the scale we typically observe (Aksnes et al., 1989; Juanes and Conover, 1995; Tommasi et al., 2014), we suggest that concatenating libraries should be done in conjunction with the sEDM approach presented here.

Although we have identified asymmetry and heterogeneity as factors that influence the efficacy of each method there are many other factors that could also play a role but were not addressed in our simulations. For example, variation in dispersal strength and distance, and differences in movement between species and age classes are likely to affect which method performs best. In light of this, it would be helpful to develop specific diagnostic tests that provide guidelines for choosing among methods (e.g. a measure of information flow between sites). In early attempts, we considered embedding dimension and spatial synchrony as potential indices, but these proved inconsistent. More work is needed on this topic.

Our main simulations also did not address how these methods perform as we predict multiple steps into the future. However, Figure A.4 shows that all methods produce better predictions than the mean up to 10 steps into the future. Changing the forecast horizon generally does not influence relative performance across methods and we expect our conclusions about complexity, asymmetry and heterogeneity extend to multi-step predictions.

We compared the six methods here because they are commonly used, they are easy to implement and they are good starting points for integrating spatial information into EDM. In the future, it would be valuable to evaluate hierarchical approaches (Munch et al., 2017; Rogers and Munch, 2020), approaches that average over several sub-optimal embeddings (Okuno et al., 2020), diffusion maps (Coifman and Lafon, 2006) or deep-learning (Chattopadhyay et al., 2020) to determine which methods are more robust.

All of our analyses made inferences from single time series. This was intended to provide proof-of-concept and determine general conditions that influence our ability to integrate spatial information. However, time series for several species are often available at multiple locations, and studies on multiview (Ye and Sugihara, 2016) and multivariate (Deyle and Sugihara, 2011; Dixon et al., 1999) embedding suggest that predictions may improve if we include information from other species and drivers. We could do this by simply including lags of other species as additional inputs in embedding vectors, we could construct a hierarchical version to share information across species, or we could combine the method of multiview embedding with sEDM. Comparing the benefits of these approaches is potential territory for future work. Additionally, many species are influenced by harvesting or other anthropogenic effects. In these cases, incorporating the past history of exploitation or anthropogenic inputs is likely to improve prediction accuracy. This is an important next step that opens up the possibility of using EDM as a tool for quantitative ecosystem management (Boettiger et al., 2015; Giron-Nava et al., 2017).

The flexibility of EDM to describe complex dynamics with incomplete data makes this a promising avenue for making strong inferences of ecological dynamics and developing multi-species management policies that are robust to structural uncertainty. However, these methods require long time series for convergence. Although typical ecological time series are fairly short, aggregating information across multiple sites—either through concatenation or sEDM—can substantially improve the utility of empirical approaches to ecological dynamics.

2.6 Acknowledgments for Chapter 2

This work was supported by funding from the Lenfest Oceans Program, NOAA's HPCC Program, and NOAA Grant #NA19OAR4170353 through the Fisheries-Sea Grant Joint Graduate Fellowship Program in Population and Ecosystem Dynamics. The statements, findings, conclusions and recommendations are those of the authors and do not necessarily reflect the views of Sea Grant. We thank Charles Perretti and Michael Fogarty for providing access to the NEFSC trawl data, and George Sugihara, Ethan Deyle, Marc Mangel, Hongyun Wang and Hao Ye for stimulating conversations about these ideas. We are grateful to Uttam Bhat and Tanya Rogers, and Adam T. Clark for many helpful comments on this manuscript.

Chapter 3

Empirical Dynamic Modeling for Missing or Irregular Data

This chapter is a reprint of Johnson and Munch (2022) as it appears in Ecological Modelling (Volume 468, Article 109948, June 2022). The dissertation author was the primary investigator and author of this paper.

3.1 Abstract

Empirical dynamic modeling (EDM) is a powerful method for forecasting and analyzing nonlinear dynamics. However, typical applications of EDM assume that samples are evenly spaced over time. This presents problems in ecology, in which data are often missing or sampled irregularly. Standard methods for handling irregularity in EDM suffer under conditions that are common in ecology, such as short time series and large dynamic fluctuations, so there is a need to adapt the framework to cope with these challenges more effectively. Here we consider a variable step-size extension of EDM, which incorporates the temporal spacing between samples into EDM delay-coordinate vectors and circumvents the challenges faced by other approaches. We evaluated the forecast accuracy of the variable step-size method along with that of two other methods: (1) exclusion of delay- coordinate vectors with missing data and (2) linear interpolation along with ordinary EDM. We tested these methods using simulated data from three chaotic ecological models with various amounts and patterns of missing data. We also evaluated them using two empirical datasets: laboratory rotifer dynamics and aphid dynamics from the field. Results showed that while exclusion and linear interpolation can produce accurate forecasts in some scenarios, the variable step-size method consistently gives accurate forecasts in a wide range of scenarios. Our analysis demonstrates that variable step-size EDM is an effective method for coping with missing or irregular samples and expands the number of datasets to which EDM can be applied. Furthermore, EDM can be extended to estimate Lyapunov exponents from irregularly sampled time series and approximate continuous dynamics from discrete-time data.

3.2 Introduction

The ability to make predictions is crucial in fields ranging from economics (Liu et al., 2016; Poon and Granger, 2003) to environmental sciences and climatology (Betts et al., 1996; Pau et al., 2011). It is particularly relevant for ecological decision makers who must make policies to mitigate impacts of pest outbreaks, natural disasters, harmful algal blooms, and diseases, or maximize economic yields from commercial fisheries, agriculture, and forestry. Since these policy decisions depend on – and influence – the future state of the ecosystem, accurate forecasts are critical to effective management (Clark, 2001; Dietze et al., 2018).

The natural world makes ecological forecasting challenging; ecosystems are typically sparsely observed complex systems that display strongly nonlinear dynamics on a wide range of time scales (Clark and Luis, 2020; May, 1976). Since scientists have imperfect understanding of nonlinear interactions among the variables in these complex systems, they are forced to make many simplifying assumptions when they use traditional mechanistic modeling approaches for forecasting. Small changes in these assumptions can lead to large differences in model predictions and the management advice they generate, which can lead to extreme uncertainty in decision making. As a consequence, linear forecasting tools (Hampton et al., 2013; Ives et al., 2003) and standard model-fitting approaches (Perretti et al., 2013) may not be optimal for forecasting ecological dynamics.

Empirical dynamic modeling (EDM) (Sugihara and May, 1990), on the other hand, avoids these problems by making minimal structural assumptions and using only observed dynamics to make forecasts. Furthermore, EDM does not require observations of all variables in a system. Takens' theorem of time delay embedding (Takens, 1981) provides the foundation for EDM by showing that an attractor (i.e. the point, set of points, or orbit to which a dynamical system converges) can be faithfully reconstructed using time lags from the time series of a single state variable, x_t, where $t = 0, 1, \ldots, T$. Given an attractor reconstruction, a variety of function approximation tools (e.g., local linear regression (Farmer and Sidorowich, 1987; Sugihara, 1994), Gaussian processes (Munch et al., 2017; Wang et al., 2008), neural networks (Bakker et al., 2000)) can be used to predict the future state of the system by fitting models of the form

$$x_{t+h} = f(x_{t-1}, x_{t-2}, \ldots, x_{t-E}) + \epsilon_t \tag{3.1}$$

where $\{x_{t-1}, x_{t-2}, \ldots, x_{t-E}\}$ are the 'delay-coordinate vectors' composed of time lags of the observed states, h is the number of time steps into the future the model predicts, f is the map that converts the past states to the future state, E is the ‘embedding dimension’, and ϵ is process or approximation error.

EDM has outperformed standard modeling approaches in predicting fish recruitment (Deyle et al., 2018; Munch et al., 2018), quantifying interactions among species (Deyle et al., 2016; Rogers et al., 2020; Ushio et al., 2018), and identifying causal interactions (Sugihara et al., 2012). Various recent developments have made EDM more applicable to ecology by incorporating information from multiple sources through hierarchical modeling (Munch et al., 2017; Rogers and Munch, 2020), using spatial replicates (Clark et al., 2015; Hsieh et al., 2008; Johnson et al., 2021), and including information from interacting species and environmental drivers (Deyle and Sugihara, 2011; Ye and Sugihara, 2016). Takens’ theorem was also extended to stochastic systems by Stark (1999), and Munch et al. (2020) demonstrated via simulation that EDM efficiently captures the conditional expectation for stochastic processes. Predictions from EDM can be used to make short-term management decisions (Deyle et al., 2013) and identify optimal management policies (Boettiger et al., 2015; Brias and Munch, 2021). More information and intuitive descriptions of EDM can be found in Ye et al. (2015a), Chang et al. (2017), and Munch et al. (2020).

While EDM shows substantial promise for contributing to management in ecology and circumvents problems associated with mechanistic models, there are still several challenges to applying it to ecological time series (Munch et al., 2020). One common obstacle arises from the sampling design of many ecological surveys, in which varying degrees of sampling effort result in missing samples or data otherwise being collected at irregular intervals. Randomly allocating sampling effort, while effective at mitigating bias (McGarvey et al., 2016), is problematic for reconstructing attractors. The resulting datasets are not amenable to standard EDM approaches, which typically rely on uniform temporal spacing between points.

There are many reasons for samples to be non-uniform in ecological data, and they represent all of the categories that are typically used to classify missing data in the statistical literature (Little and Rubin, 2002; Rubin, 1976). One common scenario is when weather conditions prevent or limit sampling during certain periods of the year. This results in regular sampling seasons followed by extended periods of missing or less frequent samples (e.g. Harrington and Woiwod (2007)). Funding constraints, lapses in funding, or convenience may also lead to data being sampled unevenly over time. For instance, experimenters may sample a collection only on weekdays (Laan and Fox, 2020). These are examples of data that are "missing at random" because the probability that a point is missing depends on an observed variable (i.e. the season or day) but not on the value itself. Equipment malfunction or interference from external factors also contributes to irregularity. For instance, unmanned aerial systems cannot detect populations that are obscured by trees or clouds (Brack et al., 2018). This is an example of "missing completely at random" because the probability that a point is missing does not depend on any measured variable. Finally, missing samples also emerge when sampling instruments have fundamental limitations. For instance, satellites cannot detect algal density when chlorophyll a concentrations lie below 1 mg/m3 (Gokul et al., 2019). Similarly, high jellyfish densities can destroy nets during trawl surveys, so sampling stations are sometimes skipped when they have too many jellyfish (Field et al., 2021). These are examples of data that are "missing not at random" because the values of the data influence the probability of missingness. Since all of these are possibilities in ecology, it is important to have a robust method to cope with them.

The statistical literature highlights two primary ways to handle nonuniform samples (Little and Rubin, 2002). One approach is to remove all missing observations from the data and analyze only the set of complete data. This is known as deletion or complete-case analysis (Little, 1992). The other approach is to fill the gaps in the data, which can be done in many ways. Missing points can be filled with the mean of the complete data, the previous value in the data, or a prediction from linear interpolation or splines. It is also possible to use more intricate imputation methods by fitting a model to the complete data and using it to estimate the missing points. A model is then fit again to all of the data including the estimated points, and the new fit is used to refine the estimates. When this procedure is iterated until convergence, it is known as the expectation-maximization (EM) algorithm (Dempster et al., 1977). Various adaptations of the EM algorithm are commonly used in ecological analyses when observations are missing. For instance, Bayesian hierarchical imputation has been used in analyses of Peary caribou (Kaluskar et al., 2020) and tuna (Horswill et al., 2019) populations. Importantly, these methods are typically used when a parametric model structure permits imputation. With EDM, however, the goal is to infer a nonparametric model with delay-coordinate vectors, and missing data are not easy to impute using standard tools because an a priori model function is not available. Since there are no examples of these imputation tools using nonparametric time lags, another way to handle missing data directly within the EDM framework is needed.

Although EM imputation approaches are not easily implemented with EDM, other standard approaches to cope with missing data have been used in EDM applications. The first is complete-case analysis, which excludes delay-coordinate vectors that have incomplete observations (Little, 1992). The exclusion method includes the following steps: 1) given a time series of length T that contains some missing values, create the lag matrix for EDM, $X = \begin{bmatrix} x_1 & ... & x_E \\ x_2 & ... & x_{E+1} \\ \vdots & \ddots & \vdots \\ x_{T-E+1} & ... & x_T \end{bmatrix}$, 2) identify and remove rows of X that have at least one missing value to obtain the submatrix X_{sub}, and 3) perform EDM using X_{sub}. The second way that missing data have been handled in EDM is to interpolate the missing observations. Linear interpolation is the most common type of interpolation used in ecological EDM analyses with irregular samples (Ness-Cohn and Braun, 2020; Ye et al., 2015b), and it has been used as a benchmark in other studies (Lekscha and Donner, 2018). Although exclusion and interpolation tend to be the default approaches for handling missing data in EDM, excluding points does not use the data efficiently, and pre-analysis gap filling may introduce bias or uncertainty to predictions. Thus, both methods are typically not robust if the time series are short and the dynamics have highly irregular fluctuations (Hummel, 1946). More efficient approaches for dealing with missing samples and irregular sample spacing will substantially expand the utility of EDM in ecological forecasting.

One promising option is to use the bundle-embedding theorem (Stark, 1999; Stark et al., 1997), which extends Takens' theorem to forced dynamical systems. Essentially, bundle-embedding expands the input space for equation (3.1) from $\{x_{t-1}, ..., x_{t-E}\}$ to $\{x_{t-1}, \ldots, x_{t-E}, \omega_{t-1}, \ldots, \omega_{t-E}\}$ where the ω_{t-i} are time lags of the forcing variable. This idea has been used repeatedly in ecology to include environmental drivers (e.g. Deyle et al. (2016), Rogers and Munch (2020)) and harvesting (Brias and Munch, 2021). Importantly, Stark (1999) notes that if we assume that the 'forcing' at time t is given by the time between sampling intervals, $\tau_i = t_{i+1} - t_i$, then the bundle-embedding theorem implies that $\{x_{i-1}, \ldots, x_{i-E}, \tau_{i-1}, \ldots, \tau_{i-E}\}$ can be used to reconstruct the dynamics of irregularly sampled systems, and we can forecast dynamics by approximating the function $x_{i+h} = f(x_{i-1}, \ldots, x_{i-E}, \omega_{i-1}, \ldots, \omega_{i-E}) + \epsilon$. To our knowledge, however, this 'variable step-size' idea has never been evaluated using simulations of realistic lengths of time series or patterns of missing data, nor has it been applied to ecological time series.

Here, we evaluate the performance of the variable step-size algorithm using Gaussian process regression for five scenarios of missing data at varying levels of severity in simulated data. We also evaluate the method using two empirical datasets: one from the laboratory and one from the field. To provide a relevant benchmark, we also evaluate the performance of exclusion and linear interpolation in each scenario. We find that, while exclusion and linear interpolation can produce accurate forecasts in some scenarios, the variable step-size method consistently gives accurate forecasts in a wide range of scenarios and is the most robust to large amounts of missing data.

3.3 Methods

3.3.1 Variable step-size EDM

Rigorous justification of the variable step-size approach to EDM is provided by Stark (1999) using the bundle-embedding theorem; however, the proof of this theorem requires topological ideas that are not particularly transparent. Fortunately, an intuitive justification is straightforward. To see this, start with an autonomous continuous-time system,

$$\dot{x} = g(x) \tag{3.2}$$

If we start at time t_i from state x_{t_i} and integrate (2) from t_i to $t_i + \tau_i$, the next state will be

$$x_{t_i+\tau_i} = x_{t_i} + \int_{t_i}^{t_i+\tau_i} g(x_s)ds = G(x_{t_i}, \tau_i) \tag{3.3}$$

That is, the next step map is a function of both the current state, x_{t_i}, and time interval τ_i. We can simplify this notation somewhat by denoting the state at the i^{th} observation time, x_{t_i}, as x_i and recognizing that the $(i+1)^{th}$ observation time is $t_{i+1} = t_i + \tau_i$, so that $x_{i+1} = G(x_i, \tau_i)$.

Variable step-size embedding combines this expanded state space with the standard EDM idea of using time lags to account for unobserved state variables. That is, we can account for variable time steps and hidden state variables when forecasting by estimating

$$x_i = f(x_{i-1}, \tau_{i-1}, ..., x_{i-E}, \tau_{i-E}) \tag{3.4}$$

for a given embedding dimension E. Note that when the sampling intervals are constant, the dependence on τ_i is irrelevant, and this model reduces to standard, fixed-lag EDM.

By approximating function f in eq. (3.4), we can make predictions for x using EDM, even with irregular sampling. We call this “variable stepsize” EDM (VS-EDM) since it incorporates the differing intervals between samples into EDM inputs.

In this analysis, we approximate f via Gaussian process (GP) regression (Rasmussen and Williams 2006; Munch et al., 2017). The GP is a continuous generalization of the multivariate normal distribution, completely specified by a mean function, $m(\mathbf{x}) = E[f(\mathbf{x})]$, and a covariance function, $C(\mathbf{x}, \mathbf{x}') = E[(f(\mathbf{x}) -$

$m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$. We use the GP to estimate f such that

$$x_i = f(\mathbf{x}_{i-1}) + \epsilon_i$$

where $\mathbf{x}_{i-1} = \{x_{i-1}, .., x_{i-E}, \tau_{i-1}, \ldots, \tau_{i-E}\}$, and ϵ_i is the approximation error, with mean 0 and variance V.

To infer f, we assume a GP prior and update the mean and covariance using the observed data. Lacking any *a priori* information on the shape of the delay-embedding map f, we set the prior mean function $m(\mathbf{x}) = 0$. In keeping with other ecological applications of GPs in EDM (e.g. Brias and Munch (2021); Munch et al. (2018, 2017); Rogers and Munch (2020)), we used a squared-exponential covariance function to control the wiggliness of f. Specifically, we set

$$C(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\sum_{i=1}^{n} \phi_i (x_i - x_i')^2\right). \tag{3.5}$$

where n is the input dimension, ϕ_i is the inverse length scale that governs how much f can vary with the i^{th} input, and σ_f^2 is the pointwise variance. In standard EDM, $n = E$, since the input includes E time lags of x. For VS-EDM, $n = 2E$ for each state, step-size pair. However, note that setting $\phi_{E+1}, \ldots, \phi_{2E}$ to 0 eliminates the dependence of f on τ, so that standard EDM is a special case of the VS-EDM.

As detailed by Munch et al. (2017), we standardize the inputs to zero mean and unit variance prior to fitting, which facilitates the use of standardized priors for the GP hyperparameters. We assign identical half-normal priors to each ϕ, with variance $\pi/(\sqrt{12})$, which regularizes f by asserting that, on average, f will have a single extremum over a unit interval of the input. This prior places most of the weight on $\phi = 0$, resulting in 'automatic relevance determination' (ARD) (Neal, 1996), which is used in the machine-learning literature to automatically remove irrelevant inputs from a model. For σ_f^2 and V, we use weakly informed prior distributions, $Beta(1.1, 1.1)$, which conservatively limits uncertainty in the next state to twice the observed total variance in order to avoid numerical artifacts during parameter tuning. With these priors, we use resilient back-propagation (Riedmiller and Braun, 1993) $(\phi_1, \ldots, \phi_n, \sigma_f^2, V)$ that maximize marginal likelihood (see Rasmussen and Williams (2006) for further details). We make predictions conditional on these maximum *a posteriori* estimates. Fully Bayesian inference with the GP is also feasible, though considerably more computationally demanding (Munch et al., 2005; Poynor and Munch, 2017; Thorson et al., 2014). For more information on GP regression, Rasmussen and Williams (2006) and Munch et al. (2017) provide overviews and examples of using GP regression and EDM.

3.3.2 Forecasting Simulated Data

To evaluate forecast performance when the correct answer is known, we used three ecological models to simulate chaotic time series: Ricker dynamics, host-parasitoid dynamics, and three-species competition dynamics (Table 3.1). For each simulation, we assigned a random initial value to each variable, iterated the model for 170 time steps, and then removed the first 100 points to avoid transient dynamics. This left time series of length 70, which we split into 50 'training' data points for fitting the GP, and 20 'testing' data points for estimating one-step-ahead forecast accuracy from out-of-sample data. For each model, we used the time series of a single variable and subsampled the training data to generate various amounts of missing data with different patterns. From the sampled time series, we constructed EDM delay-coordinate vectors with a maximum embedding dimension of $E = 4$, and used ARD to prune irrelevant inputs. Each model, sampling structure, and level of missing data was simulated 100 times with random initial conditions to produce summary statistics of forecast accuracy.

#	Name	Model	Parameters
1	Ricker	$x_{t+1} = x_t e^{r(1-x_t)+\xi_t}$	$r = 3.0$
2	Host-parasitoid	$x_{t+1} = (1-\gamma)x_t e^{r(1-x_t)+\xi_t^1}$ $+\gamma x_t e^{r(1-x_t-y_t)+\xi_t^1}$ $y_{t+1} = \gamma\beta x_t(1-e^{-y_t})e^{\xi_t^2}$	$r = 3.0$ $\gamma = 0.1$ $\beta = 12$
3	Three-species competition	$x_{t+1} = x_t e^{r(1-x_t-ay_t-bz_t)+\xi_t^1}$ $y_{t+1} = y_t e^{r(1-y_t-az_t-bx_t)+\xi_t^2}$ $z_{t+1} = z_t e^{r(1-z_t-ax_t-by_t)+\xi_t^3}$	$r = 3.0$ $a = 0.65$ $b = 0.6$

Table 3.1: Models used to generate ecological time series. Models and parameter values for chaotic dynamics of a single-species Ricker model (Model 1), a two-species host-parasitoid model (Model 2), and a three-species competition model (Model 3). Here, ξ_t^i is a noise term that is drawn independently from a normal distribution, $N\left(-\frac{s^2}{2}, s^2\right)$. We used $s = 0.05$ for Models 1 and 3, and $s = 0.1$ for Model 2.

Both the amount and pattern of missing data may influence the ability to forecast and analyze time series. To address this, we simulated five scenarios that represent the most common patterns of missing data in ecological time series: (I) points missing at random (e.g., a single site of a survey was skipped), (II) points missing during an interval (e.g., equipment malfunctioned and failed to collect samples for a period), (III) regular missing intervals (e.g., winter snow prevents data from being sampled), (IV) points missing below a threshold (e.g., equipment cannot observe a state below its detection limit), and (V) mixed-interval sampling (e.g. samples combine weekly and monthly observations). Implementation details for each scenario are as follows:

I Missing at random: For each time series, points were independently selected at random and removed from the training data. We varied the proportion of missing points from 0-0.5 in increments of 0.1.

II Missing intervals: We selected a point in the time series at random and removed subsequent observations over a randomly chosen period. The duration of this inoperative period was sampled from a uniform distribution, $U(a, b)$, in which we varied a from 0-5 in increments of 1 and simultaneously varied the associated b from 5-10 in increments of 1.

III Regular missing intervals: We simulated monthly data and varied the number of consecutive months sampled within a year from 6 to 11. Specifically, the first case had 6 sampled points followed by 6 missing points, and the last case had 11 sampled points, followed by 1 missing.

IV Missing below a threshold: We removed all points below a set percentage of the simulated range of states, which we varied from 0-15% in increments of 3%.

V Mixed intervals: We evaluated sub-scenarios in which (A) sampling occurs for seven consecutive months, then every other month for the rest of a year; (B) biweekly sampling occurs in two different seasons, resulting in eight consecutive points, a gap of seven, then five consecutive samples followed by a gap of five; and (C) five consecutive samples followed by a gap of two, corresponding to experiments in which data are not collected on weekends.

Using these simulations, we evaluated three ways of handling missing samples in EDM: (1) the exclusion method, (2) linear interpolation followed by standard EDM, and (3) VS-EDM. For each method, we measured forecast accuracy using the mean-square error scaled by variance, which we converted to an out-of-sample coefficient of determination R^2 given by

$$R^2 = 1 - \frac{\sum_{n=1}^{T}(x_{t_n} - \hat{x_{t_n}})^2}{\sum_{n=1}^{T}(x_{t_n} - \bar{x})^2},$$

where x_{t_n} for $n = 1, \ldots, T$ is the testing data, $\hat{x_{t_n}}$ is the EDM forecast for x_{t_n}, and $\bar{x}$ is the mean of testing data.

Note that a perfect forecast yields $R^2 = 1$, while a naïve forecast of predicting the mean for all points yields $R^2 = 0$. The R^2 of out-of-sample forecasts can be negative, but forecasts with a score this low are considered 'unusable' since they are worse than a naive mean predictor. As a performance benchmark, we also calculated the R^2 obtained when using the complete time series with no missing data as an upper bound of predictability. All analyses were performed in Matlab version 9.8.0.1538580 R2020a.

3.3.3 Forecasting Empirical Data

Following our simulation analysis, we tested the irregular sampling methods on one empirical dataset from the laboratory and one from the field. The laboratory data were for population dynamics of the rotifer *Brachionus calyciflorus* (Halbach, 1984). In this study, the rotifers were cultured in controlled conditions with constant water volume, temperature (20°C), and light. In two separate experiments, the population density was sampled over 55 days at irregular intervals that ranged from 0.5 to 1.5 days.

Importantly, these time series do not have clear missing values, but the intervals between samples are quite irregular. Although we framed our simulated scenarios as 'missing' data, they can be considered special cases of unevenly sampled data. Therefore, these empirical data fit with the conditions that we tested in simulations. Since the laboratory data have a relatively small range of step sizes and relatively smooth dynamics (Figure B.1), it would be possible to simply ignore the irregular time step and forecast the time series using standard EDM. This approach, which can be considered 'piecewise constant interpolation', has been used in previous studies when the data were in a similar form (Clark and Luis, 2020). Additionally, to be consistent with the simulation experiments, we also compared the methods to linear interpolation and the exclusion method using an even grid of daily time steps (i.e. samples taken on the same day were averaged, and if no sample was taken on a given day, we assigned a missing value). Since these time series were shorter than our simulated time series, we used leave-one-out cross validation to evaluate the methods. We tested a range of E from 2–4 (Cheng and Tong, 1992) for each series and reported results for the best-performing E for each method. This analysis represents a ‘best-case’ scenario in that the dynamics are simple, the time series are relatively long by ecological standards, and the abundance estimates are quite accurate.

To contrast the carefully controlled laboratory data, we also evaluated the methods using long-term field data of aphid abundance from the Rothamsted Insect Survey (Bell et al., 2015; Shortall et al., 2009). Aphid abundances were estimated using 12.2 m suction traps that sample migrating aphids at height across the United Kingdom (Macaulay et al., 1988). The traps are emptied weekly in the winter and daily in the spring, summer, and fall (i.e. the “aphid season”), and aphids are identified and counted at Rothamsted (Shortall et al., 2009).

We used the Rothamsted data to forecast the population dynamics of *Sitobion avenae*, the English grain aphid. Because field data are typically far more difficult to predict than simulated or laboratory data, we restricted our analysis to sites with few inoperative periods and treated the data prior to analysis to ensure that there were no missing points in the time series. We did this to compare our forecasts to the complete data benchmark, as we did for the simulations, which allowed us to fully evaluate the sensitivity to missing data rather than dynamic complexity. To obtain complete time series, we aggregated samples into monthly data by averaging the number of aphids caught each month from February 1972-January 2012. We focused our analysis on five sites: Brooms Barn S, Rothamsted, Writtle, Newcastle, and Hereford, because the English grain aphid is frequently found at these sites and they have operated consistently since early in the survey (Figure B.2).

Treating each site independently, each time series was split into 75% training data (February 1972-January 2002) and 25% testing data (February 2002-January 2012). We first evaluated standard EDM using the complete data and E from 4 to 12. Since higher embedding dimensions provided little predictive advantage and were more computationally demanding, we used $E = 4$ in remaining analyses. To evaluate forecasts for irregularly sampled data, we randomly selected and removed a proportion of points (from 0 to 0.5) from the training time series and repeated this procedure 100 times for each site and proportion.

3.4 Results

3.4.1 Results for Simulated Data

Based on the quartiles of R^2 for 100 replicates of scenario I, all methods perform equally well and nearly perfectly when no points are missing, which indicates that despite doubling the input size with VS-EDM, VS-EDM produces identical forecasts because the ARD effectively drops irrelevant τs out of the model when no points are missing (Figure 3.1a-c). As the proportion of data missing from the training time series increases, the accuracy of all methods decreases, but VS-EDM has the smallest rate of decline, producing usable forecasts when up to 40% of points are missing. The exclusion method provides accurate forecasts when up to 20% of points are missing but rapidly declines for higher proportions of missing data. For all three models, interpolation is the most sensitive to data that are missing at random.

All methods perform relatively well for scenario II, with intervals of missing values ranging from a maximum of 5-10 steps (Figure 3.1d-f). However, linear interpolation of missing blocks reduces forecast accuracy regardless of interval length. The exclusion method is effective for this scenario because there are few interruptions to the original time series, so the data are sufficient to cover the attractor. The decline in accuracy for interpolation is likely due to introducing biased estimates by linearly interpolating over an extended interval. In scenario III, with regular missing intervals, the forecast accuracy of all methods improves as the number of months sampled per year increases from 6-11 months (Figure 3.1g-i). VS-EDM maintains the highest accuracy over the full range of sampling intervals evaluated and is roughly 30% more accurate than the exclusion method when sampling intervals are short. In contrast, using linear interpolation reduces forecast performance greatly.

In scenario IV, when missing values are state-dependent (i.e. missing below a threshold), all methods perform poorly, although VS-EDM is considerably less sensitive than the other two methods (Figure 3.1j-l). For all three models, the interpolation and exclusion methods produce unusable forecasts ($R^2 < 0$) once the threshold exceeds 5%. Alternatively, VS-EDM tends to produce usable forecasts for thresholds up to 9%, though its forecast accuracy is low at that threshold. Importantly, all methods are extremely sensitive to state-dependent missing values because they must forecast in a region of state space that is never observed in the training data. The sensitivity may depend on which region of state space is missing and the divergence of the map in that region. Assigning missing values to other regions of the state can result in less sensitivity (Appendix B.3).

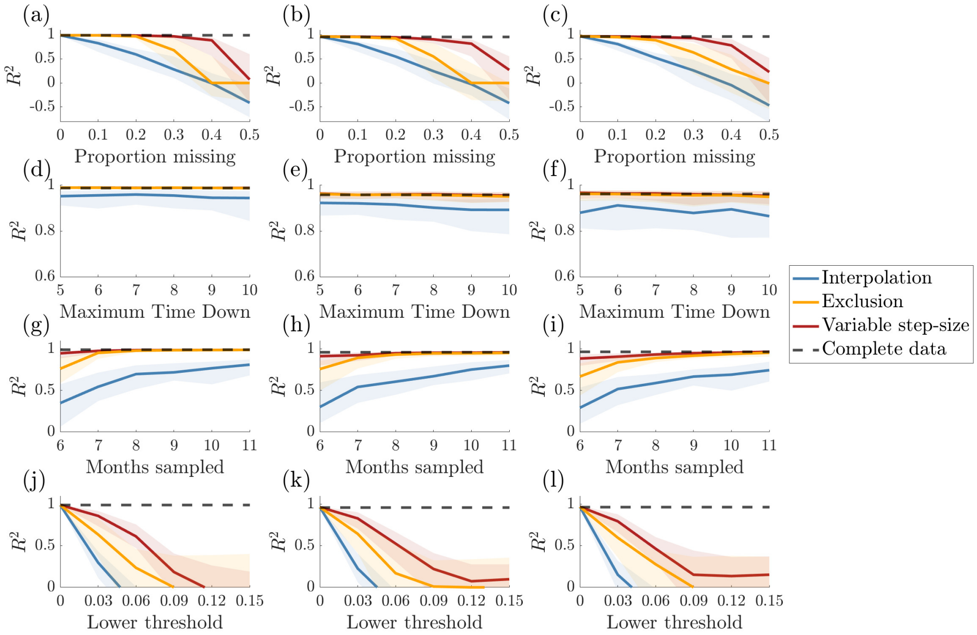

Figure 3.1: Median (solid line) and first and third quartiles (lower and upper limits of colored band, respectively) of R^2 vs. (a-c) the proportion of random missing training data, (d-f) the maximum number of consecutive time steps missing in the training data, (g-i) the number of consecutive points sampled for every 12 points in the training data, and (j-l) the lower proportional threshold at which values are missing. Results are shown for variable step-size EDM (red), the exclusion method (yellow), and the interpolation method (blue). The dashed black line shows the mean R^2 when the data have no missing points. Results are shown for (a,d,g,j) Ricker dynamics, (b,e,h,k) host-parasitoid dynamics, and (c,f,i,l) three-species competition dynamics.

Finally, in scenario V, VS-EDM provides the most accurate forecasts for all three sub-scenarios, with a median R^2 close to that of the benchmark for complete data (Figure 3.2). Generally, the interpolation method reduces the forecast accuracy significantly. The exclusion method is adequate in sub-scenarios A (mixed monthly and bimonthly sampling) and B (disjointed biweekly sampling periods) because there are usually enough data to cover the attractor. In contrast, the exclusion method suffers in sub-scenario C (weekday sampling), in which VS-EDM is 80% more accurate.

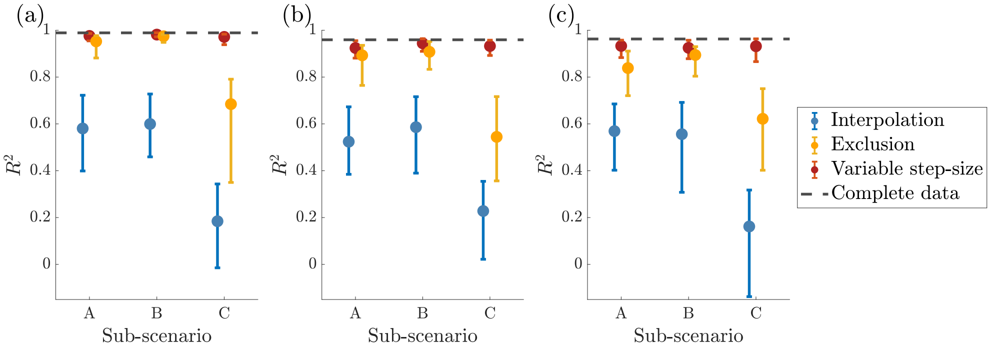

Figure 3.2: Median (dots) and first and third quartiles (lower and upper limits of error bars, respectively) for three sub-scenarios: (A) monthly sampling for seven months followed by bi-monthly sampling for the rest of a year, (B) sampling in two separate seasons of a year, and (C) sampling only on weekdays. Results are shown for variable step-size EDM (red), the exclusion method (yellow), the interpolation method (blue), and EDM of complete data (dashed black line) for (a) Ricker dynamics, (b) host-parasitoid dynamics, and (c) three-species competition dynamics.

3.4.2 Results for Empirical Data

Out-of-sample forecast accuracy (R^2) for the experimental time series from Halbach (1984) (Table 3.2) and predicted vs. observed values were evaluated (Figure B.3). Results indicate that VS-EDM can accurately forecast empirical data that are sampled irregularly. Overall, VS-EDM increased forecast accuracy for experiment 1 but made no significant difference for experiment 2. Compared to VS-EDM, linear interpolation reduced forecast accuracy by 23% for experiment 1 and 5% for experiment 2. Exclusion reduced forecast accuracy the most for experiment 1 (by 34%) but performed similarly to VS-EDM for experiment 2. Piecewise constant interpolation (i.e. ignoring the irregular step-size) performed relatively well for these data, reducing accuracy by 23% for experiment 1 and providing similar accuracy for experiment 2. Overall, since the range of step sizes was small and the largest gap was one day, all of the methods produced reasonable forecasts of these time series. We hypothesize that VS-EDM makes a larger difference for experiment 1 because its time series is not as smooth as that of experiment 2. Like for to laboratory data and simulated data, VS-EDM appears to have an advantage over the other methods for one-month-ahead forecast accuracy of English grain aphid dynamics from February 2002-January 2012 at five Rothamsted survey sites. (Figure 3.3). With 50% of the data missing, VS-EDM gives a median $R^2 > 0$ for all sites, while the other methods generally produce unusable forecasts at 50% missing. It is important to recall, however, that these results summarize 100 replicates of each scenario, and the results can depend on which points in the data are missing. Interestingly, interpolation outperforms exclusion for these laboratory data, unlike for the simulations, likely because the simulated data are highly variable and discrete-time, while the laboratory data are seasonal or smooth. When a value goes missing during a smooth period, the interpolated data are unbiased and similar to the complete data. Generally, for highly seasonal or very smooth data, interpolation is expected to provide accurate forecasts.

#	**Linear interpolation**	**Exclusion**	**VS-EDM**	**Piecewise constant interpolation**
1	0.54	0.46	0.70	0.54
2	0.63	0.67	0.66	0.66
Mean	0.59	0.57	0.68	0.60

Table 3.2: Forecast accuracy of each method on laboratory data. R^2 values for two experimental time series from Halbach (1984)

3.5 Discussion

Overall, VS-EDM handles missing data effectively in most scenarios for all simulation models we assessed. Despite doubling the effective input dimension, VS-EDM was never significantly worse, and it was often much better, than the standard approaches of interpolating the data or excluding missing inputs for all scenarios, particularly when many data points were missing. This was most clear when data were missing at random (scenario I) or had seasonal gaps (scenario III), or when the sampling design combined short and long intervals (scenario V). These results indicate that exclusion is adequate whenever data are plentiful and that linear interpolation fills short or smooth gaps effectively but creates problems when filling gaps during which substantial fluctuations occur.

All of the methods struggled when observations were missing below a threshold (scenario IV), although VS-EDM worked over a wider range of thresholds. The extreme sensitivity to missing data in this case is likely a result of the strong divergence of these models for small population sizes. When we instead removed points in a different region of the state space where the dynamics diverge less (e.g., above a given proportion of the range of states), all methods are less sensitive (Appendix B.3). Furthermore, although we treated observations below a threshold as missing, they are often recorded as zeros in real data. That is, when samples are taken but nothing is detected (e.g., low chlorophyll-a concentrations), the observed abundance is zero even though the true abundance is not. Zero-inflated models are often used to handle the large number of zeros in general linear models (Hall and Zhang, 2004), but no analogous method is available for EDM. Lacking a zero-inflated model, one may ask whether it is better to treat these zeros as missing data using VS-EDM or as ordinary observations. We tested this idea by comparing the forecast accuracy of VS-EDM when values below the threshold are treated as missing versus using standard EDM when values below the threshold are treated as zero (Appendix B.3). Broadly, we found that the results are relatively case-specific; they likely depend on the curvature in the delay embedding map over the interval treated as zero. Consequently, it may be valuable to apply both approaches when zeros are inflated in a real application.

Although we have shown that VS-EDM produces accurate forecasts in a variety of scenarios, it is useful to consider cases in which one of the other methods may be a better choice. Most obviously, when data are plentiful (scenario II or a very long time series), the exclusion method will be sufficient and less computationally expensive than VS-EDM. At the other extreme, interpolation or ignoring the irregular step may be more useful for short, smooth time series with short gaps, as exemplified by the rotifer data from Halbach (1984). In these cases, VS-EDM is likely to suffer since it requires estimating E additional hyperparameters from the limited data.

As with all EDM approaches, the performance of VS-EDM may depend on time series length. Although not a major focus of this study, we reevaluated VS-EDM for time series lengths of 20, 35, and 50 training points (Appendix B.4). As expected, shorter time series consistently yield worse predictions; however, it is generally better to have 25 points of complete data than 25 scattered points from a time series of length 50 (i.e., 50% missing). This shows that increasing the input dimension using VS-EDM can be unfavorable for extremely short time series.

Some data sets have only a few observations at a given τ (e.g., a single inoperative period, a few points missing at random), making it challenging to approximate f using the available data. In principle, it may be possible to circumvent this by inserting artificial missing values to make a longer set of training data with a wider range and more replicates of s in different areas of the attractor. We tested this approach on the three models (Table 3.1), and our preliminary results indicate that it can improve forecast accuracy greatly (Appendix B.5). Specifically, when we used standard VS-EDM with these models, forecasts were typically unusable once 50% of the data was missing. However, augmenting the training data with artificial missing values produced accurate forecasts (up to $R^2 = 0.80$) at 50% missing (Figure B.10). Although our preliminary tests show promise, there are statistical concerns with reusing the data in this way. By duplicating the time series to create artificial missing values, the outputs are not independent, which may bias GP fitting. More work is needed on this topic to determine whether inserting artificial missing values is a reliable way to supplement VS-EDM.

Having demonstrated its utility for simulated and empirical data, we now highlight two additional ideas that emerge from the VS-EDM framework but have not appeared in the literature. First, although we focus on forecasting in this study, EDM is increasingly used to understand species interactions (Deyle et al., 2016) and dynamic stability (Ushio et al., 2018). One of the most widely used metrics of dynamic stability is the Lyapunov exponent (LE). Since previous approaches to calculating LE require equal sampling intervals, we demonstrate how to calculate the dominant LE using VS-EDM (Appendix B.6). Second, although EDM is typically applied to discrete-time data in ecology, many systems are best understood in continuous time. Another facet of VS-EDM that has not appeared in the literature is that the variable step-size method could be used to extrapolate back to continuous-time dynamics from discretely sampled data. We present an outline of this framework and a preliminary test of the idea (Appendix B.7). Being able to flexibly extract continuous-time dynamics from discretely sampled data would be particularly useful for connecting theoretical models expressed using ordinary differential equations to field observations; this approach will be the subject of a future study.

To summarize, VS-EDM has several useful implications in ecology, the most obvious of which is that it expands the range of ecological datasets that can be analyzed with EDM to those in which some data are missing or irregularly sampled. Moreover, this analysis can help mitigate effects of lost sampling effort due to the COVID-19 pandemic and implies that more flexible and less expensive ecological surveys could be designed without sacrificing significant forecast accuracy. For instance, VS-EDM forecasts differ little when sampling occurs every day vs. only on weekdays. We presented several possible extensions of VS-EDM that allowed us to characterize the complexity of ecological dynamics and approximate continuous-time models. Together, these results can contribute to advancements in understanding and managing complex ecological systems.

3.6 Acknowledgments for Chapter 3

This study was supported by funding from the Lenfest Oceans Program, NOAA's HPCC Program, and NOAA Grant #NA19OAR4170353 through the Fisheries-Sea Grant Joint Graduate Fellowship Program in Population and Ecosystem Dynamics. The statements, findings, conclusions, and recommendations are those of the authors and do not necessarily reflect the views of Sea Grant. We are grateful to collaborators, including Chris Shortall, James Bell, and Lynda Alderson, who contributed data from the Rothamsted Insect Survey, a National Capability, which is funded by the Biotechnology and Biological Sciences Research Council under the Core Capability Grant BBS/E/C/000J0200. We also thank Marcella Gomez, George Sugihara, Erik Saberski, Tom Lorimer, Ethan Deyle, Tanya Rogers, Uttam Bhat, Marc Mangel, and Hongyun Wang for thought-provoking conversations about these ideas and for many helpful comments on this

manuscript.

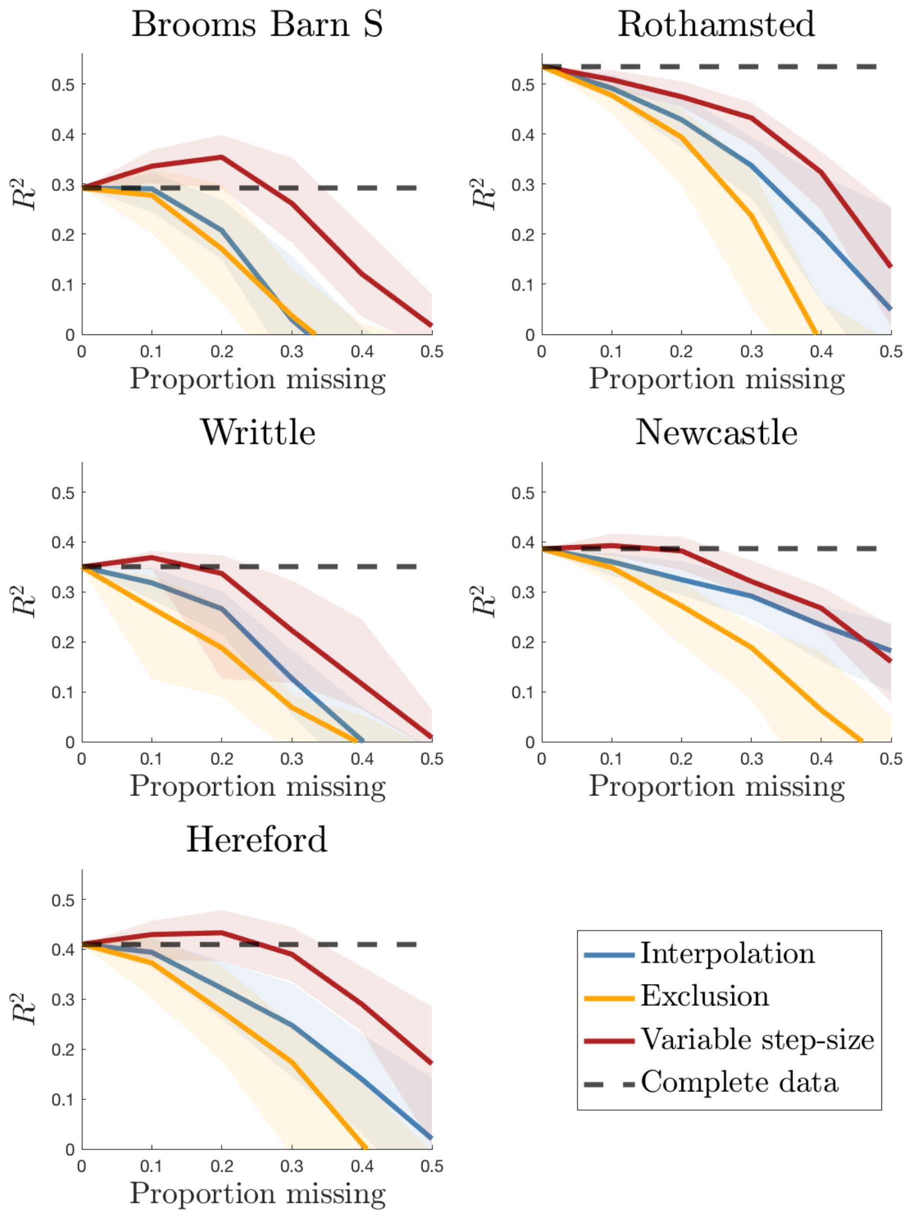

Figure 3.3: Median (solid line) and first and third quartiles (lower and upper limits of colored band, respectively) of R^2 values for out of sample forecasts of the English grain aphid from February 2002-January 2012 at five sites (Brooms Barn S, Rothamsted, Writtle, Newcastle, and Hereford). Results are shown as the proportion of missing values in the training data was varied for the interpolation method (blue), the exclusion method (yellow), variable step-size EDM (red), and EDM of the complete data (dashed black line).

Chapter 4

Empirical Dynamic Programming for Insect Pest Management

The material in this chapter will be submitted to the Proceedings of the National Academy of Sciences by Johnson, B., Gomez, M., Munch, S.B. The dissertation author was the primary investigator and author of this paper.

4.1 Abstract

Insect pests pose a major threat to humans by jeopardizing food security in agricultural systems, acting as vectors for infectious diseases, and damaging forests and other ecosystems. Despite decades of research aimed at controlling pest populations to mitigate their harmful impacts, effective pest management remains challenging in many systems. This stems, in part, from incomplete knowledge of the mechanisms that drive population dynamics, making it difficult to develop accurate models that predict insect outbreaks. Due to the challenges of mechanistic modeling and historical tendencies of pest managers, most theoretical developments in this space have yet to meet practice. Pest management is often reactive in practice, meaning control actions are taken once outbreaks have already begun, allowing for damage to occur. It is possible to improve pest management, however, by acting in anticipation of an outbreak. We show that a data-driven model can effectively predict outbreaks, thereby, circumventing the need to understand the underlying network driving population dynamics. This allows us to target pests before outbreaks occur. We also show that optimal control can be used with our data-driven model to optimize pest management strategies taking into consideration cost of application. In particular, we explore the use of empirical dynamic modeling and Gaussian process regression paired with stochastic dynamic programming to keep insect populations within acceptable bounds of tolerance. We show that this framework effectively prevents outbreaks on simulated and empirical data in a variety of scenarios. Our study provides a management framework that has potential to reduce losses from pests.

4.2 Introduction

Insect outbreaks have significant consequences. Every year, insects destroy approximately 18-20% of major crops worldwide (Sharma et al., 2017), which account for over 40% of calories consumed by humans globally (Deutsch et al., 2018; FAO, 2014). Insects also spread diseases, which cause more than 700,000 deaths annually (WHO, 2020). In addition, insects disturb up to 85 million hectares of global forest area per year (van Lierop et al., 2015) and put a large proportion of the total U.S. forest biomass at risk for invasion (Fei et al., 2019), which can have cascading effects that contribute to climate change. Together, the damages and mitigation efforts bring the total estimated cost of pest outbreaks to at least $76 billion per year globally (Bradshaw et al., 2016), and these costs are expected to increase as climate change progresses (Deutsch et al., 2018).

Given the high impact of the insect pests, there is a growing need to work toward minimizing the damage they cause and the cost of pest management. In an effort to achieve this, over the last several decades, there have been increased efforts in integrated pest management (IPM) research, which aims to use information about the state of the environmental, economic, and social system to guide decisions for the use of biological (e.g. addition of a natural enemy of the pest), behavioral (e.g. the use of pheromones to disrupt mating behavior and hinder reproduction), and chemical (e.g. application of pesticides) control recommendations to suppress pest populations (Stern et al., 1959). A successful IPM program involves carefully monitoring the pest and the damage it causes, using that observational data to set guidelines for when control is needed, and assessing the effect of the control techniques (Dara, 2019; Kogan, 1998). However, since ecological systems are highly complex with numerous interacting variables, common modeling methods cannot always accurately predict when pests pose the highest risk of an outbreak and, more importantly, how various control actions impact the pest dynamics (Garrett et al., 2013; Tonnang et al., 2017). This – among other obstacles (Parsa et al., 2014) – has led to slow and weak adoption of IPM in practice (Stenberg, 2017).

Instead, in practice, pest management often follows a reactive program or a calendar-based program. In a reactive program, farmers or local governments are advised to apply control once the population of a pest exceeds an unacceptable threshold (Stern, 1973), so interventions often occur once damage has already begun (Oliver and Roy, 2015). A calendar-based program, on the other hand, errs on the side of caution. Interventions are made at regular intervals throughout a year or growing season, often leading to the overuse of pesticides (Afun et al., 1991). Although both of these management programs are fairly common, there are limitations to each and there is likely room for improvement.

Some theoretical ecologists have proposed to improve pest management by using optimal control theory. Several examples in the literature use classic population dynamics models with control methods such as Pontryagin's Maximum Principle (Fitri et al., 2021; Kar et al., 2012; Whittle et al., 2008), model predictive control (Zangina et al., 2021), and dynamic programming (Hackett and Bonsall, 2019; Yokomizo et al., 2009) to generate pest management strategies. However, optimal control methods for pest management have primarily been developed in noise-free theoretical settings, and the studies have assumed that we have a perfect model of the system dynamics. In reality, we rarely have a model that describes system dynamics perfectly, and we must make many assumptions. Since optimal policies are highly sensitive to model structure, small changes in assumptions can lead to drastically different management advice (Wood and Thomas, 1999). Furthermore, these methods tend to yield advice that is not feasible in real applications (e.g. they suggest applying pesticides in continuous time (Whittle et al., 2008)) and theoretical studies rarely incorporate empirical data to validate their methods. Together, these issues have led to skepticism from farmers and managers (Deguine et al., 2021). In fact, one recent study claimed that although ecological theory holds promise for contributing to pest control, it has produced minimal practical value in the past (Mcevoy, 2017).

In this paper, we aim to bridge the gap between theory and practice through a data-driven approach to prediction and control. Importantly, the idea of combining data-driven forecasts with optimal control methods has been used to make short-term management decisions and identify optimal management policies in fisheries contexts (Boettiger et al., 2015; Brias and Munch, 2021). However, to the best of our knowledge, data-driven control in insect pest management is not widely used and has had limited exposure in the literature (Meisner et al., 2016). Ecosystems are highly complex, and with our current understanding, it is not feasible to take a bottom-up approach of developing mechanistic models to predict pest dynamics and intervention effects. However, it might not be necessary to construct complete ecosystem models to effectively control pests. Data-driven models have improved predictive capabilities over mechanistic models when population dynamics are complex and external perturbations are not modeled (Munch et al., 2018). In this work, using only time series for species abundance and historical control actions, we consider a Gaussian process (GP) model (Rasmussen and Williams, 2006) to forecast population dynamics and their responses to control. To take it a step closer to reality, even when we do not have data for the full system and we only have a time series for the pest, we use empirical dynamic modeling (EDM), which utilizes lags of the pest data (Sugihara and May, 1990; Takens, 1981) (Material and Methods), to make the predictions. This approach is advantageous for this application because it allows us to avoid making strong assumptions about insect dynamics, it allows us to cope with incomplete observations of the system, and it is more flexible than standard mechanistic models (Perretti et al., 2013). By leveraging the data-driven forecasts, we solve a multi-objective optimal control problem with stochastic dynamic programming (SDP) to generate control policies (Bellman, 1958; Clark and Mangel, 2000). This combination of EDM with SDP is called empirical dynamic programming (EDP)(Brias and Munch, 2021). See Figure 4.1 for an illustration of the EDP workflow.

To demonstrate this idea, we used a series of simulations that represent common pest management strategies I) biological control, II) chemical control, III)

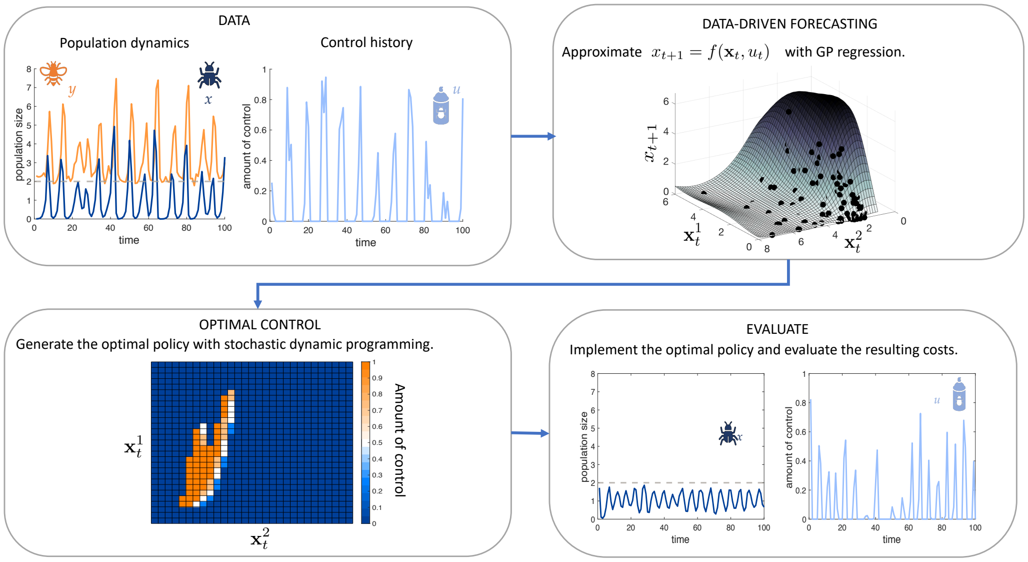

Figure 4.1: A 2D example of data-driven prediction and control for pest management. Ecosystem dynamics in the real world (top box) include the population dynamics of a pest (x), a species that interacts with the pest (y), and human interventions in the form of insecticide sprays (u) which control the pest. We consider two cases of data availability. In case 1, data for the full ecosystem (x, y, u) are available, and we perform "full-state" EDP. In case 2, data for only the pest and interventions (x, u) are available, so we implement "partial-state" EDP and use lags to account for the unobserved variable. In both cases, the first step is data-driven forecasting (second box) where we fit the function $x_{t+1} = f(\mathbf{x}_t, u_t)$ with GP regression (surface) to the available data (black dots). In the full-state scenario, $\mathbf{x}_t = [x_t, y_t]$ (i.e. $\mathbf{x}_t^1 = x_t$ and $\mathbf{x}_t^2 = y_t$). For the partial-state scenario, $\mathbf{x}_t = [x_t, x_{t-1}]$ (i.e. $\mathbf{x}_t^1 = x_t$ and $\mathbf{x}_t^2 = x_{t-1}$). Given the predictive model, the next step is to determine the optimal control policy with stochastic dynamic programming (third box). This generates the optimal action u (color) to take in any given state $\mathbf{x}_t = [\mathbf{x}_t^1, \mathbf{x}_t^2]$. Finally, we implement the optimal policy through time (bottom box). At each time step, we determine the current state, $\mathbf{x}_t$, and apply the optimal action (light blue) to the ecosystem. This results in new dynamics for the pest (dark blue), which we compare to an economic threshold (grey dotted line) to evaluate cost. The total amount of control also contributes to the cost.

behavioral control, and IV) IPM (e.g. the use of both chemical and biological controls) (Table C.1). We compared the performance of our data-driven approach against 1) the optimal policy, which uses the perfect data-generating model to generate policies and 2) a policy based on a reactive program, which only applies control after the pest has exceeded an unacceptable threshold. These benchmarks represent the best case scenario and status quo, respectively. We evaluated the importance of stochasticity by simulating all scenarios with various levels of noise. In addition, we compared trade-offs between two competing objectives, (1) minimize the cost of pest pressure and (2) minimize the cost of applying interventions.

For completeness, we followed the simulation analysis with an empirical analysis in which we applied the method to data for a cotton pest, *Lygus hesperus*, and a West Nile virus vector, *Culex pipiens*, in California. This empirical analysis allowed us demonstrate the predictive power of our models on ecological data which contains realistic levels of noise and imperfections, and it also allowed us to begin evaluating the utility of the control method in real management scenarios.

Through our simulation and empirical studies, we show that this method offers potential to improve pest management practices for a wide range of species and systems.

4.3 Material and Methods

There are two main steps required to develop data-driven pest management decisions. First, we use historical time series data and GP regression with EDM to approximate a function which takes the current state of a system along with a control variable as inputs, and outputs forecasts of the future pest population. Then, we use the approximated function to determine a control policy via stochastic dynamic programming. In this section, we will first describe our process of approximating the model with GP regression and EDM, then we discuss details of the optimal control problem and how to solve it with dynamic programming. Finally, we provide details about our simulation experiments and empirical case studies.

4.3.1 Gaussian process regression

We use GP regression (Munch et al., 2017; Rasmussen and Williams, 2006) to approximate the function, $x_t = f(\mathbf{x}_{t-1}) + \epsilon_{t-1}$ where $\epsilon_{t-1} \sim N(0, V)$, u_{t-1} is the control action at time $t-1$, and $\mathbf{x}_{t-1}$ are the inputs containing each state variable at time $t-1$ if we have data for the full system state. If the full system state is not available, EDM is necessary, so $\mathbf{x}_{t-1} = [x_{t-1}, x_{t-2}, ..., x_{t-E}]$ (see next section).

The GP is specified by a mean function $m(\mathbf{x})$, and a covariance function $C(\mathbf{x}, \mathbf{x}')$. To remain consistent with previous applications of GP regression in ecology (e.g. (Brias and Munch, 2021; Munch et al., 2018, 2017; Rogers and Munch, 2020)), we begin by standardizing the input data of the GP to have a mean of zero and a standard deviation of one. We set priors for the mean and covariance functions and update them based on the observed data. Specifically, we set the prior mean function to $m(\mathbf{x}) = \mathbf{0}$ assuming that we do not have prior knowledge about the functions characteristics, and we use the squared-exponential covariance function, $C(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp(-\sum_{i=1}^{n} \phi_i (x_i - x_i')^2)$, where σ^2 is the pointwise variance in the function f, n is the dimension of each input (i.e. dimension of the system for the full-state case or number of lags for the EDM case), and ϕ_i is the input-specific inverse length scale parameters, which govern the flexibility of the function. Note that when $\phi_i = 0$, f is constant in the i_{th} input direction. In order to avoid overfitting, we use a regularization technique is called 'automatic relevance determination' (ARD) (Neal, 1996). This encourages sparsity by assigning half-normal priors such that the prior mode is at 0. Priors for V and σ^2 come from a beta distribution $\beta(1.1, 1.1)$ to place limits on the uncertainty in the next step (see (Munch et al., 2017) for details). We update the hyperparameters using resilient back propagation (Rprop) (Blum and Riedmiller, 2013) to maximize the marginal log likelihood. Although it is possible to infer f in a fully Bayesian way (e.g. with MCMC), we make predictions by fixing the hyperparameters at their posterior modes in order to save computation time, and then predictions are made by using updating rules that come from standard formulae for conditioning in a multivariate normal distribution (Rasmussen and Williams, 2006).

4.3.2 Empirical dynamic modeling

Empirical dynamic modeling is a data-driven method that uses time series from partially observed systems to make forecasts. In cases where we do not have observations of all of the variables in a system, Takens' theorem of time delay embedding (Takens, 1981) provides the foundation for EDM by stating that the attractor (i.e. the point, set of points, or orbit to which a dynamical system converges) can be reconstructed using lags of the time series of a single state variable. From the attractor reconstruction, the future state of the system can be predicted by fitting a model of the form $x_t = f(x_{t-1}, x_{t-2}, ..., x_{t-E}) + \epsilon_{t-1}$ where x is the observed state variable, E is the 'embedding dimension', ϵ is the process error, and f is the map taking lags of the state variable to the future. Many function approximation methods can be used (e.g., local linear regression (Farmer and Sidorowich, 1987; Sugihara, 1994) or neural networks (Bakker et al., 2000). In this paper, we fit f via Gaussian process regression (Munch et al., 2017). More examples and overviews of EDM can be found in (Chang et al., 2017; Munch et al., 2020; Ye et al., 2015a), and extensions of EDM for stochastic systems can be found in (Munch et al., 2020; Stark et al., 2003).

4.3.3 Using forecasts for optimal control

Importantly, the EDM frameworks can be extended to include additional covariates. For instance, previous studies have incorporated multiple interacting species, environmental drivers, and spatial replicates (Deyle and Sugihara, 2011; Johnson et al., 2021; Rogers and Munch, 2020; Ye and Sugihara, 2016). If we include historical management actions (e.g. pesticide applications, biological control releases) as covariates, this opens up the possibility of using EDM to construct robust management policies. Concretely, suppose we would like to control the dynamics of an insect population by applying pesticides and we are given the time series of previous insect abundance (x_t) and pesticide applications (u_t) for $t = 0, 1, ..., T$. For simplicity, assume that the insect is not interacting with any other species or drivers. Using the observed time series, we can a fit a GP to model $x_t = f(x_{t-1}, u_{t-1}) + \epsilon_{t-1}$, which results in an approximation of the Markov decision process (MDP), $P(x_t|x_{t-1}, u_{t-1})$. This can then be used directly to solve an optimal control problem and generate management policies.

4.3.4 The optimal control problem

Once we have approximated the dynamics with GP regression, our goal is to minimize the cumulative discounted cost of pest management from the current time until some time in the distant future. We can express this goal as an optimal control problem given by

$$\begin{gathered}\min_{u} J(\mathbf{x},\mathbf{u}) = \sum_{t=1}^{\infty} \gamma^t c(\mathbf{x}_t, \mathbf{u}_t) \\ \text{subject to} \\ \mathbf{x}_t = f(\mathbf{x}_{t-1}, \mathbf{u}_{t-1}) \\ \mathbf{x}(0) = x_0 \end{gathered} \tag{4.1}$$

where $\mathbf{x}$ is the state variable (e.g. pest abundance and enemy abundance), $\mathbf{u}$ is the control variable (e.g. amount of pesticide applied), γ is the discount factor, f represents the system dynamics, and c is the step-wise cost of being in state $\mathbf{x}_t$ and taking control action $\mathbf{u}_t$.

One important step in setting up an applied optimal control problem is specifying the cost function. In pest control applications, there are typically two primary objectives: (1) to minimize the cost caused by pests (e.g. through yield loss), and (2) to minimize the cost of applying interventions. There is a trade-off between these two objectives, and we might adjust their priority in different scenarios. To capture these competing objectives, we define a cost function given by

$$c(x_t, u_t) = \theta \frac{x_t}{1 + e^{-10(x_t - x_{thresh})}} + (1-\theta)u_t. \tag{4.2}$$

The first term in Eq. 4.2 captures the cost associated with pest pressure where the cost is negligible if the abundance of the pest (x_t) is below a threshold (x_{thresh}) and high above it. This is similar to the "economic threshold" or “action threshold” concept that is used in practice, except that it is based on the abundance of the pest at a future point in time rather than the current state. This facilitates proactive control rather than reactive control. We use the threshold (x_{thresh}) in the first term of the cost function because it has been stated that pest management methods should aim to drive pest populations within acceptable bounds rather than eliminate them completely (Lewis et al., 1997). The second term in Eq. 4.2 captures the cost of applying control, assuming that the cost is linearly related to the amount of control applied. θ is used to tune the priority of the competing objectives by varying it from 0 to 1. For example, $\theta = 0.99$ places high priority on minimizing the cost of pests.

The IPM strategy involved two control variables, so the step-wise cost was in the form

$$c(x_t, u_t^c, u_t^b) = \theta_1 \frac{x_t}{1 + e^{-10(x_t - x_{thresh})}} + \theta_2 u_t^c + \theta_3 u_t^b \tag{4.3}$$

where u^c is the chemical control variable, u^b is the biological control variable, and θ_i places weight on the i^{th} term of the cost function.

4.3.5 Dynamic programming

Bellman's principle of optimality (Bellman, 1958) shows that the optimal control problem Eq. 4.1 can be solved via the dynamic programming equation

$$V(\mathbf{x}) = \min_{\mathbf{u}} E\left[c(\tilde{\mathbf{x}}, \mathbf{u}) + \gamma V(\tilde{\mathbf{x}}) | \mathbf{x}, \mathbf{u}\right]$$

where $V(\mathbf{x})$ is the long term discounted cost of being in state $\mathbf{x}$, $c(\tilde{\mathbf{x}}, \mathbf{u})$ is the cost of applying $\mathbf{u}$ and ending up in state $\tilde{\mathbf{x}}$, γ is the discount rate. E is the expectation over the next state given the previous state and control.

We solve the dynamic programming equation with value iteration, which is done with the following procedure. (1) Create an evenly spaced grid of possible states; we use 30 evenly spaced points between 0 and 1.25*max(x) in the training data. (2) Create an evenly spaced grid of possible actions u; we use 5 evenly spaced points between 0 and 1. 3) To solve the DP equation, we approximate the expectation as

$$E\{c(\tilde{\mathbf{x}}, \mathbf{u}_l) + V(\tilde{\mathbf{x}})|\mathbf{x}_j, \mathbf{u}_l\} \approx \sum_{i}^{\text{size(state grid)}} P_{jl}(\mathbf{x}_i)\{c(\mathbf{x}_i, \mathbf{u}_l) + \gamma V(\mathbf{x}_i)\}.$$

In our case, we estimate $P_{jl}(\mathbf{x}_i)$ with GP regression, though this can also be inferred with a mechanistic model. 4) From the dynamic programming equation, update the value function as $V(\mathbf{x}_j) = \min_{\mathbf{u}} E\{c(\tilde{\mathbf{x}}, \mathbf{u}_l) + V(\tilde{\mathbf{x}})|\mathbf{x}_j, \mathbf{u}_l\}$, and also extract the policy as $u^*(\mathbf{x}_j) = \arg\min_{\mathbf{u}} E\{c(\tilde{\mathbf{x}}, \mathbf{u}_l) + \gamma V(\tilde{\mathbf{x}})|\mathbf{x}_j, \mathbf{u}_l\}$. 5) Iterate steps 3) and 4) until the approximated value function and the optimal policy converge. Convergence is achieved when the change in the value function and optimal policy between successive iterations is below some tolerance, which we set to 0.001.

Note that as the dimensionality of the system increases, we must create a grid of possible states for each state variable. Thus, dynamic programming has exponential computational complexity making it suffer greatly from the curse of dimensionality. In high dimensional cases, where dynamic programming is not computationally tractable, other control methods such as model predictive control (i.e. ignoring future costs) can be used, though this will not yield an optimal solution.

4.3.6 Simulations

To evaluate our methods, we simulated population dynamics data with a host-parasitoid model, which in the absence of control, is given by

$$H_{t+1} = H_t e^{r(1-H_t/K)-\alpha P_t}$$

$$P_{t+1} = \beta H_t(1 - e^{-\alpha P_t}) + \gamma$$

where r is the growth rate of the host, K is the carrying capacity of host, β is the searching efficiency of the parasitoid, is the number of parasitoids that emerge from each parasitized host, and γ is a migration coefficient for the parasitoid. We assume that H represents the pest population that we aim to control and P is a natural enemy of the pest (Hassell, 2000; Jang and Yu, 2012).

There are various ways in which a manager might try to control an insect pest population. To evaluate generalizability of EDP, we simulated multiple control strategies. See Table C.1 for functional forms and parameter values of each control scenario. A description of each control is given below.

I. Biological control

There has been substantial effort to develop control techniques that can suppress pests naturally while causing minimal damage to the surrounding environment. One option is to introduce natural enemies of the pest into the system to increase pest mortality. In the context of the host-parasitoid system, biological control is modeled by adding parasitoids to the system (Table C.1).

II. Chemical control

We simulated the effects of chemical control (i.e. insecticide applications) as a direct reduction in the pest population. Since it is rare for insecticide applications to remove all pests in the system, we assumed that only a fraction of the population was removed after a pesticide application. We set this "pesticide efficiency" parameter (i.e. the maximum proportion of the population that can be reduced by spraying pesticides) to 0.6 (Table C.1).

III. Behavioral control

Another non-chemical form of pest control is behavioral control in the form of mating disruption (MD). The goal of this type of control is to decrease the growth rate of the pest population, by releasing pheromones into the pests' environment, which makes it more difficult for males to find females. Another method used to decrease the growth rate of a pest is sterile insect technique (SIT), in which sterile insects that cannot effectively reproduce with females are released into the environment. While this isn't strictly behavioral control, it does impact the pest's ability to reproduce. We modeled both these control methods as a reduction in the growth rate of the pest. The parasitoid was not effected by this method (Table C.1).

IV. Integrated pest management

Although more natural forms of control are desirable for sustainable pest management, there is a lot of experimental work that needs to be done to determine the efficacy and reliability of these methods. The development of natural pest control is still an active area of research, and as a result, many insect pest managers are unlikely to completely abandon the use of chemicals to control insects. A more reasonable approach is to use some combination of chemicals and natural approaches to optimally suppress pests. This is one of the primary goals of IPM. We simulated an IPM approach by using a combination of chemical control and biological control to suppress the pest. As a result, instead of approximating $x_t = f(\mathbf{x}_{t-1}, u_{t-1}) + \epsilon_{t-1}$ with GP regression, we approximated $x_t = f(\mathbf{x}_{t-1}, u^c_{t-1}, u^b_{t-1}) + \epsilon_{t-1}$, where u^c is chemical control, and u_b is biological control. The objective was to find the optimal combination of chemical and biological control to minimize a cost function given by 4.3, where there is a trade-off between all three therms in the cost function.

We simulated strategies I-IV above with three levels of process noise (low, moderate, and high) in the dynamics (Table C.1). In all scenarios, we simulated 600 data points. After removing 300 transient points, we split the remaining 300 points into 100 points of training data, 100 points of testing data to evaluate the R^2 forecast accuracy, and 100 points of secondary testing data to evaluate the control methods. We call this last set of points the "control" set. To generate the initial time series, we applied a random amount of control periodically (to approximately target the peaks in the pest population) in the training and testing sets and applied no control in the control set. For strategies I-III, we evaluated the methods at eleven evenly spaced values of θ between 0.001 and 0.999. For strategy IV, we evaluated methods at 21 different values of $(\theta_1, \theta_2, \theta_3)$ where $\theta_1+\theta_2+\theta_3 = 1$.

Our ability to effectively manage the pest population depends on the quality and type of data that is available. Thus, we explored two cases of available data.

Case 1: Data for both species - "full-state" EDP

In the first case, we assumed that we had access to the abundance of both the host and the parasitoid and the control variable in the training data. Given the training data, we fit a GP to approximate $[H_{t+1}, P_{t+1}] = f(H_t, P_t, u_t)$, then we tested the forecast accuracy on the subsequent testing data. Then, we performed the dynamic programming method with the steps outlined above to get the optimal policy for all possible combinations of H and P. To evaluate the dynamic programming policy, we continued the simulations in the control set by using the same parameters from the original simulation, but added a control input following the optimal policy. Note that since the simulation models gave outputs that did not fall directly on the grid that we used for the SDP algorithm, we performed a linear interpolation of the policy, to get the optimal control for each current state. We iterated this procedure in the simulation model for the 100 steps in the control set.

Case 2: Data for only the pest - "partial-state" EDP

Next we assumed that the parasitoid was unobserved, and we only had access abundance for the pest in the training data. In this case, EDM was necessary since we had incomplete observations, so we fit a GP to approximate $H_{t+1} = f(H_t, H_{t-1}, u_t)$. We performed the rest of the procedure including forecasting in the test set and policy evaluation in the control set as we did with Case 1. Contrary to Case 1, however, we did not evaluate this method with strategy IV, IPM. This is because in an IPM program, managers try to monitor and understand as much of the complex system as possible, and it is unlikely that a manager would introduce a natural enemy as a form of control without collecting data for the natural enemy first.

To evaluate the EDP policies, we compared them against two benchmarks. The first benchmark was the optimal dynamic programming policy when the true model was used to estimate the transition probabilities. This represented the best case scenario where we have perfect knowledge of the system. The second benchmark was a reactive control in which control was applied when the current state of the pest was above the specified threshold. The amount of control was proportional to θ. The comparison with these benchmarks was done by calculating average "excess cost" of EDP and the reactive method of control (See Results).

4.3.7 Empirical case studies

While the simulation analysis helped us evaluate the success of EDP in cases where we can manipulate the ground truth dynamics, in most real-world cases, there are numerous other factors that influence the dynamics which we did not include explicitly in simulations. Since these other factors can influence the control algorithms, we further tested our method on empirical datasets. The goal of this empirical analysis was two-fold. First, to evaluate whether the EDM algorithm could accurately predict the dynamics of real-world insect pest data and the impact of interventions on those pests. Second, to begin exploring whether historical control efforts in real systems are nearly optimal. Of course, it is important to note that we cannot truly validate a control policy on historical data.

Following our simulation experiments, we tested our control method on one agricultural pest and one insect that acts as a disease vector. The agricultural pest was *Lygus hesperus* from commercial cotton fields in central California. *Lygus hesperus* threatens cotton yields by damaging squares of the cotton plant early in the growing season (IPM, 2013). The data included samples from 1997 to 2008 from 567 cotton fields. Together, there were samples from 1500 field-year combinations. Approximately every week during the growing season (∼June to August), pest control advisors (PCAs) or growers took samples by swinging nets across the top of the cotton plant. The mean number of *Lygus* caught in the net per 50 swings is called a “sweep” sample and was used as the density estimate to monitor the population of *Lygus*. PCAs and growers also tracked the active ingredients, targets, and timing of various management interventions in every field.

In our analysis, we treated each field-year as an individual time series. Since EDM requires fairly long time series to make accurate forecasts, we filtered out all time series that had fewer than 15 *Lygus* samples. This resulted in 142 time series to analyze. To supplement the population density data with control data, we included insecticide sprays as a binary variable. That is, at times when *Lygus* was listed as a target in an insecticide application, the control variable was assigned a value of 1 and assigned 0 otherwise. We randomly selected 80% of the field-year time series for training the GP EDM model. The other 20% were held out for testing forecast accuracy and evaluating control policies. After square root transforming the data, we fit a EDM model with $E = 2$ and an additional input for the control variable to the training data, and evaluated the R^2 for one-step-ahead forecasts in the testing data.

Once we tested the forecast performance on out-of-sample data, we evaluated the dynamic programming policy. To do so, we defined a cost function based on the guidelines provided by the University of California Agriculture and Natural Resources Statewide Integrated Pest Management program (UC IPM) for *Lygus* management (IPM, 2013). UC IPM suggests sliding thresholds because *Lygus* densities increase steadily during the growing season, and cotton becomes more resilient against *Lygus* later in the season. Thus, we defined different cost functions for distinct periods of the growing season. All cost functions were given by Eq. 4.2, where we set $x_{thresh} = 1$ for the early squaring period of the season (∼early June), $x_{thresh} = 2$ starting in mid-June, and $x_{thresh} = 8$ during mid-squaring (∼early July). We also set $p = 10$ and varied θ from 0.4 to 0.999. We followed the dynamic programming steps as outlined above. However, in step 2, we used a grid only two points, 0 and 1, to indicate "no spray" and "spray," respectively. This resulted in an optimal policy for each state and each period of the growing season. We evaluated the policy on each field-year time series separately. Specifically, for each field-year, we initialized the input with the real data input at the beginning of the season. Using the output from the dynamic programming optimal policy, we determined the best control action given the current state. With the current state and optimal control rule, we used the optimized GP to get the next state. This was used as the new input for the next step, and we repeated that procedure for the whole time series, using the appropriate policy for different parts of the season. This process was done for all time series in the testing data. To evaluate the output, we tracked the total number of sprays and the total number of *Lygus* predicted to emerge above the sliding threshold by following the EDP policy, and added the results over all of the series in the test data. We got a separate answer for each value of θ, and used this to construct a Pareto front. To begin evaluating the historical control efforts, followed a similar procedure. For each field-year, we initialized the input as we did before. Then we plugged in the historical control action into the GP to get the next state and iterated this procedure for all time series in the testing data. With that, we were able to compare the total number of sprays and emerging *Lygus* with the dynamic programming output.

The second case study involved the mosquito species *Culex pipiens* in California. Culex pipiens is a vector of West Nile virus (WNV), which is an endemic in California and can cause long term physical and mental disabilities or death (Holcomb et al., 2021). Because of its potential harmful impacts on humans, aerial applications of pesticides are used to target adult populations of mosquitoes, to slow or prevent the transmission of WNV (Carney et al., 2008).

In our example, we used mosquito trap count data from California's Sacramento county provided by CalSurv, a mosquito data management system. In addition to the count data, we used data on historical pesticide application records from the Sacramento-Yolo Mosquito and Vector Control District (SYMVCD) from 2006 - 2017. In this analysis, split the data into six spatial regions (Figure C.5) where local dynamics were likely to occur. In each location, we took monthly aggregates of the average number of mosquitoes caught per trap per night in the data, which resulted in six time series of mosquito abundance. We supplemented this with control data in the form of a binary variable (i.e. 1 when a pesticide was sprayed in that location, zero otherwise). We used a hierarchical GP regression (Munch et al., 2017) to approximate the future growth rate of the mosquito as a function of the previous abundance and the seasonality term, i.e.$\log\left(\frac{x_{t+1}}{x_t}\right) = f(x_t, \sin\left(\frac{2\pi}{12}n + c\right), u_t)$, where n is the month of the year. We evaluated the R^2 for sequential one-step-ahead forecasts for the full time series from 2006-2017, where we initialized the input at each site with the true data and then computed one-step-ahead forecasts, using each prediction as input for the subsequent prediction. Similar to the *Lygus*, we constructed a Pareto front for the EDP policy and compared it to the historical policy by iterating the historical actions through the GP model.

4.4 Results

We compared EDP to the benchmarks by evaluating the "excess cost" of the control policies. We considered two cases for EDP, (1) full-state EDP (trained on data for all state variables) and (2) partial-state EDP (trained on lags of the pest data). Excess cost was calculated by subtracting the total cost accumulated by following the optimal policy (i.e. the best case benchmark, which uses the true mechanistic model to generate policies) from the total cost accumulated by following the method of interest. For the biological, chemical, and behavioral control strategies, the total cost was the sum over a time series of step-wise costs, which includes two competing sources of cost: the cost of pest pressure (O_1) and the cost of control (O_2). We used a weighted sum scalarization method to transform the multi-objective optimization problem into a set of single-objective optimization problems. Specifically, we varied the "importance" of each objective by sweeping through different values of a parameter $0 \leq \theta \leq 1$, which places weight on each objective. The scalar objective was to minimize the total cost given by $\theta O_1 + (1-\theta)O_2$. In the real world, we often must make multiple decisions simultaneously (e.g. how to apply multiple types of control), so we also explored a more complicated IPM scenario. The IPM scenario had three competing objectives (i.e. minimize the cost of pest pressure (O_1), cost of biological control (O_2), and cost of chemical control (O_3)), so we took a similar scalarization approach and minimized $\theta_1 O_1 + \theta_2 O_2 + \theta_3 O_3$. See Material and Methods for the functional forms (Material and Methods Eqs. 4.2 & 4.3) and more details.

We explored trade-offs between the competing objectives by sweeping through values of θ, and evaluating the individual costs (O_i) at each θ. This results in a curve in the objective space called a Pareto front, which shows the set of solutions in which it is impossible to decrease one cost without sacrificing the another (Williams and Kendall, 2017).

4.4.1 Policy generation based on data alone outperforms the status quo

Under all the scenarios tested, EDP outperformed the reactive approach and had relatively low excess costs compared to the best case scenario benchmark, even with moderate levels of noise (Table C.1) in the data (Fig. 4.2 & 4.3). On average, for the time series lengths we tested, EDP produced a policy closer to the optimal policy when data for the full state was available rather than only pest data. In general, EDP produced nearly optimal results at extreme values of θ (i.e. θ close to 0 or 1), and its excess costs increased at intermediate θ values (Fig.

4.2 a,c,e). Pareto fronts for each method (Fig. 4.2b,d,f), showed that the policies strike a balance between the competing objectives at intermediate θ values as demonstrated by the black outlined points in Fig. 4.2 b,d,f.

The results were generally similar for the more complex IPM scenario (Fig. 4.3), and EDP outperformed the status quo policy. At the extremes (i.e. when one $\theta_i = 0$), EDP performs nearly optimally, and excess cost increases at intermediate values. The reactive approach performs optimally at one extreme (i.e. when no weight is placed on the cost of pest pressure). An example of the optimal control rules and resulting dynamics of the each method (Fig. 4.4) shows that EDP and the optimal policy effectively keep the pest population within acceptable bounds, while the reactive approach does not (Fig. 4.4 a,c,e). All three methods, however, use a combination of biological control and chemical control (Fig. 4.4 b,d,f).

Overall, the level of noise in the dynamics influenced the performance of EDP, but the reactive approach was relatively consistent for the full range of noise levels we tested (Fig. 4.5). Across all scenarios, both cases of EDP reduced excess costs compared to the status quo reactive approach by a substantial amount (Table C.3). Even in the worst cases, EDP reduced excess costs relative to the reactive policy by at least 33% in the partial-state case, and it reduced costs by at least 48% in the full-state case. Both cases appeared to work best under low noise and behavioral control.

4.4.2 Policy generation based on empirical data outperforms historical policies

In our empirical case studies for *Lygus* and *Culex pipiens*, it was not possible to develop best case scenario benchmarks because we do not have perfect models for the empirical dynamics. However, we had more realistic status quo benchmarks

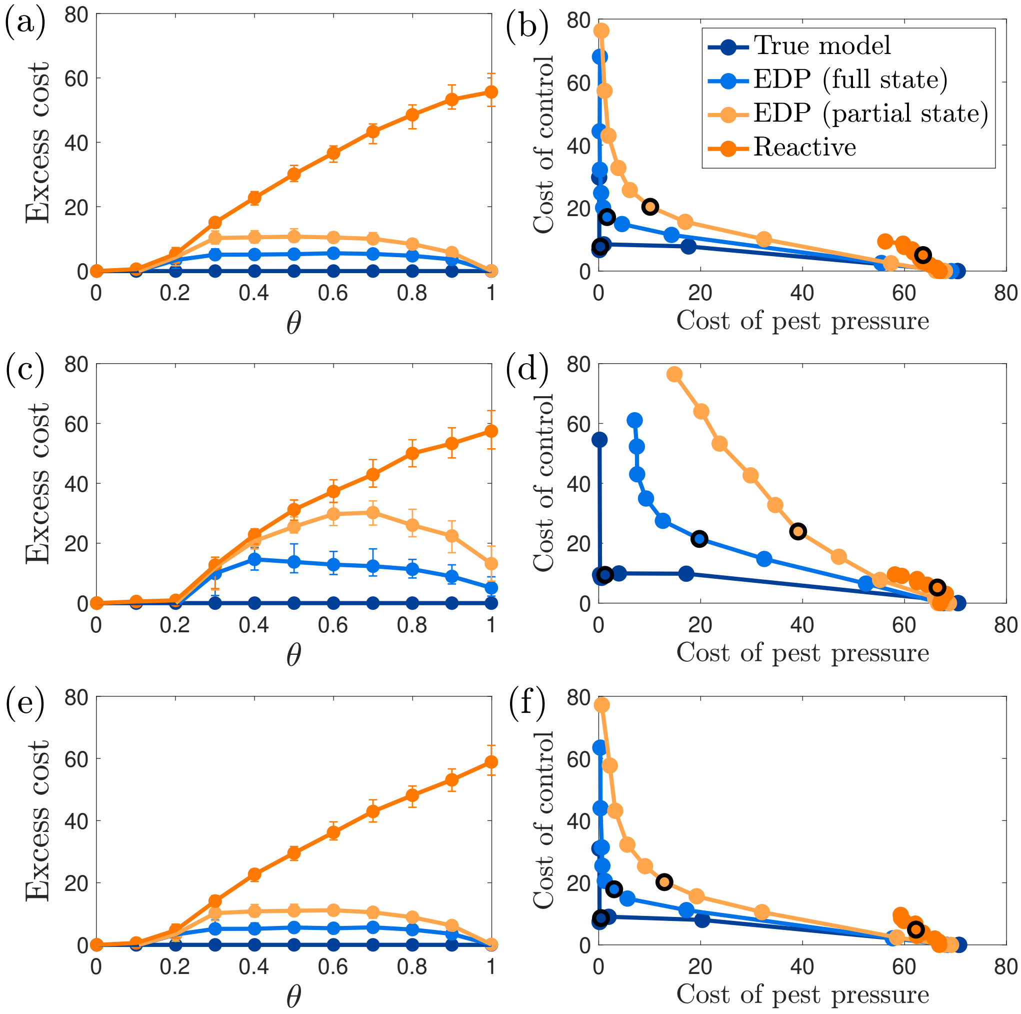

Figure 4.2: Simulation results averaged over 100 simulations of biological control (a,b), chemical control (c,d), and behavioral control (e,f) with a moderate level of noise in the data (Table C.1). Left panels show the median (dots) and lower and upper quartiles (error bars) for the excess cost of each method compared to the optimal policy (a,c,e). Pareto fronts show the trade-off between the cost of pest pressure and the cost of control (b,d,f). Each point in the Pareto fronts corresponds to a value of θ, ordered such that $\theta = 0.001$ in the lower right $\theta = 0.999$ in the upper left. Points outlined in black correspond to $\theta = 0.5$.

than the simulation analyses. In both empirical studies, we used the historical control actions from the data as the status quo benchmarks, and compared them

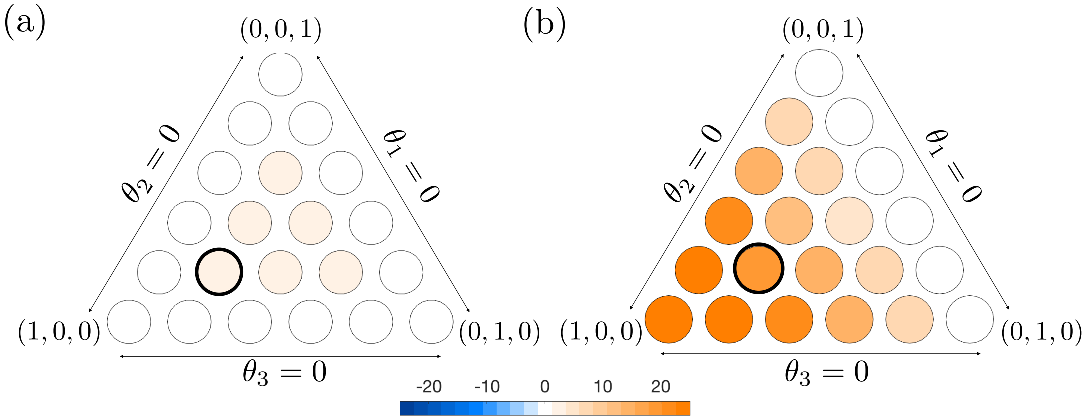

Figure 4.3: Simulation results averaged over 100 simulations of IPM control using EDP (a) and a reactive control (b) with a moderate level of noise in the data. The color in each circle is the average excess cost for a specific $(\theta_1, \theta_2, \theta_3)$ triplet. Light circles indicate low excess cost and darker circles indicates high excess cost. The circle outlined in black corresponds to $(\theta_1, \theta_2, \theta_3) = (0.6, 0.2, 0.2)$, and an example of the output dynamics for this triplet is shown in Figure 4.4.

to the EDP policy. In both cases, the EDP policy outperformed the historical policy.

In the case of *Lygus*, the out-of-sample forecast accuracy was $R^2 = 0.54$, and the historical policy fell above the EDP Pareto front (Fig. 4.6a), indicating that it may have been possible to achieve nearly 45% less pest pressure with the same number of pesticide sprays that were used historically. Alternatively, it may have been possible to achieve the same amount of pest pressure with ~ 60% fewer sprays. Similarly, the out of sample forecast accuracy was $R^2 = 0.58$ for *Culex pipiens*, and the historical control actions also fell above the EDP Pareto front (Fig. 4.6b). This suggests that it we could have achieved ~ 8% reduction in mosquito pressure with the same number of pesticide applications or the same pest pressure with ~ 60% reduction in pesticides. While we cannot truly evaluate the EDP policies on the historical data, these results serve as an initial step to demonstrating that improvement might be possible. Importantly, we can achieve

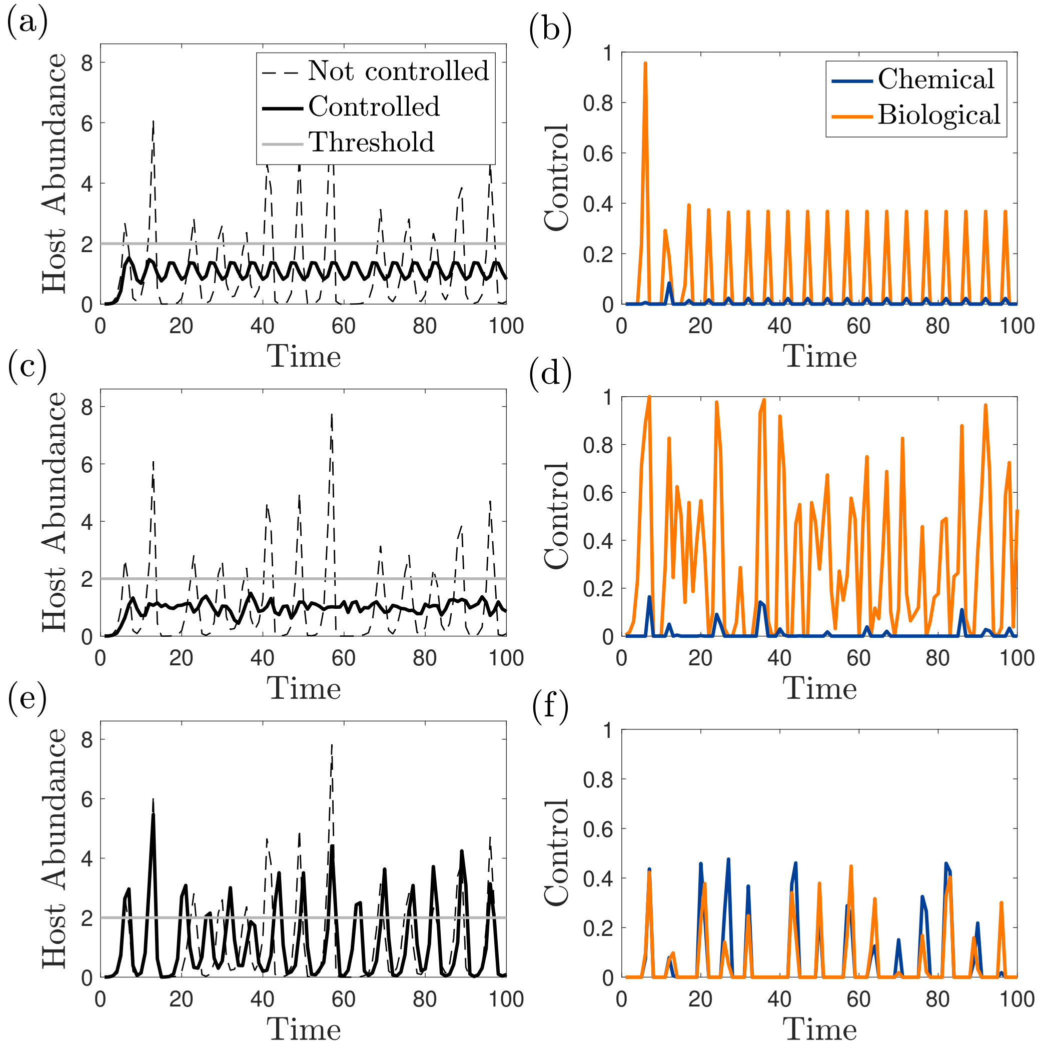

Figure 4.4: Results of a single simulation of IPM control based on the true model (a,b), EDP (c,d), and the reactive method (e,f). Comparisons between uncontrolled dynamics of the pest (dotted black lines) and controlled dynamics of the pest (solid black lines) (a,c,e). The amount of chemical (blue) and biological (orange) applied (b,d,f) to achieve the controlled dynamics.

a predictive model with the current data that is accurate enough to generate improved policies over existing ones.

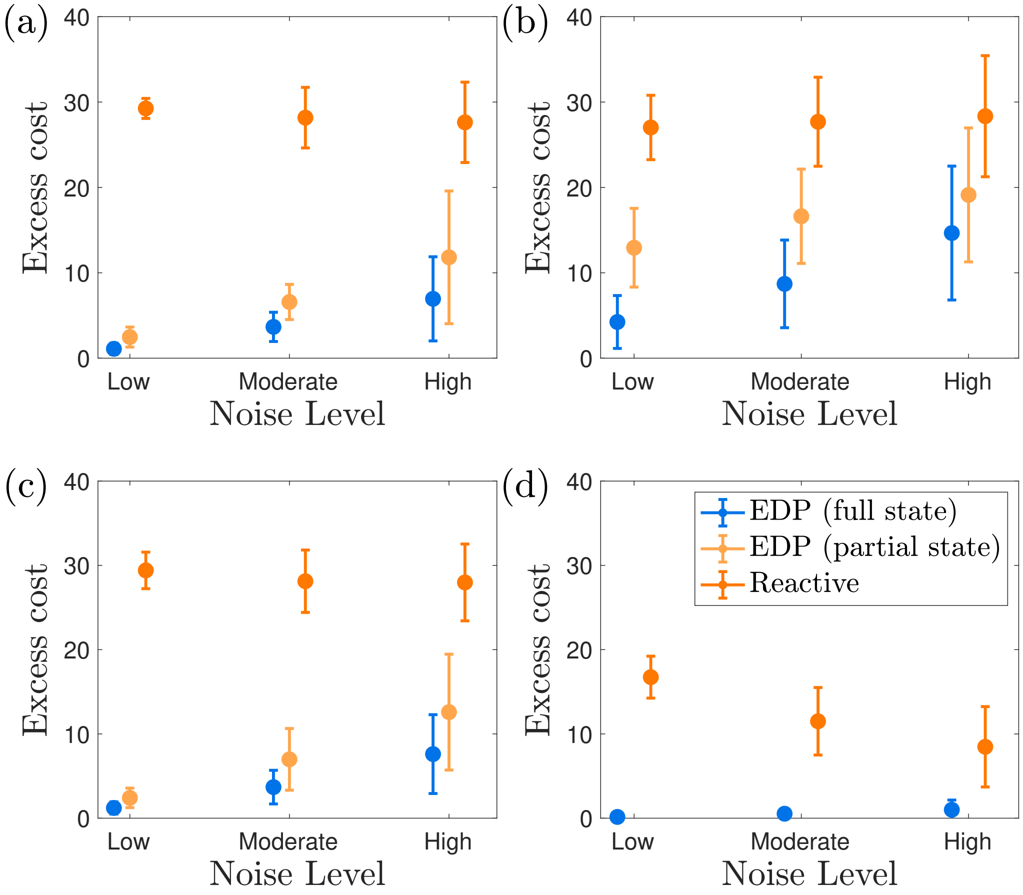

Figure 4.5: Mean (dots) and average standard deviation (error bars) of excess cost over 100 simulations of all θ for biological control (a), chemical control (b), behavioral control (c), and IPM (d).

4.5 Discussion

Our analysis suggests that this data-driven approach can mitigate the impacts of pests more effectively than the standard methods used in practice. We found that with the availability of either complete or incomplete data, EDP can produce nearly optimal management policies and outperform status quo reactive and historical approaches to pest management. Even in the worst cases, the method achieved a reduction in excess cost that could amount to millions of dollars saved

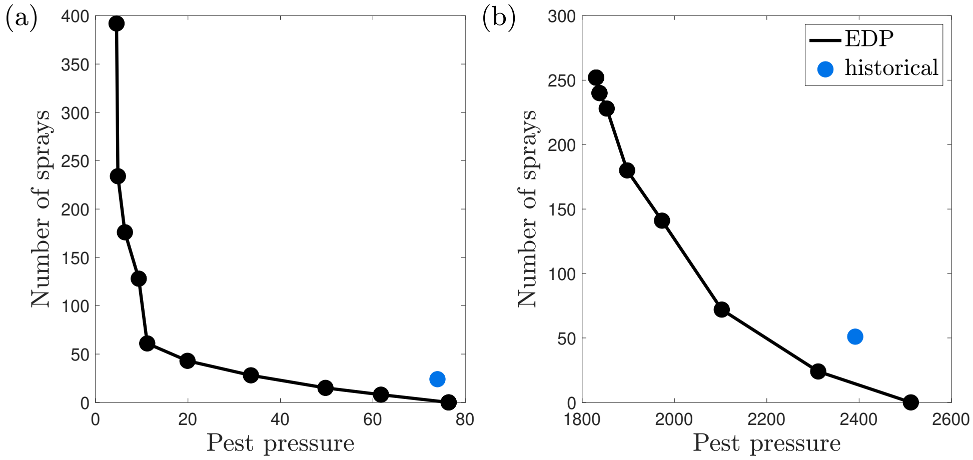

Figure 4.6: Pareto fronts of EDP and the historical policy for *Lygus* (a) and *C. pipiens* (b). Each point in the Pareto front corresponds to a value of θ, ordered such that θ is low in the lower right θ is high in the upper left.

(Bradshaw et al., 2016). We also demonstrated that EDP is fairly flexible. It can cope with a variety of control types including an IPM strategy and with data complications such as incomplete observations and high noise.

Although this method shows significant promise for improving pest control, there are a few important limitations to address. First, dynamic programming policies are sensitive to model predictions. This is evident in our results, which showed that increasing noise in the dynamics decreases forecast accuracy (Table C.2), thereby reducing the performance of EDP (Fig. 4.5). Though EDP still typically outperform a reactive policy, it is important to evaluate and carefully consider forecast accuracy before generating management advice with this method.

Second, the historical control strategy and the quality of its data can impact the performance of the EDP method. For example, if a system has been historically managed such that control has only been applied in a limited region of the state space, the GP might struggle to accurately estimate how a pest population will behave in other regions of the state space. In addition, if the data is imprecise and does not include accurate information about how much control was used historically, the method could produce suboptimal policies. Managers who are interested in using this method should be aware of which historical control strategies offer the highest probability of success (Supporting Information C.3, Table C.4) and should carefully track their intervention actions.

Finally, in cases with higher dimensions (e.g. number of species or number of EDM lags is greater than 3), EDP – as implemented here – are not computationally feasible. In these cases, similarly flexible, but more scalable methods such as approximate dynamic programming (Powell, 2011; Sutton, 1988), reinforcement learning (Sutton, 2018), or model predictive control (Morari and H. Lee, 1999) could be used. Future studies should evaluate the trade-offs of these methods in a pest management context.

There are also other important next steps. First, insect development depends on environmental conditions, especially temperature. Future iterations of this work should explicitly account for temperature or growing degree days (Naves and De Sousa, 2009) in the models by using multivariate embedding (Deyle and Sugihara, 2011) or multiview embedding (Ye and Sugihara, 2016), as this could improve policies. Similarly, in many agricultural systems, the crop is only particularly vulnerable to pests during a window of time in the season. In our empirical examples, we dealt with seasonality by generating separate policies for distinct stages of the growing season (*Lygus*) or by incorporating a seasonal term into the EDM model (*Culex pipiens*). Future work should consider more explicit and scalable ways to incorporate crop phenology and other forms of seasonal variability.

Future work should also evaluate other multi-objective cost functions. Our formulation of the cost function included only pest pressure and the cost of control. However, minimizing pesticide runoff risk (van der Werf, 1996), the risk of the insecticide resistance (Namias et al., 2021), and harmful impacts on non-targeted organisms, such as pollinators and birds, are important objectives for real-world pest management. In cases where we have data for both pests and non-target species, we can extend the multi-objective framework to identify policies that simultaneously suppress pests and protect non-target species. This approach has been used in fisheries contexts (Brias and Munch, 2021) to balance harvesting and conservation goals in multi-species settings.

Although the applications to *Lygus* and *Culex* seem promising, it is not possible for us to validate control policies *post hoc*. The ultimate test of this method should involve collaboration among empiricists, growers, or vector control agencies, and should set up controlled experiments to evaluate the efficacy of this method in a real application. Moreover, the framework we presented is quite general and has potential applications beyond insect pest management such as forecasting and mitigating the impacts of harmful algal blooms, invasive plant species, and wildfires.

Overall, our results reveal valuable insights about monitoring and managing insect pests. Under the time series lengths we studied, we found that having complete observations for the system (i.e. data for the pest and its natural enemy) had a strong advantage over partial observations (i.e. data only for the pest) (Fig. 4.5). In addition, having complete and precise data for historical control actions is important (Table C.4). Although ecological sampling is often time consuming and expensive, our study provides a clear incentive to invest in more comprehensive monitoring programs. Recent developments of automated insect sampling methods (Pegoraro et al., 2020; Rydhmer et al., 2022) and user-friendly tools to help managers track and store data (Chambers et al., 2015; Lagos-Ortiz et al., 2018) make this quite feasible and would likely yield profitable results in the long term.

Although there is more work to be done and real-world validation is still needed, the method we have proposed takes an important step toward being practically useful for management; EDP does not make idealized assumptions about noise-free dynamics or perfect models and they allow us to work directly with limited noisy data to produce useful management advice.

4.6 Acknowledgements for Chapter 4

The authors sincerely thank J. Rosenheim for providing access to the *Lygus* data and guidance on the *Lygus* biology and management practices in central California cotton. C.M. Barker and E. Lonstrup with CalSurv for providing access to mosquito trap data. K.M. Holcomb and C.M. Barker, and G. Goodman from Sacramento-Yolo Mosquito & Vector Control District for providing access to mosquito intervention data.

Chapter 5

Conclusion

This work was motivated by the need to make practical contributions to sustainable management of ecological systems. In this dissertation, we addressed this by developing generalizable methods for prediction and control in the context of ecological management for a range of terrestrial and aquatic systems. Since the traditional route of using single-species models has led to repeated failures, and since building more detailed mathematical models is challenging with our limited understanding of true ecosystem complexity, we took a more flexible data-driven approach. Empirical dynamic modeling is a powerful data-driven tool that has proven potential for making practical contributions to management. However the standard EDM framework has several limitations that greatly hinder its performance in ecological applications. Specifically, EDM cannot produce accurate predictions when time series are short or when data are sporadically sampled. In this dissertation, we explored extensions of EDM to address those limitations in an effort to improve predictions under likely ecological conditions. We additionally made EDM useful for making management decisions, by exploring a connection of EDM and an optimal control method, stochastic dynamic programming, in a pest management setting.

5.1 Summary of work

In Chapter 2, we addressed how to cope with short time series when spatial replicates are available. Prior to this project, there were several compelling but diverging methods to address this that had been proposed in the literature. In ecology, studies proposed concatenating time series from multiple locations to effectively lengthen the time series (Hsieh et al., 2008). In physics, spatial neighbors were typically directly used in embedding vectors as predictors for the dynamics at a focal site (e.g. Bialonski et al. (2015)). There was little cross-pollination between these realms of thinking, so we explored a comparison of the methods to determine what features of the data influence each approach. We also developed an approach to incorporate physical information about the dispersal of species as a Bayesian prior for the EDM framework. We tested the methods on simulated data generated with three population dynamics models with varying levels of complexity, time series length, spatial symmetry and heterogeneity. We also applied the methods to empirical fisheries data from the Northeast Fisheries Science Center.

Overall, we found that aggregating information across multiple sites – either through concatenation or the inclusion of spatial inputs – can substantially improve the utility of empirical approaches to ecological forecasting. Generally, concatenation of time series yielded the strongest improvements under common ecological conditions. However, asymmetric coupling and high spatial heterogeneity were shown to influence the utility of the methods, so there was not a single approach that universally outperformed the other. Although this analysis provided insights for how to adapt and improve the EDM framework under some common ecological conditions, there were still many datasets that were not amenable to EDM.

Therefore, in Chapter 3, we developed another extension of EDM which allows us to cope with missing or irregularly sampled data. Typical applications of EDM assume that samples are evenly spaced over time, which presents problems in ecology when data are sampled irregularly. We considered a variable step-size extension of EDM, which incorporates the temporal spacing between samples into EDM delay-coordinate vectors. We evaluated the forecast accuracy of the variable step-size method, and compared it with two other methods: (1) exclusion of delay-coordinate vectors with missing data and (2) linear interpolation along with ordinary EDM. We tested these methods using simulated data from three chaotic ecological models with various amounts and patterns of missing data, and also evaluated them using two empirical datasets: laboratory rotifer dynamics and aphid dynamics from the Rothamsted Insect Survey. Our analysis showed that while exclusion and linear interpolation produce accurate forecasts in some scenarios, the variable step-size method gives accurate forecasts in the widest range of scenarios. This analysis effectively expanded the number of datasets to which EDM can be applied.

Several additional analyses in this study revealed other interesting insights. Specifically, an adaptation of the variable step-size EDM method can be used to estimate Lyapunov exponents from irregularly sampled time series. We tested this on time series data generated by the logistic map as a proof of concept, which revealed promising results. This could help ecologists test for chaos in more ecological time series, which may have important management implications (Rogers et al., 2022). Perhaps more interestingly, we also discussed in Chapter 3 how it is possible to approximate continuous dynamics from discrete-time data using a method inspired by the variable step-size version of EDM (Appendix B.7). This could be useful for connecting theoretical models expressed in ordinary differential equations to field observations.

While it is possible that improving our ability to forecast complex ecological dynamics alone would make strong contributions to management, the key to success is determining how to optimally use the information from forecasts to make decisions. There is currently a gap in turning our understanding of ecosystems and predictions of dynamics into robust management decisions in many ecological systems. One significant example of this is in the field of pest management, where decisions are often made based on intuition, and limited observations are only used to loosely guide actions, e.g. managers spray pesticides when they observe a large number of pests. There is rarely a formal approach used to maximize long-term yield and minimize costs.

Thus, in Chapter 4, we explored the framework of empirical dynamic programming – EDM paired with stochastic dynamic programming – to make informed decisions in the context of insect pest management. Using a simulated host-parasitoid model with multiple control strategies, including biological, chemical, behavioral control, we showed that this framework effectively prevents outbreaks even with highly noisy data. It reduced costs compared to a reactive method by at least 33% in the simulated data, and it outperformed historical control strategies in empirical data from two very different systems. In the agricultural system with *Lygus*, the results from empirical dynamic programming framework suggested that it could have been possible to achieve a decrease in pest pressure with fewer pesticide applications in the past. Similar results were found for the mosquito system. This chapter demonstrated that by connecting our flexible data-driven forecasts with formal optimization strategies, we can take reasonable steps toward improve decision making in ecology.

5.2 Areas for future work

Our analysis in Chapter 2 revealed interesting insights about using this framework for inference about system characteristics. For instance, it may be possible to detect heterogeneity in the dynamics maps across sites by evaluating whether or not predictability improves by concatenating multiple sites. This can be useful for evaluating conditions in areas where we understand little about the environment, and can also help us evaluate how conditions change over time.

In Chapter 3, we highlighted one particularly interesting area for future work. We established that the ideas from variable step-size EDM can be extended to approximate continuous dynamics (i.e. $\frac{dx}{dt}$) from discretely sampled time series $(x_1, x_2, ...x_T)$. While this is an interesting result on its own, this also offers exciting possibilities in multi-species settings. Forecasting is powerful for making robust decisions in ecology, but understanding the strength and direction of species interactions can also be very useful for decision makers. There have been numerous methods proposed for estimating causal species interactions from data (e.g. Deyle et al. (2016); Sugihara et al. (2012)), but many of them measure net effects of one species on another and cannot determine if the interactions are direct or indirect. Our approach for estimating continuous dynamics from discrete data could overcome this issue. Specifically, consider a three-species system

$$\begin{aligned}
\frac{dx}{dt} &= f(x, y, z)\\
\frac{dy}{dt} &= g(x, y)\\
\frac{dz}{dt} &= h(x, z).
\end{aligned}$$

In this example, variables y and z are coupled indirectly through x. It may be possible to identify these indirect interactions by approximating $\frac{dy}{dt}$ using the approach highlighted in Section B.7, and evaluating it at multiple values of z. For indirect interactions, $\frac{dy}{dt}$ approximations should stay consistent for all z. Future studies should explore this approach and identify conditions in the data that could cause it to fail.

Finally, although we have demonstrated some level of generalizability by applying our methods to data from a range of terrestrial and marine ecosystems, there is still more work to be done to evaluate the extent of generalizability. We only tested the complete framework of prediction and control with empirical dynamic programming in one ecological scenario of insect pest management. Future studies should explore our ability to optimize decisions in other high-impact ecological systems. For instance, harmful algal blooms are detrimental to marine life and cost approximately $82 million per year (Hoagland and Scatasta, 2006). It could be interesting to study whether a similar empirical dynamic programming approach could mitigate the impacts of algal blooms. Other objectives for future studies could include predicting and preventing wildfires or maximizing long-term biodiversity in marine and forest ecosystems.

Appendix A

to Leveraging Spatial Information with Empirical Dynamic Modeling

A.1 Demonstrating lack of identifiability

We demonstrated the lack of identifiability in sEDM by simulating dynamics according to the following two-species competition model

$$x^i_{t+1} = (1-\mu)\left(x^i_t e^{r_1(1-x^i_t-ay^i_t)}\right) + \frac{\mu}{2}\left(x^{i-1}_t e^{r_1(1-x^{i-1}_t-ay^{i-1}_t)} + x^{i+1}_t e^{r_1(1-x^{i+1}_t-ay^{i+1}_t)}\right) \tag{A.1}$$

$$y^i_{t+1} = (1-\mu)\left(y^i_t e^{r_2(1-bx^i_t-y^i_t)}\right) + \frac{\mu}{2}\left(x^{i-1}_t e^{r_2(1-bx^{i-1}_t-y^{i-1}_t)} + x^{i+1}_t e^{r_2(1-bx^{i+1}_t-y^{i+1}_t)}\right) \tag{A.2}$$

where x_t^i represents the abundance of the first species at time t and location i y_t^i is the abundance of the second species at time t and location i. Parameter values were $r_1 = 3.2$, $r_2 = 3.7$, $\mu = 0.3$, $a = 0.1$, and $b = 0.9$. A time series of 20 points was simulated on a one-dimensional lattice with 7 patches. Using only species x, we fit sEDM with a naive prior and physically-informed prior and compared the length scale parameters after the GP was fit. The physically-informed prior showed more agreement across separate simulations.

A.2 Length scale prior specification

A.2.1 First-Order Approximation

We begin with an intuitive description of our rational followed by a formal derivation of priors that can be used in a wide range of sEDM applications.

In modeling $x_t^i = f(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^n$ with a GP, the length scale parameters govern how much the function f can change as their corresponding entry of $\mathbf{x}$ changes. If f is approximately linear, i.e. $x_t^i \approx \sum\limits_{k=1}^{n} a_k x_k = \mathbf{a}^\top \mathbf{x}$, it is clear that $\frac{\Delta f}{\Delta x_j} \approx a_j$ for any $x_j \in \mathbf{x}$ and corresponding $a_j \in \mathbf{a}$. Consequently, we expect length scale parameters in sEDM to be related to $\mathbf{a}$ from a linear approximation of spatiotemporal population dynamics.

Formally, consider an application in which population dynamics evolve at discrete points in time in a continuous spatial domain. The population x at time $t+1$ and spatial location u can be modeled with the first-order spatiotemporal model,

$$x_t(u) = \int_{-\infty}^{\infty} M(u-y) x_{t-1}(y) dy \tag{A.3}$$

where M is the dispersal kernel describing the probability of moving from location y to location u under an assumption of isotropy. Importantly, the form of the length scale prior we derive is specific to the dispersal kernel used to describe the network topology of the system. There is a vast literature dedicated to estimating dispersal kernels for various taxa moving in their respective environments (Kot and Schaffer, 1986; Nathan et al., 2012; Neubert et al., 1995). One of the most common dispersal kernels is the Gaussian kernel, $M(\delta) = \frac{1}{\sqrt{2\pi v}} e^{-\frac{\delta^2}{2v}}$, so we will use this for the remainder of the derivation.

Taking the Fourier transform of (A.3) in space gives

$$\begin{aligned}\int_{-\infty}^{\infty} e^{i\omega u} x_t(u) du &= \int_{-\infty}^{\infty} e^{i\omega u} \left(\int_{-\infty}^{\infty} M(u-y) x_{t-1}(y) dy \right) du \\ \tilde{x}_t(\omega) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{i\omega u} M(u-y) x_{t-1}(y) dy du\end{aligned}$$

If $\delta = u - y$, then

$$\begin{aligned}\tilde{x}_t(\omega) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{i\omega(\delta+y)} M(\delta) x_{t-1}(y) dy d\delta \\ &= \int_{-\infty}^{\infty} e^{i\omega\delta} M(\delta) d\delta \int_{-\infty}^{\infty} e^{i\omega y} x_{t-1}(y) dy \\ &= \tilde{M}(\omega) \tilde{x}_{t-1}(\omega) \\ &= \left(\tilde{M}(\omega) \right)^{\tau} \tilde{x}_{t-\tau}(\omega) \qquad \text{(A.4)}\end{aligned}$$

Under the assumption of Gaussian dispersal,

$$\left(\tilde{M}(\omega) \right)^{\tau} = \left(e^{-\frac{\omega^2 v}{2}} \right)^{\tau} = e^{-\frac{\omega^2 v \tau}{2}},$$

which implies

$$\mathcal{F}^{-1} \left[\left(\tilde{M}(\omega) \right)^{\tau} \right] = \frac{1}{\sqrt{2\pi v \tau}} e^{-\frac{\delta^2}{2v\tau}}.$$

Taking the inverse Fourier transform of (A.4), we get

$$x_t(u) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi v\tau}} e^{-\frac{\delta^2}{2v\tau}} x_{t-\tau}(u-\delta)d\delta,$$

where we have used the convolution theorem for an inverse Fourier transform.

In real applications, samples are taken at a finite set of points, so there is a finite set of distances R that contribute to the focal site. In this case,

$$x_t(u) \approx \sum_{\delta \in R} a_{\delta,\tau}\ x_{t-\tau}(u-\delta) \tag{A.5}$$

where $a_{\delta,\tau} = \frac{1}{\sqrt{2\pi v\tau}} e^{-\frac{\delta^2}{2v\tau}}$. Thus, we have an approximate linear map taking $x_t(u)$ to any of its spatiotemporal lags.

Now we show how to relate this result to the length scale parameters by considering the root mean square slope of a function $E\left(\left(\frac{\Delta f}{\Delta x_j}\right)^2\right)^{\frac{1}{2}}$, which explains how the function changes with respect to some entry x_j of $\mathbf{x}$. The intuitive description above implies that this should be equal to a_j, the approximate slope in the direction of x_j. Generally,

$$\begin{aligned} E\left(\left(\frac{\Delta f}{\Delta x_j}\right)^2\right)^{\frac{1}{2}} &= \frac{1}{\Delta x_j} E\left((f(x_j+\Delta x_j) - f(x_j))^2\right)^{\frac{1}{2}} \\ &= \frac{1}{\Delta x_j}[E(f(x_j+\Delta x_j)f(x_j+\Delta x_j)) \\ &- 2E(f(x_j+\Delta x_j)f(x_j)) \\ &+ E(f(x_j)f(x_j))]^{\frac{1}{2}} \end{aligned}$$

Note that $E(f(x), f(y)) = C(f(x), f(y)) = \sigma^2 R(\phi|x-y|)$ where $R(z)$ is the autocorrelation function, σ^2 is the variance of the data, and ϕ is the length scale parameter. Here I make an assumption that the autocorrelation is Gaussian, i.e.

$R(z) = e^{-z^2}$, so $R(0) = 1$ and

$$E\left(\left(\frac{\Delta f}{\Delta x_j}\right)^2\right)^{\frac{1}{2}} = \frac{\sqrt{2\sigma^2}}{\Delta x_j}\sqrt{1 - R(\phi \Delta x_j)}.$$

A Taylor series expansion gives

$$E\left(\left(\frac{\Delta f}{\Delta x_j}\right)^2\right)^{\frac{1}{2}} = \frac{\sqrt{2\sigma^2}}{\Delta x_j}\sqrt{1 - R(0) - \phi \Delta x_j R'(0) - \frac{(\phi \Delta x_j)^2}{2} R''(0)}.$$

Evaluating the expression above for Gaussian autocorrelation gives

$$E\left(\left(\frac{\Delta f}{\Delta x_j}\right)^2\right)^{\frac{1}{2}} = \sqrt{2\sigma^2}\phi.$$

Since $\sqrt{2\sigma^2}\phi \approx a_j$, one should set

$$\phi \approx \frac{a_j}{\sqrt{2\sigma^2}}. \tag{A.6}$$

Using the results from (A.5) and (A.6) together in sEDM applications, we set the half-normal prior distribution (2.4), and choose γ such that

$$E[\phi_{\delta,\tau}] \approx \frac{a_{\delta,\tau}}{\sqrt{2\sigma^2}}. \tag{A.7}$$

This results gives a reasonable prior specification derived from a first-order approximation. See the next section for a more general prior specification without the first-order assumption.

A.2.2 General length scale prior specification

The preceding derivation started from a scalar model of population dynamics evaluated near equilibrium. For more general settings, suppose we have the model

$$\mathbf{y}_t = A\mathbf{y}_{t-1} \qquad \text{(Process model)} \tag{A.8}$$

$$\mathbf{x}_t = K\mathbf{y}_t \qquad \text{(Observation model)} \tag{A.9}$$

Let $\mathbf{y}_t$ be a vector containing all species in the system in every location at time t. That is, in a system with N species and P locations, $\mathbf{y}_t \in \mathbb{R}^n$ with $n = P \cdot N$. Let $\mathbf{x}_t$ be the observed values at time t. Since it is often not possible to observe every species in the system, we assume K to be sparse. In fact, in all sEDM applications in this paper, we assume observations of only a single species in all locations. In other words, $\mathbf{x}_t \in \mathbb{R}^P$ and $K \in \mathbb{R}^{P\times n}$. Notice that

$$\mathbf{x}_t = K\mathbf{y}_t = KA\mathbf{y}_{t-1} = ... = KA^E\mathbf{y}_{t-E} \tag{A.10}$$

In sEDM, we would like to find

$$\mathbf{x}_t = M_1\mathbf{x}_{t-1} + M_2\mathbf{x}_{t-2} + \cdots + M_E\mathbf{x}_{t-E} \tag{A.11}$$

From the setup, it follows that

$$\begin{aligned}\mathbf{x}_t &= M_1\mathbf{x}_{t-1} + M_2\mathbf{x}_{t-2} + \cdots + M_E\mathbf{x}_{t-E}\\ &= M_1K\mathbf{y}_{t-1} + M_2K\mathbf{y}_{t-2} + \cdots + M_EK\mathbf{y}_{t-E}\\ &= M_1KA^{E-1}\mathbf{y}_{t-E} + M_2KA^{E-2}\mathbf{y}_{t-E} + \cdots + M_EK\mathbf{y}_{t-E}\end{aligned} \tag{A.12}$$

In previous sections our first-order assumption implied $M_2 = M_1^2, \ldots, M_E = M_1^E$.

Now we allow more flexibility for this to not be the case. We would like to know what $M_1, M_2, \ldots, M_E$ are, so that we can use our result from (A.11) to set prior distributions length scale parameters.

Setting (A.10) equal to (A.12), we get

$$KA^E = M_1 KA^{E-1} + M_2 KA^{E-2} + \cdots + M_E K$$

Similarly,

$$KA^{E+1} = M_1 KA^E + M_2 KA^{E-1} + \cdots + M_E KA$$

Repeating this process, we get

$$\left[KA^E \middle| \ldots \middle| KA^{2E-1} \right] = \left[M_1 \middle| \ldots \middle| M_E \right] \begin{bmatrix} KA^{E-1} & KA^E & \ldots & KA^{2E-2} \\ KA^{E-2} & KA^{E-1} & \ldots & KA^{2E-3} \\ \vdots & \vdots & \ddots & \vdots \\ K & KA & \ldots & KA^{E-1} \end{bmatrix}$$

or $V = MQ$ where we repeated the process enough times so that Q has full rank and its right inverse exists. Solving for $M_1, M_2, \ldots, M_E$, we get

$$\left[M_1 \middle| \ldots \middle| M_E \right] = \left[KA^E \middle| \ldots \middle| KA^{2E-1} \right] \begin{bmatrix} KA^{E-1} & KA^E & \ldots & KA^{2E-2} \\ KA^{E-2} & KA^{E-1} & \ldots & KA^{2E-3} \\ \vdots & \vdots & \ddots & \vdots \\ K & KA & \ldots & KA^{E-1} \end{bmatrix}^{-1}$$

These results show that if we have reasonable knowledge of A and K, we can recover a linear model relating x_t^i to any of its spatiotemporal lags by (A.11). From here, we can use the result from the previous section to set prior distributions for the length scale parameters. It is important to note that in many practical applications, it is not feasible to know A exactly and estimating it can be tedious. Additionally, the movement dynamics for all species must be the same in order for this method to yield tractable analytical solutions. Consequently, it is not likely that this method of prior specification is practical in real applications. We have simply presented it as a theoretical foundation for relating true system dynamics to the length scale parameters in sEDM.

A.3 Measuring synchrony

We hypothesize that the utility of spatial methods could be related to how synchronous the dynamics are. Intuitively, asynchronous dynamics may cause dynamics to differ widely across sites, causing sEDM to provide little advantage. We tested the consistency of this hypothesis by using three metrics to measure the degree of synchrony for each of the simulated models and each of the empirical species from the NEFSC study.

The first measure of synchrony is related to the notion of a spatial return plot (Crutchfield and Kaneko, 1987; Vasconcelos et al., 2004) which is a plot of a value of each site, x_t^i versus its nearest neighbor x_t^{i+1} at a fixed time. When dynamics are highly synchronous, the spatial return plot is concentrated around the diagonal of the plot, and as they become asynchronous, the return plot is more scattered. The first measure of synchrony computes the average distance of each point in the spatial return plot to the diagonal. We call this measure the "distance to synchrony," and it is given by

$$\bar{d} = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{1}{\sqrt{2}} \left(x_t^i - x_t^{i+1} \right) \right| . \tag{A.13}$$

Lower values of $\bar{d}$ indicate higher synchrony, with $\bar{d} = 0$ meaning full synchrony.

Note that $\bar{d} \approx 0.3$ is considered high when data are scaled to have mean 0 and standard deviation 1. For our applications, we scaled the data, measured this for every point in the time series, and reported the average $\bar{d}$ over the entire series.

The second measure of synchrony is called the "order parameter" (Kuramoto, 1984; Pinto and Viana, 2000; Vasconcelos et al., 2004) and is given by

$$z_t = \frac{1}{N} \sum_{j=1}^{N} \exp\left(2\pi i x_t^j\right). \tag{A.14}$$

The magnitude of the order parameter is 1 when dynamics are completely synchronous, and decreases as they become more asynchronous. We computed the magnitude of the order parameter for each point in the time series and reported the average $\mid z_t \mid$ over the series.

Finally, we used the distance at which dynamics become uncorrelated as a measure of synchrony. For all pairs of locations, we measured the autocorrelation given by,

$$\rho(x^i, x^j) = \frac{1}{T} \sum_{t=1}^{T} \frac{(x_t^i - \overline{x^i})(x_t^j - \overline{x^j})}{\sigma_i, \sigma_j}, \tag{A.15}$$

where $\overline{x^i}$ is the mean abundance in location i and σ_i is the standard deviation of abundance in location i. The synchrony measure was given by the distance two inputs must be from each other for the average autocorrelation to pass below a threshold of 0.5. We scaled the distance by the maximum distance so that if the autocorrelation never went below 0.5, synchrony was 1. Thus, highly synchronous dynamics are close to 1 and asynchronous dynamics are close to 0 by this measure.

Table A.1 shows the degree of synchrony for each case in the main paper. As expected, all of the simulated models with periodic dynamics are highly synchronous, and we can generate chaotic dynamics that are fairly synchronous or asynchronous. For the most part, empirical data was asynchronous.

	Distance to Synchrony	**Order Parameter**	**Autocorrelation**
Model 1 (periodic)	0.10	0.59	1
Model 2 (periodic)	0.14	0.38	1
Model 3 (periodic)	0.09	0.70	1
Model 1 (chaotic-high synch)	0.18	0.42	1
Model 2 (chaotic-high synch)	0.21	0.28	0.33
Model 3 (chaotic-high synch)	0.11	0.62	1
Model 1 (chaotic-low synch)	0.30	0.23	0.05
Model 2 (chaotic-low synch)	0.29	0.21	0.09
Model 3 (chaotic-low synch)	0.25	0.48	0.12
Longfin squid	0.46	0.12	0.04
Silver hake	0.43	0.15	0.04
Butterfish	0.45	0.13	0.04

Table A.1: Degree of synchrony in the dynamics generated with each of the simulated models and each of the species from the empirical analysis (fine resolution data). Three measures of synchrony were used. On average, the empirical species had more asynchronous dynamics than the simulated models.

A.4 More information on results

Here we give numerical values for the RMSE and the fraction of variance explained (R^2) on simulated and empirical data. Table A.2 gives results corresponding to Fig. 2.4 for simulated data in different dynamical regimes, Table A.3 corresponds to Fig. 2.5 for simulated data with different time series lengths, Table A.4 corresponds to Fig. 2.7 for simulated data in different coupling schemes.

A.5 Temporal and spatial return plots

Fig. A.1 shows the spatial and temporal return plots for 3 dynamical regimes of the models in Table 2.1. Data were simulated on a lattice of $N = 75$ with periodic boundary conditions for time series length $T = 300$. One location was

#	**Method**	**Periodic**		**Chaotic**		**Chaotic**	
				(high synch)		**(low synch)**	
		RMSE	R^2	RMSE	R^2	RMSE	R^2
1	Informed sEDM (l)	0.2782	0.92	0.4875	0.74	0.6605	0.53
	Informed sEDM (c)	0.1893	0.96	**0.2116**	**0.95**	**0.2174**	**0.95**
	Naive sEDM (l)	0.2802	0.92	0.5051	0.72	0.6876	0.49
	Naive sEDM (c)	0.1959	0.96	0.2151	0.95	0.2198	0.95
	EDM (l)	0.2102	0.95	0.3432	0.87	0.4762	0.76
	EDM (c)	**0.1749**	**0.97**	0.2126	0.95	0.2240	0.95
2	Informed sEDM (l)	0.2986	0.91	0.4833	0.76	0.6423	0.57
	Informed sEDM (c)	**0.1507**	**0.98**	**0.1483**	**0.98**	**0.1830**	**0.97**
	Naive sEDM (l)	0.3090	0.90	0.5040	0.74	0.6750	0.53
	Naive sEDM (c)	0.1507	0.98	0.1484	0.98	0.1835	0.97
	EDM (l)	0.2189	0.95	0.2903	0.91	0.4272	0.81
	EDM (c)	0.1527	0.98	0.1704	0.97	0.2276	0.95
3	Informed sEDM (l)	0.3788	0.85	0.4090	0.82	0.7494	0.43
	Informed sEDM (c)	0.2022	0.96	0.1921	0.96	0.2570	0.92
	Naive sEDM (l)	0.3926	0.83	0.4168	0.81	0.7476	0.43
	Naive sEDM (c)	0.2379	0.94	0.2364	0.94	0.3549	0.85
	EDM (l)	0.3280	0.88	0.3539	0.87	0.5936	0.63
	EDM (c)	**0.1764**	**0.97**	**0.1819**	**0.96**	**0.2190**	**0.94**

Table A.2: One-step-ahead forecast RMSE and R^2 (i.e. fraction of the variance explained) for different dynamical regimes of simulated data. RMSE are scaled such that the mean predictor has a RMSE of 1.0. Bold values indicate the method with the best forecast accuracy. Methods labeled with (l) are local methods, and methods labeled with (c) are concatenated.

chosen at random and used to generate the return plots. Generally, plots with more structure (e.g. periodic dynamics or temporal maps) are easier to predict with a GP than those with less structure (e.g. chaotic, low synchrony, spatial maps).

#	**Method**	**25**		**50**		**75**	
		RMSE	R^2	RMSE	R^2	RMSE	R^2
1	Informed sEDM (l)	0.4875	0.74	0.3303	0.88	0.2696	0.93
	Informed sEDM (c)	**0.2116**	**0.95**	**0.1895**	**0.96**	**0.1835**	**0.97**
	Naive sEDM (l)	0.5051	0.72	0.3450	0.87	0.2846	0.92
	Naive sEDM (c)	0.2151	0.95	0.1895	0.96	0.1837	0.97
	EDM (l)	0.3432	0.88	0.2668	0.93	0.2270	0.94
	EDM (c)	0.2126	0.95	0.1990	0.96	0.1925	0.96
2	Informed sEDM (l)	0.4833	0.76	0.2413	0.94	0.1919	0.96
	Informed sEDM (c)	**0.1483**	**0.98**	**0.1256**	**0.98**	**0.1176**	**0.99**
	Naive sEDM (l)	0.5040	0.74	0.2452	0.94	0.1922	0.96
	Naive sEDM (c)	0.1484	0.98	0.1258	0.98	0.1176	0.99
	EDM (l)	0.2903	0.91	0.2132	0.95	0.1886	0.96
	EDM (c)	0.1704	0.97	0.1583	0.97	0.1518	0.98
3	Informed sEDM (l)	0.4090	0.82	0.3219	0.89	0.2741	0.92
	Informed sEDM (c)	0.1921	0.96	**0.1503**	**0.98**	**0.1386**	**0.98**
	Naive sEDM (l)	0.4168	0.81	0.3526	0.86	0.3050	0.91
	Naive sEDM (c)	0.2364	0.94	0.1947	0.96	0.1756	0.97
	EDM (l)	0.3539	0.87	0.2708	0.92	0.2313	0.95
	EDM (c)	**0.1819**	**0.96**	0.1507	0.98	0.1411	0.98

Table A.3: One-step-ahead forecast RMSE and R^2 (i.e. fraction of the variance explained) for different time series lengths of simulated training data. RMSE are scaled such that the mean predictor has a RMSE of 1.0. Bold values indicate the method with the best forecast accuracy. Methods labeled with (l) are local methods, and methods labeled with (c) are concatenated.

A.6 Concatenating sites versus long time series

Since the method of concatenating libraries works drastically better than local methods on some simulations, it is reasonable to ask whether this is a result of the larger amount of data used to fit the GP with concatenated methods. Fig. A.2 shows the RMSE of forecasts on Model 1 as we increase the time series length of local methods until they have the equivalent number of data points. In this example, concatenated methods combined data from 5 randomly selected locations in the $N = 50$ lattice. Hence in a standard comparison of these methods,

#	**Method**	**Symmetric**		**Asymmetric**	
		RMSE	R^2	RMSE	R^2
1	Informed sEDM (local)	0.4875	0.74	0.5017	0.84
	Informed sEDM (concatenated)	**0.2116**	**0.95**	**0.2633**	**0.93**
	Naive sEDM (local)	0.5051	0.72	0.5138	0.83
	Naive sEDM (concatenated)	0.2151	0.95	0.2640	0.93
	EDM (local)	0.3432	0.88	0.5091	0.74
	EDM (concatenated)	0.2126	0.95	0.3243	0.89
2	Informed sEDM (local)	0.4833	0.76	0.3868	0.85
	Informed sEDM (concatenated)	**0.1483**	**0.98**	**0.1853**	**0.96**
	Naive sEDM (local)	0.5040	0.74	0.3969	0.84
	Naive sEDM (concatenated)	0.1484	0.98	0.1856	0.96
	EDM (local)	0.2903	0.91	0.4398	0.80
	EDM (concatenated)	0.1704	0.97	0.2712	0.93
3	Informed sEDM (local)	0.4090	0.82	0.4443	0.79
	Informed sEDM (concatenated)	0.1921	0.96	**0.2103**	**0.95**
	Naive sEDM (local)	0.4168	0.81	0.4495	0.79
	Naive sEDM (concatenated)	0.2364	0.94	0.2588	0.93
	EDM (local)	0.3539	0.87	0.5456	0.69
	EDM (concatenated)	**0.1819**	**0.96**	0.2245	0.95

Table A.4: One-step-ahead forecast RMSE and R^2 (i.e. fraction of the variance explained) for different dispersal coupling schemes on simulated data. RMSE are scaled such that the mean predictor has a RMSE of 1.0. Bold values indicate the method with the best forecast accuracy.

concatenated methods would be using 5× more data to fit the GP. In these simulations, we fix the length of the training time series for concatenated methods at $T = 30$ and increase the length of the time series for local methods until the two methods are trained on the same amount of data. As we intuitively expect, when we increase the time series length by 5× the predictive accuracy of local are approximately the same as the concatenated methods.

	Longfin squid		**Silver hake**		**Butterfish**	
	RMSE	R^2	RMSE	R^2	RMSE	R^2
Informed sEDM (l)	0.6327	0.60	0.5412	0.71	0.7287	0.47
Informed sEDM (c)	0.5440	0.70	**0.4552**	**0.79**	**0.6197**	**0.62**
Naive sEDM (l)	0.6407	0.59	0.5740	0.67	0.7321	0.46
Naive sEDM (c)	0.5441	0.70	0.4893	0.76	0.6289	0.60
EDM (l)	0.6208	0.61	0.6177	0.62	0.8044	0.35
EDM (c)	**0.5332**	**0.71**	0.6036	0.64	0.6272	0.61

Table A.5: Spatial data separated by major region ($N = 4$). Sequential forecast RMSE and R^2 (i.e. fraction of the variance explained) NEFSC bottom trawl survey data. RMSE are scaled such that the mean predictor has a RMSE of 1.0. Bold values indicate the method with the best forecast accuracy. Methods labeled with (l) are local methods, and methods labeled with (c) are concatenated.

	Longfin squid		**Silver hake**		**Butterfish**	
	RMSE	R^2	RMSE	R^2	RMSE	R^2
Informed sEDM (l)	0.6832	0.53	0.8805	0.22	0.9724	0.05
Informed sEDM (c)	0.5491	0.70	0.6747	0.54	**0.7763**	**0.40**
Naive sEDM (l)	0.6868	0.53	0.8600	0.26	0.9766	0.05
Naive sEDM (c)	0.5527	0.69	0.6748	0.54	0.7763	0.40
EDM (l)	0.6477	0.58	0.7481	0.4404	0.9300	0.14
EDM (c)	**0.5482**	**0.70**	**0.6225**	**0.61**	0.7780	0.39

Table A.6: Spatial data separated by survey strata ($N = 47$). Sequential forecast RMSE and R^2 (i.e. fraction of the variance explained) NEFSC bottom trawl survey data. RMSE are scaled such that the mean predictor has a RMSE of 1.0. Bold values indicate the method with the best forecast accuracy. Methods labeled with (l) are local methods, and methods labeled with (c) are concatenated.

A.7 Simultaneous effect of asymmetry and heterogeneity

Asymmetry and heterogeneity were individually shown to affect the utility of sEDM and concatenated methods. Here we show the effect when these are simultaneously present. Fig. A.3 shows that with both of these effects present,

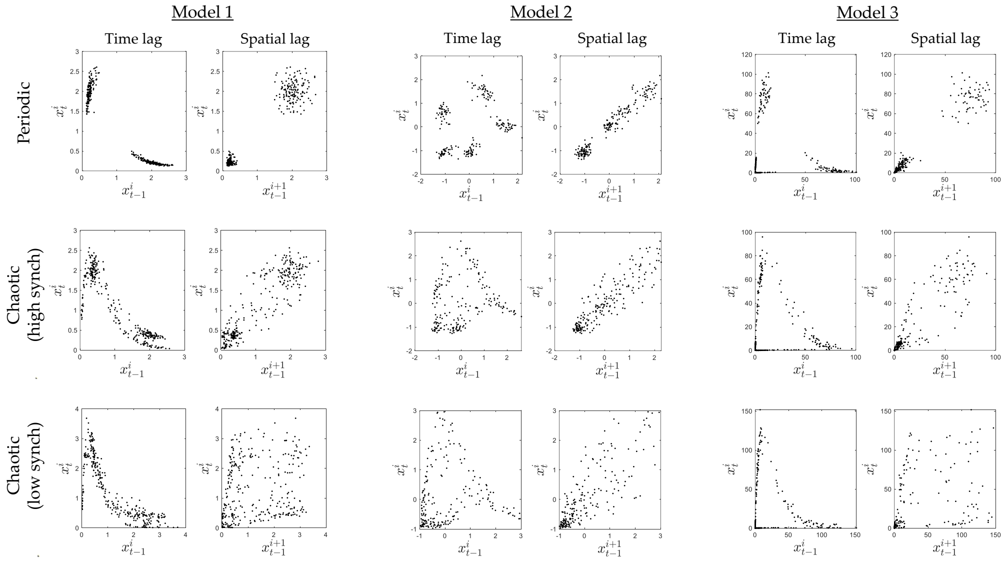

Figure A.1: Temporal and spatial return plots for all models in Table 2.1 for periodic dynamics, chaotic dynamics with high spatial synchrony, and chaotic dynamics with low spatial synchrony. Temporal return plots show x_t^i vs. its time lag of x_{t-1}^i. Spatial return plots show x_t^i as a function of its spatial lag x_t^{i+1}.

local sEDM methods outperform all other methods, and concatenated methods suffer as the degree of heterogeneity increases.

A.8 Forecast Horizon

In the main document, all simulations focused on one-step-ahead forecasts, but it is natural to consider how these methods perform as we increase the number of steps we predict into the future. Fig. A.4 shows the RMSE (averaged over 20 independent simulations) of all six methods on each model from the main paper as we increase the step-ahead forecasts. As expected, forecast accuracy decreases as we predict further into the future, but the ranking of methods does not appear to change.

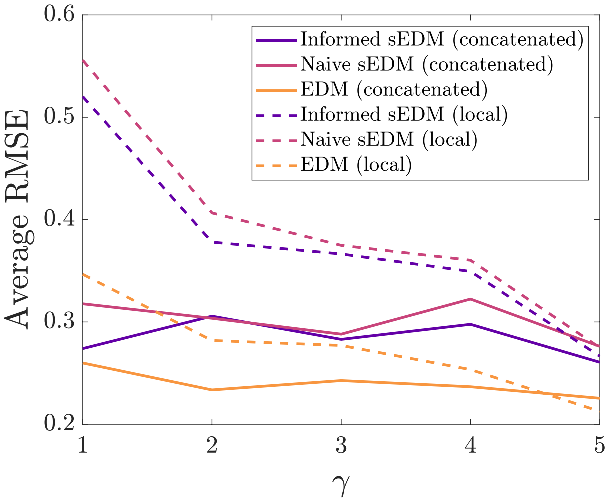

Figure A.2: Average RMSE on testing data over 100 simulations of Model 1 from Table 2.1. Results are shown for local (dotted lines) and concatenated (solid lines) informed sEDM (purple), naive sEDM (pink), and EDM (yellow). The x-axis indicates ratio of number of data points in the local methods compared to the number of data points in the concatenated methods. i.e. $\gamma = \frac{n_{local}}{n_{concatenated}}$, where n is the number of data points.

A.9 Timescale and Predictability

It is important to consider the effect of timescale on these methods. Particularly, if the periodicity of dynamics is longer than the time series or the number of time lags in embedding vectors, no form of EDM is likely to predict dynamics well. Since these methods make predictions based on the history of data, the dynamics must go through full cycles within the training period. If they do not, predictions are difficult to make with EDM. As an example, we generated dynamics for a four-species competition model given in Table A.7 on an $N = 50$ lattice with periodic boundary conditions. An example of resulting dynamics for species x are shown in Fig. A.5. Clearly, dynamics are highly irregular with a large periodicity of over 30 time points.

We used the 6 EDM and sEDM methods to these data, training on $T_{train} = 25$

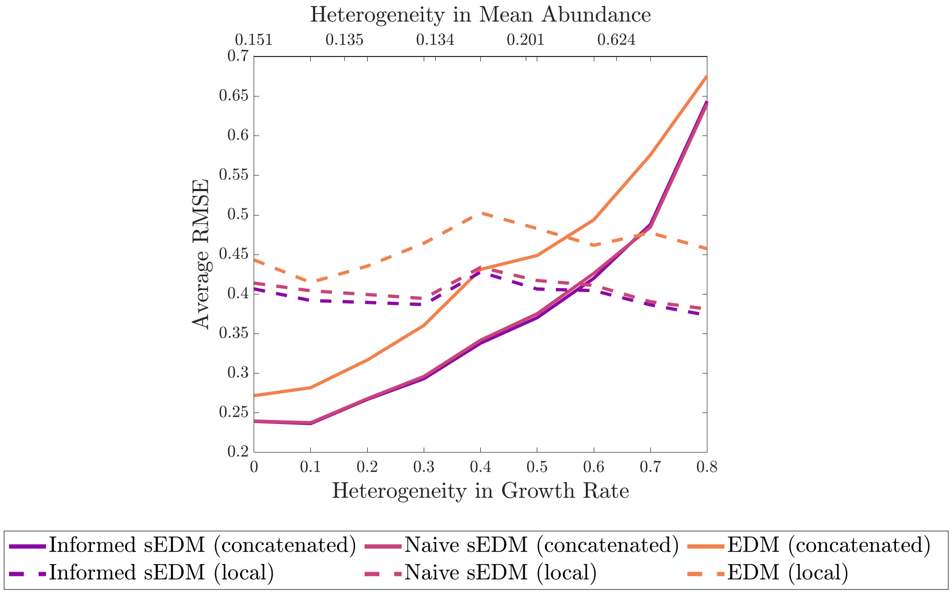

Figure A.3: Average RMSE on testing data over 50 simulations of Model 2 from Table 2.1 for various degrees of heterogeneity and asymmetric coupling with unidirectional dispersal. Results are shown for local (dotted lines) and concatenated (solid lines) informed sEDM (purple), naive sEDM (pink), and EDM (yellow). We imposed heterogeneity by varying one parameter (growth rate) across sites with a sinusoidal function. The measure of heterogeneity in growth rate is given by the amplitude of the sinusoidal function used to vary the parameter. The auxiliary x-axis at the top of each figure is the empirical measure of heterogeneity in mean abundance (see Methods).

and $T_{test} = 20$ with an embedding dimension of $E = 5$ and spatial embedding dimension $S = 5$. This resulted in the following RMSEs averaged over 100 simulations:

Informed sEDM (local):	0.9302
Informed sEDM (concatenated):	0.5081
Naive sEDM (local):	0.9322
Naive sEDM (concatenated):	0.6498
EDM (local):	0.8252

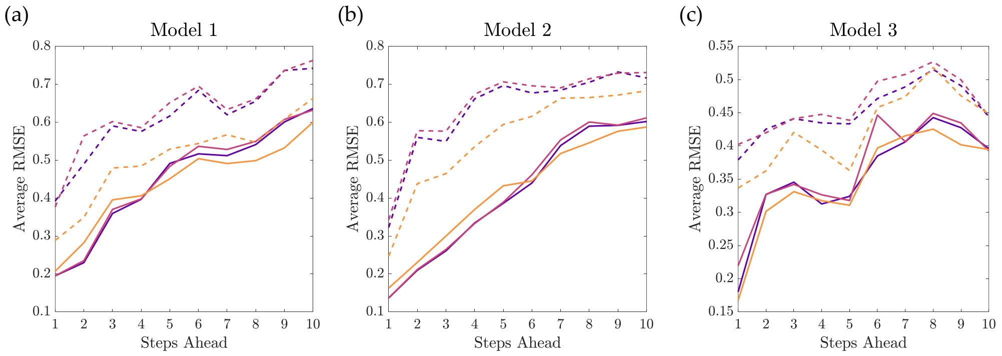

Figure A.4: Average RMSE on testing data over 20 simulations of each model from Table 2.1 for various steps ahead. Results are shown for local (dotted lines) and concatenated (solid lines) informed sEDM (purple), naive sEDM (pink), and EDM (yellow). We computed multi-step ahead dynamics by training the GPs on multi-step map.

Name	Model	Parameters
Competition	$f(\bullet) = xe^{r_1(1-A_{11}x-A_{12}y-A_{13}z-A_{14}w)}$	$r_1 = 2.75$, $r_2 = 2.8$,
(4 species: x,y,z,w)	$g(\bullet) = ye^{r_2(1-A_{21}x-A_{22}y-A_{23}z-A_{24}w)}$	$r_3 = 2.9$,$r_4 = 2.6$
	$h(\bullet) = ze^{r_3(1-A_{31}x-A_{32}y-A_{33}z-A_{34}w)}$	
	$k(\bullet) = we^{r_4(1-A_{41}x-A_{42}y-A_{43}z-A_{44}w)}$	$A = \begin{bmatrix} 1 & 1.09 & 1.52 & 0 \\ 0 & 1 & 0.44 & 1.36 \\ 2.33 & 0 & 1 & 0.47 \\ 1.21 & 0.51 & 0.35 & 1 \end{bmatrix}$

Table A.7: Functions and parameter values of a four-species Ricker competition model. Nearest neighbor dispersal on a one-dimensional lattice was generated with time series updates given by

$$x^i_{t+1} = \left[(1-\mu)f(\bullet^i_t) + \frac{\mu}{2}\left(f(\bullet^{i-1}_t) + f(\bullet^{i+1}_t)\right)\right] e^{\xi_t}$$
$$y^i_{t+1} = \left[(1-\mu)g(\bullet^i_t) + \frac{\mu}{2}\left(g(\bullet^{i-1}_t) + g(\bullet^{i+1}_t)\right)\right] e^{\xi_t}$$
$$z^i_{t+1} = \left[(1-\mu)h(\bullet^i_t) + \frac{\mu}{2}\left(h(\bullet^{i-1}_t) + h(\bullet^{i+1}_t)\right)\right] e^{\xi_t}$$
$$w^i_{t+1} = \left[(1-\mu)k(\bullet^i_t) + \frac{\mu}{2}\left(k(\bullet^{i-1}_t) + k(\bullet^{i+1}_t)\right)\right] e^{\xi_t}$$

The big dot notation indicates that the function takes x, y, z, w inputs at the specified time and spatial coordinate. For all simulations, the noise term was drawn independently from a normal distribution, i.e. $\xi_t \sim N(-s^2/2, s^2)$, with $s = 0.02$.

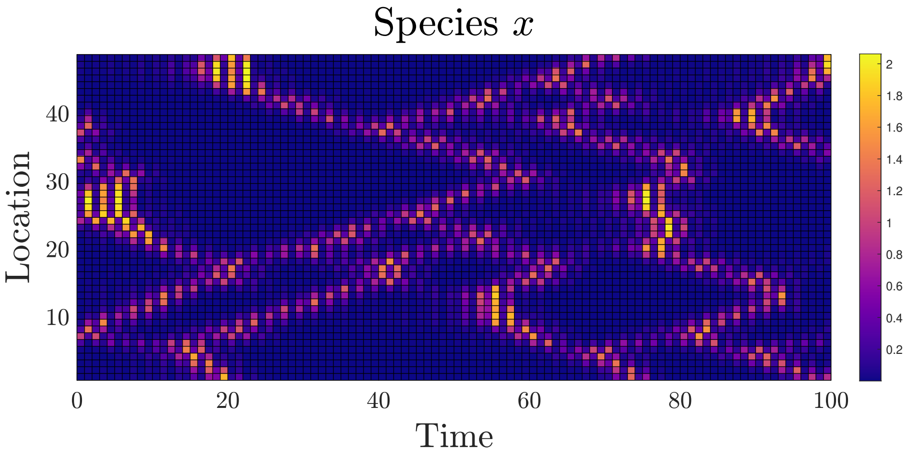

Figure A.5: Dynamics of the first species in a four-species Ricker competition model given in Table A.7.

$$\text{EDM (concatenated): } 0.6128$$

Since our EDM models typically only include 5 lags in time and the periodicity is much longer, local methods perform poorly because most lags do not contain any dynamics. The local methods were less than 20% better than the mean predictor on average, and the concatenated methods also suffered, never reducing error more than 50% from the mean predictor. This highlights the importance of scale when using these methods.

Appendix B

to Empirical Dynamic Modeling for Missing or Irregular Data

B.1 Models for simulated data

We simulated ecological dynamics using three models: (a) Ricker dynamics, (b) host-parasitoid dynamics, and (c) three-species competition dynamics (Table 3.1). These models were chosen to represent common types of species interactions that are known to exist in ecological data.

B.2 Empirical data

Laboratory data on rotifer dynamics (Figure B.1) and field data from the Rothamsted Insect Survey (Figure B.2) were used. Predicted vs. observed dynamics for both empirical examples were plotted (Figures B.3 and B.4).

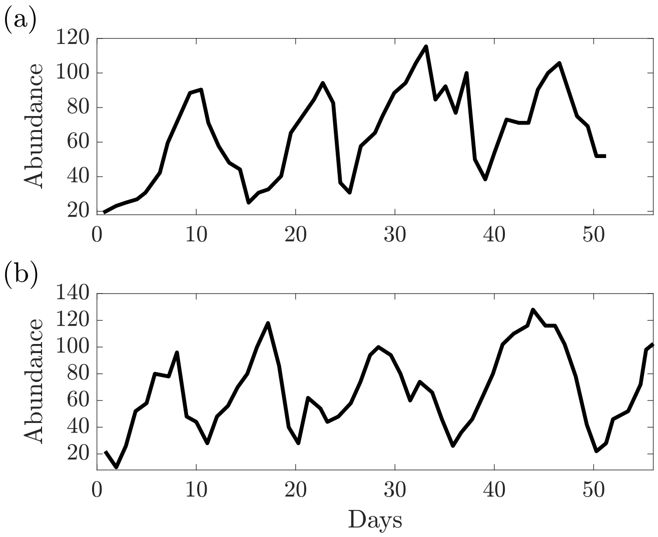

Figure B.1: Rotifer dynamics. Population dynamics of the rotifer *Brachionus calyciflorus* in experiments from Halbach (1984). Two separate experiments (experiment 1 (a) and experiment 2 (b)) were performed at 20°C. Samples were taken at irregular intervals from 0.5 to 1.5 days.

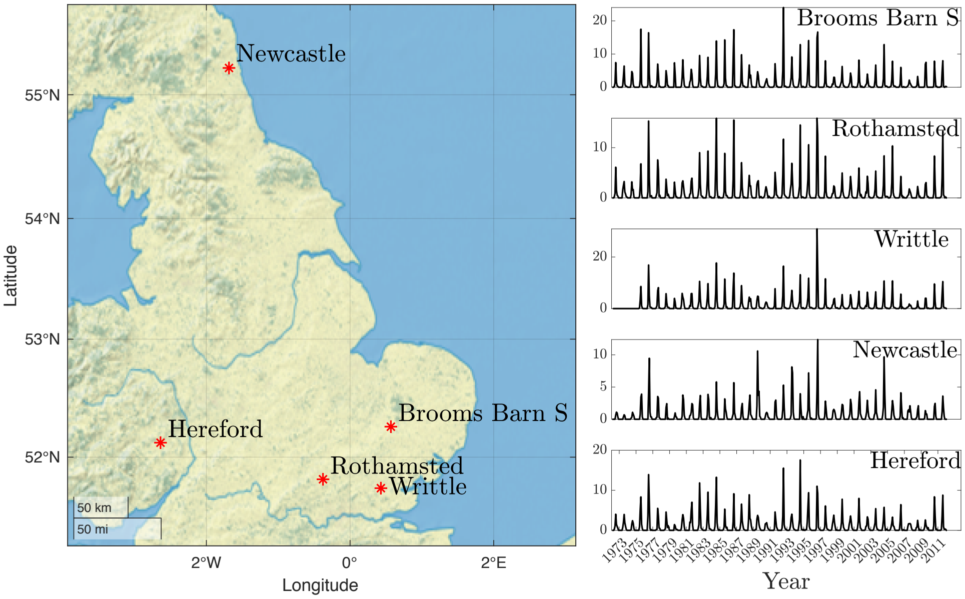

Figure B.2: Aphid survey sites and dynamics. Rothamsted insect survey sites and dynamics used to forecast dynamics of the English grain aphid (*Sitobion avenae*).

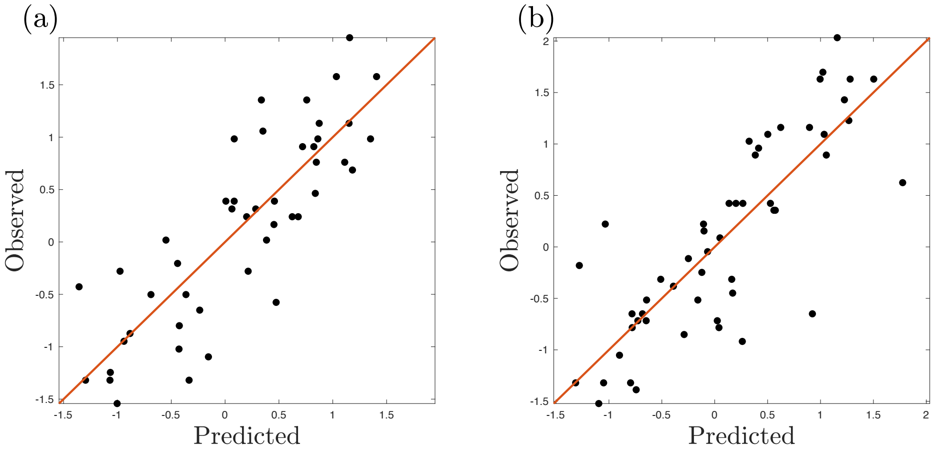

Figure B.3: Prediction accuracy for laboratory data. Predicted vs. observed values of variable step-size EDM for (a) experiment 1 and (b) experiment 2 from Halbach (1984). Black dots are predicted and observed values, and the red line is the one-to-one line.

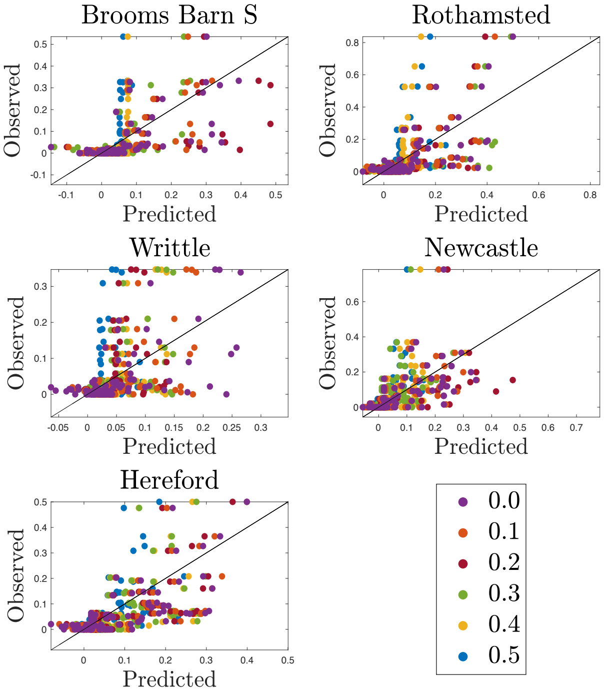

Figure B.4: Prediction accuracy on field data. Predicted vs. observed values of variable step-size EDM for the five Rothamsted insect survey sites. Different colors represent different proportions of missing data, and the solid black line is the one-to-one line.

B.3 Additional analysis on state-dependent missing values

All methods we tested were extremely sensitive to state-dependent missing values (i.e. those missing below a threshold). We hypothesized that this sensitivity was due to the models we used, all of which have strong divergence in dynamics for small population sizes. Since the models diverge less for large population sizes, we evaluated how the methods performed when we removed all points that fall above a set percentage of the simulated range of states, which we varied from 85-100% in 3% increments. As expected, when the missing values fall in this less variable region of the state space, all of the methods were considerably less sensitive to missing values (Figure B.5).

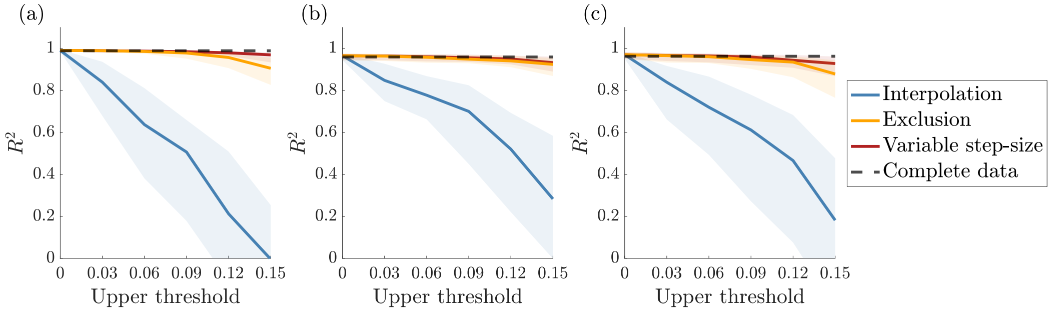

Figure B.5: Effect of state-dependent missing values on forecast accuracy. Median (solid line) and first and third quartiles (lower and upper limits of colored band, respectively) of R^2 vs. the upper proportional threshold at which values are missing. Results are shown for variable step-size EDM (red), the exclusion method (yellow), the interpolation method (blue), and EDM of complete data (dashed black) on three models: (a) Ricker dynamics, (b) host-parasitoid dynamics, and (c) three-species competition dynamics (Table 3.1).

Limitations of sampling equipment or design sometimes cause values below a detection threshold to be recorded as false zeros in ecological surveys. Similarly, when a time series is missing values, it is relatively common for EDM users to fill them with zeros (e.g. Kawatsu et al. (2021)). From an EDM point of view, these biased estimates may reduce forecast accuracy because they mischaracterize the relationship between past and future states. Although modelers do this, one may ask whether it is better to avoid the biased values entirely by treating them as missing values and performing VS-EDM.

Thus, we evaluated VS-EDM forecasts in which values below a threshold were treated as missing and compared them to standard EDM forecasts in which these values were assigned zero. We compared these two approaches for all three models

(Table 3.1) and generated the missing values using the methods described in for scenario IV (Chapter 3). Results indicate that for small detection thresholds (i.e., < 0.05), VS-EDM yields more accurate forecasts, but for larger thresholds, standard EDM of the biased data does so (Figure B.6). Importantly, in real applications it may not be clear where this intersection occurs, and the "smallness" of the detection threshold may be uncertain. Additionally, this result probably depends on the curvature of the dynamics where the data are missing or biased. In light of this, we again tested both of these methods on the same three models, but with different parameters that generate periodic dynamics. In this case, VS-EDM performed better than EDM on biased data, even for large proportions of missing values for two of the models (Figure B.7a,b), but suffered greatly for the third model (Figure B.7c). Thus, the outcome of this analysis appears to be case-specific, and in real EDM applications in which false zeros are a concern, it may be useful to apply both of these approaches.

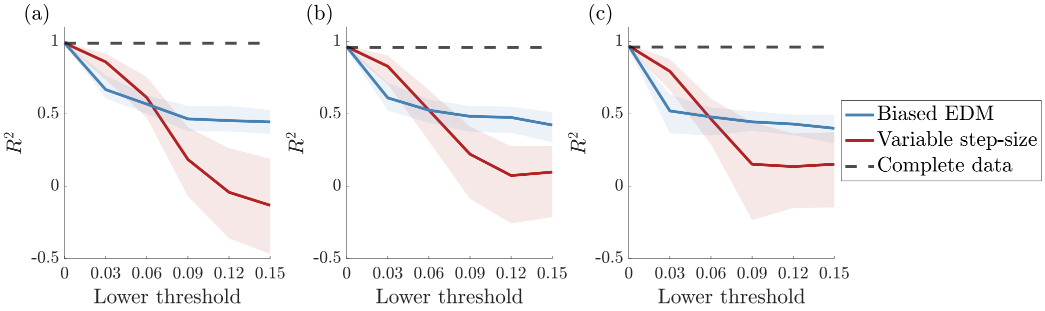

Figure B.6: Comparison of missing vs. biased values on forecast accuracy. Median (solid line) and first and third quartiles (lower and upper limits of colored band, respectively) of R^2 vs the lower proportional threshold at which values are missing or assigned a value of zero in the training data. Results are shown for variable step-size EDM when the values are missing (red), EDM when the values are zero (blue), and EDM of complete data (dashed black) for the three models, which generate chaotic (a) Ricker dynamics, (b) host-parasitoid dynamics, and (c) three-species competition dynamics (Table 3.1).

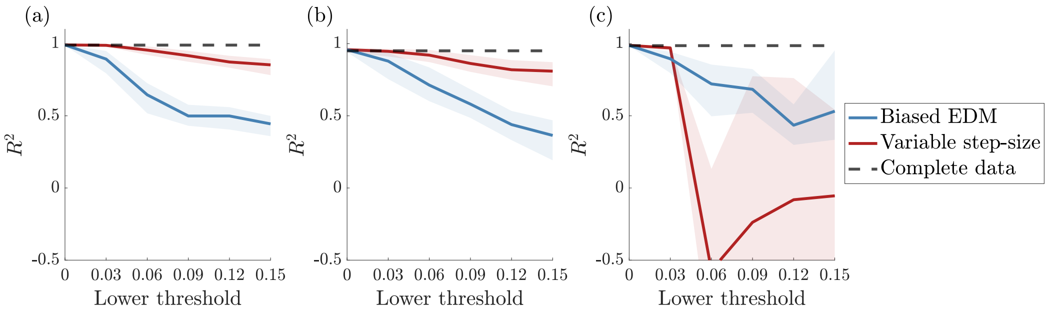

Figure B.7: Comparison of missing vs. biased values on forecast accuracy (periodic dynamics). Median (solid line) and first and third quartiles (lower and upper limits of colored band, respectively) of R^2 vs the lower proportional threshold at which values are missing or assigned a value of zero in the training data. Results are shown for variable step-size EDM when the values are missing (red), EDM when the values are zero (blue), and EDM of complete data (dashed black) for the three models, which generate periodic (a) Ricker dynamics, (b) host-parasitoid dynamics, and (c) three-species competition dynamics (Table 3.1). The parameter values to generate periodic dynamics are given in Table 3.1 with $r = 2.6$ instead of $r = 3.0$ for all models.

B.4 Sensitivity to Time Series Length

The primary disadvantage of using VS-EDM in ecological applications is that it doubles the dimension of inputs, which may be problematic for short time series, which are common in ecology. To characterize the sensitivity to short time series, we perform VS-EDM for time series of different lengths. For each model (Table 3.1), we assigned a random initial value to each variable, iterated the model for 140, 155, or 170 steps, and removed the first 100 points to avoid transient dynamics. This left time series of length 40, 55, and 70, which we split into 20, 35, or 50 'training' data points for fitting the Gaussian process (GP) hyperparameters and 20 'testing' points. For each model, we subsampled the training data to generate various amounts of randomly occurring missing data using the method described for scenario I (Chapter 3).

As expected, shorter time series yield worse predictions: with a training dataset of 50, 35, and 20 points, forecasts are usable up to 40%, 30%, and 20-25% missing, respectively (Figure B.8). Because it is well known that fewer inputs yield worse predictions, it may be more interesting to compare the forecast accuracy at an equivalent scale – the number of points that VS-EDM uses. Accordingly, we find the same results as a function of the number of inputs; given the same number of inputs, VS-EDM gives higher forecast accuracy when there are fewer missing values (Figure B.9). For instance, a time series of 50 points with 50% missing and a time series of 35 points with 30% missing both use 25 inputs, but the latter (shorter, with fewer missing values) gives much more accurate forecasts (Figure B.9). Furthermore, a regularly sampled time series of 20 points yields more accurate forecasts than 25 sparsely sampled points from a time series of 50 points. This result indicates that as new τs are introduced, longer time series to are needed produce equivalent forecasts.

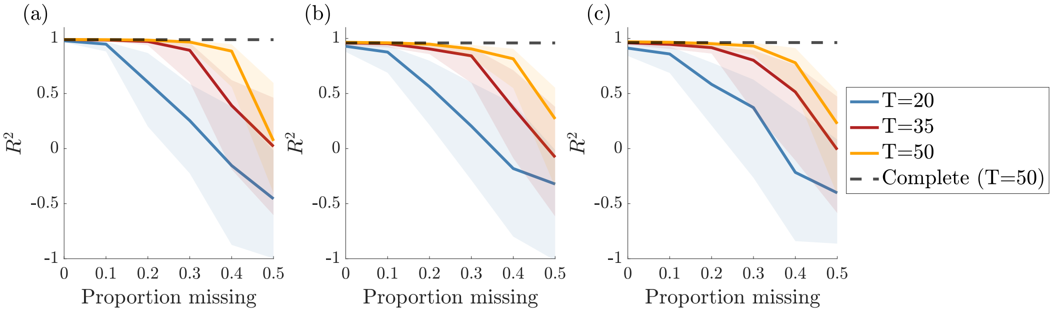

Figure B.8: Comparison of time series length. Median (solid line) and first and third quartiles (lower and upper limits of colored band, respectively) of R^2 vs. the proportion of random missing training data of variable step-time EDM with a training time series length of 20 points (blue), 35 points (red), and 50 points (yellow). The dashed black line shows the mean R^2 when the data have no missing values. Results are shown for three models: (a) Ricker dynamics, (b) host-parasitoid dynamics, and (c) three-species competition dynamics (Table 3.1).

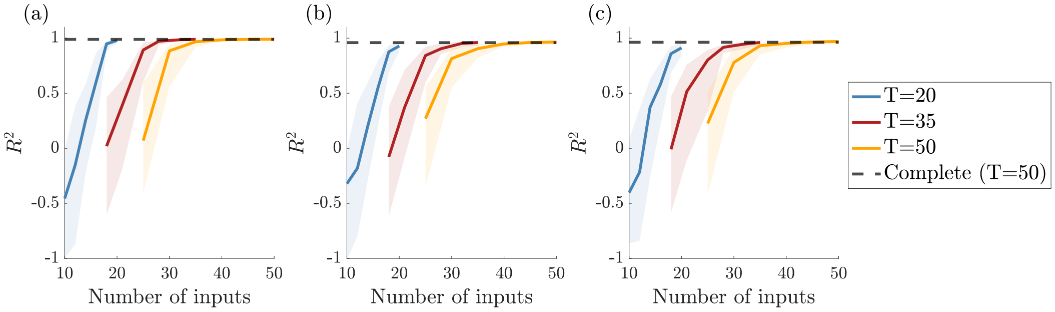

Figure B.9: Effect of number of inputs on forecast accuracy for different time series lengths. Median (solid line) and first and third quartiles (lower and upper limits of colored band, respectively) of R^2 vs. the proportion of random missing training data of variable step-size EDM with a training time series length of 20 points (blue), 35 points (red), and 50 points (yellow). The dashed black line shows the mean R^2 when the data have no missing values. Results are shown for three models: (a) Ricker dynamics, (b) host-parasitoid dynamics, and (c) three-species competition dynamics (Table 3.1).

B.5 Augmented VS-EDM with artificial missing values

As mentioned in the main text, sometimes only a few inputs exist for a given inter-sample step size, τ, which may reduce the forecast accuracy of VS-EDM. However, complete portions of the time series could be used to create more inputs corresponding to any given τ. We performed a preliminary test this idea for the three models (Table 3.1) with the following procedure. For each time series, we first created an identical replicate of the 50-point training data (which contains missing values) and randomly selected and removed an additional 10 points (i.e. an additional 20%). We stitched this new time series to the end of the original training data, which effectively doubled the length of the time series, created more replicates of existing τ's, and potentially introduced a few new values of τ. We repeated this procedure 5 times until we had a training dataset that was 6 times the length of the original time series and had more missing values. From this long time series, we constructed the VS-EDM delay-coordinate vectors, avoiding time lags that crossed over multiple replicates of the time series. Since the delay-coordinate vectors were constructed from identical time series replicates, we removed any duplicate embedding vectors to avoid redundancy and numerical issues in the GP matrix inversion step. Once the GP hyperparameters were estimated from this augmented training data, we forecasted the 20 out-of-sample testing points.

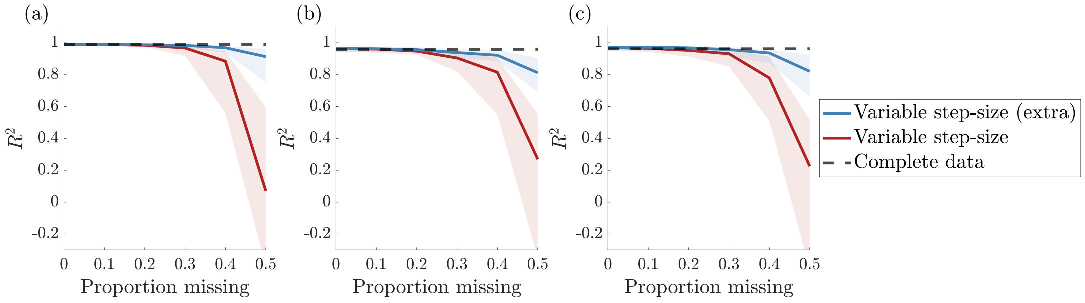

Figure B.10: Comparison of standard variable step-size EDM (VS-EDM) and an augmented version. Median (solid line) and first and third quartiles (lower and upper limits of colored band, respectively) of R^2 vs. the proportion of random missing training data of standard VS-EDM (red) and augmented VS-EDM (blue). The dashed black line shows the mean R^2 when the data have no missing values. Results are shown for three models: (a) Ricker dynamics, (b) host-parasitoid dynamics, and (c) three-species competition dynamics (Table 3.1).

The augmented version of VS-EDM greatly improved forecasts for all three models, particularly for large proportions of missing data, indicating that it may be a valuable adaptation of VS-EDM (Figure B.10). However, several statistical concerns cannot be ignored. Specifically, when time series are duplicated to create additional data, the outputs become non-independent. We propose that this may be a reasonable step forward, but more theoretical work is needed to justify its general use.

B.6 Estimating Lyapunov exponents from irregularly sampled time series

In the main text (Chapter 3), we consider the possibility of extending the concept of VS-EDM to estimate Lyapunov exponents (LE) from irregularly spaced time series. To make our argument transparent in a single-variable system, we first recall how the LE is calculated when the time step is fixed. The LE is defined as the exponential growth rate of an infinitesimal perturbation in the long-time limit (Strogatz, 1994; Wolf et al., 1985). That is, in 1-d, if $x_t = F_t[x_0]$, then the LE is given by

$$\begin{aligned}\lambda &= \lim_{t\to\infty,\Delta x_0\to 0} \frac{1}{t}\ln\left|\frac{\Delta x_t}{\Delta x_0}\right| \\ &= \lim_{t\to\infty,\Delta x_0\to 0} \frac{1}{t}\ln\left|\frac{F_t[x_0+\Delta x_0]-F_t[x_0]}{\Delta x_0}\right| \\ &= \lim_{t\to\infty} \frac{1}{t}\ln\Big|F_t'[x_0]\Big|. \end{aligned} \tag{B.1}$$

For discrete time systems, $F_t'[x_0]$ is evaluated using the chain rule, taking advantage of the fact that the t-step ahead map is the t^{th} iterate of the 1-step map, $x_{t+1} = F_1[x_t]$, so that $F_t'[x_0] = \prod_{i=0}^{t-1} F_1'[x_i]$ and so that $\lambda = \lim_{t\to\infty} \frac{1}{t}\sum_{i=0}^{t-1} \ln|F_1'[x_i]|$. See Wolf et al. (1985) for more details.

For the irregularly sampled map, we proceed analogously, noting that $x_t = F_t[x_0]$ implies $x_t = F_{t-\tau}[F_\tau[x_0]]$ so that the derivative $F_t'[x_0] = F_{t-\tau}'\left[F_\tau[x_0]\right] F_\tau'[x_0] = F_{t-\tau}'[x_\tau]F_\tau'[x_0]$. Hence, the LE may be calculated from the sequence of unequal time-step maps as $\lambda = \lim_{t\to\infty} \frac{1}{t}\sum_{i=0}^{t-1} \ln|F_{\tau_i}'[x(t_i)]|$ where τ_i is the duration of the i^{th} interval, $t_i = \sum_{j=0}^{i-1} \tau_j$ is the time at the start of the i^{th} interval, and t is the total time.

To demonstrate the utility of this approach, we simulated time series of length

200 from a logistic map ($x_{t+1} = rx_t(1-x_t)$) for a range of growth rates (r) between 3.5 and 4. From the complete time series corresponding to each r, we estimated the LE with $\lambda = \lim_{t\to\infty} \frac{1}{t}\sum_{i=0}^{t-1} \ln|F_1'[x_i]|$. Note that we approximated F with a GP and computed its derivatives from the approximation. See Rasmussen and Williams (2006) for a description of how to get derivatives from the GP. Once we estimated the LE from the complete data, we randomly selected and removed 20% of the points from the time series, leaving us with unevenly sampled time series. We used the variable step-size LE approach on these unevenly sampled time series with $\lambda = \lim_{t\to\infty} \frac{1}{t}\sum_{i=0}^{t-1} \ln|F_{\tau_i}'[x(t_i)]|$. We also evaluated interpolation and exclusion on the sparsely sampled data and quantified the accuracy of all methods by calculating their R^2s compared to the "right" answer, which uses the exact derivatives from the logistic map (i.e. $F'(x) = r - 2rx$) instead of the GP approximations. Since estimates of the derivatives from a GP are highly state-dependent and can influence the accuracy of LE estimates, we also performed the same procedure when samples below 0.1 were missing.

For a complete time series, we obtain highly accurate estimates of the LEs (R^2=0.97) for r from 3.5-4.0 (Figures B.11 and B.12). With 20% of the points missing at random from the time series, the variable step-size LE approach does well ($R^2 = 0.85$), interpolation does poorly ($R^2 = -4.55$), and exclusion does well ($R^2 = 0.84$). With points below 0.1 missing, the variable step-size LE approach still does well ($R^2 = 0.86$), interpolation gives $R^2 = -0.92$, and exclusion gives $R^2 = 0.80$. Overall, interpolation performs poorly for these data because they vary greatly. Exclusion does well because the data are plentiful and the dynamics are 1-d, so the order of derivative multiplication does not matter. In both cases, using the variable step-size method has an advantage, particularly when missing values are state-dependent. We expect that this advantage will be greater in higher-dimensional systems.

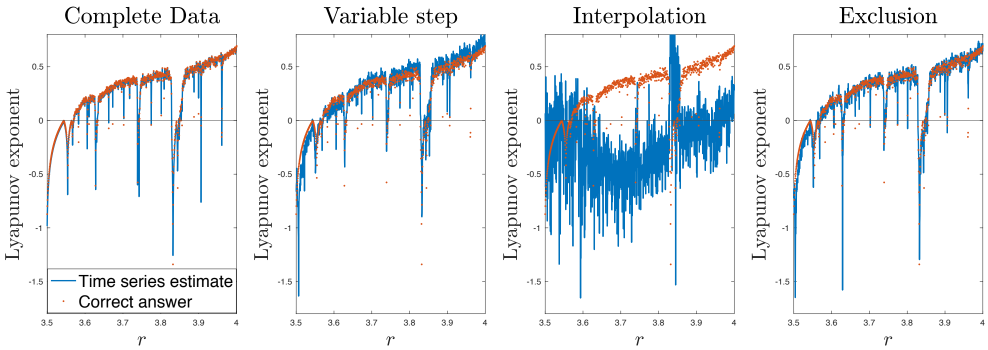

Figure B.11: Lyapunov exponent estimates from time series (missing at random): Estimated Lyapunov exponents from 200-point time series generated with the logistic map. From left to right, results are shown when using the complete time series and, for the time series with 20% missing, the variable step-size LE method, interpolation method, and exclusion method.

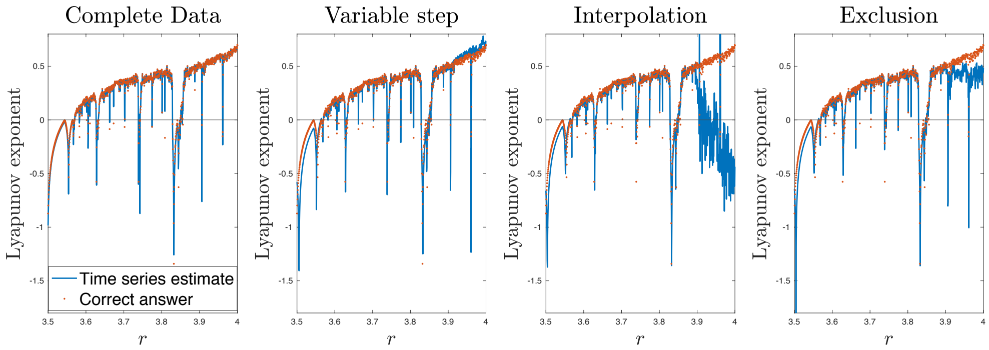

Figure B.12: Lyapunov exponent estimates from time series (missing below a threshold): Estimated Lyapunov exponents from 200-point time series generated with the logistic map. From left to right, results are shown when using the complete time series and, when points below 0.1 are missing, the variable step-size LE method, interpolation method, and exclusion method.

Importantly, the theory and example presented here applies only if there is one independent state variable. Most ecological systems are not one-dimensional, however, so the method needs to be extended to accommodate time series that come from systems with multiple interacting species.

In more than one dimension, stability is still defined by the long run growth of a perturbation to $\mathbf{x}(t) = \mathbf{F}_t[\mathbf{x}(0)]$. This is determined with the Jacobian matrix, $\mathbf{J}_t$ whose elements are $(\mathbf{J}_t)_{ij} = \frac{\partial \mathbf{F}_{it}}{\partial \mathbf{x}_j(0)}$. If ψ_t is the dominant eigenvalue of $\mathbf{J}_t$, the LE is given by $\lambda = \lim_{t\to\infty} \frac{1}{t} \ln|\psi_t|$. By analogy with the 1-d case, we calculate $\mathbf{J}_t$ via the chain rule as the product of Jacobians for each of the τ_i-step ahead maps. In the case where we are using delay coordinates, these τ_i-step ahead Jacobians have a particularly simple form,

$$J[x(t_i)] = \begin{bmatrix} \frac{\partial F}{\partial x(t_{i-1})} & \frac{\partial F}{\partial x(t_{i-2})} & \cdots & \frac{\partial F}{\partial x(t_{i-E})} \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & 0 \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Thus, to estimate the dominant LE from the time series, we would use the GP to estimate the variable time-step map, $x(t_i) = F[x(t_{i-1}), \tau_{i-1}, \ldots, x(t_{i-E}), \tau_{i-E}]$ with $\tau_j = t_{j+1} - t_j$. Then for each $x(t_i)$, assemble $\mathbf{J}[x(t_i)]$ from the estimated partial derivatives, calculate the product over the observations to obtain $\mathbf{J}_t$, and estimate the LE from the log absolute value of the dominant eigenvalue. We leave the test of this method for a future analysis.

B.7 Estimating continuous time dynamics from discrete data

As detailed in the main text (Chapter 3), the concept of incorporating variable step-sizes can be extended into delay-coordinate vectors beyond forecasting and used to estimate continuous time dynamics. This is easiest to see for 1-d dynamics, for which VS-EDM was used to estimate the map $x_{t+\tau} = f(x_t, \tau)$. Recalling that

$f(x_t, \tau) = x_t + \int_t^{t+\tau} g(x_s)ds$ (Eq. 3.3), in which continuous-time dynamics are given by $\dot{x} = g(x)$, $\frac{\partial f}{\partial \tau}$ evaluated at x_t and $\tau = 0$ is an estimate of $g(x_t)$.

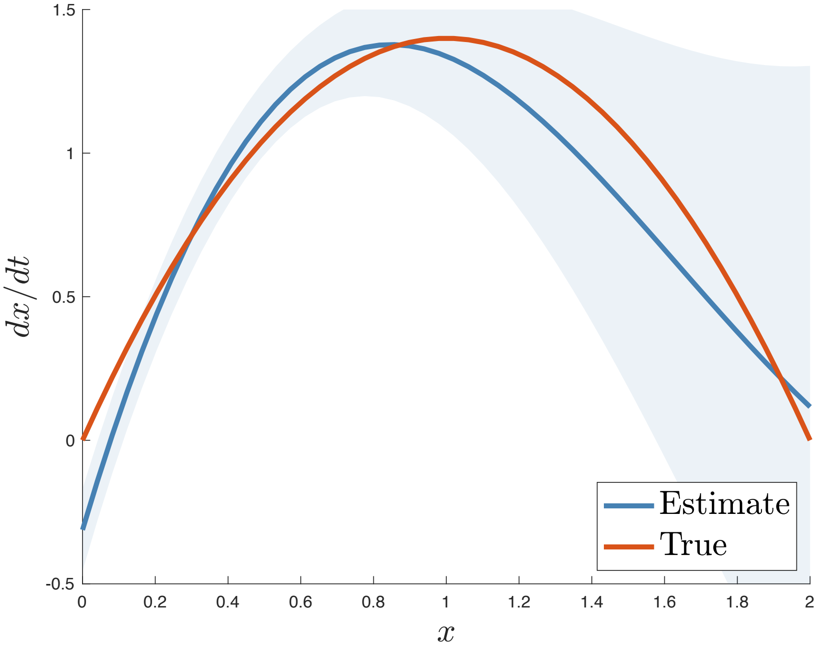

Figure B.13: Estimates of continuous-time derivatives from discrete-time data. True values (orange) and mean estimates over 1000 replicates (blue) of $\frac{dx}{dt}$ for values of x from 0-2. The blue band is the standard deviation over the 1000 replicates.

To demonstrate this idea on a simple 1-d system, we simulated dynamics from a continuous-time logistic model (i.e. $\dot{x} = 2.7x(1 - x/2)$) starting from random initial x_0 ranging from 0 to 1. We integrated the model for a random time span τ ranging from 0 to 3 and repeated this procedure 100 times to get 100 $[x_0, \tau, x_\tau]$ triplets. Taking advantage of the fact that we know the solution at $\tau = 0$ is x_0, we then augmented the set of 100 triplets with 50 additional triplets given by $[x_g, 0, x_g]$, were the x_g values were evenly spaced between 0 and 2. Given this set of 150 triplets, we estimated $x_\tau = f(x_0, \tau)$ with a GP using a mean function of $m(x_0, \tau) = x_0$ and a squared exponential covariance function of

$C(\{x,\tau\},\{x',\tau'\}) = \sigma^2 e^{-\phi_x(x-x')^2-\phi_\tau(\tau-\tau')^2}$. Then we evaluated $\frac{df}{d\tau}$ at each point on x_g, and $\tau = 0$. See Rasmussen and Williams (2006) for a description of how to estimate derivatives from the GP. We repeated this procedure 1000 times, and the results are shown in Figure B.13. The estimated continuous time derivatives are close to the true continuous derivatives with an $R^2 = 0.89$.

Appendix C

to Empirical Dynamic Programming for Insect Pest Management

C.1 Simulation Models

We generated simulated data with the models in Table C.1.

C.2 Additional results

As expected, increasing noise in the dynamics decreases the forecast accuracy of both GP regression with the full state and EDM with lags of the pest. Table C.2. shows the R^2 forecast accuracy in both cases over all the data from the simulation analysis. In general, EDM yields lower forecast accuracy than GP regression because it is missing data. The accuracy of the GP in the high noise

Type of control	Model	Parameters
Biological	$H_{t+1} = H_t e^{r(1-H_t/K)-\alpha(P_t+cu_t)}$ $P_{t+1} = \beta H_t(1 - e^{-\alpha(P_t+cu_t)}) + \gamma$	$r = 3.4$, $K = 50$, $\beta = 1.2$, $\gamma = 2$ $\alpha = 1$, $c = 0.7$
Chemical	$H_{t+1} = (1 - cu_t)H_t e^{r(1-(1-cu_t)H_t/K)-\alpha P_t}$ $P_{t+1} = \beta(1 - cu_t)H_t(1 - e^{-\alpha P_t}) + \gamma$	$r = 3.4$, $K = 50$, $\beta = 1.2$, $\gamma = 2$, $\alpha = 1$, $c = 0.6$
Behavioral	$H_{t+1} = H_t e^{(1-cu_t)r(1-H_t/K)-\alpha P_t}$ $P_{t+1} = \beta H_t(1 - e^{-\alpha P_t}) + \gamma$	$r = 3.4$, $K = 50$, $\beta = 1.2$, $\gamma = 2$, $\alpha = 1$, $c = 0.7$
IPM	$H_{t+1} = (1 - cu_t)H_t e^{r(1-(1-cu_t)H_t/K)-\alpha(P_t+dv_t)}$ $P_{t+1} = \beta(1 - cu_t)H_t(1 - e^{-\alpha(P_t+dv_t)}) + \gamma$	$r = 3.4$, $K = 50$, $\beta = 1.2$, $\gamma = 2$, $\alpha = 1$, $c = 0.6$, $d = 0.7$

Table C.1: Host-parasitoid models used to generate simulated data. Noise was included by multiplying the equations by e^{ξ}, where ξ was drawn independently from a normal distribution $\mathcal{N}(0, s^2)$. We used $s = 0.05$ to generate low noise in the data, $s = 0.1$ for moderate noise, and $s = 0.15$ for high noise.

regime is approximately 20% lower than it is in the low noise regime, while EDM accuracy is 30% worse in the high noise conditions compared to low noise.

Type of control	Method	Low noise	Moderate noise	High noise
Biological control	GP	0.98	0.90	0.82
	EDM	0.93	0.79	0.66
Chemical control	GP	0.98	0.91	0.81
	EDM	0.90	0.75	0.64
Behavioral control	GP	0.98	0.91	0.81
	EDM	0.94	0.80	0.68
IPM	GP	0.97	0.89	0.80
	EDM	-	-	-

Table C.2: Average R^2 prediction accuracy in the test set over the 100 replicate simulations of 11 values of θ.

Figs. C.1 & C.2 show the excess cost and pareto fronts for the low noise scenario and the high noise scenario of biological, chemical, and behavioral control strategies. Unsurprisingly, EDP produced nearly optimal policies in the low noise regime. While performance decreased in high noise environments, EDP in both cases outperformed the reactive policy (Table C.3). Similar results hold for the IPM control strategy (Figs. C.3 & C.4).

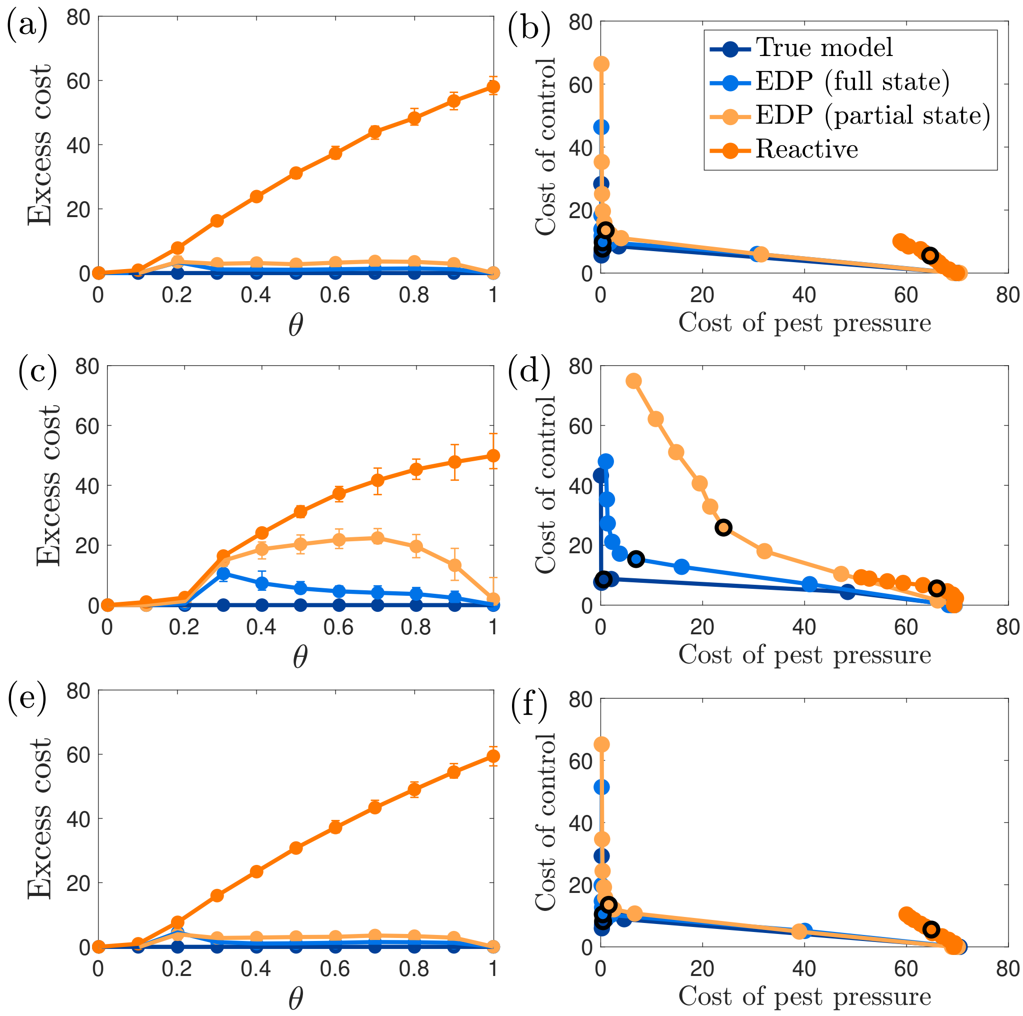

Figure C.1: Results over 100 simulations of biological control (a,b), chemical control (c,d), and behavioral control (e,f) with a low level of noise in the data. Left panels show the median (dots) and lower and upper quartiles (error bars) for the excess cost of each method compared to the optimal policy (a,c,e). Pareto fronts show the trade-off between the cost of pest pressure and the cost of control (b,d,f). Each point in the Pareto fronts corresponds to a value of θ, ordered such that $\theta = 0.001$ in the lower right $\theta = 0.999$ in the upper left.

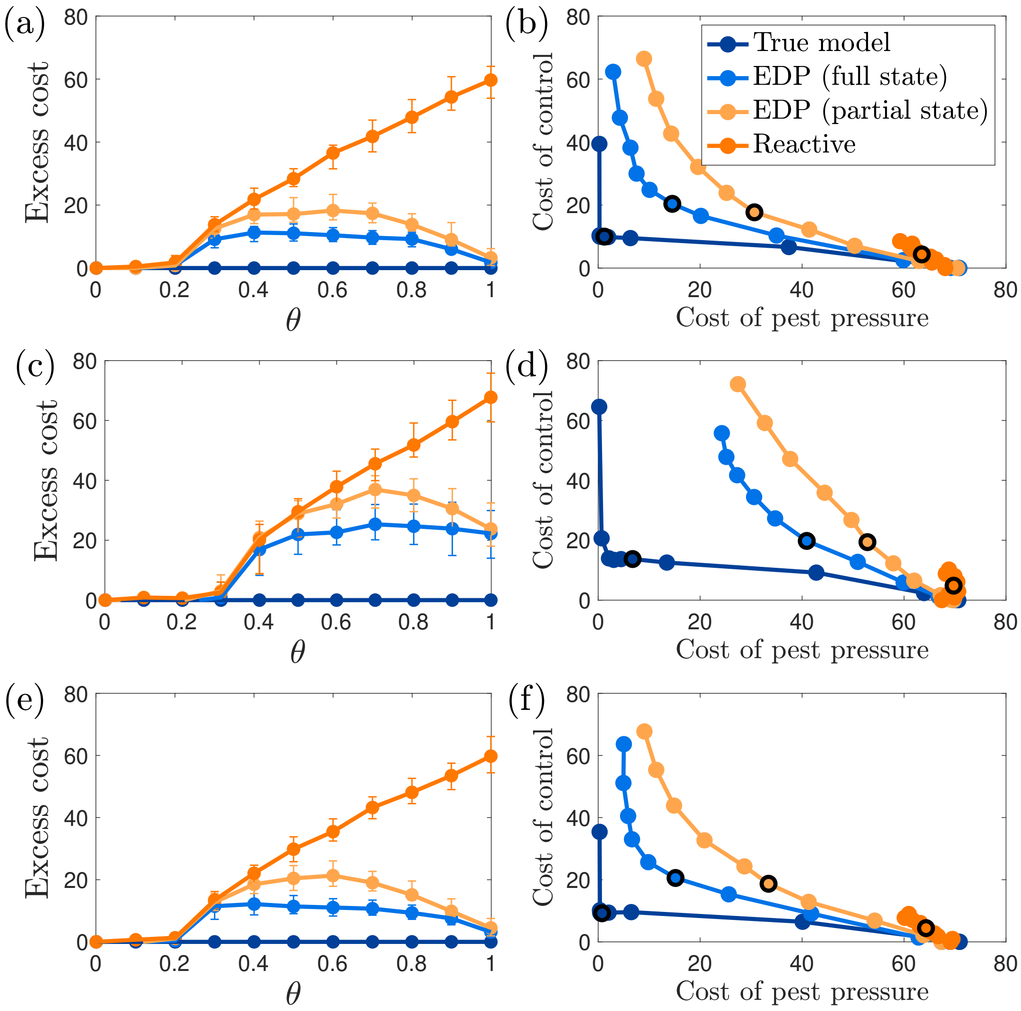

Figure C.2: Results over 100 simulations of biological control (a,b), chemical control (c,d), and behavioral control (e,f) with a high level of noise in the data. Left panels show the median (dots) and lower and upper quartiles (error bars) for the excess cost of each method compared to the optimal policy (a,c,e). Pareto fronts show the trade-off between the cost of pest pressure and the cost of control (b,d,f). Each point in the Pareto fronts corresponds to a value of θ, ordered such that $\theta = 0.001$ in the lower right $\theta = 0.999$ in the upper left.

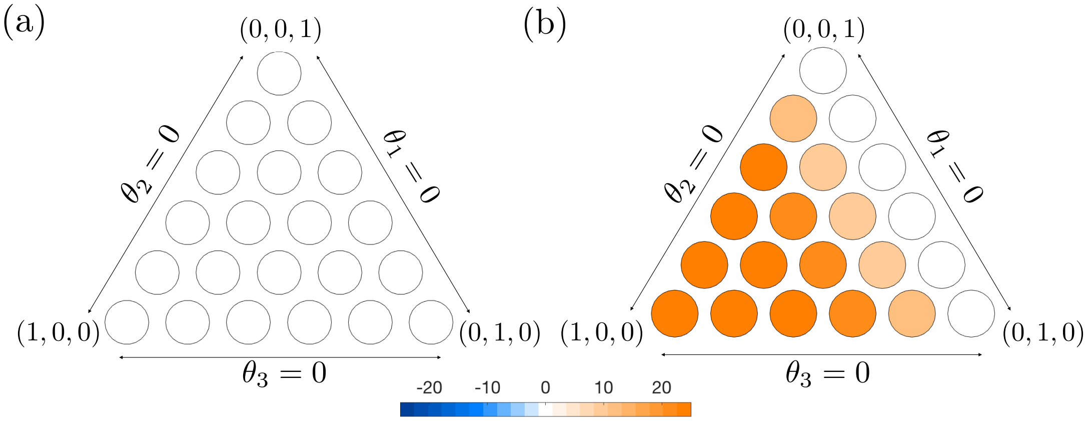

Figure C.3: Results over 100 simulations of IPM control using EDP (a) and a reactive control (b) with low levels of noise in the data. The color in each circle is the average excess cost for a specific $(\theta_1, \theta_2, \theta_3)$ triplet. Light circles indicate low excess cost and darker circles indicates high excess cost.

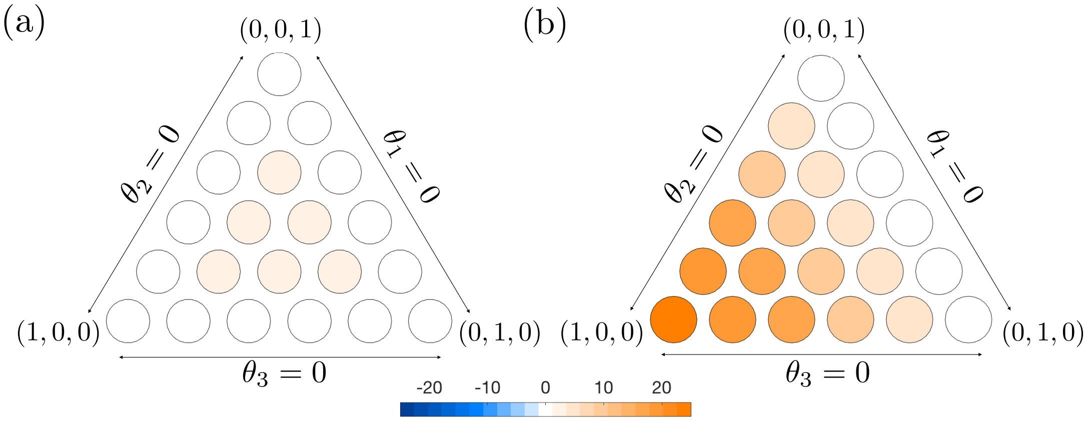

Figure C.4: Results over 100 simulations of IPM control using EDP (a) and a reactive control (b) with high levels of noise in the data. The color in each circle is the average excess cost for a specific $(\theta_1, \theta_2, \theta_3)$ triplet. Light circles indicate low excess cost and darker circles indicates high excess cost.

C.3 Effect of control history

It is possible that the amount and quality of the historical control data will influence the performance of the EDP method because control history can affect our ability to approximate dynamics. To determine the impact of control history, we explored six scenarios of control data. For each scenario, we generated time series of 600 points from the host parasitoid model with biological control and moderate noise (Table C.1). We removed the first 300 points to avoid transient dynamics, and then we split the subsequent 300 points into 100 points for training, 100 points for testing, and 100 points for evaluating the control strategies. The six scenarios are described below.

Random control history

The first scenario included a random control variable. We set the control variable to 0 for the first 50 points in the training data and set it randomly between 0 and 1 for the second 50 points of the training data. This scenario is not likely to be encountered in real applications, but it provides the best type of data for EDP, so it serves as a nice benchmark.

Reactive control history

For the reactive control history, the maximum amount of control ($u = 1$) was applied in the training data if the pest was above the economic threshold ($x_{thresh} = 2$). This scenario is likely to be encountered in real applications.

Calendar-based control history

In this scenario, we applied the maximum amount of control for three successive points periodically in the training data. The collections of sprays were spaced by twelve points every time. Since calendar-based programs are fairly common in pest management, similar data could be encountered in real systems.

Sparse control history

Sometimes control is only applied occasionally. For instance, if an insect has not presented a large threat in certain areas in the past, there may have been minimal effort to control it in the past. We simulated this scenario by randomly selecting 8 points in the training data to apply control. The rest of the control history in the training data was set to 0.

Imprecise data for control history

If our goal is to make precise decisions about when and how much control to apply, we must collect data enough data. We simulated a scenario in which a manager has only kept track of when control was applied but not how much control was applied. That is, any time control was applied, the data for the control variable was set to the maximum control.

Incomplete data for control history

In some cases, managers are responsible for multiple fields/areas and they do not have time to keep careful track of historical interventions. We simulated a scenario in which we used random control to generate the historical dynamics, but only recorded half of the historical control in the data. The control variable was set to 0 at the unrecorded times.

Results

Table C.4 shows the results of this analysis with the median and lower and upper quartiles of excess cost for all methods, and Fig. C.6 shows excess costs by θ. The scenarios in which the historical control strategy was independent of the state and frequently applied (i.e. random and calendar-based) gave the best results (Fig C.6 a, c). In contrast, with state-dependent historical policies (i.e. reactive, Fig C.6 b) or infrequent control applications (i.e. sparse, Fig C.6 d), excess cost increased. Imprecise and incomplete data for the historical control also hindered performance (Fig C.6 e,f). Although the control history had subtle impacts, the primary results remained consistent with those in the main text, and EDP outperforms the reactive policy on average.

Type of control	Method	Low noise	Moderate noise	High noise
Biological control	EDP (full state)	96%	87%	75%
	EDP (partial state)	92%	77%	57%
Chemical control	EDP (full state)	84%	69%	48%
	EDP (partial state)	52%	40%	33%
Behavioral control	EDP (full state)	96%	87%	73%
	EDP (partial state)	92%	75%	55%
IPM	EDP (full state)	99%	95%	88%
	EDP (partial state)	-	-	-

Table C.3: Average reduction in excess cost compared to the reactive policy over the 100 replicate simulations of 11 values of θ.

Scenario	EDP (full state)	EDP (partial state)	Reactive
Random	4.3 (0.1, 5.9)	7.5 (0.4, 10.5)	29.4 (7.6, 45.9)
Reactive	4.7 (0.5, 7.2)	8.3 (1.1, 12.9)	30.1 (8.6, 47.1)
Calendar	4.2 (0.2, 6.0)	7.5 (0.37, 10.5)	29.2 (7.2, 46.0)
Sparse	4.9 (0.5, 11.5)	11.3 (2.0, 29.0)	28.8 (6.1, 46.0)
Imprecise data	6.2 (0.4, 9.8)	8.6 (0.5, 13.1)	29.1 (8.2, 45.6)
Incomplete data	4.3 (0.2, 7.1)	7.6 (0.5, 12.5)	29.3 (7.2, 45.9)

Table C.4: Excess cost by scenario of historical control. The first number is the median excess cost and the parentheses are the first and third quartiles over 100 simulations of 11 values of θ.

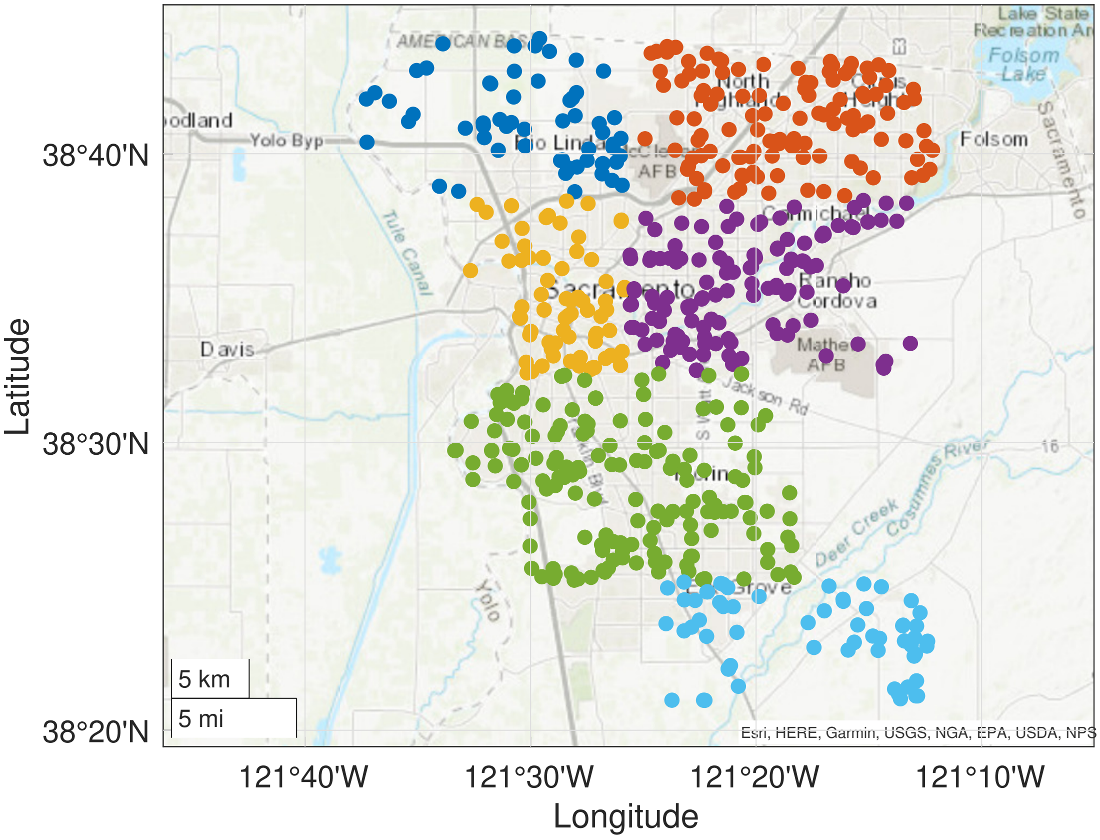

Figure C.5: Locations of mosquito traps. Colors represent the local regions.

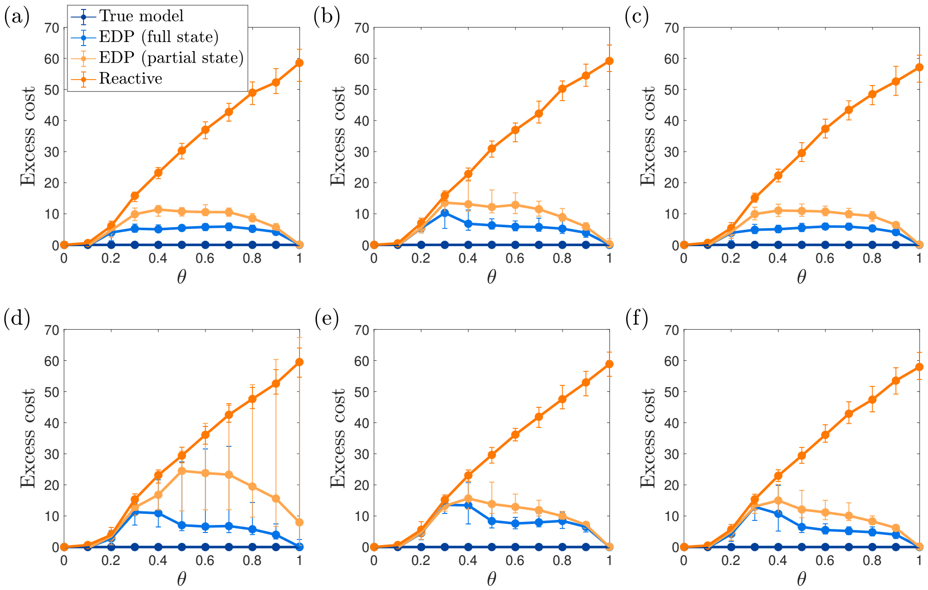

Figure C.6: Excess cost results over a 100 simulations with random (a), reactive (b), calendar-based (c), sparse (d), imprecise (e), and incomplete (f) control history data. Dots are median excess cost, error bars show the lower and upper quartiles.

Bibliography

Abarbanel, H. D. I. (1997). *Analysis of observed chaotic data*. Springer.

Afun, J., Jackai, L., and Hodgson, C. (1991). Calendar and monitored insecticide application for the control of cowpea pests. *Crop Protection*, 10(5):363–370.

Aksnes, D., Aure, J., Kaartvedt, S., Magnesen, T., and Richard, J. (1989). Significance of advection for the carrying capacities of fjord populations. *Marine Ecology Progress Series*, 50:263–274.

Arkema, K. K., Abramson, S. C., and Dewsbury, B. M. (2006). Marine ecosystem-based management: from characterization to implementation. *Frontiers in Ecology and the Environment*, 4(10):525–532.

Bakker, R., Schouten, J. C., Giles, C. L., Takens, F., and Bleek, C. M. V. D. (2000). Learning chaotic attractors by neural networks. *Neural Computation*, 12(10):2355–2383.

Banbrook, M., Ushaw, G., and Mclaughlin, S. (1997). How to extract lyapunov exponents from short and noisy time series. *IEEE Transactions on Signal Processing*, 45(5):1378–1382.

Barnett, L. A., Ward, E. J., Jannot, J. E., and Shelton, A. O. (2019). Dynamic spatial heterogeneity reveals interdependence of marine faunal density and fishery removals.

Beketov, M. A., Kefford, B. J., Schäfer, R. B., and Liess, M. (2013). Pesticides reduce regional biodiversity of stream invertebrates. *Proceedings of the National Academy of Sciences*, 110(27):11039–11043.

Bell, J. R., Alderson, L., Izera, D., Kruger, T., Parker, S., Pickup, J., Shortall, C. R., Taylor, M. S., Verrier, P., and Harrington, R. (2015). Longâterm phenological trends, species accumulation rates, aphid traits and climate: five decades of change in migrating aphids. *Journal of Animal Ecology*, 84(1):21–34. Publisher: Wiley.

Bellman, R. (1958). Dynamic programming and stochastic control processes. *Information and Control*, 1(3):228–239.

Bergstrom, D. M., Lucieer, A., Kiefer, K., Wasley, J., Belbin, L., Pedersen, T. K., and Chown, S. L. (2009). Indirect effects of invasive species removal devastate world heritage island. *Journal of Applied Ecology*, 46(1):73–81.

Betts, A. K., Ball, J. H., Beljaars, A. C. M., Miller, M. J., and Viterbo, P. A. (1996). The land surface-atmosphere interaction: A review based on observational and global modeling perspectives. *Journal of Geophysical Research: Atmospheres*, 101(D3):7209–7225.

Bialonski, S., Ansmann, G., and Kantz, H. (2015). Data-driven prediction and prevention of extreme events in a spatially extended excitable system. *Physical Review E*, 92(4).

Blum, M. and Riedmiller, M. A. (2013). Optimization of gaussian process hyperparameters using rprop. In *ESANN*.

Boettiger, C., Mangel, M., and Munch, S. (2015). Avoiding tipping points in fisheries management through gaussian process dynamic programming. *Proceedings of the Royal Society B: Biological Sciences*, 282(1801):20141631.

Brack, I. V., Kindel, A., and Oliveira, L. F. B. (2018). Detection errors in wildlife abundance estimates from Unmanned Aerial Systems (UAS) surveys: Synthesis, solutions, and challenges. *Methods in Ecology and Evolution*, 9(8):1864–1873.

Bradshaw, C. J. A., Leroy, B., Bellard, C., Roiz, D., Albert, C., Fournier, A., Barbet-Massin, M., Salles, J.-M., Simard, F., Courchamp, F., and et al. (2016). Massive yet grossly underestimated global costs of invasive insects. *Nature Communications*, 7(1):12986.

Brias, A. and Munch, S. B. (2021). Ecosystem based multi-species management using empirical dynamic programming. *Ecological Modelling*, 441:109423.

Browne, A., Jakary, A., Vinogradov, S., Fu, Y., and Deicken, R. (2008). Automatic relevance determination for identifying thalamic regions implicated in schizophrenia. *IEEE Transactions on Neural Networks*, 19(6):1101–1107.

Béné, C., Barange, M., Subasinghe, R., Pinstrup-Andersen, P., Merino, G., Hemre, G.-I., and Williams, M. (2015). Feeding 9 billion by 2050 – putting fish back on the menu. *Food Security*, 7(2):261–274.

Carney, R. M., Husted, S., Jean, C., Glaser, C., and Kramer, V. (2008). Efficacy of aerial spraying of mosquito adulticide in reducing incidence of west nile virus, california, 2005. *Emerging Infectious Diseases*, 14(5):747–754.

Chambers, U., Petit, B., and Jones, V. (2015). Wsu-das - the online pest management support system for tree fruits in washington state. *Acta Horticulturae*, (1068):27–33.

Chang, C.-W., Ushio, M., and Hsieh, C.-H. (2017). Empirical dynamic modeling for beginners. *Ecological Research*, 32(6):785–796.

Chattopadhyay, A., Hassanzadeh, P., and Pasha, S. (2020). Predicting clustered weather patterns: A test case for applications of convolutional neural networks to spatio-temporal climate data. *Scientific Reports*, 10(1).

Cheng, B. and Tong, H. (1992). On consistent nonparametric order determination and chaos. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(2):427–449.

Christensen, N. L., Bartuska, A. M., Brown, J. H., Carpenter, S., Dantonio, C., Francis, R., Franklin, J. F., Macmahon, J. A., Noss, R. F., Parsons, D. J., and et al. (1996). The report of the ecological society of america committee on the scientific basis for ecosystem management. *Ecological Applications*, 6(3):665–691.

Clark, A. T., Ye, H., Deyle, E. R., Cowles, J., Tilman, G. D., Sugihara, G., and Isbell, F. (2015). Spatial convergent cross mapping to detect causal relationships from short time series. *The Ecological Society of America.*

Clark, C. W. and Mangel, M. (2000). *Dynamic state variable models in ecology methods and applications.* Oxford University Press.

Clark, J. S. (2001). Ecological forecasts: An emerging imperative. *Science*, 293(5530):657–660.

Clark, T. J. and Luis, A. D. (2020). Nonlinear population dynamics are ubiquitous in animals. *Nature Ecology & Evolution*, 4(1):75–81.

Clarke, L., Walther, B., Munch, S., Thorrold, S., and Conover, D. (2009). Chemical signatures in the otoliths of a coastal marine fish, menidia menidia, from the northeastern united states: spatial and temporal differences. *Marine Ecology Progress Series*, 384:261–271.

Coifman, R. R. and Lafon, S. (2006). Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30.

Covas, E. (2017). Spatial-temporal forecasting the sunspot diagram. *Astronomy and Astrophysics*, 605.

Covas, E. and Benetos, E. (2019). Optimal neural network feature selection for spatial-temporal forecasting. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(6):063111.

Crutchfield, J. P. and Kaneko, K. (1987). Phenomenology of spatio-temporal chaos. *Series on Directions in Condensed Matter Physics Directions in Chaos — Volume 1*, page 272–353.

Dara, S. K. (2019). The new integrated pest management paradigm for the modern age. *Journal of Integrated Pest Management*, 10(1).

DeFries, R. and Nagendra, H. (2017). Ecosystem management as a wicked problem. *Science*, 356(6335):265–270.

Deguine, J.-P., Aubertot, J.-N., Flor, R. J., Lescourret, F., Wyckhuys, K. A., and Ratnadass, A. (2021). Integrated pest management: good intentions, hard realities. a review. *Agronomy for Sustainable Development*, 41(3).

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via theEMAlgorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22. Publisher: Wiley.

Deutsch, C. A., Tewksbury, J. J., Tigchelaar, M., Battisti, D. S., Merrill, S. C., Huey, R. B., and Naylor, R. L. (2018). Increase in crop losses to insect pests in a warming climate. *Science*, 361(6405):916–919.

Deyle, E., Schueller, A. M., Ye, H., Pao, G. M., and Sugihara, G. (2018). Ecosystem-based forecasts of recruitment in two menhaden species. *Fish and Fisheries*, 19(5):769–781.

Deyle, E. R., Fogarty, M., Hsieh, C.-H., Kaufman, L., Maccall, A. D., Munch, S. B., Perretti, C. T., Ye, H., and Sugihara, G. (2013). Predicting climate effects on pacific sardine. *Proceedings of the National Academy of Sciences*, 110(16):6430–6435.

Deyle, E. R., May, R. M., Munch, S. B., and Sugihara, G. (2016). Tracking and forecasting ecosystem interactions in real time. *Proceedings of the Royal Society B: Biological Sciences*, 283(1822):20152258.

Deyle, E. R. and Sugihara, G. (2011). Generalized theorems for nonlinear state space reconstruction. *PLoS ONE*, 6(3).

Dietze, M. C. (2017). *Ecological forecasting*. Princeton University Press.

Dietze, M. C., Fox, A., Beck-Johnson, L. M., Betancourt, J. L., Hooten, M. B., Jarnevich, C. S., Keitt, T. H., Kenney, M. A., Laney, C. M., Larsen, L. G., Loescher, H. W., Lunch, C. K., Pijanowski, B. C., Randerson, J. T., Read, E. K., Tredennick, A. T., Vargas, R., Weathers, K. C., and White, E. P. (2018).

Iterative near-term ecological forecasting: Needs, opportunities, and challenges. *Proceedings of the National Academy of Sciences*, 115(7):1424–1432.

Dixon, P. A., Milicich, M. J., and Sugihara, G. (1999). Episodic fluctuations in larval supply. *Science*, 283(5407):1528–1530.

Dyck, A. J. and Sumaila, U. R. (2010). Economic impact of ocean fish populations in the global fishery. *Journal of Bioeconomics*, 12(3):227–243.

Ellner, S. and Turchin, P. (1995). Chaos in a noisy world: new methods and evidence from time series analysis. *The American Naturalist*, 145(3):343–375.

EPAP (1999). *Ecosystem-based fishery management: a report to Congress by the Ecosystem Principles Advisory Panel.* U.S. Dept. of Commerce, National Oceanic and Atmospheric Administration, National Marine Fisheries Service.

Evans, M. R., Norris, K. J., and Benton, T. G. (2012). Predictive ecology: systems approaches. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1586):163–169.

FAO (2014). Faostat database collections. *Food and Agriculture Organization of the United Nations, Rome.*

Farmer, J. D. and Sidorowich, J. J. (1987). Predicting chaotic time series. *Physical Review Letters*, 59(8):845–848.

Fautin, D., Dalton, P., Incze, L. S., Leong, J.-A. C., Pautzke, C., Rosenberg, A., Sedberry, P. S. G., Jr, J. W. T., Abbott, I., Brainard, R. E., Brodeur, M., Eldredge, L. G., Feldman, M., Moretzsohn, F., Vroom, P. S., Wainstein, M., and Wolff, N. (2010). An overview of marine biodiversity in united states waters. *PLoS ONE*, 5(8):427–449.

Fei, S., Morin, R. S., Oswalt, C. M., and Liebhold, A. M. (2019). Biomass losses resulting from insect and disease invasions in us forests. *Proceedings of the National Academy of Sciences*, 116(35):17371–17376.

Field, J. C., Miller, R. R., Santora, J. A., Tolimieri, N., Haltuch, M. A., Brodeur, R. D., Auth, T. D., Dick, E. J., Monk, M. H., Sakuma, K. M., and Wells, B. K. (2021). Spatiotemporal patterns of variability in the abundance and distribution of winter-spawned pelagic juvenile rockfish in the California Current. *PLOS ONE*, 16(5):e0251638. Publisher: Public Library of Science (PLoS).

Fitri, I. R., Hanum, F., Kusnanto, A., and Bakhtiar, T. (2021). Optimal pest control strategies with cost-effectiveness analysis. *The Scientific World Journal*, 2021:1–17.

Garrett, K., Dobson, A., Kroschel, J., Natarajan, B., Orlandini, S., Tonnang, H., and Valdivia, C. (2013). The effects of climate variability and the color of weather time series on agricultural diseases and pests, and on decisions for their management. *Agricultural and Forest Meteorology*, 170:216–227.

Geary, W. L., Bode, M., Doherty, T. S., Fulton, E. A., Nimmo, D. G., Tulloch, A. I. T., Tulloch, V. J. D., and Ritchie, E. G. (2020). A guide to ecosystem models and their environmental applications. *Nature Ecology Evolution*, 4(11):1459–1471.

Geary, W. L., Nimmo, D. G., Doherty, T. S., Ritchie, E. G., and Tulloch, A. I. T. (2019). Threat webs: Reframing the co-occurrence and interactions of threats to biodiversity. *Journal of Applied Ecology*.

Giron-Nava, A., James, C., Johnson, A., Dannecker, D., Kolody, B., Lee, A., Nagarkar, M., Pao, G., Ye, H., Johns, D., and et al. (2017). Quantitative argument for long-term ecological monitoring. *Marine Ecology Progress Series*, 572:269–274.

Glaser, S. M., Fogarty, M. J., Liu, H., Altman, I., Hsieh, C.-H., Kaufman, L., Maccall, A. D., Rosenberg, A. A., Ye, H., Sugihara, G., and et al. (2014a). Complex dynamics may limit prediction in marine fisheries. *Fish and Fisheries*, 15(4):616–633.

Glaser, S. M., Ye, H., and Sugihara, G. (2014b). A nonlinear, low data requirement model for producing spatially explicit fishery forecasts. *Fisheries Oceanography*, 23(1):45–53.

Gokul, E. A., Raitsos, D. E., Gittings, J. A., Alkawri, A., and Hoteit, I. (2019). Remotely sensing harmful algal blooms in the Red Sea. *PLOS ONE*, 14(4):e0215463.

Grimes, D. J., Cortale, N., Baker, K., and Mcnamara, D. E. (2015). Nonlinear forecasting of intertidal shoreface evolution. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(10):103116.

Hackett, S. C. and Bonsall, M. B. (2019). Insect pest control, approximate dynamic programming, and the management of the evolution of resistance. *Ecological Applications*, 29(2).

Halbach, U. (1984). Population dynamics of rotifers and its consequences for ecotoxicology. *Hydrobiologia*, 109(1):79–96.

Hall, D. B. and Zhang, Z. (2004). Marginal models for zero inflated clustered data. *Statistical Modelling*, 4(3):161–180.

Hampton, S. E., Holmes, E. E., Scheef, L. P., Scheuerell, M. D., Katz, S. L., Pendleton, D. E., and Ward, E. J. (2013). Quantifying effects of abiotic and biotic drivers on community dynamics with multivariate autoregressive (MAR) models. *Ecology*, 94(12):2663–2669.

Harrington, R. and Woiwod, I. (2007). Foresight from hindsight: the Rothamsted Insect Survey. *Outlooks on Pest Management*, 18(1):9–14.

Hassell, M. P. (2000). Host-parasitoid population dynamics*. *Journal of Animal Ecology*, 69(4):543–566.

Hastings, A. and Powell, T. (1991). Chaos in a three-species food chain. *Ecology*, 72(3):896–903.

Herweijer, C., Evison, W., Mariam, S., Khatri, A., Albani, M., Semov, A., and Long, E. (2020). Nature risk rising: Why the crisis engulfing nature matters for business and the economy.

Hoagland, P. and Scatasta, S. (2006). *The Economic Effects of Harmful Algal Blooms*, page 391–402.

Hofmann, E. E. and Powell, T. M. (1998). Environmental variability effects on marine fisheries: Four case histories. *Ecological Applications*, 8(1).

Holcomb, K. M., Reiner, R. C., and Barker, C. M. (2021). Spatio-temporal impacts of aerial adulticide applications on populations of west nile virus vector mosquitoes. *Parasites & Vectors*, 14(1).

Horswill, C., Kindsvater, H. K., JuanâJordÃ¡, M. J., Dulvy, N. K., Mangel, M., and Matthiopoulos, J. (2019). Global reconstruction of lifeâhistory strategies: A case study using tunas. *Journal of Applied Ecology*, 56(4):855–865.

Hsieh, C., Anderson, C., and Sugihara, G. (2008). Extending nonlinear analysis to short ecological time series. *The American Naturalist*, 171(1):71–80.

Hsieh, C.-H., Glaser, S. M., Lucas, A. J., and Sugihara, G. (2005). Distinguishing random environmental fluctuations from ecological catastrophes for the north pacific ocean. *Nature*, 435(7040):336–340.

Hummel, P. M. (1946). The Accuracy of Linear Interpolation. *The American Mathematical Monthly*, 53(7):364–366.

Iatan, I. F. (2016). *Modern Neural Methods for Function Approximation*. Springer.

IPM, U. (2013). How to manage pests: Uc pest management guidelines.

Ives, A. R., Dennis, B., Cottingham, K. L., and Carpenter, S. R. (2003). Estimating community stability and ecological interactions from time-series data. *Ecological Monographs*, 73(2):301–330.

Jang, S. R.-J. and Yu, J.-L. (2012). Discrete-time host–parasitoid models with pest control. *Journal of Biological Dynamics*, 6(2):718–739.

Johnson, B., Gomez, M., and Munch, S. B. (2021). Leveraging spatial information to forecast nonlinear ecological dynamics. *Methods in Ecology and Evolution*, 12(2):266–279.

Johnson, B. and Munch, S. B. (2022). An empirical dynamic modeling framework for missing or irregular samples. *Ecological Modelling*, 468:109948.

Juanes, F. and Conover, D. (1995). Size-structured piscivory:advection and the linkage between predator and prey recruitment in young-of-the-year bluefish. *Marine Ecology Progress Series*, 128:287–304.

Judd, K. and Mees, A. (1998). Embedding as a modeling problem. *Physica D: Nonlinear Phenomena*, 120(3-4):273–286.

Judd, K. L. (1999). *Numerical methods in economics*. MIT Press.

Kaluskar, S., Blukacz-Richards, E. A., Johnson, C. A., Kim, D.-K., and Arhonditsis, G. (2020). Connecting the dots in databases of endangered species: a Bayesian hierarchical imputation strategy for missing Peary caribou (Rangifer tarandus pearyi) population data. *Ecological Complexity*, 43:100846.

Kantz, H. and Schreiber, T. (2004). *Nonlinear time series analysis.* Cambridge University Press.

Kar, T., Ghorai, A., and Jana, S. (2012). Dynamics of pest and its predator model with disease in the pest and optimal use of pesticide. *Journal of Theoretical Biology*, 310:187–198.

Kawatsu, K., Ushio, M., Veen, F. J. F., and Kondoh, M. (2021). Are networks of trophic interactions sufficient for understanding the dynamics of multiâtrophic communities? Analysis of a triâtrophic insect foodâweb timeâseries. *Ecology Letters*, 24(3):543–552.

Kogan, M. (1998). Integrated pest management: Historical perspectives and contemporary developments. *Annual Review of Entomology*, 43(1):243–270.

Kolasa, J., Pickett, S. T., and Allen, T. F. H. (1991). *Ecological heterogeneity.* Springer-Verlag.

Kot, M. and Schaffer, W. M. (1986). Discrete-time growth-dispersal models. *Mathematical Biosciences*, 80(1):109–136.

Kuramoto, Y. (1984). *Chemical oscillations, waves, and turbulence.* Springer.

Laan, E. and Fox, J. W. (2020). An experimental test of the effects of dispersal and the paradox of enrichment on metapopulation persistence. *Oikos*, 129(1):49–58.

Lagos-Ortiz, K., Medina-Moreira, J., Sinche-Guzmán, A., Garzón-Goya, M., Vergara-Lozano, V., and Valencia-García, R. (2018). *Mobile Applications for Crops Management*, page 57–69. Communications in Computer and Information Science.

Largier, J. L. (2003). Considerations in estimating larval dispersal distances from oceanographic data. *Ecological Applications*, 13(sp1):71–89.

Lekscha, J. and Donner, R. V. (2018). Phase space reconstruction for non-uniformly sampled noisy time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(8):085702.

Lewis, W. J., Van Lenteren, J. C., Phatak, S. C., and Tumlinson, J. H. (1997). A total system approach to sustainable pest management. *Proceedings of the National Academy of Sciences*, 94(23):12243–12248.

Little, R. J. A. (1992). Regression With Missing X's: A Review. *Journal of the American Statistical Association*, 87(420):1227.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data: Little/Statistical Analysis with Missing Data.* John Wiley & Sons, Inc., Hoboken, NJ, USA.

Liu, H., Fogarty, M., Glaser, S., Altman, I., Hsieh, C., Kaufman, L., Rosenberg, A., and Sugihara, G. (2012). Nonlinear dynamic features and co-predictability of the georges bank fish community. *Marine Ecology Progress Series*, 464:195–207.

Liu, X., Singh, P. V., and Srinivasan, K. (2016). A structured analysis of unstructured big data by leveraging cloud computing. *Marketing Science*, 35(3):363–388.

Lucey, S. and Nye, J. (2010). Shifting species assemblages in the northeast us continental shelf large marine ecosystem. *Marine Ecology Progress Series*, 415:23–33.

Lynch, D. R., Greenberg, D. A., Bilgili, A., McGillicuddy, D. J., Manning, J. P., and Aretxabaleta, A. L. (2014). *Particles in the coastal ocean: theory and applications*. Cambridge University Press.

Macaulay, E. D. M., Tatchell, G. M., and Taylor, L. R. (1988). The Rothamsted Insect Survey '12-metre' suction trap. *Bulletin of Entomological Research*, 78(1):121–128.

Mach, M. E., Sbrocco, E. J., Hice, L. A., Duffy, T. A., Conover, D. O., and Barber, P. H. (2011). Regional differentiation and post-glacial expansion of the atlantic silverside, menidia menidia, an annual fish with high dispersal potential. *Marine Biology*, 158(3):515–530.

MacKay, D. and Neal, R. (1994). Automatic relevance determination for neural networks. Technical report, Cambridge University.

Marquet, P. A., Allen, A. P., Brown, J. H., Dunne, J. A., Enquist, B. J., Gillooly, J. F., Gowaty, P. A., Green, J. L., Harte, J., Hubbell, S. P., and et al. (2014). On theory in ecology. *BioScience*, 64(8):701–710.

Marwala, T. (2015). *Economic modeling using artificial intelligence methods*. Springer.

Maunder, M. N. and Piner, K. R. (2015). Contemporary fisheries stock assessment: many issues still remain. *ICES Journal of Marine Science*, 72(1):7–18.

May, R. M. (1976). Simple mathematical models with very complicated dynamics. *Nature*, 261(5560):459–467.

Mcevoy, P. B. (2017). Theoretical contributions to biological control success. *BioControl*, 63(1):87–103.

McGarvey, R., Burch, P., and Matthews, J. M. (2016). Precision of systematic and random sampling in clustered populations: habitat patches and aggregating organisms. *Ecological Applications*, 26(1):233–248.

Meisner, M. H., Rosenheim, J. A., and Tagkopoulos, I. (2016). A data-driven, machine learning framework for optimal pest management in cotton. *Ecosphere*, 7(3):e01263.

Morari, M. and H. Lee, J. (1999). Model predictive control: past, present and future. *Computers Chemical Engineering*, 23(4-5):667–682.

Munch, S. B., Brias, A., Sugihara, G., and Rogers, T. L. (2020). Frequently asked questions about nonlinear dynamics and empirical dynamic modelling. *ICES Journal of Marine Science*, 77(4):1463–1479.

Munch, S. B., Giron-Nava, A., and Sugihara, G. (2018). Nonlinear dynamics and noise in fisheries recruitment: A global meta-analysis. *Fish and Fisheries*, 19(6):964–973.

Munch, S. B., Kottas, A., and Mangel, M. (2005). Bayesian nonparametric analysis of stockÂrecruitment relationships. *Canadian Journal of Fisheries and Aquatic Sciences*, 62(8):1808–1821. Publisher: Canadian Science Publishing.

Munch, S. B., Poynor, V., and Arriaza, J. L. (2017). Circumventing structural uncertainty: A bayesian perspective on nonlinear forecasting for ecology. *Ecological Complexity*, 32:134–143.

Namias, A., Jobe, N. B., Paaijmans, K. P., and Huijben, S. (2021). The need for practical insecticide-resistance guidelines to effectively inform mosquito-borne disease control programs. *eLife*, 10.

Nathan, R., Klein, E., Robledo-Arnuncio, J. J., and Revilla, E. (2012). *Dispersal kernels: review*. Oxford University Press.

Naves, P. and De Sousa, E. (2009). Threshold temperatures and degree-day estimates for development of post-dormancy larvae of monochamus galloprovincialis (coleoptera: Cerambycidae). *Journal of Pest Science*, 82(1):1–6.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer-Verlag New York.

Ness-Cohn, E. and Braun, R. (2020). TimeCycle: Topology Inspired MEthod for the Detection of Cycling Transcripts in Circadian Time-Series Data. preprint, Bioinformatics.

Neubert, M., Kot, M., and Lewis, M. (1995). Dispersal and pattern formation in a discrete-time predator-prey model. *Theoretical Population Biology*, 48(1):7–43.

Nye, J., Link, J., Hare, J., and Overholtz, W. (2009). Changing spatial distribution of fish stocks in relation to climate and population size on the northeast united states continental shelf. *Marine Ecology Progress Series*, 393:111–129.

Okuno, S., Aihara, K., and Hirata, Y. (2020). Forecasting high-dimensional dynamics exploiting suboptimal embeddings. *Scientific Reports*, 10(1).

Oliver, T. H. and Roy, D. B. (2015). The pitfalls of ecological forecasting. *Biological Journal of the Linnean Society*, 115(3):767–778.

Parlitz, U. and Merkwirth, C. (2000). Prediction of spatiotemporal time series based on reconstructed local states. *Physical Review Letters*, 84(9):1890–1893.

Parsa, S., Morse, S., Bonifacio, A., Chancellor, T. C. B., Condori, B., Crespo-Pérez, V., Hobbs, S. L. A., Kroschel, J., Ba, M. N., Rebaudo, F., and et al. (2014). Obstacles to integrated pest management adoption in developing countries. *Proceedings of the National Academy of Sciences*, 111(10):3889–3894.

Pau, S., Wolkovich, E. M., Cook, B. I., Davies, T. J., Kraft, N. J. B., Bolmgren, K., Betancourt, J. L., and Cleland, E. E. (2011). Predicting phenology by integrating ecology, evolution and climate science. *Global Change Biology*, 17(12):3633–3643.

Pegoraro, L., Hidalgo, O., Leitch, I. J., Pellicer, J., and Barlow, S. E. (2020). Automated video monitoring of insect pollinators in the field. *Emerging Topics in Life Sciences*, 4(1):87–97.

Perretti, C. T., Munch, S. B., and Sugihara, G. (2013). Model-free forecasting outperforms the correct mechanistic model for simulated and experimental data. *Proceedings of the National Academy of Sciences*, 110(13):5253–5257.

Pikitch, E. K. (2004). Ecology: Ecosystem-based fishery management. *Science*, 305(5682):346–347.

Pinto, S. E. D. S. and Viana, R. L. (2000). Synchronization plateaus in a lattice of coupled sine-circle maps. *Physical Review E*, 61(5):5154–5161.

Politis, P. J., Galbraith, J. K., Kostovick, P., and Brown, R. W. (2014). *Northeast Fisheries Science Center bottom trawl survey protocols for the NOAA Ship Henry B. Bigelow.* US Dept Commer, Northeast Fish Sci Cent Ref Doc. 14-06; 138 p. Available from: National Marine Fisheries Service, 166 Water Street, Woods Hole, MA 02543-1026.

Poon, S.-H. and Granger, C. W. (2003). Forecasting volatility in financial markets: A review. *Journal of Economic Literature*, 41(2):478–539.

Powell, W. B. (2011). *Approximate dynamic programming: Solving the curses of dimensionality*. Wiley.

Poynor, V. and Munch, S. (2017). Combining functional data with hierarchical Gaussian process models. *Environmental and Ecological Statistics*, 24(2):175–199.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian process for machine learning*. The MIT Press.

Ricker, W. E. (1954). Stock and recruitment. *Journal of the Fisheries Research Board of Canada*, 11(5):559–623.

Riedmiller, M. and Braun, H. (1993). A direct adaptive method for faster backpropagation learning: the rprop algorithm. *IEEE International Conference on Neural Networks*.

Rogers, T. L., Johnson, B. J., and Munch, S. B. (2022). Chaos is not rare in natural ecosystems. *Nature Ecology Evolution*, 6(8):1105–1111.

Rogers, T. L. and Munch, S. B. (2020). Hidden similarities in the dynamics of a weakly synchronous marine metapopulation. *PNAS*, 117(1):479–485.

Rogers, T. L., Munch, S. B., Stewart, S. D., Palkovacs, E. P., Giron-Nava, A., Matsuzaki, S. S., and Symons, C. C. (2020). Trophic control changes with season and nutrient loading in lakes. *Ecology Letters*.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rydhmer, K., Bick, E., Still, L., Strand, A., Luciano, R., Helmreich, S., Beck, B. D., Grønne, C., Malmros, L., Poulsen, K., and et al. (2022). Automating insect monitoring using unsupervised near-infrared sensors. *Scientific Reports*, 12(1).

Sacks, J. and Ylvisaker, D. (1966). Designs for regression problems with correlated errors. *The Annals of Mathematical Statistics*, 37(1):66–89.

Sharma, S., Kooner, R., and Arora, R. (2017). Insect pests and crop losses. *Breeding Insect Resistant Crops for Sustainable Agriculture*, page 45–66.

Shortall, C. R., Moore, A., Smith, E., Hall, M. J., Woiwod, I. P., and Harrington, R. (2009). Long-term changes in the abundance of flying insects. *Insect Conservation and Diversity*, 2(4):251–260.

Skern-Mauritzen, M., Ottersen, G., Handegard, N. O., Huse, G., Dingsør, G. E., Stenseth, N. C., and Kjesbu, O. S. (2016). Ecosystem processes are rarely included in tactical fisheries management. *Fish and Fisheries*, 17(1):165–175.

Stalph, P. (2014). *Introduction to Function Approximation and Regression*. Springer Fachmedien Wiesbaden GmbH.

Stark, J. (1999). Delay embeddings for forced systems. i. deterministic forcing. *Journal of Nonlinear Science*, 9(3):255–332.

Stark, J., Broomhead, D., Davies, M., and Huke, J. (1997). Takens embedding theorems for forced and stochastic systems. *Nonlinear Analysis Theory Methods Applications*, 30(8):5303–5314.

Stark, J., Broomhead, D., Davies, M., and Huke, J. (2003). Delay embeddings for forced systems. ii. stochastic forcing. *Journal of Nonlinear Science*, 13(6):519–577.

Stenberg, J. A. (2017). A conceptual framework for integrated pest management. *Trends in Plant Science*, 22(9):759–769.

Stern, V., Smith, R., van den Bosch, R., Hagen, K., et al. (1959). The integration of chemical and biological control of the spotted alfalfa aphid: the integrated control concept. *Hilgardia*, 29(2):81–101.

Stern, V. M. (1973). Economic thresholds. *Annual Review of Entomology*, 18(1):259–280.

Strogatz, S. H. (1994). *Nonlinear dynamics and Chaos: with applications to physics, biology, chemistry, and engineering*. Studies in nonlinearity. Addison-Wesley Pub, Reading, Mass.

Sugihara, G. (1994). Nonlinear forecasting for the classification of natural time series. *Philosophical Transactions of the Royal Society of London: Mathematical, Physical and Engineering Sciences*, 348.

Sugihara, G., Casdagli, M., Habjan, E., Hess, D., Dixon, P., and Holland, G. (1999). Residual delay maps unveil global patterns of atmospheric nonlinearity and produce improved local forecasts. *Proceedings of the National Academy of Sciences*, 96(25):14210–14215.

Sugihara, G., Garcia, S., Platt, T., Gulland, J. A., Rachor, E., Lawton, J. H., Rothschild, B. J., Maske, H., Ursin, E. A., Paine, R. T., and et al. (1984). Ecosystems dynamics. *Exploitation of Marine Communities*, page 131–153.

Sugihara, G., Grenfell, B., May, R. M., Chesson, P., Platt, H. M., and Williamson, M. (1990). Distinguishing error from chaos in ecological time series [and discussion]. *Philosophical Transactions of The Royal Society B Biological Sciences*, 330(1257):235–251.

Sugihara, G., May, R., Ye, H., Hsieh, C.-H., Deyle, E., Fogarty, M., and Munch, S. (2012). Detecting causality in complex ecosystems. *Science*, 338(6106):496–500.

Sugihara, G. and May, R. M. (1990). Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, 344(6268):734741.

Sutton, Richard S., a. (2018). *Reinforcement Learning: An Introduction, 2nd Edition*. Adaptive Computation and Machine Learning. The MIT Press,, Cambridge, Massachusetts ; London, England :, second edition. edition.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44.

Takens, F. (1981). Detecting strange attractors in turbulence. *Lecture Notes in Mathematics Dynamical Systems and Turbulence, Warwick 1980*, pages 366–381.

Thorson, J. T., Ono, K., and Munch, S. B. (2014). A bayesian approach to identifying and compensating for model misspecification in population models. *Ecology*, 95(2):329–341.

Tommasi, D., Hunt, B. P. V., Allen, S. E., Routledge, R., and Pakhomov, E. A. (2014). Variability in the vertical distribution and advective transport of eight mesozooplankton taxa in spring in rivers inlet, british columbia, canada. *Journal of Plankton Research*, 36(3):743–756.

Tonnang, H. E., Hervé, B. D., Biber-Freudenberger, L., Salifu, D., Subramanian, S., Ngowi, V. B., Guimapi, R. Y., Anani, B., Kakmeni, F. M., Affognon, H., and et al. (2017). Advances in crop insect modelling methods—towards a whole system approach. *Ecological Modelling*, 354:88–103.

USDA, E. (2021). Ag and food sectors and the economy.

Ushio, M., Hsieh, C.-h., Masuda, R., Deyle, E. R., Ye, H., Chang, C.-W., Sugihara, G., and Kondoh, M. (2018). Fluctuating interaction network and time-varying stability of a natural fish community. *Nature*, 554(7692):360–363.

van der Werf, H. M. (1996). Assessing the impact of pesticides on the environment. *Agriculture, Ecosystems Environment*, 60(2):81–96.

van Lierop, P., Lindquist, E., Sathyapala, S., and Franceschini, G. (2015). Global forest area disturbance from fire, insect pests, diseases and severe weather events. *Forest Ecology and Management*, 352:78–88. Changes in Global Forest Resources from 1990 to 2015.

Van Nes, E. H., Scheffer, M., Brovkin, V., Lenton, T. M., Ye, H., Deyle, E., and Sugihara, G. (2015). Causal feedbacks in climate change. *Nature Climate Change*, 5(5):445–448.

Vasconcelos, D., Viana, R., Lopes, S., Batista, A., and Pinto, S. D. S. (2004). Spatial correlations and synchronization in coupled map lattices with long-range interactions. *Physica A: Statistical Mechanics and its Applications*, 343:201–218.

Walsh, H. J., Richardson, D. E., Marancik, K. E., and Hare, J. A. (2015). Long-term changes in the distributions of larval and adult fish in the northeast u.s. shelf ecosystem. *Plos One*, 10(9).

Wang, J., Fleet, D., and Hertzmann, A. (2008). Gaussian Process Dynamical Models for Human Motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):283–298.

Whittle, A., Lenhart, S., and White, K. A. J. (2008). Optimal control of gypsy moth populations. *Bulletin of Mathematical Biology*, 70(2):398–411.

WHO (2020). Vector-borne diseases.

Williams, P. J. and Kendall, W. L. (2017). A guide to multi-objective optimization for ecological problems with an application to cackling goose management. *Ecological Modelling*, 343:54–67.

Wolf, A., Swift, J. B., Swinney, H. L., and Vastano, J. A. (1985). Determining Lyapunov exponents from a time series. *Physica D: Nonlinear Phenomena*, 16(3):285–317.

Wood, S. N. and Thomas, M. B. (1999). Super sensitivity to structure in biological models. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 266(1419):565–570.

Ye, H., Beamish, R. J., Glaser, S. M., Grant, S. C. H., Hsieh, C.-H., Richards, L. J., Schnute, J. T., and Sugihara, G. (2015a). Equation-free mechanistic ecosystem forecasting using empirical dynamic modeling. *Proceedings of the National Academy of Sciences*, 112(13).

Ye, H., Deyle, E. R., Gilarranz, L. J., and Sugihara, G. (2015b). Distinguishing time-delayed causal interactions using convergent cross mapping. *Scientific Reports*, 5(1):14750.

Ye, H. and Sugihara, G. (2016). Information leverage in interconnected ecosystems: Overcoming the curse of dimensionality. *Science*, 353(6302):922–925.

Yokomizo, H., Possingham, H. P., Thomas, M. B., and Buckley, Y. M. (2009). Managing the impact of invasive species: The value of knowing the density–impact curve. *Ecological Applications*, 19(2):376–386.

Yu, H., Zhao, M., Lv, S., and Zhu, L. (2009). Dynamic complexities in a parasitoid-host-parasitoid ecological model. *Chaos, Solitons Fractals*, 39(1):39–48.

Zangina, U., Buyamin, S., Aman, M. N., Abidin, M. S. Z., and Mahmud, M. S. A. (2021). A greedy approach to improve pesticide application for precision agriculture using model predictive control. *Computers and Electronics in Agriculture*, 182:105984.

Zavaleta, E. S., Hobbs, R. J., and Mooney, H. A. (2001). Viewing invasive species removal in a whole-ecosystem context. *Trends in Ecology Evolution*, 16(8):454–459.

Zhang, Ma, Huang, Xuebing, Gao, Zichun, Zhang, Feifan, and Cong (2018). Complex dynamics on the routes to chaos in a discrete predator-prey system with crowley-martin type functional response.

Ørstavik, S. and Stark, J. (1998). Reconstruction and cross-prediction in coupled map lattices using spatio-temporal embedding techniques. *Physics Letters A*, 247(1-2):145–160.