Dissertation

The Mathematical Modeling and Analysis of Nonlocal Ecological Invasions and Savanna Population Dynamics

Submitted by

William Christopher Strickland

Department of Mathematics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2013

Doctoral Committee:

Advisor: Gerhard Dangelmayr
Co-Advisor: Patrick Shipman

Yongcheng Zhou
Cynthia Brown

Abstract

The Mathematical Modeling and Analysis of Nonlocal Ecological Invasions and Savanna Population Dynamics

The main focus of this dissertation is the development and analysis of two new mathematical models that individually address major open problems in ecology. The first challenge is to characterize and model the processes that result in a savanna ecosystem as a stable state between grassland and forest, and the second involves modeling the non-local spread of a biological invader over heterogeneous terrain while incorporating the influence of a mass transportation network on the system. Both models utilize and compare work done in other, often more opaque, modeling paradigms to better develop transparent and application-ready solutions which can be easily adapted and inform ecological work done in the field.

Savanna is defined by the coexistence of trees and grass in seasonally dry areas of the tropics and sub-tropics, but there is no consensus as to why savanna occurs as a stable state between tropical grassland and forest. To understand the dynamics behind the tree-grass relationship, we begin by reviewing and analyzing approaches in currently available savanna models. Next, we develop a mathematical model for savanna water resource dynamics based on FLAMES, an Australian process-based software model created to capture the effects of seasonal rainfall and fire disturbance on savanna tree stands. As a mathematically explicit dynamical system represented by coupled differential equations, the new model immediately has the advantage of being concise and transparent compared to previous models, yet still robust in its ability to account for different climate and soil characteristics. Through analytical analysis of the model, we show a clear connection between climate and stand structure,

with particular emphasis on the length and severity of the dry season. As a result, we can numerically quantify the parameter space of year-by-year stochastic variability in stand structure based on rainfall and fire probabilities. This results in a characterization of savanna existence in the absence of extreme fire suppression based on the availability of water resources in the soil due to climate and ground water retention. One example of the model's success is its ability to predict a savanna environment for Darwin, Australia and a forest environment for Sydney, even though Sydney receives less annual rainfall than Darwin.

The majority of this dissertation focuses on modeling the spread of a biological invader in heterogeneous domains, where invasion often takes place non-locally, through nearby human transportation networks. Since early detection and ecological forecasting of invasive species is urgently needed for rapid response, accurately modeling invasions remains a high priority for resource managers. To achieve this goal, we begin by revisiting a particular class of deterministic contact models obtained from a stochastic birth process for invasive organisms. We then derive a deterministic integro-differential equation of a more general contact model and show that the quantity of interest may be interpreted not as population size, but rather as the probability of species occurrence. We then proceed to show how landscape heterogeneity can be included in the model by utilizing the concept of statistical habitat suitability models which condense diverse ecological data into a single statistic. Next, we develop a model for vector-based epidemic transport on a network as represented by a strongly connected, directed graph, and analytically compute the exact optimal control for suppression of the infected graph vectors. Since this model does not require any special assumptions about the underlying spatiotemporal epidemic spread process, it should prove suitable in a variety of application contexts where network based disease vector dynamics

need to be understood and properly controlled. We then discuss other methods of control for the special case of the integro-differential model developed previously and explore numerical results of applying this control. Finally, we validate model results for the *Bromus tectorum* invasion of Rocky Mountain National Park using data collected by ecologists over the past two decades, and illustrate the effect of various controls on this data.

A final chapter concerns a problem of cognitive population dynamics, namely vowel pronunciation in natural languages. We begin by developing a structured population approach to modeling changes in vowel systems, taking into account learning patterns and effects such as social trends. Our model treats vowel pronunciation as a continuous variable in vowel space and allows for continuous dependence of vowel pronunciation on time and age of the speaker. The theory of mixtures with continuous diversity provides a framework for the model, which extends the McKendrick-von Foerster equation to populations with age and phonetic structures. Numerical integrations of the model reveal how shifts in vowel pronunciation may occur in jumps or continuously given perturbations such as the influx of an immigrant population.

Acknowledgements

I would like to begin by thanking my advisers, Gerhard Dangelmayr and Patrick Shipman for their constant support and collaboration on the many projects represented in this dissertation. In particular, a large portion of the credit for Chapter 5 must go to Patrick Shipman and Sergio Faria, who were gracious enough to include me on their project after many years of collaboration. I would like to thank Tom Stohlgren and Sunil Kumar of the Colorado State University Natural Resource Ecology Laboratory for the discussions and suggestions that led to modeling the nonlocal spread of invasive species, and for providing the data on *Bromus tectorum* in Rocky Mountain National Park. Finally, I would also like to thank Adam Liedloff and Dick Williams of the CSIRO lab in Darwin, Australia for their collaboration on the savanna model, and for hosting me at the lab during the summer of 2011.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Systems modeling is an essential part of managing ecosystems and understanding large-scale population dynamics under the various forms of human and global stress. Modeling approaches are extremely varied, but in general, they can be understood as taking either a process-based approach, or a mathematical and/or statistical approach built around conservation laws or basic principles.

Modern process-based models exist primarily on computers as programs that seek to set up a virtual environment in which events occur as they do in nature, or at least in which favored processes are faithfully carried out similar to reality. Process-based models are common because they are basically a coded version of a scientist's understanding of the system in question. It is not difficult to code a process as it exists in theory, and by having that process act upon virtual organisms and environments, the computer provides a virtual model in which one can test hypotheses. However, because the system being modeled is typically complex, and opinions may vary widely as to what processes are critical for a faithful virtual representation, these programs usually grow quite large and contain many relationships between their variables. As a result, they often become black boxes and completely intractable to analyze directly. The result is that while process-based models provide some sort of structure for conducting experiments not possible in the field, it is nearly impossible to understand to what extent the virtual environment represents reality or how the underlying processes and parameters affect the system as a whole.

Mathematical and statistical approaches, on the other hand, are usually derived from clear and fundamental relations that can be written down on paper. They seek to be mathematically tractable to analyze while simultaneously capturing the critical behavior of the system as a whole. These models are not black boxes and can generally be re-implemented by a third party from scratch. On the other hand, they are not intuitively derived, understood, or implemented by anyone unfamiliar with advanced mathematics or statistics because they are written in the language of dynamical systems and differential equations. They do not seek to mimic all the details of organisms as they exist in the wild, and thus can be challenging to explain and justify to a scientist focused on processes on or below the scale of an individual or group.

In this dissertation, I will develop and analyze two new mathematical models for ecological processes, and a third model which extends the McKendrick von Foerster [1] equation with dynamics for the evolution of vowels sound systems. Both of the ecological models were inspired by existing process-based models developed by ecologists to test theoretical assumptions, and they are motivated by open ecological problems.

Chapter 2 is motivated by a history of models written primarily by ecologists to address questions about the driving relationships behind the savanna ecosystem. Savanna is defined by the coexistence of trees and grass in seasonally dry areas of the tropics and sub-tropics, but there is no consensus as to why this coexistence occurs. Since trees and grasses both use similar resources such as water, sunlight, and soil, it has been an open question for some time as to why savanna landscapes do not become either tropical grassland or forest, depending on which species has the competitive edge in the given environment.

To understand the dynamics behind the tree-grass relationship, we begin by reviewing and analyzing approaches in currently available savanna models. Next, we develop a mathematical model for savanna water resource dynamics based on FLAMES, an Australian process-based software model created to capture the effects of seasonal rainfall and fire disturbance on savanna tree stands. As a mathematically explicit dynamical system represented by coupled differential equations, this new model immediately has the advantage of being concise and transparent compared to previous models, yet still robust in its ability to account for different climate and soil characteristics.

Through analysis of this model, we show a clear connection between climate and stand structure, with particular emphasis on the length and severity of the dry season as a function of stand biomass. Using this result, it is theoretically possible for ecologists to numerically quantify year-by-year stochastic variability in a given stand structure based on rainfall history and fire probabilities. In the absence of extreme fire suppression, this results in a characterization of savanna stand structure as a function of seasonal water resource availability and ground water retention. Long-term dynamics of the model reveal an oscillating steady state for woody biomass and stand demographics based on fire disturbance and seedling survival events. One example of the model's success is its ability to predict a savanna environment for Darwin, Australia and a forest environment for Sydney, even though Sydney receives less annual rainfall than Darwin.

The majority of this dissertation, and the subject of Chapters 3 and 4, focuses on modeling the spread of a biological invader in heterogeneous domains. Invasion is assumed to take place non-locally, perhaps through the action of seeds spreading on the wind, and can reach remote locations utilizing nearby human transportation networks. Since early detection and

ecological forecasting of invasive species is urgently needed for rapid response, accurately modeling invasions remains a high priority for resource managers. Data is severely limited, however, due to a lack of resources and the spatial scales involved in an invasion. Population density information is unavailable, species presence data is sparsely collected, clustered, and incomplete, and even key functional information about the species, such as species growth rates and spread distributions, are not well understood and can be highly dependent on environmental conditions [2, 3, 4, 5].

To provide as much temporal and spatial information as possible given the limited data, we begin by revisiting a particular class of deterministic contact models obtained from a stochastic birth process for invasive organisms. We then derive a deterministic integro-differential equation of a more general contact model and show that the quantity of interest may be interpreted not as population size, but rather as the probability of species occurrence. We then proceed to show how landscape heterogeneity can be included in the model by utilizing statistical habitat suitability models which condense diverse ecological data into a single statistic [3, 6, 7].

Next, we develop a model for vector-based epidemic transport on a network as represented by a strongly connected, directed graph, and analytically compute the exact optimal control for suppression of the infected graph vectors. Since this model does not require any special assumptions about the underlying spatiotemporal epidemic spread process, it should prove suitable in a variety of application contexts where network based disease vector dynamics need to be understood and properly controlled. We then discuss other methods of control for the special case of the integro-differential model developed previously and explore numerical results of applying this control. Finally, we validate model results for the *Bromus tectorum*

invasion of Rocky Mountain National Park using data collected by ecologists over the past two decades [2, 3], and illustrate the effect of various controls on this data.

The final chapter, Chapter 5, concerns a problem of cognitive population dynamics, namely vowel pronunciation in natural languages. We begin by developing a structured population approach to modeling changes in vowel systems, taking into account learning patterns and effects such as social trends. Our model treats vowel pronunciation as a continuous variable in vowel space and allows for continuous dependence of vowel pronunciation on time and age of the speaker. The theory of mixtures with continuous diversity provides a framework for the model, which extends the McKendrick-von Foerster equation to populations with age and phonetic structures. Numerical integrations of the model reveal how shifts in vowel pronunciation may occur in jumps or continuously given perturbations such as the influx of an immigrant population [8].

CHAPTER 2

Savanna Woody Population Dynamics

2.1. Introduction

Modeling has become an essential part of understanding ecosystem dynamics, and within the savanna ecology community, models are used as a key tool to advance theories about the determinants of savanna as an ecological state between forest and grassland. The debate hinges around the fact that while savannas are defined by a co-existence between grasses and trees, there is still no consensus on the details concerning how this co-existence is maintained. The most active point of contention is between resource-based and disturbance-based theories, the first claiming that savanna exists because of a resource niche separation and/or competition with grasses while the second claims that fire suppresses woody growth in regions that would otherwise be forest, preventing closure of the canopy and maintaining a savanna state [9]. In more recent years, a middle ground has also formed, characterizing savanna as either arid or moist, the first of which is maintained through resource competition while the second is maintained through fire disturbance [10].

While models have often been a key component in arguing for a favored theory, less attention has been paid to the modeling approach itself. The result is that these models often take the form of large, complex, and opaque simulation models that cannot be readily analyzed for driving relationships and parameters, and are thus difficult to properly critique or evaluate (e.g. [11, 12, 13]). These process-based models, often referred to as Dynamic Global Vegetation models (DGVMs), are designed to simulate ecosystem composition by mimicking the author's understanding of the relevant ecological processes and how they fit together. They are difficult to understand in detail, and have often been used to advance

a favored theory or as black boxes to examine the effect of experimental conditions on the ecosystem [14, 15].

In this paper, we begin with a literature review of savanna models, theories, and field-based observations. Particular attention is paid to modeling methods and underlying assumptions, and how these have evolved over time. We then proceed to introduce a simplified model based on FLAMES, a process-based, demographic model developed by Liedloff and Cook (2007) [11] for Australian savannas. This resulting, simplified model remains tractable to mathematical analysis while still capturing key soil water resource dynamics and allowing for sophisticated fire disturbance mechanics to be explored. Our goal is to present a clearly formulated model which can easily be extended or further analyzed for theoretical implications while challenging some of the commonly held assumptions in the literature. We present an analysis of the model, including relevant parameter space approximations, and then conclude with a discussion of long-term predictions for savanna dynamics.

2.2. Review

Savannas are a biome characterized by a continuous grass layer with scattered trees. They persist in locations with strong wet and dry seasons, and cover approximately a fifth of the global land surface. Savannas also contain most of the world's rangeland, livestock, and wild herbivore biomass, and thus understanding the balance between trees and grasses in savanna locations is of critical importance for plant and livestock production, as well as maintaining ecosystem function, in the face of future changes in climate and land use [16, 17]. Despite this fact, and decades of research into the problem, there is still no consensus on the dynamics driving grass-tree coexistence in these ecosystems, though many of the factors influencing the relationship are well known [16, 18, 9].

The earliest models for grass-tree coexistence in savannas focused on competition for water resources utilizing the Lotka-Volterra equations. Since trees and grasses often have different rooting depths, competition and coexistence in these equations is defined by a niche differentiation, where trees have sole access to water in the subsoil and compete with grasses in the topsoil [19]. The differential equation model formulated by Walker and Noy-Meir (1982) [20] stands out in the literature as the prototypical articulation of this paradigm, often referred to as the "Walter hypothesis" [21, 16, 19]. Mathematical analysis of the model suggests that niche differentiation and competition for water resources results in stable grass-tree coexistence, though the extent to which the model is an accurate representation of savanna dynamics has often been questioned [20, 16, 19].

While it is not our intention to exhaustively review the factors involved in savanna composition and all of the different modeling approaches (see Scholes & Archer 1997 [16] for a comprehensive review of the subject), the Walker/Noy-Meir model deserves special attention as an analytical model focusing on the most critical driver of savanna structure: water availability [17, 18]. While there is plenty of evidence that this is not the only mechanism behind grass-tree coexistence, water availability has been shown to be the most important predictor of woody cover in African savannas [18], and rainfall seasonality remains a defining characteristic of all locations supporting the savanna biome. As a result, any model for savanna grass-tree coexistence, even those primarily focused on disturbance, must begin with a foundation in water resource dynamics. As a rooting niche differentiation model, the Walker/Noy-Meir model assumes that trees have access to water that is unavailable to grasses. This assertion may not always be valid in the form of a root niche differentiation [19], but may still hold if trees store water internally or otherwise have negligible (< 0.0001)

mortality rates when unable to access water for short periods. As such, a niche separation approach to water resource consumption continues to be the obvious place to start any attempt to model savanna composition dynamics.

The Walker/Noy-Meir analytical model includes a mixture of continuous and discrete-time dynamics. It is assumed that soil moisture is divided into a topsoil layer accessible to roots of trees and grasses, and a subsoil later accessible only to tree roots. The water is completely replenished during the rainy season and then used before the start of the next rainy season so that each year forms one time step in the discrete side of the model. During a year, dry season dynamics are described by the equations

$$\frac{dT}{dt} = -\theta_g G - \theta_w W - l \tag{1}$$

$$\frac{dS}{dt} = -\sigma_w W$$

where T is the amount of water in the topsoil, S is the amount of water in the subsoil, G represents grass biomass, W represents woody leaf biomass, and θ_g, θ_w ,l and σ_w are constants (θ_g, θ_w, and σ_w are grass and woody water usage rates and l is an evaporation loss term). It is assumed that initial values for available water in the topsoil and subsoil, T_0 and S_0 respectively, are given.

Each season, grass biomass and woody leaf biomass are updated via the equations

$$\Delta G = \epsilon_g \theta_g G t_T - M_g G \tag{2}$$

$$\Delta W = \epsilon_w(\theta_w W t_T + \sigma_w W t_S) - M_w W$$

where ϵ_g, ϵ_w and σ_w are growth rates, M_g and M_w are per season death rates, and t_T and t_S represent the amount of time that the topsoil and subsoil, respectively, contain water during a season (uniquely obtained by solving each of the equations in (1)). Walker and Noy-Meir have explored the parameter space of these equations along with certain modifications that have the effect of relaxing several basic assumptions. They find equilibrium solutions that often feature grass-tree coexistence, with varying ratios of biomass depending on rainfall and soil composition [20].

This model has several broad assumptions that are worth examining a bit further.

(1) In general, all available water in the root zone is used before the start of the next rainy season.

(2) Annual rainfall is constant.

(3) The length of the dry season is constant, and no rainfall occurs during the dry season.

(4) Annual loss of grass and woody leaf biomass is directly proportional to the amount of grass and woody leaf biomass respectively.

The first two are explicitly pointed out by Walker and Noy-Meir [20] and the last two are implicit in the construction of the model. The first assumption is very questionable for savanna structures that are not in equilibrium, and particularly in the presence of fire or grazing which reduce woody biomass below potential amounts, this will often be the case. Considering the second and third assumptions, it appears to be an open question as to how big a role stochasticity plays in savanna water resource dynamics, and this is a question

we will address in subsequent sections. Rare but severe droughts may have the potential to cause significant changes in woody composition, while occasional rains during the dry season could cause a significant change in expected woody biomass. The final assumption may be true to the extent that water resource consumption is also assumed to be directly proportional to the amount of grass and woody leaf biomass and climatic conditions are assumed to be constant, but the relationship between mortality and resource availability is not made clear in the Walker/Noy-Meir model.

It should be noted that Walker and Noy-Meir also extend their model computationally by adding a number of mechanics including sophisticated water infiltration dynamics, dependence of plant mortality on time and soil moisture, stochastic variation of mean annual rainfall, and loss of plant biomass due to grazing [20]. As a software based model, this formulation is much less explicit and difficult to review analytically. As a result, we refer the reader directly to their paper for further discussion and all computational results.

Process-based savanna models began to dominate in the literature from the year 2000 to 2009. Instead of taking a reductionist approach, these models seek to simulate a wide range of ecological processes which are then connected to yield a computational result. They are generally impossible to analyze in detail due to their complexity, and few attempts are made to understand the relative importance of each sub-process and the effects of how these processes were coupled. As a result, they are generally offered as evidence that the authors' paradigm for grass-tree coexistence is feasible, at least in so far as the model's output resembles natural savanna structure in the location that was parameterized, and are of limited value for obtaining a more detailed view of the dynamics involved.

An early example of these models can be found in Higgins *et al.* (2000) [19]. In this model, processes for rainfall, grass production, herbivory, decomposition, grass moisture content, fire (including stem mortality and stem resprouting), tree mortality, stem growth, stem neighborhoods, seed production, seed dispersal, seed bank decay, and seedling establishment are defined by a sequence of equations which are connected in various ways. Trees are modeled individually in a spatially explicit manner, and the authors promote a non-equilibrium mechanism for grass-tree coexistence based on disturbance.

Another detail to note is that rainfall was once again modeled in a rather coarse manner, using only mean annual precipitation (MAP) amounts and certain broad assumptions about its distribution during the annual seasons. Specifically, the authors assume that MAP is normally distributed about a climatic mean which is then adjusted interannually by a sinusoidal, long-term periodicity. Besides setting a probability for wet season drought, which affects seedling establishment, little attention is paid to how rainfall is distributed seasonally, which remains a defining characteristic of savanna climates.

This attention to mean annual rainfall as the primary variable for water resources also appears in some process-based models (e.g. [12]), reinforced by the observation of Sankaran *et al.* (2005) [17] that MAP is highly correlated with an upper bound on maximum woody cover in African savannas. For savannas receiving less than 650mm, this correlation translates to a linear constraint on woody cover while precipitation above that amount allows canopy closure [17]. The relative statistical importance of MAP beside several other factors (fire return interval, soil properties and browser presence) was further explored in a later paper [18], which found that mean annual precipitation was the most important determinant of woody cover below 700mm rainfall/yr. Sankaran *et al.* interpreted these data to mean

that arid (< 650mm MAP) savannas are stable, in the sense that they do not require fire to maintain grass-tree coexistence, while mesic savannas (> 650mm MAP) are unstable and require fire to prevent canopy closure. This thinking was likely influenced by existing theoretical models, which were often either resource-based or disturbance-based, and was later explored in detail by Higgins, Scheiter, and Sankaran (2010) using a new analytic model [9].

How definitive is the analysis of Sankaran *et al.* in their 2005 paper [17]? Unfortunately it raises many questions, particularly from the Australian perspective. Vast tracts of savanna with approximately 30% tree cover dominate northern Australia with mean annual precipitation between 1000 and 2000mm, directly challenging the thesis that savanna receiving more than 650mm of annual rainfall is uniformly unstable [22]. While fire is prevalent in Australia, experimentation with unburned plots supports the conclusion of savanna stability in wet Australian savannas, and even with fire present, savannas in Australia are more likely to be at or close to their climatic optimum, perhaps due in part to their higher fire tolerance compared with African tree species [23, 24, 22].

In Africa, a decades-long fire experiment across various savanna climates in Kruger National Park (mean annual precipitation between 447 and 737mm) casts further doubt on fire as a key mechanism for determining woody cover [25]. Variation in fire frequency, fire season, and total fire exclusion consistently had no significant effect on the density of trees across all rainfall amounts and soil types. Instead, fire influenced the size structure and biomass of tree populations, delivering a high mortality rate (> 0.9) for small stems less than two meters in height and a low mortality rate (< 0.05) for larger trees, though most savanna tree species also have the capacity to resprout after fire from root stocks [25].

The assumption that only mean annual precipitation need be considered, rather than intra-annual variability and seasonality, is also questionable. Models with this assumption immediately preclude dynamical comparison with climates displaying more or less rainfall seasonality at the same annual amounts, and more tellingly, wet Australian savannas with mean annual precipitation above 1000mm generally display far more seasonality than their wet African counterparts, with more than 90% falling in six contiguous months. The reverse is true when comparing arid African savannas with arid Australian savannas. In addition, Australia has far more interannual variation in rainfall than Africa, likely resulting in a fluctuation of the climatic optimum for woody cover and presenting another variable possibly masked by the use of only a static rainfall statistic in some models (e.g. [12]). These facts have led to the finding that the length of the seasonal drought, rather than mean annual precipitation, is the main mechanism by which rainfall limits tree density in Australian savannas [22].

All of these observations inform the 2007 FLAMES model of Liedloff and Cook [11]. FLAMES is a process based software model that uses an object-oriented programming approach to model individual trees in an Australian savanna stand. Precipitation amounts come directly from the daily rainfall record of the location of interest, and water is utilized by grasses and trees based upon soil infiltration and rooting depth. Fire disturbance is a key component of FLAMES, and is parameterized based on data from the Kapalga fire experiment in Kakadu National Park [24, 26]. Only small stems and very large, old stems are greatly affected by fire, and trees also possess the ability to resprout after fire damage. We will return to the mechanics of FLAMES later in this paper, as we develop a mathematical process for savanna dynamics.

One major critique of FLAMES is that it exists primarily in silico and is relatively opaque to analysis of all but the most basic sort. In contrast to agent based models, which rely on stochastic population-level simulations made up of simple individual members, the processes governing individuals in FLAMES are too complex to provide a clear explanation for their emergent behavior. This complexity also obscures the relative importance of the driving processes - some parameters and functions may be critical while others are relatively unimportant - and it is difficult to understand the strengths and weaknesses of the model in terms of functional modeling assumptions, geographical location, or climatic variables. FLAMES is hardly alone in these faults however, as several publications describing process-based savanna DGVMs have appeared throughout the decade [14, 12, 13], one requiring 33 parameters and a 29 page explanation to fully describe the model [13]. These DGVMs have primarily found use as a first evaluation of hypothetical situations [14, 15, 27] through black box model experimentation and subsequent comparison with the limited available data.

In more recent years, however, a number of simplified models have appeared in the literature that succeed in remaining mathematically tractable while including a level of detail appropriate to the system and hypothesis under consideration [28, 9, 29, 10]. In contrast with processed based simulations, these models have the ability to explore ecosystem dynamics in depth, focusing on specific processes and interactions and providing clues as to what drives savanna structure and function [28]. While these models have often focused on the dynamics of fire stem size selection or the Sankaran *et al.* (2005) [17] conclusions about ecosystem stability in Africa, Australian lessons about savanna stability in wet climates has been largely ignored.

In the following sections, we will develop a simplified model based on processes from FLAMES [11] to explore the interaction between variable resource dynamics, fire disturbance, and stand structure in Australian savannas. Our ultimate goal is to provide a basic model for savanna dynamics that can be built upon and applied globally while remaining robust to Australian phenomenon. In particular any such model must be able to account for the stability of mesic Australian savannas, so we will pay special attention to modeling water resource availability based on soil properties, rooting depth, and daily rainfall history. Fire is essentially a disturbance away from climate induced equilibrium dynamics, so understanding the basic effects of water resource limitation, variance, and seasonality is also critical for accurately capturing how fire can alter stand composition. Using data from the Kapalga fire experiment [24, 26], we can then examine the effects of fire as perturbations from an underlying climatic state.

2.3. Model formulation

We begin by considering a model similar to Walker and Noy-Meir (1982) and Higgins, Scheiter, and Sankaran (2010) [20, 9] in that soil water dynamics are characterized by a possible niche separation between grass and trees. This niche separation is generally parameterized by rooting depth, but could equivalently take the form of internal water storage or drought resistance. However, unlike the Walker and Noy-Meir model, our formulation will be less focused on competition and will ignore the effects of shade, root biomass, and relative usage of soil nutrients to suppress growth. While this approach may overestimate woody or grass biomass on the two extremes of the model (transition to forest or grassland), confounding factors such as the occasional shade tolerance of grasses, differing amounts of

woody biomass necessary for canopy closure, and complicated seedling-grass dynamics make these transitionary states difficult to model consistently across all locations and species.

Consider an Australian stand that is one hectare in size, and assume that all plants share water resources within this hectare. As rainfall enters the soil system, it first starts to fill the topsoil where it is soaked up like a sponge until the layer reaches field capacity. Any additional water added to the topsoil then begins to occupy the space between soil particles, which can continue until the topsoil reaches saturation. At the same time, any topsoil water in excess of field capacity drains quickly into the subsoil layer (it is assumed that saturated topsoil will drain back to field capacity in approximately a day), which in turn begins to similarly soak up water until it reaches field capacity. Any excess water in the subsoil is then considered lost to the system due to drainage, since the drainage rate in this layer is also relatively high on a day to day time scale. A schematic of this system, where soil layers are represented by buckets, is shown in Figure 2.1. The definition of the variables and parameters follows.

Variable definitions:

Γ: topsoil water, R: subsoil water, G: grass biomass, $\mathbf{T}$: finite vector of trees belonging to uniform woody size classes (given by basal area)

Parameter definitions:

$f(t)$: daily rainfall, $\gamma, \boldsymbol{\omega}$: water usage constants, F_Γ: topsoil field capacity, V_Γ: topsoil saturation capacity, V_R: subsoil capacity, V_S: depth of subsoil to which seedlings have access, δ: rate at which topsoil water in excess of field capacity drains to subsoil when $\Gamma > F_\Gamma$.

In our implementation of the system in Figure 2.1, we have assumed for simplicity that rainfall amounts are defined daily, e.g., via a daily rainfall record or a stochastic distribution

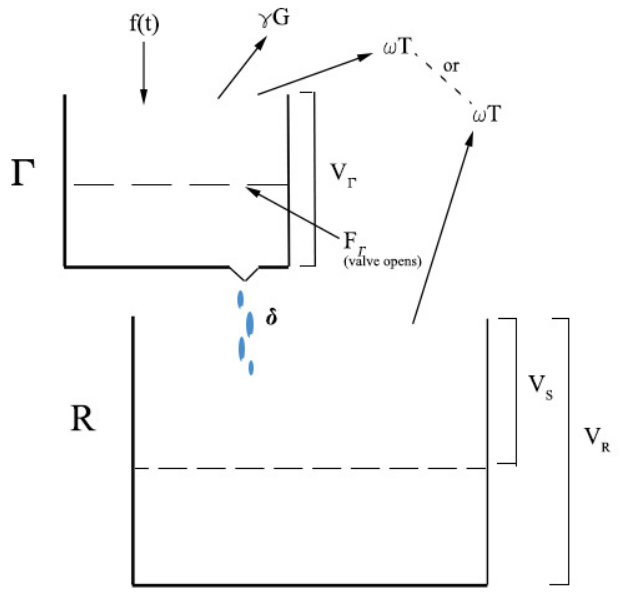

FIGURE 2.1. Schematic of water dynamics in the savanna model. Γ is the volume of water in the top (grass) layer bucket, and R is the volume of water in the bottom (tree reservoir) layer bucket.

based on such data. Water directly enters the soil as it falls, so we do not consider runoff, and the infiltration rate for water absorbed by the soil, $f(t)$, is identical to the amount of rain currently falling. This model is robust to more complicated calculations for runoff and infiltration rates by simply altering the calculation and interpretation of $f(t)$ from data. Since more nuanced infiltration rates and runoff calculations effectively just decrease the amount of water entering the system, $f(t)$ can be calculated as desired to appropriately account for intense rainfall events or other considerations.

This water resource system can be coupled to a variety of grass-tree dynamics and should prove fairly modular for use in future, alternative savanna models. In our formulation, we approximate grass dynamics by assuming that when $\Gamma > 0$ grass grows at a constant rate g (tonnes/day), and when $\Gamma = 0$, grass begins to cure (dies) at a rate c proportional to the current grass biomass [11]. Temporarily assuming that $\mathbf{T}$ is constant, we can represent the

system by the equations

$$
\begin{aligned}
\frac{d\Gamma}{dt} &= f(t)(1 - H(\Gamma - V_\Gamma)) - (H(\Gamma - F_\Gamma)\delta + \gamma G + \boldsymbol{\omega} \cdot \mathbf{T})H(\Gamma) \\
\frac{dR}{dt} &= \delta H(\Gamma - F_\Gamma)(1 - H(R - V_R)) - \boldsymbol{\omega} \cdot \mathbf{T}(1 - H(\Gamma))H(R) \\
\frac{dG}{dt} &= gH(\Gamma) - cG(1 - H(\Gamma)).
\end{aligned} \tag{3}
$$

In our effort to better capture the asymmetric demographic effects of fire, we will extend the separation of woody population used in Hanan *et al.* (2008) [28] to a vector of woody age classes representing mean population numbers. While the Hanan *et al.* approach of modeling juvenile and adult biomass differentiates between two mortality rates for fire, it loses all information about woody population numbers, implicitly making the assumption that one gigantic tree with 50 tonnes of woody biomass is identical to 100 adult trees with half a tonne biomass each. Since it has been argued that the defining effect of fire is to alter stand composition rather than total biomass [25, 22], we would like to capture the details of population demographics in our model and subsequently examine the effects of fire disturbance on stand composition to test this hypothesis.

To model the dynamics of woody biomass, we will make the simplifying assumption that basal radius increases linearly with age in the presence of water resources. However, since the vector $\mathbf{T}$ can only capture discrete classes of basal area and trees must remain in a size class for a minimum amount of time before moving on to the next one, we must either introduce a time delay for tree dynamics into our continuous-time model, or couple the continuous-time dynamics to a discretely evolving system. We found the second option

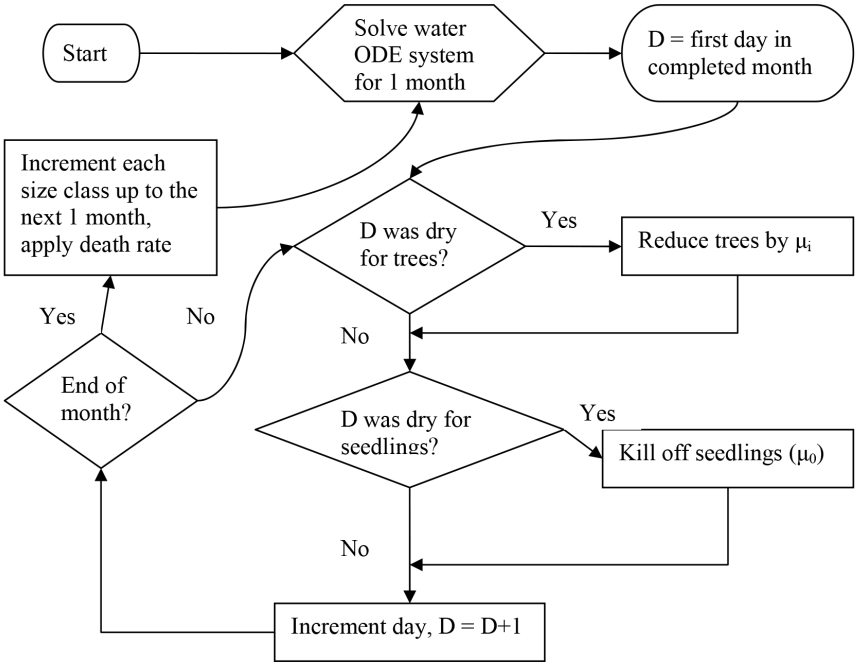

FIGURE 2.2. Flowchart for discrete/continuous numerical solution of the savanna model

to be the least difficult to implement and analyze, so we will only describe the formulation of our discrete-continuous model below.

2.3.1. WOODY DYNAMICS. In reality, trees are continuously growing through a continuum of sizes. We will discretize this process for use with a finite size (age) vector using a discrete time dynamical system, which we will then couple to the soil system described at the beginning of this section. Since trees grow on a considerably longer timescale than that which is relevant for day to day soil water dynamics, holding woody mechanics as constant on monthly intervals between discrete updates should be a good approximation to the continuous process while remaining computationally easy to implement. A schematic of the system is shown in Figure 2.2.

We consider three cases for the growth and decay of our woody population. First, either $\Gamma > 0$ or $R > V_R - V_S$ (no stress on any size classes of $\mathbf{T}$). Second, $\Gamma = 0$ and $0 < R < V_R - V_S$

(stress on seedlings, but not on mature trees). Third, $\Gamma = R = 0$ (stress on all trees). In the first case, everything grows. In the second case, seedlings begin to die, but everything else in $\mathbf{T}$ continues as normal. In the third case, there is a water shortage for all of $\mathbf{T}$, so all tree classes decrease.

Consider a size class in the vector $\mathbf{T}_t$ at time t, $T_{i,t}$, that is not near the beginning of the vector ($i > \sigma$) where trees would be considered seedlings. Let μ_i be the per-capita, per-day mortality rate of $T_{i,t}$ during times of water stress, and ν_i the death rate due to other causes (e.g., age) per month. Let Δ_t be the amount of time in days during month t that $\Gamma = R = 0$, e.g. the trees were stressed, during month t. Then we arrive at the following relation for $T_{i,t+1}$,

$$T_{i,t+1} = T_{i-1,t}(1 - \mu_{i-1})^{\Delta_t}(1 - \nu_{i-1}), \tag{4}$$

derived from solving a daily discrete-time dynamical system where either $T_{i-1,t}$ remains the same (if trees were not stressed) or is adjusted to equal $T_{i-1,t}(1 - \mu_{i-1})$ each day that water was absent.

Since the vector $\mathbf{T}_t$ is finite and everything in the last entry is lost to the system, we consider the last entry of the death rate vector $\boldsymbol{\nu}$ to be equal to 1 so that no trees live beyond reaching the last age class. The length of the vectors should be chosen such that very few trees could ever live to reach the last age class, with entries in $\boldsymbol{\nu}$ gradually increasing to 1. Ideally, $\boldsymbol{\nu}$ can be parameterized using a life table or a probability distribution describing the life-span of the woody species in question. In areas around Darwin, Australia, eucalyptus is the primary savanna tree, so we have chosen to use a gamma distribution with a mean of 400 (an estimate for the average life-span of eucalyptus) and a standard deviation of

40 (arbitrarily chosen) to represent the waiting time until natural death in our numerical simulations. This distribution was then converted into the death rate function $\nu(t)$ using conditional probabilities. Since this term is meant to model relatively rare mortality due to events such as cyclones, storms, insects and disease, the rates are generally quite low: < 0.0001/month for trees less than 300 years old, approximately 0.0017/month for trees 400 years old, and reaching a maximum of 0.0048/month just under 500 years old.

Now suppose that $T_{i,t}$ is near the beginning of the vector $\mathbf{T}_t$ ($0 < i \leq \sigma$), in an entry that would be considered a seedling size class with limited access to water resources. The growth dynamics are similar to equation (4), with the only difference being that Δ_t is replaced by a new quantity Θ_t, representing the amount of time in days during month t that $\Gamma = 0$ and $R < V_R - V_S$. This more strict condition models the seedlings greater susceptibility to drought in the soil. For new seedlings ($T_{0,t}$), we will assume a seasonal recruitment rate $s(t)$ that is independent of woody population due to the presence of latent seeds. If dependence upon woody population is desired, this dynamic can easily be incorporated into the function s, but for simplicity we will not consider such dynamics in this paper. Together with equation (4) and (3), we have the following discrete-time dynamical system for woody dynamics

$$T_{i,t+1} = \begin{cases} T_{i-1,t}(1-\mu_{i-1})^{\Delta_t}(1-\nu_{i-1}) & \text{if } i > \sigma \\ T_{i-1,t}(1-\mu_{i-1})^{\Theta_t}(1-\nu_{i-1}) & \text{if } 0 < i \leq \sigma \\ s(t) & \text{if } i = 0. \end{cases} \tag{5}$$

2.3.2. Combined model. Since we will eventually want to add fire disturbance into the model of savanna dynamics, we now include one more variable, C in equation (3) to

keep track of dead grass that has not yet decomposed. Since dry standing grass is a key seasonal fuel for fires, this quantity is important to keep track of but will have no effect on normal woody growth or water resource dynamics in our model. The full system can now be formulated by the system

$$
\begin{aligned}
\frac{d\Gamma}{dt} &= f(t)(1 - H(\Gamma - V_\Gamma)) - (H(\Gamma - F_\Gamma)\delta + \gamma G + \boldsymbol{\omega} \cdot \mathbf{T}_m)H(\Gamma) \\
\frac{dR}{dt} &= \delta H(\Gamma - F_\Gamma)(1 - H(R - V_R)) - \boldsymbol{\omega} \cdot \mathbf{T}_m(1 - H(\Gamma))H(R) \\
\frac{dG}{dt} &= gH(\Gamma) - cG(1 - H(\Gamma)) \\
\frac{dC}{dt} &= cG(1 - H(\Gamma)) - kC \\
T_{i,m+1} &= \begin{cases} T_{i-1,m}(1 - \mu_{i-1})^{\Delta_m}(1 - \nu_{i-1}) & \text{if } i > \sigma \\ T_{i-1,m}(1 - \mu_{i-1})^{\Theta_m}(1 - \nu_{i-1}) & \text{if } 0 < i \leq \sigma \\ s(m) & \text{if } i = 0 \end{cases}
\end{aligned} \tag{6}
$$

where Δ_m is the number of days in month m that $\Gamma = R = 0$ and Θ_m is the number of days in month m that $\Gamma = 0$ and $R < V_R - V_S$.

This hybrid continuous-discrete time implementation has the advantages of being simple yet thorough when a detailed solution for water resource dynamics is desired. An example of the output for this model based on data from Darwin, Australia is shown in Figure 2.3.

2.3.3. Long term dynamics. For time scales longer than about 30 years, and particularly when day-to-day water resource dynamics are unimportant in model output, we suggest moving to a fully discrete model by replacing the ODE system in equation (3) with an updating model that processes soil water content, grass biomass, and rainfall once every

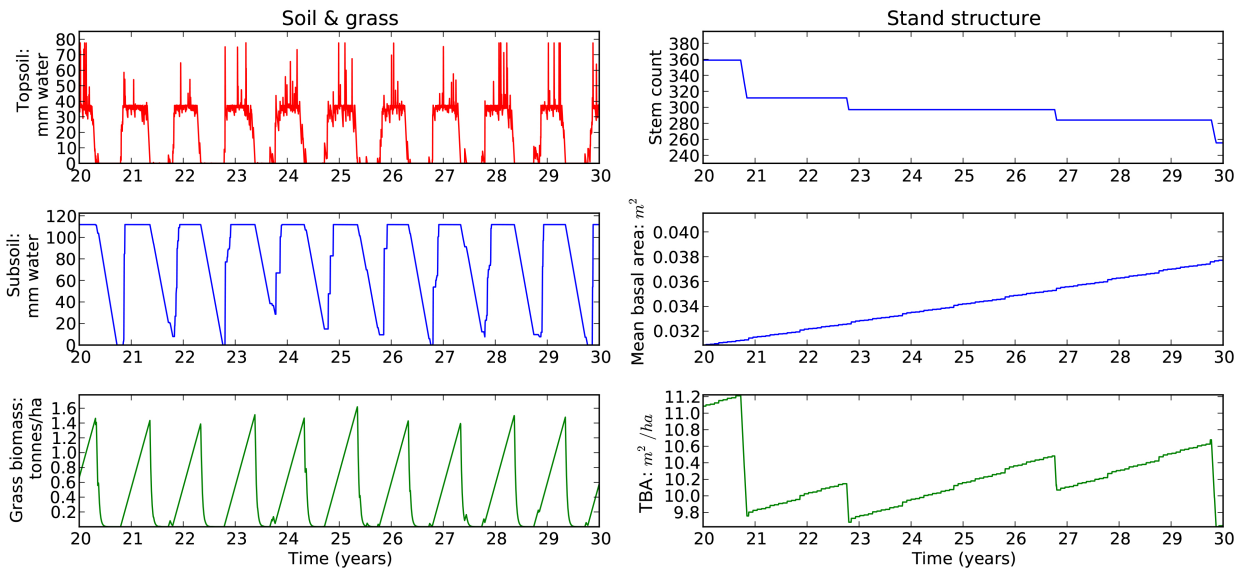

FIGURE 2.3. Soil moisture, grass biomass, and woody dynamics including total basal area (TBA) after 30 years (the last 10 years are shown) using the combined continuous/discrete approach and data from Darwin, Australia. Initial conditions were chosen to represent a current stand nearby Darwin.

day. Numerical trials comparing outputs from both of these approaches suggest that solution differences are negligible, and since the fully discrete implementation runs on a longer time scale (a day for all sections of the model), it has a huge advantage in speed and memory. An example of such long-term dynamics for Darwin, Australia is plotted in Figure 2.4 and 2.5, using the same parameters as in Figure 2.3.

Using stochastic rainfall distributions parameterized from the rainfall records of a location, arbitrarily long time periods can be simulated for long-term effects. There are many choices for forming such a distribution from data, most of them involving a two step process by which a given day is determined to be either wet or dry (using a Markov process), and then if the day was wet, an amount of rainfall is determined by sampling a probability distribution (often the Gamma distribution). Both steps can easily be parameterized using

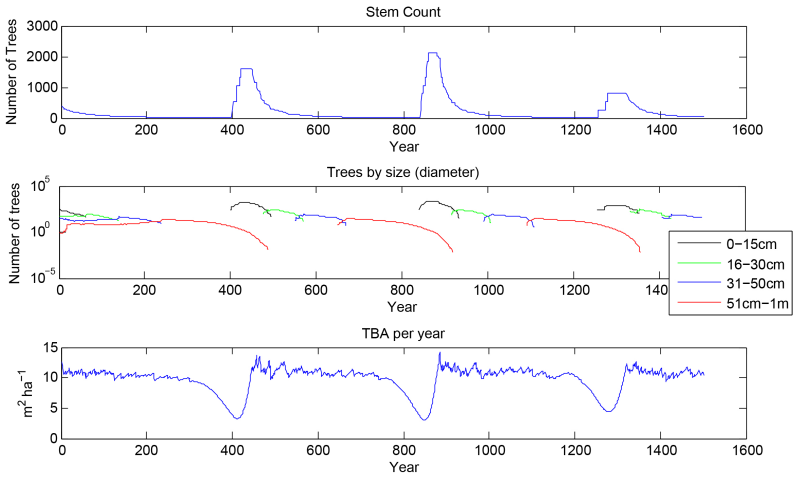

FIGURE 2.4. Model prediction for stem count, stand composition, and total basal area for Darwin, Australia without fire disturbance over a period of 1500 years. Note that the middle plot is log-linear.

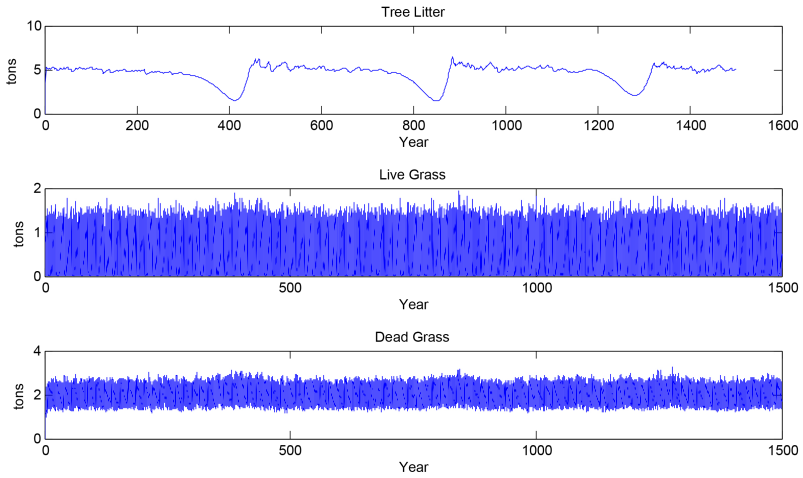

FIGURE 2.5. Model prediction for tree litter, live grass, and dead grass amounts in tonnes over a period of 1500 years. The live grass and dead grass plots are pretty typical for this model (with an additional feature that dead grass amounts drop to zero when a fire occurs), and tree litter dynamics typically mimic total basal area. The calculation from FLAMES was used to acquire these numbers, and dead grass and tree litter is used to determine fire intensity in model runs were fire dynamics are included.

a suitably long daily rainfall record, and quite a few variations on this approach and other methods have been examined, including the case where less data is available [30, 31].

In our numerics, we have chosen to use Bernoulli trials in lieu of a more complicated Markov process while sampling from a Gamma distribution for rainfall amounts. We then parameterized our combined Bernoulli-Gamma distribution, $\Phi(t_i)$, on a daily basis using a fairly long rainfall record (128 years in the case of Darwin, Australia) in order to generate stochastic, location specific, precipitation amounts. While this method provides only a very rough approximation to precipitation trends because it completely neglects the tendency for rainfall to occur on successive days, it is very easy to implement and should serve to illustrate the utility of our model and its subsequent analysis.

When fire is not present in the system, our model predicts a stand that consists of similarly aged trees dating back to a period of heavy seedling recruitment (Figure 2.4) with a relatively constant live and dead grass biomass (Figure 2.5). When enough water resources are present, monthly seedling germination events continuously introduce individuals into the model until their water usage equals the available water, after which the large quantity of seedlings begins to thin. Competition for the now limited resources between neighboring trees in turn limits woody biomass as the individual trees mature and grow old, decreasing the total stem count but maintaining a roughly constant total basal area. Finally, as the older trees begin to die out, water resources are once again unused and seedlings appear in large numbers to repeat the cycle.

2.4. Analysis

To begin our analysis of the process for water resources represented in equation (3), we will examine the dynamics of soil water during the relatively short time scale of one year. On this time scale, we assume that any changes in woody biomass are negligible and hold $\mathbf{T}$ constant, which will allow us to uncouple equation (3) from the woody growth/death

process in equation (5), and focus purely on the effect of stochastic rainfall distribution on soil moisture content and grass biomass. The goal of this approach is to gain an understanding of how stochastic variation in rainfall can influence woody demographic dynamics by stressing a given stand structure or alternatively, providing sufficient resources for additional growth or seedlings. This analysis is similar to the approach taken by D'Odorico *et al.* [32] for fire-induced savanna dynamics.

In the absence of fire, the primary environmental variable affecting tree mortality is water stress. Since adult trees become stressed when $\Gamma = R = 0$, our goal should be to quantify the likelihood and duration of this model state for the year given a current stand structure. We first note that it does not take an unusual amount of rainfall to saturate the subsoil and obtain the state $R = V_R$. This state is realized after a few days of strong rain and is actually the typical state of the system during a typical wet season rainfall pattern. Similarly, the typical state for Γ during the wet season is $F_\Gamma \leq \Gamma \leq V_\Gamma$. Figure 2.3 illustrates these points for the seasonal rainfall in Darwin, Australia.

For the purpose of analysis, we now make a few definitions regarding rainfall seasonality. We define the start of a dry season, t_d, as a day in which $\Gamma = 0$ and $R \approx V_R$. We then assume that on this day $R = V_R$, and that G decreases quickly toward 0 so that we can assume $G = 0$ for the duration of the dry season. We now define the end of the dry season as the first day in which $R = V_R$ again, though we will mistrust this date if it falls too early (less than 150 days after the start of the dry period). On the occasions when this scenario occurs, we record the date but continue to look forward in the year in case the soil dries out considerably again, signifying that the dry season is not yet over and the wet conditions

were a statistical anomaly unrelated to seasonality. In this case, we continue to look for the end of the dry season; otherwise, we keep the original date recorded.

We now suppose that the probability distribution for the daily rainfall function, which we treat as a random variable $\Phi(t_i)$ where t_i is the ith discrete day, is known. Supposing that $G = 0$ and $\boldsymbol{\omega} \cdot \mathbf{T}$ are constant, if $\Gamma \neq 0$, R and Γ decay at a constant rate as long as $\Phi(t_i) = 0$. When $\Phi(t_i) \neq 0$, we will continue to assume $G = 0$ until $R = V_R$ (this is based on the observation that grass returns relatively slowly), so that all of the water is used by the woody biomass and decays at the constant rate of $\boldsymbol{\omega} \cdot \mathbf{T}$. We will further assume that when $\Phi(t_i) > 0$ during the dry season, no water is lost to runoff. This assumption greatly simplifies our calculation and can be justified in a number of scenarios, particularly when dry season rains are known to be light and/or brief, or runoff is locally contained. Even if this assumption is not particularly justified by the physical situation, we can treat our resulting calculations as a wettest-case scenario for R and approximate downward from there.

Note that since $G = 0$ during the dry season, all incoming rainfall from Φ is either absorbed by the woody biomass at a constant rate of $\boldsymbol{\omega} \cdot \mathbf{T}$ or is stored in R for later use. More specifically, we can observe that $\Gamma \neq 0$ only for temporary, localized periods of time until the end of the dry season. As a result, we will neglect the topsoil stage of the model in our current analysis and assume that all incoming rainfall directly enters R. One way to conceptualize this simplification is that if water enters Γ and R stays constant for a certain amount of time while it is absorbed, the effect is the same as if the water had entered R instead and then is absorbed back down to the level where R would have stayed constant in the original scenario. To allow R to contain the water in Γ, we will assume that it has capacity $V_R + F_\Gamma$ and declare the dry season over when $R = V_R + F_\Gamma$.

As a result of these simplifications, we can now think about soil water R as a discrete-time Markov process with a continuous state-space and absorbing boundary conditions at $R = 0$ and $R = V_R + F_\Gamma$. The initial value of this process is $R = V_R$, and the internal transition probabilities are given by the time-dependent random variable $\Phi(t_i) - \boldsymbol{\omega} \cdot \mathbf{T}$. In the analysis of this process, we are immediately interested in two probability distributions: the amount of time spent at $R = 0$ and the first transit time at $R = V_R + F_\Gamma$, which marks the end of the dry season. These distributions quantify the amount of stress on woody biomass and the length of the dry season, respectively. We would also like to quantify the probability that in a given realization $R = V_R + F_\Gamma$ before $R < V_R - V_S$, since this is the probability that seedlings can occur despite the dry season.

Since this process is non-linear due to the absorbing boundary conditions, the most straight-forward method of analysis is numerical exploration. For a sequence of rainfall random variables $\Phi(t_i)$, we can quickly and easily generate hundreds of realizations for R and approximate the distribution for the first transit time of $R = V_R + F_\Gamma$. We can similarly approximate probabilities for the length of time when $R = 0$ and the probability that $R \not< V_R - V_S$ in a realization by recording the frequency of such visits, and then examine how these probabilities change with perturbations in the drift term, $\boldsymbol{\omega} \cdot \mathbf{T}$. The result is a comprehensive description of climate based effects on different savanna stand structures in a given location, and this information will also provide us with a basis to understand the effects of fire on available water resources.

Plots 2.6 through 2.9 were generated using this method. Note in Figure 2.9 that even with soil parameters from Darwin, Australia, Sydney's rainfall distribution produces *far*

more woody biomass than Darwin while receiving less mean annual rainfall (1610 mm/year verses 1204 mm/year).

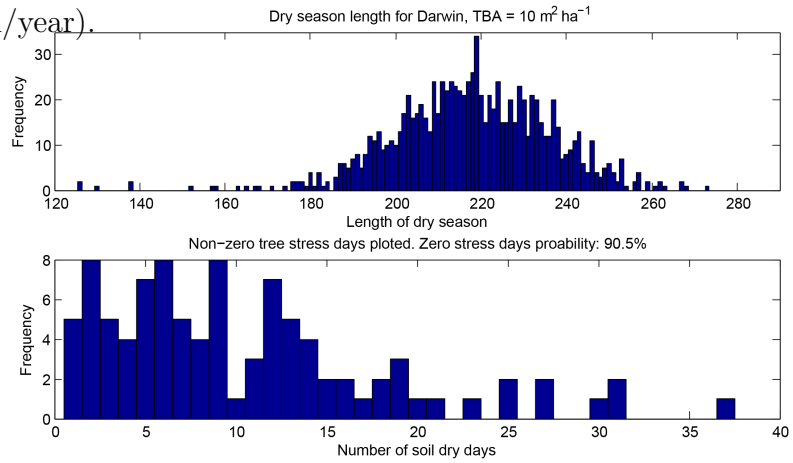

FIGURE 2.6. Approximate probability distribution of (a) dry season length and (b) non-zero tree stress days for a Darwin base tree stand with $10\text{m}^2\text{ha}^{-1}$ total basal area. Note that for 90.5% of the time, the soil was never dry.

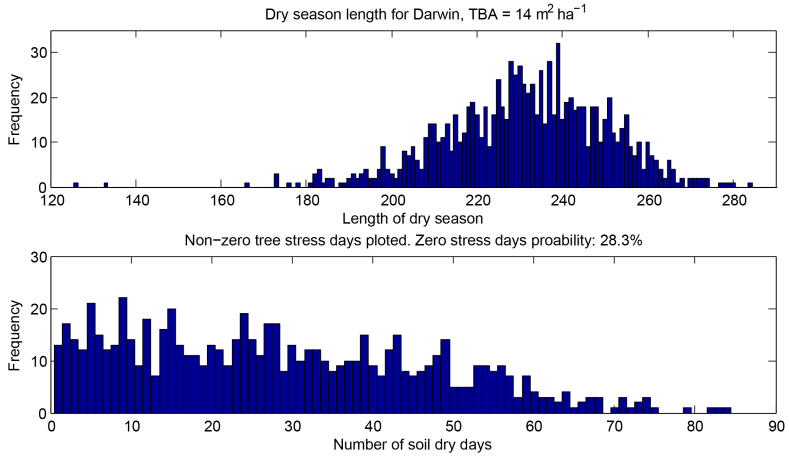

FIGURE 2.7. Approximate probability distribution of (a) dry season length and (b) non-zero tree stress days for a Darwin base tree stand with $14m^2ha^{-1}$ total basal area. Note that for 28.3% of the time, the soil was never dry.

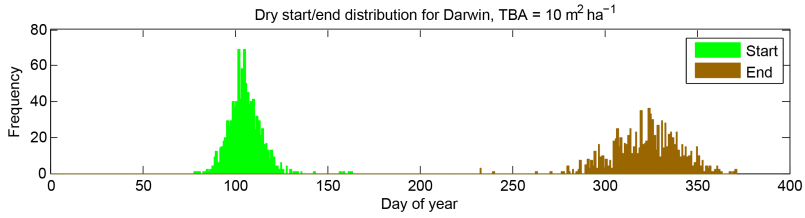

FIGURE 2.8. Distributions for the beginning and end of the dry period.

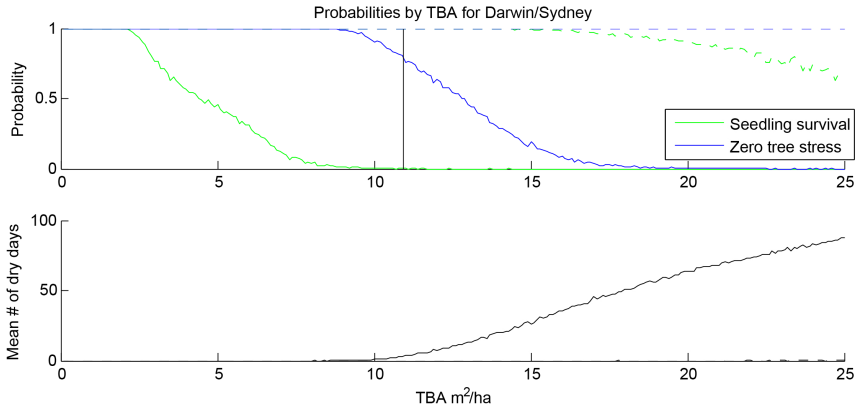

FIGURE 2.9. Stress probabilities for Darwin, Australia (solid lines) and Sydney, Australia (dashed lines) by total basal area. From top to bottom: (a) Probability that seedlings will survive in a given year, given that seedlings die if soil water level drops below a critical rooting depth during the year. Plotted with the probability that soil water is not depleted during the year. The black line marks the climatic induced average as predicted by Figure 2.4. (b) Mean number of days that the soil will be dry during the year. This last statistic is somewhat misleading, because the distribution is often heavily skewed toward zero dry days during the year. The soil parameters for Darwin, Australia were used for both locations.

2.5. The effect of savanna fires on woody dynamics

As mentioned in Section 2.2, fire can significantly alter stand structure and woody biomass away from climate induced equilibria. Since the immediate effect of fire is to reduce woody biomass, water usage can also be reduced temporarily, possibly resulting in additional, secondary changes to stand structure. We will explore some of these effects through simulation utilizing various fire regimes.

To observe the effect of fire on savanna stands, we implement the disturbance mechanics developed for FLAMES [11] based on the Kapalga Australian fire experiment [24, 26]. Fire intensity, and thus tree morality, is a function of monthly humidity averages, cured grass amounts, and stochastically generated wind conditions around a yearly average. When fire occurs, we assume that it affects the entire stand in a demographically asymmetric way. Seedlings, small trees, and old trees generally suffer significant fire damage, while

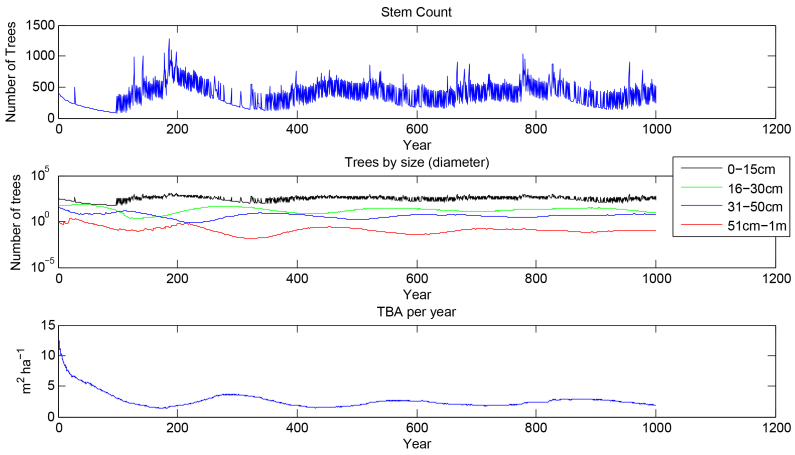

FIGURE 2.10. Model prediction for stem count, stand composition, and total basal area for Darwin, Australia with fire disturbance every two years in July.

normal adult trees have a low rate of mortality. In addition, trees burned over a certain age are allowed to resprout into a smaller age/size class 90% of the time, an assumption approximately based on observed resprout rates for eucalyptus in northern Australia.

Using the stochastic rainfall distributions described in Section 2.3.3, Figures 2.10 and 2.11 show the effects of burning every 2 and 4 years respectively during the dry season. Note that since these burns are conducted regardless of stand structure, seedlings are heavily suppressed by the burns but reappear almost continuously in the two year case, and over a prolonged period in the four year case, due to sustained resource availability caused by a reduction in woody biomass. Burns conducted every two years also heavily reduce woody biomass, eventually showing similar behavior as predicted by the model for yearly burns. Gignoux *et al.* (2009) explored the effects of frequent burning in the field in detail, and have found that forest species may be excluded at the seedling stage due to an inability to stabilize their survival probability at the following resprout stage, which is in line with our results for similar regimes [33].

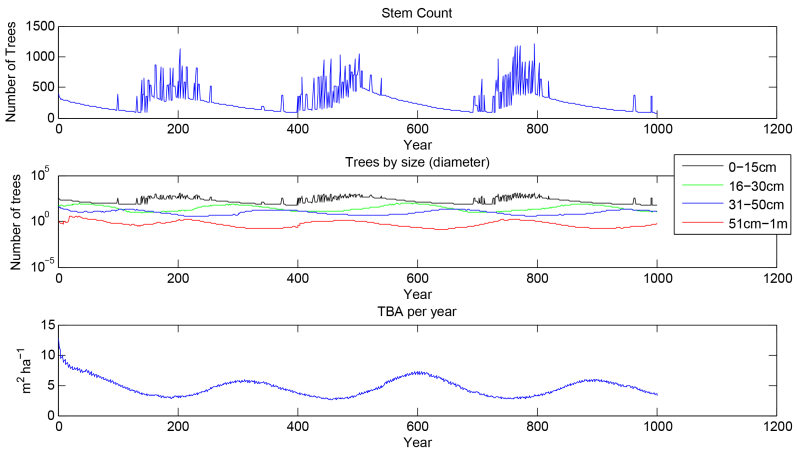

FIGURE 2.11. Model prediction for stem count, stand composition, and total basal area for Darwin, Australia with fire disturbance every four years in July.

In our model, both fire regimes also produce very specific, periodic dynamics for total basal area and reduce the period of the long-term oscillation from the 400 years seen in Figure 2.4. This observation is easily explained by noting that fire has replaced water shortage and other phenomenon as the main cause of mortality among larger trees, and trees do not typically grow as large before being burned. A typical stand will consist of trees with the same approximate age from the same seedling recruitment period. These trees continue to grow for a while, only gradually thinning due resistance to fire. Finally, as they become large, the stand has thinned sufficiently to allow seedlings to survive while the remaining trees from the previous generation die out due to higher susceptibility to fire (see, for example, the middle plot in Figure 2.11). This high susceptibility to fire in Australian savannas may be due to termite hollowing weakening old trees exposed to many fires, as observed in the Kapalga fire experiment [26, 24]. Seedlings are also quite susceptible to fire in both Africa and Australia [26, 24, 25], and it can take some time for the full stand to re-establish.

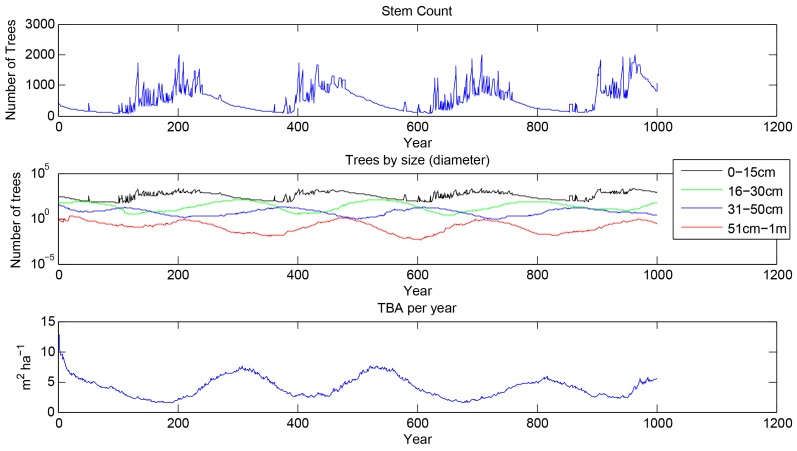

FIGURE 2.12. Model prediction for stem count, stand composition, and total basal area for Darwin, Australia with a 0.5 probability for fire disturbance each year in July. Mean total basal area for the time period shown was calculated to be $3.62\text{m}^2\text{ha}^{-1}$.

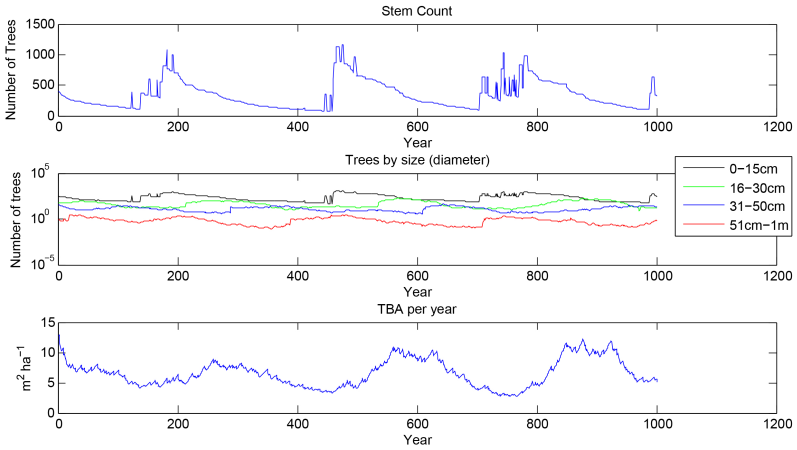

FIGURE 2.13. Model prediction for stem count, stand composition, and total basal area for Darwin, Australia with a 0.25 probability for fire disturbance each year in July. Mean total basal area for the time period shown was calculated to be $6.57\text{m}^2\text{ha}^{-1}$.

Figures 2.12 and 2.13 add some measure of stochasticity to the previous regimes seen in Figures 2.10 and 2.11. Instead of a constant burn pattern, every year has a set probability (0.5 and 0.25 respectively) that a fire will occur in July, which means that some periods will have frequent fires while others go without, possibly allowing seedlings to establish in ways that were previously impossible. We can see the effects of stochasticity in fire occurrence

on stand structure especially well when comparing Figure 2.10, with 2 years between fires, and Figure 2.12, with a mean of 2 years between fires. Figure 2.12 shows significantly larger stable seedling recruitment events than Figure 2.10, as displayed by the stem count plot, and Figure 2.12 also predicts far more woody biomass in the stand overall when compared to the equivalent plot in Figure 2.10. Comparing the two stochastic regimes, burning with a 0.25 probability every year produces almost twice the mean woody biomass as burning with a 0.5 probability, and can approximately reproduce stand structures and biomass currently seen in the Darwin, Australia area (compare the 550 year prediction in Figure 2.13 to the initial conditions in the same figure).

To examine the effect of even more stochasticity in the timing of fire events, Figure 2.14 displays the model result for a uniform, 0.05 monthly probability that a fire will occur. As a result, we expect a fire roughly every 2 years, but the fire most often will occur during a season where there is less dry grass and leaf litter available to burn. Compared with Figure 2.12, which has the same 2 year expected value for fire, we see far less seedling recruitment in Figure 2.14, but also less variation in total basal area. Since there is now only a relatively small chance of fire occurring in the summer months when there is enough unburnt fuel for a strong fire, seedlings are not generally suppressed by fire, and instead appear all at once when conditions are right. This trend continues for smaller monthly probabilities, often resulting in conditions not unlike those currently seen in the Darwin area, but the results are highly variable due to the stochastic placement and intensity of critically less frequent fire events.

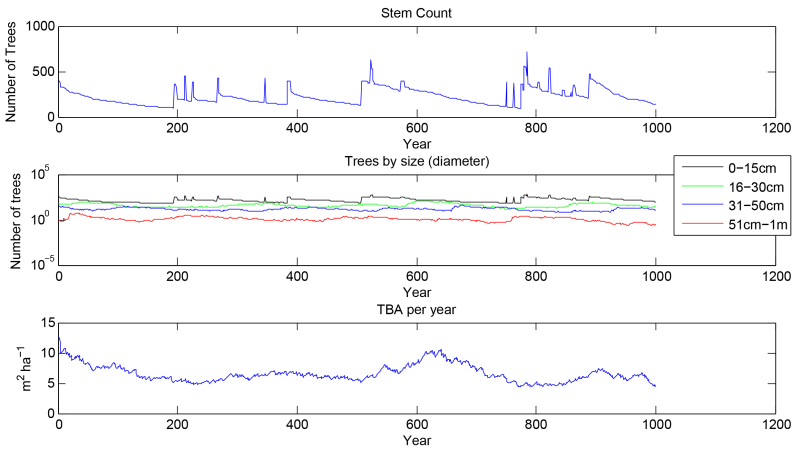

FIGURE 2.14. Model prediction for stem count, stand composition, and total basal area for Darwin, Australia with a 0.05 probability per month of fire disturbance.

2.6. DISCUSSION

With equation (3), we have introduced an explicit yet robust mathematical model for savanna water resource dynamics. This model provides transparency and tractability to the primary climatic process driving all savannas and builds on three decades of ecological understanding and theory, with the main simplifying assumption of the model being that all plants have access to the entire hectare of water. By coupling this model with the discrete-time woody growth process described in equation (5), we can then observe this model in action over the span of decades or even centuries.

Numerical simulations based on conditions outside of Darwin, Australia, suggest a very specific, climate induced cycle for savanna stands in the absence of fire disturbance. Stands tend to be roughly the same age, having sprouted in a period where conditions were unusually favorable. Mortality and growth roughly balance each other throughout the lifetime of the stand, with trees growing and dying in tandem to roughly maintain the resource induced optimum for woody biomass. Eventually, age becomes a factor for the stand, and with no new seedlings to replace occasional losses due to insects, wind, cyclones, rot, and other

factors, the stand begins to thin. This process finally frees up enough resources to allow seedlings to germinate, and with the old stand dying out, the cycle begins anew.

When fire disturbance is added back into the model, we see very different results with the outcome depending heavily upon the burn regime used. If a constant, intense, frequent burn regime is used (every one or two years in July), new woody growth is heavily suppressed due to the susceptible nature of juvenile trees to fire. The current stand slowly burns out as it gets older, and in the end, fire reduces the stand to a state that is completely determined by the specifics of the disturbance. Seedlings will almost always have plenty of resources to germinate, but fire typically removes woody growth early, so the stand consists mostly of juveniles. This observation has consequences for current frequent fire regimes, in that our model predicts long-term stand collapse when new woody growth is consistently excluded. More responsive fire regimes, however, could avoid this result by allowing seedlings to grow into fire resistent sizes as they appear.

The suppressing effect lessens as the interval between burns increases, and regular burns instead begin to primarily reduce the maximum expected age for adult trees in the stand. The outcome becomes more similar to the case without fire, only with a shortened cyclic period due to earlier removal of woody growth by fire. Similar results are observed for high intensity stochastic fire regimes (constant annual probability of burning in July), even when the annual probability for fire is quite high (0.5). In this case, stochasticity plays a big role in the resulting dynamics by allowing seedlings to mature to a fire resistant size during the occasional periods of prolonged fire absence. When fire can happen at any time of the year, there is a large chance for low intensity fires that may consume fuel and a small amount of susceptible woody biomass, but otherwise have little effect. The result of such fire regimes is

highly stochastic, with seedlings appearing at irregular intervals depending on stand specific conditions and fire history.

While these numerical simulations were kept simple with a constant woody growth rate and constant seedling recruitment when water resources allowed, the results reveal critical dynamic relationships between fire disturbance, water resource allocation, and stand structure. Most foundationally, our model demonstrates how rainfall distribution, rather than mean annual precipitation, can function as the primary driver behind climatically induced maximums on woody biomass. This is especially the case for ecosystems in high precipitation locations, as found in northern Australia, and firmly establishes in theory that seasonal availability of resources alone is enough to maintain a savanna state.

Analysis of the model resulted in a method of quantifying the probability of water stress and seedling recruitment in a given year. Given a current savanna stand structure and a probability distribution for daily precipitation amounts, we have described a simple process for generating the distribution of seasonal drought length and severity. This analysis holds regardless of stand history or fire roles and provides a useful tool to predict stand resource pressures in the short term. By coupling this process with an understanding of potential fire intensities and resulting biomass reduction (perhaps using a more detailed, probabilistic approach similar to D'Odorico *et al.* (2006) [32]), it is now possible to explore how different fire scenarios can immediately effect water consumption and seasonal stress.

Additional study is still required to understand the relative importance of shade conditions and understory on savanna conditions, both of which were not considered in this thesis, though these factors may be quite location and species specific. In our model formulation,

grass did not appear to play a significant role in woody dynamics (other than slightly reducing the total amount of water available at the beginning of the dry season), and this fact gives us further reason to believe that the paradigm of grass-tree competition and savanna as unstable state may be incorrect. Instead, savanna appears to be possible as a climatically induced state whose structure is defined primarily by woody resource dynamics, fire disturbance, and the seasonality and quantity of water resources, with grass filling in open spaces and providing fuel for fire. This dynamic directly supports the main hypothesis of Higgins *et al.* (2000) [19], who suggested that grass-tree coexistence is driven by the limited opportunities for tree seedlings to escape both drought and flame into the adult stage.

While our model was kept relatively simple for illustrative and analytic purposes, certain assumptions may be relaxed without serious implications to the model. For example, it would not be difficult to maintain two or more size class vectors for different tree species, rather than assume that all trees behaved the same, and the calculation for their age related growth rates can be altered at will. Similarly, multiple understory species may be accommodated alongside the variable for total grass biomass. More difficult challenges would include attempting to quantify woody cover and its effect on understory growth, and without a spatial component, it would be very difficult to remove the assumption that all water resources within the hectare are shared.

In future work, we hope to parameterize our model for locations both inside and outside of Australia to examine the robustness of our technique to different woody species and climatic scenarios. Since the model presented here was based on the processes underpinning the FLAMES simulation model of Liedloff and Cook (2007), stochastic realization of stands approximating the dynamics presented in section (3.2) can be found using the FLAMES

software [11], providing a more concrete, visual tool for management exploration of the scenarios described in this paper.

CHAPTER 3

Modeling the presence probability of invasive plant species with nonlocal dispersal

3.1. Introduction

Invasive species represent one of the major environmental threats of the 21$^{\text{st}}$ Century. Damage to native species and habitat as well as agricultural lands leads to economic suppression, reduced food and water security, and direct threats to human health [34]. Cheatgrass (*Bromus tectorum*), for example, has invaded over 50 million acres of rangelands, pastures, crops, prairies, and open meadows in the western states and continues to increase its range by 14% annually. It and other invasive grasses provide fuel for wildfires, and reduce plant diversity, critical wildlife habitat, and crop yield [35, 36, 37]. Although it is not possible to eradicate these species given their current spatial extent, range managers hope to minimize spread by, for example, the application of herbicides to vulnerable areas. Control costs rise dramatically with population size, so mathematical models of potential spread to inform management strategies are essential.

Ecological niche models (also called species distribution models, environmental matching models, and habitat suitability models) are increasingly being used to model and map invasive species distribution potential [38, 39]. Combining statistical algorithms with geographic information systems (GIS), ecological niche models use presence-only or presence-absence data in combination with environmental variables to predict the species' potential distribution across a landscape. These tools help to define conservation priority areas and aid in making public health decisions and in investigating the potential impacts of climate change

[38]. While niche modeling tools have been successful in estimating the potential range of a species, these programs are unable to model the *continuous spread* of a species, and so are of limited use in predicting the time scales involved, or for modeling the effects of different management strategies.

Modeling the spread and growth of biological species has been the subject of considerable mathematical interest over the last century, an interest that continues to the present day. The classical starting point (as introduced by Skellam (1951) [40] and continued to this day) is a reaction-diffusion system such as the Fisher-KPP equation, which exhibits logistic growth coupled with Fickian diffusion [41, 42, 40]. This approach has been widely criticized for numerous reasons. Dispersal is assumed to obey Fick's first law of diffusion, and to be uniform in all directions and normally distributed. In reality, plant species in particular do not "diffuse" at all - the current population remains stationary rather than physically moving to regions of lower concentration. Furthermore, dispersal can be more complicated depending on environmental and species dependent factors, and frequently fit much better with leptokurtic distributions rather than normal ones [43, 44, 45].

Another approach is to use continuous-time models based on a "contact-birth process", as introduced by Mollison (1977) [46]. Individuals are assumed to have a fixed spatial location, and each new individual born into the population is assigned a permanent location with distance from the parent given by a probability distribution. This approach is much better suited for herbaceous invasive species or other situations where the movement of present individuals is assumed to be negligible. Furthermore, such models lack the assumption of diffusion and may incorporate a variety of contact distributions, including leptokurtic ones.

They have been used in the context of epidemiology [47, 48, 49] and in more recent years, have attracted some attention for ecological applications [50].

However, Mollison and Daniels (1992) [51] point out that for the general nonlinear case of the contact birth process, a deterministic formulation in the form of a differential equation for population numbers does not follow from taking the expected value of the actual, underlying stochastic process. Rather, these authors show that the stochastic birth process admits the equation

$$\dot{y}(x,t) = \alpha\bar{y}(1-y), \tag{7}$$

where instead of population size, $y(x,t)$ is the probability that the expanding population front has passed location x at time t, and $\bar{y}$ is the convolution of y with the contact distribution. While this observation undermines the justification for studying Equation (7) as an approximation to a real (and stochastic) physical process, subsequent literature has interpreted $y(x,t)$ as population size [47, 50], citing Mollison's earlier work [48] which studied Equation (7) for results on propagation velocities. We favor the use of Equation (7) to model the spread of invasive plant species, but the aim of this paper is to derive a correct interpretation of a modification of this equation (Equation (23) in Section 3.2) from the underlying stochastic process presented by Mollison (1977) [46]. In our derivation, $y(x,t)$ represents the probability of finding a species at location x at time t.

A governing equation for the evolution of species presence probability complements available field data on species distribution which typically consist of presence-absence information over a spatial domain rather than species density. Field-condition-motivated initial conditions for such a model can be determined in spite of the lack of data on population size.

As we discuss in Section 3.3, the output of ecological niche models naturally incorporates into stochastic realizations of contact-birth processes and our deterministic equation to give spatially varying parameters.

A challenge in deriving a deterministic equation from the contact-birth process is determining of how the carrying capacity K (the ecological parameter in the contact-birth process that we focus on) carries over to parameters in the deterministic equation. We determine this dependence by comparison of the numerical simulations of the deterministic equation that we derive against stochastic realizations of the contact-birth process, and we find a nonlinear dependence of the parameters in the deterministic equation on K. Consequently, the speed of traveling wave solutions also is a nonlinear function of K. As we discuss in Section 3.2, this contrasts with calculations of wavespeed given by Medlock and Kot (2003) [47] which give wavespeed as proportional to carrying capacity in a model similar to Equation (7), but with $y(x,t)$ interpreted as population size (or number of infected individuals in an epidemiological interpretation).

This chapter is published in the Journal of Mathematical Biology [7] and is organized as follows: In Section 3.2, we derive from a contact-birth process first a spatially discrete, and then a spatially continuous evolution equation for the probability of species presence in landscapes where the carrying capacity for the species is constant. We also discuss in Section 3.2 the dependence of the speed of traveling wave solutions on the carrying capacity. We extend the model to heterogeneous landscapes of spatially varying carrying capacity (interpreted in terms of habitat suitability output of ecological niche models) in Section 3.3. Simulations of the model as well as the underlying stochastic process in two-dimensional spatial domains are presented in Section 3.4. We conclude in Section 3.5 with a discussion of

practical applications and potential extensions of our model, before examining an extension of particular interest in Chapter 4.

3.2. A Continuous-time, Contact-Birth Presence Model

We will examine in detail a generalization of the contact birth process introduced by Mollison (1977) [46]. We first consider a spatially discrete version of the model and introduce our main quantity of interest, which is the probability that the species is present at a given spatial location and time. An evolution equation for this probability is derived using a Master equation approach. We then make a continuum transition and discuss the spatially continuous version of this model.

3.2.1. Spatially Discrete Model. Consider a set of nonnegative integer-valued populations $Y_i(t)$ occupying a finite or infinite set of cells labeled by indices i from an index set $\mathcal{L}$, $i \in \mathcal{L}$. We assume that the Y_i evolve stochastically in time t according to a contact-birth process [52, 53, 54] with transition rates

$$\lim_{\tau \downarrow 0} \left(\frac{1}{\tau} \Pr[Y_i \to Y_i + 1 \text{ in } (t, t+\tau)] \right) = r \overline{Y}_i f_i(Y_i), \tag{8}$$

where $r > 0$ is the growth rate, $\overline{Y}_i = \sum_{j \in \mathcal{L}} W_{ij} Y_j$, and W_{ij} is the probability that an individual located in cell j gives birth to a new individual in cell i. The function $f_i(Y_i)$ measures how the population growth rate responds to crowding, so that typically $f_i(Y_i)$ is a decreasing function satisfying $f_i(Y_i) \geq 0$ and $f_i(0) = 1$. The entries W_{ij} of the weight matrix W are assumed to be real and nonnegative with $\sum_j W_{ij} = 1$. If cell i has a finite carrying capacity, k_i, the population is limited to $Y_i(t) \leq k_i$ and (8) holds for $0 \leq Y_i(t) < k_i$. We

set $k_i = \infty$ if cell i has no carrying capacity. If k_i is finite, we require that $f_i(k_i) = 0$. The standard choice in this case is a linearly decreasing function, $f_i(Y_i) = 1 - Y_i/k_i$.

Let $\mathbf{Y} = (Y_1, \ldots, Y_N)^T$ be the discrete random vector of populations, and let $p(\mathbf{y}, t) = \Pr[\mathbf{Y}(t) = \mathbf{y}]$ be the probability that $\mathbf{Y}(t) = \mathbf{y}$ for any possible realization $\mathbf{y} = (y_1, \ldots, y_N)^T$. The probability distribution $p(\mathbf{y}, t)$ satisfies the Master equation [55]

$$\dot{p}(\mathbf{y}, t) = \sum_{\{\mathbf{y}'\}} w(\mathbf{y}, \mathbf{y}')p(\mathbf{y}', t) - a(\mathbf{y})p(\mathbf{y}, t), \tag{9}$$

where the dot denotes $\partial/\partial t$ and the sum in (9) is a short-hand notation for summation over all possible realizations. The Master equation (9) is the general form of the evolution equation for discrete-states, continuous-time stochastic processes if the transition probability $Pr(\mathbf{Y}(t + \tau) = \mathbf{y}|\mathbf{Y}(t) = \mathbf{y}')$ for small τ has the form

$$Pr(\mathbf{Y}(t + \tau) = \mathbf{y}|\mathbf{Y}(t) = \mathbf{y}') = [1 - a(\mathbf{y})\tau]\Pi_i\delta(y_i, y_i') + w(\mathbf{y}, \mathbf{y}')\tau + O(\tau^2), \tag{10}$$

where $w(\mathbf{y}', \mathbf{y}') = 0$ and $\delta(y, y')$ is the discrete delta function, $\delta(y, y') = 1$ if $y = y'$ and $\delta(y, y') = 0$ if $y \neq y'$. Accordingly, for given $\mathbf{y}'$, $w(\mathbf{y}, \mathbf{y}')\tau + O(\tau^2)$ is the probability of the transition $\mathbf{y}' \to \mathbf{y}$ for $\mathbf{y} \neq \mathbf{y}'$, and $1 - a(\mathbf{y}')\tau + O(\tau^2)$ is the complementary probability that there is no transition, that is,

$$a(\mathbf{y}') = \sum_{\{\mathbf{y}\}} w(\mathbf{y}, \mathbf{y}').$$

The Master equation (9) follows directly from the Chapman-Kolmogorov equation in the limit $\tau \to 0$.

The specific form of $w(\mathbf{y}, \mathbf{y}')$ for our model is derived from (8). Since time is continuous and $\mathbf{Y}$ is a discrete random vector of countable length, the probability of two simultaneous

transitions is almost surely zero. Therefore, to simplify our computations, we will assume that the probability of such transitions is of the order $O(\tau^2)$ in (10). The same assumption is made for the probability of transitions by more than a unit. Then, according to (8), the transition matrix $w(\mathbf{y}, \mathbf{y}')$ is given by

$$w(\mathbf{y}, \mathbf{y}') = r \sum_{i \in \mathcal{L}} \overline{y_i'} f(y_i') \delta(y_i - 1, y_i') \prod_{j \neq i} \delta(y_j, y_j'), \tag{11}$$

where $\overline{\mathbf{y}} \equiv (\overline{y}_1, \ldots, \overline{y}_N)^T = W\mathbf{y}$, $y_i \geq 0$ for all $i \in \mathcal{L}$, and

$$a(\mathbf{y}) = \sum_{\{\mathbf{y}'\}} w(\mathbf{y}', \mathbf{y}) = r \sum_{i \in \mathcal{L}} \overline{y}_i f_i(y_i). \tag{12}$$

Notice that $a(\mathbf{y}) = 0$ if $y_i = k_i < \infty$ for all $i \in \mathcal{L}$, since in this case no transition can occur anymore. The special form of w allows us to write the first term in the Master equation more explicitly as

$$\sum_{\{\mathbf{y}'\}} w(\mathbf{y}, \mathbf{y}') p(\mathbf{y}', t) = r \sum_{i \in \mathcal{L}} \overline{(\mathbf{y} - \mathbf{e}_i)}_i f_i(y_i - 1) p(\mathbf{y} - \mathbf{e}_i, t), \tag{13}$$

where $\mathbf{e}_i$ is the i-th "coordinate vector", with components $e_{ij} = \delta(i, j)$, so that the vector $\mathbf{y}' = \mathbf{y} - \mathbf{e}_i$ has the components $y_j' = y_j$ if $j \neq i$ and $y_i' = y_i - 1$, and we set $p(\mathbf{y}, t) = 0$ if $y_i < 0$ or $y_i > k_i$ for some $i \in \mathcal{L}$.

We are particularly interested in the probability

$$p_j(t) \equiv \Pr[Y_j(t) = 0] = \sum_{\{\mathbf{y} | y_j = 0\}} p(\mathbf{y}, t), \tag{14}$$

where here the sum extends over all realizations $\mathbf{y}$ with $y_j = 0$. The evolution equation for p_j is found from the Master equation. Applying the sum in (14) to the first term in (9) using

(13) yields

$$\begin{aligned}\sum_{\{\mathbf{y}|y_j=0\}}\sum_{\{\mathbf{y}'\}} w(\mathbf{y},\mathbf{y}')p(\mathbf{y}',t) &= r\sum_{i\neq j}\sum_{y=1}^{k_i}\sum_{\{\mathbf{y}|y_j=0,y_i=y\}} f_i(y-1)\overline{(\mathbf{y}-\mathbf{e}_i)}_i p(\mathbf{y}-\mathbf{e}_i,t)\\ &= r\sum_{i\neq j}\sum_{y=0}^{k_i-1}\sum_{\{\mathbf{y}|y_j=0,y_i=y\}} f_i(y)\overline{y}_i p(\mathbf{y},t), \qquad (15)\end{aligned}$$

and the second term becomes

$$\sum_{\{\mathbf{y}|y_j=0\}} a(\mathbf{y})p(\mathbf{y},t) = r\sum_{i\neq j}\sum_{y=0}^{k_i-1}\sum_{\{\mathbf{y}|y_j=0,y_i=y\}} f_i(y)\overline{y}_i p(\mathbf{y},t) + r\sum_{\{\mathbf{y}|y_j=0\}}\overline{y}_j p(\mathbf{y},t). \qquad (16)$$

Noting that if $y_j = 0$, then $p(\mathbf{y},t)$ is the joint probability

$$p(\mathbf{y},t) = \Pr[\mathbf{Y}_j(t)=\mathbf{y}_j,\, Y_j(t)=0] = \Pr[\mathbf{Y}_j(t)=\mathbf{y}_j \,|\, Y_j(t)=0]p_j(t),$$

where $\mathbf{y}_j = (y_1,\ldots,y_{j-1},y_{j+1},\ldots,y_N)$ and analogously $\mathbf{Y}_j(t)$, the sum in the second term in (16) can be written as

$$\sum_{\{\mathbf{y}|y_j=0\}}\overline{y}_j p(\mathbf{y},t) = \sum_{\{\mathbf{y}|y_j=0\}}\overline{y}_j \Pr[\mathbf{Y}_j(t)=\mathbf{y}_j \,|\, Y_j(t)=0]p_j(t) = \mathrm{E}[\overline{Y}_j(t) \,|\, Y_j(t)=0]p_j(t),$$

where E denotes the expectation value. Thus the evolution equation for $p_j(t)$ which results upon subtracting (16) from (15) takes the form

$$\dot{p}_j(t) = -rp_j(t)\mathrm{E}[\overline{Y}_j(t) \,|\, Y_j(t)=0]. \qquad (17)$$

A quantity characteristic for the spread of the species is the complementary probability $u_j(t) = 1 - p_j(t) = \Pr[Y_j(t) > 0]$. This probability satisfies the evolution equation

$$\dot{u}_j(t) = r(1 - u_j(t))\mathrm{E}[\overline{Y}_j(t) \,|\, Y_j(t) = 0]. \tag{18}$$

In the next subsection we set up approximations and extensions for the expectation value in (18) for appropriate one- and two-dimensional continuum limits. Further extensions aimed specifically at heterogenous landscapes as required by niche models will be discussed in Section 3.3.

3.2.2. Spatially Continuous Model. We now generalize the discrete model from the preceding subsection by replacing $Y_i(t)$ with $Y(x,t)$, where $x \in \mathcal{L}$ and $\mathcal{L}$ is a measurable location space with measure μ [46]. Specifically we consider one-dimensional or two-dimensional location spaces, $\mathcal{L} = \mathbb{R}$ or $\mathcal{L} = \mathbb{R}^2$, with the standard Lebesgue-measure. In this case, the contact birth process is given by

$$\lim_{dt\to 0}(\frac{1}{dt} \cdot \Pr[Y(x,t) \to Y(x,t) + 1 \text{ in } (t, t+dt] \,]) = r\bar{Y}(x,t)f(Y(x,t),x), \tag{19}$$

where $\bar{Y}$ is given by the integral

$$\bar{Y}(x,t) = \int_{\mathcal{L}} Y(y,t)w(x,y)d\mu(y)$$

with weight function w which gives the probability density that an individual located at point y gives birth to a new individual at location x. The function w is assumed to have similar properties as W in subsection 3.2.1. If w is symmetric and constant with respect to y, $\bar{Y}$ can be taken to be a convolution, as in [46].

The function $f(Y, x)$ is also defined as in subsection 3.2.1 (usually a crowding function), so that Equation (19) is a non-decreasing, spatiotemporal process in which death plays no significant role. Such a process is well suited for modeling biological invasions in which there is no active resistance to species takeover, since we can expect that any individuals dying of natural causes will be replaced in successive generations. For heterogeneous landscapes the carrying capacity, $K(x)$, varies over $x \in \mathcal{L}$ and we assume $f(K(x), x) = 0$ as in subsection 3.2.1. Mollison has pointed out that without a crowding function $f(Y, x)$, this process results in exponential growth, as expected [46]. For homogeneous landscapes, K is constant and f depends only on Y.

It remains to interpret $x \in \mathcal{L}$ under the assumption that $\mathcal{L}$ is a continuous space. For the purpose of illustration, suppose that we wish to take $\mathcal{L} = \mathbb{R}^2$. Begin by taking a uniform square grid in $\mathbb{R}^2$ with cell dimensions given by Δx_1 and Δx_2. We can then arrange for a discretization of the space, $\mathcal{D}$, to be the set of all cells in this grid, the elements of which are then represented by the $\mathbb{R}^2$ value of the midpoints from each cell. We now have that

$$\begin{aligned}
&\lim_{dt\to 0}(\frac{1}{dt}\cdot \Pr[Y(x,t) \to Y(x,t)+1 \text{ in } (t, t+dt] \text{ and } x \in \mathcal{D}\ ]) \\
&= \lim_{dt\to 0}(\frac{1}{dt}\cdot \Pr[Y(\vec{x},t) \to Y(\vec{x},t)+1 \text{ in } (t, t+dt], \\
&\text{with } \vec{x} \in ((x_1 - \frac{\Delta x_1}{2},\ x_1 + \frac{\Delta x_1}{2}],\ (x_2 - \frac{\Delta x_2}{2},\ x_2 + \frac{\Delta x_2}{2}])\ ]) \\
&= r\bar{Y} f(Y(x,t)) \text{ for } x \in \mathcal{D},
\end{aligned}$$

where

$$\bar{Y}|_{x\in\mathcal{D}}(x) = \int_{\mathcal{D}} Y(y,t)w(x,y)d\mu(y) = \sum_{y\in\mathcal{D}} Y(y,t)w(x,y)\Delta x_1 \Delta x_2,$$

and $Y(y,t)$ is understood to be the population in the cell $((y_1-\frac{\Delta x_1}{2},\ t_1+\frac{\Delta x_1}{2}],\ (y_2-\frac{\Delta x_2}{2},\ y_2+\frac{\Delta x_2}{2}])$. But in this formulation, the choice of grid and therefore $\mathcal{D}$ was completely arbitrary. If we redefine $Y(x,t)$ to be the population in the rectangle $((x_1-\frac{\Delta x_1}{2},\ x_1+\frac{\Delta x_1}{2}],\ (x_2-\frac{\Delta x_2}{2},\ x_2+\frac{\Delta x_2}{2}])$ rather than *at* a location x in some space, then Equation (19) is valid for the continuous spatial variable x, with $\bar{Y}$ resulting from integration over $\mathbb{R}^2$. We will therefore assume that $x \in \mathbb{R}^2$ for the remainder of our discussion.

Let $u(x,t)$ be defined by $\Pr[Y(x,t) > 0]$. If we assume that Y obeys a process that matches Equation (19), then the equation for $\partial u(x,t)/\partial t$ follows form (18) by replacing the discrete variable j by x,

$$\frac{\partial u(x,t)}{\partial t} = rE[\bar{Y}(x,t)|Y(x,t)=0](1-u), \tag{20}$$

with

$$E[\bar{Y}(x,t)|Y(x,t)=0] = \int_{\mathcal{L}} E[Y(y,t)|Y(x,t)=0]w(x,y)d\mu(y). \tag{21}$$

It is generally not possible to explicitly determine $E[\bar{Y}(x,t)|Y(x,t)=0]$ from knowledge of $\Pr[Y(x,t)>0]$ only. Accordingly, we seek an approximation to this conditional expectation that will allow us to effectively model the presence probability. According to the integral representation (21), we can reduce the problem to that of approximating $E[Y(y,t)|Y(x,t)=0]$ at all locations y, after which we can take the convolution.

Different spatial values of Y are correlated to the extent that they are connected via the weight function, and since we wish to convolute the resulting expected value with this same weight function, the correlated values of Y are the most important for approximating Equation (20). Unfortunately, we have no information about the joint expected value

$E[Y(y,t), Y(x,t) = 0]$ or the joint distribution of $Y(y,t)$ and $Y(x,t)$. Treating Y as a random vector, it may be possible to look at the interdependencies of each component based on the transition probability and estimate the joint expected value, but in this paper, we will only derive rough estimates for $E[Y(y,t)|Y(x,t) = 0]$ in terms of $u(y,t)$.

It is clear that $E[Y(y,t)|Y(x,t) = 0] \leq E[Y(y,t)]$, and it may be expected that $E[Y(y,t)|Y(x,t) = 0]$ evolves similarly to $E[Y(y,t)]$ in response to different values of $u(y,t)$. The correlation between a spatially fixed stochastic variable $Y(x,t)$ and every other spatial component of Y is complicated and heavily dependent upon spatial distance and the choice of convolution kernel. In general, one cannot say that $E[Y'|Y = y] \approx E[Y']$ given two stochastic variables Y' and Y, particularly if the two variables are strongly correlated, but for kernels with relatively small standard deviation, the spatial region of high correlation among the stochastic variables $Y(x,t)$ should be relatively small. In the extreme case where the convolution kernel is a delta function, we have $E[Y(y,t)|Y(x,t) = 0] = E[Y(y,t)]$. Therefore, we will attempt to find an approximation for $E[Y(y,t)]$ through $u(y,t)$ for all locations $y \in \mathcal{L}$ and then replace $E[Y(y,t)|Y(x,t) = 0]$ by this approximation in (21), so that Equation (20) is approximated by a closed evolution equation for $u(x,t)$. Expecting that these two approximations, namely i) replacing $E[Y(y,t)|Y(x,t) = 0]$ by $E[Y(y,t)]$ and ii) approximating $E[Y(y,t)]$ for every y by a function of $u(y,t)$, results in an upper bound on the rate of change $u_t(x,t)$, we will then solve the resulting equation for $u(x,t)$ and compare with pseudo-random realizations for Y obtained from the corresponding stochastic birth process, correcting as needed. This method of approximation is naturally imperfect, but must suffice in abeyance of a further investigation which provides a better method of approximation. In the following we describe and motivate our approach to finding the

approximation of $E[Y(x,t)]$ as function of $u(x,t)$ for the case of a homogeneous landscape, where K is constant and $f(Y,x) = f(Y)$.

In the case $K = 1$, it is evident that $E[Y(x,t)] = u(x,t)$. At the other extreme, consider the limit $K = \infty$; that is, there is no carrying capacity limit to population size. We can then formally find a first-order approximation for $E[Y(x,t)]$ by introducing a new parameter $\lambda(x,t_0)$, defined to be the average rate of change for $Y(x,t)$ between times $t = 0$ and a fixed time $t = t_0$. Let $Y_{avg}(x,t)$ be a stochastic function that evolves according to the fixed rates given by $\lambda(x)$ until the time t_0. That is,

$$\lim_{d\tau\to 0}(\frac{1}{d\tau}\cdot \Pr[Y_{avg}(x,\tau) \to Y_{avg}(x,\tau)+1 \text{ in } (\tau,\tau+d\tau]\ ]) = \lambda(x), \qquad \tau \le t_0.$$

For each x, the event $Y_{avg}(x) \to Y_{avg}(x)+1$ occurs with a fixed average rate and independently of the time since the last event. Thus, by definition, $Y_{avg}(x,t)$ is Poisson distributed with parameter $\lambda(x)$ for each $x \in \mathbb{R}^2$. In essence then, we approximate the local stochastic process by a Poisson process as it is the simplest relevant stochastic process. Since we have that $P[Y(x,t) = 0] = 1 - u(x,t)$, we can now invert the probability mass function of $Y_{avg}(x,t)$ and solve for $\lambda(x)$:

$$\lambda(x) = -\ln(1-u(x)),$$

which in fact tells us that

$$E[Y(x,t)] \approx E[Y_{avg}(x,t=t_0)] = -\ln(1-u(x,t)). \tag{22}$$

3.2.2.1. *Two-parameter model.* For ecological reasons, we are naturally most interested in the situation where Y does have a finite carrying capacity, typically much larger than 1.

Interpolating between $E[Y] = u$ at $K = 1$ and $E[Y] \approx -\ln(1-u)$ at $K = \infty$, we take $E[Y] \approx -\alpha \ln(1-u) + \beta u$, where the parameters α and β are functions of the carrying capacity K. Setting $E[Y(y,t)|Y(x,t) = 0] \approx E[Y(y,t)] \approx -\alpha \ln(1-u(y,t)) + \beta u(y,t)$, the governing equation for $u(x,t)$ reads

$$\frac{\partial u(x,t)}{\partial t} = r\left[J(u(y)) * w(y)\right](1-u), \tag{23}$$

where

$$J(u) = -\alpha(K)\ln(1-u) + \beta(K)u. \tag{24}$$

We determine the functions $\alpha(K)$ and $\beta(K)$ in (24) by comparing numerical simulations of Equation (23) with stochastic results for Y obtained through pseudo-random realizations of the transition probability. For varying values of K, we compute the absolute error

$$\frac{\int_{[0,t_{end}]} ||u(t) - Pr[Y(t) > 0]||_1 \ \phi \ dt}{\int_{[0,t_{end}]} \phi \ dt}, \qquad \phi = \begin{cases} 2 & \text{if } t \leq 6 \\ 2 - (t-6)/24 & \text{if } 6 < t \leq 30 \\ 1 & \text{if } t > 30 \end{cases} \tag{25}$$

for a range of values of α and β and find the values of these parameters that minimize this error. We use a weighted absolute error in contrast to an L_∞ error because our equations are non-local. Moreover, absolute error captures deviations in both wave speed and wave shape. The choice of the weight function $\phi(t)$ gives larger weight to errors occurring at smaller values of time ($t \leq 6$), where the choice of 6 and 24 were arbitrary choices to represent short and long time periods. We choose this weighting to give priority to short-term ecological

forecasts rather than long-time forecasts that would in any case be more subject to error from changing environmental conditions. For these comparisons, we made the natural choice of $f(Y) = (1 - Y/K)$ in Equation (19) for a version of Mollison's "simple epidemic model" [46]. As this is by far the most common choice of $f(Y)$ in the literature, for the rest of this section we will continue to assume that

$$\lim_{dt \to 0} \left(\frac{1}{dt} \cdot \Pr[Y(x,t) \to Y(x,t) + 1 \text{ in } (t, t+dt] \,] \right) = r\bar{Y}(x,t)(1 - Y(x,t)/K) \tag{26}$$

is the underlying stochastic process for some large, discrete choice of $K > 1$.

In our simulations, we chose a variety of values for K ranging from 25 to 1000. For the weight function w in the convolution, we examined both a normal distribution

$$N(x;\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

and a Laplace distribution

$$L(x;b) = \frac{1}{2b} \mathrm{e}^{\frac{-|x|}{b}},$$

because the Laplace distribution exhibits heavy tails that are exponentially bounded and is one of the suggested distribution kernels for an invading organism given by Kot et al. (1996) [44]. Theoretically, any distribution kernel could be used, though Mollison [56] has shown that distributions without negative-exponentially bounded tails result in wave velocities that tend toward infinity in the limit for all non-trivial initial conditions. Initial conditions for pseudo-random realizations of Y were then chosen to be a Heaviside function, scaled such that $Y = K$ where $Y > 0$. We have generated 1000 pseudo-random realizations of the stochastic process until $t_s = 30$ (to disperse transient effects due to the initial conditions),

and $\Pr[Y(x,t) > 0]$ was estimated from the resulting data. This value for $\Pr[Y(x,t) > 0]$ was then used as the initial condition for our model ($t = 0$). We then continued to generate 1000 pseudo-random realizations of the stochastic process until $t_s = 60$ ($t = 30$). These results were again used to estimate $\Pr[Y(x,t) > 0]$, but this time for direct comparison with the output of the model as it is run from $t = 0$ to $t = 30$.

The optimal values of the parameters α and β for varying values of K are shown in Table 3.1 for the normal distribution and in Table 3.2 for the Laplace distribution. In the case of the normal distribution, for each of the three choices of the standard deviation given in Table 3.1, α (respectively β) is a generally increasing (respectively decreasing) function of K. In contrast, β increases with K in the case of the Laplace distribution, whereas α remains small. Values of α and β for more values of K are included in Tables 3.3 (for the normal distribution with $\sigma = 1$) and 3.4 (for the Laplace distribution with $b = 1$) of Section 3.2.3, where we discuss the wavespeed.

TABLE 3.1. Parameters α and β for the normal distribution kernel $w = N(u;\sigma)$.

K	$\sigma = 0.5$	α	β	$\sigma = 1$	α	β	$\sigma = 2$	α	β
100		0.000	0.82		0.000	0.82		0.000	0.86
300		0.040	0.86		0.040	0.86		0.080	0.84
500		0.120	0.72		0.120	0.72		0.190	0.72
700		0.130	0.81		0.130	0.81		0.350	0.60
900		0.260	0.55		0.260	0.55		0.340	0.63

TABLE 3.2. Parameters α and β for the Laplace distribution kernel $w = L(u;b)$.

K	$b = 0.5$	α	β	$b = 1$	α	β	$b = 2$	α	β
100		0.005	0.67		0.005	0.81		0.025	0.82
300		0.005	0.83		0.010	0.83		0.125	0.72
500		0.005	0.87		0.030	0.76		0.035	0.94
700		0.010	0.88		0.005	0.91		0.120	0.82
900		0.010	0.88		0.005	0.92		0.005	0.97

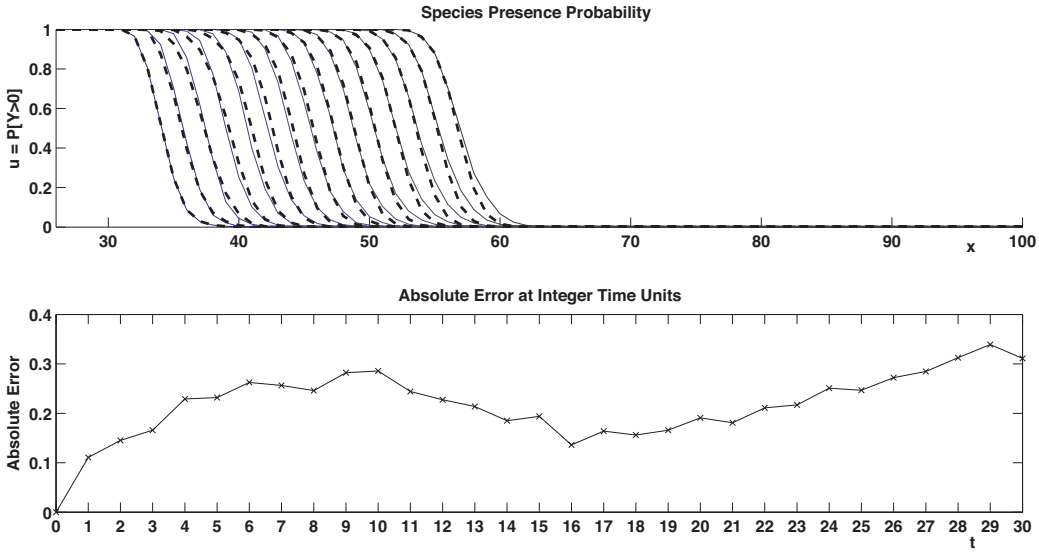

FIGURE 3.1. (a) Solution to Equation (23) plotted at even integer values of time up to time 28 (dashed lines) plotted along with $\Pr[Y > 0]$ based on data from 1000 stochastic realizations of the transition probability (solid lines). The distribution kernel is the normal distribution with $\sigma = 0.5$, and the carrying capacity is $K = 700$. (b) Absolute error as a function of time.

Results of numerical integrations of Equation (23) with the normal distribution and varying values of σ are shown in Figs. 3.1 ($\sigma = 0.5$), 3.2 ($\sigma = 1$), and 3.3 ($\sigma = 2$) for values of α and β determined by $K = 700$. The figures show good qualitative and quantitative agreement between the model and the stochastic computation, but for increasing t the leading edge of the wavefront of the model (smallest value of x for which $u = 0$) tends to lag behind the leading edge of the stochastic wavefront. The wavespeed increases with increasing σ, consistently with the discussion of wavespeed in Section 3.2.3.

Results for the Laplace distribution with $b = 1$ are shown in Fig. 3.4 ($K = 300$) and Fig. 3.5 ($K = 700$). The wavespeed is only slightly larger for $K = 700$, again consistently with calculations in Section 2.3. In all simulations, the absolute error is less than 1 for all time.

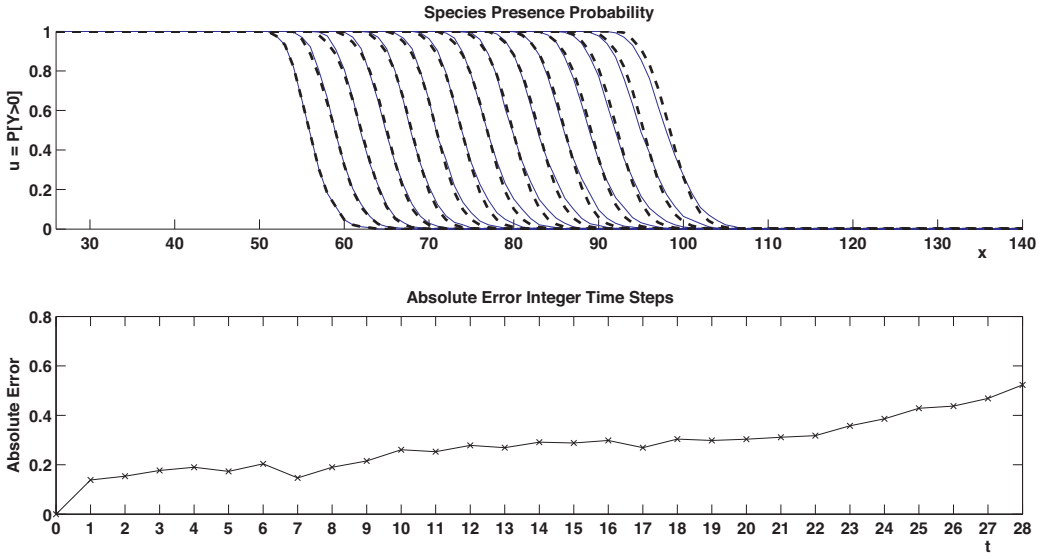

FIGURE 3.2. (a) Solution to Equation (23) plotted at even integer values of time up to time 28 (dashed lines) plotted along with $\Pr[Y > 0]$ based on data from 1000 stochastic realizations of the transition probability (solid lines). The distribution kernel is the normal distribution with $\sigma = 1$, and the carrying capacity is $K = 700$. (b) Absolute error as a function of time.

In our model, it is also necessary to specifically deal with the case where $u(y) = 1$, since $J(1)$ does not exist. Analytically, this singularity reflects the fact that we may be sure that $Y > 0$ is in a given location, but without further information about the non-local time evolution of Y, we cannot know $E[Y(x,t)|Y(x,t) > 0]$. To fix this problem, we assume that $E[Y(x,t)|Y(x,t) > 0] \approx K/2$ based on the assumption that Y grows quickly once $Y > 0$ but will not achieve carrying capacity for a significant amount of time while still in the support of $w(x)$ centered around locations with $u(x) < 1$. Numerical observations also suggest that the choice of constant does not play a significant role, so the choice of $K/2$ can also be considered as essentially arbitrary. Setting $E[Y|Y > 0]$ to a constant also fixes a key numerical concern that $J(u)$ is stiff - numerical computation can be achieved much more quickly by taking, for example, $J(1) = K/2$.

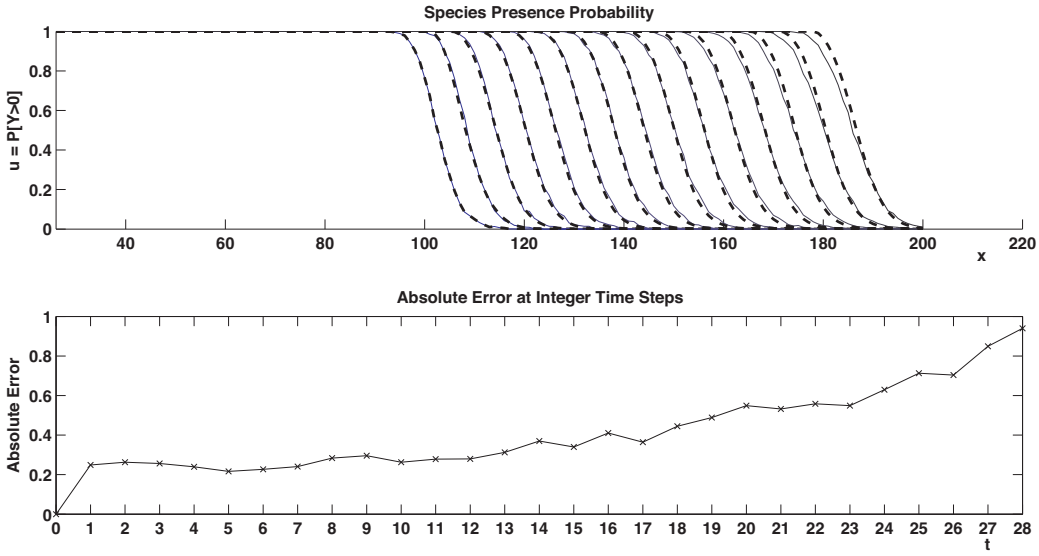

FIGURE 3.3. (a) Solution to Equation (23) plotted at even integer values of time up to time 28 (dashed lines) plotted along with $\Pr[Y > 0]$ based on data from 1000 stochastic realizations of the transition probability (solid lines). The distribution kernel is the normal distribution with $\sigma = 2$, and the carrying capacity is $K = 700$. (b) Absolute error as a function of time.

3.2.2.2. *Initial conditions and one-parameter model.* Having established parameters α and β for the two-parameter model $J(u)$ given by (24) across different choices of K and dispersal kernel w, we now consider more closely the problem of initial conditions for the model. In the field, we expect data to be given for $u(x,t)$ at time $t = 0$, but since the stochastic variable $Y(x,0)$ is unknown, it is impossible to know with any accuracy $E[Y(x,0) > 0]$. Instead, we will only know some places where $Y(x,0) > 0$, some places where $Y(x,0) = 0$, and perhaps some uncertainty in the remaining area of the form of $0 \leq u(x,0) \leq 1$. If we have full knowledge of $Y(x,0)$ in the study area, $u(x,0)$ would take on only values of 0 and 1. For simplicity, we will assume this is the case and examine Heaviside initial conditions $u(x,0)$.

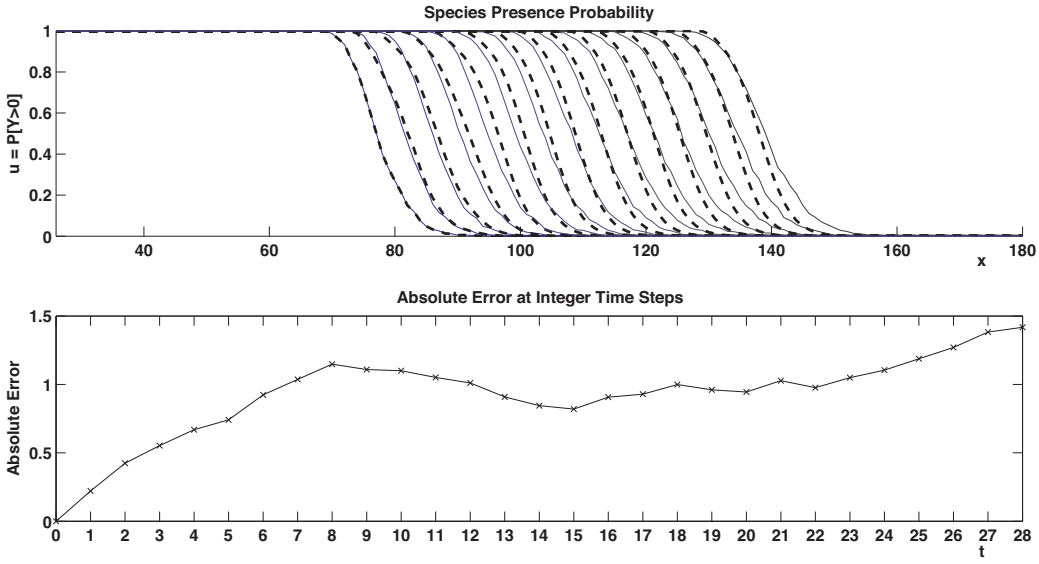

FIGURE 3.4. (a) Solution to Equation (23) plotted at even integer values of time up to time 28 (dashed lines) plotted along with $\Pr[Y > 0]$ based on data from 1000 stochastic realizations of the transition probability (solid lines). The distribution kernel is the Laplace distribution with $b = 1$, and the carrying capacity is $K = 300$. (b) Absolute error as a function of time.

Given that $u(x,t) = 1$ in some location, we must again confront the fact that $Y(x,t)$ is unknown when choosing values for $J(1)$. Since we are assuming the species represented by Y is an invader, we will assume that Y is initially large where $u(x) = 1$ and continue with our above approximation $E[Y|Y > 0] \approx K/2$. Even with this approximation, we need a further alteration of $J(u)$ for $u < 1$, since the model is critically affected by large concentrations of $E[Y]$ nearby locations where we have set $u(x,0) = 0$. In this case, the $K = \infty$ approximation of Equation (22) provides a better model. We therefore introduce in addition to the general two-parameter model (24) the model

$$J_H(u) = -\gamma \ln(1-u), \tag{27}$$

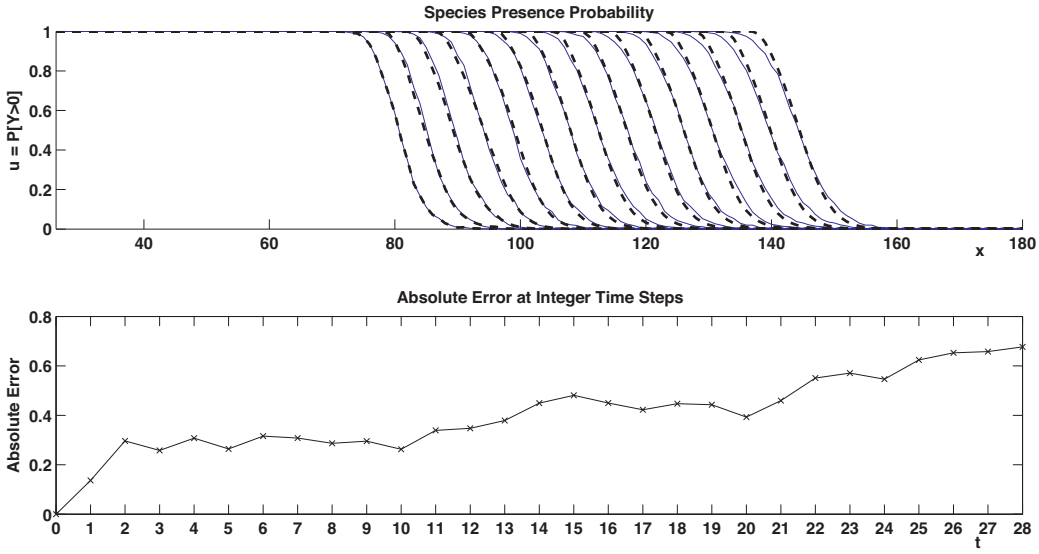

FIGURE 3.5. (a) Solution to Equation (23) plotted at even integer values of time up to time 28 (dashed lines) plotted along with $\Pr[Y > 0]$ based on data from 1000 stochastic realizations of the transition probability (solid lines). The distribution kernel is the Laplace distribution with $b = 1$, and the carrying capacity is $K = 700$. (b) Absolute error as a function of time.

where γ is determined by numerical comparison of the deterministic equation with stochastic simulations. This model serves to deal particularly with the issues of Heaviside initial conditions and is used in an initial time interval.

3.2.2.3. *Combined model.* To test the deterministic model for long time periods based on Heaviside initial conditions, we introduce a time-dependent homotopy between the one parameter model $J_H(u)$, Equation (27), and the general two-parameter model $J(u)$ of Equation (24). A linear homotopy works well, but notable improvement is achieved by perturbing the linear homotopy function with a polynomial so that, relative to the linear homotopy, higher weight is given to the model $J_H(u)$ at smaller values of time. The homotopy we use is thus

$$-(1 - h(\tau)) \ln(1 - u) + h(\tau)(-\alpha \ln(1 - u) + \beta u),$$

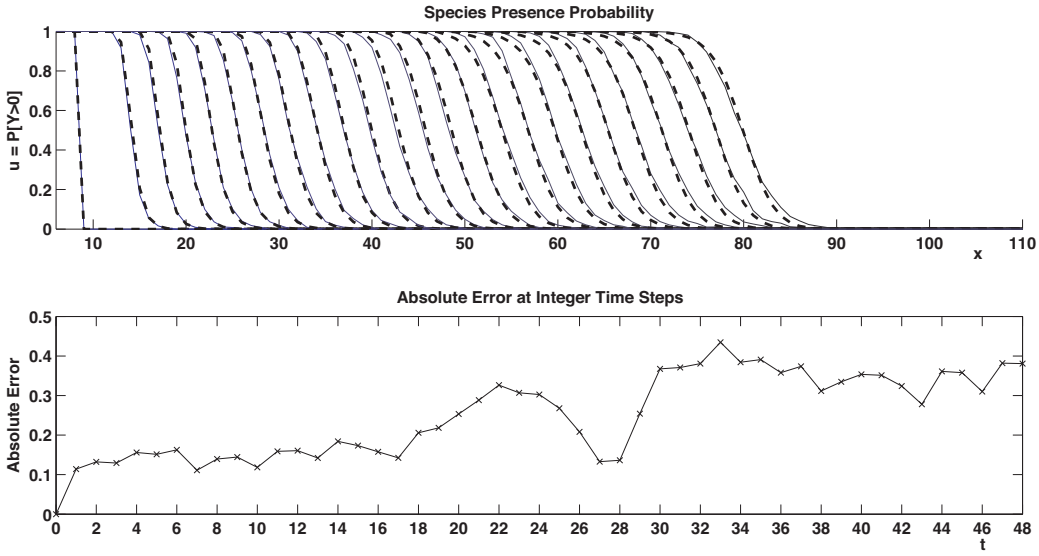

FIGURE 3.6. (a) Solution to Equation (23) plotted at even integer values of time up to time 48 (dashed lines) plotted along with $\Pr[Y > 0]$ based on data from 1000 stochastic realizations of the transition probability (solid lines). The distribution kernel is the normal distribution with $\sigma = 1$, and the carrying capacity is $K = 200$. (b) Absolute error as a function of time.

where $\tau = t/26$ (with a Heaviside initial condition, 26 time units is the approximate amount of time transients require to decay), $h(\tau) = \tau + a\tau(\tau - b)(\tau - 1)$ for $0 \leq \tau < 1$, and $h(\tau) = 1$ for $\tau \geq 1$. Figures 3.6 and 3.7 illustrate the accuracy possible when starting from Heaviside initial conditions for u with the correct temporal homotopy for $J(u)$. For $K = 200$ (Fig. 3.6), the homotopy parameters $a = 2.8$, $b = 0.44$ minimize the error, whereas for $K = 600$ (Fig. 3.7), the optimal homotopy parameters are $a = 3.1$, $b = 0.43$.

In considering the numerical size of the problem, it is immediately clear that we should begin by doing simulations in $\mathbb{R}$ rather than $\mathbb{R}^2$, since approximately 1000 stochastic realizations of Y are required for good estimations regarding $\Pr[Y(x,t) > 0]$, and each stochastic realization requires a fine time step, dt, to properly simulate the transition probability. We used a time step of $dt = 0.0001$ after finding that significantly greater values failed to produce

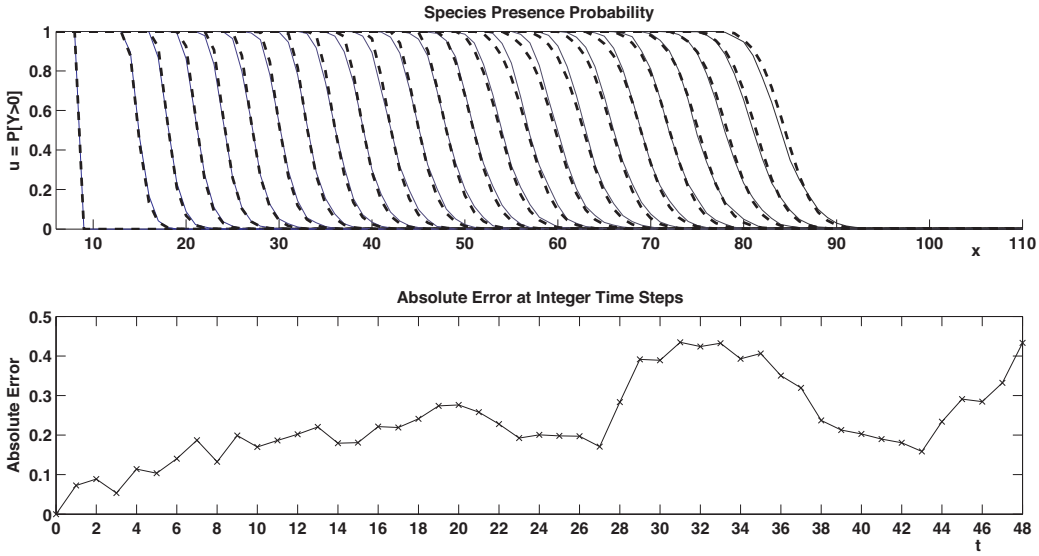

FIGURE 3.7. (a) Solution to Equation (23) plotted at even integer values of time up to time 48 (dashed lines) plotted along with $\Pr[Y > 0]$ based on data from 1000 stochastic realizations of the transition probability (solid lines). The distribution kernel is the normal distribution with $\sigma = 1$, and the carrying capacity is $K = 600$. (b) Absolute error as a function of time.

plausible results. We acquired our 1000 pseudo-random realizations for Y over 30 time units by running 18 realizations at a time in parallel, a process which took approximately 45 minutes with a 1-D spatial mesh 100 units long. Reasonably sized 2-D simulations could easily have over 100,000 spatial cells, highlighting the need for better than brute-force methods to acquire results about Y on standard desktop computers. In contrast, numerically solving our model for $u(x,t)$ takes between five seconds to 2 minutes.

In the plot of 30 time units on a larger domain, it becomes obvious that while the speed of the traveling wave is good, the shape is not quite correct. We are slightly over-estimating $u = \Pr[Y > 0]$ when $u > 0.5$, and slightly under-estimating when $u < 0.5$. This error is likely due to our use of $E[Y]$ in place of $E[Y|Y(x,t) = 0]$, but it is predictable.

3.2.3. WAVESPEED. Studies of Equation (7) and other nonlinear equations for population spread have revealed that, although nonlinear terms affect the shape of traveling wave solutions, the speed of traveling wave solutions of the nonlinear equations match the speeds of solutions to linearizations of the equations [47, 57]. Consistent with the conjecture that this should hold in general for equations governing population spread [57], numerical simulations of traveling wave solutions to our Equation (23) and its linearization

$$\frac{\partial u(x,t)}{\partial t} = r\hat{J}(u(y)) * w(y), \quad \hat{J}(u(y)) = (\alpha+\beta)u(y), \tag{28}$$

about $u = 0$ yield matching wave speeds for both normal and Laplace distributions. The linearization (28) is, up to a change in parameters, also the linearization of the equation

$$\frac{\partial u}{\partial t} = r\bar{u}(K-u) \tag{29}$$

about $u = 0$, as studied by Medlock and Kot (2003) [47], who compute wavespeeds for Equation (29) by making the traveling-wave ansatz $u(x,t) = A\mathrm{e}^{-\theta(x-ct)}$ in the linearized equation for various distribution kernels. These authors obtain wavespeeds of

$$c = \sqrt{\mathrm{e}}\sigma r\kappa \tag{30}$$

for the normal distribution kernel $N(v;\sigma)$ and

$$c = \pm\frac{3\sqrt{3}}{2}rb\kappa \tag{31}$$

for the Laplace distribution kernel $L(v;b)$, where κ is the carrying capacity K in the case of Equation (29), which translates into $\kappa = \alpha+\beta$ for our Equation (23). In Tables 3.3 and 3.4,

we give the analytical wavespeeds according to Equations (30) and (31) as well as wavespeeds calculated from numerical simulations of Equation (23). The match between analytically and numerically computed wavespeeds is fairly good, with potential error coming from the fact that the wave in the numerical calculation has not completely converged to the traveling wave, as well as from errors in numerically determining the wavespeed from profiles of u at a discrete set of times.

In contrast to the prediction given by Equations (30) and (31) with $\kappa = K$ that wavespeed is proportional to carrying capacity, our computations as given in Tables 3.3 and 3.4 suggest a nonlinear and less sensitive dependence of wavespeed on carrying capacity, a result of the nonlinear dependence of $\kappa = \alpha + \beta$ on K.

TABLE 3.3. Parameters α and β and analytical (as given by Equation (30) with $r = 1$) and numerical wave speeds for the normal distribution kernel $w = N(u; \sigma)$ with $\sigma = 1$.

K	α	β	analytical wavespeed	numerical wavespeed
100	0.000	0.82	1.352	1.3667
200	0.000	0.8	1.434	1.400
300	0.040	0.86	1.484	1.467
400	0.050	0.86	1.500	1.467
500	0.120	0.72	1.385	1.467
600	0.130	0.80	1.533	1.533
700	0.130	0.81	1.550	1.533
800	0.210	0.75	1.583	1.533
900	0.260	0.55	1.335	1.500
1000	0.160	0.71	1.434	1.500

3.3. HETEROGENEOUS LANDSCAPES

In the derivations of the previous section, we have always assumed that the landscape is homogeneous in terms of suitability for the species in question. The model takes the entire domain to be completely fertile, and no obstacles exist for establishment. We now relax this

TABLE 3.4. Parameters α and β and analytical (as given by Equation (31)) and numerical wave speeds for the Laplace distribution kernel $w = L(u; b)$ with $b = 1$.

K	α	β	analytical wavespeed	numerical wavespeed
100	0.005	0.81	2.117	2.067
200	0.005	0.85	2.221	2.167
300	0.010	0.83	2.182	2.167
400	0.010	0.85	2.234	2.233
500	0.030	0.76	2.052	2.200
600	0.005	0.93	2.429	2.300
700	0.005	0.91	2.377	2.267
800	0.005	0.94	2.455	2.300
900	0.005	0.92	2.403	2.333
1000	0.005	0.93	2.429	2.300

assumption and derive a more realistic model that takes into account available ecological data about the species, terrain, and climate.

On the surface, the challenge of creating such a model is a daunting task. Ecological data is often species- and location-dependent, both in availability and importance, and scientists will disagree over which variables should be prioritized for inclusion into the model. Furthermore, lack of data is a constant problem in ecology, and as new data are acquired, our model should be flexible to accommodate changes in ecological understanding.

As mentioned in the introduction, statistical software packages that model potential species distribution have seen much success in recent years. These programs are of particular interest to our efforts because they have the capacity to crunch a wide variety of relevant data sets to return a map of "suitability" for the species. We interpret this suitability to mean the probability that a location can support the species in question, and we will use the output of such models to parameterize our model. By outsourcing the problem of interpreting ecological data to these statistical packages, we can then concentrate on how to best incorporate the suitability of the terrain into our model. In addition, since landscape

suitability can be a function of time, we also gain the ability to model spread under changing environmental conditions by having the statistical packages predict suitability under different assumptions for temporal change. Such methods can then be used to model expected species distribution under various climate change and management scenarios. Indeed, this approach is not wholly new; recent years have seen at least a couple of numerical, process based models utilize the output from these statistical packages for similar purposes [58, 59].

We start again with the contact birth process given by Mollison (1977) [46] and assume that suitability data is defined on a discrete spatial domain at some coarse scale compared to the size of an individual from the species in question. We expect this suitability data to be a number between 0 and 1 (inclusive) which gives the probability of establishment for the species, given that its propagules have access to that location. These assumptions are reasonable for the ecological niche modeling package Maxent [6] which is a popular choice among such statistical packages, though we also expect that these assumptions are valid for other choices.

Since suitability data is coarse and discrete, we assume that within a region on which it is constant there are only areas (size dx^2 in $\mathbb{R}^2$) of binary suitability. Thus, randomly picking a location x in the domain, one finds either that x can support the species, or it cannot. We can then interpret suitability data to be the probability that a randomly chosen location x in some suitability cell will be able to support the species, that is, $\Pr[x \text{ is suitable}]$. Given that the probability that $Y(x,t) \to Y(x,t)+1$ on a location x with suitability zero is equal to zero, we examine the transition probability

$$\lim_{dt\to 0}\left(\frac{1}{dt}\Pr[Y(x,t)\to Y(x,t)+1 \text{ in } (t,t+dt] \text{ and a seed will grow at } x]\right). \tag{32}$$

Let $s(x) = \Pr[x \text{ is suitable}]$ be a function to be defined by data. Since the previous transition probability in (19) could be interpreted as acting on a spatial domain where suitability is identically 1 everywhere, we can reinterpret (19) to be the conditional probability

$$\lim_{dt\to 0}(\frac{1}{dt}\Pr[Y(x,t) \to Y(x,t)+1 \text{ in } (t,t+dt] \mid \text{a seed will grow at } x])$$

As a result, the transition probability in Equation (32) is equal to

$$\begin{aligned}
&\lim_{dt\to 0}(\frac{1}{dt}\Pr[Y(x,t) \to Y(x,t)+1 \text{ in } (t,t+dt] \text{ and a seed will grow at } x]) \\
&= \lim_{dt\to 0}(\frac{1}{dt}\Pr[Y(x,t) \to Y(x,t)+1 \text{ in } (t,t+dt] \mid \text{a seed will grow at } x])\Pr[x \text{ is suitable}] \\
&= r \cdot s(x)\bar{Y}f(Y).
\end{aligned}$$

Here, we will generally assume that $f(Y) = (1 - Y/(Ks(x)))$, since we would expect the carrying capacity to change linearly with an area's suitability. Since $s(x)$ effectively adds only a spatial component to the constants r and K, we can now set $u(x,t) \equiv \Pr[Y(x,t) > 0]$ and repeat the process described in Section 3.2 to arrive at the model

$$\frac{\partial u(x,t)}{\partial t} = rs(x)\left[J(u(y,t)) * w(y)\right](1 - u(x,t)), \tag{33}$$

where $J(u(y,t)) = -\alpha\left(Ks(y)\right)\ln(1 - u(y,t))) + \beta\left(Ks(y)\right)u(y,t)$.

When considering initial conditions, we must be more careful in heterogeneous terrains. Instead of taking $u(x,0) = 1$ whenever $Y(x,0) > K$ as in Section 3.2.2, we should instead take $u(x,0) = 1$ when $Y(x,0) > Ks(x)$, in order not to underestimate the size of initial populations in areas of relatively low suitability. While this requirement means that we

must work somewhat harder to gather initial data in some environments, all other aspects of the calculation remain unchanged.

3.4. Simulations in two spatial dimensions

For all of the 2-D simulations, we used a carrying capacity of $K = 300$ individuals per cell. Lower values of K decrease the computation time, which is important when working with longer 2-D problems. We then ran the simulation for 15 time units on a 100-by-100 mesh using Runge-Kutta 4th order with variable time steps to solve the vectorized 2-D problem. No special information about the problem was supplied to our numerical solver.

In a first simulation, we chose the suitability function to be uniform random values between 0 and 1 over the domain. Initial conditions for the population $Y(\mathbf{x}, 0)$ were five 2-D bivariate normal distributions (mean $\mu = \mathbf{0}$ and covariance matrix $\Sigma = I$), which were translated into initial conditions for $u(\mathbf{x}, 0)$ by setting $u(\mathbf{x}, 0) = 1$ when $Y(\mathbf{x}, 0) > 0.15 \cdot K$, and then smoothing the edges with a standard Gaussian distribution ($u(\mathbf{x}, 0) = \mathrm{Z}(|\mathbf{x} - \mathbf{x}_0|)/\mathrm{Z}(0)$, where $\mathbf{x}_0$ is the closest location such that $Y(\mathbf{x}_0, 0) > 0.15 \cdot K$).

Since the simulation was run for a relatively short period of time, we used the one-parameter model function $J_H(u) = -\gamma \ln(1 - u)$ in the model, scaled by an appropriate factor γ to stand in for the decay of transients. This approach works well for time periods in the range of 15 time units or less and is robust to both the bivariate normal initial conditions above and the Heaviside initial conditions discussed earlier in this chapter. Suggested scaling factors γ for Laplace and Normal distributions are given in Table 3.5 for both 1-D and 2-D problems. For simplicity, we have assumed in the 2-D simulations that the variances in the x and y directions are equal, so that the parameters satisfy $\sigma_x = \sigma_y = \sigma$.

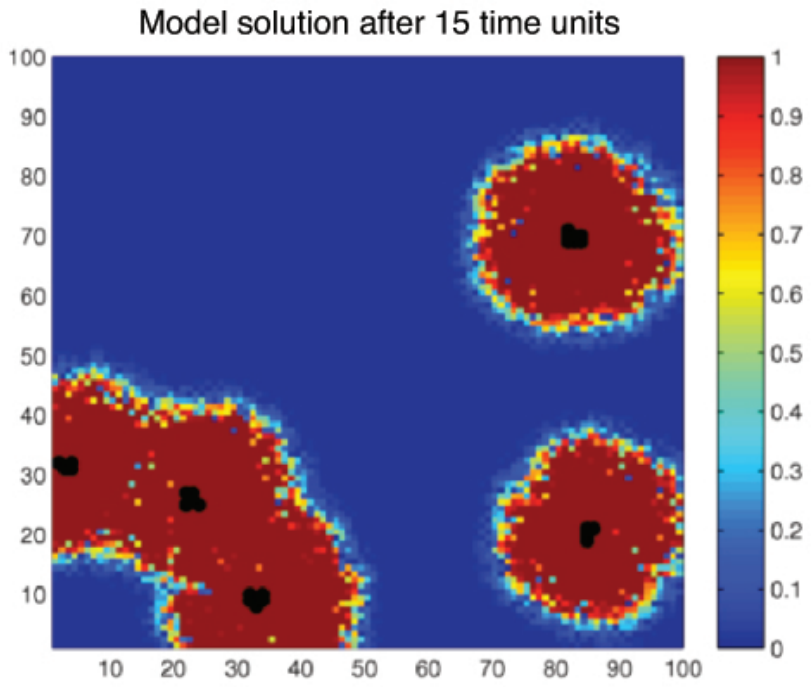

FIGURE 3.8. Solution to the model equation (33) after 15 time units with uniform random noise suitability. Initial locations where $u(\mathbf{x}, 0) = 1$ are shown in black.

The results of 2-D simulations are shown in Figures 3.8 through 3.11. The initial condition consisted of five patches where $u(x, 0) = 1$. The isotropy in the variance parameter σ and the homogeneous (random noise) suitability yield isotropic radial spreading of the population from patches where $u(x, 0) = 1$. The "ragged" edges, apparent in Fig. 3.8 and in the cutaway plot of Fig. 3.9 are due to random suitability. In general, as observed in this and other simulations, the shape of the leading edge strongly depends on the suitability function. Comparing Fig. 3.8, which shows $u(x, t = 15)$, with the graph of absolute error in Fig. 3.10, we note that the error is concentrated at the leading edge. Fig. 3.11 shows that the mean error does not increase with time.

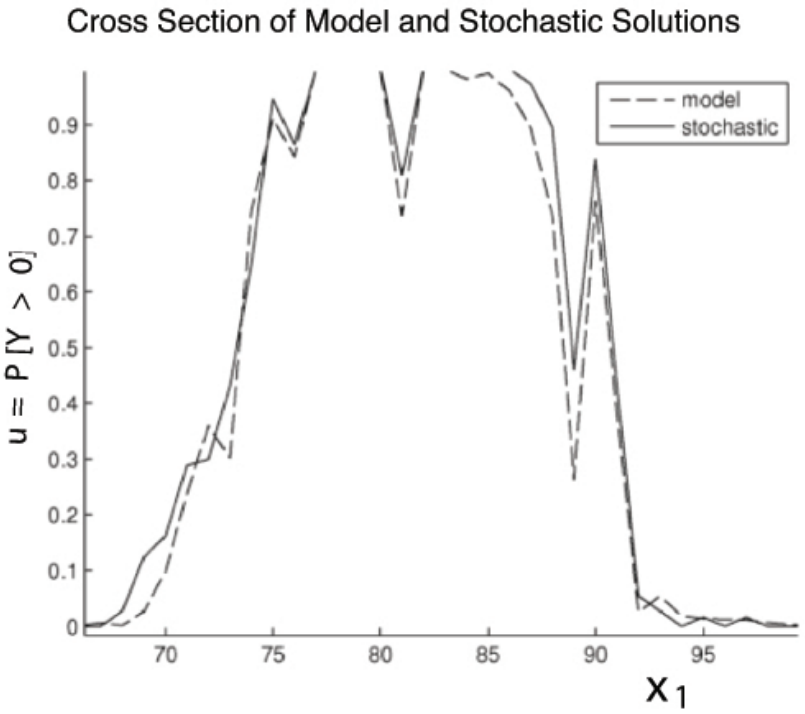

FIGURE 3.9. Cutaway view of solution to the model equation (33) (dashed line) and stochastic realization of $\Pr[Y(x_1, x_2, t) > 0]$ (solid line) at $x_2 = 80$ and time step 12. x_1 is given along the x-axis and $\Pr[Y(x_1, 80, 12)]$ is along the y-axis.

TABLE 3.5. Scaling factors γ for two distributions and various choices of parameters.

Distribution	Parameter	Scaling Factor γ	Parameter	Scaling Factor γ
1-D Normal	$\sigma = 1$	0.98	$\sigma = 2$	0.99
	$\sigma = 3$	0.98		
1-D Laplace	$b = 1$	0.86	$b = 2$	0.93
	$b = 3$	0.96		
2-D Normal	$\sigma = 1$	0.93	$\sigma = 2$	0.89
2-D Laplace	$\sigma = 1$	0.98	$\sigma = 2$	0.97

When working with 2-D domains, we must also be more careful as to how we measure the success of our model. Error is inevitable, and we can expect that it will be greatest in proximity to the leading edge. In 1-D, the area of the leading edge is more or less constant, but in 2-D, the length of the leading edge grows as it moves away from a spatially centered initial population. As a result, the total error of the system increases with time even if we

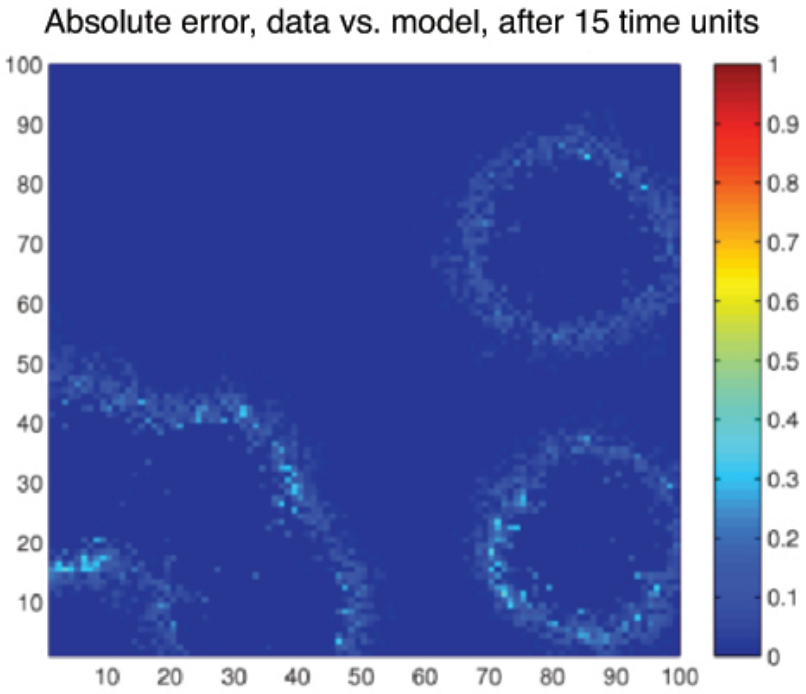

FIGURE 3.10. Absolute error between the model and stochastically generated results for $\Pr[Y > 0]$ after 15 time units with uniform random noise suitability.

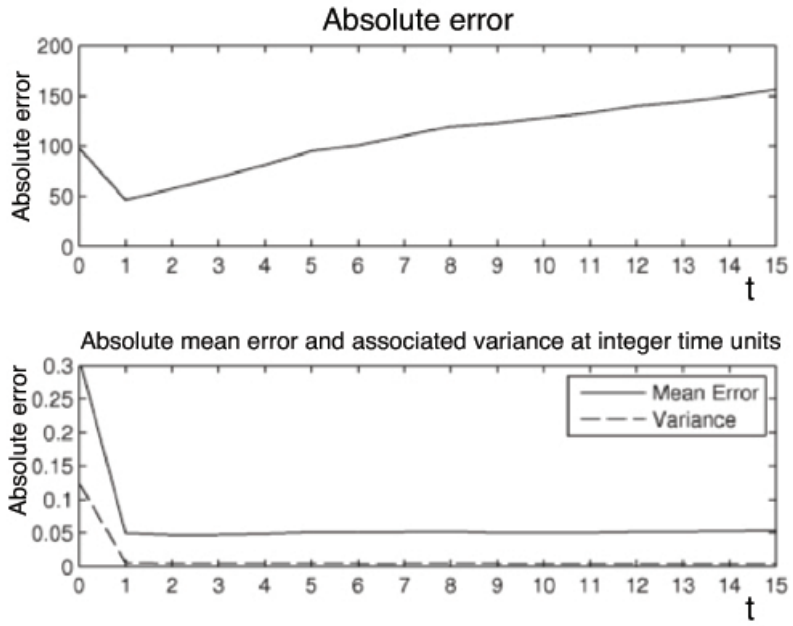

FIGURE 3.11. Error plots for the model vs. stochastically generated results for $\Pr[Y > 0]$ with uniform random noise suitability. Note that the mean error does not increase with time.

are modeling the system with the same accuracy as in the 1-D case. This suggests that total error is not necessarily a good metric by which to measure the success of our model. Instead, we will pay more attention to the mean absolute error as calculated among all cells whose absolute error is above a threshold value (0.001). For a successful model, we should observe that this mean error does not increase significantly with time.

For a second simulation, we chose a more complicated topology for the suitability, as shown in Figure 3.12. The model was then run for 15 time units with the same parameters and weight function as for the simulation of Figs. 3.8 - 3.11, and similarly compared to an expected value of $\Pr[Y > 0]$ as generated from 1000 pseudo-random realizations of Y over the same 15 time units. To give some idea of the computational advantage that the model has over repeated realizations of the stochastic process, it took 19 hours and 11 minutes to obtain the 1000 pseudo-random realizations for Y on a 100-by-100 spatial mesh using 27 processors in parallel at 2.4 ghz each (quad core) with 128 GB of memory. By contrast, the derived model did not run in parallel, and on a 2.2 ghz dual core machine with 8 GB of memory, it took 1 hour and 39 minutes.

The results are shown in Figures 3.13 through 3.15. The nonhomogeneous suitability gives rise to a nonisotropic spreading of the probability function from patches where $u(x, 0) = 1$. The error, as before, is concentrated at the edge of the propagating front.

3.5. Discussion

With Equation (33), we have derived a first model for the time-evolution of the occurrence probability for an invasive plant species. A key motivation for working with occurrence probability as our quantity of interest is the fact that field data for invasive species, and in particular data used as input to ecological niche models such as Maxent, typically come in

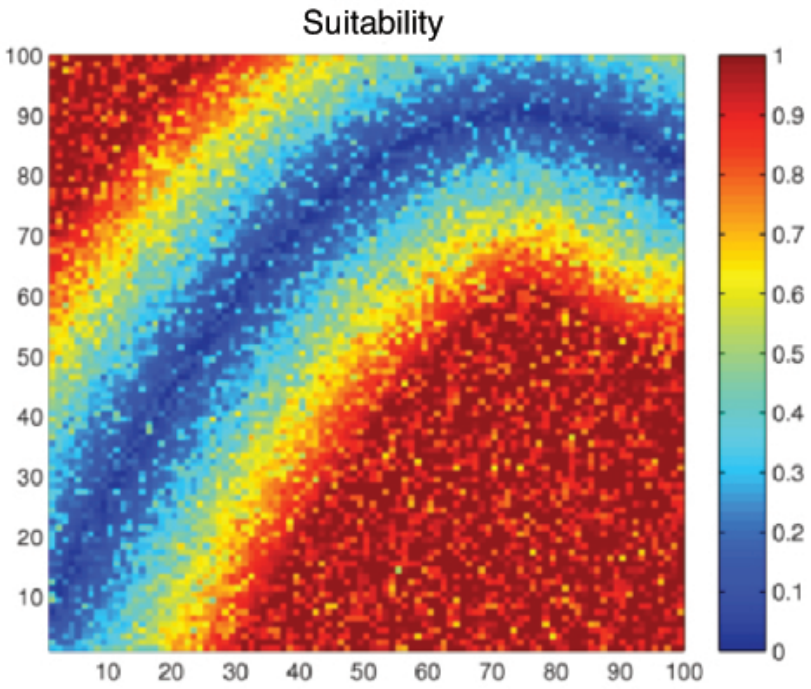

FIGURE 3.12. Plot of a suitability landscape.

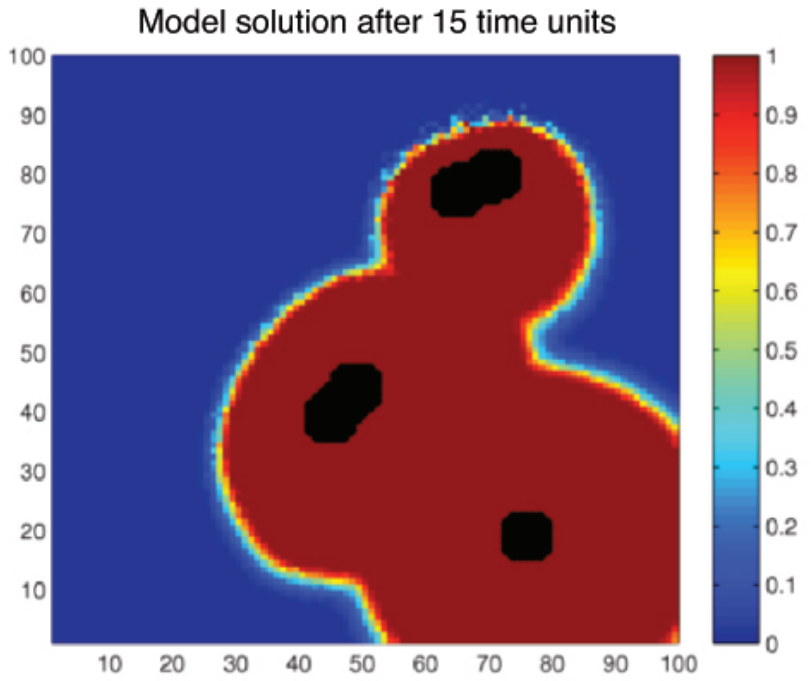

FIGURE 3.13. Model solution after 15 time units with suitability given in Figure 3.12. Initial locations where $u(\mathbf{x}, 0) = 1$ are shown in black.

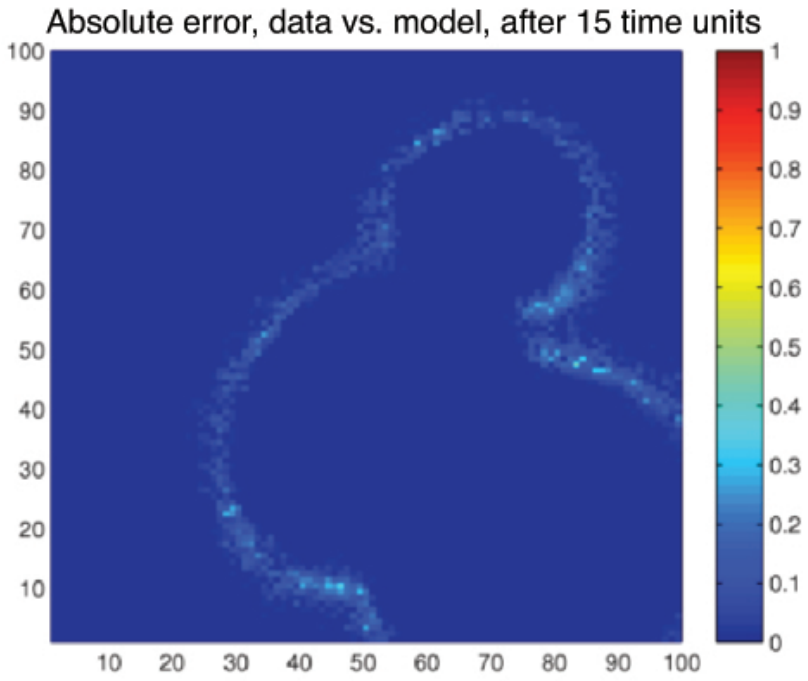

FIGURE 3.14. Absolute error between the model and stochastically generated results for $\Pr[Y > 0]$ after 15 time units with suitability given in Figure 3.12.

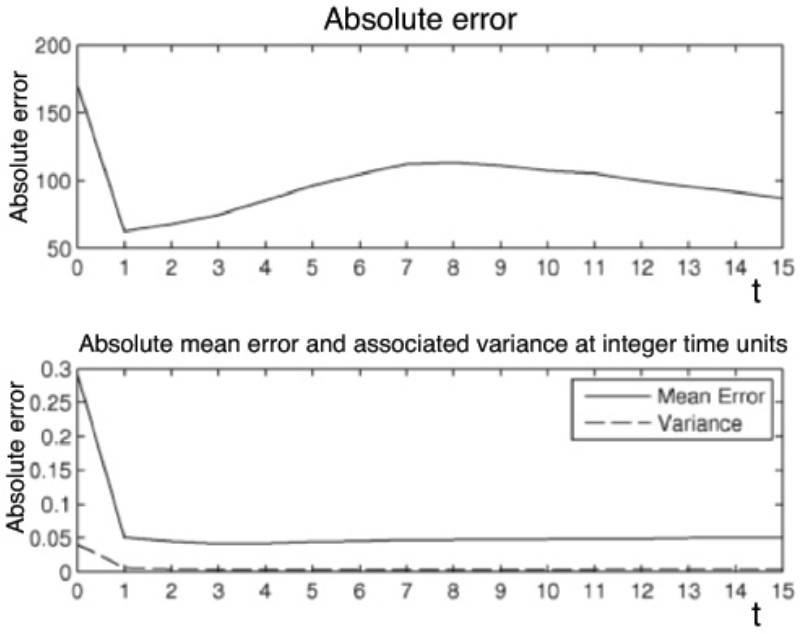

FIGURE 3.15. Error plots for the model vs. stochastically generated results for $\Pr[Y > 0]$ with suitability given in Figure 3.12.

the form of spotty species presence points. From these data, one may estimate a probability of species occurrence across the domain of interest even though one does not have species population data at hand. This estimate may then be used as a realistic initial condition for our model, but not for a model whose quantity of interest is population size.

The numerical simulations presented in this chapter were chosen to illustrate how our deterministic formulation of the model accurately captures the underlying stochastic process. The parameters α and β, which depend on the distribution kernel chosen for the model as well as the carrying capacity, must be determined once for a given kernel and carrying capacity by comparison with the stochastic process, and then the deterministic model may be run instead of the much more computationally expensive stochastic realization. The computation time saved by using the deterministic model increases as one studies larger domains of interest to ecologists and resource managers. When choosing unit values for measuring an invader and setting base carrying capacity, we suggest choosing the smallest viable, reproducing population size of the invasive organism to be the base unit in the model. Carrying capacity can then be set at the largest number of such units possible in a perfectly suitable unit area of the domain. This choice will maximize the dynamics captured by the model and give traveling wave speed results that can be described as the rate of spread for viable colonies.

As discussed in the introduction to this chapter, ecological niche models produce maps giving a suitability (a number between 0 and 1) for a given species at points in a given terrain. We have interpreted this suitability as the probability of establishment of the species and incorporated it into our model as a function $s(x)$. Suitability map outputs of these programs may now be incorporated into our model, so that ecological niche models, together with our evolution equation, can predict how a species will spread. In these models, carrying capacity

becomes a spatial function of suitability, and α and β similarly become spatial functions. In the following chapter, we will combine Maxent data for cheatgrass suitability in the Rocky Mountains with our model and compare the model output with data on cheatgrass spread. As plants (particularly in temperate climates) often have separated growth and dispersal phases, discrete analogs of Equation (7) are often used [60, 44], and for certain applications, it may be necessary to utilize a discrete-time analog of Equation (33).

The stochastic simulations of the contact-birth process (26) and the corresponding wave speeds presented in this chapter suggest a nonlinear relationship between carrying capacity K and wavespeed, with wavespeed being more sensitive to carrying capacity at lower values of K. This model, however, neglects many ecological processes that either ecological data or models suggest affect wavespeed, including Allee effects and the timing of growth and dispersal phases [61]. Cellular automaton modeling has shown that stochasticity in colonization and spatial heterogeneity can significantly increase invasion speed [62]; it will be interesting to investigate the effect of heterogeneity of carrying capacity and suitability as determined by ecological niche models on wavespeed for Equation (33).

As it stands, our model will predict how species occurrence probability changes with time under the assumption of no plant death. The model may be used to predict how a species may penetrate terrains of various suitability levels or degrees of suitability heterogeneity, but it does not currently accommodate resource management strategies such as the application of herbicides. The model may be extended in two simple ways to include such effects. By allowing the suitability function $s(x)$ to depend on time, we could model, for example, the periodic application of a herbicide that reduces the suitability of a patch for a species. The herbicide would also kill existing plants at the site, so it would also be necessary to add a

death term $-\mu u$ to the right-hand side of the governing equation. Effects such as seasonal variation and changing suitability values due to factors such as forest fire, competition with other invading species, or climate change may also be studied via the suitability function and death term that depend on time.

In the formulation of our model, we have been motivated by the study of invasive plants. As in Mollison's original contact-birth model, our model may also be appropriate for other situations, including epidemiological contexts.

CHAPTER 4

The dynamics and control of non-local invasive spread coupled with a vector-based transportation network

4.1. Introduction

In the previous chapter, we derived a model for invasive spread motivated by invasive plants, and in particular cheatgrass (*Bromus tectorum*). In order to keep our model as general as possible, we have not assumed any mechanics that are specific to *Bromus*, and we will continue to take this approach in the current chapter to the extent possible. Similarly, we do not speculate as to why the species is invasive (though plenty of theory exists on this topic, see for example [63]), but only assume that there is no significant biological competitor which reduces invasive range over short-term periods. With this general approach in mind, we now turn to the particular case of *Bromus tectorum* to evaluate the effectiveness of our modeling approach in a real world scenario, and look for ways to improve our predictions.

Bromus tectorum, also known as downy brome or cheatgrass, is an introduced annual grass that has infested more than 40 million hectares in the United States. Impacts include damage to rangeland, winter crops, hayfields, pastures, grass seed fields, and native shrub species due to increased fires [35]. In particular, there is an active *Bromus* invasion taking place in Rocky Mountain National Park, for which presence data exists intermittently since 1996 [2]. Evangelista *et al.* (2008) [3] examined the performance of various niche modeling techniques for *Bromus tectorum* in the western United States, and one of the authors, Sunil Kumar, was able to provide us with suitability data for *Bromus* in Rocky Mountain

National Park using the Maxent niche modeling package [6]. This suitability data is based on the latest *Bromus* presence data for Rocky Mountain National Park, and allows us to easily parameterize the spatial heterogeneity in the previously described presence probability model.

Dispersal distributions of *Bromus tectorum* can vary depending upon site conditions, including burn history [5]. However, allowing the distribution kernel of Chapter 3 to vary spatially represents a rather difficult numerical challenge for the area, both in terms of implementation (due to the size of the problem) and the availability of data to parameterize the spatial heterogeneity. Instead, we uniformly utilized a multivariate Laplace distribution [64] with parameters based on a mean of $(0, 0)$, variance of $(1, 1)$, and correlation 0 to roughly match the distribution data gathered in [5] for burned sites. The multivariate Laplace distribution was specifically chosen for its heavy tails and exponential shape, with most of the probability being contained inside a circle of radius 1 from the origin. While this choice certainly represents a worse-case scenario for *Bromus* spread, our modeling efforts are centered around calculating an upper bound on presence probability, and since Maxent suitability modeling often includes layers with terrain cover information, we might expect this distribution to be automatically damped in forested areas due to a suitability change (see, for example, Section 4.3.3).

Figure 4.1 displays the solution of the presence probability model in 1999 based on initial data from 1996 using the model, data, and dispersal kernel described in Chapter 3. The model growth rate was set to $r = 5$, and the carrying capacity was chosen to be $K = 323$, based on the number of square feet of cheatgrass necessary to cover a square meter (1 square

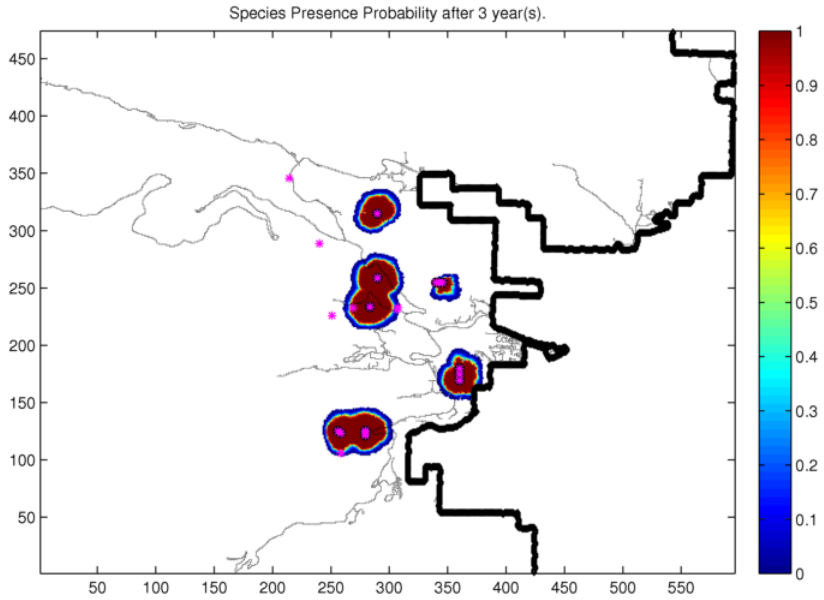

FIGURE 4.1. Predicted Rocky Mountain National Park cheatgrass presence for 1999 based on data from 1996 (black outlined pink stars). Presence data for 1999 is also plotted (pink stars). Parameters: $r = 5$, $w \sim \text{Laplace}(\mu = (0,0), \sigma = (1,1), \rho = 0)$, $K = 323$.

foot of cheatgrass was arbitrarily chosen as the minimal plant unit for presence in a 30m^2 area).

Upon inspection of Figure 4.1, one can immediately see that while the model does a fairly good job of capturing local spread with these parameters, other points are completely missed, occurring in locations nowhere close to the model output. While this may be due to a lack of search in these new areas during the 1996 census, it is well documented that trail corridors function as habitat and conduits for movement of plant species in Rocky Mountain National Park [65], and with the number of visitors, it is not hard to believe that these findings extend to non-local movement along paved roads as well. *Bromus tectorum* seeds can easily attach to clothing, and visitors would find it easy to unknowingly transport propagules down trails or to distant sites using motor vehicles. Since the additional points

in Figure 4.1 are conspicuously near roads or trails, we now seek to develop a model that can capture these long-distance events in a way that remains tractable to analysis.

This chapter is organized as follows: In Section 4.2, we introduce a graph-based population model in which individuals can become infected at nodes. The population of the graph is conserved, but every node acts like both a source and sink for individuals to leave and reenter the graph uninfected. We then conduct analysis on this network model and show how it can be coupled to a more spatially continuous infection model to obtain a more complete model of epidemic spread. In Section 4.3, we extend this model to the case of a herbaceous invader and show some numerical results for Rocky Mountain National Park. Section 4.4 introduces two different types of control to the model and discusses optimal control in the linear case. We then review numerical simulations for control in Section 4.4.5 and conclude with a discussion of our results in Section 4.5.

4.2. Infectious disease epidemic model

4.2.1. Linear graph model. To begin modeling long-distance spread, we will first consider the general case of an epidemic with intermediary carrier vectors. These carrier vectors will not be infectious amongst themselves and remain on a transportation network with well defined nodes and directional rates of flow. At each node, carrier vectors can later interact with an underlying, spatially coupled model such as the one developed in Chapter 3. If the underlying model predicts invader presence at an assigned node location, network carrier vectors will become infected and possibly spread the epidemic to other nodes, which in turn infect the underlying model. The precise mechanics of this concept will be formalized in the following sections, including some immediately relevant analysis.

Consider an individual on a strongly connected[1], directed graph, and let $X(t)$ be a stochastic variable which gives the node this individual occupies at time t. We assume that $X(t)$ satisfies the Markov property and represents a continuous-time Markov chain on the nodes of the graph. Let $g_{ij} \geq 0$ be the transition rate for the $j \to i$ node edge whenever $i \neq j$, and let

$$g_{jj} = -\sum_{i \neq j} g_{ij}.$$

We require this value for g_{jj} to assure that the rate at which individuals leave node j is equal to the sum over all the rates for entering destination nodes i. In particular, this guarantees that the number of individuals on the graph will be conserved.

Define $\mathbf{p}(t)$ to be the vector of probabilities of finding X at each node at time t: $\mathbf{p}_i(t) = \Pr[X(t) = i]$. Since the number of individuals on the graph is conserved, the matrix G such that $(G)_{ij} = g_{ij}$ gives us the evolution of these probabilities on the graph via the first-order equation

$$\frac{d\mathbf{p}}{dt} = G\mathbf{p}(t). \tag{34}$$

Assuming that there are N individuals on the graph and their initial distribution between the nodes is given by the vector $\mathbf{N}(0) = \mathbf{N}_0$, we can now set up an initial value problem

$$\frac{d\mathbf{N}}{dt} = G\mathbf{N}. \tag{35}$$

$$\mathbf{N}(0) = \mathbf{N}_0$$

for the expected distribution $\mathbf{N}$ of the individuals over the graph's nodes at time t.

[1]For every node, there exists a path through the graph to every other node. This is typical for transportation networks.

The above graph model does not explicitly contain any sources or sinks, but in a localized, real-world transportation network, individuals leave and enter the network all the time from each of the nodes. To incorporate this additional dynamic, we will suppose that the number of individuals leaving the graph is balanced by the total number of individuals entering. More specifically, suppose that individuals leave the graph node i with rate μ_i such that $g_{ii} \geq \mu_i \geq 0$, and define a vector of these rates, $\boldsymbol{\mu}$. Assuming that the total population on the graph is conserved, individuals also enter the graph at a rate given by $\sum \mu_i$ but spread out over the graph via a vector $\boldsymbol{\nu}$, with $\sum \nu_i = 1$. Then we can define a new matrix of transition rates H by the relations

$$h_{ij} = g_{ij} + (\boldsymbol{\nu}\boldsymbol{\mu}^T)_{ij}, \ \forall i \neq j \tag{36}$$

$$h_{jj} = g_{jj} - \mu_j + (\boldsymbol{\nu}\boldsymbol{\mu}^T)_{jj},$$

where $\boldsymbol{\nu}\boldsymbol{\mu}^T$ denotes the outer product, noting that

$$h_{jj} = -\sum_{i \neq j} h_{ij}$$

because $\sum \nu_i = 1$.

This definition of H is equivalent to the more shorthand relation $H = G - \text{diag}(\boldsymbol{\mu}) + \boldsymbol{\nu}\boldsymbol{\mu}^T$. Since the columns of both G and H sum to zero, we are guaranteed that these matrices are singular. While we could simply redefine the entries of G to include the case of H, the distinction between these two matrices will prove useful shortly, as we define two distinct subgroups within the graph population N.

4.2.2. Susceptible/infected vector model on a graph. Suppose now that all individuals on the graph have the potential to be carriers for some vector-borne disease or invader. We assume that these individuals are unable to contract the disease themselves, at least not in any significant way, or transmit it between members of their own population. Instead, they acquire the disease at certain nodes with a given rate which is variable in time. Define $\beta_i \geq 0$ to be the rate of infection at node i, and $\boldsymbol{\beta}(t)$ to be the vector of all such rates at time t. We will assume that once an individual has the disease, it remains a carrier for all time but may leave the graph as described in the previous section. New members that join the graph are always disease free, but can later become carriers by the same process.

Let $\mathbf{s}$ represent the expected number of suspectable but non-infected individuals at each node, and $\mathbf{c}$ the expected number of infected (carrier) individuals at each node. Using the matrix G and H of transition rates defined in the previous section, we can model the movement of these populations on the graph by the equations

$$\begin{aligned} \frac{d\mathbf{s}}{dt} &= (H - \operatorname{diag}(\boldsymbol{\beta}_t))\mathbf{s} + (\boldsymbol{\mu} \cdot \mathbf{c})\boldsymbol{\nu} \\ \frac{d\mathbf{c}}{dt} &= \operatorname{diag}(\boldsymbol{\beta}_t)\mathbf{s} + (G - \operatorname{diag}(\boldsymbol{\mu}))\mathbf{c} \\ \mathbf{s}(0) &= \mathbf{s}_0 \\ \mathbf{c}(0) &= \mathbf{c}_0 \\ \text{where } H &= G - \operatorname{diag}(\boldsymbol{\mu}) + \boldsymbol{\nu}\boldsymbol{\mu}^T, \end{aligned} \tag{37}$$

and $\operatorname{diag}(\boldsymbol{\mu})$ denotes the diagonal matrix formed from vector $\boldsymbol{\mu}$. With $\boldsymbol{\beta}(t)$ specified, these equations are linear in $\mathbf{s}$ and $\mathbf{c}$ and provide a simple model for disease vectors on a graph

which can be subsequently coupled to an underlying epidemic or ecological invasion model, as shown in Sections 4.2.3 and 4.3.2.

Before continuing, however, there are a few key observations to be made about the model given in equation (37) that will give us significant insight into the behavior of our system. First, we rewrite equation (37) in vector form

$$\frac{d}{dt}\begin{pmatrix} \mathbf{s} \\ \\ \mathbf{c} \end{pmatrix} = A_t \begin{pmatrix} \mathbf{s} \\ \\ \mathbf{c} \end{pmatrix} \tag{38}$$

$$\mathbf{s}(0) = \mathbf{s}_0 \qquad \mathbf{s} : \text{susceptible}$$

$$\mathbf{c}(0) = \mathbf{c}_0 \qquad \mathbf{c} : \text{carrier}$$

where

$$A_t = \begin{pmatrix} H - \text{diag}(\boldsymbol{\beta}_t) & \boldsymbol{\nu}\boldsymbol{\mu}^T \\ \\ \text{diag}(\boldsymbol{\beta}_t) & G - \text{diag}(\boldsymbol{\mu}) \end{pmatrix} \tag{39}$$

with $\boldsymbol{\nu}\boldsymbol{\mu}^T$ once again denoting the outer product. Since the elements of $\boldsymbol{\nu}$ sum to 1, A_t is a transition matrix of exactly the same general form as G or H, with $a_{ij} \geq 0 \; \forall i \neq j$ and

$$a_{jj} = -\sum_{i \neq j} a_{ij}.$$

In fact, we can think of the system as modeling two identical graphs, one for $\mathbf{s}$ and one for $\mathbf{c}$, with specific links between them. Every node of the $\mathbf{s}$ graph is directly reachable from any node of the $\mathbf{c}$ graph through the weighted edges given by $\boldsymbol{\nu}\boldsymbol{\mu}^T$, while nodes of the $\mathbf{c}$ graph are only directly connected to corresponding nodes of $\mathbf{s}$ graph when $\beta_i > 0$. Assuming

that at least one $\beta_i > 0$ and the graph of $\mathbf{s}$ (and thus also $\mathbf{c}$) is strongly connected, then the combined graph will also be strongly connected. This observation implies that once a single node is infected, every node becomes suspectable to infection. It also allows us to broadly apply the following lemma, and ultimately arrive at a key theorem about the dynamics of system (38).

LEMMA 1. *Let G be a matrix of transition rates for a strongly connected, directed graph, so that $g_{ij} \geq 0$ when $i \neq j$ and*

$$g_{jj} = -\sum_{i \neq j} g_{ij} < 0.$$

Then the spectrum of G is equal to $\{0, \lambda_1, ..., \lambda_{\dim(G)-1}\}$ where $\lambda_i < 0$.

PROOF. We will proceed by applying the Perron-Frobenius theorem to a linear transformation of G, and then interpreting the result for our original matrix. Let $0 < \varepsilon < 1/\max\{|g_{ij}|\}$, and define a matrix $P_\varepsilon = \varepsilon\, G + I$. Then P_ε is a column stochastic matrix - e.g., all elements of P_ε are between 0 and 1 and the columns of P_ε each sum to 1. This fact implies that the largest eigenvalue of P_ε is 1, and that this is also the spectral radius. Furthermore, since the graph associated with G is strongly connected and the positive elements of P_ε exactly correspond to the nonzero elements of G, P_ε is also associated with the strongly connected directed graph, but in the discrete-time sense (more precisely, P_ε is the transition matrix of a strongly connected, discrete-time Markov chain). Thus P_ε is an irreducible matrix, and since it is also non-negative, by the Perron-Frobenius theorem we can conclude that the maximal eigenvalue of 1 is unique. Now, since $\varepsilon\, G = P_\varepsilon - I$, it follows that $\varepsilon\, G$ has a unique largest eigenvalue of 0, and therefore so does G. □

THEOREM 1. *Let the matrix A be defined as in equation (39) from a strongly connected graph and let $\boldsymbol{\beta}$ be constant in time. Then A has a one-dimensional kernel with all other eigenvalues negative. The asymptotic behavior of any trajectory for the system (38) with initial condition $(\mathbf{s}_0, \mathbf{c}_0)$ is that it approaches a stable equilibrium $(\mathbf{s}^*, \mathbf{c}^*)$ such that $\sum \mathbf{s}^* + \sum \mathbf{c}^* = \sum \mathbf{s}_0 + \sum \mathbf{c}_0$.*

PROOF. Suppose that at least one $\beta_i > 0$. Then by construction of A and our previous observation concerning strongly connected graphs for $\mathbf{s}$ and $\mathbf{c}$, A is a matrix of transition rates for a strongly connected, directed graph that satisfies the requirements of Lemma 1. Thus A has a one-dimensional kernel with all other eigenvalues negative.

Writing out A as

$$A = \begin{pmatrix} G - \operatorname{diag}(\boldsymbol{\mu}) + \boldsymbol{\nu}\boldsymbol{\mu}^T - \operatorname{diag}(\boldsymbol{\beta}) & \boldsymbol{\nu}\boldsymbol{\mu}^T \\ \operatorname{diag}(\boldsymbol{\beta}) & G - \operatorname{diag}(\boldsymbol{\mu}) \end{pmatrix},$$

we note that the column sums of G equal 0, and since $\sum \nu_i = 1$, the column sums of $\boldsymbol{\nu}\boldsymbol{\mu}^T$ equal $\boldsymbol{\mu}$. Thus the column sums of A are zero, which implies balance in the system of differential equations (38) and therefore conservation for the sum of the elements of $\mathbf{s}$ and $\mathbf{c}$ for all time t. Thus the solution of the initial value problem approaches the kernel of A at a unique point, where the element sum of an eigenvector associated with the zero eigenvalue is equal to the sum over all the elements of $(\mathbf{s}_0, \mathbf{c}_0)$. Since all other modes have negative eigenvalues, they will decay, and the solution will asymptotically approach a stable equilibrium.

Now suppose instead that $\boldsymbol{\beta}$ is equal to the zero vector. Then A can be written as a block upper triangular matrix

$$A = \begin{pmatrix} H & \boldsymbol{\nu}\boldsymbol{\mu}^T \\ 0 & G - \mathrm{diag}(\boldsymbol{\mu}) \end{pmatrix}, \text{where } H = G - \mathrm{diag}(\boldsymbol{\mu}) + \boldsymbol{\nu}\boldsymbol{\mu}^T.$$

Recall that G is the transition matrix for a continuous-time Markov process on a strongly connected graph. By Lemma 1, G has a unique largest eigenvalue of 0.

Assume first that $\boldsymbol{\mu}$ is the zero vector. Then A is in fact block diagonal, and represents two completely uncoupled, continuous-time graph processes, both with associated transition matrix G. Since the column vectors of G all sum to zero, the differential equation system (38) is balanced and the sum of its solution vector is conserved. Along with the system decoupling, this implies that both $\mathbf{s}$ and $\mathbf{c}$ separately have families of unique, stable steady states, and thus the combined system has one unique, stable steady state for $(\mathbf{s}, \mathbf{c})$ with initial condition $(\mathbf{s}_0, \mathbf{c}_0)$.

Now assume that $\boldsymbol{\mu}$ has at least one positive entry. Since A is block upper triangular, the equation for $\frac{d\mathbf{c}}{dt}$ lacks a coupling term,

$$\frac{d\mathbf{c}}{dt} = (G - \mathrm{diag}(\boldsymbol{\mu}))\mathbf{c}. \tag{40}$$

We claim that the kernel of $(G - \mathrm{diag}(\boldsymbol{\mu}))$ is trivial. To prove this assertion, we first note that the eigenvalues of both G and $-\mathrm{diag}(\boldsymbol{\mu})$ are non-positive, since G has a unique maximal eigenvalue of 0 and $\boldsymbol{\mu}$ is always non-negative. Thus, for any vector $\mathbf{v} \in \ker(G - \mathrm{diag}(\boldsymbol{\mu}))$, $\mathbf{v} \in \ker(G)$ and $\mathbf{v} \in \ker(\mathrm{diag}(\boldsymbol{\mu}))$. Suppose that $\mathbf{v} \in \ker(\mathrm{diag}(\boldsymbol{\mu}))$. Since $\mathrm{diag}(\boldsymbol{\mu})$ is a diagonal matrix with at least one positive entry, $\mathbf{v}$ must have at least one zero element,

specifically in the row k corresponding to $\mu_k \neq 0$. Now suppose that $\mathbf{v}$ has at least one non-zero element as well.

By construction of G, including equation (34) for the time evolution of probabilities through a Markov process on a graph, it is easy to see through Lemma 1 that if $\mathbf{v} \in \ker(G)$, either $v_i \geq 0$ for all i, or $v_i \leq 0$ for all i. Since $g_{ij} \geq 0$ whenever $i \neq j$, $g_{ii} < 0$, and at least two entries of any row are nonzero, then in any row k of G such that $v_k = 0$, there must be nonzero entries only in columns j such that $v_j = 0$ as well, because all other entries have the same sign. Since G is a matrix of rates on a directed graph, this fact implies that the associated nodes j are only connected to each other. In the case that $\mathbf{v}$ has at least one nonzero element, this observation contradicts the assumption that the graph represented by G is strongly connected. Thus $\mathbf{v}$ must be the zero vector, and we have proven our claim.

With this result, we can now conclude that $\mathbf{c}$ has a single, stable steady state at $\mathbf{c} = 0$. We now need only to examine the behavior of $\mathbf{s}$ at this steady state. When $\mathbf{c} = 0$, we have

$$\frac{d\mathbf{s}}{dt} = H\mathbf{s},$$

and since the graph representation of H is equivalent to that of G up to a change of variables, we can apply Lemma 1 to show that H has a unique largest eigenvalue of 0. Since the sum of the elements in $\mathbf{s}$ and $\mathbf{c}$ are conserved, and $\mathbf{c} = 0$ at the steady state, the kernel of H intersects the solution space of $\mathbf{s}$ at exactly the point where the element sum of $\mathbf{s}$ is equal to the combined sum of the elements of $\mathbf{s}_0$ and $\mathbf{c}_0$. Thus $(\mathbf{s}, \mathbf{c})$ has exactly one stable steady state solution, which concludes the proof of the theorem. $\square$

This theorem will prove useful in the following sections to analyze different combined models and to find an optimal control. We now introduce the first of these combined models and discuss some of its applications and assumptions.

4.2.3. Coupling the graph with a spatial infection model. To combine our graph vector model with an underlying model of epidemic spread, let $Y(x,t)$ represent the infected target population (non-graph) over a location space Ω at time t. We will assume that $Y(x,t)$ evolves in time either by a deterministic process

$$\frac{\partial Y}{\partial t} = F_1(Y,x)$$

or by a stochastic transition probability

$$\lim_{dt\downarrow 0}[\frac{1}{dt}\Pr[Y(x) \to Y(x)+1 \text{ in } (t+dt]]] = F_2(Y,x)$$

in absence of the graph and its carrier vectors. To couple the graph to this model, define a vector valued function $\mathbf{w}(x)$ such that $w_i(x)$ is a probability density function centered at some location $x_i \in \Omega$ corresponding to node i on the graph. We will now say that node i on the graph is centered at location x_i in Ω, and given that a disease vector from the graph is located at node i, $w_i(x)$ is defined to be the probability of finding that vector at location x in Ω.

Next, we redefine F_1 or F_2 via

$$\tilde{F}(Y,\mathbf{c},x) = F_{1 \text{ or } 2}(Y,x) + r\mathbf{w}(x)\cdot\mathbf{c}, \tag{41}$$

where r is the infection rate of a graph disease vector. This change in the $Y(x,t)$ equation allows graph vectors to grow the infected population near assigned locations of the graph nodes, specifically by distributing the infection rate through the distribution $\mathbf{w}(x)$. Finally, given F, we explicitly define $\boldsymbol{\beta}(t)$ on the graph using the number of infected individuals Y located around each node. Let γ be the rate that an individual from Y infects a disease vector on the graph when contact is made. Then

$$\beta_i(Y) = \gamma \int_\Omega Y(x) w_i(x) dx \tag{42}$$

$$\boldsymbol{\beta}(Y) = \{\beta_i(Y)\}_{i=1}^n$$

Assuming (without loss of generality) that the underlying infection model is deterministic, our complete, coupled model can now be written as

$$\begin{aligned}
\frac{d\mathbf{s}}{dt} &= (H - \operatorname{diag}(\boldsymbol{\beta}(Y)))\mathbf{s} + (\boldsymbol{\nu}\boldsymbol{\mu}^T)\mathbf{c} \\
\frac{d\mathbf{c}}{dt} &= \operatorname{diag}(\boldsymbol{\beta}(Y))\mathbf{s} + (G - \operatorname{diag}(\boldsymbol{\mu}))\mathbf{c} \\
\frac{\partial Y}{\partial t} &= F(Y,x) + r\mathbf{w}(x)\cdot\mathbf{c} \\
\boldsymbol{\beta}(Y) &= \gamma \int_\Omega Y(x,t)\mathbf{w}(x)dx \\
\mathbf{s}(0) &= \mathbf{s}_0, \quad \mathbf{c}(0) = \mathbf{c}_0, \quad Y(x,0) = Y_0(x).
\end{aligned} \tag{43}$$

4.3. Invasive species model: multiple time-scales

4.3.1. Underlying assumptions. The model represented in equation (43) requires some underlying assumptions that make it best suited for modeling the spread of an infectious disease rather than an herbaceous, biological invader. These include

(1) Only one timescale is involved - the infection grows at a rate comparable to movement on the graph.

(2) There is no significant latency between contact with an infected graph vector and full infection of a Y individual.

(3) Graph coupling not only provides for non-local spread of the disease via graph edges, but also facilitates local infection. Infected individuals Y can infect susceptible vectors $\mathbf{s} \to \mathbf{c}$ at a node, and the newly infected vectors $\mathbf{c}$ may add to the infection rate at their node of origin before moving on.

While Assumption 3 may be appropriate for an herbaceous invader, the first two assumptions certainly are not. Plants do not germinate, grow, and release new seeds (a year-long process) in the same amount of time as an individual would typically visit a park (a day-long process). Similarly, since infection of a location corresponds to the presence of a seed bearing invasive plant in this context, there is a certain amount of latency between the time seeds are distributed by graph vectors and the resulting infection of a location. To account for this change of context, our model needs to be modified.

4.3.2. Modifying the model for an herbaceous invader. To address assumptions 1 and 2, we begin by requiring that $F(Y, x)$ is scaled appropriately by some value ϵ, $0 < \epsilon << 1$. This scaling represents the idea that our model encompasses two timescales - one on which individuals on the graph move from node to node ($\sim t$) and another on which plants reproduce and grow ($\sim \epsilon$). Additionally, we will introduce a new variable $L(x, t)$, $x \in \Omega$, representing latent seeds which have been spread by graph vectors but have not yet

established themselves in their location. $L(x)$ increases from the graph as $Y(x)$ previously did, with members leaving $L(x)$ to $Y(x)$ at a rate of order ϵ. The resulting system can be represented by the equations

$$
\begin{aligned}
\frac{d\mathbf{s}}{dt} &= (H - \mathrm{diag}(\boldsymbol{\beta}(Y)))\mathbf{s} + (\boldsymbol{\nu}\boldsymbol{\mu}^T)\mathbf{c} \\
\frac{d\mathbf{c}}{dt} &= \mathrm{diag}(\boldsymbol{\beta}(Y))\mathbf{s} + (G - \mathrm{diag}(\boldsymbol{\mu}))\mathbf{c} \\
\frac{\partial L}{\partial t} &= r\mathbf{w}(x) \cdot \mathbf{c} - \sigma L(x,t)h(Y,x) - \delta L \\
\frac{\partial Y}{\partial t} &= \epsilon F(Y,x) + \sigma L(x,t)h(Y,x) \\
\boldsymbol{\beta}(Y) &= \gamma \int_\Omega Y(x,t)\mathbf{w}(x)dx \\
\mathbf{s}(0) = \mathbf{s}_0, &\quad \mathbf{c}(0) = \mathbf{c}_0, \quad Y(x,0) = Y_0(x), \quad L(x,0) = L_0(x)
\end{aligned}
\tag{44}
$$

where $\sigma \sim \delta \sim \epsilon$, $\delta \geq 0$ is a decay term for the latent seeds, and $h(Y,x)$ is a crowding term also represented in $F(Y,x)$.

4.3.3. Numerical results: Rocky Mountain National Park. Using equation (44) and the deterministic presence probability model described in Chapter 3, we tested our model's performance on the Rocky Mountain National Park data set described in Section 4.1. The 1996-1999 model results and presence data seen in Figure 4.1 were used to parameterize the growth rate of Chapter 3 while assuming a Laplace distribution (described in Section 4.1) for the dispersal kernel $\mathbf{w}(x)$.

Nodes were chosen based on remote sensing data of the park made available through Google Earth [66], including trail heads, trail junctions, parking lots, campgrounds, picnic areas, waterfalls, stables, and ranger stations. Since traffic rates between each of the points

was unavailable, flow between graph nodes was estimated using some basic knowledge of the park's attractions, geographical proximity of the nodes, and the function of the nodes in question (e.g., if one has just finished hiking at a trail head, they are less likely to immediately hike another trail - perhaps opting instead to eat lunch at a picnic site). Similar methods were used to estimate the vector $\boldsymbol{\nu}$ for graph entry rates, and vector $\boldsymbol{\mu}$ for graph exit rates.

Since presence probability was used instead of an explicit model for $Y(x,t)$, $\boldsymbol{\beta}(Y)$ was calculated slightly different than previously described. The exact equation used was

$$\boldsymbol{\beta}(\Pr[Y(x,t)>0]) = \gamma \int_{\Omega} \Pr[Y(x,t)>0] m(x) \mathbf{w}(x) dx, \tag{45}$$

where $m(x)$ is the suitability of location x as parameterized by Maxent [6] and $\mathbf{w}(x)$, the distribution function connecting each spatial node location to the graph, was chosen to be a normal distribution with mean $(0,0)$, standard deviation $(5,5)$, and correlation 0, centered around each of the nodes. γ was chosen to be 0.1, corresponding to an estimate that if cheatgrass is present with a probability of 1 and suitability of 1, then the infection rate onto visitors' clothing is 10%.

Other parameters were chosen as follows: graph infection rate $r = 4$, carrying capacity $K = 350$, $\epsilon = 1/180$ (180 active days in the year, due to winter), $\sigma = \epsilon$, $\delta = 0.33$, and the total number of individuals on the graph at any time $N = 5000$. The model was run until 2008, starting with data from 1999 and assuming a standard normal probability distribution for $\Pr[Y(x,t)>0]$ around presence points. The results are shown in Figures 4.2 and 4.3.

As shown in Figure 4.2, the graph coupled presence probability model appears to capture all of the 2008 presence locations, with the exception of a couple of points in the lower left and one nearby the park boundary toward the center of the Figure. Both these locations illustrate

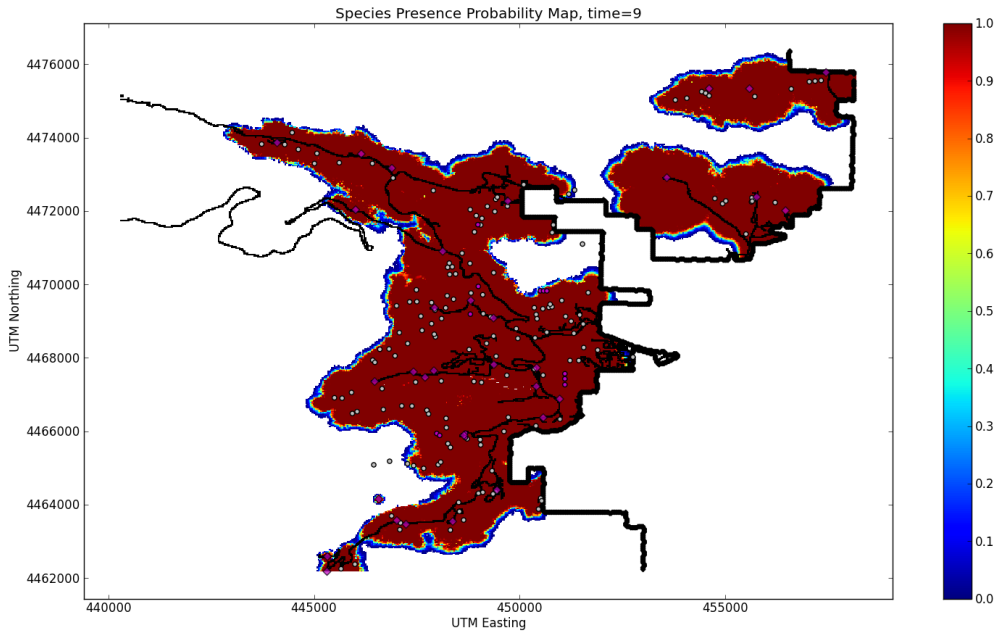

FIGURE 4.2. 2008 Rocky Mountain National Park *Bromus* presence probability, based on 1999 presence data. Grey dots are 2008 presence data, pink diamonds are node locations, and pink dots are the 1999 initial presence points.

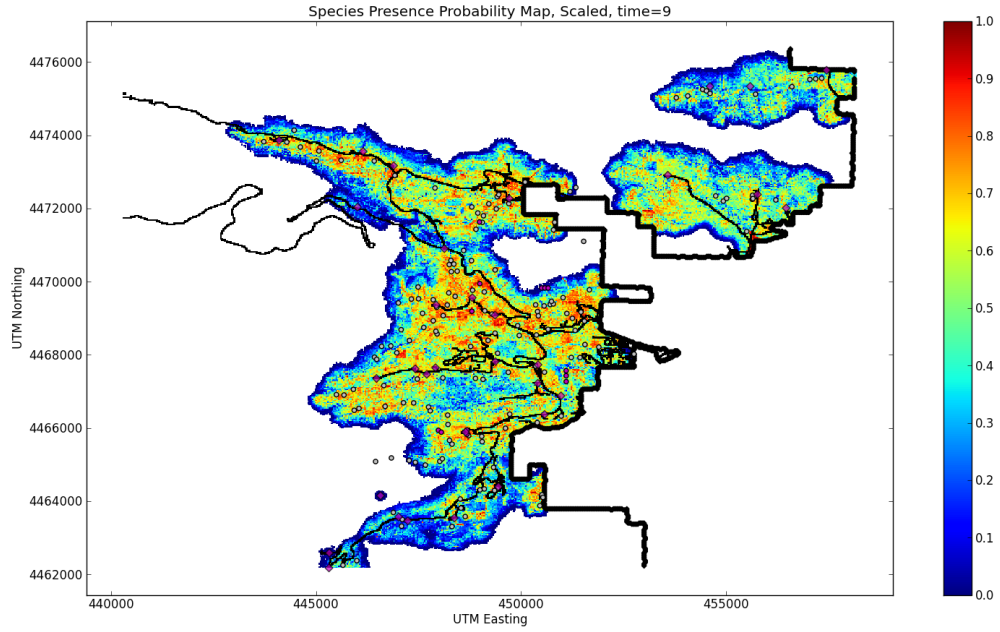

FIGURE 4.3. 2008 Rocky Mountain National Park *Bromus* presence probability, scaled by suitability. Grey dots are 2008 presence data, pink diamonds are node locations, and pink dots are the 1999 initial presence points.

important limitations to the model. The two points in the lower left lie some distance down a trail (not shown), but not particularly nearby any trail head or junction. Since the combined model was not configured to treat spread down trail corridors any different than over regular terrain (except for a possible suitability difference - trail and road data are common environmental layers to include for niche modeling), the model did not spread the invader fast enough to reach these points. In the other case, the one missed point near the center lies near the boundary of the park, outside of which no presence information was available. Since the town of Estes Park lies just outside the boundary, which includes populations of cheatgrass, it is reasonable to expect that this missed presence point was due to invasion from outside Rocky Mountain National Park boundaries.

Another feature of the model to note is that in some spots, there are relatively large areas of high presence probability but no 2008 presence data (e.g., the top right of Figure 4.2). While these locations could be due to model overestimation, we have no information about which areas of the park were surveyed in 2008, and likewise no species absence data. Since these areas are often hard to access, more information is needed to assess model performance for overestimation in these places.

Also of interest: just below the two missed points in the lower left of Figure 4.2 there is a small area of presence probability surrounding a graph node connection, but otherwise disconnected from the main mass of species presence. This node was located at a trail junction close to a lake that could possibly see a lot of travel, but because of the low suitability of the area, the invasion has been largely contained. The lack of model propagation at this node demonstrates the ability of the model to respond when, even though a location is

connected to the invasion via the transportation network, the heterogeneity of the terrain makes it difficult for the species to persist locally.

In Figure 4.3, another view of the park is shown. The model result is exactly the same as in Figure 4.2, but it has been scaled by the provided Maxent suitability data. One can immediately see the usefulness of this view, as presence data points are most often found in locations that are both highly suitable *and* predicted to have presence by the model.

4.4. Control

4.4.1. Introduction. In this section, we consider a couple methods of control that can be applied to the models described by equations (43) and (44). The first is a control on the graph, where we seek to limit the number of infectious vectors, $\mathbf{c}$. In the case of vector-borne disease, this will correspond to removing or curing the disease carrying vectors at given nodes, while in the herbaceous invasive species model, we will assume that efforts are being made to remove seeds from the humans or vehicles that carry them. The second control will be applied directly to the underlying model. In this case, we will assume that we are directly eliminating the disease in the target population. For the invasive species model, this effort corresponds to spraying a location, killing both live plants and latent seeds.

4.4.2. Graph control. To implement a control on the graph for the vector-borne disease model, we will continue to assume that the total number of graph vectors, $\mathbf{s} + \mathbf{c}$, is constant. If we also assume that infected vectors are being removed and replaced with the same distribution $\boldsymbol{\nu}$ as used before, we need only make an adjustment to the constant vector

$\boldsymbol{\mu}$ with a vector valued control function $\mathbf{u}(t)$.

$$
\begin{aligned}
\frac{d\mathbf{s}}{dt} &= [H - \text{diag}(\boldsymbol{\beta}(Y))]\mathbf{s} + [\boldsymbol{\nu}(\boldsymbol{\mu} + \mathbf{u}(t))^T]\mathbf{c} \\
\frac{d\mathbf{c}}{dt} &= \text{diag}(\boldsymbol{\beta}(Y))\mathbf{s} + [G - \text{diag}(\boldsymbol{\mu} + \mathbf{u}(t))]\mathbf{c} \\
&||\mathbf{u}(t)|| \leq K \ \forall t \text{ s.t. } t_0 \leq t \leq t_1 \\
\mathbf{s}(0) = \mathbf{s}_0, &\quad \mathbf{c}(0) = \mathbf{c}_0.
\end{aligned} \tag{46}
$$

The norm on $\mathbf{u}(t)$ is left unspecified, as it is dependent on the specific application, and this graph system can now be coupled to an underlying model as described in Sections 4.2.3 and 4.3.2.

Alternatively, if we assume that disease vectors are being cured and re-released on site, we can formulate control on the model using equations

$$
\begin{aligned}
\frac{d\mathbf{s}}{dt} &= [H - \text{diag}(\boldsymbol{\beta}(Y))]\mathbf{s} + (\boldsymbol{\nu}\boldsymbol{\mu}^T)\mathbf{c} + \text{diag}(\mathbf{u}(t))\mathbf{c} \\
\frac{d\mathbf{c}}{dt} &= \text{diag}(\boldsymbol{\beta}(Y))\mathbf{s} + [G - \text{diag}(\boldsymbol{\mu} + \mathbf{u}(t))]\mathbf{c} \\
&||\mathbf{u}(t)|| \leq K \ \forall t \text{ s.t. } t_0 \leq t \leq t_1 \\
\mathbf{s}(0) = \mathbf{s}_0, &\quad \mathbf{c}(0) = \mathbf{c}_0.
\end{aligned} \tag{47}
$$

In the case of an herbaceous invasive species, we favor system (47) over the previous formulation since the graph vectors are likely to be humans or vehicles, which after being cleaned continue to exist on the graph at the same location. As a result, we will direct our subsequent analysis to the model given in (47) and leave the other system for future work.

4.4.3. Optimal graph based control. In this section, we will consider the system of equations (47) and extrapolate our results to a combined invasive plant model. Our most important assumption will be to assume that Y is constant and omit its contribution in the analysis. Because of the presence of two time scales, this assumption is particularly mild in the plant model and will allow us to linearize the system (47) in $\mathbf{s}$ and $\mathbf{c}$, taking $\boldsymbol{\beta}$ as constant. We then look for an optimal control $\mathbf{u}$ for arbitrary $\boldsymbol{\beta}$, which will give us a method for describing an optimal control in the original model.

We begin by rewriting system (47) as an optimal control problem

$$
\begin{aligned}
&||\mathbf{c}||_{L^2(0,T)} \to \inf \\
&\frac{d\mathbf{x}}{dt} = A\mathbf{x} + B(\mathbf{u})\mathbf{x} \qquad (48) \\
&\mathbf{x}(0) = \mathbf{x}_0 \\
&||\mathbf{u}|| \leq K
\end{aligned}
$$

where

$$
\mathbf{x} = \begin{pmatrix} \mathbf{s} \\ \mathbf{c} \end{pmatrix} \quad A = \begin{pmatrix} H - \operatorname{diag}(\boldsymbol{\beta}) & \boldsymbol{\nu}\boldsymbol{\mu}^T \\ \operatorname{diag}(\boldsymbol{\beta}) & G - \operatorname{diag}(\boldsymbol{\mu}) \end{pmatrix} \quad B = \begin{pmatrix} 0 & \operatorname{diag}(\mathbf{u}) \\ 0 & -\operatorname{diag}(\mathbf{u}) \end{pmatrix}.
$$

Since $\mathbf{s}, \mathbf{c}, \mathbf{u} \in \mathbb{R}^n$ with n the number of network nodes, $\mathbf{x} \in \mathbb{R}^{2n}$, and A and B are $2n \times 2n$ matrices.

Assume that $\mathbf{u}$ is to be held constant during the time interval from 0 to T. This is likely the most realistic case since real-time information about $\mathbf{s}$ and $\mathbf{c}$ can be difficult to acquire, and management will often want to implement a constant control strategy for the season,

adjusting only periodically as further studies are carried out and completed. Consequently, we will not concern ourselves with transients in the solution, but rather attempt to minimize the unique attracting state of $||\mathbf{c}||$ for a given matrix A.

THEOREM 2. *Consider the problem*

$$||\tilde{\mathbf{C}}^*|| \to \inf$$

$$\frac{d\mathbf{x}}{dt} = A\mathbf{x} + B(\mathbf{u})\mathbf{x}$$

$$\mathbf{x}(0) = \mathbf{x}_0$$

$$||\mathbf{u}|| \leq K, \ K > 0$$

where $\mathbf{x}$*,* A*, and* B *are as defined in problem (48) and*

$$(A + B(\mathbf{u}))\tilde{\mathbf{x}}^* = (A + B(\mathbf{u})) \begin{pmatrix} \tilde{\mathbf{S}}^* \\ \\ \tilde{\mathbf{C}}^* \end{pmatrix} = 0, \ \textit{s.t.} \ \sum_{i=1}^{2n} \tilde{x}_i = N.$$

Then the optimal, constant control $\mathbf{u}$ *which minimizes* $||\tilde{\mathbf{C}}^*||$ *is given by*

$$\mathbf{u} = K\frac{\tilde{\mathbf{u}}}{||\tilde{\mathbf{u}}||} \quad \textit{where} \quad \tilde{u}_i = \begin{cases} \frac{\beta_i(C_i^* + S_i^*)}{C_i^*} & \textit{if } C_i^* \neq 0 \\ \\ 0 & \textit{if } C_i^* = 0 \end{cases}$$

where

$$A\mathbf{x}^* = A \begin{pmatrix} \mathbf{s}^* \\ \\ \mathbf{c}^* \end{pmatrix} = 0, \ \textit{s.t.} \ \sum_{i=1}^{2n} x_i = N$$

is guaranteed to have a unique solution $\mathbf{x}^*$ *by Theorem 1.*

PROOF. Let $0 < \varepsilon < 1$, and let $\tilde{\mathbf{x}} = \mathbf{x} + \varepsilon\mathbf{y}$ with $\mathbf{y_1}, \mathbf{y_2} \in \mathbb{R}^n$ so that

$$\mathbf{y} = \begin{pmatrix} \mathbf{y_1} \\ \\ \mathbf{y_2} \end{pmatrix}.$$

Our goal is to perturb the matrix A by $\varepsilon B(\mathbf{u})$, affecting a shift of $\mathbf{x}^*$ by $\varepsilon\mathbf{y}$ in the desired direction that minimizes $||\mathbf{c}^* + \varepsilon\mathbf{y_2}||$ and satisfies the equations

$$(A + \varepsilon B)(\mathbf{x}^* + \varepsilon\mathbf{y}) = 0 \tag{49}$$

and

$$\sum_{i=1}^{2n} x_i^* + \varepsilon y_i = N. \tag{50}$$

Multiplying out equation (49), we have the relation

$$B\mathbf{x}^* + A\mathbf{y} + \varepsilon B\mathbf{y} = 0. \tag{51}$$

The Fredholm alternative implies that this equation is solvable for $\mathbf{y}$ as long as for every $\mathbf{z} \in \mathbb{R}^{2n}$ such that $\mathbf{z}^T(A + \varepsilon B) = 0$, it follows that $\mathbf{z}^T B\mathbf{x}^* = 0$. But because this statement implies that $\mathbf{z}^T A = -\varepsilon\mathbf{z}^T B$, we have $\mathbf{z}^T B\mathbf{x}^* = -\mathbf{z}^T A\mathbf{x}^*/\varepsilon = 0$ and $\varepsilon > 0$, since $\mathbf{x}^* \in \ker(A)$. Thus a solution always exists, and furthermore, since the columns of $(A + \varepsilon B)$ sum to zero, the solution is not unique from this equation alone.

Now since we wish to minimize $||\mathbf{c}^* + \varepsilon\mathbf{y_2}||$ with $0 < \varepsilon < 1$, we will choose $\mathbf{y_2} = -\mathbf{c}^*$ and attempt to find $\mathbf{u}$ so that a vector $\mathbf{y}$ of this sort will satisfy equation (51). This task is made easier if we first solve for $\mathbf{y_1}$. Using the block form of A in equation (51), we can form the

system of equations

$$(H - \text{diag}(\boldsymbol{\beta}))\mathbf{y_1} + (\boldsymbol{\nu}\boldsymbol{\mu}^T)\mathbf{y_2} + \text{diag}(\mathbf{u})\mathbf{c}^* + \varepsilon\text{diag}(\mathbf{u})\mathbf{y_2} = 0 \tag{52}$$

$$\text{diag}(\boldsymbol{\beta})\mathbf{y_1} + (G - \text{diag}(\boldsymbol{\mu}))\mathbf{y_2} - \text{diag}(\mathbf{u})\mathbf{c}^* - \varepsilon\text{diag}(\mathbf{u})\mathbf{y_2} = 0.$$

Adding these two equations and recalling that $H = G - \text{diag}(\boldsymbol{\mu}) + (\boldsymbol{\nu}\boldsymbol{\mu}^T)$, we find that

$$H(\mathbf{y_1} + \mathbf{y_2}) = 0.$$

By our choice of $\mathbf{y_2} = -\mathbf{c}^*$, we now have that $\mathbf{y_1} = \mathbf{c}^* + \mathbf{v}$, where $\mathbf{v} \in \ker H$. Furthermore, by equation (50) and the fact that the elements of $\mathbf{x}^*$ sum to N due to conservation in the system given by A, $\sum y_i = 0$. Thus $\sum v_i = 0$ as well. But it was previously shown in the proof of Lemma 1 that $H + I$ satisfies the Perron-Frobenius theorem with a unique positive eigenvalue of 1, which implies that there exists a corresponding eigenvector such that all components of the vector are positive. Since this eigenvector must be in the kernel of H and the kernel of H is of dimension 1, the components of $\mathbf{v}$ must all have the same sign. Thus $\mathbf{v} = \mathbf{0}$, and $\mathbf{y_1} = \mathbf{c}^*$.

Returning to the second equation in system (52), we now have

$$\text{diag}(\boldsymbol{\beta})\mathbf{c}^* - (G - \text{diag}(\boldsymbol{\mu}))\mathbf{c}^* - \text{diag}(\mathbf{u})\mathbf{c}^* + \varepsilon\text{diag}(\mathbf{u})\mathbf{c}^* = 0.$$

Using the fact that $A\mathbf{x}^* = 0$, $\text{diag}(\boldsymbol{\beta})\mathbf{s}^* = -(G - \text{diag}(\boldsymbol{\mu}))\mathbf{c}^*$, so we arrive at

$$\text{diag}(\boldsymbol{\beta})\mathbf{c}^* + \text{diag}(\boldsymbol{\beta})\mathbf{s}^* - \text{diag}(\mathbf{u})\mathbf{c}^* + \varepsilon\text{diag}(\mathbf{u})\mathbf{c}^* = 0.$$

This system is now completely uncoupled, so whenever a $\mathbf{c}^*$ component satisfies $c_i^* \neq 0$, we arrive at the solution

$$u_i = \frac{\beta_i(c_i^* + s_i^*)}{c_i^*} \cdot \frac{1}{1-\varepsilon}.$$

Since $y_{2_i} = 0$ whenever $c_i^* = 0$, control of this component has no effect, so we can theoretically set it to anything we want. However, since we wish to maximize ε by controlling over non-zero c_i^* components, and we are constrained by the condition $||\mathbf{u}|| \leq K$, we must set $u_i = 0$ whenever $c_i^* = 0$ to minimize the effect of these components on the norm. Thus

$$||\mathbf{u}|| = ||\{u_i\}_{i=1}^n|| \leq K,$$

which we maximize to give us our result. □

4.4.4. Control on the underlying model. On the underlying model, control at minimum involves a kill function for the invader Y. Let $\varphi(x,t,Y)$ be a control to be applied to location x at time t, possibly dependent on the value of Y at that location on the underlying model. This control represents a reduction rate on the current infected population $Y(x,t)$. We incorporate this control with the equations

$$\begin{aligned}
\frac{d\mathbf{s}}{dt} &= [H - \operatorname{diag}(\boldsymbol{\beta}(Y))]\mathbf{s} + (\boldsymbol{\nu}\boldsymbol{\mu}^T)\mathbf{c} \\
\frac{d\mathbf{c}}{dt} &= \operatorname{diag}(\boldsymbol{\beta}(Y))\mathbf{s} + [G - \operatorname{diag}(\boldsymbol{\mu})]\mathbf{c} \\
\frac{\partial Y}{\partial t} &= F(Y,x) + r\mathbf{w}(x)\cdot\mathbf{c} - \varphi(x,t,Y)Y \qquad (53) \\
\boldsymbol{\beta}(Y) &= \gamma \int_\Omega Y(x,t)\mathbf{w}(x)dx \\
||\varphi(x,t)||_{L^2(\Omega)} &\leq K \ \forall t \text{ s.t. } t_0 \leq t \leq t_1
\end{aligned}$$

For the herbaceous invader model, we also apply this control to L,

$$
\begin{aligned}
\frac{d\mathbf{s}}{dt} &= [H - \operatorname{diag}(\boldsymbol{\beta}(Y))]\mathbf{s} + (\boldsymbol{\nu}\boldsymbol{\mu}^T)\mathbf{c} \\
\frac{d\mathbf{c}}{dt} &= \operatorname{diag}(\boldsymbol{\beta}(Y))\mathbf{s} + [G - \operatorname{diag}(\boldsymbol{\mu})]\mathbf{c} \\
\frac{\partial L}{\partial t} &= r\mathbf{w}(x)\cdot\mathbf{c} - \sigma L(x,t)h(Y,x) - \delta L - \varphi(x,t,Y)L \\
\frac{\partial Y}{\partial t} &= \epsilon F(Y,x) + \sigma L(x,t)h(Y,x) - \varphi(x,t,Y)Y \\
\boldsymbol{\beta}(Y) &= \gamma \int_\Omega Y(x,t)\mathbf{w}(x)dx \\
||\varphi(x,t)||_{L^2(\Omega)} &\leq K \ \forall t \text{ s.t. } t_0 \leq t \leq t_1.
\end{aligned} \tag{54}
$$

4.4.5. Numerical results. Control regimes implemented on the cheatgrass model of Rocky Mountain National Park have yielded extremely limited results. Optimized graph based control effectively reduces the amount of carrier vectors on the transportation network, but has almost no discernable affect on the underlying model output. This failure to suppress the invader is caused by early establishment of populations around the geographical node locations. Graph based control cannot hope to yield perfect results, and in realistic scenarios, some amount of propagules still make it to all the node locations, often even before control is even implemented. Once a presence probability is established at a location, graph based control has no suppressing effect on this population, and it spreads as normal. Since the invader has a high growth rate, any delay that the graph control might have caused is negligible, and no effect of the control is noticeable after as short a time as two years.

Since optimal control for the underlying model is not available, the choice of control function $\varphi(x,t,Y)$ is not immediately clear. In our numerical trials, we arbitrarily chose to implement control around the graph nodes in areas where the product of suitability

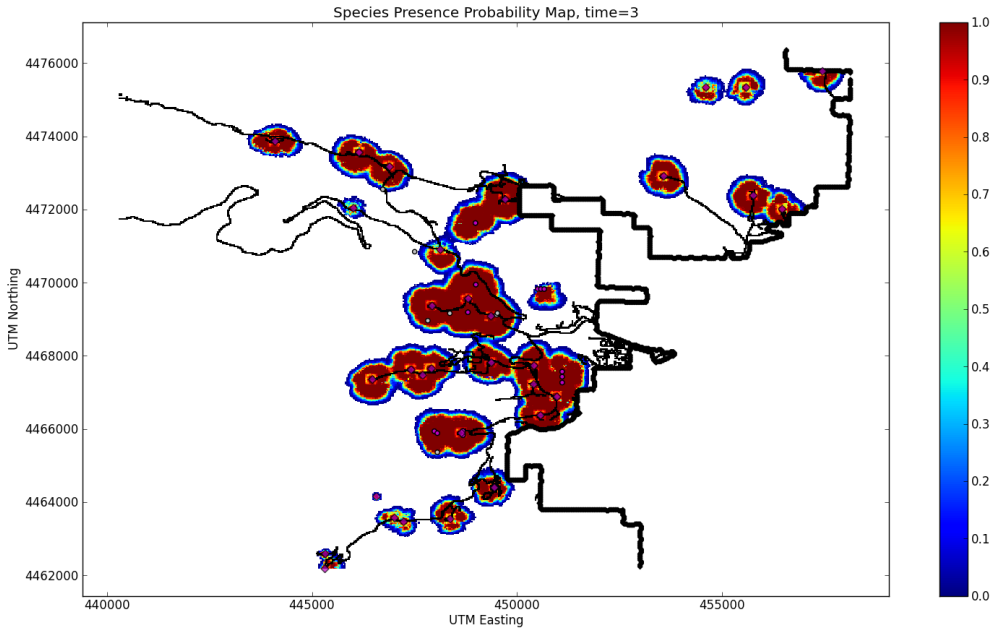

FIGURE 4.4. Projected results for the control regime φ on the underlying model in Rocky Mountain National Park. Results were based off of 1996 presence data, with the model running until 1999. Without control, the range of the species is projected to be nearly identical.

probability and species presence probability was greater than 0.1. Control decreased presence probability by a factor of 0.95 and latent seeds by a factor of 0.65 when applied, and could be used over an area the size of three American football fields. Control was applied once near the beginning of each year (one tenth of the way into the growing season), and suppressed new presence for some time, losing effect later in the growing season.

Figure 4.4 shows the results of this control regime applied on the 1996 Rocky Mountain National Park presence data until 1999. This period was chosen (rather than the 1999-2008 period of Figure 4.2) for its brevity. While the effects of the control are visible in areas nearby node locations, the breadth of the spread model is nearly identical to model output without control, suggesting that this regime was a failure at delaying the spread of the invader. One likely reason for this result is the units of the model. Since this model shows presence probability rather than population density, it is hard to know exactly what

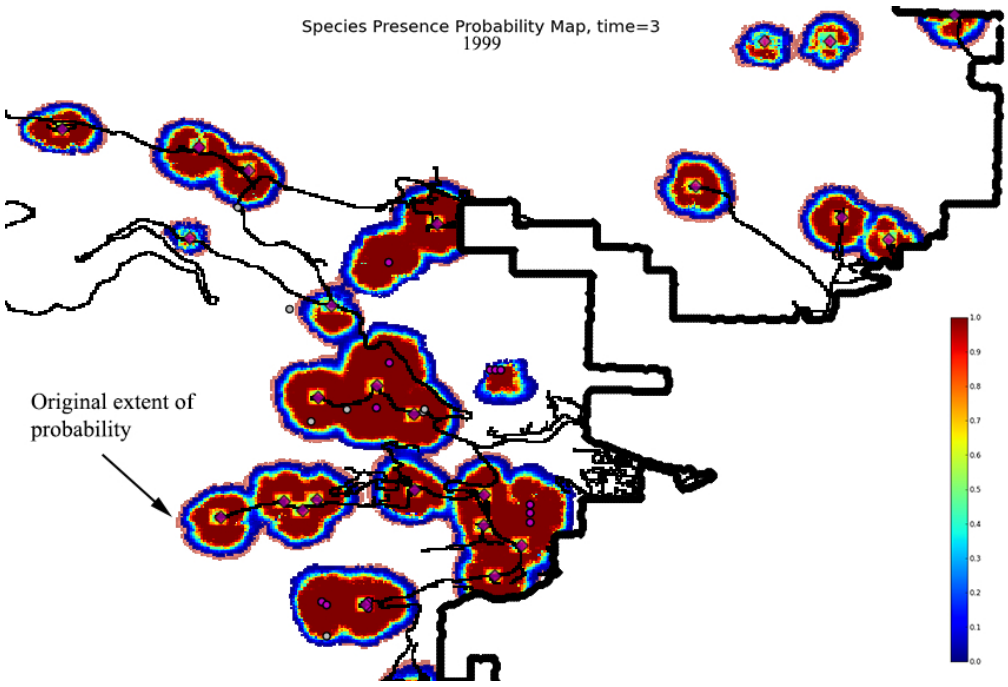

FIGURE 4.5. Projected results for the control regime φ and optimal graph based control in a central area of Rocky Mountain National Park. Results were based off of 1996 presence data, with the model running until 1999. The extent of the probability results without control is shown in pink.

happens to population numbers when control is applied. Local presence probability is heavily reduced, but this fact does not grant us much confidence that the invasion has ceased to spread nearby. Even low probabilities can continue to spread, and as they spread, they grow - only a small patch of cheatgrass is needed to quickly resume the invasion. Thus rare events can combine to continue the model's propagation through the domain.

To give the reader a better idea of what kind of invasive suppression occurs with these controls, Figure 4.5 plots the projected result of the control described above implemented alongside optimal graph based control (with $||\mathbf{u}|| \leq 200$) for an area of Rocky Mountain National Park. The pink regions around the non-zero presence probability areas show the extent of model spread without any control over the same period.

4.5. Discussion

Using the *Bromus tectorum* invasion of Rocky Mountain National Park as motivation, we have introduced with equation (43) and (44) a simple, linear model for invasive spread on a transportation network through the action of carrier vectors. Biological invasions rarely occur without human interaction, and understanding long distance dispersal events is critical for accurately assessing the exposure of remote locations to short-term infection. By using the natural graph structure of a transportation network, we can begin to model the probability of spread to remote locations, informing the steps needed to stop the invasion.

While our model predicted significant difficulty controlling *Bromus tectorum* in Rocky Mountain National Park, other methods have yet to be considered. Using the model results, it may be possible to focus control efforts on key network connections leading to as of yet uninvaded sites, especially those which are projected to be highly susceptible to transportation of *Bromus* due to a close proximity of high presence probability. Another option is to alter the transportation network itself, avoiding infected sites entirely by placing nodes elsewhere. Even if that is impossible, our results can be used in infrastructure planning to avoid placing nodes in sensitive, but as of yet uninfected, areas that would then be newly connected to the *Bromus* invasion.

In addition to the results shown in this chapter, we see relevance for our model on a continental or even global scale, where the transportation network takes the form of commercial shipping routes or airline connections. The modular form of equations (43) and (44) should allow for simple implementation with existing spatial models in both epidemiological and ecological settings, and with an optimal control available, management can quickly prioritize resources to the affected nodes for possible containment. Similarly, we envision applications

on a microbiological scale as well, including applications to cancer tumor growth and possible spread to other organs.

CHAPTER 5

Towards a continuous population model for natural language vowel shift

5.1. Introduction

> POINS: Come, your reason Jack, your reason.
>
> FALSTAFF: What, upon compulsion? No: were I at the Strappado, or all the Racks in the World, I would not tell you on compulsion. Give you a reason on compulsion? If Reasons were as plentie as Blackberries, I would give no man a Reason upon compulsion, I.

This comparison of reasons and blackberries from Shakespeare's Henry IV,[1] written in the year 1596, is better understood if one knows that *reasons* in Shakespeare's day was pronounced similarly to the current standard American pronunciation of the word *raisins*. Such difficulties in understanding Shakespeare can due to changes produced by the *Great English Vowel Shift*, which remodelled the English vowel system between the 16th and 19th centuries. But problems can occur not only in trying to read the works of the old masters. Consider the following story (adapted from [68]):

> Bernice visits her great grandmother and tells her that her baby will be a boy. "Oh!," says the grandmother, "why not name him Ian (pronounced ee-yun) after my great uncle?" "But, it is a boy!" exclaims Bernice, "I won't give him a girl's name!" Bernice's great grandmother, probably like you, is confused.

[1]Henry IV, Part I, Act 2, Scene 4, 196–200; e.g. [67] p. 139.

The explanation to Bernice's story lies in a vowel system change that is much more recent than, but equally dramatic as, the Great English Vowel Shift. First studied by William Labov *et al.* [69], this set of changes, called the *Northern Cities Vowel Shift* (NCVS) is currently underway in some cities on the United States side of the Great Lakes region. One component of the NCVS is the replacement of the vowel sound in *Ann* as spoken in Standard American English with the vowel sound at the beginning of *Ian*. As this change has affected the speech of those in Bernice's generation, she misunderstands her grandmother's older form of speech.

To understand vowel changes such as the NCVS, we need to consider the parameters that define a vowel. Vowels are classified according to how they are produced by the tongue and lips. Drawing a stylized representation of the mouth cavity as a trapezoid as in Fig. 5.1 (the teeth being the leftmost line, the roof of the mouth the top line), linguists give various vowel sounds symbols which are placed in the diagram at the tongue position used to make that sound. For example, the vowel of *Ann* as spoken in Standard American English is represented by [æ] and placed in the lower left region of the vowel trapezoid, and the initial vowel sound [i] of Ian is placed in the upper left region. One says that the NCVS "raises" the vowel of *Ann* from [æ] to [i], as the change corresponds to moving up in the trapezoid. Similarly, vowels can be "lowered," "fronted," or "backed" according to their movements within the diagram, "fronting" corresponding to moving (leftwards) towards the teeth.

A key to understanding vowel changes such as the Great English Vowel Shift or the Northern Cities Vowel Shift is to realize that the phonetic space represented by the vowel trapezoid is a continuous space. The raising of [æ] to [i] occurs phonetically gradually; at the beginning of the change, speakers may utter a vowel sound closer to the [æ] as in Fig. 5.1,

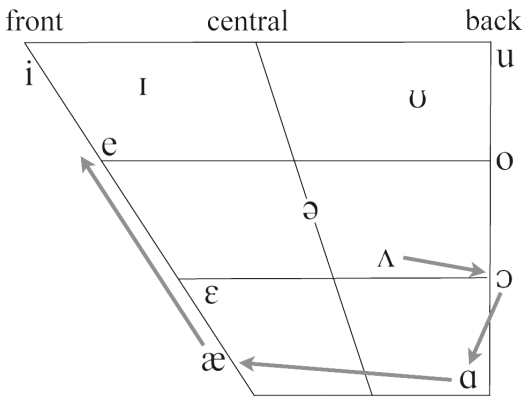

FIGURE 5.1. A vowel trapezoid. The vertical axis represents vowel height (the vertical position of the tongue relative to the mouth of the roof during articulation), and the horizontal axis represents vowel backness (the horizontal position of the tongue relative to the back of the mouth during articulation). The various symbols represent the sounds produced at the various positions. Arrows represent the Northern Cities Vowel Shift described in the text.

and as the change proceeds the typical position of the tongue moves up closer to the position of [i] in Fig. 5.1. However, the change is lexically abrupt, meaning, in our example, that the vowel [æ] will be raised in *all* words in which that vowel appears, the only possible conditioning coming from the phonetic environment. That is, the change may proceed, for example, more quickly when the vowel is uttered after a consonant [p] sound than after a [b]. Such phonetically gradual but lexically abrupt vowel changes are termed *regular* changes. Also attested are *irregular* vowel changes that are phonetically abrupt but lexically gradual—there is a sudden change from one vowel to a very different vowel, but the change first occurs in a few words where the original vowel appears, and eventually the change may spread to all words in the lexicon with the original vowel. This later type of change is also called *lexical diffusion*. According to Labov [70], regular changes are commonly associated with raising, lowering, fronting, and backing within the phonetic trapezoid, whereas changes in vowel length commonly occur through lexical diffusion.

To elucidate the factors that are involved in regular vowel changes, let us consider in more detail the NCVS. The NCVS is a change currently underway in a region of the northern United States centred around the Great Lakes, stretching from Chicago to Buffalo. The raising of [æ] to [i] mentioned above is the first step in a chain of changes associated to the NCVS. In Fig. 5.1 are diagrammed the vowels [ʌ] of *stuck*, [ɔ] of *stalk*, [ɑ] of *stock*, and [æ] of *stack* as spoken in Standard American English. The raising of [æ] to [i] leaves a "hole" in the phonetic space; there is not any more a vowel in the system between [æ] and [i]. The second step of the NCVS remedies this situation by backing [ɑ] to [æ]; so, *stock* after the change is pronounced as *stack* was before the change. The process then proceeds with the lowering of [ɔ] to [ɑ] so that *stalk* becomes pronounced as *stock* before the change, and the backing of [ʌ] to [ɔ] so that *stuck* becomes pronounced as *stalk* before the change. Such a series of movements is often referred to as a "chain shift" as the initial movement of [æ] resulted in the jostling of the other vowels in the system. The end result of this chain shift is that the original system of five vowels has moved around in a circle in phonetic space.

The work of Labov *et al.* [69] and later, for example, Eckert [71] that documents the NCVS relies on the notion of *apparent time*. That is, one assumes that the oldest speakers in a population still speak the language that they learned when they were young. If one finds in a population many great grandmothers who say *stalk* with the standard American pronunciation, many younger people such as Bernice who pronounce *stalk* as *stock*, and middle-aged speakers uttering a vowel somewhere in between the two extremes, one concludes that a change has occurred in the population over the great grandmother's lifetime. This assumption is not without complications; speakers may take on different forms of speech according to social norms as they age. Furthermore, juveniles also show developmental

differences that distinguish them from the general population; Labov [70] states children of age eight to be the appropriate youngest subjects of apparent-time phonetic change studies.

The changes associated to the NCVS have not proceeded to the same degree in all regions affected by the change; cities such as Chicago or Cleveland show the greatest advancement of the change, whereas only the first stages of the change are noticed in cities such as Pittsburgh or Indianapolis. Also, the NCVS is not reported to have affected African American English. It is also important to note that the change is not easily noticed by the communities in which it occurs; sound change is a gradual process that requires, in the case of the NCVS, three to four generations for completion [70].

Our objective in this chapter is to develop an age-structured mathematical model for vowel pronunciation in a natural language community as a basis for understanding vowel learning and regular changes in vowel systems. The model is not appropriate for changes through lexical diffusion. We develop a framework in which one may consider the question of how vowel systems that have, as in the case of American English, been stable to hundreds of years, begin to continuously change so that significant differences are noticed within a few generations. This contrasts with current models for vowel systems such as those of Liljencrants and Lindblom [72], Schwartz *et al.* [73, 74], and de Boer [75], who model the emergence of vowel systems, considering the number of vowels and their spacing in the vowel parallelogram.

Our model must be consistent with the fact that not all speakers (even of the same age) have identical vowel pronunciation. Particularly since Kuhl's studies of perception in humans and animals [76, 77, 78], this distributional aspect of of speech utterance and perception [79, 80, 81] has been of interest to linguists. Infants in particular have been shown

to be perceptive of distributional aspects of language as they learn to speak [82, 83, 84] and to be able to distinguish phonetic contrasts that adults do not distinguish [85].

Three simplifying assumptions of the model we present are the following:

(1) *The rate of change of a vowel is independent of its phonetic environment.* This means that the rate at which the pronunciation of a particular vowel changes is supposed to be the same for all words. This is not quite true in practice, for surrounding phonemes may sometimes enhance the vowel change [86]. However, as regular vowel changes typically affect, in the end, all instances of a particular vowel regardless of the phonetic environment, this assumption is justifiable for a simple model.

(2) *The chain shift is rigid.* In other words, all elements of the vowel system are supposed to undergo simultaneously the same shift. Consequently, the sound change in a single vowel defines uniquely the shift in the whole vowel system. It is evident that this is a severe simplification, which is justified only because of its resulting mathematical simplicity: in a general situation we would need to model the shift of each vowel separately, together with the vowel–vowel interactions ("knock-on effects") by drag- and push-chain mechanisms [87].

(3) *Age is the only social structure.* We will not consider, for example, differences between urban and rural speakers or social status in a community.

The object of our study is an age- and phonetic structured population, such that its individuals are distinguished by their age and vowel system sound (pronunciation). The processes of ageing, birth and death follow the standard demographic model by McKendrick

and von Foerster [1], while the evolution of language by teaching and social interaction is based on the theory of mixtures with continuous diversity proposed by Faria [88].

The structure of the remainder of this paper is as follows: The fundamental variables of the model are introduced in Sect. 5.2. General balance equations are explained in Sect. 5.3, and A.1 gives a detailed derivation of these equations. In Sects. 5.4 and 5.5 (supplemented by A.2) we propose explicit expressions for terms in the equations that describe factors such as the influence of speakers on speakers of other ages, and in Sect. 5.6 we give boundary conditions. In Sect. 5.7, we derive an approximation of a stationary solution to the balance equations, corresponding to a state of phonetic equilibrium in which speakers of all ages share the same distribution of vowel pronunciation, and we define a parameter that measures the variance in this distribution. Via a numerical simulation we show how a symmetric initial condition evolves to the equilibrium solution of the same mean vowel pronunciation. Motivated by a common suggestion of why vowel shift occurs, we model in Sect. 5.8 a situation in which the initial population consists of a majority pronunciation that is affected by an immigrant minority with a different mean vowel pronunciation. Such initial conditions lead to vowel shift in numerical simulations of the model. Finally, we close in Sect. 5.9 with perspectives on using the model to understand factors impacting vowel shift. This chapter is published in the Journal of Theoretical Biology [8].

5.2. Fundamentals

We define all sounds of a vowel system through the position of a single vowel in the two-dimensional phonetic space, represented by the vowel trapezoid of Fig. 5.1. Vowels may be quantitatively described by their formants (frequencies of highest energy), as the two dominant formants have been shown to correlate to the perception of vowels [89]. Accordingly,

in this model we assume that the phonetic space is a circle in the space of the two dominant formants. That is, sound changes occur in a closed loop—mathematically represented by the unit circle $\mathcal{S}^1$—instead of a two-dimensional region of the phonetic space. The generalization to the latter case is nonetheless straightforward. Representing vowel space as a closed loop will also lead us to take periodic boundary conditions in σ (see Sect. 5.6), but we acknowledge that the start and end of a vowel chain do not necessarily interact cyclically (cf. discussion in Sect. 5.9). We also assume that any speaker has a definite vowel pronunciation, which may be interpreted as the average over the naturally variable pronunciation of that speaker; we return to this point later, in Sect. 5.9.

All speakers of same age and utterance constitute what is called a *species*. Their number is specified by the product $n^*(\sigma, a, t)\,\mathrm{d}\,a\,\mathrm{d}\,\sigma$, which represents the number of speakers aged a to $a + \mathrm{d}\,a$ and uttering vowel sounds in the phonetic interval σ to $\sigma + \mathrm{d}\,\sigma$, at time t. Thus, the most fundamental quantity of the theory is the *speakers number density*[2]

$$n^*(\sigma, a, t) \;:\; \mathcal{S}^1 \times \mathcal{A} \times \mathbb{R} \to \mathbb{R}, \quad \text{with} \quad \mathcal{A} := [0, 1] \subset \mathbb{R}. \tag{55}$$

In the expressions above, $\sigma := \theta/(2\boldsymbol{\pi}) \in \mathcal{S}^1$ is the *vowel sound*, mathematically represented as a normalized arc length of the closed *vowel loop* $\mathcal{S}^1$, with corresponding angle $0 \leq \theta < 2\boldsymbol{\pi}$. Further, $a := x/L \in \mathcal{A}$ is the *dimensionless age*, which specifies a point in the *age space* $\mathcal{A}$, with x and L standing for the temporal (i.e. dimensional) age and the *maximum lifespan* of any speaker in the population, respectively. From these definitions, we immediately conclude

[2]In full *species density of the number of speakers*. Species densities are distribution densities on $\mathcal{S}^1 \times \mathcal{A}$.

that

$$\int_{\mathcal{S}^1} \mathrm{d}\,\sigma \,=\, \frac{1}{2\pi}\int_0^{2\pi} \mathrm{d}\,\theta \,=\, 1, \qquad \int_{\mathcal{A}} \mathrm{d}\,a \,=\, \frac{1}{L}\int_0^{L} \mathrm{d}\,x \,=\, 1, \tag{56}$$

$$\int_{\mathcal{S}^1}\int_{\mathcal{A}} \mathrm{d}\,a\,\mathrm{d}\,\sigma \,=\, \frac{1}{2\pi L}\int_0^{2\pi}\int_0^{L} \mathrm{d}\,x\,\mathrm{d}\,\theta \,=\, 1. \tag{57}$$

Additionally, the double integral

$$N\,|_{\sigma_0,a_0}^{\sigma_1,a_1}\,(t) := \int_{\sigma_0}^{\sigma_1}\int_{a_0}^{a_1} n^*(\sigma,a,t)\;\mathrm{d}\,a\,\mathrm{d}\,\sigma \tag{58}$$

specifies the number of speakers aged between a_0 and a_1 and uttering vowel sounds between σ_0 and σ_1, at time t. Likewise, we define the *population size* (i.e. the total number of speakers in the population) by

$$N := \int_{\mathcal{S}^1}\int_{\mathcal{A}} n^*(\sigma,a,t)\;\mathrm{d}\,a\,\mathrm{d}\,\sigma, \qquad N = \text{constant}. \tag{59}$$

For simplicity, we suppose N to be constant in time (i.e., there is a stable population of stationary size, with vanishing Malthusian parameter); otherwise time changes of the total number of speakers could disguise vowel shifting.

Finally, we introduce also the supplementary densities

$$n^{\triangle}(a,t) := \int_{\mathcal{S}^1} n^*(\sigma,a,t)\;\mathrm{d}\,\sigma, \qquad n^{\blacktriangle}(\sigma,t) := \int_{\mathcal{A}} n^*(\sigma,a,t)\;\mathrm{d}\,a \tag{60}$$

associated with the total number density of speakers aged a and the total number density of speakers uttering the sound σ, respectively. From this point on, we shall consider only

stable populations with *stationary age-structure*, defined by the constraint

$$\frac{\partial n^{\triangle}}{\partial t} \equiv 0 \quad \Rightarrow \quad n^{\triangle}(a). \tag{61}$$

5.3. Phonetic process and balance equations

The objective of this theory is the determination of governing equations for the time-evolution of the vowel system in a steady, age-structured population. As discussed above, the vowel sound distribution in such a population is described by the speakers number intensity n^*. Its evolution is dictated by the system of equations proposed below, which can be seen as a generalization of the celebrated McKendrick–von Foerster equation [1] according to the principles of the theory of mixtures with continuous diversity [88]. We derive in Sec. A.1 the balance equations of

- speakers number

$$\frac{\partial n^*}{\partial t} + \frac{1}{L}\frac{\partial n^*}{\partial a} + \frac{\partial n^* u^*}{\partial \sigma} = -n^*\mu^* + n^*\Gamma^*, \tag{62}$$

- transition impetus

$$\frac{\partial n^* u^*}{\partial t} + \frac{1}{L}\frac{\partial n^* u^*}{\partial a} + \frac{\partial n^* {u^*}^2}{\partial \sigma} + \frac{\partial \varphi^*}{\partial a} + \frac{\partial \phi^*}{\partial \sigma} = n^*\kappa^*. \tag{63}$$

Besides the speakers number density n^*, we identify in (62) the *phonetic transition rate* u^*, the age-dependent *mortality rate per capita* μ^*, and the *rate of abrupt sound change per capita* Γ^*. The transition rate $u^*(\sigma, a, t)$ describes the mean rate per capita at which speakers aged a and uttering the sound σ gradually change their pronunciation. Thus, u^* corresponds to a kind of "velocity" in the phonetic space; that is, it describes *regular* (continuous) sounds

changes. In contrast, the rate Γ^* models *irregular* (discontinuous) sound changes, which may be interesting in connection with certain phenomena like lexical diffusion, but are unlikely to have relevance for vowel shift. Thus for the purposes of this theory we may set[3]

$$\Gamma^*(\sigma, a, t) \equiv 0. \tag{64}$$

Equation (63) is new; it relates the evolution of the transition rate u^*—more precisely the product n^*u^*, called *phonetic transition impetus*—to the *phonetic stresses* φ^* and ϕ^*, as well as the *global stimulus rate per capita* κ^*. All these three quantities describe multi-age interactions in the phonetic space. The phonetic stresses φ^* and ϕ^* are associated to interactions between *familiar species*, i.e. between speakers with similar ages and utterances. In contrast, the global stimulus rate κ^* stands for interactions between speakers with distinct ages and utterances.

The fundamental equations of the theory are then (62) and (63), the solution $\{n^*, u^*\}$ of which defines a *phonetic process*. Nevertheless, the system (62) and (63) is not closed yet, for the quantities μ^*, φ^*, ϕ^* and κ^* are still unknown. This closure problem is solved in Sects. 5.4 and 5.5.

[3]It must be emphasized that *deliberate* sound changes, consciously adopted in particular situations, are *not* considered in the model. Rather, we are interested only in *natural* sound changes. Thus, according to the usual hypothesis that vowel shift is characterized by a regular sound change [70, 90], we conclude that (64) should hold.

5.4. Closure and phonetic functionals

The starting point to close the system (62) and (63) is to introduce general functional relations of the form

$$F_\gamma(\sigma, a, t) = \mathfrak{H}_\gamma\left(\sigma, a, t\,; \sigma',\, a',\, n^\circ,\, u^\circ\right), \qquad \gamma = 1, 2, 3, 4,$$
$$F_1 := \mu^*, F_2 := \varphi^*, F_3 := \phi^*, F_4 := \kappa^*, \tag{65}$$

where $n^\circ := n^*(\sigma', a', t)$ and $u^\circ := u^*(\sigma', a', t)$, with $\sigma' \in \mathcal{S}^1$ and $a' \in \mathcal{A}$.

Equation (65) states that the phonetic functions F_γ are determined by the (non-local) phonetic functionals $\mathfrak{H}_\gamma$, in such a way that a speaker aged a and uttering vowel σ may be influenced by individuals of all ages and pronunciations, taking into account their number and rate of pronunciation change. This kind of dependence on multiple sounds and ages suggests that the phonetic functionals $\mathfrak{H}_\gamma$ may contain intricate integrals over $\mathcal{S}^1$ and $\mathcal{A}$. It is evident that such complicated functional relations are not convenient for applications and should be simplified somehow. The objective of the remainder of this Section is to discuss such simplifications.

We begin by introducing the *hypothesis of phonetic objectivity*: the inherent behaviour of the speech community does not depend on absolute time or sound. Such an assumption is plausible as long as all speakers live in an unchanging environment and have no physiological or personal bias towards a particular pronunciation [91]. In order to apply this hypothesis to (65), one needs first to introduce the notion of *phonetic reference frame*: time and pronunciation are abstract notions that can only be specified with reference to particular time and sound, which together with the orientation (handedness) of the vowel loop $\mathcal{S}^1$ define a phonetic frame of reference. In this sense, phonetic objectivity is related to changes of

phonetic reference frame that preserve time lapses and sound intervals. The most general of such frame transformations is described by the equations

$$\hat{\sigma} = s\,\sigma + \varsigma, \qquad \hat{t} = t + \tau, \tag{66}$$

where $\hat{\sigma}$ and $\hat{t}$ are the sound and time identified with respect to the new reference frame, while $\varsigma(t) \in \mathcal{S}^1$ and $\tau \in \mathbb{R}$ denote shifts in sound and time references, respectively. The factor $s = \pm 1$ accounts for a possible change of handedness of the vowel loop $\mathcal{S}^1$, such that $s = -1$ denotes an inversion. From (66) there immediately follow the transformation rules

$$\hat{a} = a, \qquad \hat{u}^* = s\,u^* + \frac{d\varsigma}{dt}. \tag{67}$$

Seeing that the speakers number density n^* is an intrinsic property of the population, which should not depend on the choice of the phonetic reference frame, we have also

$$\hat{n}^* = n^*. \tag{68}$$

Thus, since the balance equations (62) and (63) must be invariant with respect to frame changes of the type (66), we derive

$$\hat{\mu}^* = \mu^*, \qquad \hat{\phi}^* = \phi^*, \qquad \hat{\varphi}^* = s\,\varphi^*, \qquad \hat{\kappa}^* = s\,\kappa^*. \tag{69}$$

By combining the hypothesis of phonetic objectivity with (66)–(69), we conclude that the functionals (65) cannot depend explicitly upon time t, whereas explicit dependencies on the

sound σ and the transition rate u^* are allowed only in terms of the differences

$$\sigma - \sigma' \qquad \text{and} \qquad u^* - u^\circ = u^*(\sigma, a, t) - u^*(\sigma', a', t). \tag{70}$$

Collecting all these results together, (65) reduces to

$$F_\gamma(\sigma, a, t) = \mathfrak{G}_\gamma\,(a\,; a', \sigma - \sigma',\, n^\circ, u^* - u^\circ)\,. \tag{71}$$

Further consequences of the hypothesis of phonetic objectivity will be exploited in Sect. 5.5.

An additional simplification we may invoke here is the *hypothesis of instantaneous adaptive response*: past phonetic profiles of the speech community do not affect its inherent behaviour. This hypothesis implies the absence of a collective memory, so that speakers perceive solely the current structure of the population, without prospective or retrospective dispositions. Mathematically, it simply implies the exclusion of u° from (71), leading to the reduced functional

$$F_\gamma(\sigma, a, t) = \mathfrak{F}_\gamma\,(a\,; a', \sigma - \sigma',\, n^\circ) = \mathfrak{F}_\gamma\,(a\,; a', \sigma - \sigma',\, n^*(\sigma', a', t))\,. \tag{72}$$

5.5. Explicit phonetic functions

To proceed towards explicit forms of (72) we may constrain the integrals in $\mathfrak{F}_\gamma$ to have the following general pre-defined form:

$$F_\gamma(\sigma, a, t) \;=\; \int\limits_{\mathcal{S}^1} \int\limits_{\mathcal{A}} V_\gamma\big(\sigma - \sigma'\big) W_\gamma\big(a, a'\big) \mathfrak{X}_\gamma\big(n^*(\sigma', a', t)\big)\, \mathrm{d}\, a'\, \mathrm{d}\, \sigma'. \tag{73}$$

We can determine particular expressions for $\mathfrak{X}_\gamma$ and the *influence kernels* V_γ and W_γ by exploiting the properties of the population under study as follows.

The constraints (59) and (61) of constant size and stationary structure of the population require that the *total mortality rate* M should be constant

$$M := \int_{\mathcal{S}^1} \int_{\mathcal{A}} n^* \mu^* \, \mathrm{d}\, a \, \mathrm{d}\, \sigma = \int_{\mathcal{A}} n^\triangle \mu^\triangle \, \mathrm{d}\, a = \text{constant} \, , \tag{74}$$

implying that

$$F_1(\sigma, a, t) \, := \, \mu^*(\sigma, a, t) \equiv \mu^\triangle(a), \tag{75}$$

where we assumed that the mortality rate is independent of pronunciation. Consequently, from (73) and (75) it follows

$$V_1\big(\sigma - \sigma'\big) \equiv 1, \qquad W_1\big(a, a'\big) \equiv \mu^\triangle(a), \qquad \mathfrak{X}_1\,(n^*) \equiv 1. \tag{76}$$

The explicit form of $\mu^\triangle(a)$ can be obtained from existing life tables.

Concerning the fluxes φ^* and ϕ^*, we must first realize that they describe *short-range interactions*, resulting that we can express the influence kernels V_2, W_2, V_3 and W_3 of (73) as *nascent delta functions* (e.g. narrow Gaussian bumps). It should be noticed, however, that the transformation properties (66)–(69) require φ^* to be an odd function of σ, while ϕ^* must be an even function of σ, implying that

$$V_2\big(\sigma - \sigma'\big) \equiv 0, \tag{77}$$

$$V_3\big(\sigma - \sigma'\big) \equiv \delta_\epsilon(\sigma - \sigma'), \qquad W_3\big(a, a'\big) \equiv \frac{\delta_\epsilon(a - a')}{\big[\tau_\phi(a)\big]^2}, \tag{78}$$

where $\delta_\epsilon(\cdot)$ denotes a nascent delta function and $\tau_\phi(a)$ is the *characteristic reaction time to phonetic stress* of the speakers aged a. For example, juveniles should have much shorter reaction times than elders.

In contrast, the global stimulus rate κ^* describes *long-range interactions* between speakers with distinct ages and/or pronunciations. From (66)–(69) it follows that it should be an odd function of σ, while no particular symmetry is required for its long-range age dependence. Owing to this, as well as to other reasons that will become apparent soon, we will assume here that the influence kernels V_4 and W_4 have the forms

$$V_4\big(\sigma'-\sigma\big) \equiv \sin\left[2\pi(\sigma'-\sigma)\right], \qquad W_4\big(a,a'\big) \equiv \frac{G(a,a')}{n^{\triangle}(a')\left[\tau_\kappa(a)\right]^2}, \tag{79}$$

with $\tau_\kappa(a)$ denoting the *characteristic reaction time to global stimuli*, while $G(a,a')$ is the dimensionless *inter-generational influence function*, which encodes the influence of speakers aged a' upon a speaker of age a. In particular, $G(a,a')$ should be large for $a' \simeq a$, assuming that a speaker's generation has a significant influence on the speaker's pronunciation. Another strong influence should come e.g. from the generations of parents [92, 93] and teachers.

Finally, choosing for simplicity's sake

$$\mathfrak{X}_3(n^*) \equiv n^* \equiv \mathfrak{X}_4(n^*), \tag{80}$$

we obtain from (77)–(80) the following explicit expressions for φ^*, ϕ^* and κ^*

$$F_2^*(\sigma,a,t) := \varphi^*(\sigma,a,t) = 0, \tag{81}$$

(82) $F_3^*(\sigma, a, t) := \phi^*(\sigma, a, t) =$

$$\frac{1}{[\tau_\phi(a)]^2} \int_{\mathcal{S}^1} \int_{\mathcal{A}} \delta_\epsilon(\sigma - \sigma') \delta_\epsilon(a - a')\, n^*(\sigma', a', t)\, \mathrm{d}\, a'\, \mathrm{d}\, \sigma' \simeq \tau_\phi^{-2}\, n^*,$$

(83) $F_4^*(\sigma, a, t) := \kappa^*(\sigma, a, t) =$

$$\frac{1}{[\tau_\kappa(a)]^2} \int_{\mathcal{S}^1} \int_{\mathcal{A}} \sin\left[2\pi(\sigma' - \sigma)\right] G(a, a') \frac{n^*(\sigma', a', t)}{n^\triangle(a')}\, \mathrm{d}\, a'\, \mathrm{d}\, \sigma'.$$

Note that we have taken $\delta_\epsilon(\cdot)$ to be the Dirac delta function $\delta(\cdot)$ in the final approximation of (82). While (81) and (82) are the appropriate equations for a zeroth-order theory, the construction of higher-order theories following similar arguments is straightforward. For instance, in A.2 we present the results of a linear first-order theory.

The reasons for choosing the sinusoidal expression (79) for V_4 now become evident. For instance, for a given age a_0 and time t_0, if the distribution of speakers is homogeneous, then the integral over $\mathcal{S}^1$ in (83) vanishes, that is:

(84) if $\quad n^*(\sigma', a_0, t_0) = n^\triangle(a_0), \quad$ then

$$\int_{\mathcal{S}^1} \sin\left[2\pi(\sigma' - \sigma)\right] n^*(\sigma', a_0, t_0)\, \mathrm{d}\, \sigma' = 0,$$

implying that speakers of this particular age cause no stimulus for phonetic change in the population. On the other hand, if all speakers with a particular age a_0 utter nearly the same sound σ_0 at time t_0, then they may provoke a strong stimulus of phonetic change on the

population towards the pronunciation σ_0:

(85) if $n^*(\sigma', a_0, t_0) = \delta_\epsilon(\sigma' - \sigma_0)\, n^\triangle(a_0)$, then

$$\int_{S^1} \sin\left[2\pi(\sigma' - \sigma)\right] n^*(\sigma', a_0, t_0)\, \mathrm{d}\sigma' \;\simeq\; n^\triangle(a_0)\, \sin\left[2\pi(\sigma_0 - \sigma)\right].$$

With these explicit expressions for the phonetic functions, we need now to determine the appropriate boundary conditions for the problem, in order to have a complete model. This is done in the next section.

5.6. Boundary conditions

In the phonetic space we require periodic boundary conditions with unitary period

$$\begin{gathered} n^*(\sigma, a, t) = n^*(\sigma + \Lambda, a, t), \quad u^*(\sigma, a, t) = u^*(\sigma + \Lambda, a, t), \\ \phi^*(\sigma, a, t) = \phi^*(\sigma + \Lambda, a, t), \quad \text{etc.} \quad \Lambda = 1, 2, 3, \ldots \end{gathered} \tag{86}$$

Further, since $a = 1$ represents the maximally achievable dimensionless age of any speaker, we impose the boundary conditions

$$n^*(\sigma, 1, t) \equiv 0 \equiv u^*(\sigma, 1, t). \tag{87}$$

Assuming that children are born without an intrinsic transition impetus,

$$u^*(\sigma, 0, t) \equiv 0. \tag{88}$$

It remains to determine a condition for the pronunciation profile $n^*(\sigma, 0, t)$ of children entering the speech community. New speakers learn from established speakers, so we impose

the condition

$$n^*(\sigma,0,t) = LB \int_{\mathcal{A}} G(0,a') \frac{n^*(\sigma,a',t)}{n^{\triangle}(a')}\,\mathrm{d}\,a', \tag{89}$$

where $G(0,a')$ is the inter-generational influence per capita introduced in (79), which gives in this case the collective influence of individuals aged a' on kids just entering the speech community. The constant B is the *total birth rate* of the population. The constraints (59), (61) and (75) imply that the total birth rate B should balance the (constant) total mortality rate M;

$$B := \frac{n^{\triangle}(0)}{L} = -M = \text{constant}. \tag{90}$$

Finally, assuming that the boundary condition (89) should be valid for any possible phonetic process $\{n^*, u^*\}$, it follows from (59)–(61), (89) and (90) that (choose e.g. $n^* = N$ or alternatively $n^* = n^{\triangle} n^{\blacktriangle}/N$)

$$A(0) := \int_{\mathcal{A}} G(0,a')\,\mathrm{d}\,a' = 1. \tag{91}$$

As a simple example, one may assume that children learn speech primarily from their parents, so that $G(0,a') = \delta(a'-\alpha)$, where δ is the Dirac delta function and α is the *average parental age* of individuals introducing kids into the speech community. Under this assumption, (89) reduces to

$$n^*(\sigma,0,t) = LB\,\frac{n^*(\sigma,\alpha,t)}{n^{\triangle}(\alpha)}. \tag{92}$$

5.7. Phonetic equilibrium

We define a state of *phonetic equilibrium* as any possible phonetic process $\{n^*|_{\mathrm{E}}, u^*|_{\mathrm{E}}\}$ such that the following conditions apply:

$$\frac{\partial n^*|_{\mathrm{E}}}{\partial t} \equiv 0, \qquad u^*|_{\mathrm{E}} \equiv 0, \tag{93}$$

where $|_{\mathrm{E}}$ denotes the phonetic equilibrium value of the respective quantity. In simple words, a population in phonetic equilibrium is phonetically static, that is, there is no speaker undergoing any sort of sound change.

5.7.1. Analytical equilibrium solution. We seek equilibrium solutions of (62) and (63) obeying the boundary conditions (87), (88) and (92). From (75), (81) and (93) it follows that, at equilibrium, the balance equations of speakers number (62) and transition impetus (63) reduce to

$$\frac{1}{n^*|_{\mathrm{E}}} \frac{\partial n^*|_{\mathrm{E}}}{\partial a} = -L\,\mu^{\triangle}, \tag{94}$$

$$\frac{\partial \phi^*|_{\mathrm{E}}}{\partial \sigma} = n^*|_{\mathrm{E}}\,\kappa^*|_{\mathrm{E}}. \tag{95}$$

Notice that $\mu^{\triangle}|_{\mathrm{E}}(a) \equiv \mu^{\triangle}(a)$, since this function has no phonetic dependence.

Seeing that the left-hand side of (94) depends on σ and a, while the right-hand side is solely a function of a, we conclude from (59)–(61) that (94) must have a solution of the form

$$n^*|_{\mathrm{E}}(\sigma, a) = \frac{n^{\triangle}(a)\, n^{\blacktriangle}|_{\mathrm{E}}(\sigma)}{N}, \tag{96}$$

where we have already used the fact that $n^{\triangle}|_{\mathrm{E}}(a) \equiv n^{\triangle}(a)$. The age structure $n^{\triangle}(a)$ is given by the stationary McKendrick–von Foerster equation

$$\frac{\partial n^{\triangle}}{\partial a} = -L\, n^{\triangle} \mu^{\triangle}, \tag{97}$$

which is just the outcome of (94) integrated over $\mathcal{A}$, cf. (60), (61), and (115).

Insertion of (82), (83) and (96) into (95) yields

$$\frac{1}{S|_{\mathrm{E}}\, n^{\blacktriangle}|_{\mathrm{E}}} \frac{\mathrm{d}\, n^{\blacktriangle}|_{\mathrm{E}}}{\mathrm{d}\,\sigma} \;=\; A \frac{\tau_{\phi}^2}{\tau_{\kappa}^2}, \tag{98}$$

where $\tau_{\phi}|_{\mathrm{E}}(a) \equiv \tau_{\phi}(a)$, $\tau_{\kappa}|_{\mathrm{E}}(a) \equiv \tau_{\kappa}(a)$ and

$$S|_{\mathrm{E}}(\sigma) := \frac{1}{N} \int\limits_{\mathbb{S}^1} \sin\left[2\pi(\sigma' - \sigma)\right] n^{\blacktriangle}(\sigma')\; \mathrm{d}\,\sigma', \tag{99}$$

$$A|_{\mathrm{E}}(a) \equiv A(a) := \int\limits_{\mathcal{A}} G(a, a')\; \mathrm{d}\, a'. \tag{100}$$

Since the left-hand side of (98) depends solely on σ, while the right-hand side depends only on a, we conclude that both sides of (98) must be equal to the same constant H. This constant is a dimensionless number, characteristic of the population, which gives a measure of the ratio of attractive to dispersive phonetic forces. It is defined by

$$H \;:=\; A \frac{\tau_{\phi}^2}{\tau_{\kappa}^2}, \tag{101}$$

such that (98) reduces to

$$\frac{\mathrm{d}\, n^{\blacktriangle}|_{\mathrm{E}}}{\mathrm{d}\,\sigma} \;=\; H\, n^{\blacktriangle}|_{\mathrm{E}}\, S|_{\mathrm{E}}. \tag{102}$$

It should be noted that H is also a useful measure of the stability of phonetic equilibrium for a given population. We expect H to be much larger than unity, seeing that $H \leq 1$ would correspond to a population that cannot find a consensus about pronunciation, because of the dominance of dispersive phonetic forces.

A simple, approximate (valid for large H) solution to (102) that satisfies the periodic boundary conditions is

$$n^{\blacktriangle}|_{\mathrm{E}}(\sigma) \simeq N\beta_H(\sigma - \sigma_0) = NC \exp\left(\frac{H}{2\pi}\cos\left[2\pi(\sigma - \sigma_0)\right]\right), \tag{103}$$

where

$$C^{-1} = \int_{S^1} \exp\left(\frac{H}{2\pi}\cos\left[2\pi(\sigma - \sigma_0)\right]\right) \mathrm{d}\,\sigma \tag{104}$$

is a normalization constant. Comparing $\beta_H(\sigma - \sigma_0)$ to a Gaussian nascent delta function

$$\delta_\nu(\sigma - \sigma_0) = \frac{1}{\sqrt{2\pi\nu}} \exp\left(-\frac{1}{2\nu}(\sigma - \sigma_0)^2\right) \tag{105}$$

by writing the Taylor series of $\cos\left[2\pi(\sigma - \sigma_0)\right]$ about σ_0 up to second order, we find that

$$\beta_H(\sigma - \sigma_0) \simeq \sqrt{H} \exp\left(-\pi H(\sigma - \sigma_0)^2\right), \tag{106}$$

i.e., a nascent delta function with variance $\nu = (2\pi H)^{-1}$. Two examples of β_H and δ_ν for $\nu = (2\pi H)^{-1}$ are graphed in Fig. 5.2. For large H then, the integral (99) with $n^{\blacktriangle}(\sigma)$ given

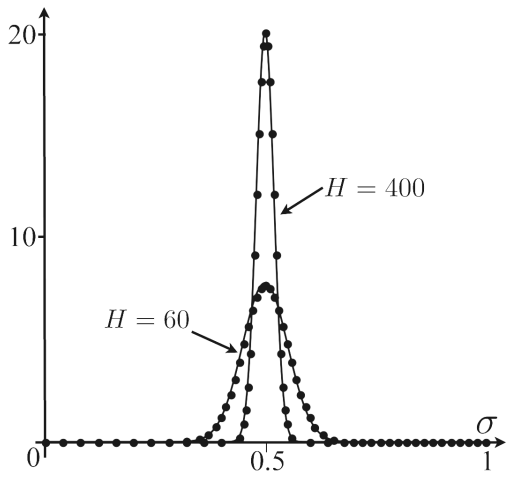

FIGURE 5.2. Graphs of $\beta_H(\sigma - \sigma_0)$ (dots) and $\delta_\nu(\sigma - \sigma_0)$ (solid lines) for $\sigma_0 = 0.5$, $\nu = (2\pi H)^{-1}$, and $H = 60$, respectively $H = 400$ as labelled.

by (103) turns into

$$S|_{\mathrm{E}}(\sigma) \,=\, \int\limits_{S^1} \sin\left[2\pi(\sigma' - \sigma)\right] \beta_H(\sigma' - \sigma_0)\, \mathrm{d}\,\sigma' \,\simeq\, \sin\left[2\pi(\sigma_0 - \sigma)\right], \tag{107}$$

and under this approximation, equation (102) holds.

To sum up, from (96) and (103) we conclude that the state of phonetic equilibrium is achieved when (93) holds and

$$\begin{gathered} n^*|_{\mathrm{E}}(\sigma, a) \,\simeq\, n^{\triangle}(a)\, \beta_H(\sigma - \sigma_0), \\ \beta_H(\sigma - \sigma_0) \,=\, C\, \mathrm{e}^{\frac{H}{2\pi}\cos[2\pi(\sigma - \sigma_0)]} \,\simeq\, H^{\frac{1}{2}}\, \mathrm{e}^{-\pi H(\sigma - \sigma_0)^2}, \end{gathered} \tag{108}$$

where C and H are given by (104) and (101), respectively. The theory gives no preference to any choice of σ_0 in the solution. The value of σ_0 is defined either by an initial condition or, if the system starts from a state out of equilibrium, by the phonetic evolution of the population.

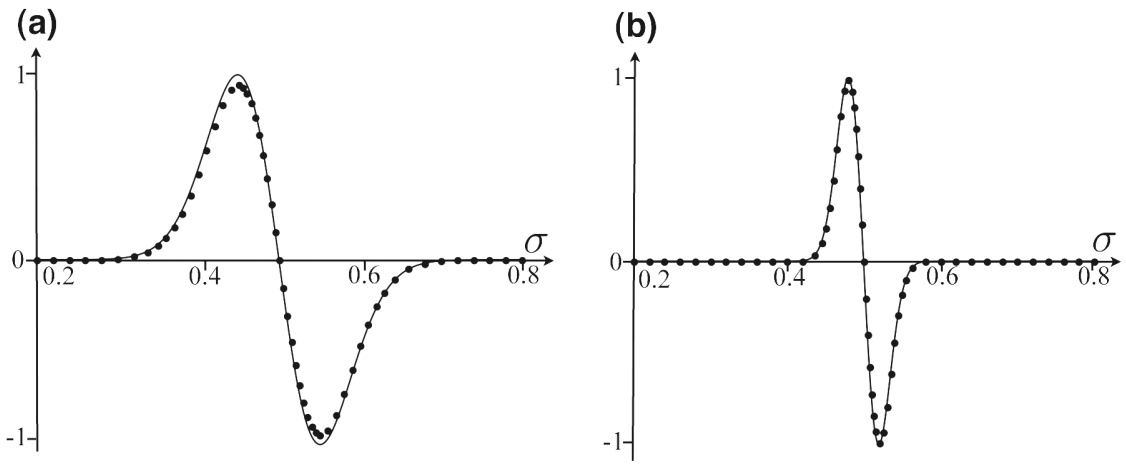

FIGURE 5.3. Graphs of the left- and right-hand sides of (102), with $n^{\blacktriangle}|_{\mathrm{E}} = N\beta_H(\sigma - \sigma_0)$ and $\sigma_0 = 0.5$, for **(a)** $H = 60$ and **(b)** $H = 400$. The vertical axes have been rescaled by the factor $f = 1/\max(\mathrm{d}\, n^{\blacktriangle}|_{\mathrm{E}}/\ \mathrm{d}\,\sigma)$ in order to accommodate all curves within the vertical range $\simeq \pm 1$. Thus, plotted are the expressions $f\,\mathrm{d}\, n^{\blacktriangle}|_{\mathrm{E}}/\ \mathrm{d}\,\sigma$ (solid lines) and $fHn^{\blacktriangle}|_{\mathrm{E}} S|_{\mathrm{E}}$ (dots), for **(a)** $f = 1/(2557N)$ and **(b)** $f = 1/(3.37 \times 10^{26} N)$. From (109) it follows that the the fractional errors between the solid and dotted curves are **(a)** $\epsilon(60) \simeq 0.0539$ and **(b)** $\epsilon(400) \simeq 0.00789$.

Seeing that the variance of (108) is given by $1/(2\pi H)$, it follows that the approximation (108) works better for large values of H, when the variance in speech is low and the phonetic attractive forces dominate over dispersive effects. In other words, the smaller the H, the broader is the phonetic equilibrium distribution $n^*|_{\mathrm{E}}$ (cf. Fig. 5.2).

As an example of how well (108) may approximate a solution to (102), we plot in Fig. 5.3 both sides of (102), for $H = 60$ and $H = 400$. The two graphs soon become nearly indistinguishable as H increases. As a measure of the error introduced by the approximation (108), we can introduce the fractional error

$$\epsilon(H) := \max\left(\frac{\mathrm{d}\, n^{\blacktriangle}|_{\mathrm{E}}}{\mathrm{d}\,\sigma} - Hn^{\blacktriangle}|_{\mathrm{E}} S|_{\mathrm{E}}\right) \Big/ \max\left(\frac{\mathrm{d}\, n^{\blacktriangle}|_{\mathrm{E}}}{\mathrm{d}\,\sigma}\right), \tag{109}$$

which is small for the examples shown in Fig. 5.3, namely $\epsilon(60) \simeq 0.0539$, and $\epsilon(400) \simeq 0.00789$.

5.7.2. Numerical test of stability. In terms of the parameter H, equation (63) with $\varphi = 0$ reads

$$\frac{\partial n^* u^*}{\partial t} + \frac{1}{L}\frac{\partial n^* u^*}{\partial a} + \frac{\partial n^* u^{*2}}{\partial \sigma} = n^* \kappa^* - \frac{1}{H}\frac{A(a)}{\tau_\kappa^2(a)}\frac{\partial n^*}{\partial \sigma}. \tag{110}$$

To run numerical simulations of the model, we must choose H, the characteristic reaction time to global stimuli $\tau_\kappa(a)$, and the intergenerational influence function $G(a, a')$ in $\kappa^*(\sigma, a, t)$ as given by (83). The choice of $G(a, a')$ also determines $A(a)$ via (100). To numerically test the stability of the equilibrium solution (103), we make explicit choices for the functions $G(a, a')$ and $\tau^2(a)$ as well as the mortality rate μ^* in (62). Assuming that a speaker's generation has maximal impact on the speaker's pronunciation, we choose a simple form for $G(a, a')$;

$$G(a, a') = \mathrm{e}^{-\gamma(a-a')^2}. \tag{111}$$

We expect the characteristic reaction time to global stimuli to be an increasing function of age, as older speakers will less likely change their utterance; that is, $\tau^{-2}(a)$ is assumed to be a decreasing function of a. In simulations, we choose functions $\tau^{-2}(a)$ of the form

$$\tau^{-2}(a) = r_1 \mathrm{e}^{-\rho a^2} + r_2. \tag{112}$$

Per capita mortality rates may be obtained from available life tables. For all simulations, we use 2007 United States Social Security population data, averaging male and female mortality rates in the actuarial life table available at [94]. From these data for μ^*, we numerically

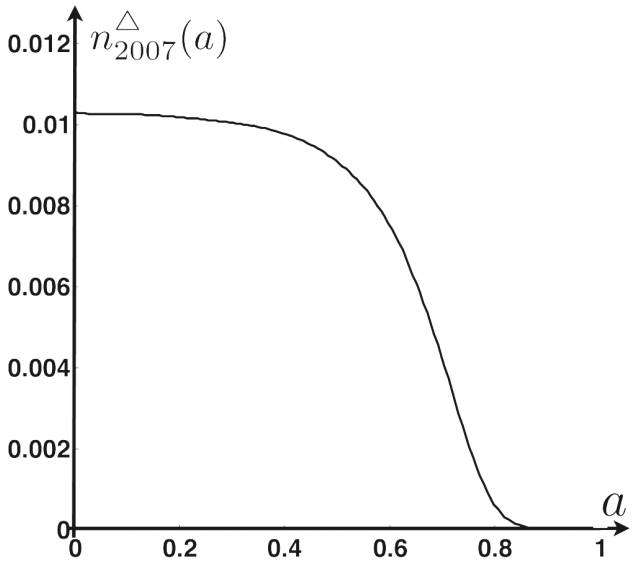

FIGURE 5.4. The stationary age structure $n^{\triangle}_{2007}(a)$.

compute the stationary age structure, which we will refer to as $n^{\triangle}_{2007}(a)$ and which is graphed in Fig. 5.4.

As a first numerical simulation of the model (with finite difference code in a 250×150 grid in (σ, a)-space and a time step of 0.25), we choose a value for H in the governing equations and start with the equilibrium solution (108) for a different value of H, viz. $H_2 \neq H$. That is,

$$n^*(\sigma, a, 0) = n^{\triangle}_{2007}(a)\beta_{H_2}(\sigma - \sigma_0), \tag{113}$$

where the mean utterance is simply chosen as $\sigma_0 = 0.5$. The initial condition as a function of σ and a, with $H_2 = 60$, is shown in Fig. 5.5a, and the state at time $t = 20$ is shown in Fig. 5.5b. Time-snapshots of cross sectional graphs of n^* for fixed a as functions of σ are shown in Fig. 5.6. Note that the evolution from the broad (in σ) curve at time $t = 0$ to the equilibrium state is slower for larger values of a, due to the fact that $\tau^{-2}(a)$ decreases with

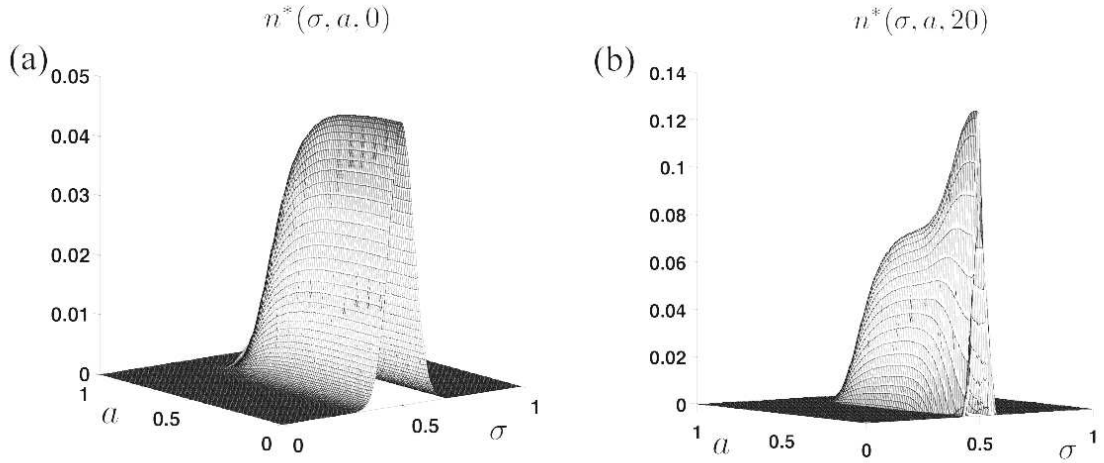

FIGURE 5.5. $n^*(\sigma, a, t = 0)$ and $n^*(\sigma, a, t = 20)$ as determined by numerical simulations of the model equations (62), (110) with parameters as given in the caption of Fig. 5.6.

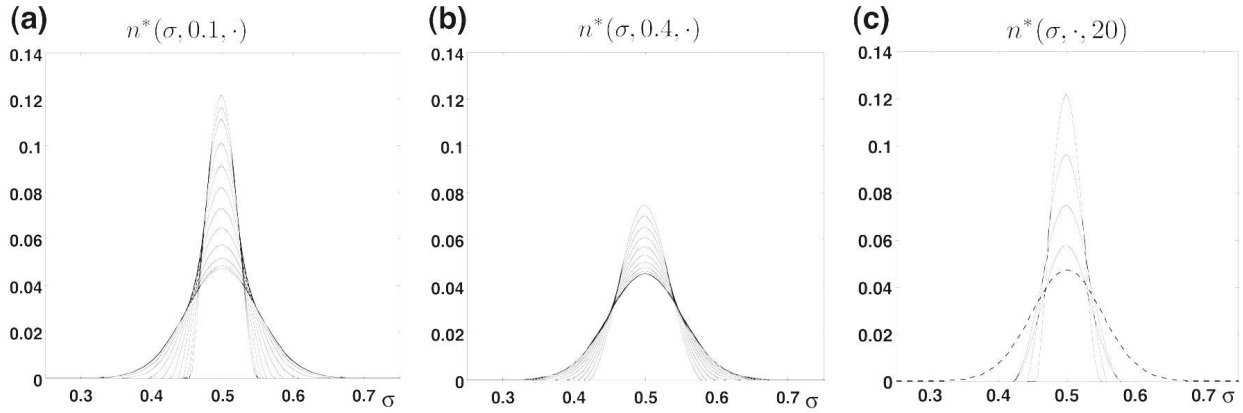

FIGURE 5.6. $n^*(\sigma, a, t)$ as a function of σ for **(a)** $a = 0.1$ and **(b)** $a = 0.4$ and values of $t = 0, 2, 4, 6 \ldots 20$ as determined by numerical simulations of the model equations (62), (110) with $H = 400$ and an initial condition with $H_2 = 60$. The peak narrows and increases in height with time. **(c)** shows $n^*(\sigma, 0.1, 0)$ (the shortest peak) and $n^*(\sigma, a, t)$ as a function of σ for $t = 20$ and $a = 0.1$, $a = 0.2$, $a = 0.4$, and $a = 0.6$ (increasing a corresponds to a taller and more narrow peak). The parameters in (112) are $r_1 = 0.8$, $r_2 = 0.2$, and $\rho = 4$.

a. $\tau^{-2}(a)$ thus provides a sort of time scale for the evolution of n^*. This evolution to the derived equilibrium solution is robust to choices of H and H_2 as well as to variations in the parameters in the choices (111) for $G(a, a')$ and (112) for $\tau^{-2}(a)$.

In these simulations, the initial condition is symmetric about a mean vowel pronunciation σ_0 and evolves to an equilibrium state with the same mean utterance σ_0. In Sect. 5.8, we simulate asymmetric perturbations of the equilibrium solution, in order to study a concrete example of vowel shift.

5.8. Asymmetric perturbations of equilibrium: Vowel shift

The initial conditions $n^*(\sigma, a, 0)$ chosen in numerical simulations of Sect. 5.7 are symmetric about the mean vowel utterance σ_0 and evolve to the equilibrium solution with the same mean utterance. In this section, we consider initial conditions that potentially change the mean utterance through vowel shift. Although there is no clear consensus on the causes of vowel shifts such as the NCVS or the Great English Vowel Shift, an aspect of common theories is that a migration of speakers into an area of a different vowel pronunciation and the subsequent desire of all speakers to accommodate their speech can give rise to vowel shift. For example, the Great English Vowel Shift may be the result of a mass migration of speakers of varying dialects into south-eastern England after the Black Death [95].

Migration of speakers into a region of previously stable mean utterance σ_0 may be modelled by an initial condition of the form

$$n^*(\sigma, a, 0) = n^{\triangle}(a)\,\frac{1}{(1+\Psi)}\left(\beta_H(\sigma-\sigma_0) + \sum_{j=1}^{p}\psi_j\beta_{H_j}(\sigma-\sigma_j)\right), \tag{114}$$

where ψ_j denotes the relative size of the j-th immigrant population (with $j = 1, \ldots, p$) and $\Psi := \sum_j \psi_j$. Each immigrant population is characterized by a a parameter H_j, a mean vowel pronunciation σ_j and an equilibrium distribution $\beta_{H_j}(\sigma - \sigma_j)$. In simulations, we take

$n^{\triangle}(a) = n^{\triangle}_{2007}(a)$ and continue to choose the functions $G(a, a')$ and $\tau^{-2}(a)$ to have the forms (111) and (112) respectively.

First consider the simplest case of a single group of immigrants: $p = 1$ in (114). Figure 5.7 shows the evolution of the vowel distribution $n^*(\sigma, \cdot, t)$ for a simulation with $H = H_1 = 400$, $\sigma_0 = 0.5$, $\sigma_1 = 0.425$, and $\psi_1 = 0.05$. Although an evolution of the speech distribution is observed for all ages, the larger values of $\tau^{-2}(a)$ for smaller a allow the younger generations to accommodate to an immigrant population more quickly (Fig. 5.8a) than the older generations (Fig. 5.8b,c). For age $a = 0.1$, the vowel distribution evolves from the initial condition with two peaks (at $\sigma = \sigma_0 = 0.5$ and $\sigma = \sigma_1 = 0.425$), to a broader distribution with only one peak, which eventually narrows into a distribution with a mean value that has shifted from the originally dominant value at $\sigma = \sigma_0$ to a value in between σ_0 and σ_1. At that time, vowel distributions at larger values of a are still centred close to the original value of $\sigma = \sigma_0$. We therefore characterize this evolution as a vowel shift.

Changing only the parameters H and H_1 to $H = H_1 = 600$, we obtain the evolution shown in Fig. 5.8. In this case, for age $a = 0.1$, the vowel distribution evolves from the initial condition with two peaks (at $\sigma = \sigma_0 = 0.5$ and $\sigma = \sigma_1 = 0.425$), to a broader distribution with only one peak, to a distribution with again two peaks, one of which eventually dominates so that distribution at time $t = 22$ has a mean value that, as in the case $H = H_1 = 400$, has shifted from the originally dominant value at $\sigma = \sigma_0$ to a value in between σ_0 and σ_1. Again, at that time, vowel distributions at larger values of a are still centred close to the original value of $\sigma = \sigma_0$.

Finally, we consider the case of two groups of immigrants: $p = 2$ in (114). Figure 5.9a shows the evolution of the vowel distribution $n^*(\sigma, 0.10, t)$ for a simulation with $H = H_1 =$

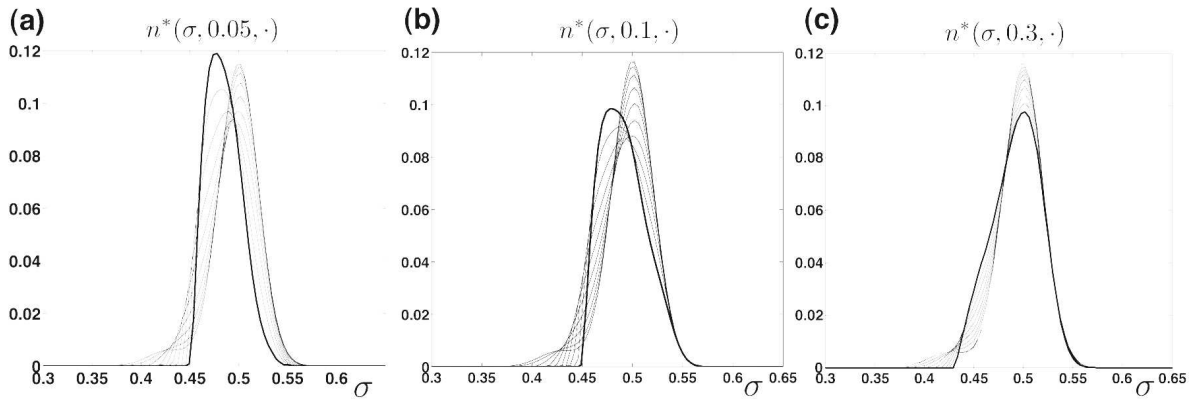

FIGURE 5.7. $n^*(\sigma, a, t)$ as a function of σ for **(a)** $a = 0.1$, **(b)** $a = 0.3$, **(c)** $a = 0.6$, and $t = 0, 2, 4, \ldots, 20$ as determined by numerical simulations of the model equations (62), (110) with $H = 400$. The initial condition (114) with $H = H_1 = 400$, $\sigma_0 = 0.5$, $\sigma_1 = 0.425$, and $\psi_1 = 0.05$ is plotted in blue. The state at time $t = 20$ is plotted as a thicker black curve. The parameters in (112) are $r_1 = 0.997$, $r_2 = 0.003$, and $\rho = 6$.

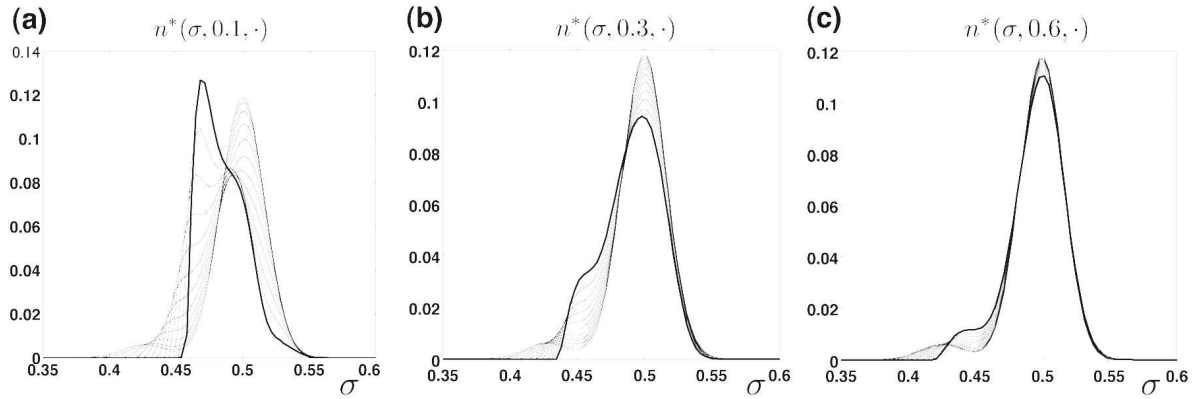

FIGURE 5.8. $n^*(\sigma, a, t)$ as a function of σ for **(a)** $a = 0.1$, **(b)** $a = 0.3$, **(c)** $a = 0.6$, and $t = 0, 2, 4, \ldots, 22$ as determined by numerical simulations of the model equations (62), (110) with $H = 600$. The initial condition (114) with $H = H_1 = 600$, $\sigma_0 = 0.5$, $\sigma_1 = 0.425$, and $\psi_1 = 0.05$ is plotted in blue. The state at time $t = 22$ is plotted as a thicker black curve. The parameters in (112) are $r_1 = 0.997$, $r_2 = 0.003$, and $\rho = 6$.

400, $\sigma_0 = 0.5$, $(\sigma_1, \psi_1) = (0.44, 0.05)$, and $(\sigma_2, \psi_2) = (0.53, 0.1)$. The second group of immigrants is larger but also has a mean vowel pronunciation σ_2 which is closer to the dominant mean vowel pronunciation σ_0: $\sigma_2 - \sigma_0 = 2(\sigma_0 - \sigma_1)$. As in the simulations of Fig. 5.7 and Fig. 5.8, the vowel distribution initially broadens. The peak first moves towards σ_2 (which corresponds to the larger value of ψ_j), but eventually moves back towards σ_1.

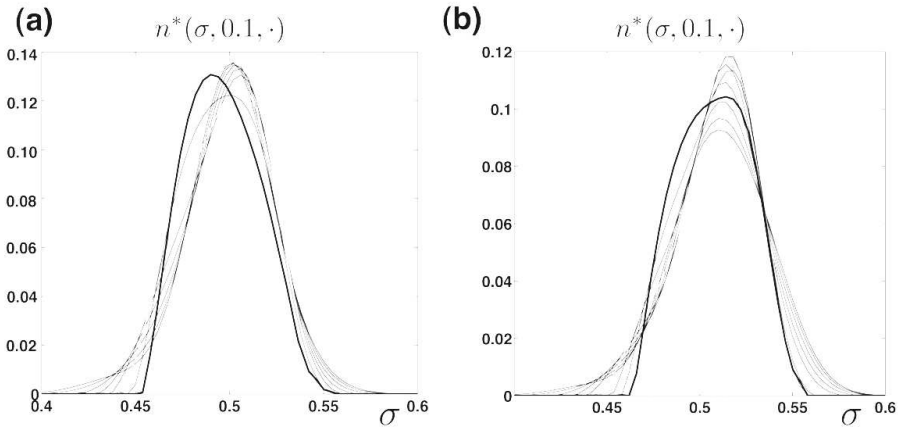

FIGURE 5.9. $n^*(\sigma, 0.10, t)$ as a function of σ for times $t = 0, 2, \ldots, 20$ as determined by numerical simulations of the model equations (62), (110) with $H = 400$. In each panel, the initial condition (114) is shown in blue. The state at time $t = 16$ is plotted in a thicker black curve. For panels, initial conditions are of the form (114) with $H = H_1 == H_2 = 400$, $\sigma_0 = 0.5$, $(\sigma_1, \psi_1) = (0.44, 0.05)$. For **(a)** $(\sigma_2, \psi_2) = (0.53, 0.1)$, and for **(b)** $(\sigma_2, \psi_2) = (0.53, 0.75)$ in (114). The parameters in (112) are $r_1 = 0.997$, $r_2 = 0.003$, and $\rho = 6$. The initial condition is graphed in blue, the state at time $t = 10$ is graphed in red, and the final state is graphed in a thicker black line.

In the long run, the mean utterance is approximately $\sigma \simeq 0.485$. Increasing ψ_2 to 0.75 while leaving other parameters constant, the long-term mean utterance is approximately equal to $\sigma_0 = 0.5$, although the peak shifts to the right before moving back; see Fig. 5.9b. Evidently a large increase in the magnitude ψ_2 is required to prevent keep the final mean vowel pronunciation from finally shifting to the left of σ_0.

5.9. Summary and perspectives

We have derived here a structured population theory for vowel learning and regular vowel change in natural languages, simplified to a one-dimensional vowel space, with age as the social structure. The proposed equations are based on the theory of mixtures with continuous diversity [88] and extend the McKendrick–von Foerster equation [1] to populations of interacting speakers. Several factors that impact the evolution of vowel pronunciation

have been taken into account, including formal and informal learning, age-dependent social trends, phonetic cohesion by affinity, and spontaneous utterance fluctuations (cf. A.1).

The resulting equations admit a stationary solution that corresponds to the typical situation of phonetic equilibrium in which all speakers have approximately the same pronunciation of a vowel. The phonetic variance of the equilibrium distribution results from a balance between the dispersive action of spontaneous, stochastic fluctuations of speech, given by the dispersive part of the phonetic stress ϕ^*, and the attractive effects of affinity, learning and social trends, modelled by the attractive part of ϕ^* and the global stimulus rate κ^*, which allows speakers of different ages to influence each others' speech.

Our equilibrium analysis suggests estimating the dimensionless number H, which is a characteristic of the population, as a step in understanding phonetic learning, variation, and change. This number measures the ratio of attractive to dispersive phonetic forces and is defined by (101). As mentioned in Sect. 5.2, the phonetic variable σ may be specified by its formants. As even an individual speaker must be expected to have stochastic variation in utterance, a measurement of σ would need to be an average over a number of a speaker's utterances. Of interest would be how variable the pronunciation of an individual speaker is relative to the variance in the population, which at phonetic equilibrium is given by $\nu = (2\pi H)^{-1}$. Formant frequencies for a vowel vary with dimensions of the vocal tract [96, 97] (see [77] for a discussion). Consequently, even if speakers have no physiological bias towards a particular pronunciation (as assumed in the present model), the stochastic variability in their speech could be distinct, therefore contributing to the dispersive factor in ϕ^*.

Simulations suggest that the parameter H also affects characteristics of vowel shift; a "double peak" occurs in the evolution for $H = 600$, but not for $H = 400$. In this chapter, we have simulated only perturbations of equilibrium states via immigrant populations. Other potential causes of vowel shift may be simulated as well. For example, a short pulse of social trend introduced from outside (e.g. by the media) can be done by imposing a temporary change in the global stimulus rate κ^*.

Our model concerns regular sound changes, which are classically characterized as being phonetically gradual but lexically abrupt. Linguists are increasingly becoming aware that this characterization is a simplification of reality, so that all sound changes are influenced by a complex interaction of lexical, social, and phonetic factors, as discussed in [98]. In light of this, the utterance parameter in our model would be best interpreted as a certain vowel pronunciation in a particular lexical context rather than in an arbitrary lexical context.

Another simplifying assumption of our model is the representation of vowel space as a loop (represented by the circle $\mathcal{S}^1$ as parameter space). Together with the hypothesis of phonetic objectivity, this leads us to treat all functions in the theory as translationally independent in σ, so that, for example, the parameter H, is independent of σ. As some vowel systems are found more often in natural languages than others, this hypothesis may fail, and our theory would need to be modified so that parameters such as H are nontrivial functions of σ. In this case, we would consider vowel space to be a line segment with boundary conditions that would depend on the position of the boundaries in vowel space.

The next step in the analysis of the simple model derived in this paper is to examine the stability of phonetic equilibrium to perturbations in the language structure. This can be done by imposing a temporary change in the global stimulus rate κ^*, representing a short pulse of

social trend introduced from outside (e.g. by the media). Another important development is to discard the hypothesis of instantaneous adaptive response (Sect. 5.4), in order to derive phonetic functionals that depend also on the transition rate u^*. Such a theory would be much more complex, allowing the study of delays caused by prospective and retrospective responses of the population, in addition to the purely adaptive behaviour considered in the current model.

Another very interesting modification of the theory would be its extension to multiple populations. Such an extension would permit studying the integration of foreign speakers with distinct pronunciations, learning and socialization skills; in other words, a phonetic counterpart to the much debated *immigrant–host problem* [99].

Further, we may address the strong simplifying assumption that the chain shift is rigid, stated in Sect. 5.1, which prevents the equations that we have derived from modeling vowel–vowel interactions ("knock-on effects") by drag- and push-chain mechanisms [87]. Real vowel shifts, such as the Northern Cities Vowel Shift, generally begin between one pair of vowels and only later propagate to other vowels in a chain [69]. Our model would need to be extended from one vowel pronunciation $\sigma \in \mathcal{S}^1$ to include a vowel system $(\sigma_1, \sigma_2, \ldots, \sigma_n) \in \mathcal{S}^1 \times \cdots \times \mathcal{S}^1$ as the domain for the functions n^* and u^*. This would not be a straightforward extension but would allow for comparison of unimodal and bimodal distributions as in the study of Maye *et al.* [84]; e.g. can a bimodal distribution in one σ_j be stable, or would it split into unimodal distributions in two utterances σ_i, σ_j?

The approach developed here for vowel systems may be applied also to other learning situations involving a continuous variable that describes an aspect of a learned behaviour. In language, this includes other aspects of phonetics (such as voicing in consonants [84,

100]) as well as semantics (such as the learning and interpretation of color words, as color is also described by continuous variables [101]). Comparable equations may also model other time-dependent processes of cognition in self-interacting populations, like opinions or perceptions. In particular, we are currently applying these ideas to model evolving scenarios of public perception of climate change. Finally, in ecology, non-human behavioural problems may also be tackled with a similar approach, seeing that continuously varying traits (e.g. vocalization, aggressiveness, etc.) are not exclusive to human communities. In this manner we are currently studying the problem of the interaction of invasive exotics with native species [102, 103].

Bibliography

[1] B. Charlesworth, *Evolution in Age-Structured Populations*. Cambridge: Cambridge UP, 2 ed., 1994.

[2] J. E. Bromberg, S. Kumar, C. S. Brown, and T. J. Stohlgren, "Distributional changes and range predictions of downy brome (bromus tectorum) in rocky mountain national park," *Invas. Plant Sci. Manag.*, vol. 4, no. 2, pp. 173–182, 2011.

[3] P. H. Evangelista, S. Kumar, T. J. Stohlgren, C. S. Jarnevich, A. W. Crall, J. B. Norman III, and D. T. Barnett, "Modelling invasion for a habitat generalist and a specialist plant species," *Diversit Distrib.*, vol. 14, pp. 808–817, 2008.

[4] S. Kumar and T. J. Stohlgren, "Maxent modeling for predicting suitable habitat for threatened and endangered tree *canacomyrica monticola* in new caledonia," *J. Ecol. Nat. Environ.*, vol. 1, no. 4, pp. 94–98, 2009.

[5] A. Monty, C. S. Brown, and D. B. Johnston, "Fire promotes downy brome (bromus tectorum l.) seed dispersal," *Biol. Invasions*, vol. 15, pp. 1113–1123, 2013.

[6] S. J. Phillips, M. Dudik, and R. E. Schapire, "A maximum entropy approach to species distribution modeling," *Proceedings of the Twenty-First International Conference on Machine Learning*, pp. 655–662, 2004.

[7] C. Strickland, G. Dangelmayr, and P. Shipman, "Modeling the presence probability of invasive plant species with nonlocal dispersal," *J. Math. Biol.*, 2013. in press.

[8] P. D. Shipman, S. H. Faria, and C. Strickland, "Towards a continuous population model for natural language vowel shift," *J. Theor. Biol.*, vol. 332, pp. 123–135, 2013.

[9] S. I. Higgins, S. Scheiter, and M. Sankaran, "The stability of African savannas: insights from the indirect estimation of the parameters of a dynamic model," *Ecology*, vol. 91,

no. 6, pp. 1682–1692, 2010.

[10] D. Donzelli, C. De Michele, and R. J. Scholes, "Competition between trees and grasses for both soil water and mineral nitrogen in dry savannas," *J. Theor. Biol.*, vol. 332, pp. 181–190, 2013.

[11] A. C. Liedloff and G. D. Cook, "Modelling the effects of rainfall variability and fire on tree populations in an Australian tropical savanna with the FLAMES simulation model," *Ecol. Model.*, vol. 201, pp. 269–282, 2007.

[12] G. Bucini and N. P. Hanan, "A continental-scale analysis of tree cover in African savannas," *Global Ecol. Biogeogr.*, vol. 16, pp. 593–605, 2007.

[13] S. Scheiter and S. I. Higgins, "Impacts of climate change on the vegetation of Africa: an adaptive dynamic vegetation modelling approach," *Glob. Change Biol.*, 2009.

[14] W. Cramer, A. Bondeau, F. I. Woodward, I. C. Prentice, R. A. Betts, V. Brovkin, P. M. Cox, V. Fisher, J. A. Foley, A. D. Friend, C. Kucharik, M. R. Lomas, N. Ramankutty, S. Sitch, B. Smith, A. White, and C. Young-Molling, "Global response of terrestrial ecosystem structure and function to CO_2 and climate change: results from six dynamic global vegetation models," *Glob. Change Biol.*, vol. 7, pp. 357–373, 2001.

[15] W. J. Bond, F. I. Woodward, and G. F. Midgley, "The global distribution of ecosystems in a world without fire," *New Phytol.*, vol. 165, pp. 525–538, 2005.

[16] R. J. Scholes and S. R. Archer, "Tree-grass interactions in savannas," *Annu. Rev. Ecol. Syst.*, vol. 28, pp. 517–44, 1997.

[17] M. Sankaran, N. P. Hanan, R. J. Scholes, J. Ratnam, D. J. Augustine, B. S. Cade, J. Gignoux, S. I. Higgins, X. L. Roux, F. Ludwig, J. Ardo, B. Feetham, A. Bronn, G. Bucini, K. K. Caylor, M. B. Coughenour, A. Diouf, W. Ekaya, C. J. Feral, E. C.

February, P. G. H. Frost, P. Hiernaux, H. Hrabar, K. L. Metzger, H. H. T. Prins, S. Ringrose, W. Sea, J. Tews, J. Worden, and N. Zambatis, "Determinants of woody cover in African savannas," *Nat.*, vol. 438, pp. 846–849, 2005.

[18] M. Sankaran, J. Ratnam, and N. Hanan, "Woody cover in African savannas: the role of resources, fire and herbivory," *Global Ecol. Biogeogr.*, vol. 17, pp. 236–245, 2008.

[19] S. I. Higgins, W. J. Bond, and W. S. W. Trollope, "Fire, resprouting and variability: a recipe for grass-tree coexistence in savanna," *J. Ecol.*, vol. 88, pp. 213–229, 2000.

[20] B. H. Walker and I. Noy-Meir, "Aspects of the stability and resilience of savanna ecosystems," *Ecol. Stu. An.*, vol. 42, pp. 557–590, 1982.

[21] H. Walter, *Ecology of Tropical and Subtropical Vegetation.* Edinburgh, UK: Oliver and Boyd, 1971.

[22] A. C. Liedloff and G. D. Cook, "Dynamics of fire and carbon in Australian tropical savannas: The Flames model," in *Ecosystem Function in Savannas: Measurement and Modeling at Landscape to Global Scales* (M. J. Hill and N. P. Hanan, eds.), ch. 15, Boca Raton: CRC Press, 2011.

[23] J. Russell-Smith, P. J. Whitehead, G. D. Cook, and J. L. Hoare, "Response of Eucalyptus-dominated savanna to frequent fires: lessons from Munmarlary, 1973-1996," *Ecol. Monogr.*, vol. 73, no. 3, pp. 349–375, 2003.

[24] A. N. Andersen, G. D. Cook, and R. J. Williams, *Fire in Tropical Savannas: The Kapalga Experiment.* New York: Springer-Verlag, 2003.

[25] S. I. Higgins, W. J. Bond, E. C. February, A. Bronn, D. I. W. Euston-Brown, B. Enslin, N. Govener, L. Rademan, S. O'Regan, A. L. F. Potgieter, S. Scheiter, R. Sowry, L. Trollope, and W. S. W. Trollope, "Effects of four decades of fire manipulation on

woody vegetation structure in savanna," *Ecology*, vol. 88, no. 5, pp. 1119–1125, 2007.

[26] A. N. Andersen, G. D. Cook, L. K. Corbett, M. M. Douglas, R. W. Eager, J. Russell-Smith, S. A. Setterfield, R. J. Williams, and J. C. Z. Woinarski, "Fire frequency and biodiversity conservation in Australian tropical savannas: implications from the kapalga fire experiment," *Austral Ecol.*, vol. 30, pp. 155–167, 2005.

[27] W. J. Bond, "What limits trees in c_4 grasslands and savannas?," *Annu. Rev. Ecol. Evol. Syst.*, vol. 39, pp. 641–659, 2008.

[28] N. P. Hanan, W. B. Sea, G. Dangelmayr, and N. Govender, "Do fires in savannas consume woody biomass? a comment on approaches to modeling savanna dynamics," *Am. Nat.*, vol. 171, pp. 851–856, 2008.

[29] A. C. Staver, S. Archibald, and S. Levin, "Tree cover in sub-Saharan Africa: Rainfall and fire constrain forest and savanna as alternative stable states," *Ecology*, vol. 92, no. 5, pp. 1063–1072, 2011.

[30] S. Geng, F. W. T. Penning de Vries, and I. Supit, "A simple method for generating daily rainfall data," *Agr. Forest Meteorol.*, vol. 36, pp. 363–376, 1986.

[31] R. Srikanthan and M. T. A, "Stochastic generation of annual, monthly and daily climate data: a review," *Hydrol. Earth Syst. Sc.*, vol. 5, no. 4, pp. 653–670, 2001.

[32] P. D'Odorico, F. Laio, and L. Ridolfi, "A probabilistic analysis of fire-induce tree-grass coexistence in savannas," *Am. Nat.*, vol. 167, no. 3, pp. 79–87, 2006.

[33] J. Gignoux, G. Lahoreau, R. Julliard, and S. Barot, "Establishment and early persistence of tree seedlings in an annually burned savanna," *J. Ecol.*, vol. 97, pp. 484–495, 2009.

[34] R. P. Keller, D. M. Lodge, M. A. Lewis, and J. F. Shogren, eds., *Bioeconomics of Invasive Species: Integrating Ecology, Economics, Policy, and Management.* Oxford UP, 2009.

[35] J. M. DiTomaso, "Invasive weeds in rangelands: Species, impacts, and management," *Weed Sci.*, vol. 48, pp. 255–265, 2000.

[36] S. T. Knick and J. T. Rotenberry, "Landscape characteristics of disturbed shrubsteepe habitats in southwestern idaho (usa)," *Landsc. Ecol.*, vol. 12, no. 5, pp. 287–297, 1997.

[37] M. G. A. van der Heijden, J. N. Klironomos, M. Ursic, P. Moutoglis, R. Steitwolf-Engel, T. Boller, A. Wiemken, and I. R. Sanders, "Mycorrhizal fungal diversity determines plant biodiversity, ecosystem variability and productivity," *Nat.*, vol. 396, pp. 69–72, 1998.

[38] S. Kumar, S. A. Spaulding, T. J. Stohlgren, K. Hermann, T. Schmidt, and L. Bahls, "Potential habitat distribution for the freshwater diatom *didymosphenia geminata* in the continental us," *Front. Ecol. Environ.*, vol. 7, no. 8, pp. 415–420, 2009.

[39] R. C. Venette, D. J. Kriticos, R. D. Magarey, F. H. Koch, R. H. A. Baker, S. P. Worner, N. N. Gòmez Raboteaux, D. W. McKenny, E. J. Dobesberger, D. Yemshanov, P. J. De Barro, W. D. Hutchinson, G. Fowler, T. M. Kalaris, and J. Pedlar, "Pest risk maps for invasive alien species: a roadmap for improvement," *Biosci.*, vol. 60, no. 5, pp. 349–362, 2010.

[40] J. G. Skellam, "Random disperal in theoretical populations," *Biom.*, vol. 38, no. 1, pp. 196–218, 1951.

[41] Y. E. Maruvka and N. M. Shnerb, "Nonlocal competition and logistic growth: patterns, defects, and fronts," *Phys. Rev. E*, vol. 73, pp. 1–12, 2006.

[42] A. Okubo and S. A. Levin, *Diffusion and ecological problems.* New York, NY: Springer-Verlag, 2001.

[43] J. Furter and M. Grinfeld, "Local vs. non-local interactions in population dynamics," *J. Math. Biol.*, vol. 27, pp. 65–80, 1989.

[44] M. Kot, M. A. Lewis, and P. van den Driessche, "Disperal data and the spread of invading organisms," *Ecology*, vol. 77, no. 7, pp. 2027–2042, 1996.

[45] N. Shigesada, K. Kawasaki, and Y. Takeda, "Modeling stratified diffusion in biological invasions," *Am. Nat.*, vol. 146, no. 2, pp. 229–251, 1995.

[46] D. Mollison, "Spatial contact models for ecological and epidemic spread," *J. R. Statist. Soc.*, vol. 39, no. 3, pp. 283–326, 1977.

[47] J. Medlock and M. Kot, "Spreading disease: integro-differential equations old and new," *Math. Biosci.*, vol. 184, pp. 201–222, 2003.

[48] D. Mollison, "Possible velocities for a simple epidemic," *Adv. Appl. Probab.*, vol. 4, pp. 233–257, 1972.

[49] S. Ruan, "Spatial-temporal dynamics in nonlocal epidemiological models," *Math. Life Sci. and Med.*, pp. 97–122, 2007.

[50] F. Lutscher, "Nonlocal dispersal and averaging in heterogeneous landscapes," *Appl. Anal.*, vol. 89, no. 7, pp. 1091–1108, 2010.

[51] D. Mollison and H. Daniels, "The 'deterministic simple epidemic' unmasked," *Math. Biosci.*, vol. 117, pp. 147–153, 1992.

[52] J. E. Besag, "Spatial interaction and the statistical analysis of lattice systems (with discussion)," *J. R. Statist. Soc. B.*, vol. 36, pp. 192–236, 1974.

[53] D. Y. Downham and R. K. B. Morgan, "Growth of abnormal cells," *Nat.*, vol. 242, pp. 528–530, 1973.

[54] M. A. Eden, "A two-dimensional growth process," *Proc. 2th Berkeley Symp. Math. Statist. Prob.*, vol. 4, pp. 223–239, 1961.

[55] J. Honerkamp, *Stochastic Dynamical Systems*. New York, Weinheim: VCH Publ., 1993.

[56] D. Mollison, "The rate of spatial propagation of simple epidemics," *Proc. Sixth Berkeley Symp. Math. Statist. and Prob.*, vol. 3, 1972.

[57] D. Mollison, "Dependence of epidemic and population velocities on basic parameters," *Math. Biosci.*, vol. 107, no. 2, p. 255, 1991.

[58] R. Engler and A. Guisan, "Migclim: predicting plant distribution and dispersal in a changing climate," *Divers. Distrib.*, vol. 15, pp. 590–601, 2009.

[59] G. F. Midgley, I. D. Davies, C. H. Albert, R. Altwegg, L. Hannah, G. O. Hughes, L. R. O'Halloran, C. Seo, J. H. Thorne, and W. Thuiller, "Biomove - an integrated platform simulating the dynamic response of species to environmental change," *Ecography*, vol. 33, pp. 612–616, 2010.

[60] M. Kot and W. M. Schaffer, "Discrete-time growth-dispersal models," *Math. Biosci.*, vol. 80, pp. 109–136, 1986.

[61] A. Hastings, K. Cuddington, K. F. Davies, C. J. Dugaw, S. Elmendorf, A. Freestone, S. Harrison, M. Holland, J. Lambrinos, U. Malvadkar, B. A. Melbourne, K. Moore, C. Taylor, and D. Thomson, "The spatial spread of invasions: new developments in theory and evidence," *Ecology*, vol. 8, no. 1, pp. 91–101, 2004.

[62] K. Kawasaki, F. Takasu, H. Caswell, and N. Shigesada, "How does stochasticity in colonization accelerate the speed of invasion in a cellular automaton model?," *Ecol.*

Res., vol. 21, pp. 334–345, 2006.

[63] J. A. Catford, R. Jansson, and C. Nilsson, "Reducing redundancy in invasion ecology by integrating hypotheses into a single theoretical framework," *Divers. Distrib.*, vol. 15, pp. 22–40, 2009.

[64] T. Eltoft, T. Kim, and T. Lee, "On the multivariate laplace distribution," *IEEE Signal Proc. Let.*, vol. 13, no. 5, 2006.

[65] M. Benninger-Truax, J. L. Vankat, and R. L. Schaefer, "Trail corridors as habitat and conduits for movement of plant species in rocky mountain national park, colorado, usa," *Landsc. Ecol.*, vol. 6, no. 4, pp. 269–278, 1992.

[66] "Rocky mountain national park," November 2004. Coord: 448000 m E and 4470000 m N. GOOGLE EARTH. Accessed: 6-22-2013.

[67] W. Shakespeare, H. S. Weil, and J. Weil, *The first part of King Henry IV, second ed.* Cambridge: Cambridge UP, 2007.

[68] M. J. Gordon, "The sounds, they are a shiftin'." PBS Essay, July 2012.

[69] W. Labov, M. Yaeger, and R. Steiner, *The Quantitative Study of Sound Change in Progress*. Philadelphia: U. S. Regional Survey, 1973.

[70] W. Labov, *Principles of Linguistic Change: Volume 1: Internal Factors*. Cambridge: Blackwell, 1994.

[71] P. Eckert, *Jocks & Burnouts: Social categories and identity in the high school*. New York: Teachers College Press, 1989.

[72] L. Liljencrants and B. Lindblom, "Numerical simulations of vowel quality systems: The role of perceptual contrast," *Language*, vol. 48, pp. 839–862, 1972.

[73] J. L. Schwartz, L.-J. Boë, N. Vallé, and C. Abry, "Major trends in vowel system inventories," *J. Phonetics*, vol. 25, pp. 233–253, 1997.

[74] J. L. Schwartz, L.-J. Boë, N. Vallé, and C. Abry, "The dispersion-focalization theory of vowel system inventories," *J. Phonetics*, vol. 25, pp. 255–286, 1997.

[75] B. de Boer, "Emergence of vowel systems through self-organization," *AI Commun*, vol. 13, pp. 27–39, 2000.

[76] P. K. Kuhl, "Discrimination of speech by nonhuman animals: Basic sensitivities conducive to the perception of speech-sound categories," *J. Acoust. Soc. Am.*, vol. 70, pp. 340–349, 1983.

[77] P. K. Kuhl, "Perception of auditory equivalence classes for speech in early infancy," *Infant Behavior and Development*, vol. 6, pp. 263–285, 1983.

[78] P. K. Kuhl, "Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques," *J. Acoust. Soc. Am.*, vol. 73, pp. 1003–1010, 1984.

[79] K. R. Kluender, A. J. Lotto, L. L. Holt, and S. L. Bloedel, "Role of experience for language-specific functional mappings of vowel sounds," *J. Acoust. Soc. Am.*, vol. 104, pp. 3568–3582, 1998.

[80] M. Ramon-Casas, D. Swingley, N. Sebastián-Gall'es, and L. Bosch, "Vowel categorization during word recognition in bilingual toddlers," *Cognitive Psychology*, vol. 59, pp. 96–121, 2009.

[81] D. Swingley and R. N. Aslin, "Spoken word recognition and lexical representation in very young children," *Cognition*, vol. 76, pp. 147–166, 2000.

[82] F. H. Guenther and M. N. Gjaja, "The perceptual magnet effect as an emergent property of neural map formation," *J. Acoust. Soc. Am.*, vol. 100, pp. 1111–1121, 1996.

[83] P. K. Kuhl, "Early linguistic experience and phonetic perception: implications for theories of developmental speech perception," *Journal of Phonetics*, vol. 21, pp. 125–139, 1993.

[84] J. Maye, J. F. Werker, and L. Gerken, "Infant sensitivity to distributional information can affect phonetic discrimination," *Cognition*, vol. 82, pp. B101–B111, 2002.

[85] R. N. Aslin, P. W. Jusczyk, and D. B. Pisoni, "Speech and auditory processing during infancy: constraints on a precursors to language," in *Handbook of Child Psychology: Cognition, Perception, and Language* (D. Kuhn and R. Siegler, eds.), pp. 147–254, New York: Wiley, 1998.

[86] M. Yaeger-Dror, "Phonetic evidence for sound change in quebec french," in *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III* (P. A. Keating, ed.), Cambridge: Cambridge UP, 1994.

[87] J. Aitchison, *Language Change: Progress or Decay?* 3 ed., 2001.

[88] S. H. Faria, "Mixtures with continuous diversity: general theory and application to polymer solutions," *Continuum Mech. Thermodyn.*, vol. 13, pp. 91–120, 2001.

[89] P. C. Delattre, A. M. Liberman, and F. S. Cooper, "Voyelles synthetiques a deux formantes et voyelles cardinales," *Maitre Phonetique*, vol. 96, pp. 30–36, 1951.

[90] A. Radford, M. Atkinson, D. Britain, H. Clahsen, and A. Spencer, *Linguistics: An Introduction*. Cambridge: Cambridge UP, 1999.

[91] R. S. Chapman, "Children's language learning: An interactionist perspective," *J. Child Pshychol. Psychiat.*, vol. 41, pp. 33–54, 2000.

[92] R. P. Cooper, J. Abraham, S. Berman, and M. Staska, "The development of infants' preference for motherese," *Infant Behavior and Development*, vol. 20, pp. 477–488,

1997.

[93] W. P. Fifer and C. Moon, "The effects of fetal experience with sound," in *Fetal Behavior: A Phychobiological Perspective* (J. P. Lecanuet, N. A. Krasnegor, W. P. Fifer, and W. P. Smotherman, eds.), pp. 351–366, Hillsdale, NJ: Erlbuam, 1995.

[94] T. O. W. of the United State Social Security Administration, "Actuarial life table," Feb. 2013.

[95] P. Wolfe, *Linguistic Change and the Great Vowel Shift*. Los Angeles: University of California Press, 1972.

[96] G. Fant, *Speech sounds and features*. Cambridge: MIT Press, 1973.

[97] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Amer.*, vol. 68, no. 33(A), 1952.

[98] J. Blevins and A. Wedel, "Inhibited sound change," *Diachronica*, vol. 26, pp. 143–183, 2009.

[99] R. Y. Bourhis, L. C. Moïse, S. Perreault, and S. Senécal, "Towards an interactive acculturation model: A social psychological approach," *Int. J. Pshychol.*, vol. 32, pp. 369–386, 1997.

[100] J. E. Pegg and J. F. Werker, "Adult and infant perception of two english phonemes," *J. Acoust. Soc. Amer.*, vol. 102, pp. 3742–3753, 1997.

[101] N. L. Komarova, K. A. Jameson, and L. Narens, "Evolutionary models of color categorization based on descrimination," *J. Math. Psychology.*, vol. 51, pp. 359–382, 2009.

[102] R. P. Keller, D. M. Lodge, M. A. Lewis, and J. F. Shogren, eds., *Bioeconomics of Invasive Species: Integrating Ecology, Economics, Policy, and Management*. Oxford: Oxford UP, 2009.

[103] M. Kot, M. A. Lewis, and P. van den Driessche, "Dispersal data and the spread of invading organisms," *Ecology*, vol. 77, pp. 2027–2042, 1996.

[104] S. H. Faria, "Creep and recrystallization of large polycrystalline masses. Part I: general continuum theory," *Proc. Roy. Soc. London A*, vol. 462, pp. 1493–1514, 2006.

[105] S. H. Faria and K. Hutter, "A systematic approach to the thermodynamics of single and mixed flowing media with microstructure. Part I: balance equations and jump conditions," *Continuum Mech. Thermodyn.*, vol. 14, pp. 459–481, 2002.

APPENDIX A

Vowel Shift Supplementary Material

A.1. Derivation of the balance equations

The starting point for the derivation of (62) and (63) is the McKendrick–von Foerster equation [1] for a homogeneous, non-stationary population with maximum lifespan L

$$\frac{\partial n^{\triangle}}{\partial t} = -\frac{1}{L}\frac{\partial n^{\triangle}}{\partial a} - n^{\triangle}\mu^{\triangle}, \tag{115}$$

where $n^{\triangle}(a,t)$ and $\mu^{\triangle}(a,t)$ denote respectively the number density and the mortality rate per capita of individuals aged $a \in \mathcal{A}$ at time $t \in \mathbb{R}$ (cf. definitions in Sect. 5.2). Equation (115) simply tells us that temporal changes in the structure of the population are caused by *ageing* and *death*. Clearly, death is interpreted as a loss of individuals and is described by the mortality rate $\mu^{\triangle}$. On the other hand, the net effect of ageing is modelled by the age derivative of $n^{\triangle}/L$, the convective flux of speakers along the a-axis. Indeed, we can think of L^{-1} as a kind of "velocity in the age space", more precisely the time derivative of the *dimensionless ageing function* h, defined as a continuous sequence of dimensionless ages in time. Let $\mathcal{P}$ denote the set of all individuals in the population. Thus, we may regard the dimensionless ageing function h of an individual $p \in \mathcal{P}$ as a map

$$h : \mathcal{P} \times \mathbb{R} \to \mathcal{A}, \qquad a = h(p,t) := \frac{t - t_{\mathrm{B}}}{L}, \tag{116}$$

where $t_{\mathrm{B}}(p)$ is the individual's birth time and L is, again, the maximum lifespan of any member of the population.

The generalization of (115) to a structured population of speakers is now straightforward. We think of the population as a structured community with a continuous diversity of speakers. A group of speakers aged a and uttering the sound $\sigma \in \mathcal{S}^1$ defines a *species* (cf. definitions in Sect. 5.2). Thus, the density number of speakers belonging to a particular species is given by $n^*(\sigma, a, t)$, and the main question is which changes occur in (115) when we replace $n^\triangle$ by n^*.

Succinctly, besides the effects of ageing and death we must now account for the effects of pronunciation changes. Such changes can generally be of two types:

(I) *Gradual changes* of pronunciation through the actions of learning, social trends, phonetic cohesion and utterance fluctuations. These changes can be described by an effective "velocity" in the phonetic space, called *phonetic transition rate* $u^*(\sigma, a, t)$. The product n^*u^* is named *phonetic transition impetus* and corresponds to a net convective flux of speakers along the σ-axis.

(II) *Abrupt changes* of pronunciation through lexical diffusion, pathological causes, deliberate sound changes, etc. Such phenomena are characterized by discontinuous transfers of speakers in the phonetic space and are modelled by the *rate of abrupt sound change per capita* Γ^*.

Thus, after replacing $n^\triangle$ by n^* in (115) and gathering together all the notions mentioned above, we obtain

$$\frac{\partial n^*}{\partial t} = -\frac{1}{L}\frac{\partial n^*}{\partial a} - n^*\mu^* + n^*\Gamma^* - \frac{\partial n^* u^*}{\partial \sigma}, \tag{117}$$

where $\mu^{\triangle}(a,t)$ has been generalized to $\mu^*(\sigma,a,t)$, for completeness. Evidently, (117) is identical to (62), with the last term interpreted as a "divergence in the phonetic space" of the transition impetus n^*u^*.

The good news about (117) is that it is the appropriate generalization of the McKendrick–von Foerster equation (62) for a population with age *and* phonetic structures. The bad news is that it is no longer possible to compute the structure of the population just by prescribing the function $\mu^{\triangle}$, as done with (62), since there are two new unknowns in (117), namely u^* and Γ^*. The latter can be eliminated by avoiding abrupt sound changes, which are not pertinent for the modelling of regular vowel chain shifts, so that for our purposes (64) holds. Concerning the transition rate u^*, the situation is more delicate. The variable u^* is essential for the modelling of vowel chain shift and therefore cannot be neglected. Instead, we may study how learning, social trends, phonetic cohesion and utterance fluctuations influence the impetus of speakers to change their pronunciations. In other words, we must construct a balance equation for the transition impetus n^*u^*.

If none of the gradual change effects listed in Item I were active, then the time change of impetus at a given point $(\sigma,a)\in\mathcal{S}^1\times\mathcal{A}$ would be simply caused by the "migration" of speakers along the age–phonetic space, i.e. there would be a convective transport of impetus according to the equation

$$\frac{\partial n^*u^*}{\partial t}=-\frac{1}{L}\frac{\partial n^*u^*}{\partial a}-\frac{\partial n^*u^{*2}}{\partial \sigma}, \tag{118}$$

such that L^{-1} and u^* represent the "velocities" at which speakers carry the impetus n^*u^* along the a- and σ-axes, respectively.

However, that several mechanisms of gradual phonetic change may affect the impetus of a speaker. We can model them as follows:

: Phonetic cohesion and utterance fluctuations. A fundamental feature of most social systems is the phenomenon of affinity: *familiar species* (viz. speakers of similar age and utterance) tend to identify with each other, generating a feedback that leads to cluster formation. Following [88, 104, 105] we may suppose that such short-range interactions between familiar species can be modelled by conductive fluxes of impetus in $\mathcal{S}^1 \times \mathcal{A}$, called *phonetic stresses* ϕ^* and φ^*. On the other hand, as in other complex systems susceptible to cluster formation, the dispersive effect of random fluctuations is unavoidable. Such utterance fluctuations can be modelled through a stochastic, conductive–diffusive flux density of impetus in the phonetic space $\mathcal{S}^1$, called *stochastic flux of impetus*, which is incorporated into the phonetic stress ϕ^*. Consequently, the phonetic stresses ϕ^* and φ^* describe the net effect of the struggle between cohesive and dispersive phonetic agents.

: Learning and social trends. The most predominant mechanism of phonetic change during childhood and early adolescence is learning. It is mainly characterized by the influence of parents and educators on the utterance of the child. As such, we expect learning to be an interaction between speakers with disparate ages and pronunciations, i.e. between *dissimilar species*. Similar nurtural effects, predominant during adolescence and early adulthood, are social trends of pronunciation. New trends determining the general utterance preferences of the whole population are created by the formation of clusters in the phonetic space $\mathcal{S}^1$, which act as phonetic attractors for speakers attempting to be understood by "everyone".

The effects of learning and social trends are introduced in (118) in the form of a long-range production of impetus called *global stimulus rate per capita* κ^*.

From the reasoning above, we can generalize (118) to the case of a structured population of speakers experiencing gradual changes of pronunciation through the actions of learning, social trends, phonetic cohesion and utterance fluctuations

$$\frac{\partial n^* u^*}{\partial t} = -\frac{1}{L}\frac{\partial n^* u^*}{\partial a} - \frac{\partial n^* u^{*2}}{\partial \sigma} - \frac{\partial \varphi^*}{\partial a} - \frac{\partial \phi^*}{\partial \sigma} + n^* \kappa^*. \tag{119}$$

A.2. Equations for a linear first-order theory

In Sect. 5.5 we derived the explicit equations for a zeroth-order theory. Here we outline the procedure to derive the explicit forms of F_γ ($\gamma = 1, 2, 3, 4$) in a higher-order theory that is linear in first-order derivatives.

First, we must realize that the arguments used to derive (75) and (83) are still valid, and so are the equations

$$F_1 := \mu^*(\sigma, a, t) \equiv \mu^{\triangle}(a), \tag{120}$$

$$F_4 := \kappa^*(\sigma, a, t) = \frac{1}{[\tau_\kappa(a)]^2} \int_{\mathcal{S}^1} \int_{\mathcal{A}} \sin\left[2\pi(\sigma' - \sigma)\right] G(a, a') n^*(\sigma', a', t) \, \mathrm{d}a' \, \mathrm{d}\sigma'. \tag{121}$$

Now, to derive the appropriate forms of F_2 and F_3 in a linear first order theory, we must first express the influence kernels V_2, W_2, V_3 and W_3 of (73) not only as nascent delta functions $\delta_\epsilon(\cdot)$, but also their first derivatives. During this process, we should respect the

symmetry properties of the respective quantities and the fact that products of derivatives are to be neglected. As already mentioned in Sect. 5.5, from (66)–(69) it follows that φ^* should be an odd function of σ, while ϕ^* must be an even function of σ. Thus, after a short inspection of (73) we arrive at

$$V_2\big(\sigma-\sigma'\big) \equiv \frac{\partial}{\partial\sigma'}\delta_\epsilon(\sigma-\sigma'), \qquad W_2\big(a,a'\big) \equiv \frac{\delta_\epsilon(a-a')}{\big[\tau_\varphi(a)\big]^2}, \tag{122}$$

$$V_3\big(\sigma-\sigma'\big) \equiv \delta_\epsilon(\sigma-\sigma'), \quad W_3\big(a,a'\big) \equiv \frac{\delta_\epsilon(a-a')}{\big[\tau_\phi(a)\big]^2} + \frac{\partial}{\partial a'}\,\frac{\delta_\epsilon(a-a')}{\big[\tau_{\phi_a}(a)\big]^2}, \tag{123}$$

where $\tau_\varphi(a)$, $\tau_\phi(a)$ and $\tau_{\phi_a}(a)$ are the characteristic reaction times of the speakers aged a.

Finally, choosing as before

$$\mathfrak{X}_\nu(n^*) \equiv n^*, \qquad \nu = 2,3, \tag{124}$$

we obtain the explicit equations for φ^* and ϕ^*

$$F_2^*(\sigma,a,t) := \varphi^*(\sigma,a,t) = \frac{1}{\big[\tau_\varphi(a)\big]^2}\int\limits_{S^1}\int\limits_{\mathcal{A}} \frac{\partial\ \delta_\epsilon(\sigma-\sigma')}{\partial\sigma'}\delta_\epsilon(a-a')\,n^*(\sigma',a',t)\,\mathrm{d}\,a'\,\mathrm{d}\,\sigma' \ \simeq\ \tau_\varphi^{-2}\,\frac{\partial n^*}{\partial\sigma}, \tag{125}$$

(126) $F_3^*(\sigma, a, t) := \phi^*(\sigma, a, t) =$

$$\frac{1}{\left[\tau_\phi(a)\right]^2} \int\limits_{\mathcal{S}^1} \int\limits_{\mathcal{A}} \delta_\epsilon(\sigma - \sigma')\delta_\epsilon(a - a')\, n^*(\sigma', a', t)\, \mathrm{d}\, a'\, \mathrm{d}\, \sigma'$$

$$+ \frac{1}{\left[\tau_{\phi_a}(a)\right]^2} \int\limits_{\mathcal{S}^1} \int\limits_{\mathcal{A}} \delta_\epsilon(\sigma - \sigma') \frac{\partial}{\partial a'}\, \delta_\epsilon(a - a')\, n^*(\sigma', a', t)\, \mathrm{d}\, a'\, \mathrm{d}\, \sigma'$$

$$\simeq \tau_\phi^{-2}\, n^* + \tau_{\phi_a}^{-2}\, \frac{\partial n^*}{\partial a},$$

where we have taken $\delta_\epsilon(\cdot)$ to be the Dirac delta function $\delta(\cdot)$ in the final approximations of (125) and (126).