Examining human-environment interactions and their impact on land-cover change during first millennia agriculture in Iberia
An agent-based modelling approach

Lane, Andrew

Awarding institution:
King's College London

Examining human-environment interactions and their impact on land-cover change during first millennia agriculture in Iberia

An agent-based modelling approach

Andrew J. Lane

PhD Thesis

Department of Geography

June 2023

Acknowledgements

Throughout the production of this thesis, I have received advice, support, and encouragement from many people, without whom this would not have been possible.

I want to thank my supervisor, Dr James Millington, for being a great companion on this journey. He is a gifted teacher, writer, and scientist and has a real knack for getting thing moving again when you're stuck.

My second supervisor, Dr Simon Miles, provided a fresh perspective on my work at key junctures during thesis writing, and helped to keep the project grounded.

Without the CANES programme and its members' commitment to cross-disciplinary research I might never have applied to do a PhD. Thanks to Professor Peter Sollich, Professor Chris Lorenz, Dr Joe Bhaseen, and Dr Alessia Annibale for your efforts over the last four years. Thanks also to my cohort on the programme. What a ride.

I want to thank Dr Francisco Seijo for reminding me that scientific research can be fun, and Dr Pooya Azarhoosh for setting a great example of what life for a scientist outside a university can look like.

Thank you to Anya for her patience, curiosity, and humour throughout this process. Thanks also to my parents for their continued supported and encouragement.

Abstract

Generative simulation models and the mathematical analysis of their outputs can help us to quantify, visualise and understand the impact of anthropogenic land cover change on terrestrial ecosystems. In this thesis I describe a socio-ecological simulation model of land-cover change in the Iberian Peninsula called AgroSuccess. This model integrates previous work from the literature describing ecological succession and ecological disturbance in the form of both fire and anthropogenic subsistence activities. AgroSuccess is an agent-based simulation model that enables users to explore the effect of agricultural land management practices and different climatic conditions on the emergent state of simulated landscapes. I demonstrate AgroSuccess by investigating the changes to land cover resulting from the introduction of agriculture during the mid-Holocene at six study sites in the Iberian Peninsula.

AgroSuccess requires input data to specify boundary conditions, as well as reference data against which to compare outputs and calibrate parameters. I have developed a collection of reusable software tools to obtain and prepare paleo-ecological pollen abundance data collected by previous researchers, as well as morphological data from remote sensing to characterise study sites. To the best of my knowledge the way in which I have synthesised these data from disparate sources is novel, and the approach I have taken can be easily replicated by others using the open source software tools that I have made available online.

To address my research questions AgroSuccess represents various ecological and anthropogenic processes. Consequently, it is an example of a complicated model that requires the collection and assimilation of multiple forms of data to parameterise and initialise simulation runs. The management, documentation, communication, and reuse of such a model is difficult. An example of how I have mitigated these challenges is the development of a software application, called Cymod, that helps users to visualise the state-and-transition model (STM) that is integral to AgroSuccess' ecological succession submodel. I argue that the measures taken to ensure the correct implementation of scientific simulation models are arbitrary, and often inadequate to provide scientists with the confidence they need to use each others' code. In response to this challenge, the software implementation of the AgroSuccess simulation model, and the scripts that process its input data, are modular and well-tested. By distributing these modular components

in public software repositories, I aim to make my work transparent and help others understand and reproduce my results.

Contents

Chapter 1

Introduction

Humans have come to dominate our planet. Following the technological developments that led to the industrial revolution in the 18th Century, global human population has surged from 1 billion in 1800 to a projected 9 billion by 2050 (Steffen et al., 2004). Increasing populations have led to increased demand for food and other products, and demand for the land to produce it on has increased accordingly. 10–15% of the earth's surface is now used for crop agriculture, and an additional 6–8% is used for pastureland (Vitousek et al., 1997). Since the first agricultural revolution, the mass of plants on Earth has decreased from two teratonnes (Tt) to one Tt, with the biomass produced by growing crops vastly offset by that lost through deforestation and other land use changes (Elhacham et al., 2020; Erb et al., 2018). These developments have altered the way terrestrial ecosystems function, including the mechanisms with which they interact with the atmosphere (Vitousek et al., 1997). Recent work (Guiot & Cramer, 2016) has predicted that unless urgent action is taken to mitigate climate change in-line with the most ambitious of the 2015 Paris Climate Agreement thresholds, this century will see dramatic climatic shifts in the Mediterranean Basin. These are likely to limit the region's capacity to provide the ecosystem services—food, timber, fibres, and other necessities—that society demands (Dearing et al., 2014).

In recent years the significance of these issues has been recognised formally by the international community. As part of the 2030 agenda set out by the UN in 2015, 17 Sustainable Development Goals (SDGs) were agreed upon (UN General Assembly, 2015). These set out global objectives to work towards for a better world. Each of these 17 goals have a number of targets associated with them, each of which are in turn associated with one or more indicators, intended to provide a way of quantifying progress towards attaining the SDGs. Of of these goals, one in particular

(SDG 15) succinctly describes the target towards which research into humans' interactions with our terrestrial environment might contribute:

> Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss.

Recent improvements in techniques for reconstructing past ecosystems (Carrión et al., 2010) could help us to understand how human land management practices in the present day influence ecosystem dynamics (Conedera et al., 2017). Pollen analysis (palynology) provides a way of studying how the abundance of pollen from different species has varied over time at a location. When pollen falls on wet sediment it can become fossilised, and by analysing the pollen found in sequential samples of an experimental sediment core it is possible to construct a set of simultaneous time series of the abundance of different species' pollen. A plot composed of several of these simultaneous time series are arranged is known as a *pollen diagram*. An example of a pollen diagram is shown in Fig. 1.1. Pollen diagrams record how the character of vegetation present in a landscape changed over periods of tens of thousands of years, and so provide a lens through which we can peer into the past. This is only possible because of efforts over the past few decades to improve the procedures used to date pollen sequences (Carrión et al., 2010).

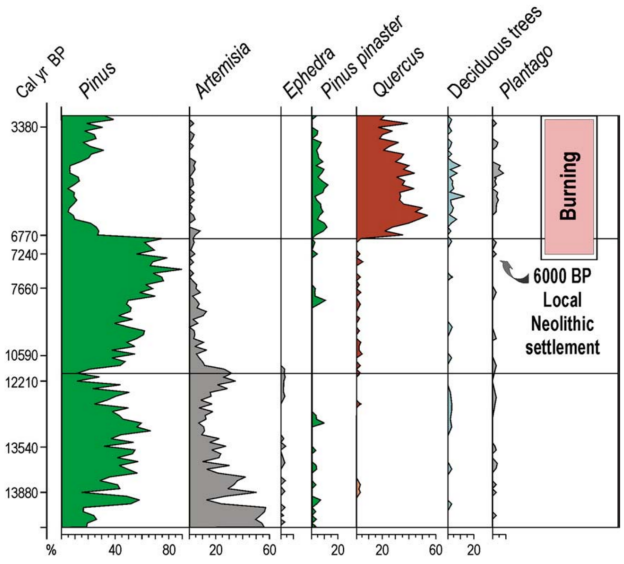


Figure 1.1: Example of a pollen diagram consisting of a series of simultaneous time series of the relative percentages of pollen contributed by various species detected at Navarrés, Spain, during the late Quaternary and Holocene. Figure taken from Carrión et al., 2010.

Of special interest for the purpose of understanding socio-ecological dynamics are the presence

of palynological indicators of the presence of humans in pollen diagrams: increases in grasses and shrub species which are hypothesised to indicate the landscape has been opened up using fire, for example (Carrión et al., 2010). Other trends in different species' pollen abundance time series might be indicative of the occurrence of a complex mixture of biophysical, ecological and climatic processes. In short, these data can be thought of as being a *signature* of the processes that created them Perry et al., 2016. The overarching aim of this thesis is the development of an agent-based model and its use to find the combinations of processes that best explain observed pollen sequences. Specifically I explore how ecological succession, anthropogenic land-use change, natural disturbance and climate interact to produce the patterns represented by pollen diagrams. This way of informing theoretical models using empirical data is known as Pattern Orientated Modelling (POM) (Grimm et al., 2005; Perry et al., 2016).

This thesis has the following aims towards the overall objective of improving understanding of the processes that drive land cover change:

Aim 1: Develop a spatially explicit agent-based simulation model that incorporates processes representing ecological succession, anthropogenic land-use change, and natural disturbance, and which is sensitive to climatic variation through its boundary conditions.

Aim 2: Ensure the model developed to meet **Aim 1** is documented and distributed such that it is useful to, and usable by, other researchers.

Aim 3: Devise and document a reproducible procedure to obtain and process the empirical data needed to provide boundary conditions and reference data for the ABM described in **Aim 1**.

Aim 4: Use the ABM described in **Aim 1** to explore counterfactual scenarios involving antrhopogenic land-cover change that are not possible using empirical data alone.

1.1 Thesis structure

In Fig. 1.2 I show how each of the chapters in this thesis build on each other to meet my research goals.

Chapter 2 introduces my approach to the scientific problem of how ecological succession, anthropogenic land-use change, natural disturbance and climate affect terrestrial ecosystem change. I

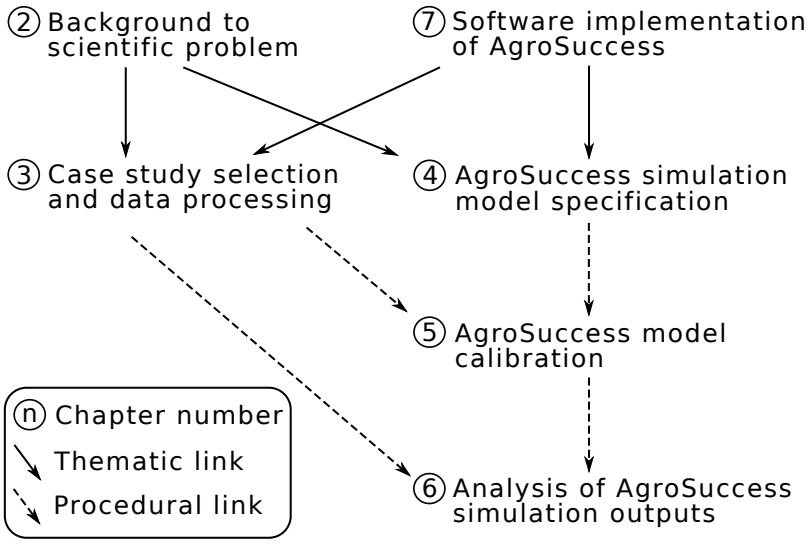


Figure 1.2: Illustration of how the chapters in this thesis relate to each other. See main text in this section for more detailed descriptions of the chapters and their relationships.

explain why understanding the respective roles of these different processes requires a view of landscape change over decadal to centennial timescales, and how the use of simulation modelling is key to overcoming the challenge of studying these processes empirically. I review the types of empirical evidence that can be used to learn about ecosystem change over the timescales of interest, and conclude the chapter with a discussion of why the human history and biogeography of the Iberian Peninsula make it a good choice of study region.

In Chapter 3 I describe the process I followed to select study sites within Iberia, contributing to **Aim 3**. These study sites serve as the basis for simulated experiments that help to address our scientific questions. I explain how I obtained empirical data characterising how land cover changed over centennial timescales, as well as morphological and climatic data for each of these study sites. I then specify how I processed this secondary data to provide boundary conditions and reference data for simulation runs representing the study sites.

The structure of the ABM described in **Aim 1**, which I have named AgroSuccess, is specified in Chapter 4. AgroSuccess combines aspects of a spatially explicit landscape fire succession model (J. D. A. Millington et al., 2009) with a model of small-holder agropastoral household agents (Ullah, 2013). These model components were previously described independently in the literature. I provide both a description and an open source implementation of an integrated model that enables users to test hypotheses concerning ecological and anthropogenic drivers of land-cover change. The model outputs are vegetation abundance time series which can be compared to

empirical pollen sequences to evaluate the degree to which models are able to explain the data.

Chapter 5 concerns the tuning of parameters belonging to the AgroSuccess model. In particular I specialise certain parameters to the study sites selected in Chapter 3 such that the simulated wildfire regimes represented during model runs correspond to empirically observed fire regime metrics described in the literature. In Chapter 6 I present outputs from simulations of the study sites selected in Chapter 3, and compare the time evolution of the proportion of the simulated landscapes occupied by different land cover types to the pollen abundance time series derived in Chapter 3. I also consider statistical properties of the simulated time series corresponding to similar analyses on empirical data described in the literature. I use AgroSuccess to explore scenarios with and without anthropogenic activity in a way that is not possible with empirical data alone. This corresponds to **Aim 4**.

In Chapter 7 I discuss how the adoption of certain software development best practices could improve confidence in other peoples' software in the scientific community. I argue that widespread adoption of testing and documentation standards will help ensure the correctness of scientific software, ensure that the results of computational experiments are reproducible, and increase trust in others' work. I describe how I applied these measures in my own work during the preparation of the empirical data described in Chapter 3, as well as in the software implementing the simulation model described in subsequent chapters. In the second part of the chapter I describe a novel application of graph database technology I have developed to help manage model complicatedness, and illustrate how I used this approach to implement the ABM described in **Aim 1**.

Chapter 2

Background to scientific problem

2.1 Spatio-temporal scale

To make progress in developing a methodology for understanding landscape scale human-environment interactions, a necessary step is to first decide on which landscapes to consider. This will limit the range of many factors—relating to geomorphology, climate, plant functional types, and human activities—to an extent which is manageable for modelling purposes. For reasons I will discuss in Section 2.2 I will focus on regions in the Mediterranean basin, and the Iberian Peninsula in particular.

It is also necessary to decide on a temporal range to study. Given my research interest in human-environment interactions, this range is determined by the period of time during which climatic conditions have been amenable to humans making longstanding changes to the landscape. This period is effectively coincident with the geological epoch known as the Holocene; that is, the stretch of time between the beginning of glacial retreat following the last glacial maximum (see e.g. Blondel and Aronson, 1999), and the present. Since glacial retreat is a global phenomena, there will clearly be some variation between the date of retreat at different sites. However, for the purpose of the selection of study sites, I will define the Holocene as the range 11,650–0 BP (Walker et al., 2009). Here the temporal unit 'years before present' (BP) is defined such that the 'present' is counted as the year 1950 by the Gregorian calendar. The use of 1950 as a reference year is an established convention in the radiocarbon dating literature (Flint & Deevey, 1961) and, as the data I am interested in makes use of radiocarbon dating, is therefore also a sensible

convention for my work.

2.2 Motivation for selecting the Mediterranean as a study region

My decision to focus on studying sites in the Iberian Peninsula within Mediterranean basin arose quite naturally out of the need to select *some* site to study on one hand, and the historical contingency of my supervisor's (James D.A. Millington) academic familiarity with the region on the other. Consequently my arguments for studying the Mediterranean which comprise this section are, admittedly, post hoc. That said, approaching the problem of study site selection as I have has given me the opportunity to think broadly about the environmental and socio-economic problems in the region, as well as the pertinent scientific questions one might ask to address them.

2.2.1 Biodiversity in the Mediterranean

The preservation of biodiversity ought to be a priority for governments around the world for many reasons. In the short term, biodiverse ecosystems are more productive because the many inter-dependent species in them can support each other by creating the finely tuned conditions each needs to thrive. Such conditions, if managed responsibly, are able to provide a secure source of the ecosystem services—such as crop yield and timber—which are required to sustain human communities (Cardinale et al., 2012).

In the longer term, biodiversity provides inspiration for human innovation. For instance, while it is possible to develop medicinal drugs using chemical combination and knowledge of the molecular target we would like to treat, the fact that 116 out of 158 new drugs licensed by the U.S. Food and Drug Administration between 1998 and 2002 were derived from natural origins highlights the advantages of studying natural medicines (Chivian & Bernstein, 2008). The successfulness of natural medicines should hardly be surprising; the myriad species on Earth interact with each other principally via chemicals, and so evolution has developed – over at least 3.5 billion years – various antibiotics, toxic peptides and other chemical defences to protect their hosts from their environment (Chivian & Bernstein, 2008). In addition to chemical innovation, evolution has also produced physical structures that we can take inspiration from for engineering applications. The

need to have their teeth withstand rasping on rocks provided the evolutionary incentive for the ancestors of modern limpets to develop exceptionally hard teeth (Barber et al., 2015). By studying their microscopic structure it is possible for scientists to exploit evolution to design the next generation of human dental implants. Since the discoveries of both natural medicines and structures with therapeutic applications are unexpected, failure to protect biodiversity runs the risk of destroying these sources of inspiration before they can even be discovered (Chivian & Bernstein, 2008).

The Mediterranean Basin has been noted to be particularly biodiverse. Conservation International maintain a list of 'Biodiversity Hotspots', defined such that they contain more than 1500 species of endemic vascular plants, and have less than 30% of their original (pre-industrial) vegetation cover (Mittermeier et al., 2004). The objective of the biodiversity hotspot concept is to focus conservation efforts on areas which are extremely valuable in terms of biodiversity, and which also bear the brunt of anthropogenic environmental pressures. The Mediterranean Basin is counted as one of these hotspots.

A recent simulation study (Guiot & Cramer, 2016) provides additional motivation to understand and protect biodiversity in the Mediterranean. With respect to the 2015 Climate Change Paris Agreement – which aims to limit global average warming to within 2.0 °C of pre-industrial levels – the authors found it likely that the Mediterranean would experience significant ecological change within the next 100 years. In particular, two simulated scenarios involving warming by 2.0 °C predict the onset of some degree of desertification in southern Iberia. This suggests that the need to understand how species key to the provision of ecosystem services will respond to such climatic shifts will become increasingly important over the coming decades.

2.2.2 A long history of human-environment interactions in Iberia

The state of the landscape, as observed at any particular point in time, is causally dependent, to some extent, on all the events which took place at that location in the past (Conedera et al., 2017). This means that in order to understand the role of humans as agents of environmental change today it is necessary to look back to the very start of the period when humans started changing the landscape. In the context of my project, this raises two questions whose answers – or relevant scholarly consensus – I will need to determine from the literature:

1. When and why did humans start manipulating vegetation cover in Iberia?

2. What factors (e.g. climate, fire regime, proximity of topological features such as rivers and the coast) determined when and where humans started changing the landscape at the local scale?

There are several recent papers which provide information relevant to these questions. In Martins et al., 2015 the authors treat the appearance of wheat, barley, sheep or goat – collectively termed the 'Neolithic package' – at a location and time as and indicator of the transition from the Mesolithic period (characterised by hunter-gatherer cultures) to the Neolithic period (characterised by an agro-pastoral lifestyle). It is widely accepted that Neolithic culture arrived in Iberia from outside (Martins et al., 2015), such that regional variation in the date of arrival of the Neolithic is expected. The authors use radiocarbon dating of artefacts of the Neolithic package to determine upper bounds on the date of regional Mesolithic-Neolithic transition, finding an earliest date for the beginning of the Neolithic in Iberia of 8500 BP. Since my working hypothesis is that it is the transition to an agro-pastoral lifestyle which led to anthropogenic landscape scale change, the existence of such analyses are essential for my project.

In addition to palaeoecological work concerning the beginnings of anthropogenic landscape change in Iberia, there has also been recent work by Gordó et al., 2015 to develop Agent Based Models to assess the likelihood of different scenarios of the spread of agriculture in Iberia. Finally Kaplan et al., 2009 provide a top-down model to help explain prehistoric deforestation in terms of land suitability to crop and pasture.

2.2.3 Disturbance and fire

The Mediterranean region is subject to regular wildfires, and has been for millennia. Plant species living in the Mediterranean today have been shaped by evolutionary processes in response to fire (Thompson, 2005). Mediterranean forests and woodlands have been described as "... perturbation dependent, nonequilibrium systems..." (Naveh, 1994), with fire playing an important role as an entropy generating process. The Mediterranean is home to both fire-resistant and fire-stimulated species – *pyrophytes*. Pyrophytes fall into two categories: sprouters and seeders. Sprouters maintain thermally-insulated underground structures that allow them to resprout after a fire. Seeders meanwhile disperse heat resistant seeds that establish quickly after a fire. There

even exist species – 'temporal dispersers' – whose seeds will remain dormant until they receive a temperature shock and then begin to germinate (Blondel & Aronson, 1999). The simple fact of the existence of pyrophytes in the Mediterranean hints at the complex relationships between ecological processes and fire, and highlights the need to better understand these relationships to develop effective policies to reduce wildfire risk.

It is likely that the changes brought about by industrialisation (including land use/ land cover change) have led to changes in the size, frequency and intensity of these fires which has, in turn, led to them becoming a serious concern for politicians in Mediterranean countries (J. M. Moreno & Oechel, 1994). Due to a legacy of unscientific (Naveh, 1994) treatment of wildfire as a purely destructive force with respect to nature conservation (akin to uncontrolled grazing), fire suppression has become the predominant policy, with large amounts of money spent on fighting fires each year (J. M. Moreno & Oechel, 1994). However, studies have found some evidence that current fire suppression policies may exacerbate the occurrence of large fires (Seijo et al., 2015).

The main factors that govern fire regimes are fuel availability, climate, and human activity. The interactions between these factors have been modelled in the literature (Seijo et al., 2016; Stephens et al., 2014) as a "mega fire triangle" (see Fig. 2.1). In a given landscape, fuel availability is influenced by the spatial distribution and structure (e.g. open vs. closed canopy forest) of the land cover. Climatic variation can lead to environmental conditions that are more or less conducive to fire ignition and spread, such as periods of drought. Human activity such as forestry operations and recreational use of areas in the wildland-urban interface (WUI) have been identified as a significant source of ignitions leading to wildfires (Stephens et al., 2014). Pausas and Fernández-Muñoz, 2012 have hypothesised that before the 1970s, the fire regime in Valencia, Spain was 'fuel limited' due to an anthropogenically-induced reduction in forest cover for agricultural purposes. By contrast, following a period of 'rural abandonment' during the 1970s, the fire regime in Valencia appears to have become 'climate limited' because in decades since the 1970s, climatic variables were better predictors of monthly summer burnt area than they were in earlier decades. This is an example of how factors represented in the fire triangle can shift to produce qualitative changes in a region's fire regime.

The apparent paradox of how fire suppression can lead to increased frequency of large fires can be explained in terms of the factors represented in the fire mega-triangle. Fire suppression can cause an accumulation of fuel load in the landscape that provides the means for wildfires to

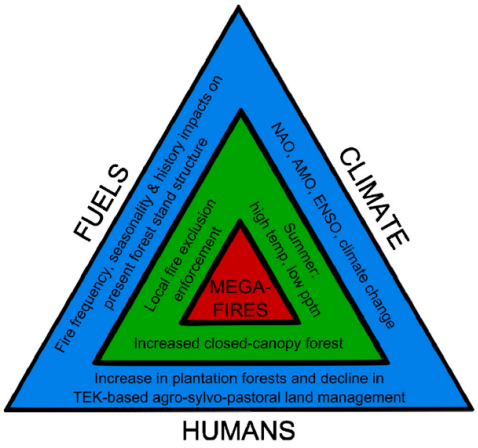


Figure 2.1: The fire “mega-triangle” illustrating the factors relating to fuel availability, anthropogenic land cover change and climate interact to create conditions suitable for extremely large fires. From Seijo et al., 2016.

spread further than they otherwise would. Multiple studies (Fernandes et al., 2013; Khabarov et al., 2016) have concluded that the use of prescribed burning (that is, the deliberate ignition of relatively small controlled fires) may be an effective way to reduce area burned by reducing fuel availability to wildfires. Seijo et al., 2016 compared the fire regimes in two Spanish municipalities in central Spain with comparable climatic and biophysical factors: Casillas and Rozas. Authorities in Rozas have implemented strict fire exclusion policies, whereas in Casillas traditional forest management practices including annual burning of litterfall are more widely tolerated. Despite the fire exclusion policies in Rozas, the observed burnt area relative to landscape size represented in official fire statistics was found to be 10 times larger in Rozas compared to Casillas for the period 1984 to 2009. While this study is small, its findings suggest that the reduction of fuel load through preventive prescribed burning may be a more effective way to reduce the incidence of large fires than fire exclusion.

2.2.4 Knowledge transfer to other Mediterranean type ecosystems

There are five regions around the world that have a ‘Mediterranean Climate’, characterised by cool, wet winters and hot, dry summers (Joffre & Rambal, 2002). These are:

- the Mediterranean Basin
- south western and southern Australia
- California

- southern Africa

- central Chile

Further, these regions are considered to be home to *Mediterranean Type Ecosystems* (MTEs) in the sense that in addition to having comparable climates, they also have similar types of vegetation community. In particular, there is a formation characterised by a predominance of woody shrubs with sclerophyllous (i.e. small, hard and leathery) leaves which is found in all five MTEs. This formation is variously referred to as *garrigue* or *maquis*, *chaparral*, *heath*, *matorral* and *fynbos* in France, California, Australia, Chile and South Africa respectively (Joffre & Rambal, 2002).

Convergent evolution in multiple regions around the world

The remarkable similarities between MTEs–combined with the accepted evolutionary isolation of Australian and South African species with respect to the other regions–provide an opportunity to test hypotheses relating to evolutionary convergence between species in different regions in response to similar climates (Joffre & Rambal, 2002).

From the perspective of my research aims, the similarities between the vegetation of the five MTEs provide an interesting and ambitious test of model performance. One could first develop a suite of models with species in the Mediterranean Basin in mind, based on life history traits such as evergreenness, sclerophylly, and disturbance response strategy (sprouter vs seeder, see Section 2.2.3). Although the species in the different MTEs differ, similarities in life history traits that have arisen as a result of evolution in response to climate could mean that the MTEs are comparable at the level of Plant Functional Types (see e.g. Rusch et al., 2003). In that case, a robust test of a model developed for the Mediterranean Basin would be to test whether it could convincingly explain empirical data collected from one of the other MTEs. Such an empirical test is outside the scope of this thesis, but is an interesting consideration for possible future work.

2.3 The Holocene

2.3.1 Overview of changes in human behaviour

The Holocene is the most recent epoch of the Earth's geological history. It is the second epoch in the Quaternary period, and is differentiated from the preceding epoch by the beginning of a clear trend of climatic warming following the final cold episode of the Pleistocene (termed the *Younger Dryas*). This period of mild climate began approximately 11,700 years ago and continues to the present day (Walker et al., 2009). In addition to being markedly warmer and wetter than the previous millennia since the last glacial maximum 20,000 years ago, the climate was also much more *consistent* during the Holocene. It is understood that this decrease in variation in temperature and precipitation over decadal timescales played an important role in providing the opportunity for our own species, *Homo sapiens*, to develop agricultural technology—and the societal changes which came with it—for the first time (Bellwood, 2004).

To put the impact which fledgling human agriculture had on terrestrial ecosystems during the Holocene into context, it is useful to reflect on humans' place in the world at the end of the Pleistocene. *H. sapiens* evolved in Africa, with fossil evidence showing that individuals' brains reached sizes comparable to those of contemporary people between 150,000 and 50,000 years ago. These individuals are considered to be anatomically modern. Anatomical modernity is distinguished from *behavioural modernity*—evidenced by the appearance of artefacts in the archaeological record which demonstrate the cognitive capacity for Culture—which is believed to have occurred between 50,000 and 40,000 years ago (Klein, 1995). This coincides with the earliest date for which there is evidence of the arrival of *H. sapiens* in Europe, which is also believed to have occurred 40,000 years ago (Hoffecker, 2009). At this time there were other species of archaic humans living in Eurasia whose ancestors had left Africa up to 1.4 million years ago, as demonstrated by archaeological finds in Israel (Klein, 1995). Notably, *H. neanderthalensis* coexisted with *H. sapiens* in Europe for millennia before going extinct within the last 28,000 years (Finlayson et al., 2006). Since *H. sapiens* were therefore the only human species extant during the Holocene, I will refer to them unambiguously as 'humans'.

Before the agricultural revolution which took place in the Holocene, humans subsisted by a combination of hunting and gathering. This intrinsically opportunistic method of subsistence would have led to a varied diet, with the risk of one source of food failing mitigated by the knowledge

of alternative sources available within a group's home range. Hunter gatherers lived in small groups whose size was allowed to fluctuate in response to food and resource availability, limiting their environmental impact. Additionally, pre-Holocene humans would have moved around often, further limiting their environmental impact and mitigating the risk of local food source failure (Moran, 2006).

The foregoing comments regarding the limitations of pre-Holocene human environmental impacts should not be taken to mean that early humans had no environmental impact. They would, for instance, have chased prey across the landscape—driving population dispersal—and influenced predator/ prey balance through their hunting (Moran, 2006). They also had access to a tool which allowed them to quickly reshape their landscape to their advantage: fire (Pausas & Keeley, 2009). Having mastered the controlled use of fire at least 500,000 years ago (James, 1989), humans were able to use fire to modify their environment in order to expedite travel, to assist in hunting, and to improve access to otherwise inaccessible edible shrubs, long before the beginning of the Holocene (Keeley, 2002; Pausas & Keeley, 2009). The use of fire for these purposes is collectively termed "fire-stick farming" (Bliege Bird et al., 2008; Pausas & Keeley, 2009).

2.3.2 The Agricultural Revolution

The 500 years between 11,000 and 10,500 yrs BP marked the beginning of a dramatic change in humans' relationship with their environment—the Agricultural Revolution. This period is also equivalently termed the Neolithic Revolution, where *Neolithic* refers to the last stage of human technological development of the stone age, characterised by the development of agriculture. For the purposes of this thesis, I will consider the Agricultural Revolution to be a series of developments that took place in an area of Southwest Asia known as the *Fertile Crescent* occupying what is now the Jordan Valley, and parts of Turkey, Syria and Iraq. This is because the technological progression to the Neolithic era in the areas I will consider as case studies (see Chapter 3) was triggered by the spread of developments out of the Fertile Crescent. It is worth noting, however, that parallel revolutions culminating in the development of agricultural technology took place at approximately the same time in East Asia, Africa and the Americas (Bellwood, 2004).

The relatively short time scale of 500 years for such a profound technological shift to occur in the Fertile Crescent, combined with the fact of the independent occurrence of similar developments elsewhere in the world raises the question of what was special about 11,000 yrs BP to facilitate

the beginning of agriculture. Bellwood, 2004 asserts that it was the shift towards the warm, wet and stable climate following the end of the Younger Dryas cold period at 11,500 yrs BP which created conditions amenable to agriculture. In particular, the increase in winter rainfall accompanying the beginning of the Holocene provided ideal growing conditions for cereals and legumes.

It has also been suggested that human population growth was an additional motivating factor in the development of agriculture (Balter, 2010; Bellwood, 2004). As climate became increasingly favourable, the amount of available wild food—including pistachio, olives, acorns and other nuts, wheat and barley—increased. This created the prospect for humans to settle down in one location all year round, or for part of the year at least, and establish settlements near abundant wild food sources without the need to travel to find enough food for sustenance; a lifestyle known as sedentism (Bellwood, 2004). It is believed that the conception rates of women in non-sedentary hunter gatherer groups are suppressed by biological factors relating to diet and the need to carry young children (Bellwood, 2004). Consequently, by adopting sedentism, the people living in the Fertile Crescent 11,000 years ago may have unwittingly amplified their birth rates. This would have created the necessity for more food to be produced from the same area of land near their settlements, providing an incentive for the development of agriculture. For instance, Bellwood, 2004 suggests that while a family of hunter gatherers might have required more than 100 ha of land for sustenance, a family of shifting agriculturalists (who farm a plot temporarily before abandoning it) would require less than 10 ha. Furthermore, a family of irrigation agriculturalists might require less than 1 ha of land to sustain them.

A foundational technology which arose 11,000 years ago was the cultivation of plants. Bellwood, 2004 defines cultivation as "... a sequence of human activity whereby crops are planted (as a seed or vegetative part), protected, harvested, then deliberately sown again, usually in a prepared plot of ground, in the following growing season.". The first agriculturalists are known to have cultivated cereals and legumes including wheat, barley, peas and lentils. They also practiced artificial selection by favouring cereals whose grains are larger than the wild type and which ripened simultaneously, and whose ears wouldn't shatter to release the grain without human intervention. These selected mutants would therefore already have depended on human intervention to complete their life cycle. In addition to crops, there is also evidence of specialised sheep and goat pastoralism, and the domestication of pigs and cattle in the Fertile Crescent before 9,000 yrs BP (Bellwood, 2004).

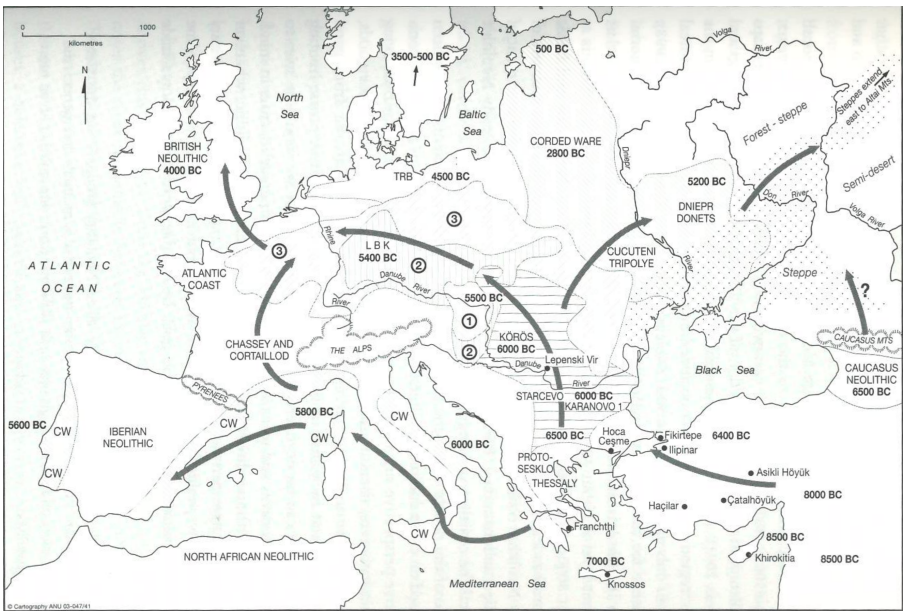


Figure 2.2: Map showing the spread of agricultural technology out of the Fertile Crescent. From Bellwood, 2004.

After a thousand year incubation period in the Fertile Crescent, the technology of the Agricultural Revolution began to spread further afield into Europe and South Asia. This is understood to have been driven by a need to find new areas to cultivate due to land degradation, and the need to grow more legumes to feed animals. Both of these factors emerged as a consequence of population growth of both plants and animals (Bellwood, 2004). They also foreshadow the impact which the subsequent millennia of human agricultural land use were to have on ecosystems world wide. Agropastoral technology reached Europe between 8500 and 6000 yrs BP (Bellwood, 2004), and reached as far as Portugal between 6800 and 6200 yrs BP (Rowley-Conwy, 1995; Zilhão, 1993) (see Fig. 2.2).

2.4 Concepts concerning terrestrial ecosystem change

This section defines concepts from the fields of ecology (including landscape ecology and historical ecology) and biogeography that are referenced in subsequent chapters. They are defined here to provide context and background for readers with an interest in simulation modelling, but without specific training in these fields.

The field of Landscape Ecology

Over the last 25 years, much of the scientific investigation into the role of humans in affecting terrestrial ecosystem change has taken place in the field of *Landscape Ecology*. This highly cross-disciplinary field has been described as "... a young branch of modern ecology that deals with the interrelationship between man and his open and built up landscapes..." (Naveh & Lieberman, 1994). It is a human ecosystem science which incorporates techniques and practices from geography, ecology, landscape planning and landscape management.

An ecosystem is a theoretical entity encompassing all the organisms in a particular area, along with all the processes which allow them to exchange energy and matter amongst themselves, and with their physical environment (Aber & Melillo, 2001; Ricklefs, 2000).

Land cover/ land use change

As human population demands increase in a particular area, land cover is converted from a relatively unmanaged state to one whose anthropogenically imposed purpose is to produce food crops, raise animals, provide space for property development, etc. (Turner & Gardner, 2015). This process of land cover/ land use change can have dramatic impacts on ecosystems both locally to where the change has taken place and further afield. For example, cutting down trees in a woodland to provide agricultural land can have multiple secondary effects on ecosystem function. First, it reduces the ability of the ground to absorb water, increasing the risk of flooding further downstream. The water retention capacity of woodland is an example of an *ecosystem service* not associated with a product. Second, loss of ground cover also contributes to erosion and loss of soil, potentially reducing the fertility of the land (Redman, 1999). Third, cutting down trees reduces the landscape's natural carbon sequestration. The beginning of agriculture represents a step change in the intensity with which humans would have caused land use/ land cover change. Before the agricultural revolution, humans subsisted as hunter gatherers who would have modified the landscape through deliberate and accidental fire and impacts on animal distributions through hunting (see Section 2.4.1). However, these land use/ land cover modifications would have been at lower intensity compared to those made in permanent settlements following the agricultural revolution (see Section 2.3.1).

Disturbance

Pickett and White, 1985 define 'ecological disturbance' as follows:

> "A disturbance is any relatively discrete event in time that disrupts ecosystem, community, or population structure and changes resources, substrate availability, or the physical environment."

Note that this definition is broad enough to incorporate a wide range of phenomena incorporating individual tree falls resulting from high winds during a storm, insect outbreaks, wildfire, and anthropogenic land management practices such as stand clearance for timber or agricultural purposes. This generality will be important in framing the conceptual model for my simulations, since the vegetation response to a lightning induced wildfire, and to the practice of anthropogenic 'slash and burn' farming methods will be essentially equivalent.

Disturbance plays a key role in the evolutionary strategies of many plant species, with some requiring disturbance to complete their life cycle. In many ecosystems (including those of the Western Mediterranean), fire is an important source of disturbance and should not necessarily be viewed as a purely destructive phenomenon (see Section 2.2.3).

Ecological succession

The species assemblage, i.e. the relative proportions of different species present at a location, are expected to change over time (Redman, 1999). This process is known as *ecological succession*. Succession is often framed in relation to a particular disturbance as the temporal starting point for a sequence of species assemblage stages. Ecological succession is a form of self-organisation (Redman, 1999).

2.4.1 Humans as part of nature

A pervasive attitude in the modern world–especially among those who would consider themselves to be an environmentalist–is that there are areas which are uninhabited by people and which are therefore entirely 'natural' and pristine. These places, and their colocated ecosystems, are

treated reverentially by environmentalists, and their special status is recognised legally by the formation of national parks (consider the Lake District and Peak District in the UK, and Yosemite and Yellowstone in the US). Such an attitude helps to create the impression that there is a clear distinction between humans and their artifices – cities, infrastructure, cultivated land – and *true* nature (or *wilderness*) – windswept mountains, impenetrable rainforests, and wild savannahs. This distinction, which Cronon refers to as "wilderness dualism", can be shown to be both ill-founded by point of historical fact in specific cases, and to be fundamentally a cultural construct more generally (Cronon, 1996).

The constructed nature of the concept of wilderness can be seen from the shift in meaning which it has undergone within the last 250 years. Before the 19th century, the wilderness was associated with desolation and waste. For example, according to the Judeo-Christian tradition, the wilderness was where Adam and Eve went after being exiled from the garden of Eden, where Moses led the Isrealites for forty years following their flight from Egypt, and where Jesus of Nazareth was tempted by the devil (Cronon, 1996; Redman, 1999). In contrast, by the time Wordsworth and his contemporaries within the romanticism movement were writing, the natural landscapes of the wilderness provided the backdrop for religious experiences which made these authors feel closer to God. Untouched nature had become culturally identified with the *sublime*. Cronon, 1996 argues that through a process of domestication which occurred over a period of decades as increasing numbers of people took to exploring natural landscapes for recreation, this recently derived sense of reverence for pristine nature gave way to the modern environmentalism movement.

Aside from the constructedness of the concept of wilderness, archaeological evidence can demonstrate that even those ecosystems which are given national park status, or are considered as otherwise important due to their perceived naturalness, have been perturbed to some extent by human action. It is known, for example, that Woolly Mammoths existed in Britain and Ireland (Kahlke, 2015) until as recently as the late Pleistocene, 14,000 yrs BP (Lister, 2009). Ongoing scholarly contention (Redman, 1999) makes it difficult to attribute the Quaternary megafaunal extinction to human action directly. However, a recent study using discriminative modelling to identify whether human or climatic influences were the best predictors of megafaunal extinctions in regions around the world, Bartlett et al., 2016 found that most of their high performing models' power in explaining the extinction of megafauna in a region was attributable to the arrival of humans. It is currently impossible to substantiate the claim that Mammoths would be roaming the

English Lake District had humans not intervened. However, the fact of their absence combined with uncertainty around the degree of human involvement in their demise serves to highlight what we don't know about the role pre-Holocene humans played in influencing the distribution of species which leave less obvious fossilised indications of their presence.

Sites that were subject to land use change since the development of agriculture provide more specific examples of how ancient humans made comprehensive changes to land cover in regions which might be naively taken to be untouched by human action. Consider, for instance, the rainforests of Central America, which were the home of the Mayan civilisation. It is estimated that, at their peak, the Maya cleared up to 75% of their lands for the purpose of agriculture. Nevertheless, by the time the Spanish arrived in Central America in the sixteenth century, they were met by unbroken rainforest (Redman, 1999).

While there is clearly cultural and aesthetic value in protecting areas of exceptional natural beauty from being subsumed into the urban sprawl, there are important practical issues raised by adopting a strong form of wilderness dualism. By revering the untouched wilderness and regarding all human uses for any non-human part of the world as abuse, we are in danger of leaving no room for people at all. Cronon, 1996 suggests that the solution to this is to recognise humans as being a part of – not aside from – nature, and to seek a responsible, principled, middle ground, whereby humans can make informed choices about how to use parts of the non-human world which is sustainable for both ourselves and the non-human world itself. A pertinent question, then, is how to derive the principles upon which this responsible middle ground should be determined. I see the field of *Historical Ecology* (Crumley, 1994; Moran, 2006) as being an important source of inspiration for addressing this question. Historical Ecology emphasises the importance of incorporating human decision making into explaining the present state of ecological systems. By understanding how humans have impacted ecosystems in the past, we might be able to learn how to limit our impacts in the present (Beller et al., 2017).

2.5 Evidence of land cover change over centennial timescales

2.5.1 Pollen

To provide evidence in support of certain models rather than others, I require empirical data against which I can compare model outputs. Due to the timescales involved in the processes I would like to model, the only evidence I can use to compare their results to is necessarily palaeoecological. In particular I will use datasets describing the abundance of fossilised pollen in sediment cores extracted by researchers in the field of palaeoecology (Carrión et al., 2010; Deza-Araujo et al., 2022; Deza-Araujo et al., 2020). Since fossilisation can only take place under specific circumstances, useful cores can only be extracted from sediment which was wet at the time of pollen deposition, such as the banks of ancient rivers or lakes (Franks, 1957).

Sediment cores can be used to build a picture of the relative proportions of different species of plants in a landscape over time—a key output for my models. By taking a small sample of sediment and treating it chemically to release the pollen and charcoal within (see e.g. Magri and Sadori, 1999 for details of this process), it is possible to inspect the fossilised plant remains under a microscope and identify the presence of different species by the characteristic appearance of their pollen. By counting the number of each species' pollen grains in the sample, researchers can calculate the proportion of pollen contributed by each species in a process known as *pollen analysis* (Franks, 1957).

Pollen analysis can be combined with radiocarbon dating techniques to determine the age of sediments at different depths below the surface at a study site of interest in a process known as pollen stratigraphy. By extracting a vertical core of sediment and performing pollen analysis on samples taken along its length, it is possible to construct a composite time series – known as a pollen diagram – representing the relative abundance of each species' pollen present around the study site over periods of millennia (Magri & Sadori, 1999). By extracting sediment cores at multiple locations in a landscape, pollen stratigraphy can be used to infer the distribution of plant taxa in both time and space.

Recently Conedera et al., 2017 set a precedent for using plant fossils for the study of the relationships between forest composition, human influences climate and disturbance, as I plan to do. These authors are concerned with the establishment of effective silvicultural practices in

the Swiss Alps, and argue for the importance of palaeoecological species distribution data in ecosystem modelling. I fully agree with them, but would also note that due to there being more precipitation in the Swiss Alps than in Iberia, one would expect there to be far fewer fossilisation-friendly sites in Iberia which can provide the high quality data Conedera et al., 2017 were able to obtain for sites in the Alps. Consequently, an important objective of my modelling efforts will be to utilise what data is available for Iberia as efficiently as possible.

A nuance of pollen analysis data which needs to be considered in the context of my project is the fact that each sample represents the *proportion* of each species' (or genus') pollen present at that (stratigraphically inferred) time. Since my models will produce output in terms of the *abundance* of each species/ genus, raw model outputs cannot be compared directly to empirical data. Due to a phenomenon known as the *Fagerlind effect* (Prentice & Webb, 1986; Reitalu et al., 2014), the relationships between the relative proportions of different pollen types present in a sediment sample and the abundance of various species/ genuses living in the area at the time are expected to be non-linear (Reitalu et al., 2014). This is a result of variation in the pollen productivity and deposition rates between different species/ genuses.

2.6 Learning from computer simulations

Simulations are a form of scientific model that can be used to encode knowledge of how a system or process evolves in time. In this way they are analogous to scientific models that are expressed as differential equations that describe how a variable of scientific interest changes over time. However, by leveraging the memory available to a computer, simulations can integrate many more data sources and processes than a human being can reason about simultaneously unassisted.

2.6.1 The role of simulation models in science

The act of creating a model formalises knowledge and provides an object of focus for future researchers. Describing a model requires the modeller to be explicit about which processes they consider important to explain the real-world phenomena their research questions commit them to explaining. They also need to be explicit about which parameters are needed to describe those

processes mathematically. Once tentative mathematical descriptions of processes are written down, parameter values tabulated, and results plotted, future researchers can scrutinise the work in detail. In particular, they can use evidence of the importance or sensitivity of a model to particular parameters to motivate subsequent empirical work to improve best-estimates of those parameters. This will, in turn, increase confidence in the outputs of models using the updated parameter values.

2.6.2 Difficulty of direct experimentation

Direct experimentation in landscape ecology is extremely challenging because of the large spatial and temporal extents over which relevant processes take place. An example of a large-scale ecological experiment is the Metatron (Haddad, 2012). This experiment contains 48 patches of $100\,\mathrm{m}^2$, and is used to perform replicated, controlled experiments on animal dispersal patterns. The study area involved in the Metatron experiment is remarkably large, and yet $100\,\mathrm{m}^2$ patches would be insufficient to address problems in landscape ecology.

An alternative approach is to develop *natural experiments*, such as the WrEN project (Watts et al., 2016). This involves the selection of study sites that exemplify particular values of ecological variables of interest. In the WrEN project, this approach is used to study biodiversity. By collecting a large enough sample of such sites, it is possible to argue that differences in the value of a dependent variable (e.g. the presence of a species of bird) can be explained by the differences in an independent variable (e.g. the proportion of the land nearby that is used for agriculture) when other independent variables are held constant across sites. This methodology limits the range of experimental treatments that can be applied because it is only possible to investigate effects that have been observed somewhere at least once.

2.6.3 Agent Based Modelling

An Agent Based Model (ABM) is a computational model that represents discrete social units (called ‘agents’) that interact with each other and their environment, and can make autonomous decisions (Axtell et al., 2002). An agent can be any entity that can be meaningfully attributed the ability to make decisions and interact in this way, such as an individual person, a household, or a nation state. ABM is a ‘bottom up’ modelling approach that allows practitioners to explore how

individual-level decisions give rise to system-level outcomes (Retzlaff et al., 2021).

Complexity

A system is complex if it exhibits emergent behaviour as a result of interactions between autonomous sub-units (often competing for a limited resource) with no centralised controller (Johnson, 2009; Sun et al., 2016). The system-level behaviour is emergent in the sense that it cannot be predicted based on knowledge of the behaviour of the sub-units alone. ABMs, and the systems they are employed to model, are complex systems. In the context of an ABM of land cover change, the autonomous agents might be people competing for access to preferred areas of land. If the agents are able to alter properties of the land they access, and those changes are persistent to some degree over time, this creates both spatial and temporal *interdependencies* between agents and their environment. Interdependencies are known to create "nonconvexities", i.e. a non-smooth functional relationship between the model's parameters and its outputs. Additionally such systems can have multiple equilibrium states, with the equilibrium state the system reaches in a given realisation of the model being dependent on its initial conditions (Parker et al., 2003). This presents a challenge for the calibration and interpretation of outputs of models of complex systems.

Scale

In ABM, a concept of scale is implied by the organisational hierarchies of agents represented in a model. In a model of land use change, for example, we might choose to represent individual people, households, and villages as agents. Decisions and behaviour at the person scale level would aggregate up to the household level, and decisions and behaviour at the household level would aggregate up to the village level. At each scale level, agent behaviour leads to emergent behaviour at the next scale up in the hierarchy, which in turn influences behaviour at the next level (Manson et al., 2012).

Advantages and disadvantages of ABM

ABM allows practitioners to specify models in terms of intuitive, real world concepts even when the behaviour of the system of a whole is not well understood. ABM can be considered in contrast to 'aggregate' or 'top-down' modelling approaches in which the behaviour of an aggregate quantity is modelled directly. One example of a top-down model in ecology is the Lotka-Volterra population model, in which in the population sizes of predator and prey species are modelled using a system of differential equations (Wangersky, 1978). In ABM, the behaviour of aggregate quantities of interest emerge as a consequence of the interactions between the entities that are represented in the model at a finer scale than the aggregate quantities themselves. This allows an ABM to provide an explanation of how fine-scale interactions cause aggregate-level behaviours in a way that a top-down model could not. Additionally, ABMs make it possible to represent heterogeneity among the represented fine-scale entities (O'Sullivan et al., 2012). For example, a land use change model in which both the area of land that is required per household *and* the spatially heterogeneous properties of that land influences each household's behaviour could not be fully realised with a top-down model.

Despite their advantages, ABMs (especially those with highly heterogeneous agents) tend to require more parameters than a comparable top-down model to represent their modelled processes. Consequently, ABMs are resource intensive in terms of both development time and computational expense (O'Sullivan et al., 2012). Moreover, the increased number of parameters required for ABMs make them more difficult to calibrate (Manson et al., 2012) and validate (Retzlaff et al., 2021) than comparable top-down models. These disadvantages need to be considered in relation to the objectives of a model before deciding to use ABM.

Modelling purpose

When developing any model (ABM or otherwise) it is important to be clear about the purpose of the model, because a model's purpose determines the criteria by which it will be evaluated, and its results understood (Edmonds et al., 2019). One such modelling purpose is *prediction*. For a model to be used for prediction it needs to be able to reliably produce estimates of one or more quantities or behaviours that were not known before the model was run. The requirement that such predictions be reliable is essential to this use case because if they are not then users would

have no confidence that any particular prediction could be relied upon (Edmonds et al., 2019). An example of a class of models that are used for prediction is those that are used to produce weather forecasts.

Alternative modelling purposes that have been proposed in the literature and are relevant to the work in this thesis are *Guide data collection*, *Illuminate core dynamics*, *Illuminate core uncertainties* (Epstein, 2008), and *Explanation* (Edmonds et al., 2019). To evaluate whether or not a model explains the observed phenomena that it is designed to represent, a modeller would use the model to run one or more scenarios (each comprising a specific set of boundary conditions and parameters) for which there is corresponding empirical data that the model's outputs can be compared to. The model can be said to explain the modelled phenomena if its outputs satisfactorily approximate the empirical data. The fact that a given model's outputs are consistent with empirical data does not imply that it is the *only* model that can explain the data. It is possible that there are other models that explain the data as well or better. This is referred to as equifinality (Poile & Safayeni, 2016). Explanatory models are useful because the process of creating them involves the specification of unambiguous logical rules and assumptions that can be scrutinised and evaluated (Epstein, 2008). This is a valuable scientific exercise because it makes assumptions explicit, and allows us to identify situations where the modelled rules do *not* satisfactorily explain observed phenomena. This, in turn, can motivate a revision of the represented processes and highlight gaps in available empirical data in the relevant field. The model presented in this thesis is intended to be used for explanation in the sense outlined here.

2.6.4 Comparable models

In this thesis I present an Agent Based Model that represents the effects of subsistence agriculture on land cover change during the mid-Holocene (see Chapter 4). An early example of the use of ABM to represent prehistoric communities' interactions with their environment is the Artificial Anasazi model (Axtell et al., 2002). This model represents demographic change and spatially explicit settlement patterns in the Long House Valley, Arizona (U.S.) from 800 CE (Janssen, 2009). Axtell et al., 2002 state their motivation for using ABM as follows:

> "Agent models offer intriguing possibilities for overcoming the experimental limitations of archaeology through systematic analyses of alternative histories. Changing

> the agents' attributes, their rules, and features of the landscape yields alternative behavioral responses to initial conditions, social relationships, and environmental forcing."

More recently, Wainwright, 2008 used spatially a explicit ABM to investigate the effect of interactions between land cover and anthropogenic activity (including agriculture) on soil erosion in Mediterranean landscapes. Additionally, modelling work by Henne et al., 2013 is comparable to the work presented in this thesis. Those authors used the LandClim model (Schumacher et al., 2004) to simulate the effect of fire and anthropogenic ungulate browsing activity on land cover from $7000\,\mathrm{yr\,BP}$ to the present. In their model, browsing intensity is represented as a 'top down' aggregate disturbance process that has the effect of reducing biomass in the landscape, rather than an agent-based model.

Chapter 3

Case study selection and data processing

This chapter begins by explaining how I selected a set of study sites against which I will calibrate the AgroSuccess model in Chapter 5, and which I will use as the subjects of simulated experiments in Chapter 6. In Section 3.2 and Section 3.3 I describe the different types of data I use to characterise the study sites, and explain the data processing steps I undertook to prepare these data for inclusion in my simulation modelling procedure.

3.1 Study sites

In this section I discuss the sources of secondary data that are available to characterise study sites, and provide the rationale by which I selected sites for inclusion in my analyses.

3.1.1 Data sources

There are two databases that I have identified as sources of pollen abundance data while selecting study sites: the European Pollen Database (EPD) (Fyfe et al., 2009; Giesecke et al., 2014) and Neotoma (Goring et al., 2015; Williams et al., 2018). Both the EPD and Neotoma collate paleoecological data relevant to this thesis and allow access through a standard interface; in the case of the EPD this is a documented PostgreSQL database structure, whereas Neotoma provides

a web-based API. A clear advantage of accessing data through a database or API (rather than by downloading data in the form of flat files for each site) is that queries can be reused. For example, one might write a script to obtain the pollen sequence for a given site from one of the databases. By simply replacing the site identifier string in this script with a variable, it is possible to obtain the corresponding data for *any* site in the database. This approach will save effort in the long run (since it will not be necessary to process data from flat files into a unified format), and enables one to write programs to search large amounts of data automatically.

Neotoma is a project borne out of an ongoing collaboration between the EPD and the North American Pollen Database (Goring et al., 2015). It contains both pollen and charcoal data, and also includes also includes data describing many other forms of palaeoecological evidence relating to the study of fauna. As one might expect given the involvement of the EPD in the Neotoma project, much of the data included in the EPD also appears in Neotoma. There is also an R package (called `neotoma`) which can be used to query Neotoma programatically via http requests, conveniently returning data in a format which is easy to manipulate as part of a larger R program.

The EPD is a database which was established in the 1980s specifically for the collation, archive and distribution of European Quaternary pollen data (Goring et al., 2015). The database also now includes plant microfossil data, including charcoal. While there is a page on the EPD's website intended to allow researchers to browse individual sites, the manual browser-based interface precludes automated filtering of study sites. Rather than rely on the browser interface, I downloaded the entire EPD as a Postgresql database. This means of distribution makes querying the EPD more difficult than querying Neotoma initially. However, I found that the EPD generally contains data on more sites within Iberia than Neotoma. See Section 3.2.1 for details of a software application I have developed to simplify interaction with the EPD.

3.1.2 Subdividing Mediterranean-type ecosystems

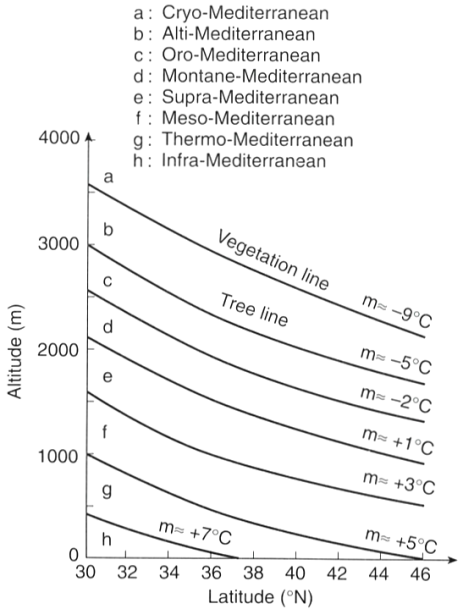
In Section 2.2 I introduced the idea of a Mediterranean-type ecosystem (MTE)—ecosystems characterised by cool, wet winters and hot, dry summers (Joffre & Rambal, 2002). As all the study sites selected will be in the Iberian Peninsula, all of them will be examples of MTEs. While selecting study sites, however, it is useful to be able to be more specific in my characterisation of the bioclimatic conditions at each site. This is because differences in bioclimatic conditions between sites are expected to cause differences in the corresponding ecological processes. To diagnose

whether AgroSuccess is able to characterise the ecological processes in locations within Iberia with slightly different bioclimatic conditions, it is important that the study sites I evaluate it against represent a range of such conditions.

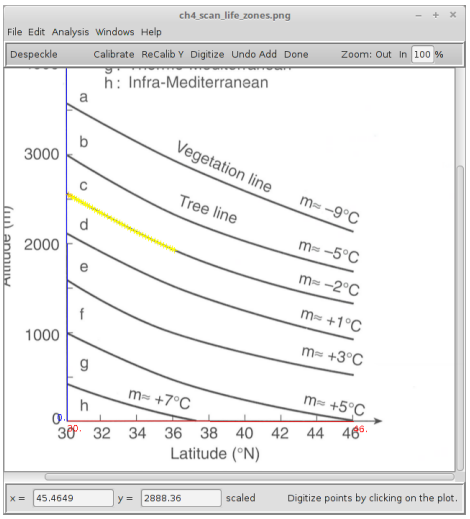
Previous work has classified the bioclimatic conditions found in Mediterranean regions into discrete bands. We can then think in terms of the flora and fauna we would expect to find in those conditions. An established way of doing this relies on the concept of **bioclimatic life zones**. Blondel and Aronson, 1999 define life zones as "elevational/latitudinal belts of plants and animals which tend to share ecological affinities and occur together." In practice, such zones can be determined by identification of the two or three dominant tree or shrub species in an area. Blondel and Aronson, 1999 define and describe eight such life zones. However, of these I will focus on three specifically. In warm regions at low altitude and latitude we have the *thermo-Mediterranean* life zone which is characterised by almost all woody plants being evergreen and sclerophyllous, and whose indicator trees are the cork oak (*Q. suber*) and the cluster pine (*P. pinaster*). At higher elevations and latitudes the *meso-Mediterranean* life zone occurs, indicated by the proliferation of the evergreen holm oak (*Q. ilex*) and the Aleppo pine (*P. halepensis*). Finally at still higher altitudes and latitudes we reach the *supra-Mediterranean* life zone which is dominated by deciduous oak forests. In Northern Spain the downy oak (*Q. humilis*) is a useful indicator species.

The use of life zones to categorise combinations of bioclimatic conditions into qualitatively differentiated categories makes it possible to abstract away the fine-grained differences between study sites. While this categorisation is somewhat artificial, it will simplify thinking and discussion about differences in socio-ecological dynamics within and between life zones. On the other hand, authors who have collected palaeoecological data from study sites are not necessarily aware of the notion of a life zone, and nor would they necessarily agree with the distinctions between them described by Blondel and Aronson, 1999. Therefore it is necessary for me to categorise sites into life zones myself, based on meta data that are reliably associated with study sites by the original authors.

Due to the qualitative nature of the distinctions between life zones, my means of classifying study sites is by my own convention. I make use of a semi-quantitative plot provided in Blondel and Aronson, 1999 (reproduced in Fig. 3.1a) which delineates boundaries between life zones as a function of latitude and altitude—data that are available for all study sites represented in the



(a) Plot of Life Zone boundaries after Blondel and Aronson, 1999



(b) Process of digitising the Life Zone boundary plot using the Plot Digitizer tool (Huwaldt, 2001)

Figure 3.1: Screen-shot of the Plot Digitizer programme in the process of digitizing the plot from Blondel and Aronson, 1999. Yellow points are points along curve d which have already been digitized, and the program's approximation of the plot's Latitude and Altitude axes are indicated in red and blue respectively.

EPD. To help me classify sites into life zones programmatically, I created a digital representation of the plot in Fig. 3.1a. First I represented each of the curves in Fig. 3.1a numerically by using Plot Digitizer (Huwaldt, 2001), an open source project that provides a GUI to sample points in an image of a plot within the plot's own coordinate system. A screen-shot of the process of sampling points along one of the curves can be seen in Fig. 3.1b. In this image, the red and blue lines are the latitude and altitude axes of the plot as represented within Plot Digitizer. The string of yellow points shows the sampled data. As can be seen, this GUI allows one to sample points from a plot image with high resolution. Repeating this process for all curves in the plot, I obtained between 53 and 120 latitude-altitude pairs for each curve (depending on the curve's total length). These data were then exported to .csv files.

Having obtained numerical samples from the life zone boundary curves, I then used the R programming language to create objects representing approximate functions of the curves. In particular, I used R's `approxfun()` function to perform linear interpolation between the sampled points for each curve. This enabled me to obtain the life zone within which any altitude-latitude pair represented on the original plot falls.

3.1.3 Study site selection criteria

I decided to focus on a selection of six study sites for the analyses described in this thesis. This is a compromise between the requirement to confront AgroSuccess with data from a variety of sources, and the need to be able to develop a qualitative understanding of the socio-ecological processes that took place at each site based on the analysis of the original authors. This latter requirement is important to ensure I am able to interpret my own results. In addition to the inclusion of sites from various life zones and locations, I also aim to also ensure that the selected sites are associated with data which covers a significant proportion of the temporal range of interest (the Holocene). Based on these requirements I selected study sites according to the following criteria:

- 4 sites should be in Mesomediterranean life zones.
- 1 site should be in a Thermomediterranean life zone.
- 1 site should be in a Supramediteranean life zone.
- Sites should be distributed as evenly as possible around Iberia to ensure adequate spatial coverage.
- There should be variety in the date of onset of human activity between sites.
- All sites should have palynological (pollen abundance) data associated with them.
- The temporal range of all sites' datasets should cover at least 50% of the Holocene.

3.1.4 Identification of candidate study sites

To create a long-list of study sites I drew on previous work by Carrión et al., 2010 in which the authors summarise trends found among 156 palaeoecological sites whose data describe the vegetation history of the Iberian Peninsula and Balearic Islands throughout the Late Glacial and the Holocene. I draw on the expert scholarship of the specialist biogeographers who wrote this paper (and the contained references) to interpret the palynological data sets associated with the study sites. The journals that published the original papers reviewed in Carrión et al., 2010 did not, in general, retain and distribute the original datasets used by the authors. Therefore, my

strategy for obtaining suitable data sets is to match datasets recorded in Neotoma and the EPD with studies described by the authors of Carrión et al., 2010.

While reviewing the work of Carrión et al., 2010 I first extracted information about the bioclimatic belt (Thermo-, Meso-, and Supra-Mediterranean) that each of the sites reviewed are situated in. I also noted if these sites had evidence of prehistoric humans in their palaeoecological record and, if so, which period. I then used a Python program to link this information to data from the paper's supplementary materials to provide geographical coordinates and references for original authors in addition to life zone and date of human onset for each site. Associating sites with references was a crucial step in ensuring sites found in the EPD and Neotoma really did correspond with sites discussed in Carrión et al., 2010.

Because of possible variations in the rendering of site names that might exist between the review article and the databases (especially in light of accents associated with the Spanish language), I used geographical coordinates as the principle method of cross referencing sites. I first queried Neotoma for sites within a $4\,\mathrm{km}^2$ bounding box around those studied in Carrión et al., 2010. This procedure identified 5 sites in Neotoma. Performing the same search on the EPD yielded the same sites as were found in Neotoma, plus an additional 4 sites. From these I selected two of the sites that I ultimately selected as study sites—San Rafael and Navarrés (see Table 3.1) . Of the other 7 sites found in the EPD and Neotoma, some covered a temporal range of less than 50% of the Holocene, and the others didn't satisfy my objectives specified in Section 3.1.3.

To obtain additional sites, I broadened my search by querying both Neotoma and the EPD to obtain a list of all sites within a geographical bounding box encompassing Iberia. The life zone for each site was calculated using its latitude and altitude as stated in the source database. I also retrieved the earliest and latest radiocarbon dates associated with the data from each site, in order to determine which datasets covered a sufficient proportion of the Holocene. From this search I obtained a long-list of 59 sites. Of these 9 corresponded to the studies referenced in Carrión et al., 2010, and 50 were not considered in that review but are included in the EPD or Neotoma.

3.1.5 Final selection of study sites from long-list

To make a final selection of study sites from the list of candidates identified in the EPD and Neotoma I:

1. Plotted all 59 potential study sites on a map of Iberia to establish an intuition for their distribution in space.
2. Preferentially selected sites whose database entries indicated their data covered a large proportion of the Holocene (and excluding completely sites whose data covers less than 50% of the Holocene).

In Table 3.1 I summarise the final selection of study sites, including pertinent data relating the the onset of human activity (if known based on the summary analysis in Carrión et al., 2010), the temporal range of the associated data, and references to the original studies that produced the data. Fig. 3.2 shows a map of Iberia indicating the locations of the 59 study sites considered, with the six sites referenced in the remainder of this thesis highlighted. Of the selected sites, four (Algendar, Atxuri, Algendar, and San Rafael) are within $20\,\mathrm{km}$ of the coast. Consequently, readers should note that there is a bias towards coastal regions in the study sites analysed in this thesis. I produced Fig. 3.2 by applying the bioclimatic life zone model described above (see Section 3.1.2) to SRTM30 Digital Elevation Model (DEM) data (Farr et al., 2007) to associate each pixel in the DEM with a life zone category using a Python program. I then plotted the resulting image as a map using QGIS software. The locations of study sites were extracted from the EPD.

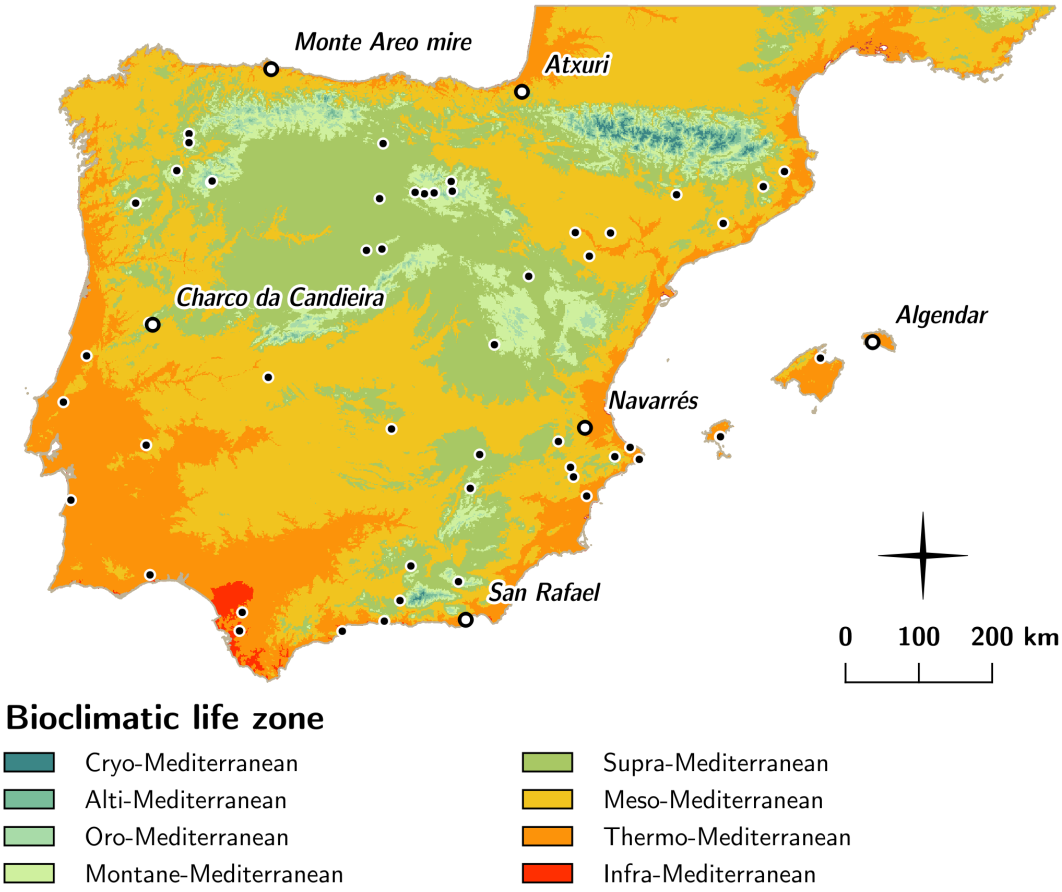


Figure 3.2: Map of locations of study sites within Iberia. Open circles indicate sites that were selected for subsequent analysis. Closed circles indicate sites that were considered but not selected.

	Elevation [m]	Life zone	Hypothesis	Prediction in terms of model attributes	Citation
Charco da Candieira	1,400	Montane	First anthropogenic impact on vegetation occurred 6500 yr BP.	Expect models in which date of arrival of first agriculturalist, t_0^a, is ca. 6500 yr BP will be deemed more likely in light of palynological data than otherwise.	van der Knaap and van Leeuwen, 1995
			Regime change from climate dominated vegetation trends to human domination occurred in the latter part of the Holocene (after 5500 yr BP).	Models in which anthropogenic disturbance is excluded will perform worse in explaining pollen sequences after 5500 yr BP.	van der Knaap and van Leeuwen, 1995
			Humans had a modest impact on forest composition through small-scale deforestation and grazing until 4500 yr BP, after which over-grazing occurred.	Model runs in which incidence of pasturalism increases after 4500 yr BP will be better predictors of pollen trends than those in which incidence of pasturalism does not increase around this time.	van der Knaap and van Leeuwen, 1995
			Regime change from small-scale to large-scale deforestation ocurred 3200 yr BP associated with the development of a ‘cultural landscape’.	Model’s ability to explain pollen trends will decrease significantly after 3200 yr BP due to lack of representation of cultural landscapes.	van der Knaap and van Leeuwen, 1995

Atxuri	500	Meso	First human settlement occurred around 5000 yr BP.	Expect models in which date of arrival of first agriculturalist, t_0^a, is ca. 5000 yr BP will be deemed more likely in light of palynological data than otherwise.	Penalba, 1994
			The absence of *Fagus* (Beech) in north-west Spain is explained by its east-west expansion being interrupted by anthropogenic disturbance.	Models in which anthropogenic disturbance is excluded will tend to predict greater prevalence of fire intolerant species (such as Beech) than those in which human impact is considered.	Penalba, 1994
Monte Areo mire	200	Meso	Human pastoralism was established in the area by 7300 yr BP.	Expect models in which date of arrival of first pastoralist, t_0^p, is around 7300 yrs BP will be deemed more likely in light of empirical data than otherwise.	López-Merino et al., 2010
			Cereal pollen dated to 6700 yrs BP indicates agriculture was established in the area by this time.	Expect models in which date of arrival of first agriculturalist, t_0^a, is before 6700 yrs BP will be deemed more likely in light of palynological data than otherwise.	López-Merino et al., 2010

			Climate and anthropogenic drivers interacted together to produce observed trends.	Models in which both anthropogenic and climatic changes occur simultaneously will better predict observed palynological data than those models in which either occur individually.	López-Merino et al., 2010
Navarrés	225	Thermo	Agricultural practices emerged in the region ca. 7000 yr BP, and intense agricultural activity in the vicinity of Navarrés at least as early as 5000 yr BP.	Expect models in which date of arrival of first agriculturalist, t_0^a, is between 5000 yr BP and 7000 yr BP will be deemed more likely in light of palynological data than otherwise.	Carrión and Dupré, 1996
			Human disturbance led to abrupt change from Pinus to Quercus dominance ca. 5930 yr BP.	Trend of switching from Pinus to Quercus more likely to emerge in models including anthropogenic disturbance than those which don't.	Carrión and Dupré, 1996
			Pinus to Quercus replacement driven primarily by increase in wildfire frequency after 7000 yr BP.	Pinus to Quercus transition could be explained by increasing fire frequency parameter, p_F, after 7000 yr BP in model runs independently of anthropogenic activity.	Carrión and Van Geel, 1999

Algendar	21	Thermo	First human settlement occurred between 5000 and 4000 yr BP.	Expect models in which date of arrival of first agriculturalist, t_0^a, is between 4000 and 5000 yr BP will be deemed more likely in light of palynological data than otherwise.	Yll et al., 1997
			Abrupt change from predominantly mesophilous communities to olea dominated maquis (open landscape) between 5000 and 4000 yr BP was climatically triggered and then exascerbated by human settlement.	Models in which anthropogenic disturbance is excluded will perform worse in explaining pollen sequences after 4000 yr BP.	Yll et al., 1997
San Rafael	0	Infra	Rapid deforestation and spread of steppe communities around 5000 - 4500 yr BP occurred in response to emergence of the semi-arid conditions which persist in the present.	Deforestation trend beginning ca. 4500 yr BP can be explained by models which represent climatic changes, but no anthropogenic influence.	Pantaleon-Cano et al., 2003

	Otherwise expected trends associated with anthropogenic disturbance in the area are not visible in the San Rafael sequence because they were overshadowed by the effects of aridification due to climatic change.	Including anthropogenic influence into models in addition to climatic driving should have a negligible affect on vegetation trend.	Pantaleon-Cano et al., 2003

Table 3.1: Study sites and key characteristics considered in their selection. Also shown are summaries of hypotheses which have been proposed to explain observed trends in pollen sequences – in most cases invoking anthropogenic influence – along with corresponding predictions in terms of model attributes.

3.2 Empirical pollen abundance reference data

This section explains how I obtained pollen abundance data for each of the study sites selected in Section 3.1.4. I describe the steps undertaken to process the data into a form that can be readily compared to the outputs of simulation outputs. In Chapter 6, each simulation run will generate data that allows us to understand how the proportion of the simulated landscape occupied by different types of land cover changes over time. The reference data described this section links how the land cover at each of the study sites changed over time according to physical evidence on one hand, to the outputs of simulation runs on the other. This reference data plays a central role in grounding the model in reality. By manipulating the model and seeing how different settings make simulation outputs more or less like the reference data we can 'tune' the model parameters to allow its outputs to most closely reproduce empirical data. The code used to perform the analyses described in this section can be found in the `pollen-abundance` directory of the agrosuccess-data repository (see Appendix G.S4).

It is necessary to use pollen abundance data as an empirical proxy of the land cover of the landscapes simulated by AgroSuccess because the target study period for all study sites considered is the mid-Holocene (see Section 2.5). It is not appropriate to use contemporary land cover data from remote sensing (e.g. the CORINE dataset from the Copernicus Land Monitoring Service) because mid-Holocene land cover is expected to be significantly different to contemporary land cover due to the impacts of agriculture and industrialisation that occurred in the intervening millennia. Similarly, I do not expect to be able to meaningfully compare AgroSuccess's outputs to contemporary land cover data sets.

3.2.1 Obtain and clean species-level pollen abundance data

Pollen abundance data was extracted from the European Pollen Database (EPD) (Fyfe et al., 2009) (see Section 3.1.1). The only file format in which the EPD is distributed that does not require proprietary software to read is a dump from a Postgres database. To access this data, users need to:

1. Set up a Postgres database on their system
2. Consult the EPD's published documentation to learn how the data they need is organised in

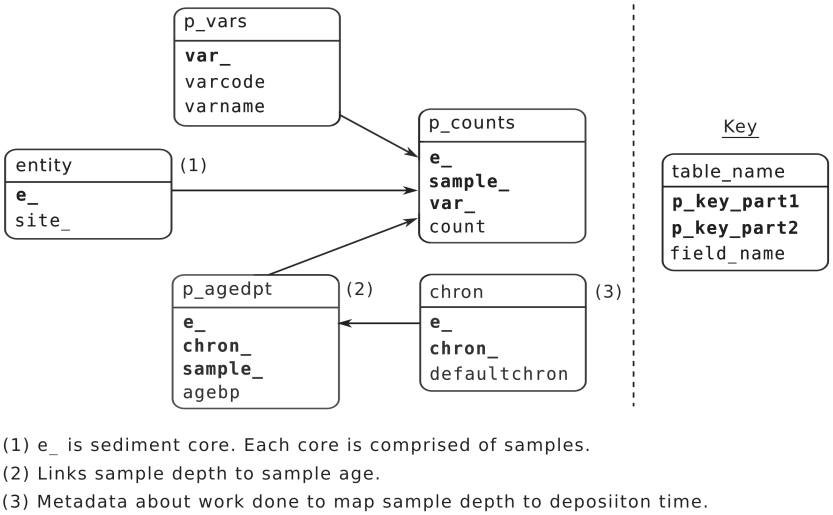


Figure 3.3: Example of a schema diagram used during the construction of a SQL query for extracting data from the European Pollen Database.

the database

3. Construct a SQL query to relate the database tables containing the required data in the appropriate way, possibly making use of a schema diagram such as that shown in Fig. 3.3.

Even for simple use cases, such as extracting pollen abundance time series for a list of study sites, users are required to know how to administer a database (and have the access rights on their system to do so), as well as how to write SQL. This limits the accessibility of the data, in the sense of the FAIR guiding principles for data management (Wilkinson et al., 2016), to people with specific technical skills.

These challenges created an obstacle to making the data processing steps described in this section transparent and reproducible. In response I developed an MIT licensed application called `epd-query` (Lane, 2019) that abstracts away the complexity of administering database software and specifying SQL queries on the EPD. `epd-query` allows users to retrieve pollen abundance time series as `.csv` files by simply providing the application with a copy of the EPD Postgres database dump, and specifying the ID numbers of study sites for which to extract data. The configuration file used to extract the data for the study sites selected in Section 3.1 is shown in Listing 3.1.

```
# config.yml
queries:
  - site_location_info
  - site_pollen_abundance_ts

sites:
  - name: Charco da Candieira
    epd_number: 762
    settlement_period: Neolithic
  - name: San Rafael
    epd_number: 486
  - name: Atxuri
    epd_number : 76
    settlement_period: Neolithic
  - name: Monte Areo mire
    epd_number: 1252
  - name: Navarres
    epd_number: 396
  - name: Algendar
    epd_number: 55
    settlement_period: Bronze age
```

Listing 3.1: Configuration file for used to extract site location data and pollen abundance time series for the study sites selected in Section 3.1 from the European Pollen Database using the `epd-query` application.

Having extracted `.csv` files containing pollen abundance time series for each study site, I carried out data cleansing to identify potential problems with the data and ensure comparability between study sites. The EPD contains data for three cores from Navarrés, but two of these contained 30 or fewer samples so these were excluded them from further analysis. I noted that the cores for Navarrés, Monte Areo mire, and Charco da Candieira contained a significant amount of 'pollen spike' or 'Lycopodium spike'. This is related to a method whereby a known quantity of exotic pollen is added to a sample, enabling workers to calculate the absolute number of pollen grains deposited per unit area (pollen concentration), in addition to being able to compare the relative abundance of different species (Bonny, 1972). Reviewing the literature associated with the cores from Navarrés (Carrión & Dupré, 1996), Monte Areo mire (López-Merino et al., 2010), and Charco da Candieira (van der Knaap & van Leeuwen, 1995), I confirmed that Lycopodium spike was indeed added to these samples. As my objective is to report the *proportion* of landscape occupied by different species, my analysis does not depend on the ability to calculate absolute

pollen concentration. Furthermore, a later data processing step (see Section 3.2.2) will depend on being able to discern whether at least 90% of pollen for each core has been related to a particular Plant Functional Type (PFT) corresponding to a land cover type in simulation run outputs. As the presence of added pollen spike would otherwise complicate this calculation, I excluded all pollen spike from the pollen abundance time series. I similarly exclude pollen originating, moss and fungal spores, and any pollen whose species could not be identified.

Pollen from aquatic species was also excluded from the pollen abundance time series used in subsequent analyses. It is possible that changes in the presence of aquatic species could be signals of processes that are caused by, or drive, anthropogenic land use change. For example, agriculturalists diverting water away from the tributaries of a lake to irrigate crops could cause the lake to dry out. Alternatively, a nearby lake drying out might cause agriculturalists who had previously settled nearby to move somewhere else with an available water source. However, processes that would cause a lake to dry out are not represented in the simulation model described in Chapter 4. Additionally, the date at which humans start practicing agriculture at each study site is treated as a boundary condition of the model (see Section 3.1.4), rather than an emergent phenomenon that responds to the availability of water sources (for example). I therefore focus on changes in the abundance of terrestrial species in this analysis.

Pollen diagrams for all study sites using raw data extracted from the EPD are shown in figures listed in Table 3.2.

Table 3.2: Listing of figures containing pollen diagrams for all study sites.

Study site	Pollen diagram
Charco da Candieira	Fig. 3.4
Atxuri	Fig. 3.5
Monte Areo mire	Fig. 3.6
Navarrés	Fig. 3.7
Algendar	Fig. 3.8
San Rafael	Fig. 3.9

3.2.2 Aggregate species-level data to categorical land-cover types

Having excluded data relating to pollen not expected to contribute to the land cover types represented in the AgroSuccess simulation model, I aggregated species-level pollen counts to the amounts contributed by each land cover type (LCT). In particular I map pollen species to the following LCTs:

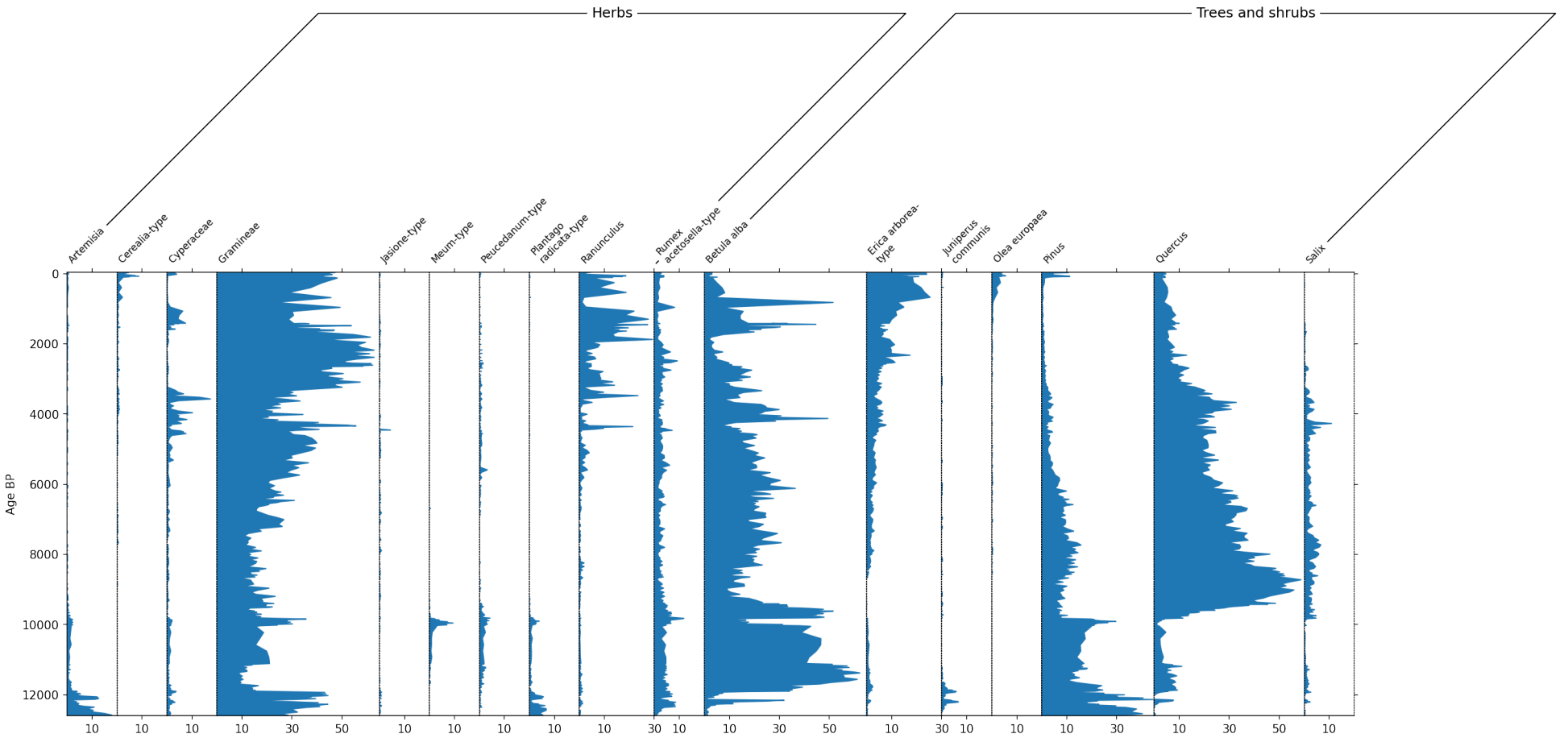


Figure 3.4: Pollen diagram for the Charco da Candieira study site using data from the European Pollen Database (EPD) (Fyfe et al., 2009). Time series are shown for all taxa included in the reported pollen counts that contributed at least $4\,\%$ to the total pollen count at the site. Species groupings correspond to those listed for the relevant taxa in the EPD.

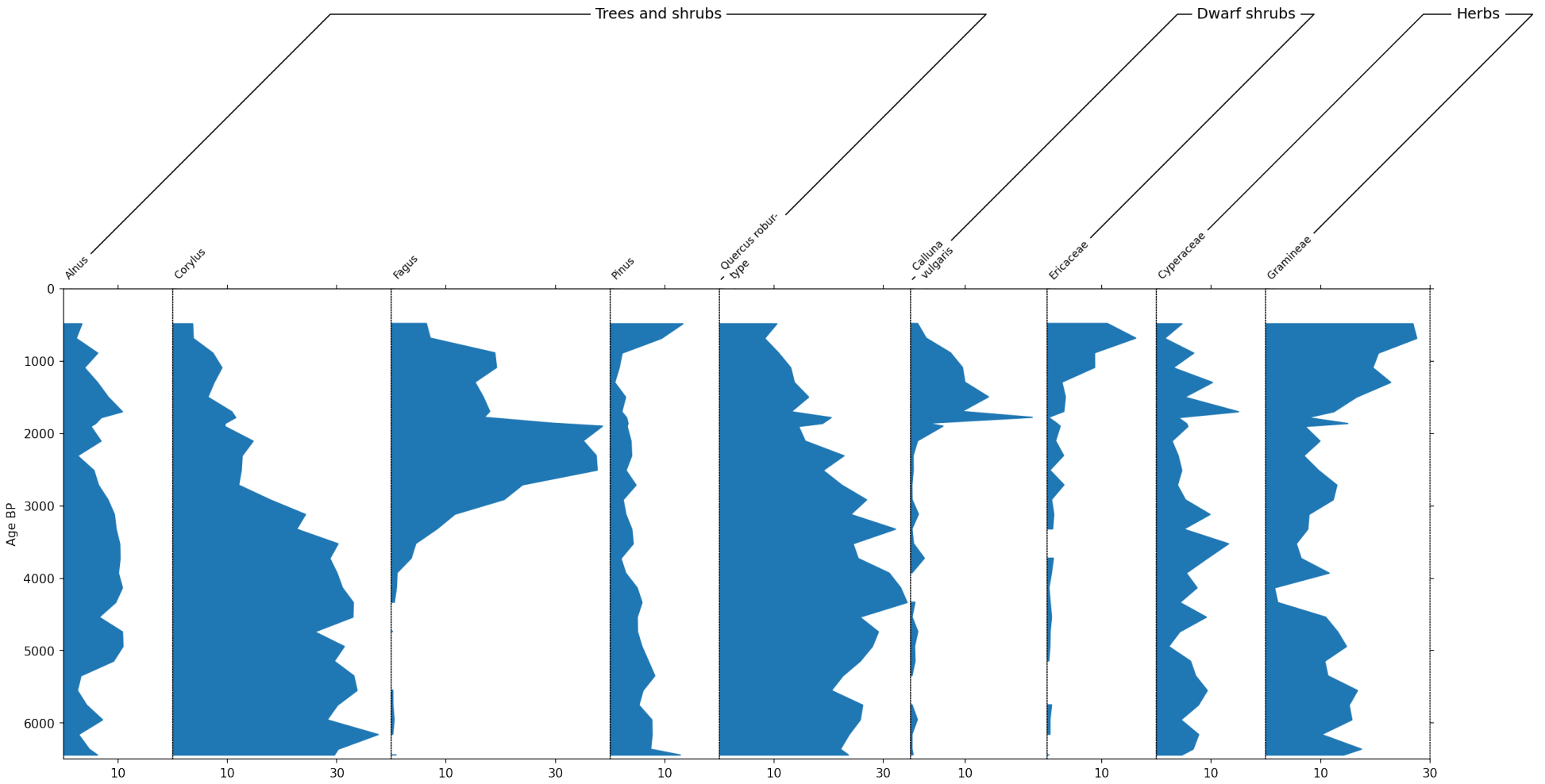


Figure 3.5: Pollen diagram for the Atxuri study site using data from the European Pollen Database (EPD) (Fyfe et al., 2009). Time series are shown for all taxa included in the reported pollen counts that contributed at least $4\,\%$ to the total pollen count at the site. Species groupings correspond to those listed for the relevant taxa in the EPD.

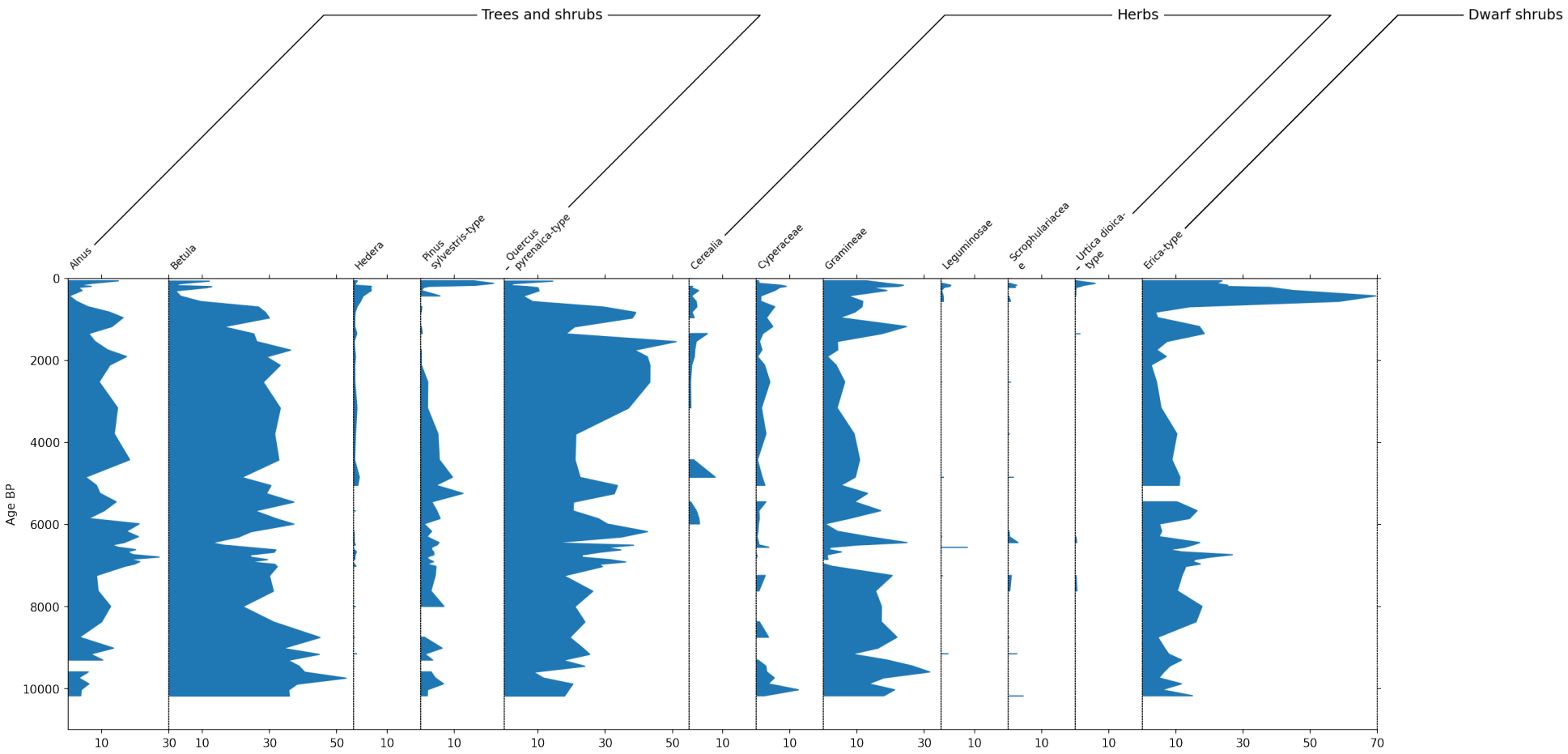


Figure 3.6: Pollen diagram for the Monte Areo mire study site using data from the European Pollen Database (EPD) (Fyfe et al., 2009). Time series are shown for all taxa included in the reported pollen counts that contributed at least 4% to the total pollen count at the site. Species groupings correspond to those listed for the relevant taxa in the EPD.

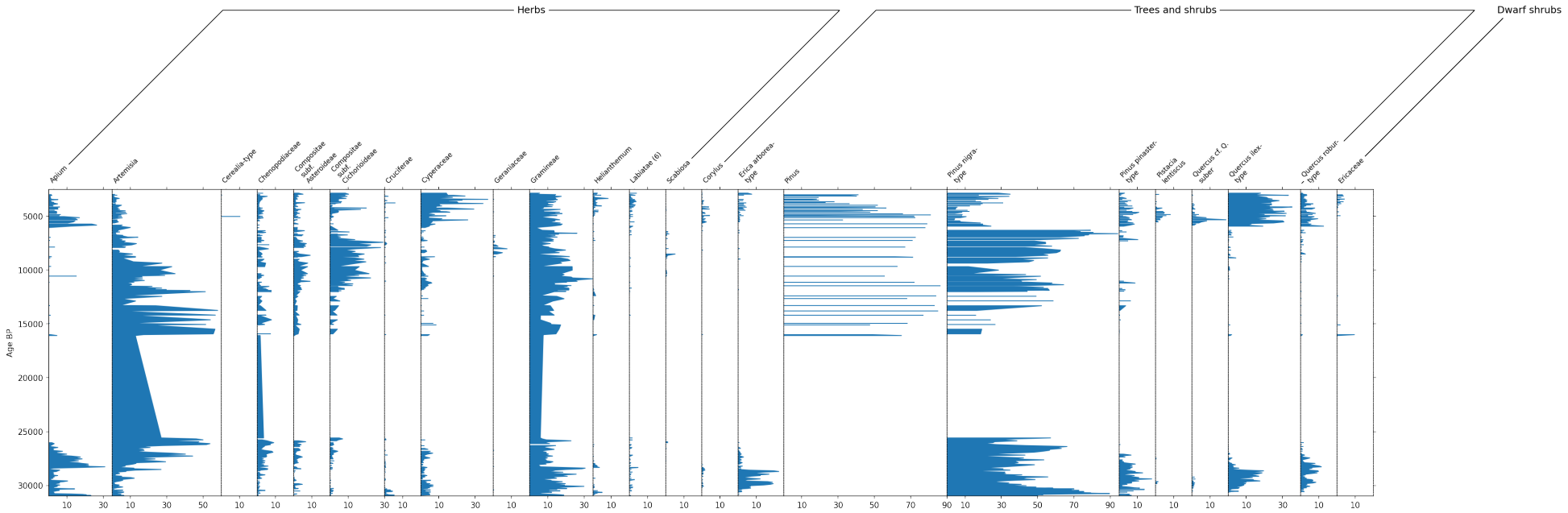


Figure 3.7: Pollen diagram for the Navarrés study site using data from the European Pollen Database (EPD) (Fyfe et al., 2009). Time series are shown for all taxa included in the reported pollen counts that contributed at least $4\,\%$ to the total pollen count at the site. Species groupings correspond to those listed for the relevant taxa in the EPD.

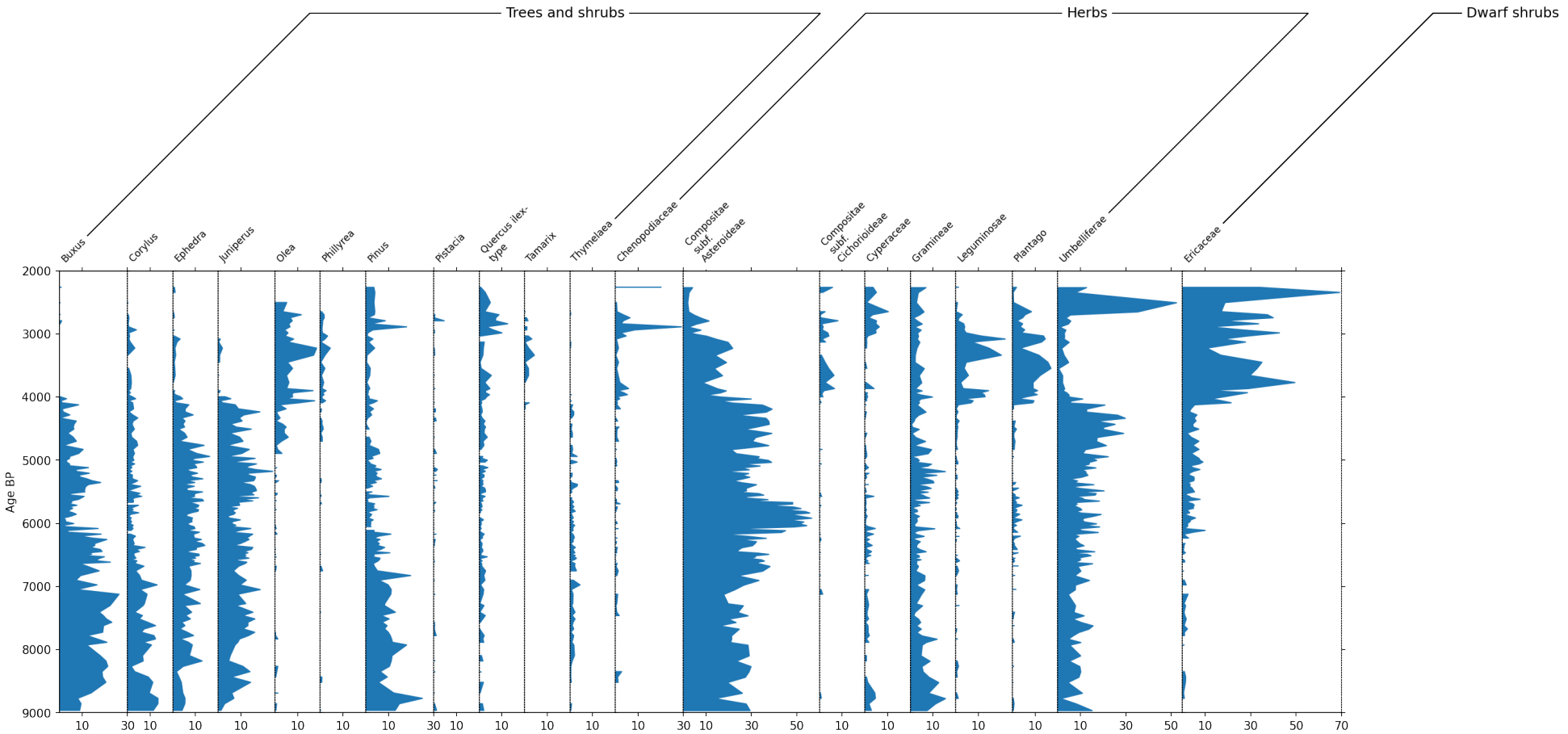


Figure 3.8: Pollen diagram for the Algendar study site using data from the European Pollen Database (EPD) (Fyfe et al., 2009). Time series are shown for all taxa included in the reported pollen counts that contributed at least 4% to the total pollen count at the site. Species groupings correspond to those listed for the relevant taxa in the EPD.

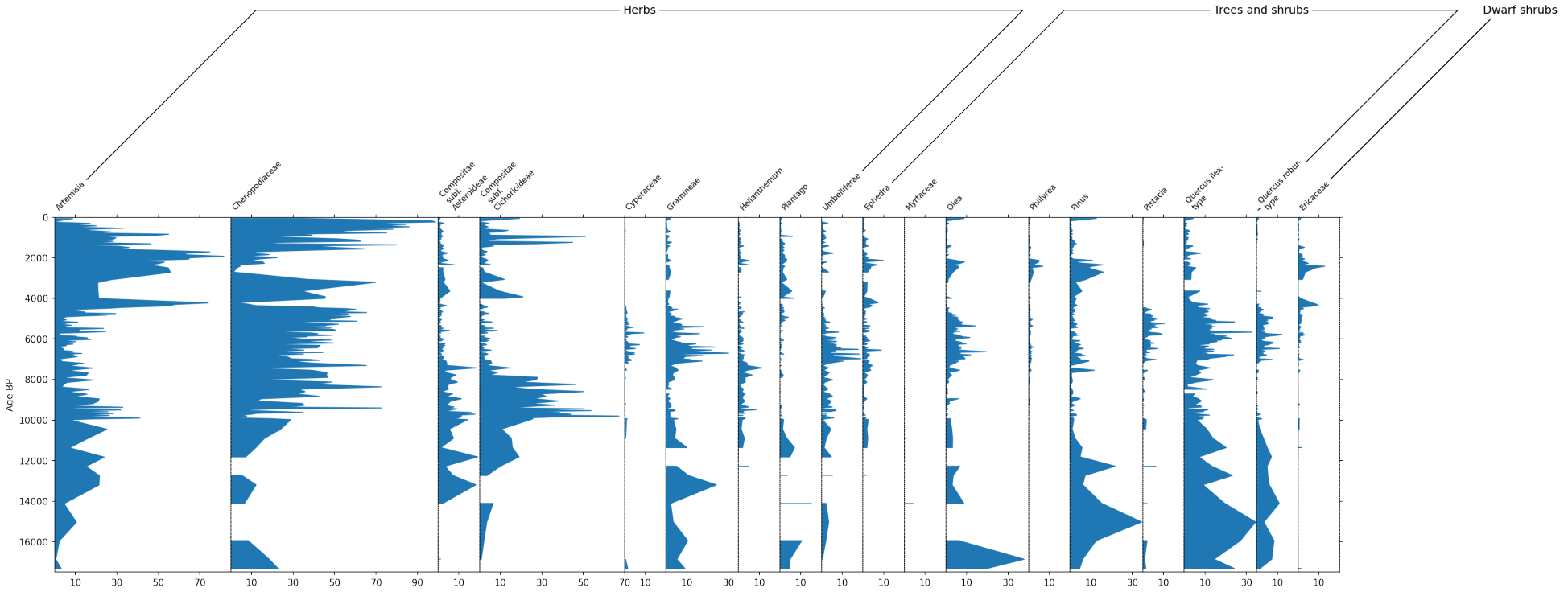


Figure 3.9: Pollen diagram for the San Rafael study site using data from the European Pollen Database (EPD) (Fyfe et al., 2009). Time series are shown for all taxa included in the reported pollen counts that contributed at least 4% to the total pollen count at the site. Species groupings correspond to those listed for the relevant taxa in the EPD.

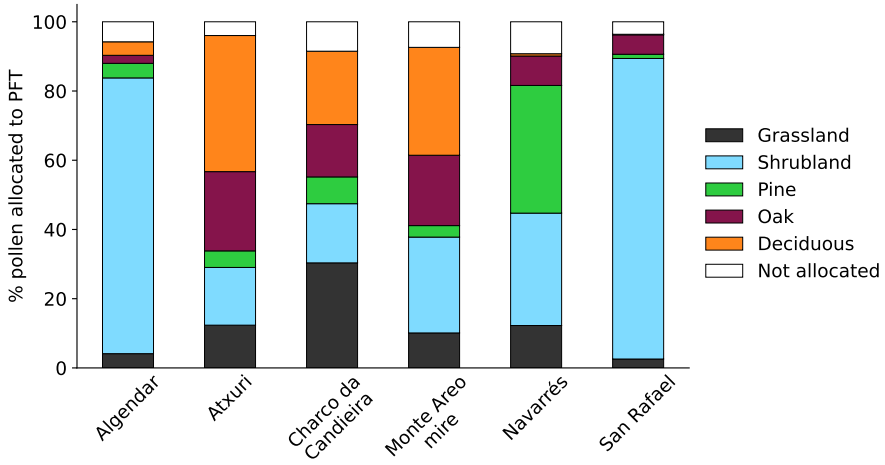


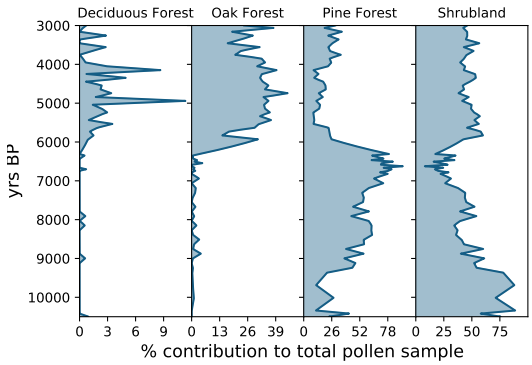
Figure 3.10: Proportion of the pollen allocated to plant functional types for the duration of each study site's pollen record. We see that for all sites the amount of pollen which has not been allocated to any plant functional type is less than 10%.

- Grassland
- Shrubland
- Deciduous forest
- Pine forest
- Oak forest

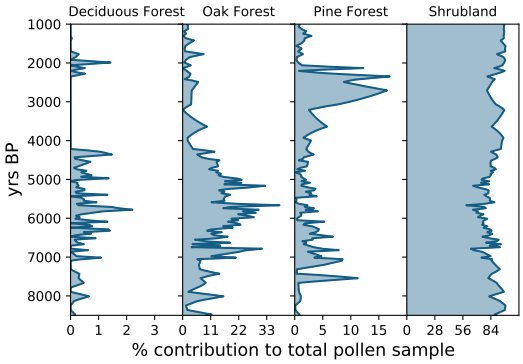
The EPD includes a `varname` field (e.g. in the `p_vars` table) containing taxonomic names of the species which produced the pollen identified in each sample. The rank specificity of the names given to pollen in each sample ranges from species to family. To map from species to LCTs, I constructed a set of regular expressions that match taxonomic names, and associated these search patterns with LCTs. See Table A.1 in Appendix A for the expression/ LCT pairings used, as well as the notebook `make_lct_timeseries.ipynb` for the implementation. In Fig. 3.10 I show the proportion of the total pollen for each study site that was allocated to each of the LCTs. Note that at least 90% of the pollen identified in the sediment cores for each of the study sites was successfully allocated to one of the LCTs.

3.2.3 Temporal interpolation

The AgroSuccess simulation model has an annual timestep, so the pollen abundance reference data from the study sites should have a value associated with each year between the earliest and



(a) **Navarrés**: Notice the approx. 200 year period oscillation in percentages of shrubland and pine forest 6400–6800 yrs BP, followed by sudden and sustained increase in oak forest after 6400 yrs BP.



(b) **San Rafael**: We see large variation in shrubland and oak forest around the time it is thought agriculture reached Iberia (6500 yrs BP).

Figure 3.11: Pollen diagrams for the Navarrés and San Rafael study sites derived from the pollen abundance time series described in this section.

latest dates of the chronology. The dates of the samples extracted from the EPD are determined by radiocarbon dating, and do not correspond one-to-one, or even linearly with time steps. To achieve an annual time step in the reference data I performed linear interpolation on the time series, subject to the constraints that: i. no land cover type ever makes up a negative percentage of land cover, and ii. total contributions from all four land cover types must total 100%. See Fig. 3.11 for plots of these interpolated pollen abundance time series for the Navarrés and San Rafael study sites.

3.2.4 Pollen abundance and landscape reconstruction

Note that the approach taken in this section relies on the use of pollen abundance as a proxy for the proportion of the landscape that was occupied by vegetation of a particular type. As noted in Section 2.5.1, the Fagerlind effect is likely to limit the correspondence between pollen abundance and land cover proportion. There is ongoing work in the literature to develop quantitative methods that correct for variation in pollen productivity and dispersal between species, most notably the Landscape Reconstruction Algorithm (LRA) (Sugita, 2007), and the more recent MARCO POLO tool that was developed with the aim of being simpler to use than the LRA (Mrotzek et al., 2017). The application of these algorithms is beyond the scope of this thesis, but should be considered an area of investigation for future work (see Section 8.3.1). We should be mindful of the biases that are likely to exist in the pollen abundance data as a result of the Fagerlind effect when comparing it to simulated land cover proportion outputs.

3.3 Model input data

Whereas the previous section described the steps needed to produce the pollen abundance reference data against which I will compare simulation outputs to evaluate model performance, in this section I describe the steps taken to produce the data that will serve as model *inputs*. The reason the production of reference data was described before the production of input data is that the reference data will be used below in Section 3.3.3 to establish initial conditions for land cover types.

This section introduces some ideas from geographic data analysis that might not be familiar to a wider audience. In particular I frequently refer to **raster** data. This is one of the two broad categories of geographic data, the other being vector data. Vector data describes geometric objects—points, lines, and polygons—that exist in geographic space. In the context of this thesis, the geographical coordinates at which pollen-containing sediment cores were extracted are examples of points whose vectors are specified by their latitude and longitude coordinates. Raster data, by contrast, are used to model how quantities vary across space. The geographical area of interest is represented as a 2-dimensional grid of pixels representing the quantity of interest. The value of each pixel encodes the value of some quantity that the raster models as being representatives of all points falling within that pixel (O'Sullivan & Unwin, 2010, pp. 188–191).

Another aspect of geographical data analysis that is relevant in this section is the idea of a map projection, or coordinate reference system (CRS). The geographic coordinates specified by latitude and longitude are convenient for locating a point on the globe. However, because the units of geographic coordinates are degrees of arc, their use complicates the calculation of distances and slope between points. I therefore re-project the data from geographic coordinates into a projected coordinate reference system whose unit of measurement is a measure of distance, such as metres.

I have used the GeoTiff file format to store raster data that will be used in AgroSuccess simulation models, as this format is widely supported by GIS software and libraries. It also allows users to store information about the map projection, the geographic transform (i.e. where the grid is located with respect to the origin of the CRS), and the size of each grid cell information in the same file as the raster grid.

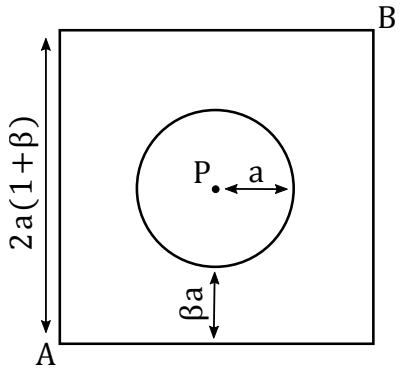


Figure 3.12: Geometric construction of the problem of finding the bounding box around study site locations needed to provide the required raster layers for my simulations. Point P is the location from which the sediment core used to derive pollen time series for the site was extracted according to the EPD. a is the radius of the circle from which it is assumed pollen has contributed to the sediment core – the *experimental zone*. β is a buffer parameter which controls the area around the experimental zone which will also be included in the simulation to help account for edge effects. Points A and B are, respectively, the points of minimum and maximum latitude and longitude defining the bounding box around the study site.

3.3.1 Digital Elevation Model and derived layers

To determine the size of the raster layers to produce for inclusion in AgroSuccess simulation runs, we first consider the geometry of the region around the point where each study site's sediment core was extracted. If P is the location at which the sediment core was extracted, we will consider the circle with area $A = 30\,\mathrm{km}^2$ centred on P as the *experimental zone* within which land cover proportions will be calculated from simulation models. This circle has radius $a = \sqrt{30/\pi}\,\mathrm{km} \simeq 3.1\,\mathrm{km}$. Let β be a parameter that controls the size of a buffer around the experimental zone, such that a buffer of size βa will be added to the experimental zone. The inclusion of this buffer region means it will be possible for seeds and fires to enter the experimental zone from outside during simulation runs, reducing the impact of edge effects on the quantities measured within the experimental zone. By setting $\beta = 1$ we introduce a buffer which is the same size as the radius of the experimental zone. This implies that the input raster layers for simulations should be squares with edge length $4a = 12.4\,\mathrm{km}^2$ (see diagram in Fig. 3.12).

Digital Elevation Models (DEMs) are raster data sets in which each cell encodes the elevation above sea level of the area of land represented each pixel. Within simulation runs, the DEM will be used by the fire spread model to influence the relative probability of a fire spreading uphill compared to downhill. DEMs for each study site will also be used as the basis from which to calculate other raster layers (see below). The specific DEM dataset I use is called SRTM30 (Farr et al., 2007), which contains 30 m^2 pixel size elevation data for the entire Earth. See code in the notebook `download_site_elevation_data.ipynb` in the `dem-derived` directory in

Appendix G.S4 for the procedure used to download the SRTM30 data, convert it from geographic coordinates to a projected coordinate system suitable for study sites on the Iberian Peninsula—Madrid 1870 (Madrid) / Spain (EPSG:2062)—and crop it to the geometry with respect to the sediment core extraction points described in the previous section.

A final step taken to prepare DEMs before subsequent data processing and consumption in simulation runs is to remove 'sinks'. These are raster cells that have a lower elevation than any of their neighbours. It is common practice in hydrological analysis to treat such pixels as spurious data artefacts, and automated procedures to remove (or 'fill') them are routinely applied (Sharma & Tiwari, 2019). I used the TauDEM application to remove sinks from the downloaded DEMs. Heat maps of the downloaded elevation datasets for each of the study sites are included in Fig. 3.13.

Three additional raster layers based on the DEM for each study site were produced:

- **Slope**—used by agricultural agents to preferentially select flat land patches to convert to crop land cover types.
- **Flow direction**—used to drive soil moisture calculations which depend on the direction in which water flows over the landscape. Soil moisture then influences ecological succession in turn.
- **Binary aspect**—encodes whether a patch of land is on a northerly or southerly facing slope. This enters directly into ecological succession logic to reflect the fact that southerly facing slopes receive more sunlight than northern slopes.

All of these raster layers were generated using an MIT licensed software tool I developed called `demproc` (see Appendix G.S5). This provides an API on top of the TauDEM and gdal applications to generate the DEM-derived raster layers described above.

3.3.2 Soil type

The type of soil at each point in the landscape can influence the degree to which water flowing over the landscape is absorbed by the soil as opposed to forming surface runoff. This difference is represented in AgroSuccess by considering local soil type along with flow direction in the soil moisture calculations that are performed in each time step (see Section 4.2.5). Following Ferrér

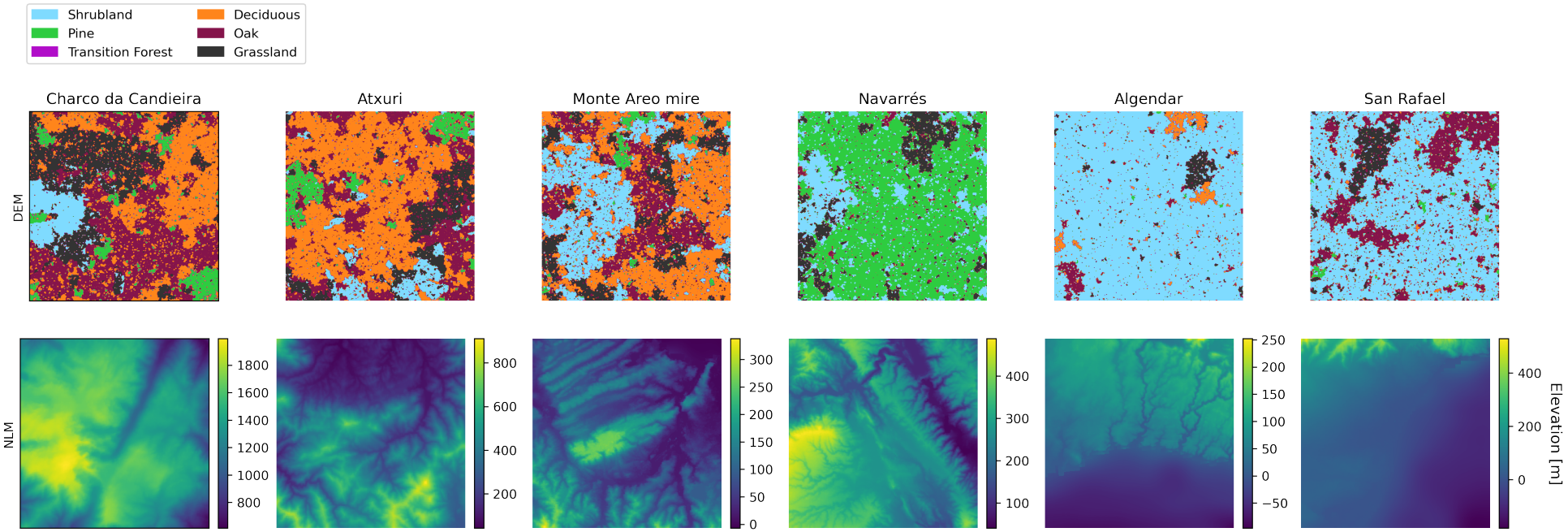


Figure 3.13: Example Neutral landscape models (NLMs, top) and digital elevation models (DEMs, bottom) for each of the study sites considered in this report. Elevation shown for the DEMs indicates metres above sea level.

et al., 1995; J. D. A. Millington et al., 2009 I represent four discrete classes of soil type labelled A to D. These are characterised by the amount of water they are able to absorb, with soil type A able to hold the most water, and soil type D the least. Equivalently soil type A produces the lowest runoff, and type D the highest. In the version of AgroSuccess presented in this thesis, I assume that all landscapes have spatially uniform soil type A. The capacity to explore sensitivity to heterogeneous study sites is included in AgroSuccess, and could be explored in future versions (see Section 8.3.4).

3.3.3 Initial spatial distribution of land cover

As AgroSuccess simulation runs represent how land cover evolves over time, it is necessary to provide each simulation with a raster map representing the distribution of land cover at the beginning of the simulation as a boundary condition. The pollen abundance reference data whose production are described in Section 3.2.2 allow us to estimate, for a given year during the time series duration, the proportion of the landscape occupied by shrubland, pine forest, deciduous forest, and oak forest. However, these reference data do not provide any information about the *spatial* distribution of these land cover types.

As a solution to the need to generate an initial spatial configuration of land cover, I used the modified random cluster (MRC) algorithm (Saura & Martínez-Millan, 2000) to produce spatially random neutral landscape models. These neutral landscape models (NLMs) are constrained such that i. the shapes of the clusters resemble an areal view of vegetation cover, and ii. the proportion of the generated landscape occupied by different classes correspond to the proportions of land cover type pollen present in the reference data. I used the Use NLMpy (Etherington et al., 2015) implementation of the MRC algorithm for our analysis. See the `landcover-nlms` directory in the `agrosuccess-data` repository (Appendix G.S4) for the implementation. Note that because of the inclusion of a buffer zone beyond the experimental zone in the raster layers, there is an implicit assumption that the area whose vegetation proportions can be estimated from the pollen abundance reference data is representative of the larger area around it.

The implementation of the MRC algorithm provided in NLMpy enables us to generate an NLM in which the proportions of different land cover types are constrained across the whole landscape. However, it might be useful to be able to specify that all the land cover above the treeline is shrubland, and then ensure that the lowland land cover was such that the overall land cover

proportions matched the reference data. Here I derive some expressions to enable the implementation of such an approach.

Define N^{tot} as the total number of cells in the model. We can separate these cells into components contributed by each of the land-cover classes represented in the model such that

$$N^{\text{tot}} = N^{\text{tot}}(\mathbf{C}) = \sum_{c} N_c(\mathbf{C}) \tag{3.1}$$

where we define $N_c(\mathbf{C})$ as the number of cells in class $c \in \{\text{oak forest}, \text{shrubland} \dots\}$. The land-cover class matrix $\mathbf{C}$ has elements c_{ij} encoding the land cover class of the cell in position (i, j) : $i \in (1, \dots, N_y)$, $j \in (1, \dots, N_x)$. In symbols,

$$N_c(\mathbf{C}) = \sum_{i=1}^{N_y} \sum_{j=1}^{N_x} \delta_{c_{ij},c} \tag{3.2}$$

The total number of cells in class c, $N_c(\mathbf{C})$, can be further divided into $N_c^{\text{hi}}(\mathbf{C}, \mathbf{E}; \epsilon)$ and $N_c^{\text{lo}}(\mathbf{C}, \mathbf{E}; \epsilon)$ – the number of cells in class c which are above and below the treeline respectively. The matrix $\mathbf{E}$ has elements e_{ij} which encode the elevation of the DEM cell in position $(i, j) : i \in (1, \dots, N_y)$, $j \in (1, \dots, N_x)$. The parameter ϵ is the elevation of the treeline and is study site dependent.

$$N_c(\mathbf{C}) = N_c^{\text{hi}}(\mathbf{C}, \mathbf{E}; \epsilon) + N_c^{\text{lo}}(\mathbf{C}, \mathbf{E}; \epsilon) \tag{3.3}$$

$$N_c^{\text{hi}}(\mathbf{C}, \mathbf{E}; \epsilon) = \sum_{i=1}^{N_y} \sum_{j=1}^{N_x} \delta_{c_{ij},c}\, \Theta(e_{ij} - \epsilon) \tag{3.4}$$

$$N_c^{\text{lo}}(\mathbf{C}, \mathbf{E}; \epsilon) = \sum_{i=1}^{N_y} \sum_{j=1}^{N_x} \delta_{c_{ij},c}\, [1 - \Theta(e_{ij} - \epsilon)] \tag{3.5}$$

In the above, $\Theta(n)$ is the Heaviside step function defined such that

$$\Theta(n) = \begin{cases} 0, & n < 0 \\ \\ 1, & n \geq 0 \end{cases} . \tag{3.6}$$

Since our data is expressed in terms of *proportions* of land-cover occupied by each class, we define $\rho_c = N_c / N^{\text{tot}}$ (total proportion of land-cover occupied by class c), and $\rho_c^{\text{hi}} = N_c^{\text{hi}} / N^{\text{hi}}$ and

$\rho_c^{\text{lo}} = N_c^{\text{lo}}/N^{\text{lo}}$ (proportions of cells in class c above and below the treeline respectively). Here $N^{\text{hi}} = \sum_c N_c^{\text{hi}}$ and $N^{\text{lo}} = \sum_c N_c^{\text{lo}}$. Note $\sum_c \rho_c^{\text{hi}} = \sum_c \rho_c^{\text{lo}} = 1$. We have

$$\rho_c^{\text{hi}}(\mathbf{C}, \mathbf{E}; \epsilon) = \frac{1}{N^{\text{hi}}} \sum_{i=1}^{N_y} \sum_{j=1}^{N_x} \delta_{c_{ij},c}\, \Theta(e_{ij} - \epsilon) \tag{3.7}$$

$$\rho_c^{\text{lo}}(\mathbf{C}, \mathbf{E}; \epsilon) = \frac{1}{N^{\text{lo}}} \sum_{i=1}^{N_y} \sum_{j=1}^{N_x} \delta_{c_{ij},c}\, [1 - \Theta(e_{ij} - \epsilon)] \tag{3.8}$$

We know the value of ρ_c for each c from our data. It will be useful to be able to specify, as part of our model, the relatively simple proportion of each land-cover type occupying the area above the treeline (e.g. 100% shrubland), and calculate the lowland proportions which preserve our target global land cover proportions, ρ_c. We can derive an equation to do this based on the quantities defined above:

$$N_c = N_c^{\text{hi}} + N_c^{\text{lo}} \tag{3.9}$$

$$N^{\text{tot}} \rho_c = N^{\text{hi}} \rho_c^{\text{hi}} + N^{\text{lo}} \rho_c^{\text{lo}} \tag{3.10}$$

$$\implies \rho_c^{\text{lo}}(\mathbf{E}, \rho_c, \rho_c^{\text{hi}}; \epsilon) = \frac{N^{\text{tot}}(\mathbf{E})\, \rho_c - N^{\text{hi}}(\mathbf{E}; \epsilon)\, \rho_c^{\text{hi}}}{N^{\text{lo}}(\mathbf{E}; \epsilon)} \tag{3.11}$$

Also note

$$N^{\text{hi}}(\mathbf{E}; \epsilon) = \sum_{i=1}^{N_y} \sum_{j=1}^{N_x} \Theta(e_{ij} - \epsilon) \tag{3.12}$$

$$N^{\text{lo}}(\mathbf{E}; \epsilon) = \sum_{i=1}^{N_y} \sum_{j=1}^{N_x} [1 - \Theta(e_{ij} - \epsilon)] \tag{3.13}$$

Example NLMs generated using the MRC algorithm for all study sites are shown in Fig. 3.13.

3.3.4 Climate data

The preceding subsections describe how I obtained DEM, derived slope, flow direction, aspect, soil type, and NLM data sets. These are all raster data sets that, among the collection obtained for each study site, all share the same spatial extent and pixel size. In addition to this raster data, the simulation model described in Chapter 4 also requires climatic information to reflect differences in temperature and rainfall between the study sites. The amount of rainfall in each

simulated year will drive calculations characterising how soil moisture varies throughout the landscape. Both temperature and precipitation will influence the expected number of fires that occur in a given year, as well as the likelihood that fires spread from cell-to-cell.

To obtain estimates of temperature and precipitation at my selected study sites during the mid-Holocene I rely on the outputs of Global Climate Models (GCMs). Specifically, I used total annual precipitation and mean annual temperature values produced by the BCC-CSM1-1 GCM distributed as part of the WorldClim 1.4 downscaled paleo climate dataset (Hijmans et al., 2005). This dataset comprises 'climate surfaces' (i.e. time series of raster grids describing spatial variation of climatic variables) derived from paleo-simulations provided by the CMIP5 project (Harrison et al., 2015). The BCC-CSM1-1 GCM was selected because it was readily available as part of the WorldClim 1.4 distribution and has was evaluated as 'Satisfactory' (the highest ranking) in its ability to estimate temperature and precipitation in Europe by (McSweeney et al., 2015) in an analysis of all available CMIP5 models. Note that the WorldClim 1.4 dataset does not include mean wind speed or direction data. For this reason I used contemporary data from weather stations closest to my selected study sites (see Section 3.3.5).

To prepare the WorldClim 1.4 BCC-CSM1-1 data for use in my simulation model, I downloaded the corresponding 30 arc s resolution 'bioclimatic variables' file from the WorldClim website, which contains GeoTiff files covering the globe for various bioclimatic variables. I then used the script `total_precip_and_temp.py` (see the `climate` directory in Appendix G.S4) to extract values from the annual precipitation and mean annual temperature GeoTiff files for the raster cells corresponding to my study sites.

As a first approximation during simulation runs I will consider annual precipitation and mean annual temperature to remain fixed throughout simulation runs. A possible future extension could be to incorporate time series of mean annual precipitation and mean annual temperature from Holocene climate models. While climate models are able to produce time series at monthly (or higher) resolution, the simulation model described in Chapter 4 has an annual time step, so accommodating higher resolution climate data would require revisions to the model structure. Incorporation of temporally varying temperature and precipitation data would lead to greater realism in the model, as it would allow simulated subsistence agriculturalists to react to droughts and floods, as well as influencing the wildfire regime.

3.3.5 Wind speed and direction

The final piece of data required as input into the AgroSuccess model is wind speed—specifically the probability of observing low, medium, or high wind speed on a given day—and wind direction data. These categorical wind speed classes correspond to values on the Beaufort scale as shown in Table 3.3.

Table 3.3: Beaufort numbers corresponding to different wind speed classes represented in simulation models.

Wind speed class	Beaufort No. range
Low	0–2
Medium	3–5
High	≥ 6

This wind speed and direction data is used to drive the fire spread sub-model during simulation runs, reflecting the fact that fire is more likely to spread in the same direction as the wind is blowing, and that higher wind speed increases the risk of fire spread in general.

To obtain estimates of wind speed and direction probabilities for the study sites I have assumed that these probabilities in the present day are comparable to those in the mid-Holocene. This enables me to collect daily wind speed and direction observations from contemporary weather stations close to the selected study sites, and use these to calculate the probability of observing different wind speeds and directions. The main source of this data is the Spanish State Meteorological Agency (AEMET) which provides a REST API to access daily weather observations.

To support the acquisition of daily wind data I wrote a Python package called `aemet_api` (available in the aemet-wind repository, see Appendix G.S2). This enables users to identify weather stations managed by the Spanish Met office (AEMET) that are close to specific target locations, and download daily wind speed and direction observations for those stations via the AEMET REST API. I used this package to download wind speed and direction data collected between 1st January 1990 and 1st November 2019 for weather stations within 50 km of all of the selected study sites that are within Spain. See Fig. 3.14.A for the location of these weather stations with respect to the study sites they represent.

I was unable to locate a source of daily wind data for the Portuguese study site, Charco da Candieira. To work around this limitation I sought AEMET weather stations within Spain whose locations make them suitable to represent Charco da Candieira. I reasoned that the most im-

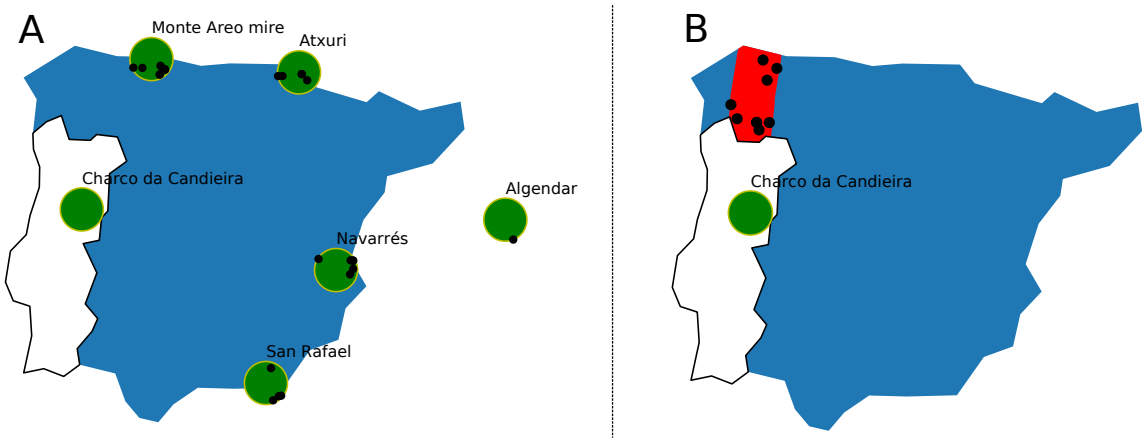


Figure 3.14: (A) Location of weather stations (black dots) selected to represent study sites located within Spain (green circles). (B): Black dots show the AEMET weather stations selected to represent the Charco da Candieira which is located within Portugal, and for which data from nearby weather stations were unavailable. The red region shows the region in Spain with the same distance to the Atlantic coast as Charco da Candieira.

portant factor influencing wind speed and direction at Charco da Candieira is likely to be its proximity to the Atlantic coast. I identified the region in Spain which is within a 50 km buffer of the line that is the same distance from the Atlantic coast as Charco da Candieira (this region is shown in red in Fig. 3.14.B). I also obtained a single reported average wind speed for the Portuguese weather station closest to Charco da Candieira (Porto) from the UN data portal. Finally I compared the average wind speed reported between 1st January 1990 and 1st November 2019 at each of the weather stations in the Spanish Atlantic coast region to the average wind speed at the Porto station, and chose the station with the smallest difference to represent Charco da Candieira.

Having identified AEMET weather stations to represent each study site, I used `aemet_api` to download wind speed and direction data for all sites on all dates for which it was available between 1st January 1990 and 1st November 2019. The numbers of daily observations for each study site across all representative weather stations and dates ranged from 10,577 at Algendar to 50,097 at Navarrés. Finally I calculated wind speed (see Fig. 3.15) and direction (see Fig. 3.16) probabilities for each study site.

In this analysis I have implicitly assumed that wind speed and direction are independent. A development of this approach might consider the joint distribution of wind speed and direction. Simulation models could then sample from this joint distribution to determine wind speed and direction on a given day.

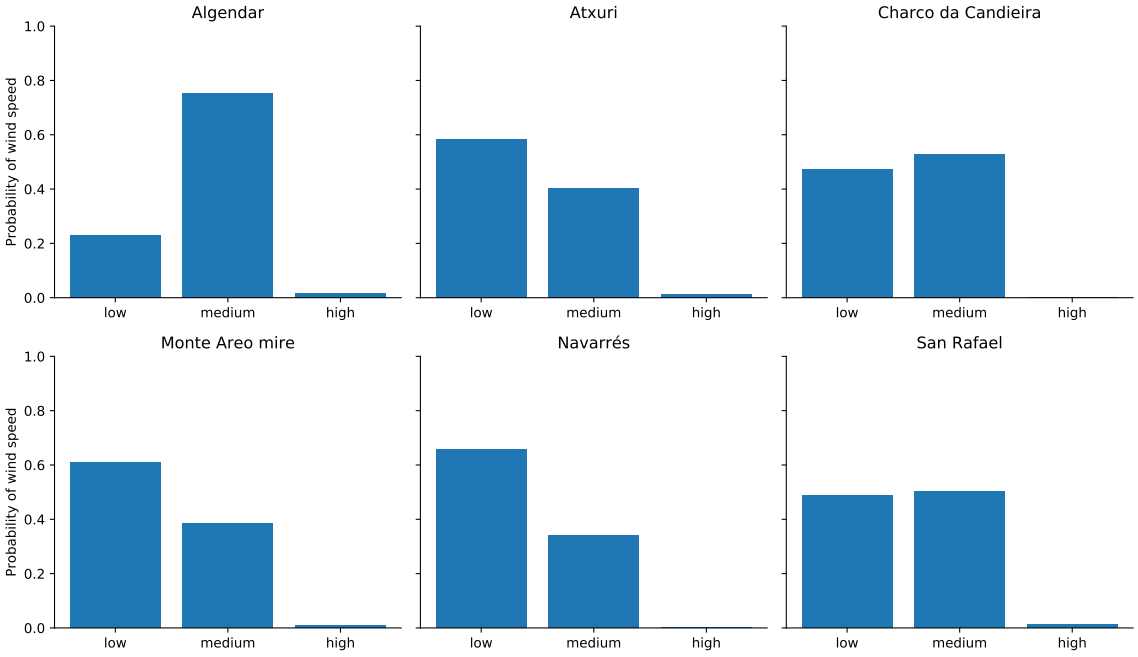


Figure 3.15: Probability of observing low, medium, or high wind speed (as defined by the Beaufort scale) at each of the study sites. These probabilities are calculated from daily observations taken between 1st January 1990 and 1st November 1991 (where available) collected from AEMET weather stations within 50 km of the study sites.

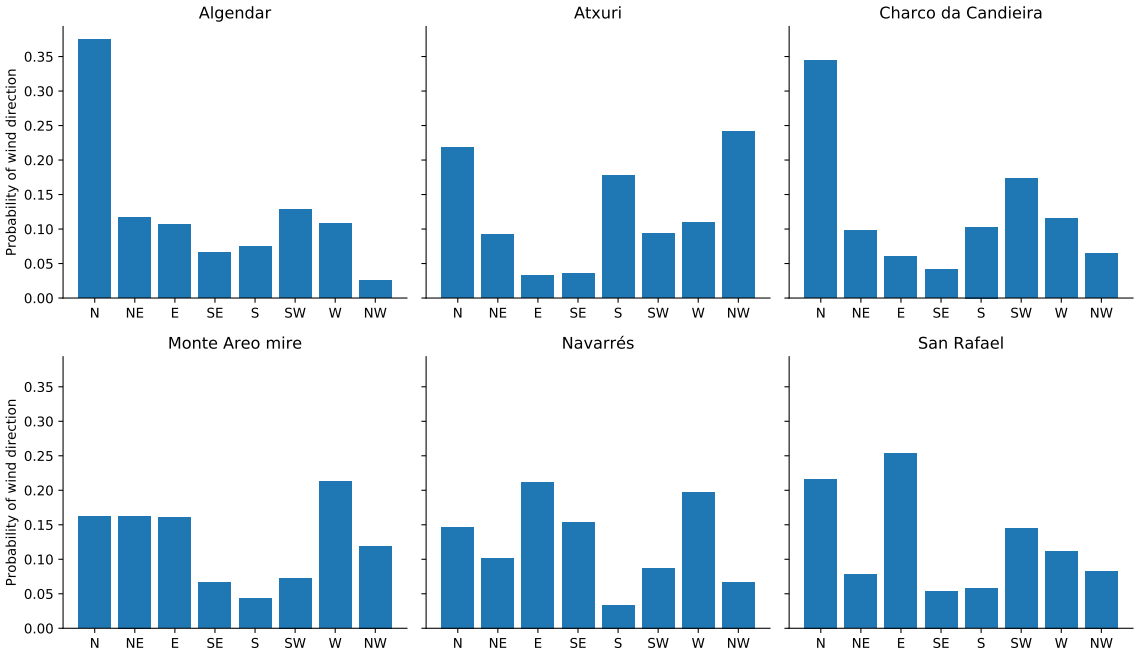


Figure 3.16: Probability of the wind blowing in each cardinal direction at each of the study sites. These probabilities are calculated from daily observations taken between 1st January 1990 and 1st November 1991 (where available) collected from AEMET weather stations within 50 km of the study sites.

Chapter 4

AgroSuccess simulation model specification

In this chapter I first provide an overview of the ecological and anthropological processes that are represented in AgroSuccess, before describing in detail the various rules that were used to implement these processes as a computer simulation model. I calibrate the AgroSuccess model in Chapter 5, and use the calibrated model to perform experimental simulation runs to evaluate archaeologically motivated hypotheses explaining pollen abundance change discussed in the literature in Chapter 6.

4.1 Model overview

To contextualise the AgroSuccess model is useful to consider the spatio-temporal grain and extent required to meet my objectives. Having established my justification for the spatio-temporal scale of the model, I present a table containing an overview of the model's structure. Finally, I present the various submodels that together constitute the AgroSuccess agent-based simulation model.

AgroSuccess has been designed to represent landscape areas on the order of $100\,\mathrm{km}^2$. This ensures that the spatial extent of simulated scenarios is large enough to capture the dynamics of the disturbance processes that AgroSuccess is intended to be used to investigate (wildfire and anthropogenic land-use change). An upper limit of $100\,\mathrm{km}^2$ has been used by other models used to simulate landscape-scale land-cover change in response to disturbance events, including

LandClim (Colombaroli et al., 2010) and LANDIS (Schumacher et al., 2004). The outputs of the version of AgroSuccess presented in this thesis are compared to pollen abundance time-series. In future I may develop this work to instead compare the simulated outputs to empirical land-cover proportion estimates derived from landscape reconstruction algorithms such as the LRA (Sugita, 2007) or MARCO POLO (Mrotzek et al., 2017) (see discussion in Section 3.2.4). Previous authors have argued that these techniques are suitable for reconstructing landscapes from pollen abundance for areas up to $100\,\mathrm{km}^2$ surrounding the location that sedimentary pollen cores were extracted from (Sugita, 2007). Therefore, the outputs of AgroSuccess simulations with extents of up to $100\,\mathrm{km}^2$ would also be comparable to land-cover proportion estimates produced by these techniques if they become available for study sites of interest in the future.

Since I use AgroSuccess to explain trends in pollen diagrams in terms of both ecological factors and anthropogenic decision making around land-cover change, the spatial and temporal grain of the model are influenced by my approach to modelling these processes. The ecological submodel aims to represent the changing character of vegetation cover in the landscape as a consequence of interacting succession and disturbance processes (i.e. wildfire). The process of a wildfire converting areas of the landscape previously occupied by vegetation to burnt land takes hours or days to occur. Meanwhile, the vegetative colonisation and growth processes that lead to ecological succession take years or decades to become apparent. The anthropogenic submodel in AgroSuccess aims to represent the interaction between anthropogenic actions to modify the landscape to support subsistence agriculture and ecological succession processes. I assume that subsistence agriculturalists would have determined the area of land to modify on an annual basis, taking into account the number of members of their household and knowledge of whether or not the area of land farmed in the preceding year was sufficient to meet the household's needs. They would then increase or decrease the planned area to farm in the next year accordingly.

These modelling considerations motivate the use of an annual time step in AgroSuccess. Although each wildfire take less than a year to affect land-cover, the response of the slower coupled vegetative growth process takes years to materialise. Since there are no other processes with a sub-annual effect represented in the model, it is sufficient to aggregate the effects of all wildfires in a given year into an annual time step. This approach to aggregating relatively fast disturbance processes is similar to that found in other ecological succession models at annual (J. D. A. Millington et al., 2009; Perry, Wilmshurst, McGlone, McWethy et al., 2012) and even decadal (Henne et al., 2013; Schumacher et al., 2004) temporal resolution. Additionally, an annual time

step is directly compatible with the annual anthropogenic decision-making process represented in the model. There is precedence for use of an annual time step in similar models of subsistence agriculture in the literature (Ullah, 2013).

Because AgroSuccess is intended to be run for temporal extents of hundreds of years, it is possible that the agricultural practices and decision-making criteria employed within a community or household would change during the time period represented by an individual simulation run. For example, there could be a change in the conservativeness with which households evaluate their subsistence needs, causing them to work harder by cultivating more land to reduce risk of food deficiency. This type of change in anthropogenic decision-making *strategy* could be modelled as taking place over generational timescales (e.g. evaluated every 40 simulated years). Investigation of this type of scenario is out of scope for this thesis, but would be possible with modest changes to the AgroSuccess model code.

4.2 Environmental submodel

4.2.1 Rule Based Community Level modelling

The approach to modelling ecological succession used in AgroSuccess is known as rule-based community-level modelling (RBCLM) (McIntosh, 2003; J. D. A. Millington et al., 2009). This is an approach to succession modelling designed to address situations in which modellers have qualitative understanding of the land-cover transitions one would expect to occur under specific conditions in a modelling scenario. RBCLM avoids the need for quantitative data to parameterise models of inter-species competition such as that presented by M. A. Zavala and Zea, 2004.

In RBCLM, knowledge about the ecological system in question is divided into ‘declarative’ and ‘procedural’ knowledge (McIntosh, 2003). Declarative knowledge about the system is expressed as a set of discrete states and a corresponding set of possible transitions between those states. Each transition specifies the start and end states of the transition (the transitions are directed), the combination of environmental conditions that cause the transition to occur, and the time required for the transition to conclude (McIntosh, 2003). The declarative knowledge component of an RBCLM is an example of a state-and-transition model (see Section 7.2 and McIntosh et al., 2003; J. D. A. Millington et al., 2009). Procedural knowledge about the system modelled in an

RBCLM specifies how the state of a particular location changes over time consistently with the transitions that are permitted in the model.

RBCLM as a modelling paradigm is not inherently spatial (McIntosh et al., 2003), however, in AgroSuccess I follow J. D. A. Millington et al., 2009 in using a spatial variant of RBCLM in which each of the cells in the simulation grid representing geographic space are characterised as being in one of a finite set of land-cover states. For example, a landscape might be represented as a mosaic of *oak forest*, *pine forest*, and *shrubland* cells. Each cell is also associated with a set of variables representing the cell's environmental state. The RBCLM interacts with this data structure to produce landscape evolution dynamics. The direction of state transition for a particular cell can be influenced by its own environmental conditions, as well as by the state of the cells that surround it. For example, a patch of shrubland might develop into pine forest under dry conditions provided there is a source of pine seeds nearby. Alternatively, the same cell might develop into an oak forest cell under hydric conditions, provided there is a nearby source of acorns.

4.2.2 Land-cover states and environmental conditions

As an RBCLM, the core of AgroSuccess is a set of discrete states that characterise the land-cover in each patch of land represented in the model, and a set of *environmental conditions* that govern the circumstances under which transitions between the states occur. Environmental conditions comprise anything about the state of the land patch that is relevant to the model and that isn't specifically related to its land-cover. For example, two land patches might be characterised as pine forest, but one of the patches contains acorns while the other doesn't. Knowledge of the presence of acorns is relevant to understand the likely future succession pathways of the two patches but isn't part of the land-cover type of the patch. The presence or absence of acorns is therefore an example of an environmental condition. In the following sub-sections I specify the states and environmental conditions that are included in AgroSuccess.

Land-cover states

While the environmental conditions represented in the model are mainly relevant to the ecological succession submodel (see Section 4.2.3) the land-cover types are integral to all the sub-

models that together comprise AgroSuccess. The flammability of each land-cover type influences the likelihood of a fire spreading into a patch of that type in the wildfire submodel (see Section 4.2.6), and the fertility and ease of conversion to agriculture of each of the land-cover types influence how agricultural household agents interact with land patches in the agent-based model of subsistence agriculture submodel (see Section 4.3.4). Additionally, the model outputs that contribute to my research questions are directly related to the land-cover types. At the end of each simulation timestep, the model software counts the number of grid cells in each land-cover type and reports the proportion of the simulated landscape in each state. These outputs are collated across multiple stochastically varying model runs and compared to the empirical pollen abundance time series described in Section 3.2 to evaluate different model parametrisations' ability to reproduce the observed patterns.

The land-cover types represented in AgroSuccess are:

- **Water/ Quarry**: area with no land-cover, included to support the representation of areas that cannot be colonised by vegetation that causes land-cover type to change over time.
- **Burnt**: area that has been recently subject to fire. Land-cover represented by this type may or may not include vegetable matter that is able to resprout, depending on the fire frequency experienced by the patch (see Section 4.2.3, Eq. (S5)).
- **Grassland**: grasses in the *Poaceae* (or *Gramineae*) family, including *Cerealia*-type.
- **Wheat**: area used for wheat agriculture (*Cerealia*-type *Poaceae*).
- **DAL**: Depleted Agricultural Land (see below).
- **Shrubland**: area dominated by shrubby plants including both resprouting species (e.g. *Quercus coccifera* and *Pistacia lentiscus*) and species that regenerate from seed (e.g. *Ulex parviflorus*, *Rosmarinus officinalis* and *Cistus* species) (Baeza et al., 2007).
- **Pine**: area dominated by *Pinus* forest.
- **Transition forest**: area occupied by a mixture of *Pinus*, *Quercus*, and *Juniperus* species. This state represents land in the transitional stage between being predominantly *Pinus* forest and being predominantly *Quercus* forest.
- **Deciduous**: area occupied by mature individuals of a mixture of deciduous species including *Castanea sativa*, deciduous oak (e.g. *Quercus pyrenaica*), *Alnus* species, and *Populus*

species.

- **Oak**: area dominated by evergreen oak forest, e.g. *Quercus ilex*.

Table 4.1 specifies the fertility, wood value, and land-cover conversion cost attributes that are associated with each land-cover type. These are discussed more detail in the remainder of this section. The land-cover types in AgroSuccess are similar to those used by J. D. A. Millington et al., 2009, but also include 'Wheat', 'Depleted Agricultural Land' (DAL), and 'Grassland'. The DAL land-cover type represents land whose soil's nutrient and organic matter content has been reduced by agricultural use, reflecting the lack of access to artificial fertiliser in pre-industrial times. The Wheat and DAL land-cover types allow AgroSuccess to represent the effects of anthropogenic agricultural activity on the landscape, and to explore the impact this has on the processes represented by the ecological succession and wildfire submodels.

While developing an STM it is necessary to limit the set of states and associated transition rules to make the resulting model tractable. First, AgroSuccess includes four land-cover types that represent land-cover that is predominantly occupied by mature Pine, Transition forest, Deciduous, and Oak. These are 'forest types'. Our classification here assumes that individuals in the associated land patches have reached maturity and are dominated by the respective type. This is a simplification of the real-world situation where there will be other species (both arboreal and non-arboreal) present, but in the model their presence is implicit as they are not explicitly represented in the names of the states. This is an explicit modelling simplification associated with using a state-and-transition model. This is similar to the approach taken by Henne et al., 2013 in the use of discrete categories of land-cover type 'Evergreen shrubs', 'Pinus type', and 'Other deciduous' in their simulation study incorporating paleoecological pollen analysis data. Second, I intend to inclusively represent *all* evergreen oak species in the 'Oak' land-cover type. While I acknowledge that there are evergreen oak species other than *Quercus ilex* in the Iberian landscapes under consideration, these are usually in shrub form, and therefore don't meet the maturity criterion to justify their own 'forest type' (see previous point). Additionally, the use of coarse-grained 'Oak' and 'Pine' land-cover types is sufficient to investigate the oak/ pine competition dynamic that has been of interest to others in the literature for some time (Barbero et al., 1990; M. A. Zavala & Zea, 2004). This is an important issue, and therefore valuable for AgroSuccess to be able to explore. Third, AgroSuccess includes only a single Shrubland land-cover type. There is evidence that grassland land-cover can transition into qualitatively distinct shrubland

land-cover types subject to variation in soil type. For example, Baeza et al., 2007 found that grassland transitioned to *Quercus coccifera* shrubland on limestone soil, or to *Ulex* shrubland and then onto *Rosmarius* or Mixed *Rosmarius* shrubland on marl soil. However, I do not treat soil type as an environmental gradient in AgroSuccess. Therefore, I was unable to distinguish the dynamics that would lead to different shrubland types, and so collapse all shrubland types into a single categorical land-cover type.

Table 4.1: Land-cover types represented in AgroSuccess. Every land patch in an AgroSuccess simulation is in one of these states. *Fertility* scores influence the productivity of land patches of each type. *Wood value* scores influence the utility of each land-cover type as a source of fire wood. *Conversion cost* indicates the relative difficulty of using a patch of a given type for wheat agriculture.

Land-cover type, L_c	Fertility, F_{L_c}	Wood value, W_{L_c}	Conversion cost, C_{L_c}
Water/ Quarry	0	-1	-1
Burnt	5	1	1
Grassland	4	-1	1
Wheat	4	-1	1
Depleted Agricultural Land	1	-1	5
Shrubland	2	-1	2
Pine	2	5	3
Transition Forest	3	4	3
Deciduous	3	4	3
Oak	3	3	4

I have used a *semi-qualitative* approach to characterising soil fertility, wood value and land-cover conversion cost for each land-cover type. Land-cover types are assigned dimensionless scores on a scale of 0–5 as opposed to objectively measurable quantities with physical units. For comparison, I could have attempted to characterise soil fertility in terms of the percentage of organic matter by mass, or $\mathrm{mg/kg}$ of nutrients such as phosphorus and potassium (Khresat et al., 2008). However, within the context of AgroSuccess the use of objective measurements for land-cover fertility and conversion cost is unjustified in comparison to the inherently subjective nature of the modelled anthropogenic decision-making processes that they inform. When an individual or collective makes a decision, they aim to maximise their benefits and minimise their costs. They do this using heuristics, rather than by exhaustively analysing all available data (Parker et al., 2003). The use of semi-qualitative scores for fertility, wood value and land-cover conversion cost reflects the decision to model household agents in AgroSuccess as using inductive reasoning based on past experience to evaluate the costs and benefits of using different patches of land for agriculture.

My overall approach to allocating fertility, F_{L_c}, wood value, W_{L_c}, and land-cover conversion cost, C_{L_c}, scores for land-cover types was the following procedure:

1. Sort the land cover types into ascending order with respect to each quantity
2. Decide whether the quantity should vary continuously between land-cover types, or whether there should be some clustering
3. Assign values based on understanding derived from the literature (see below).

The semi-qualitative approach described here is intended to be flexible enough to support modifications following discussion with subject matter experts. In the remainder of this sub-section I justify my choices of fertility and land-cover cost scores specified in Table 4.1, using relevant literature where available. However, in recognition of the uncertainty inherent in the model, this is an aspect of the model that might be focused on in future for development with other experts in the field.

In AgroSuccess, the fertility of a land-cover patch of type L_c is specified by F_{L_c}. This is a dimensionless score on a scale of 0–5. A fertility score of $F_{L_\mathrm{c}} = 0$ indicates that land-cover type L_c is not fertile, and a score of $F_{L_\mathrm{c}} = 5$ indicates the patch is maximally fertile. When a land plot is farmed its fertility decreases as a consequence of the disruption of the natural interactions between vegetation and soil, which in turn decreases the availability of nutrients to crops (Khresat et al., 2008; Ullah, 2013).

The rate at which the fertility of a patch of land decreases during the course of its use for agriculture in Mediterranean-type ecosystems has been estimated to be 0.5 to $1.0\,\%/\mathrm{yr}$ (Ullah, 2013, p. 92). I integrate this information into the discrete representation of land-cover states in AgroSuccess by imposing a rule that after T_F years of use for wheat agriculture, a patch of land will transition to a land-cover state called 'Depleted Agricultural Land' (DAL). Here T_F is a model parameter whose default value I choose to be $T_F = 50\,\mathrm{yr}$, representing the time taken for soil fertility to be reduced by approximately $40\,\%$ at a rate of $1.0\,\%/\mathrm{yr}$. The assumption that soils are depleted of nutrients after $50\,\mathrm{yr}$ of cropping is consistent with Ullah, 2013, p. 130. Note that a fertility depletion of $1\,\%/\mathrm{yr}$ produces a geometric rate of change in fertility such that the fertility of a land patch after t years, F_t, is

$$F_t = (1 - 0.01)^t F_0$$

where F_0 is the initial fertility of the land patch. After a patch has become DAL, succession pro-

cesses will cause it to transition into shrubland, and from there into each of the other possible land-cover types represented in the model subject to local environmental conditions. Each of the discrete land-cover states has a fertility value associated with it such that land-cover states representing land patches occupied by later successional species are associated with larger fertility values. This is to represent the process by which organic matter and nutrients are returned to the soil by the natural vegetation occupying the patch during the time it is left fallow.

The fact that soil fertility decreases when it is used for agriculture doesn't necessarily mean that the households don't make any effort to improve soil fertility, rather that the *net* effect of agriculture is to decrease soil fertility over time. This is consistent with MedLand, in which model scenarios are characterised in terms of whether the represented households are 'good' or 'greedy', differentiated by how intensively they work the land. 'Good' agriculturalists produce fertility depletion at a rate of $1\,\%/\mathrm{yr}$, wheres 'greedy' agriculturalists produce fertility depletion at a rate of $2\,\%/\mathrm{yr}$. The assumption that households don't improve soil fertility by default in AgroSuccess is analogous to the MedLand model. The default AgroSuccess fertile time of $T_F = 50\,\mathrm{yr}$ corresponds to the $1\,\%/\mathrm{yr}$ depletion rate of 'good' agriculturalists in MedLand. 'Greedy' agriculturalists could be modelled in AgroSuccess by setting $T_F = 25\,\mathrm{yr}$.

Scenarios in which households work to decrease the rate of soil nutrient depletion further could depend on increasing T_F, as this would imply a reduction in the rate of soil depletion as a consequence of the household's efforts to fertilise the soil. I explore the model's sensitivity to $T_F = 50\,\mathrm{yr}$ in Chapter 5.

The possibility of household agents working to replenish soil fertility in DAL is represented by the DAL land-cover type's land-cover conversion cost of 5 (see Table 4.1). This represents the effort needed to convert DAL back to productive wheat-producing land e.g. using intensive artificial fertilisation with animal manure. The MedLand authors discuss soil fertility regain in MedLand, specifically "... the rates of reduction and regain [of soil fertility] are set as constants at the start of the model..." (Ullah, 2013, p. 97). The rates of soil fertility reduction are specified as aspects of the household behaviour scenarios (Ullah, 2013, p. 161), but I have not been able to find explicit statement of the fertility regain rate in the literature around MedLand. However, I assume based on the above quote that fertility regain is some fixed percentage that is independent of human agricultural activities. I aim to capture this in the ecological succession processes that transition a DAL patch to shrubland etc.

A possible future improvement to AgroSuccess would be to investigate the use of fire to improve soil fertility (Perry, Wilmshurst, McGlone & Napier, 2012). This could be implemented by manipulating the counter already implemented in AgroSuccess to track whether or not a cell that is used for agriculture has exceeded its fertile time, T_F. Reintroduction of nutrients into the soil through burning could be represented by decreasing or resetting this counter.

An additional practical argument against the use of objective physical measurements to characterise land patch fertility, wood value and conversion cost in AgroSuccess is that such measures are likely to be specific to a particular study site. Unlike the number of calories an individual can obtain from consuming $1\,\mathrm{kg}$ of wheat (see Table 4.9), the nutrient content of the soil, or perceived effort to cut down oak forest to grow crops is likely to differ between study sites and groups of agents. The acquisition of data to provide measured quantities for each study site would be prohibitively difficult and an inefficient use of resources considering that it is only the *relative* difference between land-cover types at each study site that is relevant for agent decision-making.

The use of objective physical measurements as inputs as a means to persuade the reader of the precision or accuracy of the model's outputs would be misleading given the scale of uncertainty present in the model's structure and parameters. By presenting these aspects of the model as subjective judgements in this way I provide the reader with a framework in which they can scrutinise our proposals and suggest reasoned improvements.

Environmental conditions

The *environmental conditions* represented in AgroSuccess are all the variables that characterise the state of a land patch throughout the course of a model run *apart from* its land-cover type. These are manipulated and consumed by the succession submodel (see Section 4.2.3) to determine how land-cover state should evolve in response to ecological processes.

`water`: The local soil moisture class. These are, in order of increasing soil moisture, 'xeric', 'mesic', and 'hydric'. See Section 4.2.5 for details of how the thresholds between these classes are specified.

`aspect`: The binary aspect of the land patch, either 'north' or 'south'.

`succession`: Indicates whether the land patch is on a 'secondary' or 'regeneration' succession

pathway. See Section 4.2.3 for a detailed desciption of succession pathways.

pine: Binary variable indicating whether pine seeds are present in the patch, J_{pine}.

deciduous: Binary variable indicating whether seeds for non-resprouting deciduous species are present in the patch, J_{dec}.

oak: Binary variable indicating whether acorns are present in the patch, J_{oak}.

delta_t: The amount of time (in yr) the patch has been in its current land-cover state, L_{c}.

These environmental conditions are the same as those used by J. D. A. Millington et al., 2009.

4.2.3 Ecological succession

The succession submodel's role is to represent the transitions between land-cover states that occur due to the interactions between different plant species, and with their environment. It encodes qualitative knowledge about ecology which has been discovered by observation and experimentation at a finer spatio-temporal scale than that represented in AgroSuccess. For example, if the supply of water in a landscape is limited, a patch of shrubland in that landscape might develop into pine forest rather than oak forest as it might have done if the supply of water was plentiful. The states and possible combinations of environmental conditions that are possible in AgroSuccess are described in the preceding section (Section 4.2.2). As discussed in Section 4.2.1, the environmental succession submodel of AgroSuccess is an RBCLM, meaning that the possible transitions are defined in terms of start and end states, combinations of environmental conditions, and the time each transition takes. Here I explain how specific combinations of environmental conditions, start states and end states, and transition times are grouped together to form *succession rules* in AgroSuccess. All possible transitions due to ecological succession are shown in Fig. 4.1, and the time for each possible transition to occur (depending on specific environmental conditions in the grid cell) is shown in Table 4.2.

Land-cover state update rules

The succession rules in AgroSuccess are adapted from those used by J. D. A. Millington et al., 2009 in the Millington LFSM (Wainwright & Millington, 2010). In the supplementary materials

Table 4.2: Land-cover transitions represented in AgroSuccess, and the time taken for each transition. Some transitions can take a range of time to complete as they can occur faster or slower depending on the specific combination of environmental conditions.

Initial land-cover type	Final land-cover type	Transition time
Burnt	Grassland	1 yr
Wheat	Shrubland	3 yr
Depleted Agricultural Land	Shrubland	3 yr
Grassland	Shrubland	1 yr
Shrubland	Oak	30–50 yr
Shrubland	Deciduous	15–20 yr
Shrubland	Transition Forest	15 yr
Shrubland	Pine	10–15 yr
Pine	Transition Forest	15–40 yr
Pine	Deciduous	20 yr
Transition Forest	Deciduous	20–25 yr
Transition Forest	Oak	20–50 yr
Transition Forest	Pine	20–30 yr
Deciduous	Pine	20–30 yr
Deciduous	Transition Forest	30–40 yr
Oak	Transition Forest	30 yr

of their paper describing the Millington LFSM, J. D. A. Millington et al., 2009 provide a table that specifies all the possible combinations of environmental conditions that can lead to a land-cover patch transitioning from one state to another state in their model, including the time required for each transition to occur. In the `agrosuccess-graph` software repository (Appendix G.S7) I provide programs that extract the data from the original paper's supplementary materials specifying the succession rules (`scripts/clean_millington_trans_table.py`), and generate a new succession rule table that is modified to account for the differences between the Millington LFSM and AgroSuccess (`scripts/repurpose_trans_rules_agrosuccess.py`). Notably I replace the 'cropland' land-cover state represented in the Millington LFSM with wheat and DAL land-cover states such that any transition from cropland to any other state in the Millington LFSM enables the same transition from *either* wheat or DAL in AgroSuccess.

The transition rules that form the state-and-transition model underlying AgroSuccess are designed to reflect the effect that a range of possible environmental gradients would have on the competitive advantage of different species and, by extension, the land-cover types those species produce. For example, pine performs better in drier soils than oak, and vice versa. This is modelled by the transition rules in the AgroSuccess STM by having hydric soil moisture cause Shrubland cells to transition to Oak, and xeric soil moisture causes Shrubland to transition to Pine (all else being equal). These rules are encoded in a property graph (see Section 7.2) and constitute the *declarative* knowledge component of the state-and-transition model that underpins

the AgroSuccess ecological succession submodel. The other component of the STM is *procedural* knowledge about how land-cover transitions take place (McIntosh, 2003). This is expressed in terms of a set of logical statements that together specify the behaviour (though not the implementation) of the algorithm that is used in AgroSuccess to update the land-cover state of grid cells during the course of a simulation run. Every transition in AgroSuccess is defined in terms of a start state, C, a target state, ΔD, and a transition time, ΔT. ΔT is the total transition time for a grid cell to transition from its current state to its target state under current environmental conditions according to the STM. I also define another transition time, $\Delta\hat{T}$, that represents the total transition time for a particular grid cell in the simulation. It is not necessarily true that $\Delta\hat{T} = \Delta T$ because while ΔT refers to the default time a transition from state C to state ΔD takes to occur, situations can arise during a simulation run (discussed later in this section) that cause the transition time to deviate from the default on a cell-by-cell basis. $\Delta\hat{T}$ represents an endogenous variable specific to a particular cell. I formally define C, ΔD, ΔT, and $\Delta\hat{T}$ as functions of time

$$C = C(t) \in L \tag{4.1}$$

$$\Delta D = \Delta D(t) \in L \cup \{\varnothing\} \tag{4.2}$$

$$\Delta\hat{T} = \Delta\hat{T}(t) \in \mathbb{N} \cup \{\varnothing\} \tag{4.3}$$

$$\Delta T = \Delta T(t) \in \mathbb{N} \cup \{\varnothing\}\,. \tag{4.4}$$

Here

$$L = \{\text{Burnt, Wheat, DAL, Grassland, Shrubland, Pine, Transition Forest, Deciduous, Oak}\}$$

is the set of possible land-cover types in the succession submodel. This includes all the land-cover types presented in Table 4.1 except 'Water/ Quarry' because it does not participate in ecological processes. $\mathbb{N} = \{0, 1, 2, \dots\}$ denotes the set of natural numbers, and $\varnothing = \{x : x \neq x\}$ is the empty set. It is used in this context to represent situations where the STM underlying AgroSuccess does not specify a target state for a combination of current state and physical attributes of the cell (see

Section 4.2.2). This occurs when the environmental conditions in a cell favour the current land-cover state persisting indefinitely.

The rules specifying how the land-cover types of grid cells are updated are expressed using a series of logical sentences presented in this section. The number and complexity of the rules that define the ecological succession model make it impractical to state them precisely in natural language. The use of logical symbols makes it possible to express all rules concisely and unequivocally, allowing them to be scrutinised to confirm that all possible simulation states have been accounted for with corresponding rules. This will make it easier for future readers to accurately reproduce my work than it would be if the rules were expressed in natural language. In this sense, the use of logical expressions is the clearest way to describe the (dynamic) state update rules analogously to how a property graph is the clearest way to describe the related (structural) land-cover state and transition model (see Section 7.2). To assist readers who are not familiar with the notation used in this section, I provide a set of descriptive examples in Appendix D.

The following sentences define predicates in terms of the current simulation time step, t.

$$Ct \leftrightarrow C(t) = C(t-1) \tag{4.5}$$

$$Dt \leftrightarrow \Delta D(t) = \Delta D(t-1) \tag{4.6}$$

$$D_{\varnothing}t \leftrightarrow \Delta D(t) = \varnothing \leftrightarrow \Delta T(t) = \varnothing \tag{4.7}$$

$$T_{\varnothing}t \leftrightarrow \Delta \hat{T}(t-1) = \varnothing \tag{4.8}$$

Here Eq. (4.5) states that at time step t the land-cover state in the cell is the same as it was in the previous time step. Eq. (4.6) states that at time step t the land-cover state the cell is in the process of transitioning to is the same state it was transitioning to in the previous time step. Eq. (4.7) states that at time step t, there is no target state or corresponding transition time specified for the cell's combination of environmental conditions in the STM. Eq. (4.8) states that at time step t the transition time for the cell in the previous time step, $\Delta\hat{T}(t-1)$, was unspecified, i.e. the cell was on a trajectory to remain in the current state indefinitely.

The sentences used to determine the number of years a simulation cell has been in its current state, $T_{\text{in}}(t)$, are

$$\forall t \in \mathbb{N} \setminus \{0\} \, (\neg Ct \rightarrow T_{\text{in}}(t) = 1) \tag{S1a}$$

$$\forall t \in \mathbb{N} \setminus \{0\} \, (Ct \wedge Dt \rightarrow T_{\text{in}}(t) = T_{\text{in}}(t-1) + 1) \tag{S1b}$$

$$\forall t \in \mathbb{N} \setminus \{0\} \, (Ct \wedge \neg Dt \rightarrow T_{\text{in}}(t) = 1) \tag{S1c}$$

The following sentences specify how simulation transition times, $\Delta\hat{T}(t)$, are updated

$$\forall t \in \mathbb{N} \setminus \{0\} \, (Ct \wedge \neg Dt \wedge \neg D_{\varnothing} t \wedge \neg T_{\varnothing} t \rightarrow \Delta\hat{T}(t) = \text{round}([\Delta\hat{T}(t-1) + \Delta T(t)]/2)) \tag{S2a}$$

$$\forall t \in \mathbb{N} \setminus \{0\} \, (Ct \wedge \neg Dt \wedge \neg D_{\varnothing} t \wedge T_{\varnothing} t \rightarrow \Delta\hat{T}(t) = \text{round}([1 + \Delta T(t)]/2)) \tag{S2b}$$

$$\forall t \in \mathbb{N} \setminus \{0\} \, (Ct \wedge Dt \wedge \neg D_{\varnothing} t \rightarrow \Delta\hat{T}(t) = \Delta\hat{T}(t-1)) \tag{S2c}$$

$$\forall t \in \mathbb{N} \setminus \{0\} \, (Ct \wedge D_{\varnothing} t \rightarrow \Delta\hat{T}(t) = \varnothing) \tag{S2d}$$

$$\forall t \in \mathbb{N} \setminus \{0\} \, (\neg Ct \rightarrow \Delta\hat{T}(t) = \Delta T(t)) \tag{S2e}$$

S2a and S2b account for situations where the target land-cover state has changed part-way through a transition. If the previous simulation transition time was unspecified, I assume the previous transition time was $1\,\text{yr}$. This is consistent with the approach taken to using null transitions in the Millington LFSM (see Section 4.2.7). S2c ensures that if neither the current state nor the target state have changed since the last time step, the transition time will remain the same. This prevents a situation where the transition time calculated by S2a or S2b is overwritten by the native transition time of ΔD.

The following statements are used to determine the grid cell's land-cover state in time step t, $C(t)$.

$$\forall t \in \mathbb{N} \setminus \{0\} \, (T_{\text{in}}(t-1) \geq \Delta\hat{T}(t-1) \wedge \neg T_{\varnothing} t \rightarrow C(t) = \Delta D(t-1)) \tag{S3a}$$

$$\forall t \in \mathbb{N} \setminus \{0\} \, (T_{\text{in}}(t-1) < \Delta\hat{T}(t-1) \vee T_{\varnothing} t \rightarrow Ct) \tag{S3b}$$

S3a causes a cell that has been in its current state for a sufficient amount of time to transition to its target state according to the STM. S3b states that if a cell has not been in its current state

for long enough to transition to a new land-cover type according to its succession trajectory, it remains in its current state. If the current combination of physical conditions imply a simulation cell won't transition to a different land cover type, its land cover type in the current time step will remain the same as its land-cover type in the previous time step.

Remove juvenile individuals in areas transitioning to mature vegetation

I consider the Pine, Oak, and Deciduous land-cover types to represent 'mature' vegetation communities. By this I mean that in areas occupied by each of these types, individuals that are representative of the type have successfully out-competed individuals of other land-cover types. For example, we expect individuals in mature oak woodland to shade out pine saplings. Under these circumstances we do not expect there to be juvenile vegetation belonging to species representative of other land-cover types present in the understory. To represent this ecologically motivated constraint I have included a rule in AgroSuccess stating that when a simulation cell transitions to a mature vegetation state, any juvenile plants present in that cell (pine, oak and deciduous) are removed.

$$\forall t \in \mathbb{N} \setminus \{0\} \left(\neg Ct \wedge C(t) \in L_{\text{mature}} \rightarrow \neg(J_{\text{pine}} \vee J_{\text{oak}} \vee J_{\text{dec}})\right) \tag{S4}$$

where $L_{\text{mature}} = \{\text{pine}, \text{oak}, \text{deciduous}\}$. This rule interacts with the colonisation submodel whose role is to distribute juvenile vegetation in the landscape. Since juvenile vegetation is otherwise persistent (that is, if seedlings are deposited in grid cell i at time t, they will stay there in subsequent time steps $t+1, t+2, \ldots$), without S4 there would be no mechanism to remove juvenile plants from the landscape (see Section 4.2.4).

Determine succession pathway

Grid cells have a 'succession pathway' physical attribute that determines whether they are on a secondary or regeneration succession pathway. Secondary succession occurs when mature vegetation develops on an established soil that previously lacked mature resprouter species (e.g. oak). Regeneration succession occurs when mature vegetation develops at a site where resprouting

species had been established previously, but had been destroyed by a disturbance event leaving some resprouter vegetative matter intact (J. D. A. Millington et al., 2009). I denote the fact that a grid cell is on a secondary regeneration pathway by P_{sec}, and the fact that a grid cell is on a regeneration pathway by P_{reg}.

In the landscapes represented by AgroSuccess, the key resprouting species is evergreen oak (e.g. *Quercus ilex*), and the relevant disturbance process that leads to regenerative succession is burning due to wildfire. A grid cell transitions *to* the regeneration pathway when it first contains reproductively mature oak, i.e. individuals capable of producing acorns. Both the Oak and Transition Forest land-cover types contain reproductively mature oak. For each grid cell I maintain a count of the number of years it has remained in *either* of these states, T_{oak}. If the cell transitions to any state other than Oak or Transition Forest, T_{oak} for that cell is set to 0. A grid cell can transition from the regeneration succession pathway to the secondary succession pathway if it experiences sufficiently frequent fire. This is referred to as 'oak mortality.'

We define

$$T'_{\text{oak}} = \begin{cases} T_{\text{oak}} & \text{if} \quad T_{\text{oak}} < 100\,\text{yr} \\ 100\,\text{yr} & \text{otherwise} \end{cases}$$

encoding the assumption that if oak individuals have been established for $100\,\text{yr}$ their resilience to fire becomes constant. If the frequency of fire in the cell is sufficiently high and it is not already on the secondary succession pathway, the cell will transition to it.

$$f_{\text{fire}} > T'_{\text{oak}}/\Omega_{\text{oak}} \rightarrow P_{\text{sec}} \tag{S5}$$

where f_{fire} (fire/yr) is the number of fires in the cell per year (an endogenous model variable), and $\Omega_{\text{oak}} = 200\,\text{yr}^2/\text{fire}$ is an oak mortality scaling parameter (J. D. A. Millington et al., 2009). If the antecedent of S5 is not true, the cell will remain in its current succession pathway.

Comparison of land-cover state update rules to the Millington LFSM

The rules specifying the procedural knowledge in AgroSuccess described in the preceding paragraphs correspond closely to rules in the Millington LFSM presented in J. D. A. Millington et al., 2009 with some modifications described here. The succession rules in both AgroSuccess and the Millington LFSM are expressed as sets of logical formulas. Table 4.3 shows the formulas in AgroSuccess that correspond to equivalent statements in the Millington LFSM. During the development of AgroSuccess I augmented the logical statements specified in J. D. A. Millington et al., 2009 to make features of the model that were treated as implementation details in the Millington LFSM explicit. This is important to facilitate informed discussion about the structure of AgroSuccess without readers needing to scrutinise the relevant source code.

Table 4.3: Mapping between model rules in AgroSuccess specified in Section 4.2.3 and equivalent rules in the Millington LFSM (J. D. A. Millington et al., 2009). The 'Rule purpose' column indicates the aspect of the model each rule contributes to (see main text for details). There are serveral rules in AgroSuccess with no counterpart in the Millington LFSM, denoted by '-' in the 'Millington LFSM rule' column. See main text for the motivation and description of these additional rules. Rule 2a in AgroSuccess differs from rule 4 in the Millington LFSM by making it explicit that derived transition times arising from changes in target state part way through a transition are rounded to the nearest whole year.

Rule purpose	AgroSuccess rule	Millington LFSM rule
Update time in state, T_{in}	S1a	Statement 1
	S1b	Statement 2
	S1c	Statement 3
Update transition time, $\Delta\hat{T}$	S2a	Statement 4*
	S2b	-
	S2c	-
	S2d	-
	S2e	-
Update land-cover state, C	S3a	Statement 5
	S3b	Statement 6
Remove juvenile vegetation	S4	-
Determine succession pathway	S5	Statement 7a, 7b

To account for the difference in my approach to representing null transitions compared to J. D. A. Millington et al., 2009 (see Section 4.2.7 for a description of 'null transitions') I have added the logical statements S2b, S2c, and S2d. Together these emulate the outcome of the null transitions in the Millington LFSM, where the associated transition time, $\Delta\hat{T}$, is set at $1\,\text{yr}$.

Additionally, I have included S4 in the AgroSuccess succession submodel to link the process of a patch of land-cover reaching a state of mature vegetation cover with the removal of juvenile individuals in that land patch. This is a rule that was implemented in the Millington LFSM, but was not formally documented in J. D. A. Millington, 2007 or J. D. A. Millington et al., 2009 (J. D. A.

Millington 2020, personal communication, 14 July). By expressing this process as a formal rule I aim to make the structure of the model more explicit.

4.2.4 Land-cover colonisation

An important part of the life cycle of most plant species is the production and distribution of seeds. The presence of seeds in a particular location is a necessary condition for these species to be able to establish dominance in that location at some future time. This includes resprouter species such as *Quercus spp.* because, even in areas undergoing regeneration succession, the presence of resprouting vegetative matter implies the colonisation of the area with oak seeds (acorns) at some point in the past. In AgroSuccess the locations in the landscape of juvenile individuals corresponding to the mature vegetation land-cover types (pine, oak, and deciduous) are encoded in the environmental condition attributes of each simulation grid cell (see Section 4.2.2). Each grid cell has three binary variables that indicate the presence or absence of juvenile individuals corresponding to the Pine, Oak, and Deciduous land-cover types—J_{pine}, J_{oak}, and J_{dec} respectively. These variables influence the ecological succession submodel by enabling patches containing juvenile vegetation to transition to the corresponding land-cover type (see Section 4.2.3). The succession submodel also accounts for the case of regeneration succession using a separate 'succession pathway' environmental condition attribute, symbolised by P_{sec} for secondary succession and P_{reg} for regeneration succession. In this way a patch can transition to the Oak land-cover type if it contains regenerative oak vegetative matter even if it does not contain acorns or juvenile oak.

In this section I present the land-cover colonisation submodel used in AgroSuccess. I depart from J. D. A. Millington et al., 2009 by referring to the modelled process as 'land-cover colonisation' rather than 'seed dispersal' to reflect the greater level of abstraction with which I have modelled the establishment of juvenile individuals. In AgroSuccess the land-cover colonisation model is intended to represent the aggregate effect of all factors that culminate in successfully germinated seedlings becoming established at locations in the simulated landscape.

I characterise 'colonisation' as the culmination of the processes of seed dispersal, seed germination, seedling establishment, and seedling survival (Thompson, 2005). The land-cover colonisation model in AgroSuccess takes the location of simulation grid cells containing Oak, Pine, and Deciduous land-cover as input, and returns the locations of cells containing juvenile vegetation

corresponding to each of those land-cover types. Colonisation implies the successful establishment and survival of saplings by definition. Correspondingly, once a cell has had juvenile vegetation corresponding to a land-cover type deposited inside it, that juvenile vegetation will remain there until the cell transitions to one of the mature vegetation land-cover states (Pine, Oak, or Deciduous), at which point it is removed (see Section 4.2.3). The removal of juvenile vegetation upon transition to a mature land-cover type represents either the successful development of juveniles into mature individuals corresponding to the same land-cover type, or the situation where juveniles of one land-cover type are out-competed by mature individuals corresponding to a different land-cover type.

Any cell containing mature vegetation corresponding to land-cover type $\sigma \in \{\text{Pine}, \text{Oak}, \text{Deciduous}\}$ is assumed to contain juvenile vegetation corresponding to that type. Additionally in each time step, t, I randomly sample

$$N_{\sigma,t} = aN_{\sigma,t} + bN_{\text{tot}} \tag{4.9}$$

grid cells and add juvenile vegetation corresponding to land-cover type $\sigma \in \{\text{Pine}, \text{Oak}, \text{Deciduous}\}$ to these cells. Here the parameter $a \geq 0$ is the proportion of the number of cells in the simulated landscape containing mature vegetation which produces juvenile vegetation, and $b \geq 0$ is the 'background rate' of juvenile vegetation for each land-cover type in the model. The background rate represents juvenile vegetation due to seeds entering the landscape from outside the simulation grid, which avoids certain land-cover types going 'extinct' in the simulation. $N_{\sigma,t}$ is the number of cells containing mature vegetation corresponding to land-cover type σ at time t, and N_{tot} is the total number of grid cells in the model. Note that juveniles are not added to additional cells if the cells sampled already contain juvenile vegetation of type σ.

Land-cover colonisation model complicatedness

The land-cover colonisation model described above is *completely spatially random* in the sense that the locations of cells containing juvenile vegetation are not correlated with the locations of cells containing corresponding mature vegetation. This is a modelling approximation which is simpler than the approach used by J. D. A. Millington et al., 2009 in the Millington LFSM (see Section 4.2.7). However, any approach to modelling the spatial distribution of land-cover colon-

isation is likely to be subject to simplifying assumptions. Thompson, 2005 found descriptions of colonisation processes in the literature that suggest that the approach to modelling colonisation using seed dispersal kernels found in the Millington LFSM is unlikely to capture the true spatial variability of seedling establishment. For example, although many seeds of fleshy-fruited species are dropped close to the seed source, there are peaks in the seed shadow around isolated trees (potentially of species different to the seeds themselves) that birds carrying the seeds use as perches (Thompson, 2005, pp. 133–134). I see no reason to believe the spatial distribution of acorns would not follow a similar pattern (with multiple peaks centred on potential perches) and note that the log-normal distribution used in the Millington LFSM J. D. A. Millington et al., 2009 cannot capture this pattern. Additionally, Thompson, 2005, p. 134 describes how wind-dispersed species colonise open areas through a process called *nucleation*, whereby a colony emerges from an isolated individual established from a seed dispersed over long distance. Taking the example of *Pinus sylvestris*, such an individual will produce a greater abundance of cones upon reaching maturity than a comparable individual in a stand (see e.g. Debain et al., 2003), accelerating the establishment of a new colony. Both of these examples show that the culmination of the colonising processes that lead to the establishment of either wind- or animal-dispersed species are not random in the way assumed by the seed dispersal kernel method used in the Millington LFSM. Furthermore, to model these processes would require a more fine grained (perhaps individual-based) representation of vegetation cover than is used in either the Millington LFSM or AgroSuccess.

The above discussion highlights how the level of detail represented in a model of plant colonisation is a trade-off between realism and tractability, constrained by the overarching scientific objectives of the model. I argue that the representation of colonisation in AgroSuccess should be kept as simple as possible in terms of both processes and parameters, while satisfying the objective that the greater the area occupied by a land-cover type, the more opportunities for colonisation by species corresponding to that land-cover type are created in each time step.

4.2.5 Soil moisture

Soil moisture in the simulation grid is calculated for each grid cell in each time step. This is relevant to the ecological succession submodel because the land-cover transitions available to grid cells depends on the available soil moisture (see Section 4.2.2). AgroSuccess follows the

Millington LFSM in using the United States Department of Agriculture's (USDA) approach to calculating runoff following precipitation (United States Department of Agriculture, 2004). The USDA relate runoff, Q, to precipitation, P, as follows

$$Q = \frac{(P - 0.2S)^2}{P + 0.8S} \,. \tag{4.10}$$

S is the maximum potential water retention which depends on the local slope, land-cover and soil type through a dimensionless *curve number*, CN, such that

$$S = 25.4 \left(\frac{1000}{CN} - 10 \right) \,. \tag{4.11}$$

See United States Department of Agriculture, 2004 for details of the approach to estimating runoff using curve numbers. The mapping between slope, soil type and land-cover type is given in Appendix C after J. D. A. Millington et al., 2009. P, Q, and S are all in mm.

In AgroSuccess soil moisture is calculated on a per-pixel basis such that the soil moisture of grid cell i at time t is given by

$$M_{i,t} = T_{i,t} - Q_{i,t} \tag{4.12}$$

where

$$T_{i,t} = P_t + \sum_{j \in D_i} Q_{j,t} \tag{4.13}$$

$$Q_{i,t} = \frac{(T_{i,t} - 0.2S_{i,t})^2}{T_{i,t} + 0.8S_{i,t}} \,. \tag{4.14}$$

$T_{i,t}$ is the total amount of water in mm entering cell i in time step t accounting for both precipitation, P_t, and runoff from neighbouring cells. The scenarios discussed in this thesis set P_t to be the mid-Holocene mean annual precipitation for each study site (see Section 3.3) for all years. However, more elaborate scenarios where annual precipitation varies over time could be explored in future. D_i is the set of neighbouring cells that drain into cell i, and is determined

using the flow direction map for each study site (see Section 3.3.1). Note that while each cell drains into exactly one neighbour (or out of the grid in the case of edge cells), up to seven cells can drain into a single cell depending on the topology of the terrain. This is because in AgroSuccess each cell has eight neighbours (the 'Queen's neighbourhood'), and I have preprocessed the digital elevation models used for our study sites to ensure that no cell has an elevation less than all of its neighbours (this is known as 'sink' removal, see Section 3.3.1). Eq. (4.14) corresponds to Eq. (4.10) provided by United States Department of Agriculture, 2004, but with $T_{i,t}$ replacing P. This reflects the fact that in AgroSuccess, the water entering each grid cell includes both precipitation and runoff from neighbouring cells. $S_{i,t}$ is the maximum water retention in cell i at time t calculated using Eq. (4.11). The dependence of $S_{i,t}$ on t arises because the curve number of cell i, $CN_{i,t}$, can change from one time step to the next as the cell's land-cover type evolves according to the ecological succession submodel.

In the ecological model, continuous soil moisture values in each cell expressed in mm are converted to discrete soil moisture classes. The boundaries for these classes used in AgroSuccess correspond to those used in the Millington LFSM and are shown in Table 4.4.

Table 4.4: Mapping of continuous soil moisture values to discrete soil moisture classes used in the ecological succession submodel.

Soil moisture in cell i at time t, $M_{i,t}$ [mm]	Soil moisture class
$M_{i,t} \leq 500$	xeric
$500 < M_{i,t} \leq 1000$	mesic
$M_{i,t} > 1000$	hydric

4.2.6 Fire

In addition to ecological dynamics, I also follow the example of J. D. A. Millington et al., 2009 in their implementation of a Cellular Automata model of fire spread (see also Perry, Wilmshurst, McGlone, McWethy et al., 2012; Peterson, 2002). This is a straight-forward algorithm whereby, in each simulated year, a specified number of fire ignitions are made. Each cell has a certain probability of catching fire influenced by its land-cover class. During each fire event, burning is allowed to spread from neighbouring cell to cell until there are no more active fires. While simple, J. D. A. Millington et al., 2009 showed that this model is able to reproduce wildfire frequency-size statistics comparable to those found in empirical wildfire data characterising the fire regime in areas of the US within a Mediterranean-type ecoregion (Malamud et al., 2005). In the following subsections, I describe in detail how the number of ignitions per simulated year is determined

and the algorithm used to spread these simulated fires in the landscape.

Number of ignitions per year

$\lambda = \lambda(\bar{T}, \bar{P}; m)$ is the average number of ignitions in the study region per year, given by Eq. (4.15). $\bar{T}$ is the average temperature taken across the year in degrees Celsius, and $\bar{P}$ is total annual precipitation in millimetres. Both $\bar{T}$ and $\bar{P}$ are treated as empirical data that vary by study site. m is a climate ignition scaling parameter with units $\mathrm{mm}/^{\circ}\mathrm{C}$. m is treated as a calibrated model parameter. See Section 5.1 for details of how m was calibrated for each study site.

$$\lambda(\bar{T}, \bar{P}; m) = m\frac{\bar{T}}{\bar{P}} \tag{4.15}$$

The number of fires successfully started per year follows a Poisson distribution given in Eq. (4.16). For each fire a cell will be randomly drawn from the landscape and an ignition will be attempted. If the selected cell is not a flammable land cover type (if it has already been burnt or is occupied by water, for example) another cell will be drawn. This re-drawing of cells will continue until an ignition is successful, or until there have been 1000 unsuccessful attempts to start a fire.

$$p(x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \tag{4.16}$$

Fire spread

Once a fire has started, its spread is simulated by a Monte Carlo process in which the probability of fire spreading from a cell to its neighbours depends on the neighbouring cells' biophysical properties. The probability of a fire spreading from an actively burning cell to a neighbouring cell, i is given by

$$p_i = \mathrm{LCF}_i \cdot \mathrm{SR}_i \cdot \mathrm{FMR} \cdot \mathrm{WR}_i \tag{4.17}$$

where the multiplicative risk factors are defined in the remainder of this section.

Each land cover type represented in the model contains different quantities and distribution of combustible fuel. Consequently, the likelihood that an active fire will spread to a particular patch depends on the relative flammability of the land cover in that patch. The land cover flammability value for land cover type L_c is denoted by LCF_{L_c}. In recognition of the uncertainty in the quantification of LCF_{L_c}, and the possibility that LCF_{L_c} may vary across study sites, I consider 14 *land cover flammability scenarios* where each scenario specifies a set of LCF_{L_c}, each with one LCF_{L_c} value for each land cover type. These are shown in Table 4.5. The land cover flammability scenario used for a particular simulation, S_{LCF}, is treated as a model parameter that is calibrated for each study site (see Section 5.1).

Table 4.5: Land-cover flammability values for land-cover types in AgroSuccess, LCF_{L_c}, after J. D. A. Millington et al., 2009. Each row corresponds to a different scenario, with overall land-cover flammability increasing further down the table.

L_c Scenario, S_{LCF}	Grassland	Wheat	DAL	Shrubland	Pine	T. Forest	Deciduous	Oak	Mean
TFN4	0.15	0.15	0.15	0.15	0.14	0.14	0.13	0.13	0.14
TFN3	0.16	0.16	0.16	0.16	0.15	0.15	0.14	0.14	0.15
TFN2	0.17	0.17	0.17	0.17	0.16	0.16	0.15	0.15	0.16
TFN1	0.18	0.18	0.18	0.18	0.17	0.17	0.16	0.16	0.17
TF0	0.19	0.19	0.19	0.19	0.18	0.18	0.17	0.17	0.18
TF1	0.20	0.20	0.20	0.20	0.19	0.19	0.18	0.18	0.19
TF2	0.21	0.21	0.21	0.21	0.20	0.20	0.19	0.19	0.20
TF3	0.22	0.22	0.22	0.22	0.21	0.21	0.20	0.20	0.21
TF4	0.23	0.23	0.23	0.23	0.22	0.22	0.21	0.21	0.22
Default	0.24	0.24	0.24	0.24	0.23	0.23	0.22	0.22	0.23
TF5	0.25	0.25	0.25	0.25	0.24	0.24	0.23	0.23	0.24
TF6	0.26	0.26	0.26	0.26	0.25	0.25	0.24	0.24	0.25
TF7	0.27	0.27	0.27	0.27	0.26	0.26	0.25	0.25	0.26
TF8	0.28	0.28	0.28	0.28	0.27	0.27	0.26	0.26	0.27

To reflect the fact that fire spreads more easily uphill than downhill I include a slope risk factor, SR_i. The values for SR_i for grid cell i with percentage slope $S_i^{(\%)}$ are given in Table 4.6.

Table 4.6: Slope risk value for grid cell i, SR_i given a slope (grade) between cell i and a burning cell of $S_i^{(\%)}$. SR_i is a multiplicative factor used to determine the likelihood that a fire spreads from a cell that is already burning to cell i.

Slope, $S_i^{(\%)}$ [%]	Slope risk, SR_i
$S_i^{(\%)} < -25$	0.80
$-25 \leq S_i^{(\%)} < -15$	0.90
$-15 \leq S_i^{(\%)} < -5$	0.95
$-5 \leq S_i^{(\%)} < 5$	1.00
$5 \leq S_i^{(\%)} < 15$	1.05
$15 \leq S_i^{(\%)} < 25$	1.10
$S_i^{(\%)} \geq 25$	1.20

The expected number of fires per year, λ, is used to determine the *fuel-moisture risk* of the landscape. From Eq. (4.15) we see that λ is a linear function of the ratio of temperature to precipitation, such that for greater values of λ we expect the likelihood of vegetation in the landscape to catch fire when exposed to a source of ignition to increase. Consequently λ can also be thought of as a vegetation moisture parameter. Table 4.7 shows the mapping between vegetation moisture classes and fuel-moisture risk factors. These factors are the same as those used in the Millington LFSM and are used to calculate the probability that a fire spreads from one simulation grid cell to another. Notice that while the other multiplicative factors in Eq. (4.17) depend on the local biophysical conditions of the cell the fire might spread to, FMR is common to all cells in the landscape.

Table 4.7: Mapping from vegetation moisture class, λ, (see Eq. (4.15)) and fuel-moisture risk factors used in the simulation model. Fuel-moisture risk factors increase the risk of fires spreading in hotter, drier landscapes.

Vegetation moisture class, λ	Fuel-moisture risk (FMR)
$\lambda < 0.2$	0.8
$0.2 \leq \lambda < 0.3$	0.9
$0.3 \leq \lambda < 0.5$	1.0
$0.5 \leq \lambda < 0.6$	1.1
$\lambda \geq 0.6$	1.2

Wind speed and direction influences wildfire spread because fire is more likely to spread into a cell that is downwind of a burning cell, and fire is more likely to spread if wind speed is high than if wind speed is low. Cell i is 'downwind' of cell j if the vector from cell j to cell i is parallel to the direction of the wind. Following J. D. A. Millington et al., 2009 and Perry and Enright, 2002 I use a model that allocates wind risk factors, WR_i to grid cells adjacent to cells that are already burning. These are shown in Fig. 4.2.

For each study site I derived empirical probability distributions that quantify the probability of observing a particular wind speed category and direction on a given day (see Section 3.3). Each time a fire occurs during a simulation run, a specific wind speed and direction are sampled from these distributions and remain constant for the duration of the fire.

4.2.7 Relationship of AgroSuccess to the Millington LFSM

J. D. A. Millington et al., 2009 deployed RBCLM in a Landscape Fire Succession Model (LFSM) in a manner similar to that described in Section 4.2.1. This LFSM model was subsequently integrated with an agent-based model of contemporary agricultural land-use/cover change to form

a coupled model called SPASIMv1 (Wainwright & Millington, 2010). SPASIMv1 was successfully used to study land-cover change in an $830\,\mathrm{km}^2$ area of Spain (J. Millington et al., 2008), an area approximately 20 times larger than the $40\,\mathrm{km}^2$ spatial extent of AgroSuccess. While the agent-based component of SPASIMv1 focuses on contemporary (rather than prehistoric) land-use change, the LFSM described in J. D. A. Millington et al., 2009 represents Mediterranean ecosystems at similar spatio-temporal grains to those required in AgroSuccess. Motivated by these similarities I took considerable inspiration from the LFSM developed in J. D. A. Millington et al., 2009—which I will refer to from now on as 'the Millington LFSM'—during our development of AgroSuccess. In particular, the design of both the ecological succession and fire spread modules used in AgroSuccess are based on those used in the Millington LFSM. In this section I describe the similarities and differences between AgroSuccess and the Millington LFSM, as well as how the succession mechanism implemented in AgroSuccess is tailored to address our research objective of investigating the influence of early agriculturalists on land cover change.

Soil fertility and land-cover conversion cost land-cover type attributes

A key way in which AgroSuccess differs from the Millington LFSM is in its association of land-cover states with fertility and land-cover conversion cost attributes. These attributes do not contribute directly to AgroSuccess' ecological succession pathways, but are used by simulated agriculturalist households to evaluate different land cover patches' suitability for agriculture. This enables AgroSuccess to represent agricultural household decision-making processes similar to those in MedLand consistently with the STM underlying its ecological succession submodel (see Section 4.3).

Succession submodel

The RBCLM succession submodel in AgroSuccess has two components:

1. A state-and-transition model (STM) that specifies which land-cover types are represented in the model and the transitions which occur between them under specific sets of environmental conditions and time scales (declarative knowledge).

2. A set of rules that govern how simulation grid cells transition from one land-cover state to

another (procedural knowledge).

Like the Millington LFSM, declarative knowledge about which land-cover states are represented and which transitions are possible in AgroSuccess are encoded in an STM. However, there are differences in the land-cover states which are represented in the two models (see above) leading to differences in the structure of their STMs. Consequently, the behaviour of the succession submodels will differ between the two models due to differences in the structure of the respective STMs underlying them.

Apart from differences in structure between the STMs in AgroSuccess and the Millington LFSM, I have also made changes to the way the STM is implemented. The STM in the Millington LFSM is implemented as a transition table where each row encodes a single transition under a specific set of environmental conditions (see supplementary materials in J. D. A. Millington et al., 2009). The transition table includes a row for every combination of environmental conditions for each pair of states for which any transition pathway exists, including 'null transitions'. Null transitions occur when, for a particular combination of environmental conditions, grid cells should remain in their current land-cover type indefinitely. This reflects ecologically realistic situations where environmental conditions persistently favour the existing land-cover type over any other. Null transitions contribute a significant fraction of all transitions. In AgroSuccess grid cells are always in one of 96 possible states in respect to their environmental conditions (see Section 4.2.2). The proportion of these states that cause a cell to remain in its current state varies by land-cover type, but is greater than one third for all types. For example, $36\,\%$ of possible combinations of environmental conditions would cause a grid cell to remain in the Pine land-cover state until those conditions changed, whereas the equivalent proportion of combinations of environmental conditions for the Oak land-cover type is $57\,\%$. Rows corresponding to null transitions in the Millington LFSM STM transition table specify that grid cells with the corresponding land-cover state and combination of environmental conditions should 'transition' to the same land-cover state after a period of $1\,\mathrm{yr}$. This design causes a cell which has been in its current state for 5 time steps to be indistinguishable from one which has remained in its present state for 500 time steps. I argue that these two scenarios are qualitatively different, and that it is diagnostically useful from a modelling perspective to be able to detect when cells have remained in a particular state for a large proportion of the simulation run time. Such a finding might motivate a reconsideration of additional transition rules which were previously dismissed.

The STM in AgroSuccess is implemented as a property graph (see Section 7.2) rather than as a transition table. The existence of null transitions is implied by the absence of transition pathways through the property graph encoding the STM for particular combinations of environmental conditions. By implementing the STM in a form which is easy to visualise (i.e. a graph) and by not explicitly representing null transitions, the AgroSuccess STM is easier to specify and visually inspect than the Millington LFSM STM. These changes to the way the declarative component of the AgroSuccess STM is implemented compared to the Millington LFSM are not intended to produce any difference in the behaviour of the model. However, to maintain equivalence of behaviour it is necessary to make a corresponding modification to the rules specifying procedural knowledge about the system. Briefly, it is necessary to introduce a *null state* that grid cells can be classified as being in the process of transitioning towards (as opposed to any particular land-cover state). In addition to enabling the improvements to the representation of the STM described above, the introduction of the null state simplifies the identification of those grid cells that have been involved in a null transition for a large proportion of a simulation run time. See Section 4.2.3 for details of the differences in the representation of procedural knowledge in AgroSuccess compared to the Millington LFSM.

Seed distribution submodel

The Millington LFSM contains a seed dispersal submodel (see Section 3.3 in J. D. A. Millington et al., 2009). This model simulates the distribution of seeds in the landscape by using empirically motivated probability distributions to relate the distance of a cell from its nearest seed source to the probability of finding seeds in it. In general, cells that are closer to a seed source have a higher probability of having seeds in them according to the probability distributions, such that the locations of cells containing seeds become spatially correlated with the locations of seed sources. As noted above, the spatial extent of AgroSuccess is approximately 20 times smaller than the Millington LFSM. I argue that over the $40\,\mathrm{km}^2$ landscapes which are simulated in AgroSuccess, the presence of some seeds of a particular type in the landscape is likely to mean such seeds could be found anywhere in that area. On this basis I have chosen to implement a simpler *spatially random* land-cover colonisation model (see Section 4.2.4) similar to the seed distribution submodel used in an earlier unpublished version of the Millington LFSM (J. D. A. Millington, 2007). This model is much cheaper to run computationally, simpler to implement efficiently, and easier to reason about and evaluate due to the smaller number of model parameters.

Graph representation of succession submodel

The succession model in the Millington LFSM is encoded and distributed (see J. D. A. Millington et al., 2009, supplementary material) as a 751-line table in which each row specifies an individual transition, and the columns encode start and end states, an exhaustive set of environmental conditions, and the transition time. During the development of AgroSuccess, I capitalised on the inherently networked structure of RBCLMs to encode and interact with the succession submodel as a graph database (see Section 7.2). Being able to visualise the model as a graph rather than a table enabled me to identify some transcription errors that occurred during the manual transfer of the details of the Millington LFSM from an unpublished PhD thesis (J. D. A. Millington, 2007) to a published journal article (J. D. A. Millington et al., 2009). The succession submodel in AgroSuccess is encoded using automated data processing steps which can be reproduced using code distributed in Appendix G.S7. See Section 4.2.3 for further details.

4.3 Agent-based model of subsistence agriculture

In this section I describe the agent-based subsistence agriculture component of AgroSuccess. This submodel builds on modelling of land-cover change from subsistence agriculture developed previously in the *MedLand* model (Barton et al., 2010). See Section 4.3.2 for an overview of the relationship between AgroSuccess' subsistence agriculture submodel and MedLand. In Section 4.3.3 I give an overview of the entities and processes represented in the submodel, before specifying model behaviours in detail in Section 4.3.4.

4.3.1 The ODD protocol

I have chosen to use the ODD+D protocol (Müller et al., 2013) to formally describe the human decision-making component of AgroSuccess. This is a framework designed to assist in the clear description of agent-based models that incorporate human decision making. It is an evolution of the original ODD protocol for specifying agent-based models in general (Grimm et al., 2006), and shares with its predecessor the ambition of making descriptions of ABMs more consistent, and therefore quicker to understand and easier to reproduce. See Table B.1 in Appendix B for the AgroSuccess ODD model specification.

4.3.2 Relationship between AgroSuccess and MedLand

Many of the concepts, rules, and parameters that govern the subsistence agriculture submodel in AgroSuccess are inspired by MedLand directly, as described by Ullah, 2013. MedLand—the 'Mediterranean Landscape Dynamics project' (Barton et al., 2010)—is a simulation modelling laboratory originally developed by a team in the School of Human Evolution and Social Change at Arizona State University to investigate the impact of Neolithic agropastoral land use practices on soil erosion. At its heart is an agent-based model that represents households of agriculturalists who interact with the landscape in the vicinity of their village to produce the food and firewood they need to subsist.

Incorporating aspects of a model developed by other researchers into AgroSuccess is an efficient use of research time. By leveraging work performed by experts in the field, I was able to combine existing knowledge with my own understanding of ecological processes to address novel research questions. A model like MedLand represents a formalisation of knowledge about Neolithic land management practices. Its construction required a detailed literature review of archaeological and ethnographic theory to specify a plausible description of life in the prehistoric Mediterranean. The manner in which I have been able to reuse ideas from MedLand is an example how formalising theoretical knowledge about the mechanisms and parameters that drive complex systems into a model helps to increase the accessibility of that knowledge to future researchers (see Section 2.6.1).

Despite the shared features between the subsistence agriculture submodel of AgroSuccess on one hand, and MedLand on the other, I emphasise that AgroSuccess builds on MedLand, and present the subsistence agriculture mechanisms included in the AgroSuccess submodel as a new model. To avoid multiple digressions from the presentation of the subsistence agriculture submodel in AgroSuccess to explain how and why specific mechanisms differ from their counterparts in MedLand, Section 4.3.3 and Section 4.3.4 describe the subsistence agriculture model as it is implemented in AgroSuccess without reference to MedLand. Summaries of MedLand's development history and how AgroSuccess differs from MedLand are now presented.

MedLand publication history

In Fig. 4.3 I show a selection of papers describing the MedLanD model that have been published throughout a period of over 12 years. The model was first described in conference proceedings by Mayer et al., 2006, which provides an overview of the modelled processes and describes preliminary results but does not specify any formal decision rules (e.g. in the form of equations) for the household agents. Barton et al., 2010 summarise contributions to the model at the time of publication in 2010 (Mayer, 2009; Mayer & Sarjoughian, 2007, 2009; Mayer et al., 2006), and describe the agent-based component of the model as "currently in development" at that time. Ullah, 2013 provides a full description of the household agents' decision rules in an unpublished PhD thesis. That thesis is the only source specifying the model decision rules that I have been able to identify, and excerpts from it form the basis of the documentation which the authors distribute with the model software (Barton et al., 2017). Although unpublished, I am confident Ullah, 2013 is the authoritative description of the MedLanD ABM decision rules because a subsequent paper with several authors previously involved in developing the model cites this thesis as a reference for the details of the MedLanD model (D. T. Robinson et al., 2018).

Summary of differences between AgroSuccess and MedLand

As stated above, the subsistence agriculture submodel of AgroSuccess is a new model with respect to MedLand. No code is shared between the two models' implementations and, because they are each designed to answer different research questions, they emphasise different processes in their respective representations of prehistoric subsistence agriculture. Here I give an overview of the differences between AgroSuccess and MedLand, paying particular attention to the reasons for the high-level differences between the two models. For specific details of how the decision rules followed by the agropastoralist household agents in AgroSuccess differ from those in MedLand see Appendix E.

A key difference between AgroSuccess and MedLand is the decision to exclude the soil erosion and deposition processes represented in MedLand from AgroSuccess. This is a direct reflection of the different research objectives of the two models. The authors of MedLand wanted to investigate the impact of Neolithic subsistence farming and pastoralism on land degradation via soil erosion in the Mediterranean. To do so they allow anthropogenic agents in the model to modify

land cover. They then map each land-cover state represented in the model to a quantity (called a 'C-factor') which determines the rate of soil erosion, i.e the degree to which different land cover types help to reduce the effects of erosion (Ullah, 2013, pp. 99–101). These C-factors are then fed into equations that enable MedLand simulations to represent how soil depth changes at different points in the landscape in response to anthropogenic land-cover change. Importantly, the tracking of land-cover state is an intermediate step to enable the modellers to calculate the dependent variable of principal interest—soil depth.

Conversely, the research questions that motivate the development of AgroSuccess do not explicitly concern soil erosion. I consider soil erosion to be a factor that is held constant while studying the effects of other factors—human agricultural practices, the fire regime, climate, and succession—on ecological succession and emergent land-cover state. This change of modelling focus justifies the exclusion of soil erosion calculations from AgroSuccess, and results in the need to modify other aspects of the model to maintain consistency. In particular, I have modified the MedLand equations used to calculate crop productivity because their MedLand equivalents depended on soil depth. See Appendix E.2 for details of this change.

Like AgroSuccess, MedLand incorporates a representation of spatial variation in land-cover using a raster grid (see Section 3.3). In both models the numerical value associated with the raster cell representing an area of the landscape during a particular time step specifies the type of land-cover at that location and time. In MedLand there are 38 different land-cover states a patch of land can occupy at each time. Raster cells progress sequentially through these states such that if a patch of 'bare land' is allowed to develop without disturbance for 50 simulated years, it will eventually transition into 'fully matured woodland' (Ullah, 2013, pp. 108–109).

The simple, linear representation of ecological succession in MedLand is inadequate to address my research questions. The purpose of AgroSuccess is to investigate the impact of various factors on emergent land-cover state, so I have developed an ecological succession submodel that makes it possible for various different succession pathways to occur in response to differing environmental conditions, as well as natural and anthropogenic disturbance. In contrast to the 38 land-cover states represented in MedLand, AgroSuccess focuses on 10 qualitatively distinct land-cover states (see Section 4.2.3). This is done to enable AgroSuccess to represent more detail about the states themselves and, crucially, the transitions between the states. The land-cover states in AgroSuccess are coarse-grained, making it possible to include more information in the

model about the *relationships* between states. In MedLand, land-cover state is calculated as an intermediate step during the calculation of changes to soil erosion depth, whereas in AgroSuccess land-cover state is the dependent variable of interest, motivating a shift in focus towards complex ecological succession pathways.

Model implementation

The model of subsistence agriculture within MedLand is coupled to the other model components using a framework called *DEVs Suite* which was developed in-house by the MedLand team. In the early stages of developing AgroSuccess I assessed the feasibility of integrating my work into DEVs Suite by reading the available source code and documentation, and contacting the authors of the model with queries which arose from our preliminary work. However, I found that the idiosyncrasies of the original implementation of MedLand were such that it would be implausible to build on it without establishing a close ongoing collaboration with the original authors. Due to the differences in focus between the projects underlying the AgroSuccess and MedLand models (see Section 4.3.2 above), I made the project management decision to develop AgroSuccess as an independent model without input from the team behind MedLand, rather than attempt to establish a collaboration with them.

Rather than develop the software for AgroSuccess completely from scratch, I decided to use an established agent-based modelling framework called Repast Simphony (North et al., 2013). Using a framework while developing a model provides the advantage of some initial structure into which modellers can insert their model-specific components, as well as a community of other modellers who can help to resolve problems with the software.

4.3.3 Overview of subsistence agriculture submodel

Here I present a high-level overview of the concepts entering into the subsistence agriculture submodel of AgroSuccess as recommended in the ODD Protocol (see Section 4.3.1).

Entities, state variables and scales

The most important entity in the subsistence agriculture submodel is the *household*. These are the decision-making agents that interact with the model environment (i.e. the simulated landscape) to affect land-use change. Each Household agent represents a family of prehistoric subsistence agriculturalists who cooperate with each other to obtain the resources they need to survive. Households have a *population* variable that tracks how many people live in the household, and a memory of the mass of wheat per hectare they were able to produce in the previous year. For ABMs other than MedLand that use the household as an anthropogenic decision-making unit see e.g. An et al., 2005; Clark and Crabtree, 2015.

Households are organised into *villages*, and each Household knows which Village it belongs to. Village agents have an immutable location variable that situates it in geographical space, and a mutable set of Household agents that belong to the village.

A Land Patch Allocator abstract agent represents the process by which land patches are allocated to households, and keeps track of which patches have been allocated to which household. It can be thought of as a model of the social contract according to which households recognise ownership of land patches by specific households.

In the preceding paragraphs I have made a distinction between ‘agents’ and ‘abstract agents’. By ‘agent’ I mean a concrete object that exists in the model and corresponds to something tangible in the world, whereas an ‘abstract agent’ encapsulates a conceptualisation of a mechanism or process that occurs in the world and which we want to represent in the model. Naming the abstract agents *as agents* simplifies the communication of the important mechanisms they represent, because within the semantics of agent-based modelling it is necessary for each action to have a *thing* to do it. If one were to avoid explicit discussion of abstract agents, it would become necessary to make concrete agents responsible for affecting processes or tracking variables in a way that doesn’t naturally correspond to the real-world entities they represent in the model. For example, I could have made the Village agent responsible for the process of allocating land patches to Household agents, and tracking which patches had been allocated to which household. This is contrived, because while a village can certainly *be located* somewhere, the claim that a village *allocated* something might signal the speaker was using the word ‘village’ in a figurative or otherwise non-standard way. Instead, I provided an abstract Land Patch Allocator to represent

this modelled process.

In Table 4.8 I show the exogenous drivers of the agent-based subsistence agriculture submodel. Note that with the exception of precipitation and temperature (the climatic boundary conditions), all of these factors are *outputs* of the biophysical model components described in Section 4.2. Precipitation drives the submodel directly through a mechanism that increases wheat yield on patches receiving more rainfall (and decreases yield on patches experiencing drought). Other exogenous drivers influence the subsistence agriculture submodel indirectly through their impact on ecological succession dynamics. For example, a patch of shrubland containing oak seeds with high soil moisture will transition into a patch of oak forest, wheres an otherwise similar patch with low soil moisture and pine seeds will transition into pine forest. Oak forest has a higher land cover conversion cost than pine forest (see Table 4.1 in Section 4.2.2) so higher soil moisture indirectly causes the patch to become more difficult for households to convert to agriculture.

Table 4.8: Exogenous drivers of the AgroSuccess subsistence agriculture submodel.

Driver	Direct	Indirect (fire regime)	Indirect (succession)
Precipitation	×	×	×
Temperature		×	×
Fire spread		×	×
Seed dispersal			×

The *temporal grain* of AgroSuccess is annual, and its *temporal extent* is 200–1000 years. This is consistent with temporal scale (Ullah, 2013) and grain (Barton et al., 2016) for which MedLanD has been used previously.

The *spatial grain* of AgroSuccess is approximately $25\,\mathrm{m}^2$ ($0.0025\,\mathrm{ha}$), and its *spatial extent* approximately $40\,\mathrm{km}^2$. Note that the spatial grain and extent will vary slightly depending on study site.

Process overview and scheduling

At the beginning of each simulation time step, each household decides on a *subsistence plan* that specifies:

1. The number of land cover patches they need to farm in the coming simulated year to produce enough calories to support the members of the household, $N_t^{(w)}$.

2. The number of land cover patches they need to use to gather sufficient firewood to meet their heating and cooking needs, $N_t^{(f)}$.

The larger a household is, the more labour they are able to dedicate to cultivating land patches and, correspondingly, the more calories they need to generate in order to survive. Each household's objective is to harvest enough food and firewood to survive each simulated year.

Each Village agent sorts the land cover patches in its vicinity in order of their suitability for both wheat farming and wood gathering. A patch's suitability with respect to farming or firewood gathering can change from one time step to the next, because different land cover types have different land-cover suitability attributes and the land-cover type of each simulation grid cell will change throughout the simulation run. All households within the same village value land patches the same as each other, but households in different villages will favour land patches that are closer to their own village's location.

Once all households have determined their subsistence plans and the households have appraised the grid cells in the vicinity of their village for wheat farming and firewood gathering, the Land Patch Allocator abstract agent determines which land patches each household is allocated to farm for the upcoming simulated year. Land patches are allocated by randomly iterating through households until all households have satisfied their subsistence plans. Households do not retain land patches between simulated years. Instead all patches are reallocated at the beginning of each year. This is intended to reflect an egalitarian society in which all households get an opportunity to farm the highest value land patches. See Section 4.3.4 for details of the land patch allocation algorithm.

Finally, Households calculate their wheat and firewood yields for the current year's subsistence activity and report this value to their respective Population Update Manager abstract agents. In the version of AgroSuccess presented in this thesis, the populations of households remain constant throughout the simulation run.

This completes the steps required of the subsistence agriculture submodel for the current time step. The simulated year ends with the calculation of soil moisture and the simulation of seed dispersal, ecological succession, fire ignition and spread. See Section 4.2 for details of these processes. See Fig. 4.4 for an overview of the simulation process sequence. For details of the rules by which agents decide their subsistence plans and express preference for some land patches

over others, see equations in Section 4.3.4.

4.3.4 Details

Households determine wheat subsistence plan

To determine their subsistence plan, all households calculate the number of land patches they will need to farm to satisfy the calorie requirements of their members. The number of wheat plots required by a household in time step t is given by

$$N_t^{(w)} = \frac{3.65 \times 10^6 \cdot E_{\text{tot}}\, P_t^{(h)}\, (1 + p_s)}{E^{(w)}\, \mu_{t-1}^{(w)}\, A_r\, C} \tag{4.18}$$

where $P_t^{(h)}$ is the number of members of the household at time t, p_s is the proportion of the crop households hold back to reseed the next year, $E^{(w)}$ is the energy household members can extract from $1\,\text{kg}$ of wheat, $\mu_{t-1}^{(w)}$ is the wheat yield in kg/ha the household obtained in year $t-1$, and C is the farmer conservativeness scalar. The factor 3.65×10^6 accounts for how E_{tot} is expressed in terms of d rather than yr, and how $\mu^{(w)}$ is expressed in terms of ha rather than m^2. See Table 4.9 and Table 4.10 for an overview of these parameters and variables, and Appendix E.1 for details of how this equation differs from its counterpart in MedLand.

In addition to the desired number of wheat patches, households also consider whether or not they are able to supply the amount of labour required to farm this quantity of wheat. This is done by calculating

$$L_{\text{h}}(t) = P_t^{(h)} L_{\text{pers}} \tag{4.19}$$

and

$$L_{\text{h}}^{(w)}(t) = 10^{-4} N_t^{(w)} A_r L^{(w)} \tag{4.20}$$

Here $L_{\text{h}}(t)$ is the labour the household is able to supply in the current time step, and $L_{\text{h}}^{(w)}(t)$ is

the labour that would be required to farm $N_t^{(w)}$ plots of wheat (note that $10^{-4}A_r$ is the raster cell area in ha). If $L_{\mathrm{h}}^{(w)}(t) \leq L_{\mathrm{h}}(t)$ then $N_t^{(w)}$ will be requested. Otherwise, the maximum number of patches the household can farm will be requested.

The parameters entering into Eq. (4.18) are summarised in Table 4.9, and relevant endogenous variables are shown in Table 4.10.

Table 4.9: Symbols representing parameters used in subsistence plan and yield equations. Unless otherwise cited, values are derived from Ullah, 2013, p. 160.

Symbol	Value	Description
E_{tot}	$2500\,\mathrm{kcal}/(\mathrm{pers\,d})$	Total number of kilocalories needed per person per day (Ullah, 2013, p. 155)
$E^{(w)}$	$3540\,\mathrm{kcal/kg}$	Energy yield per kilogram of wheat
p_s	0.15	Proportion of crop held back to seed next year's plots
$\mu_0^{(w)}$	$3500\,\mathrm{kg/ha}$	Initial mass of wheat grown per hectare, assuming maximum possible yield (Ullah, 2013, p. 162)
C	0.75	Expectation scalar determining farmer conservativeness (Ullah, 2013, p. 162)
L_{pers}	$300\,\mathrm{pers\,d/yr}$	Total labour availability per person per year (McCall, 1985)
$L^{(w)}$	$50\,\mathrm{pers\,d}/(\mathrm{ha\,yr})$	Labour requirement for wheat (Ullah, 2013, p. 160) agriculture
A_r	$\approx 625\,\mathrm{m}^2$	Area of each raster grid cell (exact value is study site dependent)

Table 4.10: Endogenous variables that occur within running models which are involved in calculating subsistence plans.

Symbol	Units	Description
$P_t^{(h)}$	pers	Population of household at time t
$\mu_t^{(w)}$	kg/ha	Mass of wheat per ha grown by household in year t

Households determine wood gathering plan

Households determine a number of land patches to use to gather firewood that they will use for both cooking and heating. Like the patches they use for wheat agriculture, in each time step households request a set of patches to use to gather firewood that only they will use for the simulated year. However, unlike the case of patches used for agriculture, households do not modify the land-cover type of patches used for firewood gathering. Instead, I assume that households gather wood from patches of one of the mature land-cover types at a *sustainable*

intensity, such that the land-cover type doesn't change. The mature land-cover types are: Pine, Transition Forest, Deciduous, and Oak (see Section 4.2.2). This process relies on the following assumptions about the household agents:

1. Households are aware of what constitutes a sustainable harvest.

2. Households choose to harvest sustainably even at the cost of travelling further from the village, rather than gathering from the closest available patches.

I define a sustainable harvest as the removal of $10\,\%$ of biomass from a mature forest land patch (Grabher, 2021; Schulze et al., 2012). Further, I assume the maximum biomass density of each mature forest type to be $300\,\mathrm{t/ha}$ (Schumacher et al., 2004, p. 189). This was estimated by inspecting yield tables for large trees from Italian forests (see also Schumacher et al., 2004 supplementary materials). By definition, patches occupied by mature forest land-cover have maximum biomass.

To estimate the mass of firewood required per year by households in AgroSuccess, I assume that neolithic households would have consumed firewood at a similar rate to modern communities that use firewood for subsistence. This is necessary because I am not aware of any studies that have investigated firewood use in the Neolithic period. Kirkland et al., 2007; Twine, 2003 estimate firewood requirement per person as $1\text{–}1.1\,\mathrm{t/pers/yr}$.

Using these assumptions, the number of wood gathering patches required by households in AgroSuccess are calculated as

$$N_t^{(f)} = \left\lceil \frac{P_t^{(h)} m_f}{r D A_r} \right\rceil \tag{4.21}$$

Here m_f is the mass of wood required per person per year, r is the proportion of the total biomass to be removed each year, and D is the climax forest biomass density. For example, a household of 5 using the parameter values described above and summarised in Table 4.11 would require $N_=^{(f)} 3$ patches per year.

Table 4.11: Symbols representing parameters used in the firewood gathering plan equation.

Symbol	Value	Description
m_f	$1.1\,\mathrm{t/pers/yr}$	Firewood requirement per capita per year (Kirkland et al., 2007; Twine, 2003)
D	$300\,\mathrm{t/ha}$	Climax forest biomass density (Schumacher et al., 2004, pp. 189)
r	0.1	Proportion of biomass per unit area to remove when gathering firewood
A_r	$\approx 625\,\mathrm{m}^2$	Area of each raster grid cell (exact value is study site dependent)
$P^{(h)}$	$5\,\mathrm{pers}$	Household population

Households in each village appraise farm plots

When appraising land patches in the vicinity of their village, Household agents favour land patches that are close to their village, are relatively flat, have fertile soil, and are currently occupied by a land cover type that is easy to convert to wheat agriculture. Here I describe how these factors are represented as model quantities in AgroSuccess.

To quantify the distance between a village and a given land patch it is necessary to specify a distance metric. In AgroSuccess I use the euclidean distance between grid cells

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{4.22}$$

where D_{ij} is the euclidean distance between cell i and the village whose center is located in cell j. Here x_k and y_k are the 2-dimensional Cartesian coordinates of the center of cell k. It is important that the units of these coordinates are distances as opposed to angular units used in geographic coordinate systems.

When comparing land patches, households need to be able to evaluate the relative distance between different land patches and their village. To facilitate this I introduce a *distance factor* defined as

$$\tilde{D}_{ij} =: \frac{D_{ij}}{\sup_i D_{ij}} \tag{4.23}$$

where $0 \leq \tilde{D}_{ij} \leq 1$ is a distance measure between a grid cell i and the center of the village located a j. The denominator in Eq. (4.23) is the maximum distance it is possible to travel from the village j to any other grid cell in the simulated landscape. If cell k is the furthest cell in the simulated landscape from the village in cell j then $\tilde{D}_{kj} = 1$ and $\tilde{D}_{jj} = 0$. The use of $\tilde{D}_{ij}$ rather than the raw distance D_{ij} in the land patch evaluation mechanism has the advantage of giving grid cells a dimensionless distance 'score' that can be weighted in comparison with other factors determining land-cover value to give grid cells and overall value score.

In addition to considering land patches' proximity to their village, households preferentially select flat land patches to use for wheat agriculture. The spatially varying slope of the land is provided as a raster data set for each simulation run (see Section 3.3.1). Following Ullah, 2013 I assign each land patch, i, a slope modification value SV_i that depends on its slope (percentage grade) such that steeper slopes are given a smaller slope modification value. See Table 4.12 for the mapping of slope to slope modification value used. This approach to mapping slope ranges to discrete classes is analogous to that used to classify slopes into different slope risk classes in the fire spread algorithm described in Section 4.2.6, as well as in e.g. J. D. A. Millington et al., 2009; Perry and Enright, 2002.

Table 4.12: Slope modification value for grid cell i, SV_i given a slope (grade) of $S_i^{(\%)}$. The classification boundary values of $0\,\%$, $18\,\%$, $36\,\%$, and $173\,\%$ correspond to $0°$, $10°$, $20°$, and $60°$ respectively.

Slope, $S_i^{(\%)}$ [%]	Slope modification value, SV_i
$0 \leq S_i^{(\%)} < 18$	1.00
$18 \leq S_i^{(\%)} < 36$	0.75
$36 \leq S_i^{(\%)} < 173$	0.25
$S_i^{(\%)} > 173$	0.00

The remaining attributes of land patches that household agents consider when determining which ones to convert to agriculture are their fertility and land cover conversion cost scores. These are intrinsically linked to the land cover type of each land patch at the time the household is evaluating them. See Section 4.2.2 for details of the fertility and land cover conversion cost scores assigned to each land cover type represented in the model.

In summary, agents consider the slope, distance from village, land cover fertility and land cover conversion cost when selecting land patches to convert to agriculture. The relative importance of distance, land cover fertility, and land cover conversion cost in household decision making is controlled by three dimensionless model parameters: $\delta > 0$ specifies the relative importance of a cell's distance to the center of the village, $\phi > 0$ specifies the relative importance of soil fertility,

and $\lambda > 0$ specifies the relative importance of land cover conversion cost.

For a household belonging to a village centered in cell j, the farming value of a land patch in cell i is given by:

$$FV_i^{(j)}(\tau;\, \delta, \phi, \lambda) = SV_i \cdot \phi\, F_i(\tau) - \delta \tilde{D}_{ij} - \lambda\, C_i \tau \tag{4.24}$$

Table 4.13 summarises the parameters that control the relative influence of different factors in households' decision to select particular land patches for farming purposes.

Table 4.13: Parameters which control the degree to which fertility, cost to convert to agriculture, and distance from their village influence households' evaluation of land patches. All parameters are dimensionless.

Symbol	Description
ϕ	Soil fertility decision weighting
λ	Land conversion cost decision weighting
δ	Patch distance decision weighting

Households in each village appraise wood-gathering plots

The value that all households in a village assign to a land patch for the purpose of gathering firewood is given in Eq. (4.25). Here $WV_i^{(j)}(\tau;\, \delta)$ is the wood value given to patch i at time τ by households in a village centred in cell j. $\tilde{D}_i^{(j)}$ has the same meaning as is defined in Eq. (4.23). The meaning of δ is given in Table 4.13. $WV_i(\tau) = WV_i(LC_i(\tau))$ is the value of a patch of land i at time τ for gathering firewood. $IWV_i(\tau) = IWV_i(LC_i(\tau))$, intrinsic wood value, is the value attached to a patch of land has for firewood collecting purposes. This depends only on the land cover class at that point. The intrinsic wood value for each land-cover type, W_{L_c} is specified in Table 4.1.

$$WV_i^{(j)}(\tau;\, \delta) = \frac{1}{1+\delta}\left[IWV_i(\tau) + \delta\left(1 - \tilde{D}_i^{(j)}\right)\right] \tag{4.25}$$

I assume that each household selects the patches its occupants will harvest for firewood each year independently and in isolation from every other household in the village. It is plausible that, in reality, households within a village would cooperate to designate an area of land close to the

village to use communally for wood-gathering purposes. However, to simplify the behavioural rules required in AgroSuccess, I assume that households gather firewood independently of each other.

Allocate land patches to agents

Once each household has determined how many land patches it requires to meet its subsistence and firewood gathering needs for the year, it is necessary to decide *which* patches will be allocated to each household. In general, agents want to minimise the distance from their village that they will need to travel to access the land they manage. When choosing land patches for agriculture, agents favour flatter land that has fertile soil and whose current land cover can be converted to agriculture with little effort. When choosing land patches for firewood gathering, agents prefer land patches with forest land-cover types.

Following Ullah, 2013, p. 91 this process is carried out by the land patch allocator abstract agent. The land patch allocator abstract agent follows a 'round-robin' schedule to decide the order in which agents are allowed to request land patches. Each time step a random ordering of households is generated within each village. Then, drawing villages at random, AgroSuccess loops through households in each village allowing them to choose one land cover patch to manage for the year in each iteration of the loop until all households have satisfied their subsistence and firewood gathering plans, before moving on to the next village.

Each time a household is given the opportunity to claim a land-cover patch, they must decide whether to claim a patch for agricultural or firewood gathering use. They do this by calculating the proportions of the planned numbers of agricultural and wood gathering plots they require for their subsistence plan that are already satisfied. They then prioritise selecting a patch to satisfy the resource type for which the smallest proportion is already satisfied. In the event or a tie, a resource type is selected to prioritise randomly with equal probability.

Households do not retain the same land patches from one year to the next, and all patches are available for agents to request at the start of each simulated year. This modelling approach is justified by my modelling objectives and assumptions:

1. The purpose of AgroSuccess is to explore the emergent impact of anthropogenic land-cover

change (among other factors) on landscape evolution. While it is important to record *which land patches* are converted to agriculture, my modelling objectives do not depend on knowledge of *which households* control those patches.

2. The procedure of randomly generating the order in which villages and households are allowed to express preference for land-cover patches reflects a modelling assumption that the modelled households constitute an egalitarian society (Ullah, 2013, p. 91).

If a household is allocated a land patch whose land cover type is not already Wheat, the household will convert the cover type of that land patch from whatever its original state was to wheat agriculture. Households' use of land patches for firewood gathering is assumed to leave the land-cover type of the patch unchanged.

Calculate wheat gathering returns

The approach to calculating farming returns follows Ullah, 2013, pp. 94–95. Wheat returns in kg for grid cell i are calculated using Eq. (4.26). The parameters in this equation are: wheat production rate for grid cell i, $r_i^{(w)} = r_i^{(w)}(P, F_i) \in [0, 1]$ (see Eq. (4.27)), the slope modification value in cell i, SV_i (see Table 4.12), maximum wheat yield in $\mathrm{kg/ha}$ under ideal conditions, $M_{\max}^{(w)}$, and raster cell area in m^2, A_r, such that $A_r/10,000$ gives the area of each raster cell in ha.

$$R_i^{(w)} = \frac{r_i^{(w)} \cdot SV_i \cdot M_{\max}^{(w)} \cdot A_r}{10,000} \tag{4.26}$$

The expression for $r_i^{(w)}(P, F_i)$ is given in Eq. (4.27). This is modified from Eq. (4.10) in Ullah, 2013, p. 95, and differs from the original authors' approach by dropping dependence on soil depth. This is done because, unlike MedLanD, AgroSuccess does not model soil depth. Here P is annual precipitation in mm and $F_i \in [0, 5]$ is the land cover fertility value of the cell's land cover type. $F_i = 5$ implies cell i is maximally fertile.

$$r_i^{(w)} = \frac{(0.51 \ln(P/1000) + 1.03) + (0.19 \ln(F_i/5) + 1)}{2} \tag{4.27}$$

Calculate wood gathering returns

The approach to the calculation of the mass of firewood collected from each grid cell follows Ullah, 2013, p. 96. The mass of fire wood returned from raster cell i in kg is given by

$$R_i^{(f)} = I^{(f)} \cdot A_r \tag{4.28}$$

where $I^{(f)}$ is the firewood gathering intensity in $\mathrm{kg/m^2}$, and A_r is the area of each raster cell in $\mathrm{m^2}$. Note that Eq. (4.28) assumes that the land cover type at cell i contains woody vegetation.

Household population

In AgroSuccess, household populations are held at a fixed size of six members. This is consistent with the average Neolithic household size determined by Ullah, 2013. The overall human population in a simulated landscape is controlled by modifying a parameter controlling the number of households in the village, N_h. This is a simple model of population size, but still allows users of AgroSuccess to experiment with the impact of different human population sizes on land cover change by running simulated scenarios with different values of N_h.

I experimented with the use of a *dynamic* model of household population change similar to that described by Ullah, 2013. In this model, household populations changed over time as members gave birth and died. The birth and death rates of each household were influenced by the extent to which the household was able to satisfy the calorie requirements of its members through its farming activity. The number of land patches that a household could farm was limited by their labour availability in comparison to the labour required to farm the desired number of wheat land patches (see Eq. (4.19) and Eq. (4.20)). However, I found that this model produced uncontrolled population growth, and implausibly large proportions of the landscape being converted to wheat agriculture. This can be explained by the fact that under this model, each household member contributes both additional labour availability and the possibility of reproducing. In the absence of any negative feedback, it is reasonable to expect exponential population growth. It is possible that a more sophisticated model of population change that produces realistic outcomes could be developed, for example by representing an ageing process that prevents some members from reproducing or contributing to farming effort. However, such a model would require several

additional parameters and assumptions, and is not essential to allow AgroSuccess to address my research objectives.

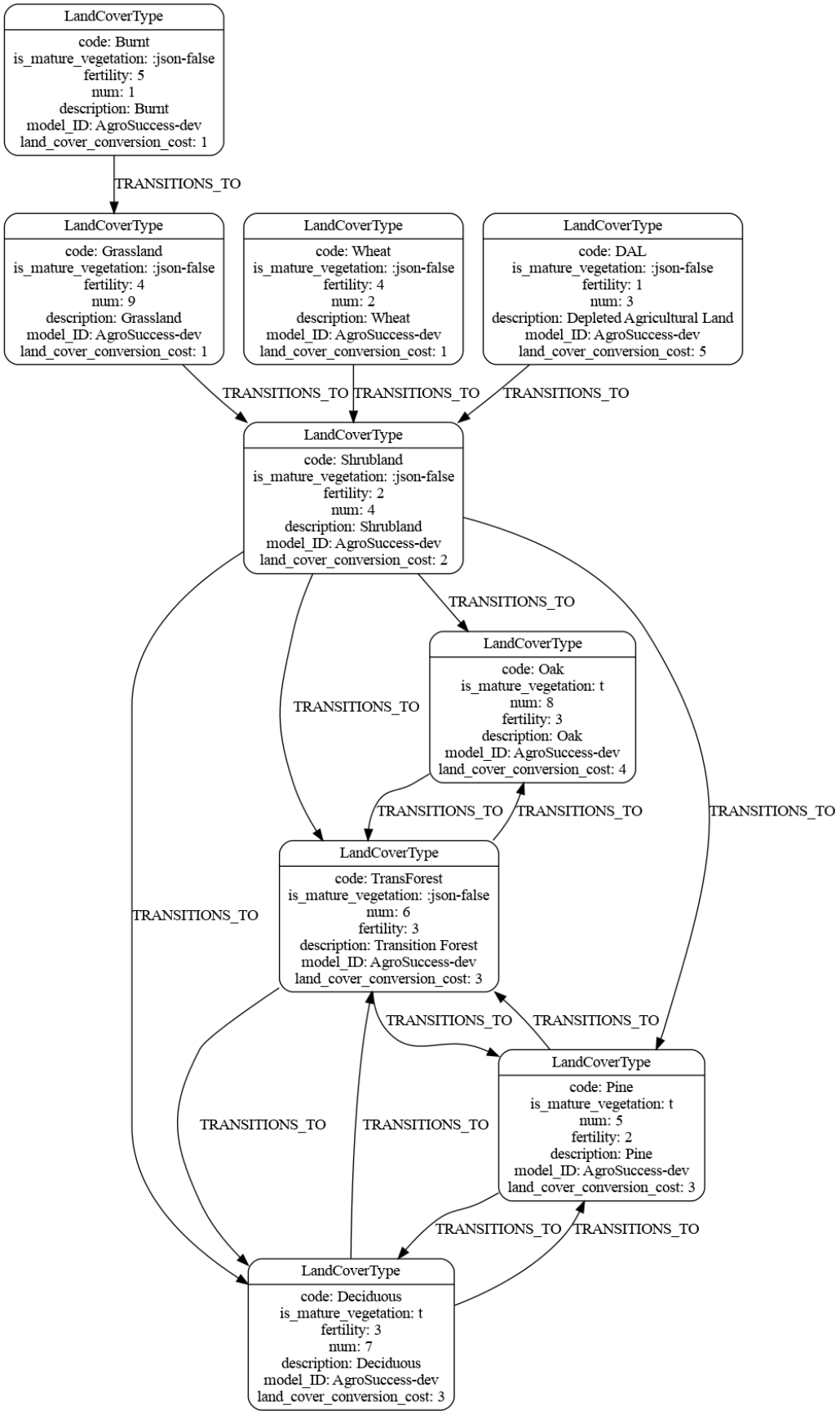


Figure 4.1: States and transitions represented in AgroSuccess. Self-links show land-cover states with combinations of environmental conditions that result in a land patch remaining in the indicated state until the environmental conditions change, or until the patch is subject to an external disturbance (anthropogenic or fire). This figure was automatically generated from a Neo4j graph database containing the AgroSuccess succession model using the Graphviz package.

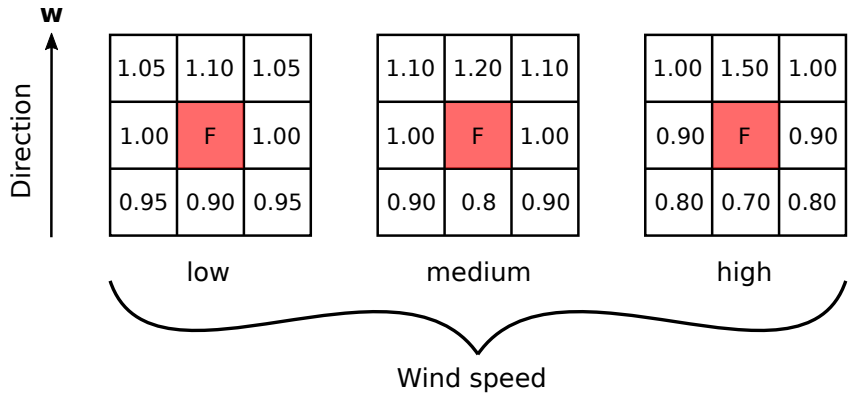


Figure 4.2: Wind risk factors, WR_i, for grid cells neighbouring an active fire. The risk of a fire spreading from a cell containing a fire (indicated by an 'F') to one of its neighbours depends on both wind speed and the angle between the path from the fire to the neighbouring cell and the wind direction.

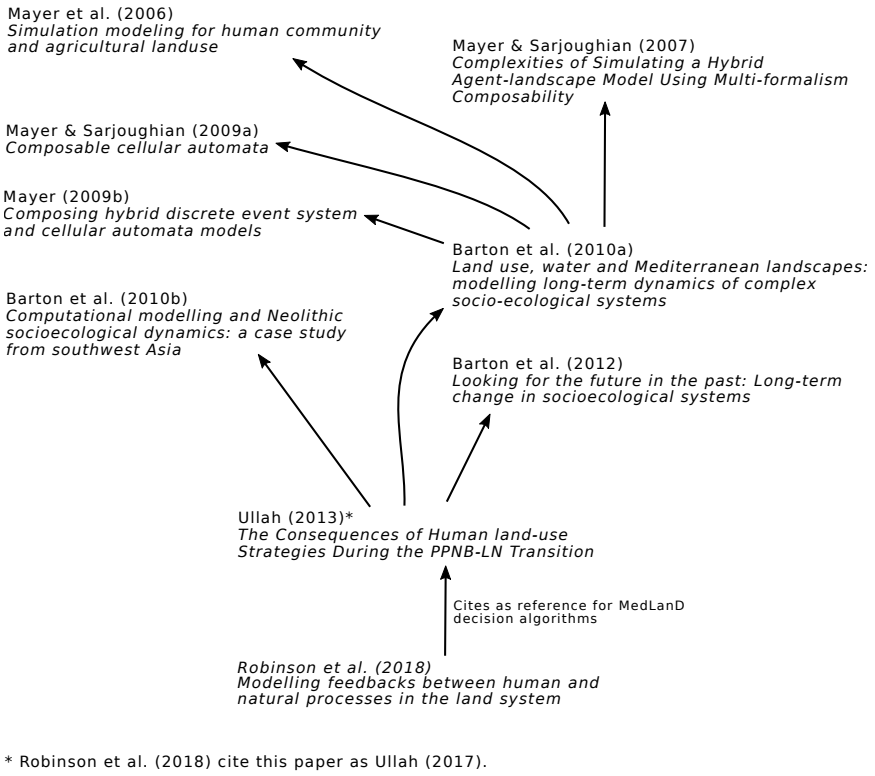


Figure 4.3: Citation relationships between papers describing the evolution of the MedLanD model

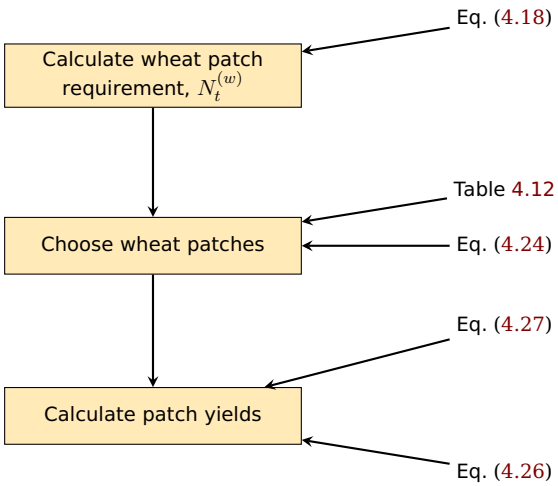


Figure 4.4: Process overview of the household agent decision-making process within each time step.

Chapter 5

AgroSuccess model calibration

5.1 Wildfire submodel calibration

The AgroSuccess model includes two parameters that directly influence the action of the wildfire ignition and spread model. These are the mean number of fires that occur in the simulation grid per simulated year, λ (fires/year), and the land-cover flammability scenario, S_{LCF}. Here I describe my strategy for determining the most appropriate values for these parameters for each of the study sites considered in this thesis. I present the results from simulated experiments conducted to determine these values, and discuss the limitations of our approach. I argue that the limitations identified during the course of this analysis motivate and justify ongoing efforts to improve access to high frequency sedimentary charcoal data sets.

5.1.1 Wildfire submodel parameters

The mean number of fires to take place in the simulation grid per simulated year, λ, is treated as a boundary condition and determined for each study site using empirical data (see Table 5.1). The land-cover flammability scenario, S_{LCF}, is a discrete parameter that controls the likelihood of an active fire spreading to a neighbouring simulation cell, in consideration of the neighbouring cell's land-cover type. The specific land cover flammability values for each land cover type within all considered land cover flammability scenarios are given in Table 4.5. The value of S_{LCF} is selected for each study site by calibration against empirical reference data that characterise the wildfire regime at each study site.

5.1.2 Strategy for calibrating the wildfire submodel

The wildfire regime of a given region over a period of time can be characterised by analysing the frequency-area distribution of wildfires and the average burned area per year in that region. Previous authors (Malamud et al., 2005; M. V. Moreno et al., 2011) have found that the distributions of wildfire burnt areas derived from empirical data sets are well-modelled by power-laws. These are a family of probability distributions with cumulative distribution functions of the form

$$F_X(x) = F_X(x; \beta, x_{\min}) = \left(\frac{x}{x_{\min}} \right)^{-\beta+1} . \tag{5.1}$$

Here $x_{\min} > 0$ is a parameter that sets area of the smallest fire that can be represented by the statistical model (Clauset et al., 2009). It is a normalisation constant that emerges from the assumption of a distribution with a PDF of the form $f(x) \propto x^{-\beta}$. β is the frequency-area distribution power-law exponent. It quantifies the relative likelihood of large fires compared to small ones. If $\beta = 0$ large and small fires are equally likely. As β increases large fires become rarer with respect to small fires (Malamud et al., 2005). Successfully calibrated AgroSuccess models will produce simulated wildfire regimes whose wildfire frequency-area distributions have similar power-law exponents and average burned area per year to those that have been observed empirically in the areas the simulations represent.

Wildfire regime reference data

I characterise the fire regime empirically for each study site using three statistics: the wildfire frequency-size distribution power-law exponent parameter, β, the average burned area per year per hectare, and the fire frequency. To estimate these statistics for the study sites in Spain, I use the work of M. V. Moreno et al., 2011. They analysed a database of wildfires in Spain spanning the time period 1988-2007 and report derived wildfire frequency-area statistics for the following climate regions:

- Mediterranean Continental
- Mediterranean Mountain
- Mediterranean Coast

- Oceanic
- Mountain

By overlaying the locations of the study sites identified in Chapter 3 on the map of climate region boundaries that M. V. Moreno et al., 2011 identified, I associated Navarrés, San Rafael and Algendar with the Mediterranean Coast climate region, and Atxuri and Monte Areo mire with the Oceanic climate region.

The Charco da Candieira study site is located in Portugal, and is sufficiently close to the Atlantic that I did not believe it was reasonable to assume it would be in the same climate region as any of the Spanish study sites. Seeking a data-driven approach, I obtained fire inventory data for the parish containing Charco da Candieira from the Portuguese Rural Fire Database (Pereira et al., 2011). I fitted a power-law frequency-size distribution to this data to estimate a value of β, and similarly calculated an estimate for the burned area per year and fire frequency in the area surrounding Charco da Candieira. The fire regime reference data for all study sites is given in Table 5.1.

Table 5.1: Power-law exponent parameter, burned area density, and fire frequency density values expected at each study site. The values for the study sites located in Spain are derived from the wildfire frequency-area statistics reported by M. V. Moreno et al., 2011. I estimated the corresponding values for the Charco da Candieira study site–which is located in Portugal–using data from the Portuguese Rural Fire Database (Pereira et al., 2011).

Study site	Power-law parameter, β	Burned area [$\mathrm{ha\,ha^{-1}\,yr^{-1}}$]	Fire frequency [$\mathrm{yr^{-1}\,ha^{-1}}$]
Algendar	1.59 ± 0.25	4.65×10^{-3}	2.693×10^{-4}
Atxuri	1.99 ± 0.28	11.3×10^{-3}	1.616×10^{-3}
Charco da Candieira	1.69	27.5×10^{-3}	9.384×10^{-4}
Monte Areo mire	1.99 ± 0.28	11.3×10^{-3}	1.616×10^{-3}
Navarrés	1.59 ± 0.25	4.65×10^{-3}	2.693×10^{-4}
San Rafael	1.59 ± 0.25	4.65×10^{-3}	2.693×10^{-4}

In the following sections, the power-law parameter and burned area statistics listed in Table 5.1 are used as calibration targets. That is, simulations producing similar statistics are judged as accurately reproducing the empirically observed fire regime. The mean number of fires per year model parameter, λ, for each study site is calculated by multiplying the fire frequency values in Table 5.1 by the total simulation grid area.

A limitation that any attempt to model Iberian fire regimes in the mid-Holocene is likely to encounter is the lack of availability of data to quantify fire regimes at that time. Because of the great extent to which the land cover of the Iberian Peninsula has been modified by human activity over the intervening millennia, there are no 'pristine' landscapes (Redman, 1999) in the study

region that we can use as contemporary proxies to infer fire regime characteristics during the mid-Holocene. This situation could be improved by advances in the analysis of sedimentary charcoal to reconstruct past fire regimes (Whitlock & Larsen, 2001).

5.1.3 Specification of simulated experiments

To focus on the fire spread dynamics during fire model calibration, I held all parameters except the land cover flammability scenario, S_{LCF}, constant. Anthropogenic agents were excluded, and all other parameters were kept at their default values, as described in Section 6.1. To reduce the impact of edge effects (e.g. fires appearing smaller than they would have been because they spread to the edge of the simulation grid), only fires that started within each study site's *experimental zone* were included in subsequent data analysis. See Section 3.3.1 for details of how experimental zones are defined.

I ran 10 calibration scenarios, one for each land cover flammability scenario with mean land cover flammability ranging from 0.14–0.23 (scenarios TFN4–Default in Table 4.5). For each calibration scenario I ran 10 simulation replicas. All simulations were run for 400 simulated years.

5.1.4 Calibration results

Calibration against power-law exponent parameter, β

To calculate the power-law exponent parameters that characterise the fire regimes produced by the calibration scenarios, it is first necessary to evaluate the range of fire sizes that can be plausibly modelled by a power law in the calibration scenario simulation outputs. Fig. 5.1 shows the empirical survival function for all fires observed in all calibration scenarios for each study site. If the distribution of fire sizes in these scenarios was well modelled by a power-law distribution, these functions could be approximated by a straight line on log-log axes. Here we see that this can be achieved for Algendar, Navarres, and San Rafael for fire sizes of up to approximately $6000\,\mathrm{ha}$. The roll-off for fires larger than $6000\,\mathrm{ha}$ is likely to be an edge effect resulting from the finite size of the simulation grid. However, the results for Atxuri, Charco da Candieira, and Monte Areo mire show considerable curvature throughout the range of fire sizes observed in their calibration scenarios simulations. This is evidence that the fire frequency-size distributions

produced by AgroSuccess may not be well modelled by a power-law for all study sites of interest.

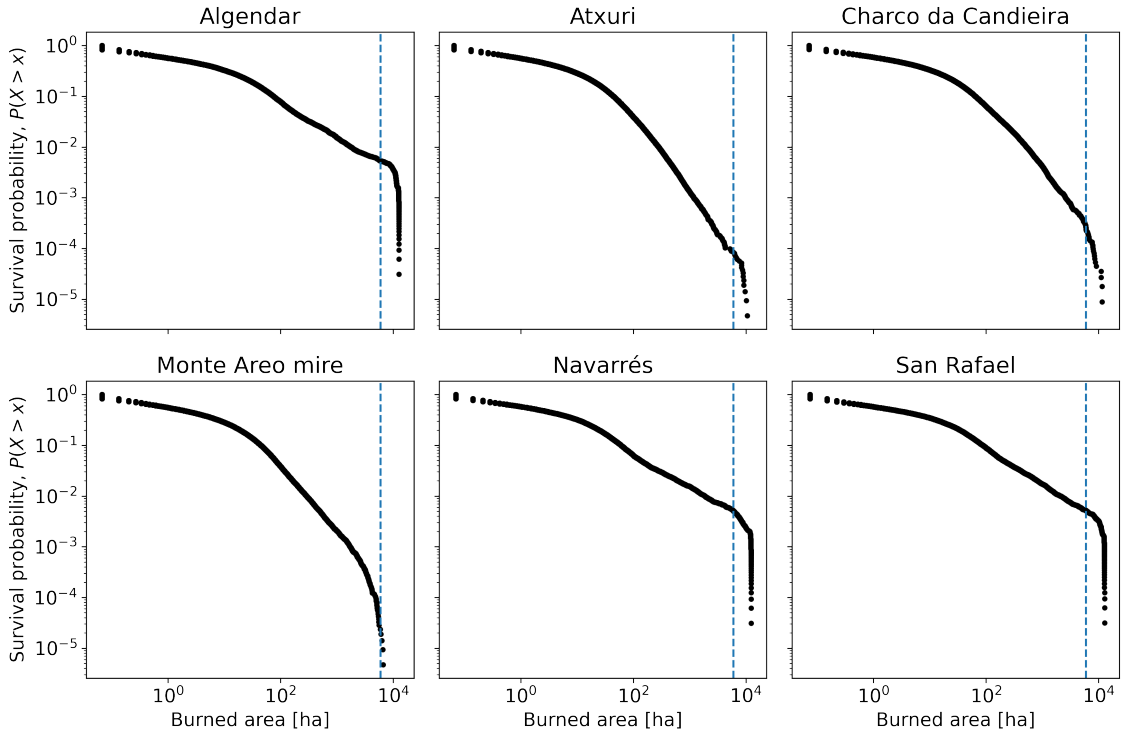


Figure 5.1: Empirical survival function, $F_X(x)$, for all simulated fires within the experimental zone for each study site. Dotted blue lines indicate the $6000\,\mathrm{ha}$ cut-off below which wildfire frequency-size can be modelled with a power-law. When plotted on log-log axes, the survival function for a power-law distributed variable appears as a straight line.

Based on the considerations above, I took the maximum fire size that can be modelled by a power-law distribution to be $6000\,\mathrm{ha}$, and excluded larger fires from subsequent analysis. I calculated an estimate of β for each of the calibration scenario simulations. For a set of fire size observations, $\boldsymbol{x}$, an estimate of β can be obtained using the following expression for the maximum likelihood estimator (Clauset et al., 2009).

$$\hat{\beta} = 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}} \right]^{-1} \tag{5.2}$$

where $\hat{\beta}$ is the estimate of β given the data derived from the simulations, and n is the number of samples (i.e. the number of simulated fires included in the analysis for a given simulation). As in Eq. (5.1), $x_{\min}$ is the area of the smallest fire that is included in the power-law model. I set $x_{\min} = 0.074\,\mathrm{ha}$ corresponding to the area of a single simulation cell, which is the smallest fire that can occur in an AgroSuccess simulation model.

The results in Fig. 5.2 show the spread of $\hat{\beta}$ values obtained for each land cover flammability

scenario across the 10 simulation replicas performed for each scenario. As noted above, as β increases, the likelihood of large fires decreases relative to small fires assuming the underlying data follow a power-law distribution. The results in Fig. 5.2 appear to show that as the average land cover flammability increases, estimates of β *also* increase. That is, increasing land cover flammability decreases the likelihood of large fires. This result is contrary to the logic specified in the wildfire spread model (see Section 4.2) and is explained by a power-law being an inappropriate model of the fire frequency-size distributions seen in the calibration scenario simulations. This conclusion is also supported by the observation of curvature in the empirical survival function in three of the six study sites seen in Fig. 5.1. The fact that AgroSuccess fire sizes do not follow a power-law is an interesting finding, but forces me to conclude that it is not possible to calibrate the land cover flammability scenario parameter by comparing simulated values to empirically derived values for β.

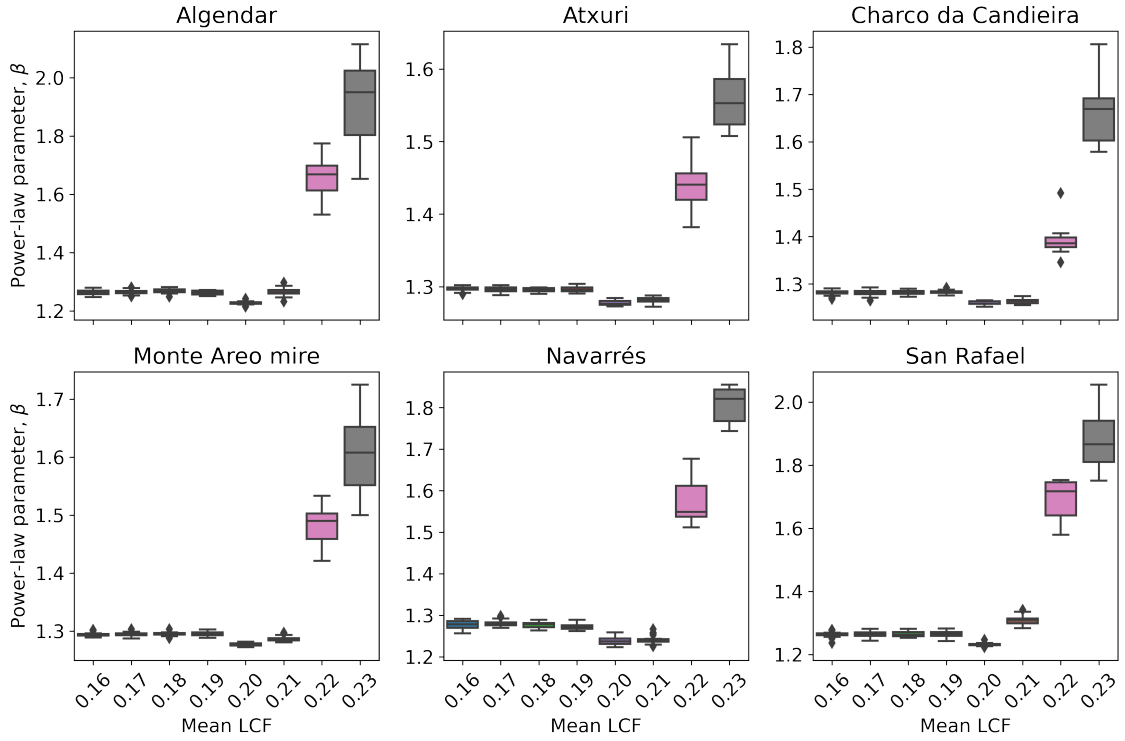


Figure 5.2: Best fit power-law exponent parameters determined by maximum likelihood for different land-cover flammability scenarios. Boxes show the interquartile range (IQR) of values for β estimated from 10 simulation runs of 400 time steps for each LCF scenario. Whiskers extend to 150% of the IQR, and diamonds indicate outliers.

Calibration against burnt area

To calibrate the land cover flammability scenario for each study site against burnt area, I calculated the average burnt area per hectare per year from fires starting within the experimental zone

for each calibration scenario simulation. Fig. 5.3 shows the simulated burnt area per hectare per year obtained by averaging over 10 simulations for each study site. The empirically derived 'target' burnt area values from Table 5.1 are shown as dotted red lines. I find average land-cover flammability of 0.20, corresponding to land-cover scenario TF2, is the optimal land-cover flammability for Charco da Candieira. For all other study sites the optimal land-cover flammability scenario is TF1 (with mean land-cover flammability of 0.19). I attempted to produce land-cover flammability scenarios with lower burned area per hectare per year than was produced by TF1 (TFN4–TF0, see Table 4.5). However, I found that land-cover flammability scenarios with smaller average land-cover flammability than 0.19 do not produce smaller simulated burned areas. This is a notable limitation of the wildfire spread model. Consequently the average burned areas in simulations of the Monte Areo mire and Atxuri study sites are expected to slightly exceed the corresponding empirical reference values.

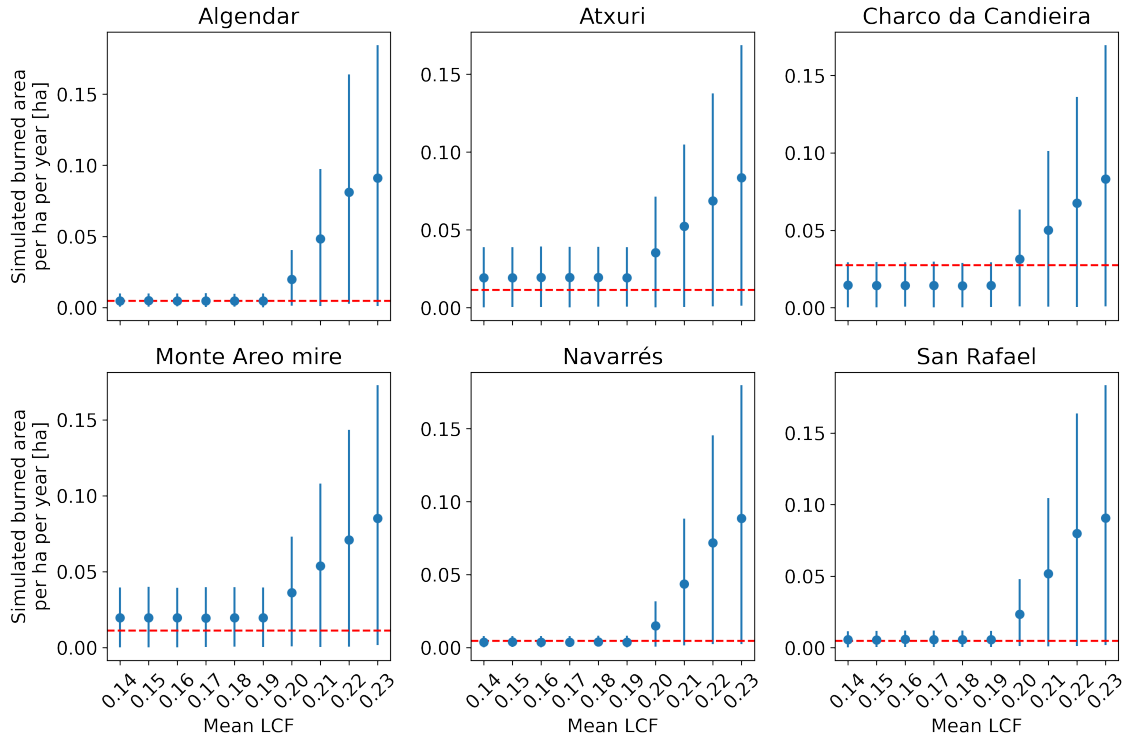


Figure 5.3: simulated burned area per hectare per year observed in simulations for different land-cover flammability scenarios. The dotted red lines indicate the empirical reference values expected for each study site. Plotted points are median values for burned area per hectare per year across 10 simulations per land-cover flammability scenario for all study sites. Error bars indicate 95% confidence intervals.

Fig. 5.4 shows examples of simulated landscapes generated from calibrated models for each study site.

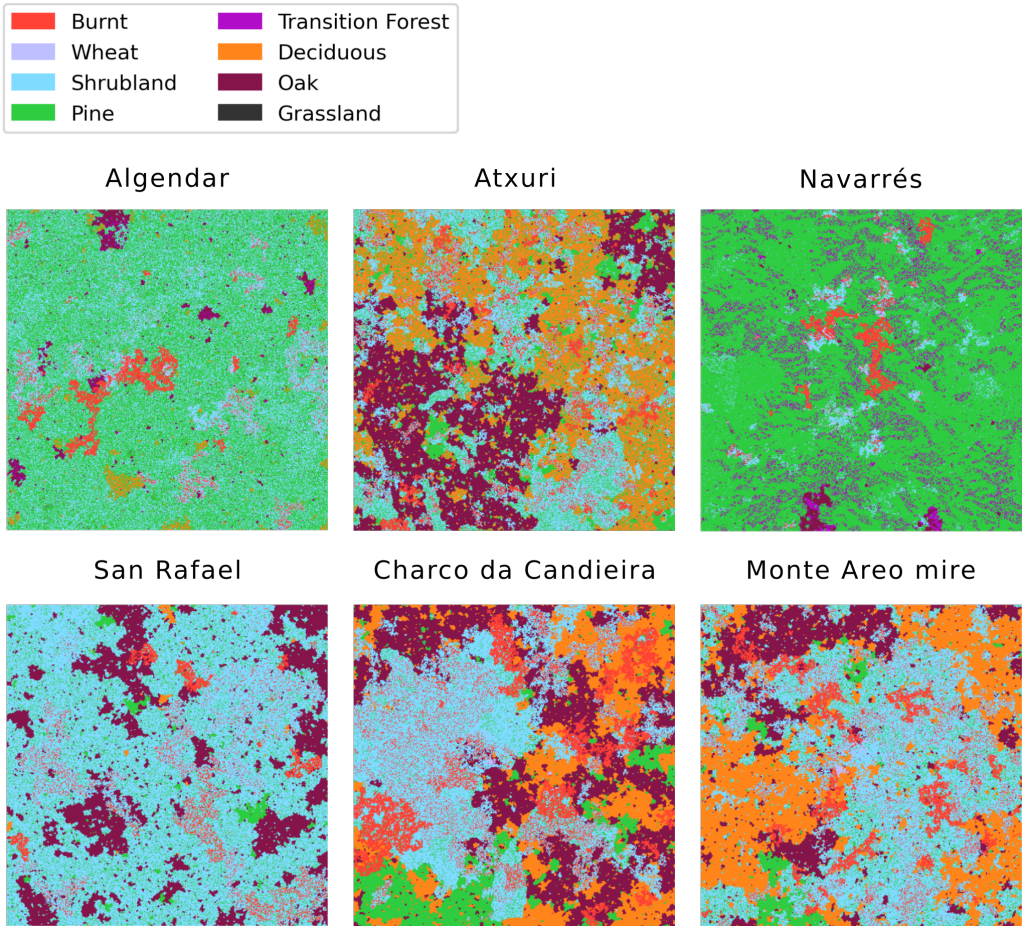


Figure 5.4: Snapshots of simulated landscapes from calibrated models for each study site extracted from the Repast Simphony graphical interface. Burnt cells are in red, showing the spatial pattern emerging from recent fires. Light blue cells show shrubland, green cells show pine woodland, light purple cells show transition forest, dark purple cells show oak forest, and orange cells show deciduous forest.

Discussion

By comparing the average burnt area per hectare per year in a suite of AgroSuccess simulations to corresponding empirical values, I determined that the optimal land cover flammability scenario parameter, S_{LCF}, for all study sites is TF1, apart from the Charco da Candieira site for which the optimal scenario is TF2. These values will be used in the AgroSuccess simulation scenarios explored in Chapter 6.

A notable finding from the analyses performed to calibrate the wildfire model is that the distribution of fire sizes generated by AgroSuccess do not follow a power-law. This is surprising because AgroSuccess uses a mechanistic wildfire spread model that is similar to that used in the Millington LFSM, and J. D. A. Millington et al., 2009 were able to find power-law distributed fire sizes in the outputs of that model. This may be explained by AgroSuccess' inclusion of additional empirical data about the study sites under investigation that J. D. A. Millington et al., 2009 did not. For example, in AgroSuccess wind speed and direction are drawn from an empirical distribution (Section 3.3.5) at the start of each fire whereas in the Millington LFSM these are completely random. Meanwhile, I have kept other aspects of the wildfire spread model the same as in the Millington LFSM (e.g. the soil moisture threshold parameters, Section 4.2.5). This motley approach to simulation modelling (Winsberg, 1999, 2009) has precedents in the literature, and I argue that the inclusion of additional empirical data is important. In the Millington LFSM, parameter values were chosen for their phenomenological plausibility. In AgroSuccess I have endeavoured to distinguish study sites from each other by using empirical data where it is available, and fall back to selecting parameters on the basis of phenomenological plausibility where necessary.

Another reason that AgroSuccess may not be producing power-law distributed fire sizes is that the simulation grid may not be large enough for this phenomenon to be observed. J. D. A. Millington et al., 2009 used a simulated area of $83\,000\,\mathrm{ha}$. The simulation grids used on the simulations discussed above are all approximately $15\,000\,\mathrm{ha}$. These are the same order of magnitude, but it would still be informative to investigate whether increasing the area of the simulation grid led to power-law distributed frequency-area distributions in the AgroSuccess simulation outputs.

Chapter 6

Analysis of AgroSuccess simulation outputs

6.1 Sensitivity analysis

Sensitivity analysis was performed on the AgroSuccess simulation model to allow me to understand how changes in the model parameters lead to changes in a selection of simulation outputs. This analysis allows me to evaluate the extent to which the model is behaving as expected with respect to the parameters that influence the modelled processes. Additionally, by demonstrating high model sensitivity to parameters that are subject to relatively high uncertainty, I am able to argue in favour of future empirical work that could decrease uncertainty in those parameters and, in turn, decrease overall uncertainty in the model outputs.

AgroSuccess has 17 model parameters that are described in Chapter 4. The sensitivity analysis results presented here are based on the results of running 10 simulation replicas for 200 simulated years under each of the following scenarios:

- A *default parameters* scenario for each study site
- A -10% scenario for each parameter and study site
- A +10% scenario for each parameter and study site

For the -10% and +10% scenarios, the corresponding parameter is decreased or increased with

respect to its default value, and the value of all *other* parameters held at their default values. This scheme allows me to isolate the effect of each individual parameter on the model outputs in turn.

Each simulation performed for this sensitivity analysis was allowed to run for 200 simulated years. I found that AgroSuccess simulations for all study sites tended to undergo an initial transient period during which land cover proportions diverged rapidly from their initial conditions (See Section 6.3). I determined that 200 simulated years was long enough for simulations to reach an equilibrium state by visual inspection of aggregate time series of land cover proportions from a sample of simulations. By running simulations for long enough to enter an equilibrium state, I sought to avoid having the sensitivity analysis results affected by fluctuations in the outputs during the initial transient period. Running the sensitivity analysis simulations for longer than 200 years would have increased the simulation run times with no analytical benefit.

The choice of 10 runs per simulated scenario was determined by the practical availability of computational resources. The sensitivity analysis described here required a total of 306 unique parameter combinations to be run across all study sites and parameter values. Additionally, each simulation took approximately 10 minutes to run. Under these circumstances, 10 runs per parameter combination was the most I was able to perform with the computational resources available to me. Similar computational resource constraints prevented me from exploring a greater range of values for each parameter (compared to the ±10% of default value scenarios presented here). See Section 8.3.3 for proposed future work to make the methodology for selecting the number of simulation runs per parameter combination more robust.

6.1.1 Selection of output variables

I explore changes to the following aggregated model output variables under the different scenarios:

- *Shrubland*, the proportion of the simulated landscape in the Shrubland land cover state in the final simulated year of the simulation
- *Mature Forest*, the proportion of the simulated landscape in one of the forest land cover state (Pine, Transition forest, Deciduous, or Oak) in the final simulated year of the simula-

tion

- *Farmed Area*, total simulated landscape area (in ha) used by households for wheat agriculture in the final simulated year of the simulation
- *Wood Area*, total simulated landscape area (in ha) used by households for firewood gathering in the final simulated year of the simulation
- *Burned Area*, mean proportion of the landscape burned per year (in yr^{-1}) over the final 10 years of the simulation

These variables are selected to provide an insight into the effect of varying model parameters on a selection of land cover types reflecting both early and late successional stages (*Shrubland* and *Mature Forest*), human land use (*Farmed Area* and *Wood Area*), and the fire regime (*Burned Area*). All output variables are intended to quantify the state of the simulated landscape at the end of the simulation run. For all outputs except *Burned Area*, the value in the final simulated year is used. The decision to use an average over the last 10 simulated years when summarising burned area was informed by an analysis of the variation in time of each of the output parameters in the final 10 years of all simulation runs used in the sensitivity analysis. For each simulation run and output variable, I calculated the coefficient of variation, c_v for the output in the final 10 years. Here

$$c_v = \frac{\sigma_x}{\mu_x} \tag{6.1}$$

where μ_x and σ_x are the sample mean and sample standard deviation of the output variable in the final 10 years of each simulation. I then calculated the mean of c_v for each output across all simulations in the sensitivity analysis for each study site to obtain an estimate of the relative variation in time for each variable. The use of c_v is appropriate for comparing variation *across* variables because, unlike σ_x, c_v is normalised by each variable's mean.

The results of the output variable time variation analysis are shown in Fig. 6.1. I found that, on average across simulations in the sensitivity analysis, *Burned Area* has coefficient of variation in time that is an order of magnitude greater than that of the other output variables considered. This is to be expected as a consequence of the non-linear relationship between the number of ignitions in a given year and the burned area as a result of the wildfire spread model (see Sec-

tion 4.2.6). Using the 10 year average burned area as a statistic to characterise burned area at the end of each simulation helps to limit the impact of non-representative very large fires in the final simulated year skewing the results of the sensitivity analysis. Note that c_v for *Wood Area* is 0 for all simulations because this is fixed for a given household size, and both the number and size of households within each simulation is held constant (see Eq. (4.21)). There is also very small variation found in *Farmed Area*. This is reasonable, as I would not expect to see large inter-annual variation in the behaviour of agricultural households whose populations are fixed. When a household experiences a 'shortage' in the sense that it does not receive the anticipated wheat returns due to a lack of precipitation or being forced to farm patches with low soil fertility (see Eq. (4.26)), this causes them to request more wheat patches to farm in the subsequent simulated year through the $\mu_{t-1}^{(w)}$ term (mass of wheat per ha collected in the previous year) in Eq. (4.18). However, at the population densities represented in the simulations analysed for this sensitivity analysis, households are unlikely to need to resort to farming patches with low soil fertility. Additionally, in the version of AgroSuccess presented here (Lane, 2023), annual precipitation is held fixed. Consequently, there is little source of interannual variability in households' farming returns which, in turn, leads to little variability in the number of land patches they require to farm.

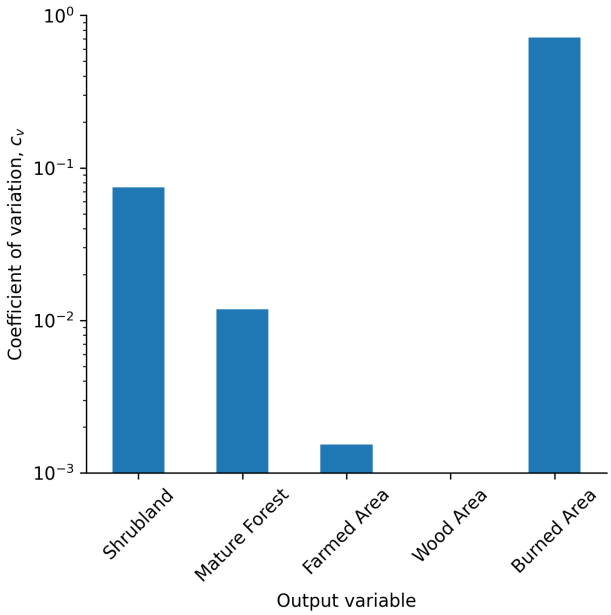


Figure 6.1: Coefficient of variation, c_v, across the final 10 simulated years for each sensitivity analysis output variable. Values shown are the mean c_v taken across all simulations used in the sensitivity analysis. *Burned Area* exhibits an average c_v that is an order of magnitude greater than that for the other output variables.

6.1.2 Statistical significance

It is important to be able to quantify the amount by which an output variable would need to vary in the results from a given parameter combination in comparison to those from the default scenario to be considered *statistically significant*. This can be done by calculating the standard error of the mean,

$$\hat{\sigma}_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \tag{6.2}$$

for each combination of output variable and study site. In Eq. (6.2), σ_x is the sample standard deviation for the output variable taken across n simulation replicas with default parameter values for each study site. In the analysis presented here, I used $n = 10$ replicas per study site. The results of this analysis are shown in Table 6.1.

Table 6.1: Standard error of the mean, $\hat{\sigma}_{\bar{x}}$, for model output variables used in the sensitivity analysis across 10 replicas using default parameter values for each study site.

Study site	Shrubland	Mature Forest	Farmed Area	Wood Area	Burned Area
Algendar	0.00407	0.00480	0.00000	0.00000	0.00055
Atxuri	0.00479	0.00670	0.00000	0.00000	0.00058
Charco da Candieira	0.00414	0.00816	0.00000	0.00000	0.00103
Monte Areo mire	0.00382	0.00651	0.00000	0.00000	0.00062
Navarrés	0.00289	0.00318	0.00000	0.00000	0.00037
San Rafael	0.00507	0.00608	0.00000	0.00000	0.00049

I use the $\hat{\sigma}_{\bar{x}}$ values shown in Table 6.1 to construct a confidence interval (CI) for each output variable. Choosing a confidence level of 66% (significance level, $\alpha = 0.34$), the margin of error, ϵ_x, for each variable is given by

$$\epsilon_x = t_{(n-1),\alpha}\, \hat{\sigma}_{\bar{x}} = t_{9,0.34}\, \hat{\sigma}_{\bar{x}} \simeq \hat{\sigma}_{\bar{x}} \tag{6.3}$$

where $t_{(n-1),\alpha}$ is the critical value of the two-tailed Student's t-distribution as a function of n and α. If an infinite number of batches of 10 default parameter value simulations were run, and ϵ_x calculated for each batch using Eq. (6.3), the population mean for each output variable, μ_x, would be in the range $\bar{x} - \epsilon_x \leq \mu_x \leq \bar{x} + \epsilon_x$ in 66% of the batches (where $\bar{x}$ is the sample mean of variable x for a given batch) (Cumming & Finch, 2005). Consequently, for each output variable under the default parameter scenario, x, the population mean is likely to be in the range

$\bar{x} - \epsilon_x \leq \mu_x \leq \bar{x} + \epsilon_x$. In the following analysis, I treat the difference between mean values for each output variable in the non-default scenarios and the corresponding mean values in the default scenario as statistically significant if they differ by more than ϵ_x. Note that because this approach relies on the comparison of means (taken from batches of simulations run using default and non-default parameter combinations) I rely on the fact that the sample mean tends towards the population mean for large enough sample sizes even if the underlying distribution is not normal under the central limit theorem. See Section 8.3.3 for a discussion of how the approach to statistical significance testing described here could be improved in future work.

The standard error values for the land cover output variables (Mature Forest and Shrubland) in Table 6.1 are broadly similar to each other and across study sites. Standard error values for Burned Area are small relative to the land cover output variables reflecting the fact that the units for Burned Area are yr^{-1} whereas the land cover output variables are unitless landscape proportions, such that their values are not directly comparable. Farmed Area and Wood Area have $\hat{\sigma}_{\bar{x}} = 0$ for all study sites because their values were constant across all default parameter value replicas. This is because the number of wheat and wood plots required by each household are deterministically linked to the size of the household (see Eq. (4.18) and Eq. (4.21) respectively). Since the number and size of households are held constant in AgroSuccess, it is expected that there would not be any variation across replicas for a given combination of model parameters.

6.1.3 Discussion

Here I review the results of the sensitivity analysis for the Navarrés study site in detail. Analogous results data are shown for the remaining five study sites in Appendix F. At the end of this section I highlight common themes that can be seen across study sites, as well as some identified differences.

The sensitivity analysis results for Navarrés are shown in Table 6.2, with statistically significant results shown in **bold**. All results for the *Farmed Area* and *Wood Area* output variables would be statistically significant according to the significance levels determined in Section 6.1.2 because the standard error of the mean, $\hat{\sigma}_{\bar{x}}$, for those variables was found to be zero. We therefore treat changes in *Farmed Area* and *Wood Area* of $\leq -10\%$ or $\geq +10\%$ relative to the default scenario as significant. Several *Mature Forest* values are reported as statistically significant despite representing less than 0.5% change with respect to the default parameter scenario. This is because,

while the output might represent a small proportional change relative to the default scenario, only a small absolute difference of $0.003\,18$ is required for the difference to be statistically significant (see Table 6.1). This, in turn, is due to there being little variation in the proportion of the landscape occupied by Mature Forest land cover types in the last simulated year of the default parameter simulation results.

Table 6.2: Sensitivity analysis results for the Navarrés study site. The 'Value' column specifies the default value for each parameter, and the 'Value +10%' and 'Value -10%' columns give the corresponding values for parameters used in the +10% and -10% scenarios. The remaining columns give the mean values under the +10% and -10% scenarios for the *Shrubland*, *Mature Forest*, *Farmed Area* *Wood Area*, and *Burned Area* output variables averaged over 10 simulation replicas, along with the corresponding percentage change relative to the default scenario in parentheses. Values in **bold** show statistically significant differences with respect to the default scenario (see Section 6.1.2). Output variables indicated by * showed 0 variance in default scenario runs, so outputs that differ from the default scenario values by $\leq -10\%$ or $\geq +10\%$ are shown as significant.

Parameter	Value	Value -10%	Value +10%	Shrubland, -10%	Shrubland, +10%	Mature Forest, -10%	Mature Forest, +10%	Farmed Area, -10%*	Farmed Area, +10%*	Wood Area, -10%*	Wood Area, +10%*	Burned Area, -10%	Burned Area, +10%
Number of households per village	10.00	9.000	11.000	**0.042 (-10%)**	**0.041 (-12%)**	**0.949 (1%)**	**0.947 (0%)**	6.996 (-10%)	**8.628 (11%)**	2.544 (-10%)	**3.109 (10%)**	**0.004 (-9%)**	**0.004 (-11%)**
Labour availability	300.00	270.000	330.000	**0.043 (-8%)**	**0.043 (-8%)**	**0.946 (0%)**	0.943 (0%)	7.843 (1%)	7.773 (0%)	2.826 (0%)	2.826 (0%)	0.004 (-2%)	0.004 (5%)
Wheat farming labour requirement	50.00	45.000	55.000	0.044 (-5%)	0.049 (5%)	0.941 (-0%)	0.941 (-0%)	7.773 (0%)	7.773 (0%)	2.826 (0%)	2.826 (0%)	0.004 (3%)	0.004 (2%)
Maximum patch farm time	50.00	45.000	55.000	**0.052 (11%)**	**0.041 (-12%)**	**0.934 (-1%)**	0.945 (0%)	7.773 (0%)	7.914 (2%)	2.826 (0%)	2.826 (0%)	**0.005 (20%)**	0.004 (-4%)
Farmer conservativeness scalar	0.75	0.675	0.825	**0.041 (-12%)**	0.047 (1%)	**0.947 (1%)**	**0.938 (-0%)**	**9.186 (18%)**	7.137 (-8%)	2.826 (0%)	2.826 (0%)	**0.003 (-17%)**	**0.005 (13%)**
Crop reseed proportion	0.15	0.135	0.165	0.048 (3%)	0.048 (1%)	0.940 (-0%)	0.940 (-0%)	7.773 (0%)	8.479 (9%)	2.826 (0%)	2.826 (0%)	0.004 (6%)	0.004 (7%)
Farm value distance parameter	1.00	0.900	1.100	**0.042 (-10%)**	**0.040 (-14%)**	**0.947 (1%)**	**0.947 (0%)**	7.773 (0%)	7.843 (1%)	2.826 (0%)	2.826 (0%)	0.004 (-6%)	0.004 (-6%)
Farm value fertility parameter	1.00	0.900	1.100	**0.043 (-9%)**	**0.052 (12%)**	**0.946 (0%)**	**0.934 (-1%)**	7.843 (1%)	7.773 (0%)	2.826 (0%)	2.826 (0%)	0.004 (-4%)	**0.005 (21%)**
Farm value land cover conversion parameter	1.00	0.900	1.100	0.044 (-6%)	**0.050 (7%)**	0.944 (0%)	0.940 (-0%)	7.773 (0%)	7.843 (1%)	2.826 (0%)	2.826 (0%)	0.004 (-4%)	0.004 (6%)
Firewood required per capita per year	1100.00	990.000	1210.000	**0.042 (-10%)**	0.047 (-0%)	0.945 (0%)	0.942 (-0%)	7.794 (0%)	7.843 (1%)	**2.120 (-25%)**	2.826 (0%)	**0.004 (-10%)**	0.004 (7%)
Firewood biomass removal rate	0.10	0.090	0.110	0.048 (3%)	0.047 (1%)	0.941 (-0%)	0.940 (-0%)	7.773 (0%)	7.773 (0%)	2.826 (0%)	**2.120 (-25%)**	0.004 (7%)	0.004 (0%)
Wood value distance parameter	1.00	0.900	1.100	**0.043 (-9%)**	0.048 (3%)	0.942 (-0%)	0.941 (-0%)	7.773 (0%)	7.773 (0%)	2.826 (0%)	2.826 (0%)	0.004 (3%)	0.004 (1%)
Climax forest biomass density	300000.00	270000.000	330000.000	0.048 (3%)	**0.044 (-7%)**	**0.938 (-0%)**	0.942 (-0%)	7.773 (0%)	7.773 (0%)	2.826 (0%)	**2.120 (-25%)**	0.004 (4%)	0.004 (3%)
Land cover colonisation base rate	0.05	0.045	0.055	0.048 (2%)	**0.042 (-11%)**	0.941 (-0%)	**0.948 (1%)**	7.773 (0%)	7.914 (2%)	2.826 (0%)	2.826 (0%)	0.004 (-7%)	**0.003 (-16%)**
Land cover colonisation spread rate	0.20	0.180	0.220	**0.052 (11%)**	0.047 (1%)	**0.931 (-1%)**	**0.937 (-1%)**	7.773 (0%)	7.843 (1%)	2.826 (0%)	2.826 (0%)	**0.005 (31%)**	**0.005 (12%)**
Mesic threshold	500.00	450.000	550.000	0.046 (-3%)	0.046 (-2%)	0.940 (-0%)	0.941 (-0%)	7.773 (0%)	7.843 (1%)	2.826 (0%)	2.826 (0%)	0.004 (-4%)	0.004 (5%)
Hydric threshold	1000.00	900.000	1100.000	**0.039 (-16%)**	**0.036 (-22%)**	**0.951 (1%)**	**0.952 (1%)**	7.964 (2%)	7.985 (3%)	2.826 (0%)	2.826 (0%)	**0.003 (-25%)**	**0.003 (-26%)**

The results show expected sensitivity to the main parameters that drive the subsistence agriculturalist submodel. Increasing 'Number of households per village' by 10% increases both *Farmed Area* and *Wood Area* by approximately 10%, and decreasing 'Number of households per village' produces an equal and opposite effect. Increasing 'Farmer conservativeness scalar' decreases *Farmed Area*, and the converse increases it. This is expected because the number of wheat patches a household requires is inversely proportional to the conservativeness scalar parameter (see Eq. (4.18)). Increasing 'Crop reseed proportion' (the proportion of wheat held back for reseeding the subsequent year) increases *Farmed Area* as expected. *Wood Area* is strongly sensitive to 'Firewood required per capita per year', 'Firewood biomass removal rate', and 'Climax forest biomass density'. 'Increasing Firewood biomass removal rate' or 'Climax forest biomass density' decreases *Wood Area*, and decreasing 'Firewood per capita per year' also decreases *Wood Area*. The fact that AgroSuccess is so sensitive to 'Number of households per village' and 'Climax forest biomass density' is noteworthy because both parameters are subject to significant empirical uncertainty (see Section 4.3.4). AgroSuccess (and other models like it) would need to be provided with accurate estimates of the number of agriculturalists within the simulated area, as well as key biophysical data like 'Climax forest biomass density', to produce realistic results. This is an example of how modelling studies such as this help to identify gaps in available empirical data and motivate future data collection areas.

The results show some unexpected patterns in the behaviour of the land cover type output variables, especially *Shrubland*. There are several parameters for which both +10% and -10% scenarios show a significant decrease in *Shrubland* (e.g. 'Number of households per village', 'Farm value distance parameter', and 'Hydric threshold'). This could be explained by the possibility that we are observing the secondary effect of random fluctuations in *Burnt Area* through its causal effect on the proportion of *Shrubland* in the landscape, rather than the direct effect of the corresponding parameters on these outputs. Additionally, there are several instances of 'asymmetric effects' characterised by a change in a parameter in one direction having the expected effect on an output (e.g. increasing 'Crop reseed proportion' increasing *Farmed Area* or increasing 'Firewood required per capita per year' increasing *Wood Area*), but the corresponding change in the opposite direction not resulting in any significant change in the outputs. This is explained by the $0.071\,\mathrm{ha}$ resolution of the simulation grid. Households always use a discrete number of grid cells for their farming and wood gathering activities. The 25% decrease in *Wood Area* to $2.120\,\mathrm{ha}$ resulting from a 10% decrease in 'Firewood required per capita per year', for example, implies an absolute reduction of $0.71\,\mathrm{ha}$, corresponding to each of the 10 households in the village using

one less land patch for firewood gathering. If the fractional increase in the number of wood patches to gather within the ceiling function in Eq. (4.21) due to a 10% increase in 'Firewood required per capita per year' is not enough to increase the result by 1, then households will not claim an additional land patch. This effect could be mitigated by increasing the resolution of the simulation grid at the cost of additional computational expense.

Sensitivity analysis has demonstrated that there are some parameters that AgroSuccess is not sensitive to in the parameter regime considered here. I found that no land patches were converted to Depleted Agricultural Land (DAL) in any of the simulations performed for this analysis. This demonstrates that for a community with up to 11 households, the anthropogenic pressure on the landscape is sufficiently low that it is not necessary to farm the same land patch 45 or more times in a 200 year time frame. Consequently, for the range of 'Maximum patch farm time' and 'Number of households per village' values considered in this analysis, AgroSuccess is not sensitive to 'Maximum patch farm time'. However, this parameter could become useful when exploring landscapes with larger human populations. The results show that AgroSuccess is not sensitive to 'Labour availability' or 'Wheat farming labour requirement', indicating that for the household sizes represented here, farming labour is not a limiting factor. Additionally, the output variables analysed here are not sensitive to the 'Farm value distance', 'Farm value fertility', 'Farm value land cover conversion', or 'Wood value distance' parameters. This is expected, because these parameters are used to influence the spatial location of land patches used by households, rather than the total area as represented by *Farmed Area* and *Wood Area*.

Several of the patterns observed in the sensitivity analysis results for Navarrés are also present in the results for the other study sites considered in this thesis (see Appendix F). This includes the effects of 'Number of households per village' and 'Farmer con scalar' on *Farmed Area* that are described above. However, there are two key differences. First, an increase in 'Crop reseed proportion' does not produce an observed increase in *Farmed Area* in the results for Algendar, Charco da Candieira, Monte Areo mire, or San Rafael. Second, while Algendar, Charco da Candieira, and Monte Areo mire show analogous effects of 'Firewood required per capita per year', 'Firewood biomass removal rate', and 'Climax forest biomass density' on *Wood Area* in comparison to Navarrés, San Rafael shows a different 'asymmetric effect'–i.e. no change for increasing 'Firewood biomass removal rate' or 'Climax forest biomass density', or decreasing 'Firewood required per capita per year', but an observed change in the expected direction when **decreasing** 'Firewood biomass removal rate' and 'Climax forest biomass density', or **increasing** 'Firewood

required per capita per year'. Additionally, Atxuri doesn't show any sensitivity to these parameters intended to influence *Wood Area*. These differences can be explained by the same simulation grid resolution issue that is discussed above in the context of 'asymmetric effects'.

The above sensitivity analysis demonstrates that AgroSuccess is sensitive to key parameters controlling the anthropogenic model components. As anthropogenic activity and fire are two main drivers of land cover change, this is evidence that AgroSuccess is useful for investigating interactions between humans, fire, and climate. This analysis also showed that the range of human populations explored here (9–11) are not enough to apply significant anthropogenic pressure to the landscape. Although we don't see consistent statistically significant effects on land cover when modifying anthropogenically related parameters, this is not unexpected. For example, *Mature Forest* occupies such a large proportion of the simulated landscapes that it is not surprising that human populations have a low proportional impact on it. An important finding is that a simulation grid resolution of $0.071\,\mathrm{ha}$ may be too coarse to capture detailed information about the effects of small changes in parameter values on anthropogenic impacts. This is because the change in parameter value may factor into household agents' calculations, but not enough to require them to change their subsistence plans up or down by a whole simulation cell (see discussion of 'asymmetric effects' above).

6.2 Results: Counterfactual Scenarios

Here I present and analyse results from sets of simulations for counterfactual scenarios that examine ecological dynamics and human activity over centennial timescales. I also investigate whether or not the anthropogenic component of the model is able to produce statistically significant effects in the simulated landscapes' land cover state. In the sensitivity analysis reported Section 6.1 I found no clear statistical signal that the number of households in the simulation had a meaningful effect on the proportion of the landscape occupied by early-successional (Shrubland) or late-successional (Mature Forest) land-cover types, or the proportion of Burned Area in the landscape. This is explained by the the range of 9–11 households explored in the sensitivity analysis only placing low anthropogenic pressure on the landscape. In the analysis presented in this section I also include results of simulations that include 100 households per simulation to allow me to demonstrate the effect of increased anthropogenic pressure on land cover proportion and burned area time series, and to determine whether or not this increase in number of

households produces a statistically significant effect on land cover proportions.

For each of the six study sites, I ran simulations for three counterfactual scenarios with 20 replicas each:

1. No households, representing a baseline scenario with no anthropogenic activity
2. 10 households per simulation
3. 100 households per simulation

All scenarios were run for 400 simulated years, with all model parameters except the number of households held at the default values used in the sensitivity analysis (see Section 6.1). A simulation length of 400 years was selected because all study sites have at least 400 years of pollen abundance data available following the date at which it is believed humans began practicing agriculture at each site (see Section 3.1.4). Running these simulations for 400 simulated years maximised the amount of simulated land cover proportion data that could be compared to empirical pollen abundance time series. For each study site, the first simulated year is the hypothesised date that humans began practicing agriculture at the site (see Section 3.1). The initial conditions for each simulated run comprise a randomly generated land cover map in which the relative proportion of each land cover type is constrained to match the land cover proportions at the appropriate year in the corresponding empirical pollen abundance data (see Section 3.3.3 for details of how these maps were generated).

6.2.1 Comparison to empirical pollen abundance data

Fig. 6.2 shows land cover proportion time series generated from the scenarios described above, alongside corresponding time series derived from empirical pollen abundance data (see Section 3.2.2).

Initial transient and equilibrium states

Results for all study sites show a large initial transient during the first 150 simulated years, before reaching a stable 'equilibrium' state. The equilibrium state for all study sites is significantly

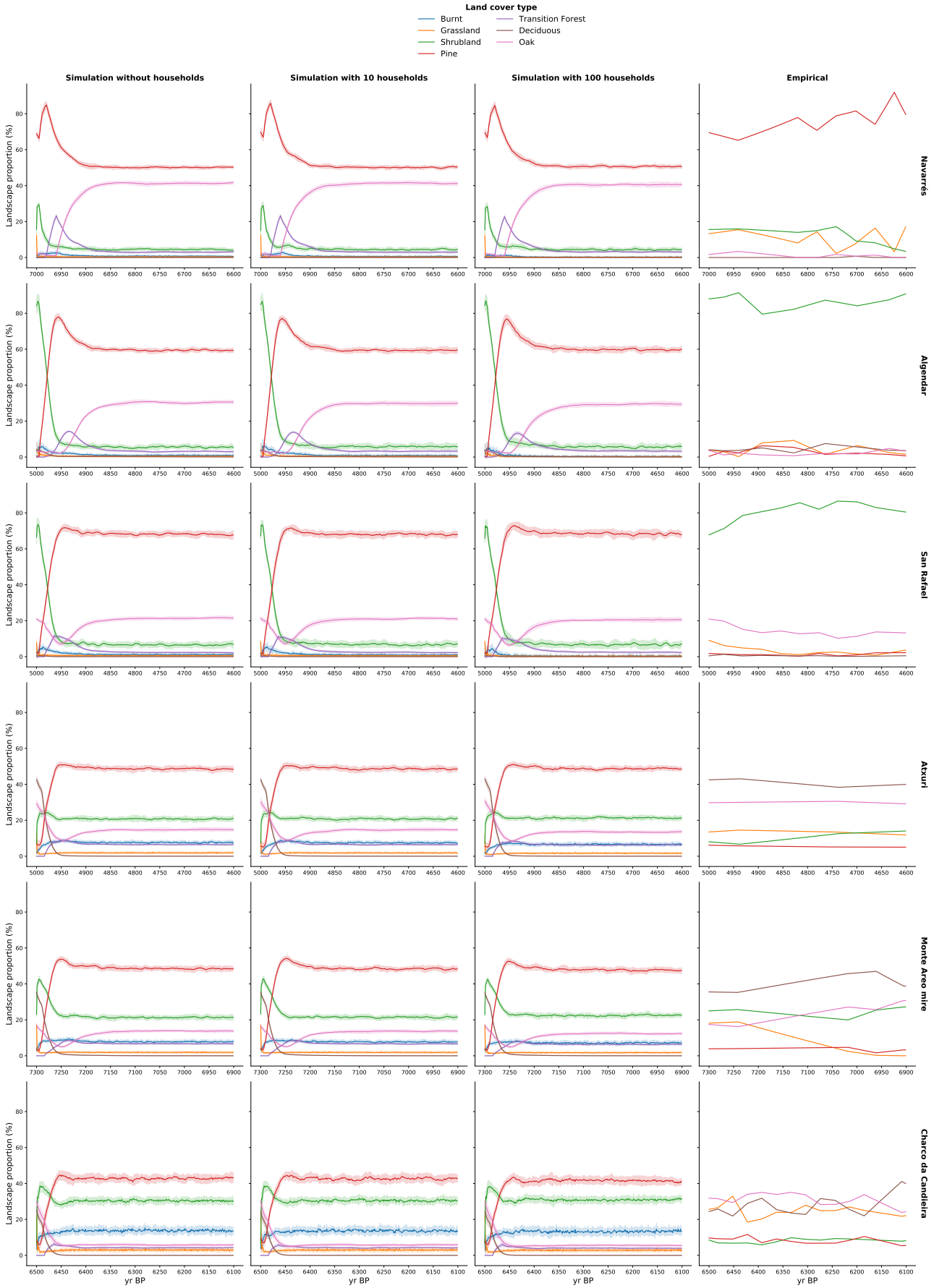


Figure 6.2: Time series of land cover proportion from empirical pollen abundance and three simulated scenarios (no households, 10 households, and 100 households per simulation) for each of the six study sites. Confidence intervals for simulated results are 1 standard deviation around the mean calculated across 20 simulation replicas.

different than the land cover state represented in the empirical pollen abundance time series. This demonstrates that there is a mismatch between the initial conditions inferred from the pollen abundance time series and the model dynamics resulting from the simulation rules described in Chapter 4.

Within each simulation, once the equilibrium land cover proportions have been reached, they remain approximately constant in time and are subject to only minor fluctuations. Consequently, the model fails to reproduce changes in land cover proportion observed in the empirical data. For example, in the empirical data:

1. At the Navarrés study site, there is a decrease in Pine and increase in Grassland at $6675\,\mathrm{yr\,BP}$
2. At the Monte Areo mire study site, there is a decrease in Deciduous and increase in Oak and Pine at $6955\,\mathrm{yr\,BP}$
3. At the Charco da Candieira study site, there is a decrease in Oak and Pine and an increase in Deciduous at $1680\,\mathrm{yr\,BP}$

The behaviour in 1. could be reproduced by the model if there was an increase in fire or agricultural activity (or both) at this point in the simulation. This is because frequent burning (through fire) or conversion of Pine land cover to Wheat (through agricultural activity) would maintain a greater proportion of the landscape in early-successional states, including Shrubland (see Fig. 4.1). The effect of decreasing Deciduous in favour of increase in Oak and Pine observed in 2., and the inverse that process observed in 3., could be reproduced by the model by modifying the availability of Deciduous seeds at the appropriate points in the simulation. This is because Pine and Oak (via Transition Forest) grid cells transition to Deciduous over time given the presence of Deciduous seeds. Conversely, the absence of Deciduous seeds would have the opposite effect.

Variability across scenarios and study sites

For each study site, all simulated scenarios show visually indiscernible patterns in the aggregated land cover proportion outputs. This demonstrates that, even with 100 households per simulation, land cover changes resulting from anthropogenic activity do not produce large changes in the

observed land cover. However, it is worth noting that because the simulation grids for all study sites have an area of $15\,279\,\mathrm{ha}$, a human population of 100 households with 6 members each corresponds to a population density of only $0.04\,\mathrm{pers/ha}$. By comparison, the contemporary agricultural municipality of Casillas in central Spain had a population density of $0.70\,\mathrm{pers/ha}$ in 2011 (Seijo et al., 2015). While a population density of $6\,\%$ of contemporary values seems plausible (or even high) for a community in the mid-Holocene, it is unsurprising that a population of this size would not significantly alter the character of the landscape.

Fig. 6.3 shows example land cover maps obtained after 400 simulated years for the Navarrés and Charco da Candieira study sites with 0, 10, and 100 simulated households. These are taken from individual runs of the simulations whose outputs were used to produce the aggregated time series in Fig. 6.2. The examples for Navarrés show a broadly consistent spatial pattern of land cover across simulated household scenarios. The variation that is apparent is mainly in the location of burnt and shrubland patches, and can be explained by the occurrence of stochastic wildfires. The examples for Charco da Candieira show relatively high variation in spatial land cover pattern across simulation runs compared to Navarrés. This is due to differences in the fire regime between the two sites (see below). The effect of the presence of agriculturalist households is apparent from the presence of wheat patches in the 100 household scenarios for both study sites, especially in the bottom left of frame for Navarrés and around the centre of the frame for Charco da Candieira. However, in both cases wheat makes up only a small proportion of the overall land cover.

In this section I have focused on analysing simulation outputs that have been aggregated to produce time series, rather than the analysis of spatial difference in land cover maps across simulation runs and between scenarios. This is because the empirical pollen abundance data that I have used to evaluate model performance is inherently aspatial. That is, land cover change at each study site is characterised by a time series. Consequently, I have aggregated simulation outputs to produce time series that are comparable with the empirical data.

Across study sites we see similar patterns in the rank order of the land cover types that dominate the landscapes in the equilibrium states of the simulation outputs:

1. Navarrés, Algendar, and San Rafael are dominated by Pine, Oak, and Shrubland

2. Atxuri and Monte Areo mire dominated by Pine, Shrubland, and Oak

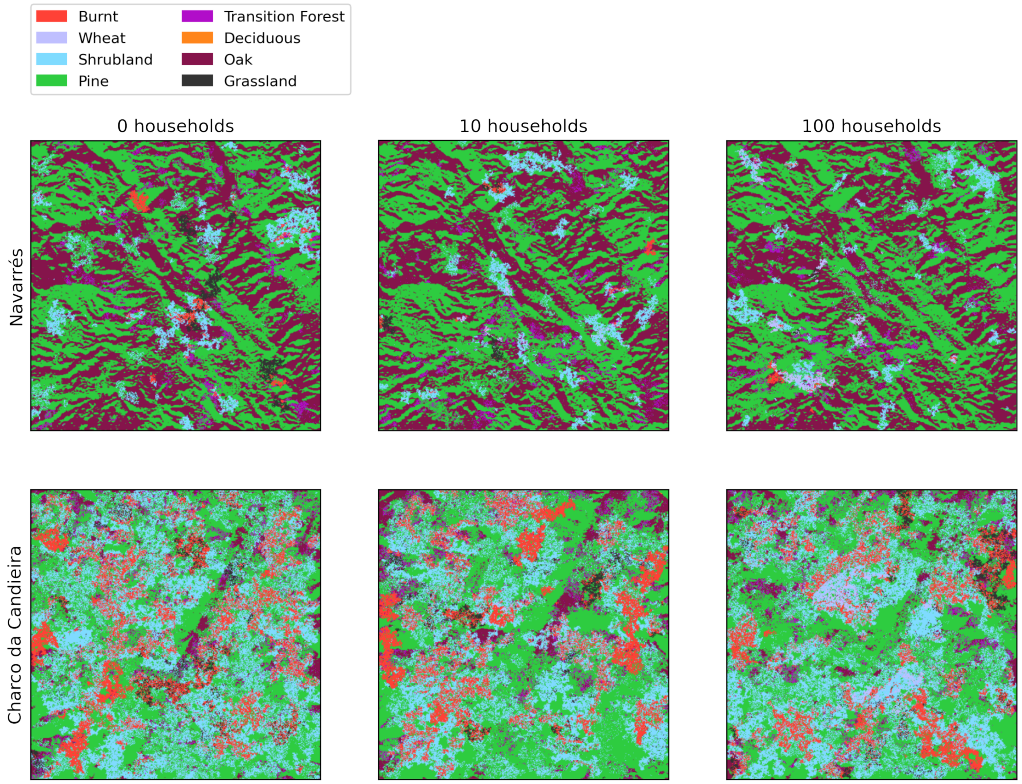


Figure 6.3: Example land cover maps produced by individual runs of AgroSuccess simulations after 400 simulated years for the Navarrés and Charco da Candieira study sites, and for scenarios that include 0, 10, and 100 simulated households.

3. Charco da Candieira is dominated by Pine, Shrubland, and Burnt area

These differences can be explained by the effect of fire in the simulated landscapes. The first group above all have an average of 4.1 fires per year and calibrated land-cover flammability scenario TF1, the second group have 27.2 fires per year and calibrated land-cover flammability scenario TF1, and the third group have 14.3 fires per year and land-cover flammability scenario TF2 (see Section 5.1.2). Increased fire frequency for the second and third groups, and increased flammability in the case of the third group, allows early-successional Shrubland to outcompete late-successional Oak in the equilibrium state.

Stochastic variation

Across all study sites there is limited and approximately constant stochastic variation within simulation outputs for each land cover type demonstrating that the model dynamics tend to produce a stable equilibrium state. Among the different land cover types, we consistently see highest stochastic variation in the proportion of the landscape occupied by Pine and Shrubland. This

can be understood in terms of the modelled ecological dynamics. Shrubland has high land-cover flammability (see Table 4.5) and is therefore relatively susceptible to short-term fluctuations due to fire occurrence. The proportion of the landscape occupied by Pine is coupled to the proportion occupied by Shrubland because Shrubland grid cells transition to Pine in 10–15 simulated years (depending on the cell's aspect and soil moisture) when Pine seeds are present. Because the landscapes simulated here are dominated by Pine, there is also high availability of Pine seeds (see rules specified in Section 4.2.4) resulting in persistent Pine dominance.

6.2.2 Signal of anthropogenic change

To evaluate whether AgroSuccess is able to produce a statistically significant signal of anthropogenic change, I performed a similar analysis to that used to quantify statistical significance in the sensitivity analysis in Section 6.1.2 but using the same simulations as those analysed in the previous section. Table 6.3 shows the standard error of the mean, $\hat{\sigma}_{\bar{x}}$, for the *Shrubland*, *Mature Forest* and *Burned Area* output variables across the 20 replicas of the simulations without any anthropogenic agents for each study site (see Section 6.1.1 for definitions of these variables). This defines a 'baseline' scenario that I will compare the corresponding output statistics from the simulations that do include anthropogenic agents to. Note that unlike in Section 6.1.2 I do not consider the effect of adding household agents on the *Wood Area* and *Farm Area* output variables because the baseline scenario does not include anthropogenic agents so these outputs will be 0, making a meaningful comparison impossible.

Table 6.3: Standard error of the mean, $\hat{\sigma}_{\bar{x}}$, for model output variables used to determine statistical significance of anthropogenic effects across 20 replicas using default parameter values for each study site.

Study site	Shrubland	Mature Forest	Burned Area
Algendar	0.00312	0.00381	0.00037
Atxuri	0.00352	0.00465	0.00045
Charco da Candieira	0.00456	0.00688	0.00082
Monte Areo mire	0.00299	0.00350	0.00041
Navarrés	0.00242	0.00283	0.00019
San Rafael	0.00498	0.00582	0.00052

The effects of adding anthropogenic agents on *Shrubland*, *Mature Forest* and *Burned Area* simulation outputs are summarised in Table 6.4. While inclusion of 10 households does not produce a clear statistical effect on land cover (as I found in the sensitivity analysis), simulations with 100 households do tend to show statistically significant differences in the outputs considered compared to the baseline scenario (no households).

Table 6.4: Summary of the effect of adding anthropogenic agents to AgroSuccess simulations on *Shrubland*, *Mature Forest*, and *Burned Area* output variables. For each output variable, scenario (10 households and 100 households), and study site, mean values across 20 simulation replicas are shown. Values in parentheses are percentage change relative to the baseline scenario (no households). Values in **bold** show statistically significant differences with respect to the baseline scenario.

Variable	Shrubland		Mature Forest		Burned Area	
No. Households	10	100	10	100	10	100
Study Site						
Algendar	**0.058 (7%)**	**0.058 (8%)**	0.929 (-0%)	0.927 (-0%)	0.005 (-2%)	**0.005 (12%)**
Atxuri	0.211 (0%)	0.212 (1%)	0.698 (0%)	**0.685 (-2%)**	0.019 (0%)	**0.021 (9%)**
Charco da Candieira	0.304 (2%)	**0.317 (6%)**	0.528 (-1%)	**0.505 (-5%)**	**0.032 (5%)**	**0.035 (14%)**
Monte Areo mire	**0.216 (2%)**	**0.227 (7%)**	0.687 (-0%)	**0.662 (-4%)**	**0.020 (3%)**	**0.023 (15%)**
Navarrés	**0.043 (9%)**	**0.044 (11%)**	**0.945 (-1%)**	**0.942 (-1%)**	**0.004 (20%)**	**0.004 (35%)**
San Rafael	0.071 (4%)	0.072 (5%)	0.912 (-0%)	0.908 (-1%)	0.006 (5%)	**0.007 (23%)**

The statistically significant decrease in *Mature Forest* seen in four of the six study sites is explained by the increased anthropogenic pressure on the landscape causing agents to resort to converting forest land cover types to agricultural land. Note that due to the rules used to select land patches to convert to wheat agriculture specified in Section 4.3.4, households preferentially select Burnt, Grassland, Shrubland, and existing Wheat patches to convert to Wheat over forest types due to their relatively small land cover conversion cost (see Table 4.1 for details). The observed increase in proportion of Shrubland in simulations with 100 households is likely caused by the ecological succession rules causing land that agents had converted to Wheat subsequently progressing to Shrubland. Similarly, the increase in *Burned Area* seen in simulations with 10 households is explained by the fact that, once households have converted a land patch to Wheat, it will take $3\,\mathrm{yr}$ to transition to Shrubland, and then a further 10–15 years to transition to a forest type, e.g. Pine. Because Wheat and Shrubland are have relatively high land-cover flammability values (see Table 4.5) this has the effect of increasing the proportion of the landscape that fire can most easily spread to.

6.3 Discussion

As seen in Section 6.2.1, AgroSuccess simulations for the study sites considered in this thesis exhibit an initial transient period in which land cover proportion outputs diverge rapidly from their boundary conditions before converging on a stable equilibrium state. Additionally, we see from the confidence intervals in the simulation output time series in Fig. 6.2 that stochastic variability in time series outputs is quite limited. This is likely because the wildfire submodel (see Section 4.2.6) fire is the only source of stochasticity in the model. While anthropogenic

activity is a source of modelled ecological disturbance, the model processes that represent it are deterministic because the human population is held fixed, and the rules governing subsistence planning are deterministic. J. D. A. Millington et al., 2009 saw a similar pattern of initial transient period followed by stable equilibrium when using the Millington LFSM to run a 'no disturbance' scenario (i.e. without fire). This points to the possibility that the AgroSuccess model may not include sufficiently strong disturbance effects (from fire and anthropogenic activity) to produce qualitative changes in the land cover state.

Additionally, the fact that the aggregate outputs of AgroSuccess simulations diverge rapidly from their initial conditions implies that the model cannot be sensitive to the initial *spatial configuration* of the landscape that is provided by the neutral landscape models (NLMs, Section 3.3.3). Instead, the model dynamics cause the evolution of a compatible spatial configuration of land cover during the course of the model run. Rather than using NLMs, it may be more appropriate to use a model 'burn-in period' in which a simulation is allowed to run for sufficient time to produce a landscape configuration that corresponds to the model dynamics, before introducing scenario-orientated perturbations (e.g. the introduction of anthropogenic agents). Under this methodology, it would be sufficient to use a non-informative uniform Burnt area land cover map as an initial condition, removing the need for the generation of NLMs.

There are a number of sources of residual uncertainty that could explain why AgroSuccess' model dynamics do not match those of the empirically derived land cover type data that its outputs are intended to be comparable to. First, the ecological succession model is able to incorporate spatially heterogeneous soil type maps, but in the analyses presented here a uniform map was used (see Section 3.3.2). Soil type indirectly influences the land cover by altering soil moisture. Second, each simulation uses constant values for wind speed, temperature and precipitation. In reality these are time varying quantities. Third, seeds corresponding to species represented in the forest land cover types are 'imported' into the simulation grid at a fixed rate throughout each simulation (controlled by the 'Land cover colonisation base rate' parameter), and moreover at the *same* rate for all seed types (Pine, Oak, and Deciduous). It is possible that each seed type would require its own 'import rate' representing differences in the quantity of seeds of each type entering the simulation grid for the model to be able to accurately reproduce ecological dynamics. See Section 8.3.4 for discussion of how these residual uncertainties could be addressed in future work.

An outcome of the sensitivity analysis discussed in Section 6.1.3 is that, while AgroSuccess is sensitive to many of the parameters related to the anthropogenic component of the model, there are some that it is not sensitive to at all. These include 'Labour availability' and 'Wheat farming labour requirement'. Additionally, I found in the sensitivity analysis that no land patches were converted to the Depleted Agricultural Land (DAL) land cover type. Given that the model was not sensitive to these aspects of the model, they might be considered for removal to facilitate model simplification. However, it is important to note that the sensitivity analysis described in Section 6.1 did not account for possible parameter interactions. Future work on sensitivity analysis (see Section 8.3.3) may identify such interactions. Future work may also lead to the inclusion of dynamically evolving household populations through the addition of birth and death processes. In this scenario, the inclusion of 'Labour availability' and 'Wheat farming labour requirement' parameters would be useful. Consequently, removal of these parameters could be counterproductive to ongoing model development. While the model did not appear to be strongly sensitive to the anthropogenically related parameters 'Farm value distance', 'Farm value fertility', 'Farm value land cover conversion', and 'Wood value distance', I believe these should be retained because of their influence on the spatial distribution of land use in the landscape.

A key finding from the analyses presented in this chapter is the high sensitivity of the model to the 'Number of households per village' parameter. I showed in Section 6.2.2 that adding anthropogenic agents has a statistically significant effect on the land cover proportions of the landscape, but only if a sufficient number of households are included in the simulation. For AgroSuccess to be used to reproduce landscape dynamics, it would be necessary to obtain realistic estimates of the human population of each study site during the simulated period. See Section 8.3.5 for a proposal of how this might be achieved as a piece of future work.

Chapter 7

Software implementation of AgroSuccess

Here I discuss the general principals I followed during the implementation of the AgroSuccess model. The implementation of a socio-ecological model as piece of software is a distinct process with respect to the development of a conceptual model. Whereas the conceptual model is a design (which can be analysed and critiqued in its own right), the implementation is a specific realisation of that design which can be subject to errors or other issues that limit its usefulness to both the original developer and prospective future users. I have endeavoured to make the model implementation *open* in the sense that is available for others to scrutinise and reuse (see Section 7.1.2). Having described my overall approach to developing the AgroSuccess model implementation, I present a novel approach to managing the complex ecological succession rules in the model using a graph database in Section 7.2. In Section 7.3 I summarise some of the challenges that arose during the implementation of the model.

7.1 Simulation model implementation

7.1.1 Development framework

AgroSuccess is implemented in the Java programming language and uses the Repast Simphony Agent Based Modeling framework (North et al., 2013). Repast Simphony was chosen because

it provides a set of tools including a graphical user interface (GUI) to view running simulations interactively, utilities for outputting data from running models, utilities for scheduling events to occur in simulations (e.g. agents performing decision making activities), and the capacity to run simulations in 'batch mode' including on remote machines such as those provided by Cloud providers including AWS. Fig. 7.1 shows a screenshot of AgroSuccess running in the Repast Simphony GUI, illustrating how it allows for interactive modification of parameter values and visualisation of model outputs in a running simulation.

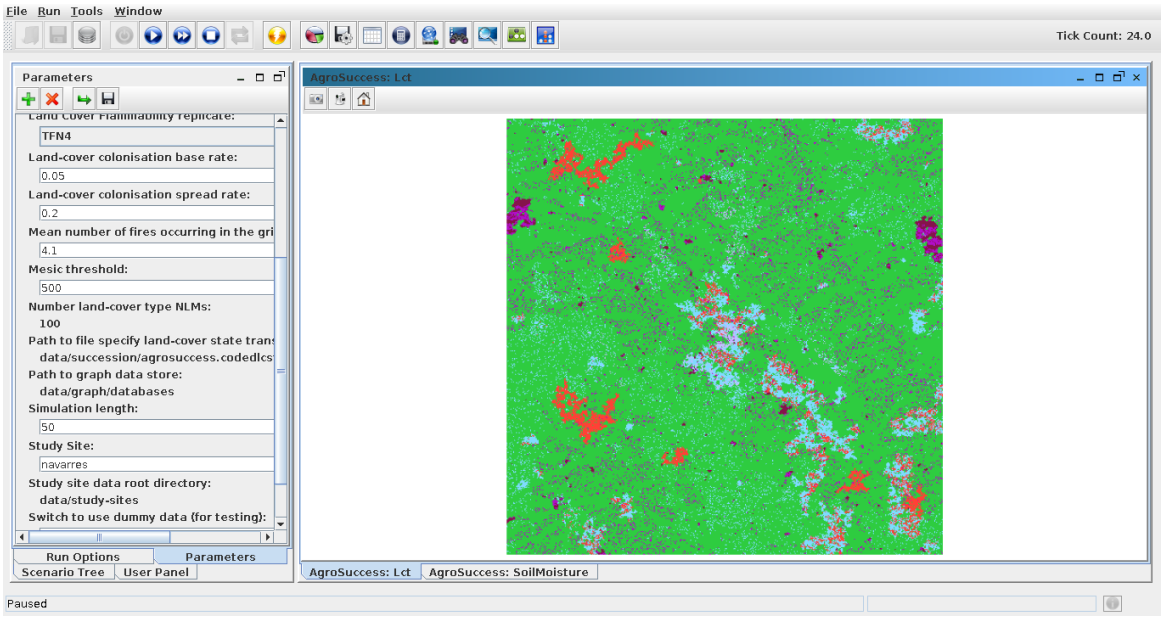


Figure 7.1: Screenshot of the AgroSuccess model running using the graphical interface provided by Repast Simphony (North et al., 2013). The panel on the left provides interactive controls to experimentally adjust model parameters and select study sites. The panel on the left shows a spatially explicit view of the time evolving land-cover in the simulated landscape. See Appendix G.S8 for access to the code needed to reun the AgroSuccess model.

There are other agent-based modelling frameworks available that could have been used in place of Repast, including NetLogo (Tisue & Wilensky, 2004), GAMA (Taillandier et al., 2019), MASON (Luke et al., 2005), and Mesa (Kazil et al., 2020). These options all provide an environment for developing agent-based models, with support for expressing concepts relevant to a given model in a formal programming language, running simulations, and visualising simulation outputs. NetLogo and GAMA provide their own domain-specific languages (DSLs) for specifying ABMs. Conversely MASON, Mesa, and Repast are simulation tool-kits that provide reusable software components useful for specifying agent-based models in a general purpose programming language (Java for MASON and Repast, and Python for Mesa). The use of a DSL helps to reduce the time required for modellers who are not already familiar with a programming language to start developing models. Frameworks that use a DSL are good choices for modellers who are new to programming and want to develop a model quickly without needing to learn a general-purpose language (many of

whose features they may not need to use). Additionally, the documentation for these frameworks often includes self-contained example models that new users can refer to while learning the language. However, because DSLs are niche languages that are designed to support a particular modelling framework, they lack the mature ecosystem of tools for developing and systematically testing the implementation of models written in them that are typically available for established general-purpose programming languages. The additional software tools that are available for general-purpose programming languages compared to DSLs make tool-kits like MASON, Repast, and Mesa better suited for developing modular, well tested software.

Of the simulation tool-kits considered, I favoured the Java-based frameworks over Mesa, which implemented in Python, because of the ease of portability of Java programs. It is desirable for AgroSuccess to be able to run in various remote execution environments (e.g. in the cloud and on institutional HPC infrastructure) to make it convenient to perform large numbers of simulations in parallel. In my experience, it can be difficult to exactly reproduce Python environments on different infrastructures without using a containerisation technology such as Docker (Merkel, 2014) which is not necessarily available on institutional HPC systems. By contrast, Java programs can be packaged with everything they need to run on any system where the Java Virtual Machine (JVM) is available using standard tools. As simulation tool-kits written in Java, Repast and MASON are broadly comparable. However, I found Repast easier to install and start working with. Additionally, the Repast user community and associated forums are active and responsive, making it easy to get help during the development process.

Java programs typically perform well in terms of both run time and memory utilisation compared to an equivalent implementation in an interpreted language such as Python. This stems from the fact that Java programs are compiled into an optimised bytecode format that can then be executed by the JVM. By contrast, programs written in an interpreted language don't benefit from compile-time optimisations. Other compiled languages such as C++ may be able to achieve better run time performance than equivalent programs written in Java. This is because in C++ the programmer is responsible for allocating and deallocating memory, whereas in Java the JVM does this automatically at the cost of additional computational overhead. However, there is a trade off between the run time efficiency of C++ and the additional burden on the developer to manually manage memory in the program. Other modern programming languages such as Rust (Perkel, 2020) my be able to achieve comparable performance to C++ without requiring the programmer to independently guarantee that all memory is managed appropriately. Rust does

this by enforcing correct memory management at compile time. Additionally, Rust simplifies the process of writing *concurrent* programs (that is, programs that use procedures that run simultaneously on different processor cores). This may enable developers of ABMs to better utilise multi-core CPUs to decrease the time needed for each model run.

7.1.2 Adherence to model development standards

Modelling standards help to define a working framework to help scientific modellers create software that is able to meet its scientific research goals and useful to others in the scientific community. An overarching objective of such standards is that of *openness* in the sense that models should be transparent and available for scrutiny by others. An example of the problems caused by lack of openness by modelling scientists came to the public attention during the COVID-19 pandemic. At this time it became apparent that researchers modelling the public health implications of COVID-19 were not universally willing to share their model source code due to concerns about giving away a competitive advantage to other researchers. This resulted in the publication of an open letter in the journal *Science*, arguing that the practice of guarding model source code in this way is fundamentally unscientific and against the public interest (Sills et al., 2020).

The Open Modeling Foundation is an international community of modelling scientists that have established a set of standards (The Open Modeling Foundation, 2023) around Accessibility, Documentation, Interoperability, and Reusability for other scientists to follow to maximise the usefulness of their work. Such a framework helps reduce the arbitrariness of the standards to which scientific software is developed, and to make it easier to discern and evaluate the rigour with which scientific models have been implemented. Improved confidence in the software developed by other scientists, combined with the availability of high quality documentation of that software, is beneficial for the community as a whole because it reduces the perceived need to re-implement models developed by others. The issues addressed by the OMF have been raised by others in the ABM community. Shin, 2021 presents the advantages of openness and transparency around source code used to implement scientific software to both the individual researcher and the community as a whole (‘Accessibility’), Müller et al., 2013 advocate the development of standardised formats of natural language descriptions of models to accompany source code (‘Documentation’), and Polhill and Edmonds, 2007 highlight the importance of providing software licenses for simulation model implementations (‘Reusability’). In the following subsections I describe the ways in

which my implementation of AgroSuccess aspires to address the OMF standards.

Accessibility

All software used to both run AgroSuccess simulation models and generate the required input data from publicly available data sets is available both in the Zenodo software repository and on GitHub. Zenodo is an immutable software repository specifically designed for sharing open scientific software. See listing of available software in Table 7.1 (available software is also listed in Appendix G for convenience of reference). All software is under version control (Scheller et al., 2010; G. Wilson et al., 2014), and the citations given for each piece of software reference a specific version.

Table 7.1: Summary of open source software tools developed to support users working with AgroSuccess, including the AgroSuccess simulation model code itself (`agrosuccess-sim`).

Name	Description	Reference	Appendix
`epd-query`	An application to help extract data from the European Pollen Database in a reproducible and open way	(Lane, 2019)	Appendix G.S1
`aemet-wind`	A Python library for working with wind speed and direction data from the Spanish State Meteorological Agency's weather data REST API	(Lane, 2021a)	Appendix G.S2
`aslib`	A Python library containing software objects used in other code for working with AgroSuccess	(Lane, 2021d)	Appendix G.S3
`agrosuccess-data`	Code used to generate input data for the AgroSuccess simulation model	(Lane, 2021b)	Appendix G.S4
`demproc`	A Python package used to derive raster layers based on a Digital Elevation Model	(Lane, 2021e)	Appendix G.S5
`cymod`	Application used to convert complicated state-and-transition models described in the Cypher graph query languages into Neo4j graph data stores	(Lane, 2020)	Appendix G.S6
`agrosuccess-graph`	A Python package that uses Cymod to represent the AgroSuccess STM	(Lane, 2021c)	Appendix G.S7
`agrosuccess-sim`	The AgroSuccess simulation model code	(Lane, 2023)	Appendix G.S8

Documentation

In addition to the `README` document in the AgroSuccess software repository, the AgroSuccess source code also includes detailed 'Javadoc' documentation of the constituent software compon-

ents (classes and packages). Including documentation as part of the source code in this way can be considered a form of literate programming, i.e. code that is intended to be easy for humans to read (Scheller et al., 2010).

Interoperability

The source code implementing AgroSuccess (Lane, 2023) is organised as a collection of related Java packages. For example, there are separate packages for the succession, fire, and subsistence agriculturalist models. These can be used together as a coherent model, as I did in the preparation of this thesis, or can be recombined to form parts of other models. This improves the likelihood of code reuse. However, I argue a more important outcome of this method of organisation is that it makes the code easier to read. This is because the reader can focus on the software artefacts that are mainly within a single package, making it easier to understand.

Type safety in Java helps improve the interoperability of the source code, because it forces interfaces to be clearly and unambiguously defined. Users of dynamically typed languages such as Python need to rely much more heavily on documentation of software components developed in such languages, and the documentation is not guaranteed to match the implementation.

Reusability

AgroSuccess is implemented using the well known Repast Simphony modelling framework, making it easy for new users to install and use it. It includes a `README` document that describes the required input data, and provides references to the ancillary software tools that I have developed to generate required input data for the study sites considered in this thesis, or potentially other study sites that are of interest to the user. AgroSuccess is licensed under a permissive MIT license so other perspective users should have no legal concerns about reusing the software.

7.1.3 Maintainability and extensibility

In addition to the measures described in Section 7.1.2 to ensure that my work on AgroSuccess is useful to and understandable by other researchers, I have also taken steps to ensure that the software implementation of AgroSuccess itself is *maintainable* and *extensible*. Maintaining

software is the process of modifying existing software to correct faults, improve performance, or adapt it to reflect changes in the specification of the real-world system it represents. Extending software involves adding functionality that was not planned at the time that the software was originally developed. Both of these activities are relevant to the development of software that implements a scientific model. Software implementations of scientific models represent scientific understanding that is itself subject to change. Correspondingly, software implementations of such models should be amenable to modification and extension to enable them to remain useful as scientific understanding evolves.

To assure the extensibility and maintainability of the software implementation of AgroSuccess, I have organised its source code using the SOLID design principles (Martin, 2017). The SOLID principles are a well known set of design principles commonly applied to object-orientated software that are intended to help developers write code that is easy to maintain and extend. These are the 'Single Responsibility Principle, 'Open-Closed Principle', 'Liskov Substitution Principle', 'Interface Segregation Principle', and 'Dependency Inversion Principle'. The following subsections describe these principles and provide examples of how I have used them in AgroSuccess, where applicable.

Single Responsibility Principle

The Single Responsibility Principle (SRP) states that each software component should have exactly one reason to change. In the context of a scientific model, the clearest example of a reason for a component needing to change is an update to the understanding of the aspect of the model that the component implements. In AgroSuccess, software 'components' are implemented as Java classes. For example, the `FarmingPlanCalculator` class is responsible for calculating the number of farm patches required by a household, and the `FarmingReturnCalculator` class is responsible for calculating the mass of wheat obtained by farming a particular patch (see Section 4.3.4). Each of these classes would only need to change if the modelling rules within their well-defined scopes changed, with all other classes remaining unchanged.

Following the SRP helps to promote modularity (Bugmann, 1994) in the sense that all code related to each concept in the conceptual model is logically grouped together, and can be readily scrutinised and verified independently of other model components. This results in software that is easier to understand because each class forms its own unit of functionality.

Open-Closed Principle

The Open-Closed Principle (OCP) states that software components should be open to extension, but closed to modification. Following this principle allows code at a higher level of abstraction to be reliably reused by (potentially) multiple other pieces of code at a lower level of abstraction. Examples of classes that follow this principle in AgroSuccess are `SeedDisperser` and `SpatiallyRandomSeedDisperser`. The `SeedDisperser` class implements the logic needed to manage the data structures that are used to track the location of seeds in the simulation grid (a relatively high level of abstraction). The `SpatiallyRandomSeedDisperser` class *extends* `SeedDisperser` and contains the logic needed to implement the specific land-cover colonisation process described in Section 4.2.4. When developing and testing `SpatiallyRandomSeedDisperser`, the modeller can focus on the logic relating to *how* seeds are to be randomly distributed, while relying on the logic in `SeedDisperser` to manage the required data structures. In future, other modellers or I could add a new class that implements an alternative seed dispersal algorithm that also extends `SeedDisperser`, promoting code reuse.

Liskov Substitution Principle

The Liskov Substitution Principle (LSP) concerns the circumstances under which an object within a program (i.e. an instance of a class) of a given type, T, should be interchangeable with objects that are subtypes of T. It is closely related to the concept of *inheritance* in object-orientated programming. AgroSuccess does not currently make use of inheritance in a way that facilitates an illustration of the LSP. I refer the reader to the literature on the topic (e.g. Martin, 2017) for further information.

Interface Segregation Principle

The Interface Segregation Principle (ISP) states that classes should only depend on interfaces that they actually use. In Java, an 'interface' can be thought of as an abstract contract that specifies the set of methods which a class that implements a given interface must provide. That is, an interface specifies *what* an implementing class does. On the other hand, a class is a specific, concrete implementation that specifies *how* it achieves what the interfaces it implements require (see e.g. Schildt, 2007 for details of the relationship between interfaces and classes in Java).

Consider, for example, the `SiteAllData` interface in AgroSuccess. Classes implementing this interface must provide methods for retrieving all data associated with a study site (slope, flow direction, and land cover type maps, wind data, precipitation data, etc.). This interface is implemented by the `SiteDataLoader` class which reads required data from the file system. No classes that consume site data require access to all the data provided by the `SiteAllData` interface, so it is *segregated* into distinct `SiteClimateData`, `SiteMetaData`, `SiteRasterData`, and `SiteWindData` interfaces. This allows, for example, the `initSeedDisperser` method in the `AgroSuccessContextBuilder` class to depend only on the data provided by the narrowly scoped `SiteRasterData` interface. This means that even if other parts of the `SiteAllData` interface change in future (e.g. to change the way that wind data is consumed by the model), the `initSeedDisperser` method can remain unchanged, aiding maintainability.

Dependency Inversion Principle

The Dependency Inversion Principle (DIP) states that, where possible, classes should depend on abstract interfaces rather than concrete implementations (classes). As explained above in the description of ISP, interfaces specify what implementing classes do, whereas classes specify how they achieve what the interfaces they implement require. Following this principle simplifies extensibility by making it straightforward to exchange one class that implements a particular interface for another that implements the same interface, but using different logic.

In AgroSuccess, the `DefaultVillage` class requires a set of `Household` objects as input (i.e. the households that make up the village). `Household` is an interface that specifies that implementing classes must have methods to calculate a subsistence plan, claim land patches to use to fulfil that subsistence plan, and update their population. In the current version of AgroSuccess, we use the `DefaultHousehold` class to implement the logic described in Section 4.3.3. If I wanted to extend the model by including pastoralism as well as agriculture in households' subsistence planning, I could add a new `PastoralHousehold` class that implements the additional logic required for pastoralism and also conforms to the `Household` interface. Because the `DefaultVillage` class depends on the `Household` interface rather than the `DefaultHousehold` class (in accordance with DIP), I could simply provide the `DefaultVillage` class with `PastoralHousehold` objects instead of `DefaultHousehold` objects without making any changes to the `DefaultVillage` class. It would also be possible to specify that a particular model run should use either `DafaultHousehold` or

`PastoralHousehold` objects through a model parameter, simplifying the execution of simulated experiments to explore the impact of the addition of pastoralism on simulation outputs.

7.1.4 Model implementation testing

To provide assurance that the software used to implement the AgroSuccess simulation model is free from programming errors, I have made use of extensive automated testing of its source code. Use of automated testing is a form of model verification, specifically what Augusiak et al., 2014 call 'implementation verification'. Model verification is one aspect of the broader exercise of model evaluation (Manson et al., 2012) or 'evaludation' (Augusiak et al., 2014). Another aspect of model evaluation is *validation*. The terms 'verification' and 'validation' are used inconsistently or even interchangeably in the literature (Augusiak et al., 2014; Oreskes et al., 1994). To be explicit, by 'verification' I mean any activity that provides assurance that the software implementation of a model accurately reflects the conceptual model that it is intended to represent, and is behaving as intended. By 'validation' I mean the comparison of model outputs to real-world empirical data to determine whether the model outputs are sufficiently accurate for it to be used for a particular purpose (Augusiak et al., 2014; Manson et al., 2012; Schmolke et al., 2010). In addition to the work towards implementation verification discussed in this section, the sensitivity analysis presented in Section 6.1 is also a form of model verification.

Automated testing in scientific software

Automated testing is a routine practice in modern software development that helps developers ensure that their software works as intended. Others in the literature have noted that automated testing has been historically underutilised by scientists who develop software (Scheller et al., 2010; G. Wilson et al., 2014; G. V. Wilson, 2006).

There are two main categories of automated tests (Scheller et al., 2010; G. Wilson et al., 2014). *Unit tests* are used to verify that individual logical units of code, such as functions or methods, are behaving as expected. For example, a unit test might call a class method that implements an equation defined in the conceptual model and confirm that it returns the correct value for a range of input values. Unit tests are often written at the same time as (or even before) the code they are designed to test, and serve as a permanent record of how the code under test is

intended to behave. *Integration tests* are used to confirm that multiple software components, such as classes, are working together as expected. For example, they might be used to test that data is being correctly communicated from one software object to another. Both unit tests and integration tests should be run frequently. During development, it is possible for a new change to cause existing code to stop working as expected. Running automated tests can help to identify these issues as soon as they occur, at such a time that it is clear to the developer which change caused the code to stop working. Timely identification of problems like this helps to decrease development time and improve code correctness. This mode of using automated tests is known as *regression testing*.

Scheller et al., 2010 also identify *system testing* in which the software system (i.e. model) as a whole is tested to confirm that it is behaving as expected, e.g. by analysing the model outputs. However, I have found that this type of testing usually involves some degree of critical evaluation on the part of the modeller and therefore, while important, is difficult to incorporate into an *automated* testing workflow.

The development of automated tests leads to two additional benefits besides the timely identification of defects. First, writing code that is easy to test naturally leads to more modular designs because it encourages the developer to create each component (class or function) in isolation and confirm that it is behaving as expected before integrating it with other components (G. Wilson et al., 2014). Such modular designs lead to the production of components that are easier to reuse, and the tests themselves serve as examples to other developers of how these components are used. Second, the test code itself can be reviewed by collaborators to confirm that the model code has been confirmed to work in the ways required by the conceptual model specification. For example, if a collaborator notices that the code that tests a method implementing an equation doesn't test the full range of valid input values to that equation, they could request that the test code be updated to confirm that the method under test continues to behave as expected for all valid input values.

Testing in AgroSuccess

The AgroSuccess model implementation is a large piece of software (over 12,000 lines of Java), with many interdependent software objects. It represents a complex system whose outputs are emergent quantities resulting from the interactions between its constituent components. It is a

good example of a project in which the correctness of its implementation cannot be inferred from analysis of its outputs alone. It therefore benefits from the use of automated testing to provide verification of its implementation.

All code used to run the automated tests for AgroSuccess can be found in the `src/test/java` directory of its software repository on Zenodo (Lane, 2023). AgroSuccess uses the popular JUnit testing framework for Java to implement both its unit tests and integration tests. It includes a total of 264 individual automated tests that verify the implementations of 86 Java classes. An example of a unit test can be seen in `FarmingPlanCalculatorTest`. This demonstrates the correctness of the implementation of Eq. (4.18) as specified in Section 4.3.4. An example of an integration test can be seen in `AgroSuccessLcsUpdaterTest`. This tests that land cover types of patches in a 3x3 test grid evolve from a known start state in the expected way, and verifies the correct interaction between four separate classes: `SeedStateUpdater`, `SuccessionPathwayUpdater`, `AgroSuccessLcsUpdateDecider`, and `AgroSuccessLcsUpdater`.

Writing software to implement automated tests in addition to the software needed to implement the model itself requires additional effort compared to writing the model implementation alone. However, I argue that the increased speed with which automated tests allow a developer to identify and resolve defects in the software could plausibly decrease overall development time. Additionally, the process of writing tests encourages the development of modular software that is easier to understand, reuse, and verify.

7.2 Graph database representation of land-cover STM

7.2.1 Introduction to Cymod

State-and-transition models (STMs) are a type of conceptual model used in the environmental sciences to organise information about ecosystem and landscape change (Bestelmeyer et al., 2017). STMs have been used to model ecological (Batllori et al., 2017; Daniel et al., 2018; McIntosh et al., 2003; J. D. A. Millington et al., 2009; Spooner & Allcock, 2006) and geomorphic (Phillips & Van Dyke, 2017) systems, and can encode knowledge about both natural (e.g. McIntosh et al., 2003) and anthropogenic (e.g. Spooner and Allcock, 2006) drivers of environmental state change. This modelling approach is particularly useful to explore how vegetation communities

change between discrete states over time in situations where quantitative data are scarce, or where causes and constraints of change are incompletely known but qualitative understanding about change is available (McIntosh, 2003; McIntosh et al., 2003), or both.

The basis of all STMs is a set of discrete states and a specification of all possible transition pathways between those states (Daniel et al., 2016), often summarised as a diagram. These diagrammatic representations naturally take the form of graphs, with nodes representing states and edges representing transitions (e.g. Figure 1 in McIntosh et al., 2003). Although most modellers using STMs instinctively use diagrams depicting graphs, many do not explicitly state that they are doing so, nor do they exploit the possibility of performing graph theoretical analysis on their models (e.g. Daniel et al., 2018; Spooner and Allcock, 2006). While sketches using a graph structure may be used during the development of an STM (e.g. McIntosh et al., 2003), myriad relationships between the states due to alternate processes of change can produce a complicated model structure that resists coherent rendering on a 2-dimensional surface such as a piece of paper or a computer screen. Consequently, attempts to visualise (and therefore reason about) the entire integrated STM can be challenging. By making it easier to formalise STMs as graphs that are queryable data (rather than as graphics), researchers will be able to develop more sophisticated visualisations and further analyse their models.

With the aim of supporting modellers in expressing STMs as computer-readable graph data structures, I have developed an open source Python package called Cymod (Lane, 2020). Available to download from PyPI, Cymod provides an Application Programming Interface to encode knowledge of STMs to facilitate modelling and analysis (Fig. 7.2). It prompts a workflow for translating the kinds of sketch diagrams already used as conceptual models when developing STMs into a formal, queryable and visualisable data structure. In turn, this makes the iterative development of STMs (potentially with non-scientific stakeholders, such as land managers) quicker, while also helping to avoid logical errors.

The motivating idea behind Cymod is that STMs can be represented using a property graph data structure (I. Robinson et al., 2015). Property graphs generalise the concept of a graph by associating data with its nodes and relationships over and above what can be expressed in a conventional directed, weighted graph. Their utility for organising data describing related concepts in a flexible way make property graphs a natural choice of data structure to capture the ontological richness of complex ecological systems. Focusing on the concept of a ‘view’, Cymod helps mod-

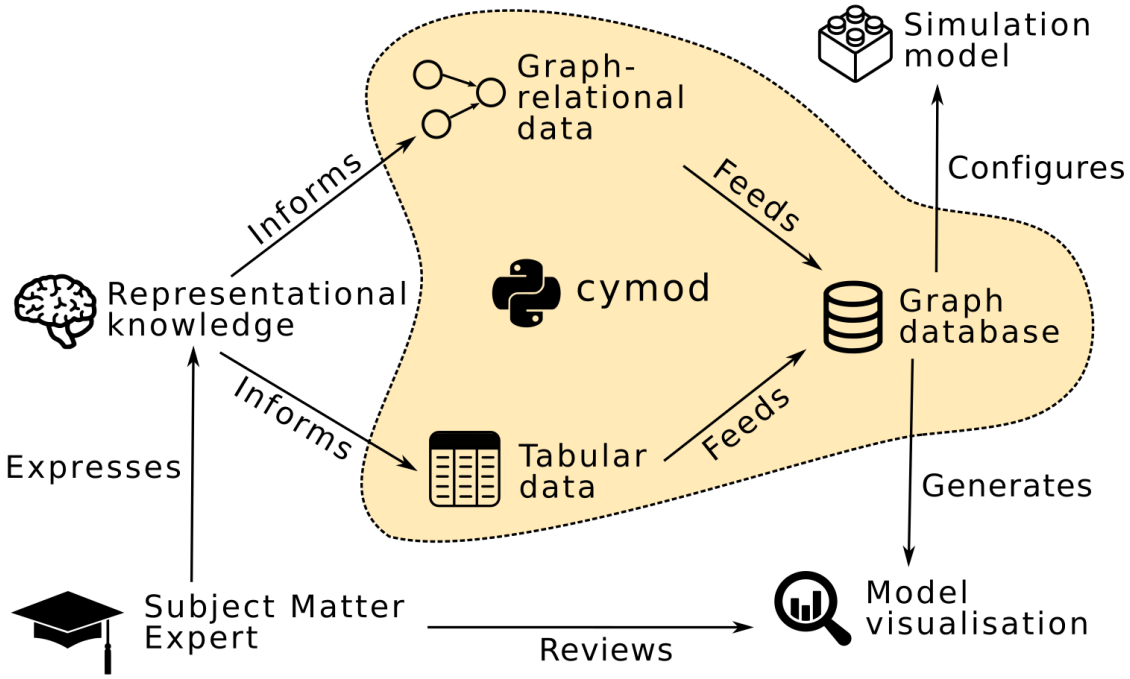


Figure 7.2: The relationship between Cymod, subject matter knowledge, and graph database tools.

ellers transcribe diagrammatic and narrative understanding into a structured, queryable graph database. This database structure also simplifies the steps needed to derive statistics about model structure (Drescher & Perera, 2010; Phillips, 2011), and to use the knowledge encoded in the STM in simulation models (Daniel et al., 2016; McIntosh et al., 2003; J. D. A. Millington et al., 2009). Here, I present a description of what Cymod does, demonstrate its use through a case study, and then discuss further how it is useful for facilitating research using STMs and simulation models based upon them.

7.2.2 What does Cymod do?

Cymod prompts a workflow for specifying complicated STMs

Models describing systems which exhibit complexity often have complicated structures. The Cymod workflow helps modellers manage model complicatedness by breaking the model down into individual 'views', where each view represents an individual aspect of the model which modellers can comfortably reason about and sketch (e.g. on a piece of paper). For a relatively simple model, a view might specify the set of all possible land cover types in a model (e.g. Figure 3 in Daniel et al., 2016), whereas for a more complicated model views might correspond to different

```
MERGE (:LandCoverType {name: "shrubland"})-[:TRANSITIONS_TO]→
      (:LandCoverType {name: "pine forest"})
```

Figure 7.3: Cypher query expressing the concept that shrubland transitions into pine forest.

transition pathways in different locations of a study area due to variations in processes of change (e.g. Figure 7 in McIntosh et al., 2003). The set of views which is appropriate for any particular model is entirely at the discretion of the modeller. This approach to representing complicated models as collections of simpler components improves understanding of the model, as well as facilitating communication during the conception phase of model development.

Cymod supports the use of Cypher as a modelling language

With the conceptually manageable model views established, Cymod then supports mapping these views directly into a computer readable format via the Cypher language. Cypher is a declarative language designed specifically for describing connected data, such as STMs. Cypher uses ASCII art syntax so that the textual representation of the model matches closely to the 'box-and-arrow' drawings which evolve naturally when describing models on paper (McConnell et al., 2011). In Cypher, nodes are represented with pairs of opening and closing parentheses, and edges are represented with arrows drawn with hyphens and angle brackets—see Fig. 7.3 as example. The approach of using a declarative formal language in this way means that the computer model uses the same terminology and semantics as the practitioners in the subject domain.

While Cypher helps the modeller to organise information about related model concepts into code snippets (or 'queries' in database terminology), Cymod aims to support the modeller in organising the sets of interrelated queries needed to describe complicated models. Thus, Cypher queries representing individual transition pathways are in an organisational level below Cymod views (which, in turn, represent multiple pathways).

In a typical Cymod workflow, a modeller will write all the Cypher queries needed to produce a single view of the model in a plain text file. In this way an arbitrarily complicated model can be described piecemeal. Cymod then wires the individual files (views) together to form a single cohesive model within the graph database. Consequently, while the integrated model may be too complicated to represent clearly in its entirety on a 2-dimensional surface, the modeller does not need to contend with that complicatedness while they are describing the model. A further advantage of using separate files to specify each model view is that each file can be individually

placed under version control using software such as Git, enabling modellers to see which parts of their models have changed over time. This simplifies the process of retrieving historical versions of models, and supports reproducibility (G. Wilson et al., 2017). Finally, as models are stored in a queryable database, it is possible to design automated tests which confirm that the implemented model corresponds with the modellers' intentions.

Cymod helps modellers leverage the Neo4j graph database platform

Using Cymod in the workflow outlined above, modellers will have expressed their model in a way that can be consumed by graph database and analysis software. Cymod produces outputs that can be used by a Neo4j database and associated tools in the Neo4j platform to visualise and further analyse an STM. The Neo4j Community Edition graph database system is cross platform (running on the Java Virtual Machine) and open source (distributed under GPL v3). The Neo4j Browser is particularly useful for STM development as it provides an interactive graphical interface to the Neo4j database that enables users to visualise the results of Cypher queries entered into the browser, and to interactively 'click-and-drag' the returned nodes and relationships. In particular, users can drill down into data associated with nodes and relationships and expand nodes to explore their connections.

The utility of being able to visualise the data associated with nodes and relationships highlights the significance of Neo4j databases storing property graphs and demonstrates advantages over the 'tables and rows' structure of traditional relational databases (e.g. I. Robinson et al., 2015). Within an ecological modelling context, a node might represent a land cover state, an edge might represent a transition, and each transition might be mediated by multiple environmental factors (e.g. soil moisture, seed presence/absence). Representing such relationships as a graph is more intuitive and allows more powerful querying than using a table (e.g.J. D. A. Millington et al., 2009; Spooner and Allcock, 2006) or relational database. Furthermore, whereas a weighted graph would only allow encoding of a single number associated with an edge (e.g. transition probability or time to transition), using a property graph enables modellers to encode the full ontological richness of the system as understood in the subject domain.

7.2.3 Illustrative application

In this section I discuss how I used Cymod during the development of the AgroSuccess model (Chapter 4). This provides an example of how Cymod can be used in practice, including concrete examples of the visualisations that can be produced using the Neo4j platform. The RBCLM ecological succession model in AgroSuccess includes 15 possible transitions between pairs of land-cover type states. Additionally there are 8 land-cover type transitions that occur due to anthropogenic 'ecosystem engineering activities'. That is, agricultural households modifying land-cover to convert patches of land to wheat agriculture.

Specifying the AgroSuccess model using Cymod

As noted above, model 'views' can be established in different ways for any particular model, at the discretion of the modeller. I use four views to specify the AgroSuccess model in Cymod. Of these two specify the attributes of the modelled anthropogenic agents (`Agent_w.cql`) and land-cover types (`LandCoverType_w.cql`), one specifies the land-cover modifying activities of the anthropogenic agents (`activities_w.cql`), and one provides a summary visualisation of all transitions due to ecological succession (`visualisation_summary_w.cql`). This conceptualisation suggests a natural hierarchy of views insofar as views representing the means by which transitions between states occur depend on concepts expressed at a lower level of abstraction (the land cover states themselves). I summarise the conceptual hierarchy as follows:

- Level 1: All the land cover states and anthropogenic agents involved in the model including both unique identifier codes and detailed human readable descriptions of the land cover states.
- Level 2: Transitions between land cover states (both natural and anthropogenic), e.g. pine forest to transition forest due to sufficiently hydric conditions.
- Level 3: Summary nodes useful for visualising the completed model.

These levels and the model components expressed within them can be mapped into Cypher files such that all the code used to specify a view are contained within a single Cypher file (with extension `.cql`, see Fig. 7.4). None of the Cymod views named above encode the ecological

```
views
├── abstract
│   ├── Agent_w.cql
│   └── LandCoverType_w.cql
├── landcover_change
│   └── AgroPastoralist
│       └── activities_w.cql
└── succession
    └── visualisation_summary_w.cql
scripts
├── clean_millington_trans_table.py
├── load_agrosuccess_model.py
├── repurpose_trans_rules_agrosuccess.py
└── summarise_millington_table.py
```

Figure 7.4: Listing of files (provided in supplementary materials, see Appendix G.S7) used to specify the AgroSuccess AgroSuccess RBCLM model and load it into a Neo4j database using Cymod.

succession rules. AgroSuccess uses a version of the RBCLM developed for the Millington LFSM, modified to include the additional required wheat agriculture-related land-cover states. We used Python scripts (`clean_millington_trans_table.py`, `summarise_millington_table.py`, and `repurpose_trans_rules_agrosuccess.py`) to parse the transition rules expressed in the supplementary materials provided by J. D. A. Millington et al., 2009, transform them appropriately. Finally a Python script (`load_agrosuccess_model.py` in Fig. 7.4 and Appendix G.S7) using Cymod as a library loads these files into a Neo4j database.

Visualising and analysing the model with the Neo4j platform

Once the Cypher files have been written and loaded into a Neo4j database using Cymod to process and load the model files into the database, we can leverage the Neo4j Browser to interact with the database via a graphical interface. In Fig. 7.5 I show several examples of graphical visualisations extracted from the tool. Creating visualisations (through database queries) that present subsets of the whole model should facilitate better analysis and communication. For example, Fig. 7.5.B shows an example of how a user exploring the model can search for anthropogenic activities involving a particular land-cover type (pine forest, in this case). Alternatively, Fig. 7.5.C illustrates how the tool enables modellers to visualise ecological and anthropogenic land cover change processes simultaneously, while limiting the view to transitions involving the wheat land-cover type. This helps to reduce the number of elements in the resulting diagram sufficiently that it can be easily studied.

As stated above, the AgroSuccess STM derives its ecological state transition rules from those

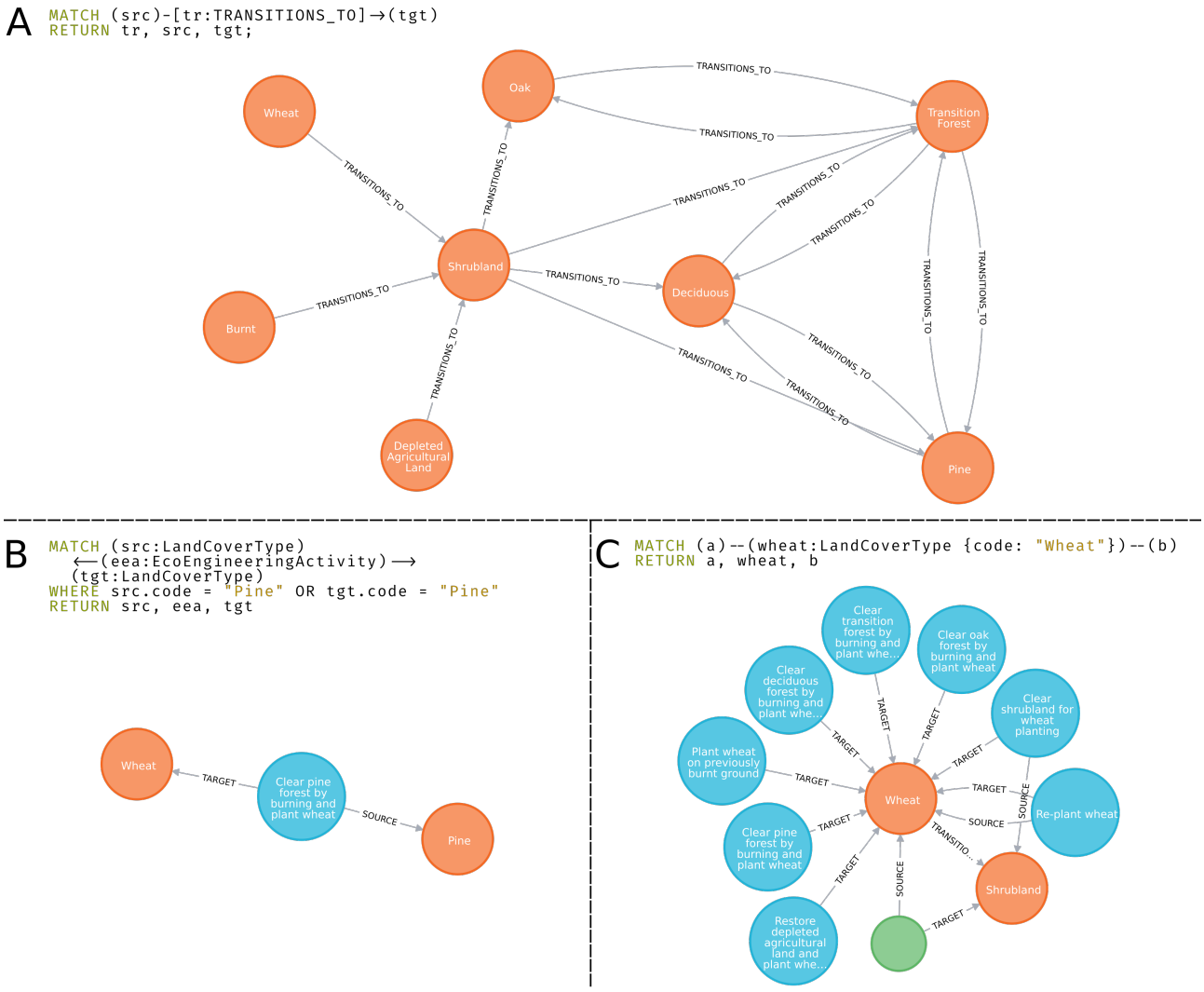
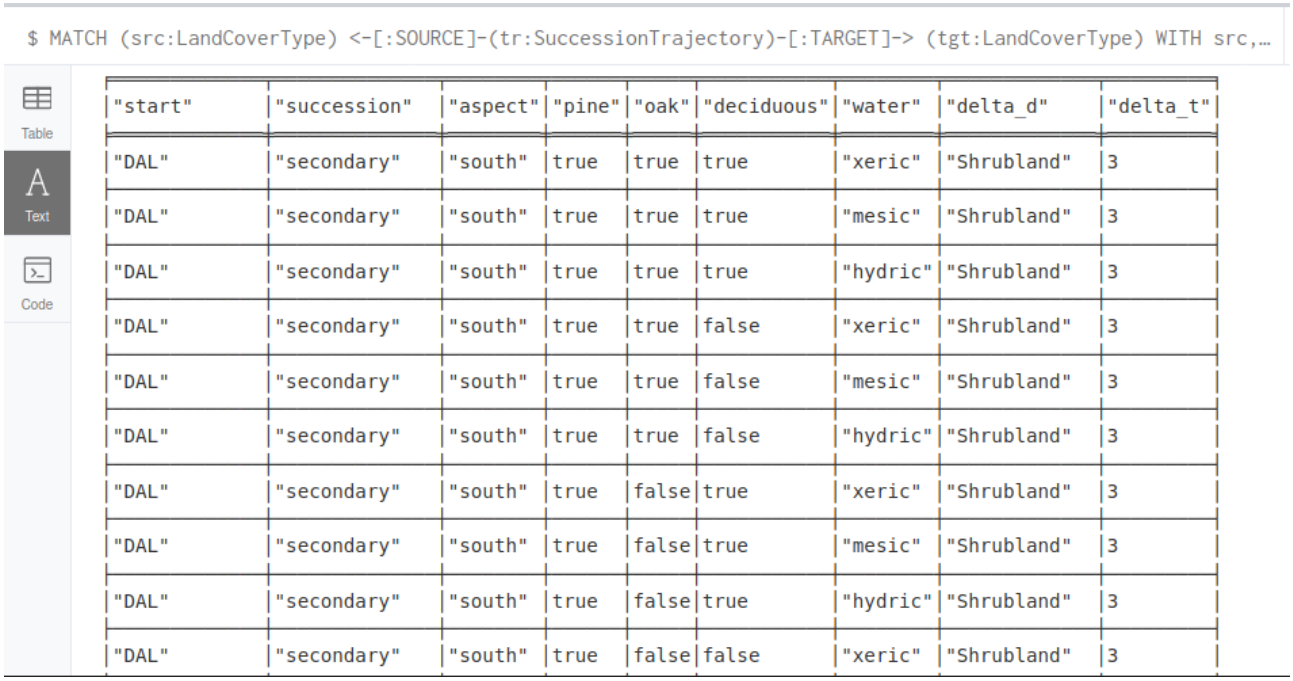


Figure 7.5: (A) A summary visualisation generated via a user query in the Neo4j Browser tool, showing all land-cover transitions due to ecological succession represented in the AgroSuccess STM. (B) A visualisation showing anthropogenic activities that modify land-cover involving transition either to or from Pine forest. (C) A visualisation showing how ecological succession and anthropogenic activities interact to alternately to cause land-cover to transition to or from wheat agriculture.

```
MATCH (src:LandCoverType)
  <-[:SOURCE]-(tr:SuccessionTrajectory)-[:TARGET]->
  (tgt:LandCoverType)
WITH src, tr, tgt
MATCH (cond:EnvironCondition)-[:CAUSES]->(tr)
RETURN src.code AS start, cond.succession AS succession, cond.aspect AS aspect,
  cond.pine AS pine, cond.oak AS oak, cond.deciduous AS deciduous,
  cond.water AS water, tgt.code AS delta_d, cond.delta_t AS delta_t
```

Figure 7.6: Cypher code used to query the graph database to produce a tabular structure where each row specifies a possible ecological succession pathway. This is analogous to the look-up table that specifies the ecological succession model of the Millington LFSM provided in the supplementary materials of J. D. A. Millington et al., 2009.



"start"	"succession"	"aspect"	"pine"	"oak"	"deciduous"	"water"	"delta_d"	"delta_t"
"DAL"	"secondary"	"south"	true	true	true	"xeric"	"Shrubland"	3
"DAL"	"secondary"	"south"	true	true	true	"mesic"	"Shrubland"	3
"DAL"	"secondary"	"south"	true	true	true	"hydric"	"Shrubland"	3
"DAL"	"secondary"	"south"	true	true	false	"xeric"	"Shrubland"	3
"DAL"	"secondary"	"south"	true	true	false	"mesic"	"Shrubland"	3
"DAL"	"secondary"	"south"	true	true	false	"hydric"	"Shrubland"	3
"DAL"	"secondary"	"south"	true	false	true	"xeric"	"Shrubland"	3
"DAL"	"secondary"	"south"	true	false	true	"mesic"	"Shrubland"	3
"DAL"	"secondary"	"south"	true	false	true	"hydric"	"Shrubland"	3
"DAL"	"secondary"	"south"	true	false	false	"xeric"	"Shrubland"	3

Figure 7.7: Output from Neo4j Browser resulting from the query given in Fig. 7.6

used in the Millington LFSM (J. D. A. Millington et al., 2009). The state transition rules for that model were recorded in a tabular format and published as supplementary materials. The visualisation we have developed showing all ecologically driven state transitions (Fig. 7.5.A) presents a summarised view of the information contained in such a table but offers an alternative perspective that can bring potential insights that are difficult to observe from a tabular format. For example, Fig. 7.5.A highlights the cyclical relationship between the transition forest land-cover type and the oak, deciduous, and pine land-cover types, as well as the high betweenness centrality of the shurbland land-cover type. Furthermore, the original tabular structure can be recovered from the model implemented in Cypher by entering the Cypher code given in Fig. 7.6 into the Neo4j Browser, producing the output shown in Fig. 7.7.

7.2.4 Discussion

Cymod for facilitating iterative STM development

Much of the utility of state-and-transition models (STMs) stems from their representational flexibility, allowing them to capture detailed knowledge about the causes of environmental change. Bestelmeyer et al., 2017 suggested there are three emerging ways in which STMs are being used that exploit this flexibility: i) incorporation of participatory approaches into STM development (often for community-based management), ii) using STM for structured decision-making (by ecosystem managers), and iii) combined use of STMs with new spatial data. Alone, Cymod offers clear benefits to modellers working with non-modellers either to develop or use STMs (contributing to the first two ways), and when linked to a simulation model can contribute to the use of spatial data (addressing iii; see the following section).

Local community members and ecosystem managers are often able to provide both specific information (e.g. which states and transitions pathways exist in the system) and quantitative estimates based on expert judgement (e.g. how long each specific transition is expected to take). The flexibility of STMs makes them extremely useful during the stage of the modelling process in which modellers elicit expert knowledge using explorative "what if. . . ?" questioning while sketching 'box-and-arrow' conceptual models (e.g. see McConnell et al., 2011; McIntosh, 2003). These sketches can then serve as the starting point for subsequent thought and discussions as part of an iterative model development process which culminates in a completed STM. The Cymod workflow prompts iterative model development in two key ways. First, the suggested manner of breaking down detailed knowledge about complicated systems into simpler components which are easy to understand and visualise helps stimulate clarifying discussions between modellers and non-modellers, encouraging consistency of understanding. Second, as a consequence of the close mapping between the conceptual model and the graph data structure which can be achieved by expressing the model in Cypher, incremental changes decided upon as a consequence of these conversations can be easily transferred from paper sketches to the model code (and these changes formally tracked via versioning tools such as Git).

Cymod for translating STMs into simulation models

STMs can be used as the basis for computer simulation models whose outputs can be interpreted as time-evolving realisations of the modelled system. These simulation models are known as state-and-transition simulation models (STSMs) and are distinct from the underlying conceptual model (the STM). Cymod itself does not enable the execution or running of simulations (STSMs) but focuses instead on supporting the development and visualisation of STMs. Numerous examples of STSMs exist in the literature (e.g. Batllori et al., 2017; J. D. A. Millington et al., 2009) and frameworks, such as ST-Sim (Daniel et al., 2016), have been developed to help modellers implement such simulations. Tools such as ST-Sim are valuable to help modellers implement their STMs and run them as STSMs, but they provide only limited structure for the development of the STM itself (e.g. indicating which transition probabilities need to be specified once states have been established).

By capitalising on the natural correspondence between STMs and graph data structures, Cymod not only provides new ways to visualise STMs (as demonstrated above) but also provides a formal means to implement STMs as STSMs. This is because by representing STMs formally as graphs, Cymod creates data structures that computer programs—including simulation models—can interact with. Thus, although Cymod does not by itself enable simulation of STMs, its role in loading models specified using Cypher into a Neo4j database helps modellers interact with their STM programmatically via a database connection. This is particularly desirable when an objective of simulation is to integrate spatial data for a specific geographic location (e.g. to investigate land cover change scenarios). Such functionality is possible because considering an STM as a graph allows it to be queried, manipulated and visualized in flexible ways using multiple software tools. The Cymod package supports a workflow for STMs considered as graphs, enabling modellers to do the conceptual work in advance (as usual) and then facilitating a rendering of the model as a graph structure that can readily be consumed by other tools such as the Neo4j system for further visualisation and analysis.

Comparison to other graph-based modelling approaches

In the preceding sections I described how Cymod helps users to visualise STMs and to integrate them into simulation models to produce STSMs. STMs are just one example of how graph

data structures have been used in ecological modelling. Here, I review three other modelling approaches that use graph data structures (Bayesian belief networks, decision trees, and causal loop diagrams) and contrast their use cases with STMs that can be developed with the assistance of Cymod.

Bayesian belief networks (BBNs) are directed acyclic graphs (DAGs) that encode the probabilistic causal relationships between a set of variables. Nodes in the graph represent variables, and an edge between a pair of nodes indicates that the target node's variable is statistically dependent on the source node's variable. Each node is associated with a conditional probability distribution that describes the likelihood of observing each of the possible values of the node's variable *conditioned on* all of the variables it is statistically dependent on. BBNs can be thought of as having two conceptual components (Aguilera et al., 2011): a qualitative component that is the overall graph structure, and a quantitative component that is the probability distribution associated with each node. Once a BBN has been designed, it can be used to predict the value of a 'goal variable' in a way that takes into account all available information about each of the other variables in the BBN that the goal variable is statistically dependent on (Aguilera et al., 2011).

BBNs have been used in ecological modelling to represent complex systems in which the effects of variables on other variables in the system are uncertain and acknowledged as such by relevant informed stakeholders. Under these circumstances, BBNs can be used as part of a participatory modelling approach in which both the qualitative structure of the BBN (i.e. which variables are statistically dependent on other variables) and the quantitative relationships between the variables are subject to discussion among various stakeholders. For example, Salliou et al., 2017 developed a BBN to represent the causal relationships between use of pesticides, prevalence of aphids, and crop yields.

In an STM, each of the nodes represents a possible state of a vegetation community in a particular location, and edges represent the set of environmental circumstances under which a state transition will occur with certainty between the two nodes connected by the edge. Transitions between states is *deterministic* with respect to the specific combination of environmental conditions associated with the edge. Stochastic variation in an STSM that is integrated with the STM arises from the particular simulated processes represented in the STSM that lead to changes in the cell's environmental conditions. This is distinct from a possible alternative modelling approach involving a BBN in which vegetation communities *probabilistically* transition between

land-cover states according to conditional distributions over the environmental conditions in the area occupied by the vegetation community. In that scenario, the land-cover state of the vegetation community would be a single variable in the BBN that is statistically dependent on variables representing all the environmental conditions represented in the model.

Decision trees are predicative models that can be used to predict the value of one or more target variables given a set of input variable values (attributes). They are a type of supervised machine learning technique and can be used for both classification and regression problems. As in other supervised machine learning approaches, decision trees are trained by providing an algorithm with a dataset that includes both attributes and their corresponding target variables as input. The output of the training algorithm is a tree-like graph in which each non-leaf node is a ‘decision‘, and each leaf node is a possible predicted value of the target variable (or set of values if multiple target variables are sought). A decision is a function that takes the value of a particular data attribute as input and returns another node in the graph as output. When using a trained decision tree for prediction given new data, the prediction algorithm starts by inspecting the root node and comparing the data attribute associated with its decision to the corresponding attribute in the input data to determine the next decision in the sequence determined by the graph. This process is repeated until a leaf-node that is associated with a predicted target value is reached (Debeljak & Džeroski, 2011).

The structure of the graph produced by the decision tree algorithm, including which decision is applied at each node, is ‘learned‘ automatically by the algorithm during the training step. This makes decision trees useful in scenarios where modellers don’t have prior expert knowledge about how attributes influence the values of target variables, and they wish to infer this information from training data. Additionally, the structure of generated decision trees can be inspected by modellers and decision makers, enabling them to ‘sense check‘ the decisions that are used during the prediction process. Because of this, decision trees are *understandable* in a way that other supervised machine learning approaches (e.g. neural networks) are not (Kotsiantis, 2013). Decision trees have been used in ecology to model, for example, population dynamics (Debeljak & Džeroski, 2011).

Decision trees provide a way to infer the relationships between input attributes and target variables in an automated data-driven way that makes use of a graph-based data structure. Their primary purpose is to *predict* the value of target variables given new data attributes. This is

fundamentally different to the objective of an STM, that is, to encode existing expert knowledge about state transitions among vegetation communities in a form that is structured and easy to visualise.

Causal loop diagrams (CLDs) are analytical tools from the field of systems dynamics (Lin et al., 2020). In the context of systems dynamics, a system is a collection of components that interact with each other to produce feedbacks. A useful systems dynamics model should include all the components that are required to produce the dynamic behaviour of interest. That is, such models should be self-contained and not rely on any external drivers to produce their dynamics (Richardson, 2011). One example of such a system is a Lotka–Volterra predator-prey model in which the components are the populations of predator and prey species (Wangersky, 1978). A CLD of this system would take the form of a graph with two nodes–the number of predators and the number of prey–and two directed edges connecting the number of predators to the number of prey and vice versa. The edge flowing from prey to predator would be associated with a 'positive polarity' because increasing prey numbers *causes* increasing predator numbers. Conversely, the edge flowing from predator to prey would have a 'negative polarity' because increasing predator numbers decreases prey numbers (Lin et al., 2020). Constructing a CLD of this system helps to visualise the balancing feedback loop that exists between predator and prey numbers in the model. CLDs can be used to configure simulations of the modelled systems directly using commercial software tools such as AnyLogic® or Vensim®, for example.

In a CLD, a graph structure is used to describe how the values of all the components in a system change over time. Conversely, an STM specifies how a vegetation community could transition between each of its possible states at any given point in time. In an STM there is only one 'variable' under consideration–the state of the vegetation community in a particular location. An STM can, however, be combined with a simulation model to influence system-level dynamics.

7.3 Challenges arising during model implementation

Here I summarise some key challenges that arose during the implementation of AgroSuccess. These relate to the inherent difficulty in reusing concepts and source code from complicated models that have not been developed with the kind of modelling standards described in Section 7.1.

7.3.1 Integrating features of the MedLand model

The availability of the description of the MedLand model in Ullah, 2013 was of great importance during the development of this thesis. However, as noted in Section 4.3.2, to develop a full understanding of the model it was necessary to trace its history through nine separate publications. This highlights a major advantage of centralising model documentation in the way that is advocated by the OMF (The Open Modeling Foundation, 2023).

Unfortunately, I was not able to run the MedLand model or obtain up-to-date source code for it, despite contacting the original authors. There is a listing for the MedLand model on CoMSES Catalogue (CoMSES Net, 2020), a cross-disciplinary catalogue of agent-based and individual-based models that is maintained by the CoMSES Network. However, it was not possible to reconcile the version of the source code in the CoMSES Catalogue with the model description in Ullah, 2013. This is an example of how immutable software repositories like Zenodo that provide specific references for specific software versions are valuable. The inability to run or access the source code for MedLand led to serious inefficiencies in my work, because it was necessary to reimplement all model concepts myself. I intend for my efforts towards assuring the interoperability of AgroSuccess (see Section 7.1.2) to avoid a similar situation arising with respect to AgroSuccess in the future.

7.3.2 Interpreting the Millington LFSM

While implementing the environmental submodel in AgroSuccess, I noted a number of aspects of the description of the Millington LFSM that required clarification from the original author or correction. First, the specification of the soil moisture model in J. D. A. Millington et al., 2009 does not make explicit the spatial dependency between grid cells (how water flows from one cell to another). E.g. in Eq. (4) in J. D. A. Millington et al., 2009 the runoff is expressed in terms of the precipitation entering the cell, not the total water entering the cell. Clarification of this issue required a review of the original (unpublished) model source code in collaboration with the original author. I also found there was an inconsistency in the derivation of Eq. (5) in J. D. A. Millington et al., 2009 from the relevant equations in United States Department of Agriculture, 2004. This is corrected in Eq. (4.11) in this thesis. Second, in the description of the Millington LFSM in J. D. A. Millington et al., 2009, the ecological succession rules are specified as a table in

the supplementary materials for the published paper. I found that the numerical codes used to identify land cover types in the paper did not match the codes for their intended land cover types in the supplementary materials. Third, the description of the way in which juvenile individuals of forest land cover types (seeds) persist over time in grid cells could not be inferred from the model description in J. D. A. Millington et al., 2009, requiring detailed discussion with the original author.

These issues highlight how difficult it is to accurately and unambiguously describe complicated models. The first and third points highlight the importance of making model source code available for readers to review. Model rules specified in a formal programming language will always be less ambiguous than the corresponding verbal description in the model documentation. Additionally, the second point illustrates the advantage of finding ways to describe and visualise complicated models in a form that is easy for humans to understand. It would have been impossible to diagnose the issue of mismatched land cover type codes between the paper and supplementary materials by scrutinising the tabular representation of the ecological succession rules alone. However, when I loaded these rules into a graph database using Cymod (see Section 7.2) the issue became immediately apparent.

Chapter 8

Conclusions

Here, I review the contributions made in this thesis in relation to the aims specified in Chapter 1. I highlight the major challenges faced during the production of this work, and indicate the most fruitful directions for future work. Finally, I reflect on how my work can be generalised and applied to other problems in landscape ecology, as well as in other scientific fields where complicated simulation models are used.

8.1 Key outcomes

In Chapter 4 I described the AgroSuccess agent-based simulation model. This spatially explicit model represents ecological succession, wildfire and anthropogenic land-use change affected by Neolithic subsistence agriculturalists. It can be used to represent specific geographical regions in the Mediterranean by providing it with morphological, ecological, and climatic data that characterise those regions. AgroSuccess can be used as a computational laboratory to study the emergent ecological dynamics that arise though the interaction of anthropogenic and natural disturbance processes over centennial timescales. This makes it a valuable tool to develop our understanding of terrestrial ecosystem change. The development of AgroSuccess addresses **Aim 1** identified in Chapter 1. Furthermore, I have published the AgroSuccess simulation model implementation's source code in an immutable software repository (Lane, 2023) to ensure it is available for other scientists to use and build on. This contributes to **Aim 2**.

AgroSuccess combines features of two models that have been reported previously in the literat-

ure: the Millington LFSM (J. D. A. Millington et al., 2009) and MedLand (Ullah, 2013). These are both complicated models in the sense that they are comprised of many interacting submodels (Sun et al., 2016). During model development I identified shortcomings in the description of both of these models that needed to be resolved before we understood the relevant submodels sufficiently well to integrate them into our own work (see Section 7.3). Additionally, I was unable to reconcile the detailed description of the MedLand model to the model code that we were able to obtain. These difficulties point to serious threats to reproducibility in scientific fields that depend on complicated models. I argued in Chapter 7 that modellers should observe some software development best practices–especially software documentation and automated testing–to ensure that others can understand their model code and confirm the code is behaving as expected. This will increase confidence in models and promote code reuse. I have invested significant time and effort to follow these recommendations in our implementation of AgroSuccess, helping to ensure it satisfies **Aim 2**.

In addition to documenting and testing the AgroSuccess simulation model code, I also developed the Cymod package to help other modellers visualise the state-and-transition (STM) model that underpins the ecological succession submodel in AgroSuccess. Cymod is described in Section 7.2 and is publicly available to download (Lane, 2020). I published the scripts we used alongside Cymod to adapt the STM used in the Millington LFSM for AgroSuccess as the `agrosuccess-graph` package (Lane, 2021c).

As noted above, AgroSuccess requires various input data to configure it to represent a particular geographical region. The work I did to obtain this data for the six study sites considered in this thesis is reported in Chapter 3. To make the steps I took to obtain and process this data explicit and reproducible I have published the software tools I developed and used for the task as software repositories on Zenodo (Lane, 2019, 2021b, 2021e). This addresses **Aim 3**.

In Chapter 5 I described how I calibrated the wildfire submodel to tailor AgroSuccess for each study site. I used the calibrated AgroSuccess model in Chapter 6 to examine its sensitivity to input parameters. I also ran AgroSuccess for counterfactual scenarios of human activity, addressing **Aim 4**. The outputs from these simulations demonstrate that AgroSuccess is ready for use and capable of generating land-cover abundance time series for simulated scenarios that include or exclude Neolithic subsistence agriculturalists.

8.2 Challenges

Calibrating complicated ABMs is inherently challenging. Complicated models typically have a large number of parameters, and ABMs typically take at least 10s of minutes to run. Consequently, exploration of parameter space is expensive. Furthermore, in the case of AgroSuccess, the only empirical data available to calibrate against is pollen abundance–a proxy for land-cover proportion. The development of additional empirical data sets to calibrate against would be extremely valuable to the development of complicated models of land-cover change in fire-prone ecosystems. This provides support for ongoing work to develop new high frequency sedimentary charcoal data sets (see Section 5.1).

While AgroSuccess draws extensively from the Millington LFSM J. D. A. Millington et al., 2009 and MedLand Ullah, 2013 models, I decided to implement it as a new model using only the written descriptions of the earlier models as an exercise in scientific reproducibility. It transpired that this was informative, as it allowed me to identify shortcomings in the description of the earlier models that we were able to rectify in AgroSuccess. Consequently, AgroSuccess doesn't share any code with either the Millington LFSM or MedLand. Designing code that is easy for others to read, documented, and well testing takes time. In a 3-year funded PhD it is ambitious to both implement a model from scratch and apply it to an empirical problem, while maintaining sufficient programming standards to provide assurance that the model behaves as expected.

8.3 Outlook for future work

8.3.1 Improved approach to landscape reconstruction

In the analysis presented in this thesis, the time evolution of the proportion of the landscape occupied by different types of land cover at the included study sites was estimated using raw pollen abundance time series from the European Pollen Database (EPD) (Fyfe et al., 2009). As noted in Section 8.3.1 this is likely to introduce biases because pollen productivity varies by species (the Fagerlind effect), so the proportion of pollen from species contributing to a land cover type in a sample does not map directly to the proportion of the landscape occupied by that land cover type at the time corresponding to the sample. These biases could be addressed in future work by applying the Landscape Reconstruction Algorithm (LRA) (Sugita, 2007) or the

more recent MARCO POLO tool that was developed with the aim of being simpler to use than the LRA (Mrotzek et al., 2017) to the pollen abundance time series for the considered study sites. The LRA and MARCO POLO attempt to correct for biases introduced by variation in pollen abundance by using pollen data from a single large lake (or alternatively, in the case of the LRA, multiple smaller lakes) in the same region as the study site to construct a picture of vegetation composition (i.e. proportion of area occupied by specific genera) in a *regional* area of $10^4\,\mathrm{km}^2$–$10^5\,\mathrm{km}^2$. By comparing pollen data from a study site of interest to the regional background, the LRA and MARCO POLO aim to discern *local* changes to vegetation composition in areas up to $100\,\mathrm{km}^2$ (Sugita, 2007). The LRA has been applied and deemed to perform well in studies in Michigan and Wisonsin (Sugita et al., 2010), Norway (Hjelle et al., 2015), and the Czech Republic (Abraham et al., 2017), and MARCO POLO is claimed to perform similarly to the LRA (Mrotzek et al., 2017).

8.3.2 Further analysis of fire frequency-size statistics

In Section 5.1.4 I attempted to calibrate simulated wildfire regimes to empirical observations using a methodology that assumed the burned area of simulated fires had a frequency-size distribution that is well modelled by a power-law. However, I found that fire sizes in the scenarios I considered did not follow a power-law distribution. It would be informative to spend more time investigating the reason that I did not observe power-law distributed wildfire burned areas in AgroSuccess. For example, the model might be improved by revisiting and revising the phenomenologically motivated parameter values that originated in the Millington LFSM (J. D. A. Millington et al., 2009). In particular, I might treat parameters that were previously held fixed during calibration as calibration parameters to see if this increases the incidence of large fires. The under-representation of large fires seen during my calibration procedure indicates there is not enough flammability in the AgroSuccess model to produce expected fire statistics. Any parameter that affects flammability indirectly could influence the ratio of large fires to small ones and, therefore, the value of the power-law exponent parameter, β. Second, future work should investigate whether the reason we didn't see power-law distributed fire sizes in the simulation outputs was due to 'edge effects' arising from the finite area of the simulation grid. This could be done by running AgroSuccess simulations with larger simulation grids.

8.3.3 Improved sensitivity analysis

During the production of the sensitivity analysis described in Section 6.1, I was limited in the number of simulations I could run per parameter value by computational resource constraints. Consequently, my results are based on a set of simulations with 10 runs per parameter value. Future work should include rerunning this analysis with a number of simulations per parameter value that is determined by a principled statistical analysis of the simulation outputs, rather than resource constraints. This analysis should take particular care to ensure that burned area is sufficiently constrained.

Sensitivity analysis should be improved in future work by using a more formal test of statistical significance compared to that presented in Section 6.1. For example, I could use a two-sample trimmed t-test (Yuen, 1974). The confidence level of 66% used to identify scenarios that had had a statistically significant effect on output variables is non-standard and relatively low. Future work to improve sensitivity analysis should use a 95% confidence level when determining statistically significant effects, which is in line with confidence levels often used in the literature (Cumming & Finch, 2005).

The approach to sensitivity analysis presented in Section 6.1 is an example of 'local' or 'one-factor-at-a-time' sensitivity analysis (Bar Massada & Carmel, 2008; Thiele et al., 2014). Under this approach, a set of simulations are run over a range of values for each parameter while holding the values of all other parameters constant. Local sensitivity analysis enables modellers to identify parameters that a model is highly sensitive to relative to other parameters. The sampling method used in local sensitivity analysis as applied to AgroSuccess in Section 6.1 was to explore the model's sensitivity to *all* parameters subject to changes of ±10% relative to their default values. This is a simple approach that enabled me to identify parameters that individually have a large net effect on the simulation outputs, while requiring relatively modest computational expense. Local sensitivity analysis acts as a way of verifying that the model is behaving as expected (e.g. increasing the number of households increases the farmed area in the landscape as expected) and highlights parameters that should be prioritised when attempting to reduce uncertainty in their values. It also allowed me to identify parameters that the model was not sensitive to, which could help guide model simplification (Bar Massada & Carmel, 2008).

A more robust approach to sensitivity analysis that should be considered in future work is to

use a ‘global’ sensitivity analysis in which groups of parameters are varied simultaneously, enabling the identification of the effects of interactions between parameters on simulation outputs (Thiele et al., 2014). For example, it would be informative to quantify the strength of interaction between the ‘Climax forest biomass density’ and ‘Number of households per village’ parameters in AgroSuccess. This is because a given utilisation of area for firewood gathering in a landscape could be explained by either a large number of households sourcing firewood from high density woodland, or a small number of households collecting firewood from low density woodland. It is therefore plausible that there would be an interaction between these two parameters. The number of parameters in AgroSuccess (17) combined with its non-trivial run time (approximately 5 minutes to simulate enough time steps to reach an equilibrium state) means that global sensitivity analysis would be computationally expensive. This could be mitigated by using a screening method, such as Morris’s elementary effects screening (Morris, 1991; Thiele et al., 2014), to systematically identify parameters that the model is sensitive to, and progress only those parameters to the detailed global sensitivity analysis.

8.3.4 Correspondence between empirical data and simulation outputs

In Section 6.2.1 I found that AgroSuccess tends to produce an initial transient period in which the land cover proportions diverge from the boundary conditions followed by a stable equilibrium state that does not match those of the target empirical data. This undermines its ability to reproduce ecological dynamics because it shows that there is a mismatch between the model dynamics and the target empirical land cover proportions. This should be addressed in future work by diagnosing the aspects of the model dynamics that are causing the model outputs to diverge from the target empirical patterns. Specifically, I should investigate whether the model’s divergence from target land cover proportions is due to one or more of the following sources of residual uncertainty.

Soil type

AgroSuccess is able to model the effect of spatially heterogeneous soil type on surface runoff and soil moisture. However, in the version of AgroSuccess presented in this thesis, the model’s sensitivity to this data input is not explored, since all models are assumed to have a uniform categorical soil type (see Section 3.3). Future work should include the development of soil type

maps that are compatible with AgroSuccess for the study sites used in this thesis. AgroSuccess accepts categorical soil type maps in which all simulation cells have type A, B, C, or D, where type A produces the least runoff and type D the most (Ferrér et al., 1995; J. D. A. Millington et al., 2009). Soil type maps of this form could be produced by analysing soil maps provided by the European Soil Data Centre (ESDAC), for example.

Temperature and precipitation

As noted in Section 6.3, AgroSuccess's sensitivity to wind speed and direction, temperature, and precipitation input data is a source of residual uncertainty and would benefit from further investigation. In the analysis described in this thesis, annual mean temperature and precipitation were estimated from outputs of the BCC-CSM1-1 GCM model only (see Section 3.3.4). Future work could include a systematic analysis of the variance of mean annual temperature and precipitation in the set of 25 GCMs that McSweeney et al., 2015 determined to provide 'Satisfactory' performance with respect to estimates of temperature and precipitation in Europe. This would provide a confidence interval around which to explore AgroSuccess's sensitivity to GCM-derived temperature and precipitation inputs.

Seed availability

In the current version of AgroSuccess, seeds corresponding to the forest land cover types are 'imported' into the simulation grid at a fixed rate that is common to all seed types for the duration of each simulation (see Section 4.2.4). Consequently, AgroSuccess is not able to represent temporal variation in the quantity of seeds entering the simulated area from the wider reason. This, in turn, may explain why we see lack of correspondence with empirical data in the simulation outputs. Future work should explore the effect of introducing more fine-grained control over the rate at which seeds of different type are 'imported' into the simulated area from the wider region. This could be done by making the 'background rate' parameter, b, time dependent and including separate rates for each seed type.

8.3.5 Generate model outputs for additional scenarios

AgroSuccess could be further analysed by running simulations of additional anthropogenic land-use change scenarios. The pollen abundance data that I have obtained for each study site and described in Chapter 3 may contain evidence of the earliest date of human agriculture at each site. A natural use-case for AgroSuccess would be to investigate whether simulations that include subsistence agriculturalists in the landscape (beginning at a certain date) produce land-cover proportion time series that are more similar to the empirical pollen abundance data than simulations that do not include subsistence agriculturalists. Each alternative date for the beginning of agriculture would constitute a possible scenario.

I noted in Section 6.3 that to produce realistic ecological dynamics, AgroSuccess would need to be provided with accurate estimates of the size of human populations in the model. These estimates could be obtained using AgroSuccess itself by covarying parameters controlling both the number of households in the simulation and various ecological factors. In this way, future users of AgroSuccess could treat human populations as a 'driver' of ecological change, and search for the size of human population that causes the land cover proportions in the model outputs to best approximate empirical data.

8.4 Generalising the work in this thesis

A principle that has emerged during the development of this thesis is that inadequate documentation and transparency around the software used to implement scientific models creates challenges for reproducibility. If we, as the scientific community, don't review and evaluate code quality, and if high quality code isn't rewarded, then there is no incentive for scientists to produce high quality code. This issue is especially important in fields that involve the use of complicated models, such as in landscape ecology.

The software that implements AgroSuccess is designed to be modular and extensible. I have provided extensive in-code documentation and a suite of unit tests that both verify the correctness of the code's implementation, and illustrate how the included software components can be used. These contributions enable other modellers to implement alternative anthropogenic land-cover change behaviours to address specific research problems. Doing so they could investigate

scenarios involving anthropogenic land-cover change other than those involving Neolithic subsistence agriculture. Additionally, the software tools I developed and used to obtain empirical data to configure AgroSuccess for study sites in the Iberian Peninsula could be used by other researchers to obtain equivalent data for other study sites.

Bibliography

Aber, J. D., & Melillo, J. M. (2001). *Terrestrial ecosystems* (2nd ed.). Harcourt Academic Press.

Abraham, V., Novák, J., Houfková, P., Petr, L., & Dudová, L. (2017). A landscape reconstruction algorithm and pedoanthracological data reveal late holocene woodland history in the lowlands of the ne czech republic. *Review of Palaeobotany and Palynology*, *244*, 54–64. https://doi.org/10.1016/j.revpalbo.2017.04.009

Aguilera, P., Fernández, A., Fernández, R., Rumí, R., & Salmerón, A. (2011). Bayesian networks in environmental modelling. *Environmental Modelling & Software*, *26*(12), 1376–1388. https://doi.org/https://doi.org/10.1016/j.envsoft.2011.06.004

An, L., Linderman, M., Qi, J., Shortridge, A., & Liu, J. (2005). Exploring complexity in a human-environment system: An agent-based spatial model for multidisciplinary and multiscale integration. *Annals of the Association of American Geographers*, *95*(1), 54–79. https://doi.org/10.1111/j.1467-8306.2005.00450.x

Augusiak, J., Van den Brink, P. J., & Grimm, V. (2014). Merging validation and evaluation of ecological models to 'evaludation': A review of terminology and a practical approach. *Ecological Modelling*, *280*, 117–128. https://doi.org/10.1016/j.ecolmodel.2013.11.009

Axtell, R. L., Epstein, J. M., Dean, J. S., Gumerman, G. J., Swedlund, A. C., Harburger, J., Chakravarty, S., Hammond, R., Parker, J., & Parker, M. (2002). Population growth and collapse in a multiagent model of the kayenta anasazi in long house valley. *Proceedings of the National Academy of Sciences*, *99*(Suppl 3), 7275–7279. https://doi.org/10.1073/pnas.092080799

Baeza, M., Valdecantos, A., Alloza, J., & Vallejo, V. (2007). Human disturbance and environmental factors as drivers of long-term post-fire regeneration patterns in Mediterranean forests. *Journal of Vegetation Science*, *18*(2), 243–252. https://doi.org/10.1111/j.1654-1103.2007.tb02535.x

Balter, M. (2010). The Tangled Roots of Agriculture. *Science*, *327*(January), 404–407. https://doi.org/10.1126/science.327.5964.404

Bar Massada, A., & Carmel, Y. (2008). Incorporating output variance in local sensitivity analysis for stochastic models. *Ecological Modelling*, *213*(3), 463–467. https://doi.org/https://doi.org/10.1016/j.ecolmodel.2008.01.021

Barber, A. H., Lu, D., & Pugno, N. M. (2015). Extreme strength observed in limpet teeth. *Journal of the Royal Society, Interface*, *12*(105), 20141326–. https://doi.org/10.1098/rsif.2014.1326

Barbero, M., Bonin, G., Loisel, R., & Quézel, P. (1990). Changes and disturbances of forest ecosystems caused by human activities in the western part of the mediterranean basin. *Vegetatio*, *87*(2), 151–173. https://doi.org/10.1007/BF00042952

Bartlett, L. J., Williams, D. R., Prescott, G. W., Balmford, A., Green, R. E., Eriksson, A., Valdes, P. J., Singarayer, J. S., & Manica, A. (2016). Robustness despite uncertainty: Regional climate data reveal the dominant role of humans in explaining global extinctions of Late Quaternary megafauna. *Ecography*, *39*(2), 152–161. https://doi.org/10.1111/ecog.01566

Barton, C. M., Ullah, I., Mayer, G., Bergin, S., Sarjoughian, H., & Mitasova, H. (2017). MedLanD Modeling Laboratory v.1. URL: https://www.comses.net/codebases/4609/releases/1.1.0/

Barton, C. M., Ullah, I. I., & Bergin, S. (2010). Land use, water and Mediterranean landscapes: modelling long-term dynamics of complex socio-ecological systems. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, *368*(1931), 5275–5297. https://doi.org/10.1098/rsta.2010.0193

Barton, C. M., Ullah, I. I., Bergin, S. M., Sarjoughian, H. S., Mayer, G. R., Bernabeu-Auban, J. E., Heimsath, A. M., Acevedo, M. F., Riel-Salvatore, J. G., & Arrowsmith, J. R. (2016). Experimental socioecology: Integrative science for anthropocene landscape dynamics. *Anthropocene*, *13*(March), 34–45. https://doi.org/10.1016/j.ancene.2015.12.004

Batllori, E., De Cáceres, M., Brotons, L., Ackerly, D. D., Moritz, M. A., & Lloret, F. (2017). Cumulative effects of fire and drought in Mediterranean ecosystems. *Ecosphere*, *8*(8), e01906. https://doi.org/10.1002/ecs2.1906

Beller, E., McClenachan, L., Trant, A., Sanderson, E. W., Rhemtulla, J., Guerrini, A., Grossinger, R., & Higgs, E. (2017). Toward principles of historical ecology. *American Journal of Botany*, *104*(5), 1–4. https://doi.org/10.3732/ajb.1700070

Bellwood, P. (2004). *First Farmers: The Origins of Agricultural Societies*. Wiley-Blackwell.

Bestelmeyer, B. T., Ash, A., Brown, J. R., Densambuu, B., Fernández-Giménez, M., Johanson, J., Levi, M., Lopez, D., Peinetti, R., Rumpff, L., & Shaver, P. (2017). State and Transition Models: Theory, Applications, and Challenges. In D. D. Briske (Ed.), *Rangeland systems:*

Processes, management and challenges (pp. 303–345). Springer, Cham. https://doi.org/10.1007/978-3-319-46709-2_9

Bliege Bird, R., Bird, D. W., Codding, B. F., Parker, C. H., & Jones, J. H. (2008). The "fire stick farming" hypothesis: Australian Aboriginal foraging strategies, biodiversity, and anthropogenic fire mosaics. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(39), 14796–14801. https://doi.org/10.1073/pnas.0804757105

Blondel, J., & Aronson, J. (1999). *Biology and Wildlife of the Mediterranean Region*. Oxford University Press.

Bonny, A. P. (1972). A method for determining absolute pollen frequencies in lake sediments. *New Phytologist*, *71*(2), 393–405. https://doi.org/10.1111/j.1469-8137.1972.tb04086.x

Bugmann, H. (1994). *On the Ecology of Mountainous Forests in a Changing Climate: A Simulation Study* (Doctoral dissertation). SWISS FEDERAL INSTITUTE OF TECHNOLOGY ZüRICH. https://doi.org/10.3929/ethz-a-000946508

Cardinale, B. J., Duffy, J. E., Gonzalez, A., Hooper, D. U., Perrings, C., Venail, P., Narwani, A., Mace, G. M., Tilman, D., A.Wardle, D., Kinzig, A. P., Daily, G. C., Loreau, M., Grace, J. B., Larigauderie, A., Srivastava, D. S., & Naeem, S. (2012). Biodiversity loss and its impact on humanity. *Nature*, *489*(7415), 326–326. https://doi.org/10.1038/nature11373

Carrión, J. S., & Dupré, M. (1996). Late Quaternary vegetational history at Navarres, Eastern Spain. A two core approach. *New Phytology*, *134*, 177–191. https://doi.org/10.1111/j.1469-8137.1996.tb01157.x

Carrión, J. S., & Van Geel, B. (1999). Fine-resolution Upper Weichselian and Holocene palynological record from Navarres (Valencia, Spain) and a discussion about factors of Mediterranean forest succession. *Review of Palaeobotany and Palynology*, *106*(3-4), 209–236. https://doi.org/10.1016/S0034-6667(99)00009-3

Carrión, J. S., Fernández, S., González-Sampériz, P., Gil-Romera, G., Badal, E., Carrión-Marco, Y., López-Merino, L., López-Sáez, J. A., Fierro, E., & Burjachs, F. (2010). Expected trends and surprises in the Lateglacial and Holocene vegetation history of the Iberian Peninsula and Balearic Islands. *Review of Palaeobotany and Palynology*, *162*(3), 458–475. https://doi.org/http://dx.doi.org/10.1016/j.revpalbo.2009.12.007

Chivian, E., & Bernstein, A. (Eds.). (2008). *Sustaining Life: How Human Health Depends on Biodiversity*. Oxford University Press.

Clark, J. K., & Crabtree, S. A. (2015). Examining social adaptations in a volatile landscape in Northern Mongolia via the agent-based model Ger Grouper. *Land*, *4*, 157–181. https://doi.org/10.3390/land4010157

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-Law Distributions in Empirical Data. *SIAM Review*, *51*(4), 661–703. https://doi.org/10.1137/070710111

Colombaroli, D., Henne, P. D., Kaltenrieder, P., Gobet, E., & Tinner, W. (2010). Species responses to fire, climate and human impact at tree line in the Alps as evidenced by palaeo-environmental records and a dynamic simulation model. *Journal of Ecology*, *98*(6), 1346–1357. https://doi.org/10.1111/j.1365-2745.2010.01723.x

CoMSES Net. (2020). CoMSES Catalog. Retrieved June 1, 2020, from URL: https://catalog.comses.net/

Conedera, M., Colombaroli, D., Tinner, W., Krebs, P., & Whitlock, C. (2017). Insights about past forest dynamics as a tool for present and future forest management in Switzerland. *Forest Ecology and Management*, *388*, 100–112. https://doi.org/10.1016/j.foreco.2016.10.027

Cronon, W. (1996). The Trouble with Wilderness: Or, Getting Back to the Wrong Nature. *Environmental History*, *1*(1), 7–28. https://doi.org/10.2307/3985059

Crumley, C. L. (Ed.). (1994). *Historical Ecology: Cultural Knowledge and Changing Landscapes*. School of American Research Press.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, *60*(2), 170–180. https://doi.org/10.1037/0003-066X.60.2.170

Daniel, C. J., Frid, L., Sleeter, B. M., & Fortin, M.-J. (2016). State-and-transition simulation models: a framework for forecasting landscape change. *Methods in Ecology and Evolution*, *7*, 1413–1423. https://doi.org/10.1111/2041-210X.12597

Daniel, C. J., Sleeter, B. M., Frid, L., & Fortin, M. J. (2018). Integrating continuous stocks and flows into state-and-transition simulation models of landscape change. *Methods in Ecology and Evolution*, *9*(4), 1133–1143. https://doi.org/10.1111/2041-210X.12952

Dearing, J. A., Wang, R., Zhang, K., Dyke, J. G., Haberl, H., Hossain, M. S., Langdon, P. G., Lenton, T. M., Raworth, K., Brown, S., Carstensen, J., Cole, M. J., Cornell, S. E., Dawson, T. P., Doncaster, C. P., Eigenbrod, F., Flörke, M., Jeffers, E., Mackay, A. W., ... Poppy, G. M. (2014). Safe and just operating spaces for regional social-ecological systems. *Global Environmental Change*, *28*(1), 227–238. https://doi.org/10.1016/j.gloenvcha.2014.06.012

Debain, S., Curt, T., Lepart, J., & Prevosto, B. (2003). Reproductive variability in Pinus sylvestris in southern France: Implications for invasion. *Journal of Vegetation Science*, *14*, 509–516. https://doi.org/10.1111/j.1654-1103.2003.tb02177.x

Debeljak, M., & Džeroski, S. (2011). Decision trees in ecological modelling. In F. Jopp, H. Reuter & B. Breckling (Eds.), *Modelling complex ecological dynamics: An introduction into ecological modelling for students, teachers & scientists* (pp. 197–209). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-05029-9_14

Deza-Araujo, M., Morales-Molino, C., Conedera, M., Henne, P. D., Krebs, P., Hinz, M., Heitz, C., Hafner, A., & Tinner, W. (2022). A new indicator approach to reconstruct agricultural land use in europe from sedimentary pollen assemblages. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *599*, 111051. https://doi.org/https://doi.org/10.1016/j.palaeo.2022.111051

Deza-Araujo, M., Morales-Molino, C., Tinner, W., Henne, P. D., Heitz, C., Pezzatti, G. B., Hafner, A., & Conedera, M. (2020). A critical assessment of human-impact indices based on anthropogenic pollen indicators. *Quaternary Science Reviews*, *236*, 106291. https://doi.org/https://doi.org/10.1016/j.quascirev.2020.106291

Drescher, M., & Perera, A. H. (2010). A network approach for evaluating and communicating forest change models. *Journal of Applied Ecology*, *47*, 57–66. https://doi.org/10.1111/j.1365-2664.2009.01754.x

Edmonds, B., Le Page, C., Bithell, M., Chattoe-Brown, E., Grimm, V., Meyer, R., Montañola-Sales, C., Ormerod, P., Root, H., & Squazzoni, F. (2019). Different modelling purposes. *Journal of Artificial Societies and Social Simulation*, *22*(3), 6. https://doi.org/10.18564/jasss.3993

Elhacham, E., Ben-Uri, L., Grozovski, J., Bar-On, Y. M., & Milo, R. (2020). Global human-made mass exceeds all living biomass. *Nature*, *588*(7838), 442–444. https://doi.org/10.1038/s41586-020-3010-5

Epstein, J. M. (2008). Why model? *Journal of Artificial Societies and Social Simulation*, *11*(4), 12.

Erb, K.-H., Kastner, T., Plutzar, C., Bais, A. L. S., Carvalhais, N., Fetzel, T., Gingrich, S., Haberl, H., Lauk, C., Niedertscheider, M., Pongratz, J., Thurner, M., & Luyssaert, S. (2018). Unexpectedly large impact of forest management and grazing on global vegetation biomass. *Nature*, *553*(7686), 73–76. https://doi.org/10.1038/nature25138

Etherington, T. R., Holland, E. P., & O'Sullivan, D. (2015). NLMpy: A python software package for the creation of neutral landscape models within a general numerical framework. *Methods in Ecology and Evolution*, *6*(2), 164–168. https://doi.org/10.1111/2041-210X.12308

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., & Alsdorf, D. (2007). The Shuttle Radar Topography Mission. *Reviews of Geophysics*, *45*(2). https://doi.org/10.1029/2005RG000183

Fernandes, P. M., Davies, G. M., Ascoli, D., Fernández, C., Moreira, F., Rigolot, E., Stoof, C. R., Vega, J. A., & Molina, D. (2013). Prescribed burning in southern Europe: Developing fire management in a dynamic landscape. *Frontiers in Ecology and the Environment*, *11*(SUPPL. 1). https://doi.org/10.1890/120298

Ferrér, M., Rodríguez, J., & Estrela, T. (1995). Generación automática del número de curva con sistemas de información geográfica. *Ingeniería del agua*, *2*(4), 43–58. https://doi.org/10.4995/ia.1995.2686

Finlayson, C., Giles Pacheco, F., Rodríguez-Vidal, J., Fa, D. A., María Gutierrez López, J., Santiago Pérez, A., Finlayson, G., Allue, E., Baena Preysler, J., Cáceres, I., Carrión, J. S., Fernández Jalvo, Y., Gleed-Owen, C. P., Jimenez Espejo, F. J., López, P., Antonio López Sáez, J., Antonio Riquelme Cantal, J., Sánchez Marco, A., Giles Guzman, F., . . . Sakamoto, T. (2006). Late survival of Neanderthals at the southernmost extreme of Europe. *Nature*, *443*(7113), 850–853. https://doi.org/10.1038/nature05195

Flint, R. F., & Deevey, E. S. (1961). Editorial Statement. *Radiocarbon*, *4*(1).

Franks, J. W. (1957). Pollen Analysis: a technique for investigating early agrarian history. *Agricultural History Review*, *5*(1), 2–11.

Fyfe, R., de Beaulieu, J., Binney, H., Bradshaw, R., Brewer, S., Le Flao, A., Finsinger, W., Gaillard, M., Giesecke, T., Gil-Romera, G., Grimm, E., Huntley, B., Kunes, P., Kuhl, N., Leydet, M., Lotter, A., Tarasov, P., & Tonkov, S. (2009). The European Pollen Database: Past efforts and current activities. *Vegetation History and Archaeobotany*, *18*(5), 417–424. https://doi.org/10.1007/s00334-009-0215-9

Giesecke, T., Davis, B., Brewer, S., Finsinger, W., Wolters, S., Blaauw, M., de Beaulieu, J.-L., Binney, H., Fyfe, R. M., Gaillard, M.-J., Gil-Romera, G., van der Knaap, W. O., Kuneš, P., Kühl, N., van Leeuwen, J. F. N., Leydet, M., Lotter, A. F., Ortu, E., Semmler, M., & Bradshaw, R. H. W. (2014). Towards mapping the late quaternary vegetation change of europe. *Vegetation History and Archaeobotany*, *23*(1), 75–86. https://doi.org/10.1007/s00334-012-0390-y

Gordó, S. P., Bernabeu Aubán, J., García Puchol, O., Barton, M., & Bergin, S. M. (2015). The origins of agriculture in Iberia: a computational model. *Documenta Praehistorica*, *42*, 117. https://doi.org/10.4312/dp.42.7

Goring, S., Dawson, A., Simpson, G. L., Ram, K., Graham, R. W., Grimm, E. C., & Williams, J. W. (2015). neotoma: A Programmatic Interface to the Neotoma Paleoecological Database. *OpenQuaternary*, *1*(2), 1–17. https://doi.org/10.5334/oq.ab

Grabher, H. (2021). HANPP trajectories for Ethiopia reveal recent agricultural efficiency gains but high grazing intensity. *Environment, Development and Sustainability*, *23*(4), 5277–5296. https://doi.org/10.1007/s10668-020-00814-x

Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., Goss-Custard, J., Grand, T., Heinz, S. K., Huse, G., Huth, A., Jepsen, J. U., Jørgensen, C., Mooij, W. M., Müller, B., Pe'er, G., Piou, C., Railsback, S. F., Robbins, A. M., . . . DeAngelis, D. L. (2006). A standard protocol for describing individual-based and agent-based models. *Ecological Modelling*, *198*(1-2), 115–126. https://doi.org/10.1016/j.ecolmodel.2006.04.023

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W. M., Railsback, S. F., Thulke, H.-H., Weiner, J., Wiegand, T., & DeAngelis, D. L. (2005). Pattern-Oriented Modeling of Agent-Based Complex Systems: Lessons from Ecology. *Science*, *310*(5750), 987–991. https://doi.org/10.1126/science.1116681

Guiot, J., & Cramer, W. (2016). Climate change: The 2015 Paris Agreement thresholds and Mediterranean basin ecosystems. *Science*, *354*(6311), 465–468. https://doi.org/10.1126/science.aah5015

Haddad, N. M. (2012). Connecting ecology and conservation through experiment. *Nature Methods*, *9*(8), 794–795. https://doi.org/10.1038/nmeth.2107

Harrison, S. P., Bartlein, P. J., Izumi, K., Li, G., Annan, J., Hargreaves, J., Braconnot, P., & Kageyama, M. (2015). Evaluation of cmip5 palaeo-simulations to improve climate projections. *Nature Climate Change*, *5*(8), 735–743. https://doi.org/10.1038/nclimate2649

Henne, P. D., Elkin, C., Colombaroli, D., Samartin, S., Bugmann, H., Heiri, O., & Tinner, W. (2013). Impacts of changing climate and land use on vegetation dynamics in a Mediterranean ecosystem: Insights from paleoecology and dynamic modeling. *Landscape Ecology*, *28*(5), 819–833. https://doi.org/10.1007/s10980-012-9782-8

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, *25*(15), 1965–1978. https://doi.org/10.1002/joc.1276

Hjelle, K. L., Mehl, I. K., Sugita, S., & Andersen, G. L. (2015). From pollen percentage to vegetation cover: Evaluation of the Landscape Reconstruction Algorithm in western Norway. *Journal of Quaternary Science*, *30*(4), 312–324. https://doi.org/10.1002/jqs.2769

Hoffecker, J. F. (2009). The spread of modern humans in Europe. *Proceedings of the National Academy of Sciences*, *106*(38), 16040–16045. https://doi.org/10.1073/pnas.0903446106

Huwaldt, J. A. (2001). Plot Digitizer. Retrieved November 1, 2016, from URL: http://plotdigitizer.sourceforge.net/

James, S. R. (1989). Hominid Use of Fire in the Lower and Middle Pleistocene: A Review of the Evidence. *Current Anthropology*, *30*(1), 1–26. https://doi.org/10.1086/203705

Janssen, M. A. (2009). Understanding artificial anasazi. *Journal of Artificial Societies and Social Simulation*, *12*(4), 13.

Joffre, R., & Rambal, S. (2002). Mediterranean Ecosystems. *eLS*, 1–7. https://doi.org/10.1038/npg.els.0003196

Johnson, N. F. (2009). *Simply Complexity*. Oneworld Publications.

Kahlke, R.-D. (2015). The maximum geographic extension of Late Pleistocene Mammuthus primigenius (Proboscidea , Mammalia) and its limiting factors. *Quaternary International*, *379*, 147–154. https://doi.org/10.1016/j.quaint.2015.03.023

Kaplan, J. O., Krumhardt, K. M., & Zimmermann, N. (2009). The prehistoric and preindustrial deforestation of Europe. *Quaternary Science Reviews*, *28*(27-28), 3016–3034. https://doi.org/10.1016/j.quascirev.2009.09.028

Kazil, J., Masad, D., & Crooks, A. (2020). Utilizing python for agent-based modeling: The mesa framework. In R. Thomson, H. Bisgin, C. Dancy, A. Hyder & M. Hussain (Eds.), *Social, cultural, and behavioral modeling* (pp. 308–317). Springer International Publishing.

Keeley, J. (2002). Native American impacs on fire regimes of the California coastal ranges. *Journal of Biogeography*, *29*(3), 303–320. https://doi.org/10.1046/j.1365-2699.2002.00676.x

Khabarov, N., Krasovskii, A., Obersteiner, M., Swart, R., Dosio, A., San-Miguel-Ayanz, J., Durrant, T., Camia, A., & Migliavacca, M. (2016). Forest fires and adaptation options in europe. *Regional Environmental Change*, *16*(1), 21–30. https://doi.org/10.1007/s10113-014-0621-0

Khresat, S., Al-bakri, J., & Al-Tahhan, R. (2008). Impacts of land use/cover change on soil properties in the Mediterranean region of northwestern Jordan. *Land Degradation & Development*, *19*, 397–407. https://doi.org/10.1002/ldr.847

Kirkland, T., Hunter, L. M., & Twine, W. (2007). “The Bush is No More”: Insights on Institutional Change and Natural Resource Availability in Rural South Africa. *Society & Natural Resources*, *20*(4), 337–350. https://doi.org/10.1080/08941920601161353

Klein, R. G. (1995). Anatomy, behavior, and modern human origins. *Journal of World Prehistory*, *9*(2), 167–198. https://doi.org/10.1007/BF02221838

Kotsiantis, S. B. (2013). Decision trees: A recent overview. *Artificial Intelligence Review*, *39*(4), 261–283. https://doi.org/10.1007/s10462-011-9272-4

Lane, A. J. (2019). *Lanecodes/epd-query: Initial release* (Version v0.0.1). Zenodo. https://doi.org/10.5281/zenodo.3560683

Lane, A. J. (2020). *Lanecodes/cymod: First public release* (Version v0.0.5). Zenodo. https://doi.org/10.5281/zenodo.3630631

Lane, A. J. (2021a). *Lanecodes/aemet-wind* (Version v0.1). Zenodo. https://doi.org/10.5281/zenodo.4641908

Lane, A. J. (2021b). *Lanecodes/agrosuccess-data: Initial release* (Version v0.1). Zenodo. https://doi.org/10.5281/zenodo.4641446

Lane, A. J. (2021c). *Lanecodes/agrosuccess-graph: Initial release* (Version v0,). Zenodo. https://doi.org/10.5281/zenodo.4641458

Lane, A. J. (2021d). *Lanecodes/aslib: Initial release* (Version v0.1). Zenodo. https://doi.org/10.5281/zenodo.4641910

Lane, A. J. (2021e). *Lanecodes/demproc: Initial release* (Version v0.0.2). Zenodo. https://doi.org/10.5281/zenodo.4641912

Lane, A. J. (2023). *Lanecodes/agrosuccess-sim: Version 0.2* (Version v0.2). Zenodo. https://doi.org/10.5281/zenodo.8079621

Lin, G., Palopoli, M., & Dadwal, V. (2020). From causal loop diagrams to system dynamics models in a data-rich ecosystem. In L. A. Celi, M. S. Majumder, P. Ordóñez, J. S. Osorio, K. E. Paik & M. Somai (Eds.), *Leveraging data science for global health* (pp. 77–98). Springer International Publishing. https://doi.org/10.1007/978-3-030-47994-7_6

Lister, A. M. (2009). Late-glacial mammoth skeletons (Mammuthus primigenius) from Condover (Shropshire, UK): anatomy, pathology, taphonomy and chronological significance. *Geological Journal*, *44*, 447–479.

López-Merino, L., Cortizas, A. M., & López-Sáez, J. A. (2010). Early agriculture and palaeoenvironmental history in the North of the Iberian Peninsula: a multi-proxy analysis of the

Monte Areo mire (Asturias, Spain). *Journal of Archaeological Science*, *37*(8), 1978–1988. https://doi.org/10.1016/j.jas.2010.03.003

Luke, S., Cioffi-Revilla, C., Panait, L., Sullivan, K., & Balan, G. (2005). Mason: A multiagent simulation environment. *SIMULATION*, *81*(7), 517–527. https://doi.org/10.1177/0037549705058073

Magri, D., & Sadori, L. (1999). Late Pleistocene and Holocene pollen stratigraphy at Lago di Vico, central Italy. *Vegetation History and Archaeobotany*, *8*(4), 247–260. https://doi.org/10.1007/BF01291777

Malamud, B., Millington, J., & Perry, G. (2005). Characterizing wildfire regimes in the United States. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(13), 4694–4699. https://doi.org/10.1073/pnas.0500880102

Manson, S. M., Sun, S., & Bonsal, D. (2012). Agent-based modeling and complexity. In A. J. Heppenstall, A. T. Crooks, L. M. See & M. Batty (Eds.), *Agent-based models of geographical systems* (pp. 125–139). Springer Netherlands. https://doi.org/10.1007/978-90-481-8927-4_7

Martin, R. C. (2017). *Clean architecture: A craftsman's guide to software structure and design* (1st). Prentice Hall Press.

Martins, H., Oms Arias, F. X., Pereira, L., Pike, A. W. G., Rowsell, K., & Zilhão, J. (2015). Radiocarbon dating the beginning of the Neolithcis in Iberia: new results, new problems. *Journal of Mediterranean Archaeology*, *28*(1), 105–131. https://doi.org/10.1558/jmea.v28i1.27503

Mayer, G. R. (2009). *Composing hybrid discrete event system and cellular automata models* (Doctoral dissertation August). Arizona State University, Tempe, AZ.

Mayer, G. R., & Sarjoughian, H. S. (2007). Complexities of Simulating a Hybrid Agent-landscape Model Using Multi-formalism Composability. *Proceedings of the 2007 Spring Simulation Multiconference - Volume 2, Norfolk, V*, 161–168.

Mayer, G. R., & Sarjoughian, H. S. (2009). Composable cellular automata. *Simulation*, *85*(11-12), 735–749. https://doi.org/10.1177/0037549709106341

Mayer, G. R., Sarjoughian, H. S., Allen, E. K., Falconer, S. E., & Barton, C. M. (2006). Simulation modeling for human community and agricultural landuse. *Proceedings of the Agent-Directed Simulation Multi-Conference, Huntsville, Alabama*, (January).

McCall, M. K. (1985). The significance of distance constraints in peasant farming systems with special reference to sub-Saharan Africa. *Applied Geography*, *5*, 325–345.

McConnell, W. J., Millington, J. D. A., Reo, N. J., Alberti, M., Asbjornsen, H., Baker, L. A., Brozović, N., Drinkwater, L. E., Drzyzga, S. A., Fragoso, J., Holland, D. S., Jantz, C. A., Kohler, T. A., Maschner, H. D. G., Monticino, M., Podestá, G., Pontius Jr., R. G., Redman, C. L., Sailor, D., . . . Liu, J. (2011). Research on Coupled Human and Natural Systems (CHANS): Approach, Challenges, and Strategies. *The Bulletin of the Ecological Society of America*, *92*(2), 218–228. https://doi.org/10.1890/0012-9623-92.2.218

McGuffie, K., & Henderson-Sellers, A. (2005). *A Climate Modelling Primer* (3rd ed.). Wiley.

McIntosh, B. S. (2003). Qualitative modelling with imprecise ecological knowledge: A framework for simulation. *Environmental Modelling and Software*, *18*(4), 295–307. https://doi.org/10.1016/S1364-8152(03)00002-1

McIntosh, B. S., Muetzelfeldt, R. I., Legg, C. J., Mazzoleni, S., & Csontos, P. (2003). Reasoning with direction and rate of change in vegetation state transition modelling. *Environmental Modelling and Software*, *18*(10), 915–927. https://doi.org/10.1016/S1364-8152(03)00055-0

McSweeney, C. F., Jones, R. G., Lee, R. W., & Rowell, D. P. (2015). Selecting cmip5 gcms for downscaling over multiple regions. *Climate Dynamics*, *44*(11), 3237–3260. https://doi.org/10.1007/s00382-014-2418-8

Merkel, D. (2014). Docker: lightweight linux containers for consistent development and deployment. *Linux journal*, *2014*(239), 2.

Millington, J., Romero-Calcerrada, R., Wainwright, J., & Perry, G. (2008). An agent-based model of mediterranean agricultural land-use/cover change for examining wildfire risk. *JASSS*, *11*(4), 4.

Millington, J. D. A. (2007). *Modelling land-use/cover change and wildfire regimes in a Mediterranean landscape* (Doctoral dissertation). University of London.

Millington, J. D. A., Wainwright, J., Perry, G. L. W., Romero-Calcerrada, R., & Malamud, B. D. (2009). Modelling Mediterranean landscape succession-disturbance dynamics: A landscape fire-succession model. *Environmental Modelling and Software*, *24*(10), 1196–1208. https://doi.org/10.1016/j.envsoft.2009.03.013

Mittermeier, R. A., Gil, P. R., Hoffmann, M., Pilgrim, J., Brooks, T., Mittermeier, C. G., John, L., & da Fonseca, G. A. (2004). *Hotspots Revisited: Earth's Biologically Richest and Most Endangered Ecoregions*. Conservation International. https://doi.org/10.2744/ccab-14-01-2-10.1

Moran, E. F. (2006). *People and Nature: An Introduction to Human Ecological Relations*. Wiley.

Moreno, J. M., & Oechel, W. C. (Eds.). (1994). *The Role of Fire in Mediterranean-Type Ecosystems*. Springer-Verlag.

Moreno, M. V., Malamud, B. D., & Chuvieco, E. (2011). Wildfire frequency-area statistics in Spain. *Procedia Environmental Sciences*, *7*, 182–187. https://doi.org/10.1016/j.proenv.2011.07.032

Morris, M. D. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics*, *33*(2), 161–174. https://doi.org/10.1080/00401706.1991.10484804

Mrotzek, A., Couwenberg, J., Theuerkauf, M., & Joosten, H. (2017). MARCO POLO – A new and simple tool for pollen-based stand-scale vegetation reconstruction. *Holocene*, *27*(3), 321–330. https://doi.org/10.1177/0959683616660171

Müller, B., Bohn, F., Dreßler, G., Groeneveld, J., Klassert, C., Martin, R., Schlüter, M., Schulze, J., Weise, H., & Schwarz, N. (2013). Describing human decisions in agent-based models - ODD+D, an extension of the ODD protocol. *Environmental Modelling and Software*, *48*, 37–48. https://doi.org/10.1016/j.envsoft.2013.06.003

Naveh, Z. (1994). The Role of Fire and Its Management in the Conservation of Mediterranean Ecosystems and Landscapes. In J. M. Moreno & W. C. Oechel (Eds.), *The role of fire in mediterranean-type ecosystems* (pp. 163–182). Springer-Verlag.

Naveh, Z., & Lieberman, A. S. (1994). *Landscape Ecology: Theory and Application*. Springer-Verlag. https://doi.org/10.1007/978-1-4757-2331-1

North, M. J., Collier, N. T., Ozik, J., Tatara, E. R., Macal, C. M., Bragen, M., & Sydelko, P. (2013). Complex adaptive systems modeling with repast simphony. *Complex Adaptive Systems Modeling*, *1*(1), 1–26. https://doi.org/10.1186/2194-3206-1-3

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. *Science*, *263*(February), 641–646. https://doi.org/10.1126/science.263.5147.641

O'Sullivan, D., Millington, J., Perry, G., & Wainwright, J. (2012). Agent-based models – because they're worth it? In *Agent-based models of geographical systems* (pp. 1–15). https://doi.org/10.1007/978-90-481-8927-4

O'Sullivan, D., & Unwin, D. J. (2010). *Geographic Information Analysis* (2nd ed.). Wiley.

Pantaleon-Cano, J., Yll, E. I., Perez-Obiol, R., & Roure, J. M. (2003). Palynological evidence for vegetational history in semi-arid areas of the western Mediterranean (Almeria, Spain). *Holocene*, *13*(1), 109–119. https://doi.org/10.1191/0959683603hl598rp

Parker, D. C., Manson, S. M., Janssen, M. A., Hoffmann, M. J., & Deadman, P. (2003). Multi-agent systems for the simulation of land-use and land-cover change: A review. *Annals of the Association of American Geographers*, *93*(2), 314–337. https://doi.org/10.1111/1467-8306.9302004

Pausas, J. G., & Fernández-Muñoz, S. (2012). Fire regime changes in the western mediterranean basin: From fuel-limited to drought-driven fire regime. *Climatic Change*, *110*, 215–216. https://doi.org/10.1007/s10584-011-0060-6

Pausas, J. G., & Keeley, J. E. (2009). A Burning Story: The Role of Fire in the History of Life. *BioScience*, *59*(7), 593–601. https://doi.org/10.1525/bio.2009.59.7.10

Penalba, M. C. (1994). The History of the Holocene Vegetation in Northern Spain from Pollen Analysis. *Journal of Ecology*, *82*, 815–832.

Pereira, M. G., Malamud, B. D., Trigo, R. M., & Alves, P. I. (2011). The history and characteristics of the 1980-2005 Portuguese rural fire database. *Natural Hazards and Earth System Science*, *11*(12), 3343–3358. https://doi.org/10.5194/nhess-11-3343-2011

Perkel, J. M. (2020). Why scientists are turning to rust. *Nature*, (588), 185–186. https://doi.org/10.1038/d41586-020-03382-2

Perry, G. L. W., Wainwright, J., Etherington, T. R., & Wilmshurst, J. M. (2016). Experimental simulation: using generative modeling and palaeoecological data to understand human-environment interactions. *Frontiers in Ecology and Evolution*, *4*(October), 1–14. https://doi.org/10.3389/fevo.2016.00109

Perry, G. L. W., Wilmshurst, J. M., McGlone, M. S., McWethy, D. B., & Whitlock, C. (2012). Explaining fire-driven landscape transformation during the Initial Burning Period of New Zealand's prehistory. *Global Change Biology*. https://doi.org/10.1111/j.1365-2486.2011.02631.x

Perry, G. L. W., Wilmshurst, J. M., McGlone, M. S., & Napier, A. (2012). Reconstructing spatial vulnerability to forest loss by fire in pre-historic New Zealand. *Global Ecology and Biogeography*, *21*(10), 1029–1041. https://doi.org/10.1111/j.1466-8238.2011.00745.x

Perry, G. L., & Enright, N. J. (2002). Spatial modelling of landscape composition and pattern in a maquis-forest complex, Mont Do, New Caledonia. *Ecological Modelling*, *152*(2-3), 279–302. https://doi.org/10.1016/S0304-3800(02)00004-2

Peterson, G. (2002). Contagious disturbance, ecological memory, and the emergence of landscape pattern. *Ecosystems*, *5*(4), 329–338. https://doi.org/10.1007/s10021-001-0077-1

Phillips, J. D. (2011). Predicting modes of spatial change from state-and-transition models. *Ecological Modelling*, *222*(3), 475–484. https://doi.org/10.1016/j.ecolmodel.2010.11.018

Phillips, J. D., & Van Dyke, C. (2017). State-and-transition models in geomorphology. *Catena*, *153*, 168–181. https://doi.org/10.1016/j.catena.2017.02.009

Pickett, S., & White, P. (1985). *The Ecology of Natural Disturbance and Patch Dynamics*. Academic Press.

Poile, C., & Safayeni, F. (2016). Using computational modeling for building theory: A double edged sword. *Journal of Artificial Societies and Social Simulation*, *19*(3), 8.

Polhill, J. G., & Edmonds, B. (2007). Open access for social simulation. *Journal of Artificial Societies and Social Simulation*, *10*(3), 10.

Prentice, I. C., & Webb, T. I. (1986). Pollen percentages, tree abundances and the Fagerlind effect. *Journal of Quaternary Science*, *1*(1), 35–43. https://doi.org/10.1002/jqs.3390010105

Redman, C. L. (1999). *Human Impact on Ancient Environments*. University of Arizona Press.

Reitalu, T., Kunes, P., & Giesecke, T. (2014). Closing the gap between plant ecology and Quaternary palaeoecology (R. Kalamees, Ed.). *Journal of Vegetation Science*, *25*(5), 1188–1194. https://doi.org/10.1111/jvs.12187

Retzlaff, C. O., Ziefle, M., & Calero Valdez, A. (2021). The history of agent-based modeling in the social sciences. In V. G. Duffy (Ed.), *Digital human modeling and applications in health, safety, ergonomics and risk management. human body, motion and behavior* (pp. 304–319). Springer International Publishing.

Richardson, G. P. (2011). Reflections on the foundations of system dynamics. *System Dynamics Review*, *27*(3), 219–243. https://doi.org/https://doi.org/10.1002/sdr.462

Ricklefs, R. E. (2000). *Ecology* (R. E. Ricklefs & G. L. Miller, Eds.; 4th). W.H. Freeman.

Robinson, D. T., Di Vittorio, A., Alexander, P., Arneth, A., Michael Barton, C., Brown, D. G., Kettner, A., Lemmen, C., O'Neill, B. C., Janssen, M., Pugh, T. A., Rabin, S. S., Rounsevell, M., Syvitski, J. P., Ullah, I., & Verburg, P. H. (2018). Modelling feedbacks between human and natural processes in the land system. *Earth System Dynamics*, *9*(2), 895–914. https://doi.org/10.5194/esd-9-895-2018

Robinson, I., Webber, J., & Eifrem, E. (2015). *Graph Databases* (2nd ed.). O'Reilly.

Rowley-Conwy, P. (1995). Making first farmers younger: the west European evidence. *Current Anthropology*, *36*(2), 346–353.

Rusch, G. M., Pausas, J. G., & Lepš, J. (2003). Plant Functional Types in relation to disturbance and land use. *Journal of Vegetation Science*, *14*, 307–310. https://doi.org/10.1658/1100-9233(2003)014[0307:PFTIRT]2.0.CO;2

Salliou, N., Barnaud, C., Vialatte, A., & Monteil, C. (2017). A participatory bayesian belief network approach to explore ambiguity among stakeholders about socio-ecological systems. *Environmental Modelling & Software*, *96*, 199–209. https://doi.org/https://doi.org/10.1016/j.envsoft.2017.06.050

Saura, S., & Martínez-Millan, J. (2000). Landscape patterns simulation with a modified random clusters method. *Landscape Ecology*, *15*(7), 661–678. https://doi.org/10.1023/A:1008107902848

Scheller, R. M., Sturtevant, B. R., Gustafson, E. J., Ward, B. C., & Mladenoff, D. J. (2010). Increasing the reliability of ecological models using modern software engineering techniques. *Frontiers in Ecology and the Environment*, *8*(5), 253–260. https://doi.org/10.1890/080141

Schildt, H. (2007). *Java: The Complete Reference, Seventh Edition*. McGraw-Hill.

Schmolke, A., Thorbek, P., DeAngelis, D. L., & Grimm, V. (2010). Ecological models supporting environmental decision making: A strategy for the future. *Trends in Ecology and Evolution*, *25*(8), 479–486. https://doi.org/10.1016/j.tree.2010.05.001

Schulze, E.-D., Körner, C., Law, B. E., Haberl, H., & Luyssaert, S. (2012). Large-scale bioenergy from additional harvest of forest biomass is neither sustainable nor greenhouse gas neutral. *GCB Bioenergy*, *4*(6), 611–616. https://doi.org/10.1111/j.1757-1707.2012.01169.x

Schumacher, S., Bugmann, H., & Mladenoff, D. J. (2004). Improving the formulation of tree growth and succession in a spatially explicit landscape model. *Ecological Modelling*, *180*(1), 175–194. https://doi.org/10.1016/j.ecolmodel.2003.12.055

Seijo, F., Millington, J. D. A., Gray, R., Sanz, V., Lozano, J., García-Serrano, F., Sangüesa-Barreda, G., & Julio Camarero, J. (2015). Forgetting fire: Traditional fire knowledge in two chestnut forest ecosystems of the Iberian Peninsula and its implications for European fire management policy. *Land Use Policy*, *47*, 130–144. https://doi.org/10.1016/j.landusepol.2015.03.006

Seijo, F., Millington, J. D., Gray, R., Mateo, L. H., Sangüesa-Barreda, G., & Camarero, J. J. (2016). Divergent Fire Regimes in Two Contrasting Mediterranean Chestnut Forest Landscapes. *Human Ecology*. https://doi.org/10.1007/s10745-016-9879-9

Sharma, A., & Tiwari, K. N. (2019). Sink removal from digital elevation model - a necessary evil for hydrological analysis. *Current Science*, *117*(9), 1512–1515. https://doi.org/10.18520/cs/v117/i9/1512-1515

Shin, H. (2021). Benefits of open research in social simulation: An early-career researcher's perspective [https://rofasss.org/2021/11/23/benefits-open-research/]. *Review of Artificial Societies and Social Simulation*.

Sills, J., Barton, C. M., Alberti, M., Ames, D., Atkinson, J.-A., Bales, J., Burke, E., Chen, M., Diallo, S. Y., Earn, D. J. D., Fath, B., Feng, Z., Gibbons, C., Hammond, R., Heffernan, J., Houser, H., Hovmand, P. S., Kopainsky, B., Mabry, P. L., . . . Tucker, G. (2020). Call for transparency of covid-19 models. *Science*, *368*(6490), 482–483. https://doi.org/10.1126/science.abb8637

Spooner, P. G., & Allcock, K. G. (2006). Using a state-and-transition approach to manage endangered Eucalyptus albens (White Box) woodlands. *Environmental Management*, *38*(5), 771–783. https://doi.org/10.1007/s00267-005-0133-2

Steffen, W., Sanderson, A., Tyson, P., Jäger, J., Matson, P., Moore III, B., Oldfield, F., Richardson, K., Schellnhuber, H.-J., Turner II, B., & Wasson, R. (2004). *Global Change and the Earth System: A Planet Under Pressure*. Springer. https://doi.org/10.1007/978-3-540-32730-1_16

Stephens, S. L., Burrows, N., Buyantuyev, A., Gray, R. W., Keane, R. E., Kubian, R., Liu, S., Seijo, F., Shu, L., Tolhurst, K. G., & van Wagtendonk, J. W. (2014). Temperate and boreal forest mega-fires: Characteristics and challenges. *Frontiers in Ecology and the Environment*, *12*(2), 115–122. https://doi.org/10.1890/120332

Sugita, S. (2007). Theory of quantitative reconstruction of vegetation I: pollen from large sites REVEALS regional vegetation composition. *The Holocene*, *17*, 229–241. https://doi.org/10.1177/0959683607075837

Sugita, S., Parshall, T., Calcote, R., & Walker, K. (2010). Testing the Landscape Reconstruction Algorithm for spatially explicit reconstruction of vegetation in northern Michigan and Wisconsin. *Quaternary Research*, *74*(2), 289–300. https://doi.org/10.1016/j.yqres.2010.07.008

Sun, Z., Lorscheid, I., Millington, J. D., Lauf, S., Magliocca, N. R., Groeneveld, J., Balbi, S., Nolzen, H., M?ller, B., Schulze, J., & Buchmann, C. M. (2016). Simple or complicated agent-based models? A complicated issue. *Environmental Modelling and Software*, *86*(3), 56–67. https://doi.org/10.1016/j.envsoft.2016.09.006

Symeonakis, E., Koukoulas, S., Calvo-Cases, A., Arnau-Rosalen, E., & Makris, I. (2004). A landuse change and land degradation study in Spain and Greece using remote sensing and GIS. In O. Altan (Ed.), *XXth ISPRS Congress, Istanbul, Turkey* (pp. 553–558).

Taillandier, P., Gaudou, B., Grignard, A., Huynh, Q.-N., Marilleau, N., Caillou, P., Philippon, D., & Drogoul, A. (2019). Building, composing and experimenting complex spatial models with the gama platform. *GeoInformatica*, *23*(2), 299–322. https://doi.org/10.1007/s10707-018-00339-6

The Open Modeling Foundation. (2023). Standards. URL: https://www.openmodelingfoundation.org/standards

Thiele, J. C., Kurth, W., & Grimm, V. (2014). Facilitating Parameter Estimation and Sensitivity Analysis of Agent-Based Models: A Cookbook Using NetLogo and R. *Journal of Artificial Societies and Social Simulation*, *17*(3), 11. https://doi.org/10.18564/jasss.2503

Thompson, J. D. (2005). *Plant Evolution in the Mediterranean*. Oxford University Press.

Tisue, S., & Wilensky, U. (2004). Netlogo: A simple environment for modeling complexity. *International Conference on Complex Systems*, 16–21.

Turner, M. G., & Gardner, R. H. (2015). *Landscape Ecology in Theory and Practice* (Second). Springer. https://doi.org/10.1007/978-1-4939-2794-4

Turner, M. G., Romme, W. H., Gardner, R. H., O'Neill, R. V., & Kratz, T. K. (1993). A revised concept of landscape equilibrium: Disturbance and stability on scaled landscapes. *Landscape Ecology*, *8*(3), 213–227. https://doi.org/10.1007/BF00125352

Twine, W. (2003). Consumption and direct-use values of savanna bio-resources used by rural households in Mametja, a semi-arid area of Limpopo province, South Africa. *South African Journal of Science*, *99*(September/October), 467–473.

Ullah, I. (2013). *The Consequences of Human land-use Strategies During the PPNB-LN Transition* (Doctoral dissertation May). Arizona State University.

UN General Assembly. (2015). *Transforming our world: The 2030 agenda for sustainable development* (tech. rep.). https://doi.org/10.1007/s13398-014-0173-7.2

United States Department of Agriculture. (2004). Estimation of Direct Runoff from Storm Rainfall. In *National engineering handbook part 630 (hydrology)*.

van der Knaap, W., & van Leeuwen, J. (1995). Holocene vegetation succession and degradation as responses to climatic change and human activity in the Serra de Estrela, Portugal. *Review of Paleobotany and Palynology*, 153–211.

van Benthem, J., van Ditmarsch, H., van Eijck, J., & Jaspars, J. (2016). Logic in Action. URL: http://www.logicinaction.org

Vitousek, P. M., Mooney, H. A., Lubchenco, J., & Melillo, J. M. (1997). Human Domination of Earth's Ecosystems. *Science*, *277*(5325), 494–499. https://doi.org/10.1126/science.277.5325.494

Wainwright, J., & Millington, J. (2010). Mind, the gap in landscape-evolution modelling. *Earth Surface Processes and Landforms*, *35*(7), 842–855. https://doi.org/10.1002/esp.2008

Wainwright, J. (2008). Can modelling enable us to understand the rôle of humans in landscape evolution? *Geoforum*, *39*(2), 659–674. https://doi.org/10.1016/j.geoforum.2006.09.011

Walker, M., Johnsen, S., Rasmussen, S. O., Popp, T., Steffensen, J. P., Gibbard, P., Hoek, W., Lowe, J., Andrews, J., Björck, S., Cwynar, L. C., Hughen, K., Kershaw, P., Kromer, B., Litt, T., Lowe, D. J., Nakagawa, T., Newnham, R., & Schwander, J. (2009). Formal definition and dating of the GSSP (Global Stratotype Section and Point) for the base of the Holocene using the Greenland NGRIP ice core, and selected auxiliary records. *Journal of Quaternary Science*, *24*(1), 3–17. https://doi.org/10.1002/jqs.1227

Wangersky, P. J. (1978). Lotka-volterra population models. *Annual Review of Ecology and Systematics*, *9*, 189–218.

Watts, K., Fuentes-Montemayor, E., Macgregor, N. A., Peredo-Alvarez, V., Ferryman, M., Bellamy, C., Brown, N., & Park, K. J. (2016). Using historical woodland creation to construct a long-term, large-scale natural experiment: the WrEN project. *Ecology and Evolution*, *6*(9), 3012–3025. https://doi.org/10.1002/ece3.2066

Whitlock, C., & Larsen, C. (2001). Charcoal as a fire proxy. In J. P. Smol, H. Birks & W. M. Last (Eds.), *Tracking environmental change using lake sediments* (pp. 75–97). Springer Netherlands. https://doi.org/10.1007/0-306-47668-1_5

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., . . . Mons, B. (2016). Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(160018), 1–9. https://doi.org/10.1038/sdata.2016.18

Williams, J. W., Grimm, E. C., Blois, J. L., Charles, D. F., Davis, E. B., Goring, S. J., Graham, R. W., Smith, A. J., Anderson, M., Arroyo-Cabrales, J., & et al. (2018). The neotoma paleoecology database, a multiproxy, international, community-curated data resource. *Quaternary Research*, *89*(1), 156–177. https://doi.org/10.1017/qua.2017.105

Wilson, G., Aruliah, D. A., Brown, C. T., Chue Hong, N. P., Davis, M., Guy, R. T., Haddock, S. H., Huff, K. D., Mitchell, I. M., Plumbley, M. D., Waugh, B., White, E. P., & Wilson, P. (2014). Best Practices for Scientific Computing. *PLoS Biology*, *12*(1). https://doi.org/10.1371/journal.pbio.1001745

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., & Teal, T. K. (2017). Good enough practices in scientific computing. *PLOS Computational Biology*, *13*(6), 1–20. https://doi.org/10.1371/journal.pcbi.1005510

Wilson, G. V. (2006). Where's the Real Bottleneck in Scientific Computing? *American Scientist*, *94*(1), 5–6.

Winsberg, E. (1999). Sanctioning Models: The Epistemology of Simulation. *Science in Context*, *12*(02), 275–292. https://doi.org/10.1017/S0269889700003422

Winsberg, E. (2009). Computer Simulation and the Philosophy of Science. *Philosophy Compass*, *4*(5), 835–845. https://doi.org/10.1111/j.1747-9991.2009.00236.x

Yll, E.-I., Perez-Obiol, R., Pantaleon-Cano, J., & Roure, J. M. (1997). Palynological Evidence for Climatic Change and Human Activity during the Holocene on Minorca (Balearic Islands). *Quaternary Research*, *48*, 339–347. https://doi.org/10.1006/qres.1997.1925

Yuen, K. K. (1974). The two-sample trimmed t for unequal population variances. *Biometrika*, *61*(1), 165–170.

Zavala, M., Espelta, J., & Retana, J. (2000). Constraints and trade-offs in Mediterranean plant communities: the case of holm oak - aleppo pine forests. *The Botanical Review*, *66*(1), 119–149. https://doi.org/10.1007/bf02857785

Zavala, M. A., & Zea, E. (2004). Mechanisms maintaining biodiversity in Mediterranean pine-oak forests: Insights from a spatial simulation model. *Plant Ecology*, *171*(1-2), 197–207. https://doi.org/10.1023/B:VEGE.0000029387.15947.b7

Zilhão, J. (1993). The spread of agro-pastoral economies across Mediterranean Europe: a view from the far west. *Journal of Mediterranean Archaeology*, *6*(1), 5–63. https://doi.org/10.1558/jmea.v6i1.5

Appendix A

Regular expressions for plant functional types table

Functional type	Description	Regular expression
Grassland	Grasses	.*Poaceae|.*Gramineae|.*Cerealia
Shrubland	Box plant (shrubby tree)	.*Buxus
	Bracken/ ferns	.*Pteridium|.*Polypodium|.*Filicales
	Celery and marthwort genus	.*Apium
	Celery, carrot, parsley family	.*Umbelliferae
	Cypress family	.*Cupressaceae
	Doc/ sorrel genus	.*Rumex
	Family of shrubby plants	.*Asteroideae
	Flowering plants found in wet regions	.*Sparganium|.*Typha angustifolia
	Flowering plants in lettuce/ dendelion family	.*Cichorioideae
	Flowering, perennial shrubs	.*Cistus
	Genus of flowering plants in the cashew family	.*Pistacia
	Genus of flowering plants in the family Fabaceae. Thorny, evergreen shrubs.	.*Ulex
	Genus of flowering plants including buttercup	.*Ranunculus
	Genus of gymnosperm shrubs	.*Ephedra
	Goosefoot family	.*Chenopodiaceae
	Heather family	.*Calluna vulgaris|.*Erica(ceae|-type|\s)
	Juniper	.*Juniperus

Continued on next page

Functional type	Description	Regular expression
	Mugwort genus	.*Artemisia
	Olive genus	.*Olea
	Plantain/ fleawort genus	.*Plantago
	Quillwort	.*Isoetes
	Sedge family (superficially resemble grasses)	.*Cyperaceae
Pine forest	Pine genus	\s?Pinus\s?
Deciduous forest	Alder genus	.*Alnus
	Beech family	.*Fagaceae
	Beech genus	.*Fagus
	Birch genus	.*Betula
	Chestnut genus	.*Castanea
	Hazel genus	.*Corylus
	Hornbeam genus	.*Carpinus
	Willow genus	.*Salix
Oak forest	Oak genus	\s?Quercus\s?

Table A.1: Regular expressions (’regexs’) used to map the names of plant species to plant functional types as described in Section 3.2.2.

Appendix B

Model description table

Structural elements	Sub-elements	Question No.	Guiding questions	Response
Overview	Purpose	1.1.1	What is the purpose of the study?	1. Test hypotheses which attempt to explain changes to vegetation cover during the first millenia of agriculture in Iberia in terms of human agricultural activities, ecological succession, natural disturbance, and climate. 2. Identify qualitatively distinct landscape states, defined in terms of both social and ecological variables, and which are emergent properties of the model.
		1.1.2	For whom is the model designed?	Scientists interested in one or more of paleoecology, landscape ecology, complex adaptive systems, and nonequilibrium systems.

Entities, state variables, and scales	1.2.1	What kinds of entities are in the model?	**Anthropogenic:** Each agent consists of a troop of humans of variable size (enivsaged as an extended family group) all of whom have opted to specialise in a particular Resource Management Strategy (RMS). Each RMS is chosen to represent a lifestyle available to humans during the first millenia of agriculture: - Hunter-gatherer - Pastoralist - Agriculturalist **Ecological:** Grid cell characterised by a particular land cover type (e.g. Oak forest, Pine forest, grassland).

1.2.2	By what attributes (i.e. state variables and parameters) are these entities characterized?	**Agents:** - ID number, i - Position of home cell, (i, j) - Resource Management Strategy, RMS_i - Reproductive rate, r_i - Size of group, N_i - Land occupancy (list of grid cells), O_i - Resource Requirement Quota, $\text{RRQ}_i = \text{RRQ}_i(N_i)$ - Generational score for resource management strategy, S_i. **Grid cells:** - Position in space, (i, j) - ID of controlling agent, A_{ij} - RMS of controlling agent, A_{ij}^{RMS} - Land-cover class, L_{ij}^{c} - Land-cover flammability, $L_{ij}^{f} = L_{ij}^{f}(L_{ij}^{c})$
1.2.3	What are the exogenous factors / drivers of the model?	Climatic variables (temperature and precipitation), and the time of arrival of first agent which is not a hunter-gatherer (i.e. either a pastoralist or an agriculturalist).
1.2.4	If applicable, how is space included in the model?	Space is included explicitly in the model, and is linked via georeference to a Digital Elevation Model (DEM) and spatially and temporally varying climatic data.

1.2.5	What are the temporal and spatial resolutions and extents of the model?	One timestep represents a 40 year human generation, and simulations are run for 5000 simulated years (125 timesteps). Grid cells represent 30 m^2 (0.003 ha) and the modelled landscape comprises 30 km^2 (3000 ha), giving a total of 10^6 grid cells.

Process overview and scheduling	1.3.1	What entity does what, and in what order?	Each timestep of the anthropogenic land cover change model consists of the following steps which are repeated until simulation timesteps are completed. **Evolve vegetation in model cells** - 40 years of succession and fire regime - Calculate annual food yields for each cell **Evaluate agent performance in preceeding generation** - Analyse annual food yields - Wealthier neighbours share resources to help others - Assign generation scores to agents **Update agents in model** - Agents leave model grid if score < threshold - Remaining agents decide whether or not to update their RMS by comparing their score with neighbours - New agents arrive in the model grid, claiming a patch of contiguous land large enough to support them given their initial RMS **Update model grid** - Change cell land-cover classes according to occupying agents' RMSs

Design concepts	Theoretical and Empirical Background	2.1.1	Which general concepts, theories or hypotheses are underlying the model's design at the system level or at the level(s) of the submodel(s) (apart from the decision model)? What is the link to complexity and the purpose of the model?	**Anthropogenic:** Sedentism and transition to agriculture allowed for increase in family size resulting in overall population growth Bellwood, 2004; Moran, 2006. Pre-historic humans engaged in strategic food sharing to mitigate risk of food source failure Moran, 2006. **Succession submodel:** Ecological succession (Redman, 1999) and disturbance (Pickett & White, 1985; Turner & Gardner, 2015; Turner et al., 1993), especially the work of Barbero et al., 1990 on the role of Mediterranean species' life history traits and environmental conditions in determining succession pathways realised at a location. Another key concept is that of a plant functional type as a way to group together species which respond similarly to environmental conditions, including disturbance regimes (M. Zavala et al., 2000). **Fire submodel:** Fire spread is assumed to be adequately modelled by mechanistic cellular automata based approaches, such as that discussed in J. D. A. Millington et al., 2009.

2.1.2	On what assumptions is/are the agents' decision model(s) based?	The overarching assumption governing agent decision-making is that upon observing that, over the previous generation, neighbouring agents had significantly more to eat than you did, you would likely choose to change your own resource management strategy to that which turned out to be most successful amongst your neighbours.
2.1.3	Why is a/are certain decision model(s) chosen?	Because of the nature of archaeological research – namely its focus on the retrieval and study of material artefacts – it has proved extremely difficult to obtain data or references to theories describing something as intangible as Neolithic /culture/. The chosen model of decision making around land use change is presented as a general and intuitively appealing theory about how land use decisions were made among early agriculturalists. I will use the model as part of my wider Pattern Orientated Modelling strategy to evaluate the accuracy of this decision model, and propose possible improvements.

	2.1.4	If the model / a submodel (e.g. the decision model) is based on empirical data, where does the data come from?	The succession submodel uses climate data derived generated by General Circulation Models running climate simulations for the Holocene (Hijmans et al., 2005). The fire spread submodel will use a Digital Elevation Model (DEM) based on Shuttle Radar Topography Mission (STRM) data at 1 arc-sec resolution.
	2.1.5	At which level of aggregation were the data available?	Climate data is available at 900 m resolution. DEM data is available at 30 m resolution (the same resolution as model grid cells).
Individual Decision Making	2.2.1	What are the subjects and objects of decision-making? On which level of aggregation is decision-making modeled? Are multiple levels of decision making included?	The human agents are the subjects of decision making, with the choice of land use being the object.
	2.2.2	What is the basic rationality behind agents' decision-making in the model? Do agents pursue an explicit objective or have other success criteria?	The agents attempt to optimize their food availability with respect to the amount of effort required to obtain it, and are assumed to behave rationally in the situation of observing that other strategies have worked better than their own.

2.2.3	How do agents make their decisions?	Decision tree.
2.2.4	Do the agents adapt their behavior to changing endogenous and exogenous state variables? And if yes, how?	Yes. Changing climatic conditions during the course of model runs will cause different succession pathways to become prevalent at different times, resulting in changes in the productivity of different cells for agents with different Resporce Management Strategies. Additionally, the presence of fire in the landscape means areas in the model grid can change very quickly. Agents adapt to these endogenous environmental changes by choosing an RMS which worked successfully for their neighbours in the previous generation.
2.2.6	Do spatial aspects play a role in the decision process?	Space plays a role in the decision process only insofar as agents make decisions based on the performance of their nearest N_n neighbours.
2.2.7	Do temporal aspects play a role in the decision process?	There is no spatial aspect to the decision process.
2.2.8	To which extent and how is uncertainty included in the agents' decision rules?	Not at all, decision making is deterministic.

Individual Sensing	2.4.1	What endogenous and exogenous state variables are individuals assumed to sense and consider in their decisions? Is the sensing process erroneous?	Agents are able to sense whether or not a grid cell is occupied by any other agent, which facilitates agents settling on an unoccupied home cell. This sensing process is not erroneous.
	2.4.2	What state variables of which other individuals can an individual perceive? Is the sensing process erroneous?	Agents can perceive resource management strategy, RMS_j and score S_j of neighbouring N_n agents. This sensing process is not erroneous.
	2.4.3	What is the spatial scale of sensing?	Knowledge of grid cell occupation is global, but each agent's knowledge of neighbours' Resource Management Strategies and generational score is limited to their N_n nearest neighbours.
	2.4.4	Are the mechanisms by which agents obtain information modeled explicitly, or are individuals simply assumed to know these variables?	Agents are assumed to know state variables relavent to their decision making.

Individual Prediction	2.5.1	Which data uses the agent to predict future conditions?	Observation of which of their neighbours had successful Resource Management Strategies in the previous generation. Agents use this knowledge to infer that a similar RMS will perform well in the subsequent generation.
	2.5.2	What internal models are agents assumed to use to estimate future conditions or consequences of their decisions?	Because timesteps represent human generations, it is assumed that agents base their decision of which RMS to follow exclusively on the performance of their own RMS in the previous generation compared to that of their neighbours.
Interaction	2.6.1	Are interactions among agents and entities assumed as direct or indirect?	Agents interact both directly through strategic food sharing, and and indirectly through the environment, via seed spread from neighbouring patches managed by different agents.
	2.6.2	On what do the interactions depend?	Neighbourhood (agents detect the wealth of their N_n neighbours).

Heterogeneity	2.8.1	Are the agents heterogeneous? If yes, which state variables and/or processes differ between the agents?	Yes, the main difference being how agents with different resource management strategies (hunter gatherer, pasturalist and agriculturalist) make systematic changes to the land under their control. Whereas pasturalists and agriculturalists will use fire and practice cultivation to modify land cover, HG are assumed to obtain their sustenance without directly changing the vegetative composition of the landscape (albeit making use of much more land).
	2.8.2	Are the agents heterogeneous in their decision-making? If yes, which decision models or decision objects differ between the agents?	No, the decision rules for transitioning between different resource management strategies is common to all agents.
Stochasticity	2.9.1	What processes (including initialization) are modeled by assuming they are random or partly random?	- Number of new agents from outside the model grid in each generation, and their initial RMS. - Number of fires occurring naturally in each simulated year. - Fire spread mechanism - Seed dispersal

	Observation	2.10.1	What data are collected from the ABM for testing, understanding, and analyzing it, and how and when are they collected?	At the end of each simulated generation, the model will write to disk: - Number of cells in each land-cover class - Spatial statistics, including autocorrelations for each pair of land-cover classes - Sizes of each fire initialised in the 40 year period For debugging and heuristic purposes, the model will be able to output raster grids specifying the land-cover class of every model cell at each generation time-step. However, as this is likely to be computationally expensive, this will be implemented as an option which will be deselected for most model runs.
		2.10.2	What key results, outputs or characteristics of the model are emerging from the individuals? (Emergence)	At the end of each simulated generation, the model will write to disk the RMS and generation score of each agent in the model.
Details	Implementation details	3.1.1	How has the model been implemented?	The model will be implemented in Java using the Repast framework.

Initialization	3.2.1	What is the initial state of the model world, i.e. at time $t = 0$ of a simulation run?	The initial state of the model grid will be determined by manually creating an artificial landscape in which the abundance of different vegetation types is consistent with that found in pollen assemblage data 100 simulated years *prior* to the postulated date of the arrival of the first agriculturalist or pastoralist. I will then perform a 100 year 'Spin-up' period (McGuffie & Henderson-Sellers, 2005) in which the artificial landscape will be allowed to evolve away from its initial conditions consistently with the model's succession submodel. At this point, the model will be counted as being at $t = 0$, and N_0 agents will be introduced into the manner in the same manner as described under 'Update agents in model' in response to question 1.3.1.
	3.2.2	Is initialization always the same, or is it allowed to vary among simulations?	Initialisation procedure is always the same, but a variety of artificial landscapes will be used, and the differences in results compared.
	3.2.3	Are the initial values chosen arbitrarily or based on data?	Arbitrarily.

Table B.1: Model specification within the ODD+D protocol proposed by Müller et al., 2013

Appendix C

Soil moisture curve numbers

The amount of water retained in grid cells following precipitation is determined by a *curve number* which depends on slope, soil type, and land-cover type. After Ferrér et al., 1995; J. D. A. Millington et al., 2009; Symeonakis et al., 2004.

Table C.1: Curve numbers for various slopes, soil types, and land-cover types in AgroSuccess.

		LCT							
Soil	Slope [%]	Pine	T. Forest	Wheat	DAL	Deciduous	Shrubland	Oak	Burnt
A	< 3	35	35	62	62	35	46	35	91
B	< 3	54	54	72	72	54	68	54	91
C	< 3	69	69	78	78	69	78	69	91
D	< 3	77	77	82	82	77	83	77	91
A	≥ 3	39	39	65	65	39	46	39	94
B	≥ 3	60	60	76	76	60	68	60	94
C	≥ 3	73	73	84	84	73	78	73	94
D	≥ 3	78	78	87	87	78	83	78	94

Appendix D

Logical expressions

The land-cover state update rules expressed in Section 4.2.3 are provided as expressions using formal logical notation. To assist readers who are not familiar with the notation used in the main text, I here provide verbal interpretations of examples of the three forms of these expressions used in the text, along with an explanation of the symbols used. See also e.g. van Benthem et al., 2016 for further background on the logical formalism used here.

Eq. (4.2) is

$$\Delta D = \Delta D(t) \in L \cup \{\varnothing\}$$

This specifies that the symbol for a simulation cell's target state, ΔD, is implicitly a time dependent variable since $\Delta D = \Delta D(t)$. The value of ΔD is *either* one of the elements of the set L (one of the land-cover types included in the model) *or* no land-cover type. This corresponds to the fragment $\Delta D(t) \in L \cup \{\varnothing\}$. Here $\in$ says the value on the left hand side is one of the elements specified by the set on the right hand side. $\cup$ is the union operator, such that $L \cup \{\varnothing\}$ equals the set L plus 'nothing' (expressed here as $\varnothing$).

Eq. (4.6) is

$$Dt \leftrightarrow \Delta D(t) = \Delta D(t-1)$$

which could be stated as "Dt is true if and only if a simulation cell's target state in the current time step, $\Delta D(t)$, is the same as it was in the previous time step, $\Delta D(t-1)$". This allows us to use the symbol Dt in subsequent expressions to assert that a cell's target state hasn't changed, or its logical negation, $\neg Dt$, to assert that a cell's target state has changed. The symbol $\leftrightarrow$ is the logical biconditional operator, and is used to assert that the expressions on its left and right are logically equivalent.

Eq. (S1c) is

$$\forall t \in \mathbb{N} \setminus \{0\} \, (Ct \wedge \neg Dt \rightarrow T_{\text{in}}(t) = 1)$$

This can be read as, “For any time step (specified by a natural number that is not 0), if a simulation cell’s start state is the same as it was in the previous time step, *and* the cell’s target state has changed, then the simulation cell’s ‘time in state’ value is equal to 1”. The ‘universal quantifier’, $\forall t$, communicates that the rule applies to any value of t, and the set notation $\in \mathbb{N} \setminus \{0\}$ limits the values of t considered to those that represent valid time steps in the simulation $(1, 2, 3, \ldots)$. The logical predicate inside the parentheses then specifies the rule that applies to those values of t. Ct and Dt are defined in the sentences Eq. (4.5) and Eq. (4.6) respectively. The conditional operator, $\rightarrow$, specifies that if the expression to its left is true, then the expression right hand is also true.

Appendix E

Differences between AgroSuccess and MedLanD

Here I provide details of how the decision rules used by agriculturalist agents in AgroSuccess differ from those in MedLanD.

E.1 Reformulation of equation for number of required wheat patches

Ullah, 2013, p. 89, gives Eq. (E.1) as the number of wheat plots required per household in year t.

$$N_t^{(w)} = \frac{P_t^{(h)}\, M^{(w)}\, (1 + p_s)}{\mu_{t-1}^{(w)}\, C} \tag{E.1}$$

As $N_t^{(w)}$ is a unitless quantity, the right-hand side of Eq. (E.1) should also be unitless. However, dimensional analysis shows its units are in fact ha. This is consistent with the fact that we expect the number of wheat plots required to be inversely proportional to the area of each grid cell. Eq. (E.1) can be corrected by normalising with respect to the area of each grid cell expressed in ha, $A_r/10000$ such that Eq. (E.1) becomes

$$N_t^{(w)} = \frac{10000 \cdot P_t^{(h)}\, M^{(w)}\, (1 + p_s)}{\mu_{t-1}^{(w)}\, A_r\, C} \tag{E.2}$$

Although logically consistent, Eq. (E.2) is not as clear as it could be because the mass of wheat needed per person per year, $M^{(w)}$ is dependent on two other parameters: the total number of calories required per person per day, E_{tot}, and the energy yield per kilogram of wheat, $E^{(w)}$.

Specifically

$$M^{(w)} = \frac{365 \cdot E_{\text{tot}}}{E^{(w)}} \,. \tag{E.3}$$

To simplify reasoning about $N_t^{(w)}$, Eq. (4.18) expresses $N_t^{(w)}$ in terms of these basic parameters explicitly.

E.2 Exclusion of calculation of soil depth

Ullah, 2013, p. 95 use "the average of three regressions against precipitation, soil depth, and soil fertility" (their Eq. 4.10) to calculate wheat returns with respect to annual precipitation, soil depth, and soil fertility as follows:

$$r_i^{(w)} = \frac{(0.51 \ln(P') + 1.03) + (0.28 \ln(\text{SD}_i) + 0.87) + (0.19 \ln(F_i) + 1)}{3}$$

here, annual precipitation, P', and soil depth in cell i, SD_i, are in metres, and soil fertility in cell i, F_i, is a proportion between 0 and 1.

Since I do not model soil depth in AgroSuccess, I drop the term representing the regression over soil depth and take the average over the remaining regressions over annual precipitation and soil fertility. This results in Eq. (4.27) as described in the main text.

E.3 Exclusion of pastoralism

Households in MedLanD farm barley which they use as fodder to feed to sheep and goats in addition to wheat, which is farmed for direct human consumption. While it would be interesting to add barley farming to support pastoralism to AgroSuccess, I was unable to obtain sufficient detail to implement this mechanism from Ullah, 2013 for the following reason.

Ullah, 2013, p. 90 used Eq. (E.4) to calculate the number of patches required for grazing.

$$N_g = \frac{P_h \, P_{oc} \, DM_{\text{tot}} \, (1 + p_b) - DM_{\text{stub}}}{\mu_g} \tag{E.4}$$

The symbol DM_{stub} represents the contribution to herd diet from grazing on the stubble of harvested fields controlled by a household. However, I have not been able to find a description of how DM_{stub} is calculated in Ullah, 2013. This highlights the challenge of documenting models as complicated as MedLand. Because it is expected that wild fodder will make up a significant fraction of ovicaprid diets, it is not appropriate to simply assume $DM_{\text{stub}} = 0$. Therefore, for the version of AgroSuccess described in this thesis, I omit any pastoralism mechanism and assume all household calories are derived from wheat agriculture only.

Appendix F

Additional sensitivity analysis results

Table F.1: Sensitivity analysis results for the Algendar study site. The 'Value' column specifies the default value for each parameter, and the 'Value +10%' and 'Value -10%' columns give the corresponding values for parameters used in the +10% and -10% scenarios. The remaining columns give the mean values under the +10% and -10% scenarios for the *Shrubland, Mature Forest, Farmed Area Wood Area,* and *Burned Area* output variables averaged over 10 simulation replicas, along with the corresponding percentage change relative to the default scenario in parentheses. Values in **bold** show statistically significant differences with respect to the default scenario (see Section 6.1.2). Output variables indicated by * showed 0 variance in default scenario runs, so outputs that differ from the default scenario values by $\leq -10\%$ or $\geq +10\%$ are shown as significant.

Parameter	Value	Value -10%	Value +10%	Shrubland, -10%	Shrubland, +10%	Mature Forest, -10%	Mature Forest, +10%	Farmed Area, -10%*	Farmed Area, +10%*	Wood Area, -10%*	Wood Area, +10%*	Burned Area, -10%	Burned Area, +10%
Number of households per village	10.00	9.000	11.000	**0.057 (13%)**	**0.066 (31%)**	**0.930 (-1%)**	**0.918 (-2%)**	**6.704 (-10%)**	8.194 (10%)	2.438 (-10%)	2.980 (10%)	0.005 (2%)	**0.006 (19%)**
Labour availability	300.00	270.000	330.000	**0.056 (11%)**	0.053 (6%)	**0.930 (-1%)**	0.932 (-0%)	7.449 (0%)	7.449 (0%)	2.709 (0%)	2.709 (0%)	**0.005 (14%)**	0.005 (10%)
Wheat farming labour requirement	50.00	45.000	55.000	0.053 (4%)	0.054 (7%)	0.932 (-0%)	0.934 (-0%)	7.449 (0%)	7.449 (0%)	2.709 (0%)	2.709 (0%)	0.005 (5%)	0.004 (-10%)
Maximum patch farm time	50.00	45.000	55.000	**0.056 (10%)**	**0.061 (21%)**	**0.931 (-1%)**	**0.924 (-1%)**	7.449 (0%)	7.449 (0%)	2.709 (0%)	2.709 (0%)	0.005 (4%)	0.005 (6%)
Farmer conservativeness scalar	0.75	0.675	0.825	**0.062 (23%)**	**0.055 (10%)**	**0.924 (-1%)**	**0.926 (-1%)**	8.126 (9%)	6.772 (-9%)	2.709 (0%)	2.709 (0%)	0.005 (9%)	**0.005 (16%)**
Crop reseed proportion	0.15	0.135	0.165	**0.062 (22%)**	0.054 (7%)	**0.922 (-2%)**	**0.930 (-1%)**	7.449 (0%)	7.449 (0%)	2.709 (0%)	2.709 (0%)	**0.006 (19%)**	0.004 (-4%)
Farm value distance parameter	1.00	0.900	1.100	**0.062 (22%)**	**0.061 (20%)**	**0.917 (-2%)**	**0.925 (-1%)**	**8.939 (20%)**	7.449 (0%)	**3.250 (20%)**	2.709 (0%)	**0.007 (54%)**	0.005 (-0%)
Farm value fertility parameter	1.00	0.900	1.100	**0.064 (26%)**	**0.055 (8%)**	**0.918 (-2%)**	0.932 (-0%)	7.449 (0%)	7.449 (0%)	2.709 (0%)	2.709 (0%)	**0.006 (23%)**	0.004 (-6%)
Farm value land cover conversion parameter	1.00	0.900	1.100	0.052 (3%)	**0.062 (22%)**	0.932 (-0%)	**0.924 (-1%)**	7.449 (0%)	7.449 (0%)	2.709 (0%)	2.709 (0%)	0.005 (-1%)	**0.006 (19%)**
Firewood required per capita per year	1100.00	990.000	1210.000	**0.066 (31%)**	0.051 (2%)	**0.920 (-2%)**	0.932 (-0%)	7.449 (0%)	7.449 (0%)	**2.032 (-25%)**	2.709 (0%)	0.005 (8%)	0.005 (1%)
Firewood biomass removal rate	0.10	0.090	0.110	**0.063 (24%)**	**0.062 (23%)**	**0.921 (-2%)**	**0.923 (-1%)**	7.449 (0%)	7.449 (0%)	2.709 (0%)	**2.032 (-25%)**	0.005 (10%)	0.005 (10%)
Wood value distance parameter	1.00	0.900	1.100	0.054 (7%)	0.054 (6%)	0.933 (-0%)	**0.929 (-1%)**	7.449 (0%)	7.449 (0%)	2.709 (0%)	2.709 (0%)	0.005 (-2%)	0.005 (9%)
Climax forest biomass density	300000.00	270000.000	330000.000	0.052 (2%)	0.052 (2%)	**0.928 (-1%)**	**0.929 (-1%)**	7.449 (0%)	7.449 (0%)	2.709 (0%)	**2.032 (-25%)**	0.005 (3%)	0.004 (-5%)
Land cover colonisation base rate	0.05	0.045	0.055	**0.060 (18%)**	0.053 (5%)	**0.925 (-1%)**	0.933 (-0%)	7.449 (0%)	7.449 (0%)	2.709 (0%)	2.709 (0%)	0.005 (7%)	0.004 (-4%)
Land cover colonisation spread rate	0.20	0.180	0.220	**0.060 (19%)**	**0.056 (11%)**	**0.927 (-1%)**	0.931 (-0%)	7.449 (0%)	7.449 (0%)	2.709 (0%)	2.709 (0%)	0.005 (-2%)	0.004 (-4%)
Mesic threshold	500.00	450.000	550.000	0.050 (-0%)	**0.061 (21%)**	0.934 (-0%)	**0.926 (-1%)**	7.449 (0%)	7.449 (0%)	2.709 (0%)	2.709 (0%)	**0.004 (-15%)**	0.005 (-1%)
Hydric threshold	1000.00	900.000	1100.000	**0.059 (18%)**	**0.059 (17%)**	**0.926 (-1%)**	**0.928 (-1%)**	7.449 (0%)	7.449 (0%)	2.709 (0%)	2.709 (0%)	0.005 (-1%)	0.004 (-5%)

Table F.2: Sensitivity analysis results for the Atxuri study site. The 'Value' column specifies the default value for each parameter, and the 'Value +10%' and 'Value -10%' columns give the corresponding values for parameters used in the +10% and -10% scenarios. The remaining columns give the mean values under the +10% and -10% scenarios for the *Shrubland, Mature Forest, Farmed Area Wood Area,* and *Burned Area* output variables averaged over 10 simulation replicas, along with the corresponding percentage change relative to the default scenario in parentheses. Values in **bold** show statistically significant differences with respect to the default scenario (see Section 6.1.2). Output variables indicated by * showed 0 variance in default scenario runs, so outputs that differ from the default scenario values by $\leq -10\%$ or $\geq +10\%$ are shown as significant.

Parameter	Value	Value -10%	Value +10%	Shrubland, -10%	Shrubland, +10%	Mature Forest, -10%	Mature Forest, +10%	Farmed Area, -10%*	Farmed Area, +10%*	Wood Area, -10%*	Wood Area, +10%*	Burned Area, -10%	Burned Area, +10%
Number of households per village	10.00	9.000	11.000	0.207 (-1%)	0.209 (-0%)	**0.691 (-1%)**	0.695 (-1%)	5.305 (-10%)	6.484 (10%)	**2.358 (-10%)**	2.882 (10%)	0.020 (3%)	0.020 (3%)
Labour availability	300.00	270.000	330.000	**0.204 (-2%)**	**0.199 (-5%)**	**0.707 (1%)**	**0.708 (1%)**	5.894 (0%)	5.894 (0%)	2.620 (0%)	2.620 (0%)	**0.018 (-6%)**	0.020 (2%)
Wheat farming labour requirement	50.00	45.000	55.000	**0.204 (-2%)**	0.208 (-0%)	0.699 (-0%)	0.697 (-0%)	5.894 (0%)	5.894 (0%)	2.620 (0%)	2.620 (0%)	**0.020 (3%)**	0.020 (2%)
Maximum patch farm time	50.00	45.000	55.000	0.206 (-1%)	0.208 (-0%)	0.701 (0%)	**0.689 (-2%)**	5.894 (0%)	5.894 (0%)	2.620 (0%)	2.620 (0%)	0.019 (-1%)	**0.020 (3%)**
Farmer conservativeness scalar	0.75	0.675	0.825	0.206 (-1%)	0.206 (-1%)	0.694 (-1%)	0.696 (-0%)	**6.549 (11%)**	5.894 (0%)	2.620 (0%)	2.620 (0%)	0.020 (2%)	0.020 (1%)
Crop reseed proportion	0.15	0.135	0.165	0.206 (-1%)	0.205 (-2%)	0.701 (0%)	0.702 (0%)	5.894 (0%)	**6.549 (11%)**	2.620 (0%)	2.620 (0%)	0.020 (1%)	0.019 (-2%)
Farm value distance parameter	1.00	0.900	1.100	0.208 (-0%)	**0.202 (-3%)**	0.696 (-1%)	0.701 (0%)	5.894 (0%)	5.894 (0%)	2.620 (0%)	2.620 (0%)	0.020 (3%)	0.020 (1%)
Farm value fertility parameter	1.00	0.900	1.100	0.211 (1%)	0.212 (2%)	**0.693 (-1%)**	**0.692 (-1%)**	5.894 (0%)	5.894 (0%)	2.620 (0%)	2.620 (0%)	0.020 (1%)	0.020 (1%)
Farm value land cover conversion parameter	1.00	0.900	1.100	**0.195 (-7%)**	0.209 (0%)	**0.710 (1%)**	0.698 (-0%)	5.894 (0%)	5.894 (0%)	2.620 (0%)	2.620 (0%)	0.019 (-3%)	**0.019 (-4%)**
Firewood required per capita per year	1100.00	990.000	1210.000	**0.204 (-2%)**	0.206 (-1%)	0.699 (-0%)	0.700 (0%)	5.894 (0%)	5.894 (0%)	2.620 (0%)	2.620 (0%)	**0.020 (3%)**	0.019 (-1%)
Firewood biomass removal rate	0.10	0.090	0.110	0.209 (0%)	0.208 (-0%)	0.697 (-0%)	0.697 (-0%)	5.894 (0%)	5.894 (0%)	2.620 (0%)	2.620 (0%)	0.019 (1%)	0.020 (2%)
Wood value distance parameter	1.00	0.900	1.100	**0.204 (-2%)**	0.211 (1%)	0.696 (-1%)	0.696 (-1%)	5.894 (0%)	5.894 (0%)	2.620 (0%)	2.620 (0%)	**0.021 (6%)**	0.019 (-3%)
Climax forest biomass density	300000.00	270000.000	330000.000	0.210 (1%)	0.213 (2%)	0.699 (-0%)	**0.693 (-1%)**	5.894 (0%)	5.894 (0%)	2.620 (0%)	2.620 (0%)	0.020 (1%)	0.020 (2%)
Land cover colonisation base rate	0.05	0.045	0.055	0.207 (-1%)	0.209 (0%)	0.698 (-0%)	**0.692 (-1%)**	5.894 (0%)	5.894 (0%)	2.620 (0%)	2.620 (0%)	0.019 (-2%)	**0.020 (5%)**
Land cover colonisation spread rate	0.20	0.180	0.220	**0.215 (3%)**	0.210 (1%)	**0.691 (-1%)**	**0.692 (-1%)**	5.894 (0%)	5.894 (0%)	2.620 (0%)	2.620 (0%)	0.019 (-1%)	**0.020 (3%)**
Mesic threshold	500.00	450.000	550.000	**0.186 (-11%)**	**0.200 (-4%)**	**0.728 (4%)**	**0.707 (1%)**	5.894 (0%)	5.894 (0%)	2.620 (0%)	2.620 (0%)	**0.016 (-16%)**	0.019 (-2%)
Hydric threshold	1000.00	900.000	1100.000	0.211 (1%)	0.207 (-1%)	**0.691 (-1%)**	**0.707 (1%)**	5.894 (0%)	5.894 (0%)	2.620 (0%)	2.620 (0%)	0.019 (0%)	0.019 (-2%)

Table F.3: Sensitivity analysis results for the Charco da Candieira study site. The 'Value' column specifies the default value for each parameter, and the 'Value +10%' and 'Value -10%' columns give the corresponding values for parameters used in the +10% and -10% scenarios. The remaining columns give the mean values under the +10% and -10% scenarios for the *Shrubland, Mature Forest, Farmed Area Wood Area,* and *Burned Area* output variables averaged over 10 simulation replicas, along with the corresponding percentage change relative to the default scenario in parentheses. Values in **bold** show statistically significant differences with respect to the default scenario (see Section 6.1.2). Output variables indicated by * showed 0 variance in default scenario runs, so outputs that differ from the default scenario values by $\leq -10\%$ or $\geq +10\%$ are shown as significant.

Parameter	Value	Value -10%	Value +10%	Shrubland, -10%	Shrubland, +10%	Mature Forest, -10%	Mature Forest, +10%	Farmed Area, -10%*	Farmed Area, +10%*	Wood Area, -10%*	Wood Area, +10%*	Burned Area, -10%	Burned Area, +10%
Number of households per village	10.00	9.000	11.000	**0.312 (2%)**	0.305 (-0%)	0.516 (-0%)	**0.527 (2%)**	5.176 (-10%)	6.327 (10%)	2.588 (-10%)	3.163 (10%)	0.033 (-2%)	**0.031 (-7%)**
Labour availability	300.00	270.000	330.000	0.308 (1%)	0.308 (1%)	0.516 (-0%)	0.525 (1%)	5.751 (0%)	5.751 (0%)	2.876 (0%)	2.876 (0%)	**0.032 (-4%)**	**0.032 (-4%)**
Wheat farming labour requirement	50.00	45.000	55.000	**0.313 (2%)**	**0.289 (-6%)**	**0.531 (3%)**	**0.547 (6%)**	5.751 (0%)	5.751 (0%)	2.876 (0%)	2.876 (0%)	**0.031 (-7%)**	**0.030 (-10%)**
Maximum patch farm time	50.00	45.000	55.000	**0.313 (2%)**	0.309 (1%)	0.519 (0%)	0.518 (0%)	5.751 (0%)	5.751 (0%)	2.876 (0%)	2.876 (0%)	**0.032 (-4%)**	0.033 (-2%)
Farmer conservativeness scalar	0.75	0.675	0.825	0.307 (0%)	**0.294 (-4%)**	0.516 (-0%)	**0.536 (4%)**	**6.470 (13%)**	**5.033 (-12%)**	2.876 (0%)	2.876 (0%)	0.034 (1%)	**0.032 (-5%)**
Crop reseed proportion	0.15	0.135	0.165	**0.293 (-4%)**	**0.300 (-2%)**	**0.526 (2%)**	**0.534 (3%)**	5.751 (0%)	5.751 (0%)	2.876 (0%)	2.876 (0%)	0.033 (-0%)	**0.029 (-13%)**
Farm value distance parameter	1.00	0.900	1.100	0.304 (-1%)	0.303 (-1%)	0.522 (1%)	0.525 (2%)	5.751 (0%)	5.751 (0%)	2.876 (0%)	2.876 (0%)	0.033 (-3%)	0.033 (-3%)
Farm value fertility parameter	1.00	0.900	1.100	0.307 (0%)	**0.301 (-2%)**	0.524 (1%)	**0.530 (3%)**	5.751 (0%)	5.751 (0%)	2.876 (0%)	2.876 (0%)	**0.032 (-4%)**	**0.032 (-4%)**
Farm value land cover conversion parameter	1.00	0.900	1.100	0.304 (-1%)	0.307 (0%)	**0.535 (4%)**	0.510 (-1%)	5.751 (0%)	5.751 (0%)	2.876 (0%)	2.876 (0%)	**0.031 (-9%)**	0.033 (-1%)
Firewood required per capita per year	1100.00	990.000	1210.000	**0.317 (4%)**	**0.299 (-2%)**	0.513 (-1%)	0.518 (0%)	5.751 (0%)	5.751 (0%)	**2.157 (-25%)**	2.876 (0%)	**0.032 (-4%)**	0.034 (2%)
Firewood biomass removal rate	0.10	0.090	0.110	**0.300 (-2%)**	**0.311 (2%)**	**0.534 (3%)**	0.521 (1%)	5.751 (0%)	5.751 (0%)	2.876 (0%)	**2.157 (-25%)**	**0.031 (-7%)**	0.033 (-2%)
Climax forest biomass density	300000.00	270000.000	330000.000	0.310 (1%)	0.303 (-1%)	**0.526 (2%)**	0.517 (-0%)	5.751 (0%)	5.751 (0%)	2.876 (0%)	**2.157 (-25%)**	**0.032 (-6%)**	0.033 (-2%)
Land cover colonisation base rate	0.05	0.045	0.055	**0.314 (3%)**	0.303 (-1%)	0.522 (1%)	**0.528 (2%)**	5.751 (0%)	5.751 (0%)	2.876 (0%)	2.876 (0%)	**0.032 (-4%)**	0.033 (-3%)
Land cover colonisation spread rate	0.20	0.180	0.220	0.308 (0%)	**0.287 (-6%)**	**0.533 (3%)**	**0.551 (7%)**	5.751 (0%)	5.751 (0%)	2.876 (0%)	2.876 (0%)	**0.031 (-8%)**	**0.029 (-14%)**
Mesic threshold	500.00	450.000	550.000	**0.260 (-15%)**	0.310 (1%)	**0.583 (13%)**	**0.529 (2%)**	5.751 (0%)	5.751 (0%)	2.876 (0%)	2.876 (0%)	**0.026 (-22%)**	**0.032 (-5%)**
Hydric threshold	1000.00	900.000	1100.000	0.302 (-1%)	**0.312 (2%)**	0.521 (1%)	0.523 (1%)	5.751 (0%)	5.751 (0%)	2.876 (0%)	2.876 (0%)	**0.032 (-4%)**	**0.032 (-4%)**

Table F.4: Sensitivity analysis results for the Monte Areo mire study site. The 'Value' column specifies the default value for each parameter, and the 'Value +10%' and 'Value -10%' columns give the corresponding values for parameters used in the +10% and -10% scenarios. The remaining columns give the mean values under the +10% and -10% scenarios for the *Shrubland, Mature Forest, Farmed Area Wood Area,* and *Burned Area* output variables averaged over 10 simulation replicas, along with the corresponding percentage change relative to the default scenario in parentheses. Values in **bold** show statistically significant differences with respect to the default scenario (see Section 6.1.2). Output variables indicated by * showed 0 variance in default scenario runs, so outputs that differ from the default scenario values by $\leq -10\%$ or $\geq +10\%$ are shown as significant.

Parameter	Value	Value -10%	Value +10%	Shrubland, -10%	Shrubland, +10%	Mature Forest, -10%	Mature Forest, +10%	Farmed Area, -10%*	Farmed Area, +10%*	Wood Area, -10%*	Wood Area, +10%*	Burned Area, -10%	Burned Area, +10%
Number of households per village	10.00	9.000	11.000	**0.212 (-5%)**	**0.213 (-4%)**	**0.690 (2%)**	**0.693 (2%)**	**6.044 (-10%)**	**7.387 (10%)**	2.417 (-10%)	**2.955 (10%)**	**0.020 (-5%)**	**0.019 (-6%)**
Labour availability	300.00	270.000	330.000	**0.215 (-4%)**	**0.213 (-4%)**	**0.688 (1%)**	0.680 (0%)	6.715 (0%)	6.715 (0%)	2.686 (0%)	2.686 (0%)	**0.019 (-8%)**	0.021 (-1%)
Wheat farming labour requirement	50.00	45.000	55.000	**0.206 (-7%)**	**0.217 (-2%)**	**0.696 (3%)**	0.680 (0%)	6.715 (0%)	6.715 (0%)	2.686 (0%)	2.686 (0%)	**0.019 (-6%)**	**0.020 (-5%)**
Maximum patch farm time	50.00	45.000	55.000	**0.209 (-6%)**	**0.213 (-4%)**	**0.686 (1%)**	0.685 (1%)	6.715 (0%)	6.715 (0%)	2.686 (0%)	2.686 (0%)	0.020 (-3%)	0.021 (-1%)
Farmer conservativeness scalar	0.75	0.675	0.825	**0.215 (-3%)**	0.223 (0%)	0.684 (1%)	0.682 (0%)	**7.387 (10%)**	**6.044 (-10%)**	2.686 (0%)	2.686 (0%)	0.020 (-2%)	**0.020 (-5%)**
Crop reseed proportion	0.15	0.135	0.165	**0.214 (-4%)**	**0.200 (-10%)**	0.684 (1%)	**0.704 (4%)**	6.715 (0%)	6.715 (0%)	2.686 (0%)	2.686 (0%)	0.021 (1%)	**0.019 (-10%)**
Farm value distance parameter	1.00	0.900	1.100	**0.209 (-6%)**	**0.213 (-4%)**	**0.696 (3%)**	**0.685 (1%)**	6.715 (0%)	6.715 (0%)	2.686 (0%)	2.686 (0%)	**0.020 (-5%)**	0.021 (-0%)
Farm value fertility parameter	1.00	0.900	1.100	**0.214 (-4%)**	**0.215 (-3%)**	**0.693 (2%)**	**0.687 (1%)**	6.715 (0%)	6.715 (0%)	2.686 (0%)	2.686 (0%)	**0.019 (-9%)**	**0.020 (-4%)**
Farm value land cover conversion parameter	1.00	0.900	1.100	**0.211 (-5%)**	**0.216 (-3%)**	**0.692 (2%)**	0.677 (-0%)	6.715 (0%)	6.715 (0%)	2.686 (0%)	2.686 (0%)	**0.019 (-6%)**	0.021 (1%)
Firewood required per capita per year	1100.00	990.000	1210.000	0.221 (-1%)	0.219 (-2%)	0.677 (-0%)	0.681 (0%)	6.715 (0%)	6.715 (0%)	**2.015 (-25%)**	2.686 (0%)	0.021 (0%)	0.020 (-2%)
Firewood biomass removal rate	0.10	0.090	0.110	**0.217 (-2%)**	**0.216 (-3%)**	0.681 (0%)	0.682 (0%)	6.715 (0%)	6.715 (0%)	2.686 (0%)	**2.015 (-25%)**	0.020 (-2%)	0.021 (1%)
Wood value distance parameter	1.00	0.900	1.100	0.221 (-1%)	**0.215 (-3%)**	0.674 (-1%)	0.685 (1%)	6.715 (0%)	6.715 (0%)	2.686 (0%)	2.686 (0%)	0.021 (2%)	0.021 (-1%)
Climax forest biomass density	300000.00	270000.000	330000.000	**0.228 (3%)**	**0.215 (-4%)**	**0.670 (-1%)**	0.684 (1%)	6.715 (0%)	6.715 (0%)	2.686 (0%)	**2.015 (-25%)**	0.021 (1%)	0.021 (-1%)
Land cover colonisation base rate	0.05	0.045	0.055	**0.217 (-2%)**	**0.211 (-5%)**	0.685 (1%)	**0.688 (1%)**	6.715 (0%)	6.715 (0%)	2.686 (0%)	2.686 (0%)	**0.020 (-4%)**	**0.020 (-5%)**
Land cover colonisation spread rate	0.20	0.180	0.220	**0.203 (-9%)**	**0.212 (-5%)**	**0.705 (4%)**	0.680 (0%)	6.715 (0%)	6.715 (0%)	2.686 (0%)	2.686 (0%)	**0.019 (-9%)**	0.021 (1%)
Mesic threshold	500.00	450.000	550.000	**0.196 (-12%)**	0.220 (-1%)	**0.716 (5%)**	0.684 (1%)	6.715 (0%)	6.715 (0%)	2.686 (0%)	2.686 (0%)	**0.017 (-18%)**	**0.020 (-5%)**
Hydric threshold	1000.00	900.000	1100.000	**0.217 (-3%)**	**0.217 (-2%)**	**0.690 (2%)**	**0.687 (1%)**	6.715 (0%)	6.715 (0%)	2.686 (0%)	2.686 (0%)	**0.019 (-10%)**	**0.020 (-4%)**

Table F.5: Sensitivity analysis results for the San Rafael study site. The 'Value' column specifies the default value for each parameter, and the 'Value +10%' and 'Value -10%' columns give the corresponding values for parameters used in the +10% and -10% scenarios. The remaining columns give the mean values under the +10% and -10% scenarios for the *Shrubland*, *Mature Forest*, *Farmed Area* *Wood Area*, and *Burned Area* output variables averaged over 10 simulation replicas, along with the corresponding percentage change relative to the default scenario in parentheses. Values in **bold** show statistically significant differences with respect to the default scenario (see Section 6.1.2). Output variables indicated by * showed 0 variance in default scenario runs, so outputs that differ from the default scenario values by $\leq -10\%$ or $\geq +10\%$ are shown as significant.

Parameter	Value	Value -10%	Value +10%	Shrubland, -10%	Shrubland, +10%	Mature Forest, -10%	Mature Forest, +10%	Farmed Area, -10%*	Farmed Area, +10%*	Wood Area, -10%*	Wood Area, +10%*	Burned Area, -10%	Burned Area, +10%
Number of households per village	10.00	9.000	11.000	0.072 (5%)	0.069 (0%)	**0.906 (-1%)**	0.914 (0%)	**8.711 (-10%)**	10.647 (10%)	2.010 (-10%)	**2.457 (10%)**	**0.007 (18%)**	0.006 (4%)
Labour availability	300.00	270.000	330.000	0.073 (7%)	0.066 (-3%)	0.909 (-0%)	0.917 (0%)	9.679 (0%)	9.679 (0%)	2.234 (0%)	2.234 (0%)	0.006 (8%)	**0.005 (-11%)**
Wheat farming labour requirement	50.00	45.000	55.000	**0.056 (-19%)**	0.065 (-5%)	**0.926 (1%)**	0.913 (0%)	9.679 (0%)	9.679 (0%)	2.234 (0%)	2.234 (0%)	**0.005 (-18%)**	**0.006 (10%)**
Maximum patch farm time	50.00	45.000	55.000	0.070 (1%)	**0.062 (-9%)**	0.916 (0%)	**0.921 (1%)**	9.679 (0%)	9.679 (0%)	2.234 (0%)	2.234 (0%)	0.006 (-0%)	0.006 (-4%)
Farmer conservativeness scalar	0.75	0.675	0.825	0.071 (4%)	0.072 (4%)	0.912 (-0%)	0.913 (0%)	10.424 (8%)	8.935 (-8%)	2.234 (0%)	2.234 (0%)	0.006 (-1%)	**0.005 (-11%)**
Crop reseed proportion	0.15	0.135	0.165	**0.063 (-8%)**	0.070 (2%)	0.919 (1%)	0.909 (-0%)	9.679 (0%)	9.679 (0%)	2.234 (0%)	2.234 (0%)	0.005 (-6%)	**0.007 (14%)**
Farm value distance parameter	1.00	0.900	1.100	0.073 (6%)	**0.061 (-11%)**	0.912 (-0%)	**0.924 (1%)**	9.679 (0%)	9.679 (0%)	2.234 (0%)	2.234 (0%)	0.006 (8%)	0.005 (-7%)
Farm value fertility parameter	1.00	0.900	1.100	0.065 (-5%)	**0.080 (16%)**	**0.920 (1%)**	**0.899 (-1%)**	9.679 (0%)	9.679 (0%)	2.234 (0%)	2.234 (0%)	**0.005 (-13%)**	**0.007 (19%)**
Farm value land cover conversion parameter	1.00	0.900	1.100	**0.062 (-9%)**	**0.076 (11%)**	0.918 (1%)	**0.904 (-1%)**	9.679 (0%)	9.679 (0%)	2.234 (0%)	2.234 (0%)	0.006 (-1%)	**0.006 (12%)**
Firewood required per capita per year	1100.00	990.000	1210.000	0.069 (1%)	0.068 (-2%)	**0.920 (1%)**	0.914 (0%)	9.679 (0%)	9.679 (0%)	2.234 (0%)	**2.978 (33%)**	0.005 (-6%)	**0.007 (13%)**
Firewood biomass removal rate	0.10	0.090	0.110	**0.062 (-10%)**	**0.058 (-16%)**	**0.919 (1%)**	**0.923 (1%)**	9.679 (0%)	9.679 (0%)	**2.978 (33%)**	2.234 (0%)	0.006 (-3%)	0.006 (-4%)
Wood value distance parameter	1.00	0.900	1.100	0.070 (2%)	0.072 (5%)	0.912 (-0%)	0.911 (-0%)	9.679 (0%)	9.679 (0%)	2.234 (0%)	2.234 (0%)	0.006 (4%)	0.006 (4%)
Climax forest biomass density	300000.00	270000.000	330000.000	0.069 (1%)	0.072 (5%)	0.910 (-0%)	0.910 (-0%)	9.679 (0%)	9.679 (0%)	**2.978 (33%)**	2.234 (0%)	**0.006 (10%)**	**0.007 (16%)**
Land cover colonisation base rate	0.05	0.045	0.055	0.066 (-4%)	0.072 (5%)	0.916 (0%)	**0.906 (-1%)**	9.679 (0%)	9.679 (0%)	2.234 (0%)	2.234 (0%)	**0.005 (-14%)**	**0.007 (28%)**
Land cover colonisation spread rate	0.20	0.180	0.220	0.065 (-5%)	**0.063 (-8%)**	0.915 (0%)	0.917 (0%)	9.679 (0%)	9.679 (0%)	2.234 (0%)	2.234 (0%)	0.006 (8%)	0.005 (-7%)
Mesic threshold	500.00	450.000	550.000	**0.079 (16%)**	0.066 (-4%)	**0.897 (-2%)**	0.917 (0%)	9.679 (0%)	9.679 (0%)	2.234 (0%)	2.234 (0%)	**0.007 (26%)**	0.006 (0%)
Hydric threshold	1000.00	900.000	1100.000	0.070 (3%)	**0.058 (-15%)**	0.911 (-0%)	**0.925 (1%)**	9.679 (0%)	9.679 (0%)	2.234 (0%)	2.234 (0%)	**0.006 (11%)**	**0.005 (-12%)**

Appendix G

Listing of online supplementary materials

The following is a list of the software packages that were developed to meet the goals of this thesis. The references provide links to pages in the Zenodo software repository from which the packages can be explored and downloaded.

- S1: `epd-query`, an application to help extract data from the European Pollen Database in a reproducible and open way (Lane, 2019).
- S2: `aemet-wind`, a Python library for working with wind speed and direction data from the Spanish State Meteorological Agency's weather data REST API (Lane, 2021a).
- S3: `aslib`, a Python library containing software objects used in other code for working with AgroSuccess (Lane, 2021d).
- S4: `agrosuccess-data`, code used to generate input data for the AgroSuccess simulation model (Lane, 2021b).
- S5: `demproc`, a Python package used to derive raster layers based on a Digital Elevation Model (Lane, 2021e).
- S6: Cymod application used to convert complicated state-and-transition models described in the Cypher graph query languages into Neo4j graph data stores (Lane, 2020).
- S7: `agrosuccess-graph`, a Python package that uses Cymod to represent the AgroSuccess STM (Lane, 2021c).
- S8: `agrosuccess-sim`, the AgroSuccess simulation model code (Lane, 2023).