1

Introduction

1.1 WHY FORECAST?

Humanity has long depended on natural resources and ecosystem services to survive. However, those same natural systems are increasingly becoming dependent upon humanity for their survival. We live in an era of rapid and interacting changes in the natural world: climate is changing; atmospheric CO_2, CH_4, N_2O, and O_3 are rising; land-use change is making habitats smaller and more fragmented; native species are shifting their ranges; exotic species are invading; new diseases are emerging; and agriculture, forestry, and fisheries are depleting resources unsustainably. Driving all these changes is human population, which has grown by a billion people every 12 to 13 years since the 1970s, and is projected to keep rising (United Nations 2014). That's equivalent to adding another US population to the world every 4 years. Furthermore, the economic activities of humankind have changed radically, with more and more of the world enjoying a higher, but often more resource-intensive, standard of living.

Within this context, decisions are being made every day, at levels from individuals to nations, that affect the maintenance of biodiversity and the sustainability of ecosystem services. Furthermore, the ecological questions being asked by policymakers, managers, and everyday citizens *are fundamentally about the future*. They want to know what's going to happen in response to decisions versus what's going to happen if they do nothing. Those decisions are being made with or without the input of ecologists, and far too often these questions are not being answered with the best available science and data. Ecologists are being asked to respond to unprecedented environmental challenges. So how do we, as ecologists, provide the best available scientific predictions of what will happen in the future? In a nutshell, how do we make ecological forecasts?

The foundational paper "Ecological Forecasting: An Emerging Imperative" (Clark et al. 2001) defines ecological forecasting as

> the process of predicting the state of ecosystems, ecosystem services, and natural capital, with fully specified uncertainties, and is contingent on explicit scenarios for climate, land use, human population, technologies, and economic activity.

However, the knowledge and tools for making ecological forecasts lies outside the training of most ecologists. This information is scattered across the literature, with different ecological subdisciplines often unaware of progress being made in others,

and with many forecasting techniques being coopted from other disciplines, such as meteorology and statistics. This book aims to synthesize this literature, both distilling key concepts and highlighting case studies from different ecological disciplines.

The importance of learning to make ecological forecasts goes beyond decision making. The ability to make quantitative, testable predictions, and then confront them with data, is at the heart of the scientific method and advances our basic scientific understanding about ecological processes. In general, a successful forecast into the future, to a new location, or under novel conditions provides much stronger support for a hypothesis than its success in explaining the original data used to develop it. Indeed, it is possible for alternative models to produce similar fits to data, but make very different predictions. By making frequent forecasts, checking those forecasts against data, and then updating our projections, we have the potential to accelerate the pace of our science. Indeed, a key attribute of science is the emphasis on updating our ideas as new information becomes available. Forecasting puts this idea front and center, updating projections routinely as new information becomes available. Furthermore, reframing an ecological question into a forecasting problem can often allow the problem to be seen in a new light or highlight deficiencies in existing theories and models.

Despite the importance of forecasting to making our science more relevant and robust, there are ecologists who will rightfully point out that in ecology our understanding of many processes is often coarse, data are noisy, and dynamics can be idiosyncratic, varying from system to system or site to site in ways that defy our current understanding. These are all valid points, and because of that, much of this book will focus on quantifying, partitioning, and propagating uncertainties and sources of variability. These challenges also highlight the need to update projections in light of new data—for example, adjusting projections after low-probability events, such as disturbances, occur. Indeed, making ecology a more predictive science will rely heavily on understanding probability and uncertainty—ecological forecasts need to be probabilistic to capture these uncertainties.

As ecologists we appreciate the complexities and idiosyncrasies of the systems we study, and at times the multitude of possible interactions and outcomes can seem overwhelming. When faced with questions about how ecological systems will respond to change, it is far too easy to answer these questions with "it depends." However, while the idea that everything in an ecosystem is connected to everything else makes for a profound and deep mythos, the reality is that not all things are connected equally. The responses of complex systems are indeed capable of surprises, but more often their responses are driven by a small subset of interactions and processes. Fundamentally, I believe that ecology is more than just a collection of case studies and just-so stories. The extent to which ecological systems will prove to be forecastable remains a critical, but ultimately empirical, question. One that at the moment is unanswered, but profoundly important to the future of our science and its relevance to society. How we might go about answering that question is the subject of this book.

Figure 1.1 illustrates a general conceptual workflow for how we might develop ecological forecasts and maps the chapters of this book, and how they relate to one another, onto that workflow. In brief, making forecasts ultimately depends on models (center), but models are dependent on data in many ways (chapters 3, 4). Data are used as real-world drivers, constrain parameters (chapters 5, 6, 8, 9), and provide observations of system states to both initialize (chapters 13, 14) and validate (chap-

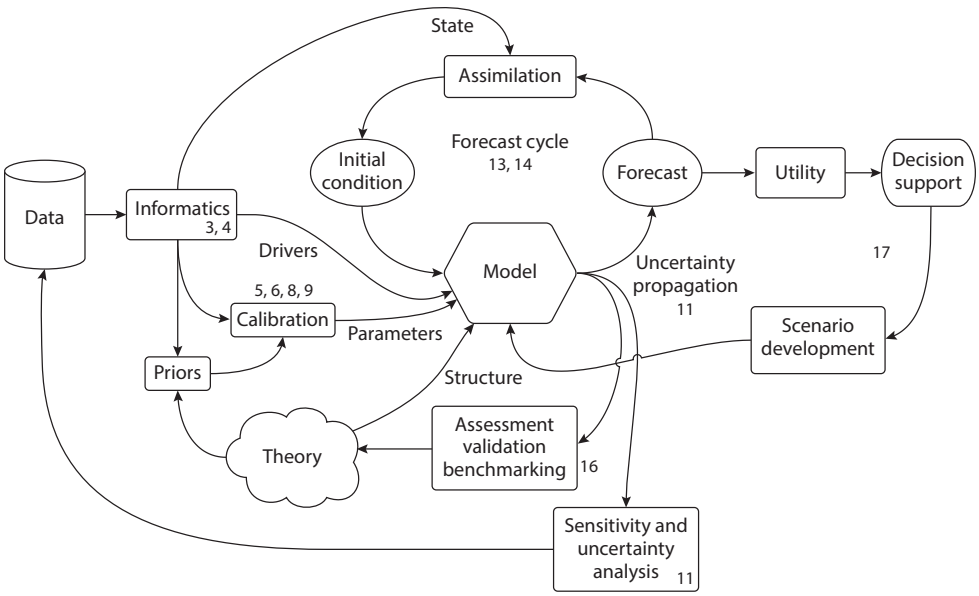

FIGURE 1.1. Ecological Forecasting conceptual workflow. Relevant chapters identified by number.

ter 16) models. Uncertainties enter at all stages, and need to be partitioned into different sources (chapter 6) and propagated into forecasts (chapter 11). Improving forecasts relies critically upon a number of feedback loops: sensitivity and uncertainty analyses (chapter 11) identify key parameters and processes driving uncertainties; model assessment, validation, and benchmarking (chapter 16) identify model structural errors; and data assimilation, through the forecast cycle (chapters 13, 14), addresses model process error. While some forecasts are done to advance basic science, many are done to provide explicit decision support (chapter 17), evaluating specific scenarios and allowing decision makers to assess trade-offs among alternatives.

Beyond this conceptual workflow, there are a number of major themes cutting across the book, which are summarized in the following sections.

1.2 THE INFORMATICS CHALLENGE IN FORECASTING

Ecological forecasts are generally focused on making projections for specific places and systems in the real world. This focus on the real world means that much of the day-to-day effort in producing ecological forecasts may end up being devoted to data management and processing. Since we generally want to update forecasts as new data become available, this puts a premium on the documentation and automation of workflows. Furthermore, since many forecasts are developed explicitly for policy and management goals, forecasts must be transparent, repeatable, and adhere to best practices. Ultimately, up-front attention to the informatics challenges of forecasting not only makes for better science, but, by reducing error and increasing automation, it also can allow you to do more science. Unfortunately, these informatics challenges are outside the core training of many ecologists. While the field of informatics is rapidly evolving, with new tools being developed daily, many of the core concepts are

persistent. Thus the informatics of forecasting is discussed up front in this book (chapters 3 and 4) as a foundation that will be leveraged when performing syntheses and making forecasts.

1.3 THE MODEL-DATA LOOP

In response to the multitude of changes occurring in the natural world, ecologists have invested considerable effort in monitoring. Ecological monitoring occurs across a wide range of scales and systems, from individual organisms being tagged to satellites measuring the globe, and has become a fundamental mission of many federal agencies and nongovernmental organizations (NGOs) around the world. However, *the core aim of monitoring is to detect change after it has happened, rather than to anticipate such change*. As such, monitoring is a primarily data-driven exercise, and can actually occur in a relatively model-free manner. By contrast, forecasting, by its nature, requires that we embrace models, as they are our only way to project our current understanding into the future, to new locations, or into new conditions.

Because they are the backbone of forecasting, I will be discussing models throughout this book. If the thought of that makes you anxious, I'll ask you to stick with me because this is most definitely not a book about modeling. Furthermore, models don't have to be complex and impenetrable. Even basic statistical models, such as ANOVAs and regressions, are models, the former predicting that different groups have different means and the latter predicting a straight line. For almost everything I discuss in this book, *the underlying principles are the same regardless of how simple or complex the model*. Indeed, I will often rely on very simple models, such as fitting a mean, to illustrate concepts that are applicable to all models regardless of complexity.

Across this spectrum of model complexity, the common denominator is that models remain a quantitative distillation and formalization of our hypotheses about how a system works. Because they embody our current working hypotheses, models are relevant to all ecologists, not just modelers and theoreticians. Unfortunately, in ecology there has often been a disconnect between the empirical and modeling portions of the community. In truth, all ecologists are quantitative ecologists, because we all use numbers to tell things apart. Similarly, any ecologist who has ever done statistics is a modeler. Or better yet, we should recognize than none of us are "modelers" any more than a geneticist is a "PCRer"—models are tools we all use, not a separate discipline, and we're all just ecologists.

While models are a critical part of any forecast, so are data. Unlike many theoretical modeling exercises, forecasts focus on projecting real systems, and working in real systems requires data about those systems. Therefore, *any approach to forecasting, from the most simple to the most complex, requires the combination of models and data*. In combining models and data, forecasting places a premium on quantifying the uncertainties in both. The estimation of uncertainty in a forecast is critical to decision support, as underestimating uncertainty can lead to overconfident decisions, while overestimating uncertainty can lead to excessive, and expensive, levels of caution. Once uncertainties are quantified, we will sometimes find that effective decision making (or hypothesis testing) requires that these uncertainties be reduced. The next job of the forecaster is thus to identify the sources of new data that will have the greatest or most cost-effective impact on reducing model uncertainty.

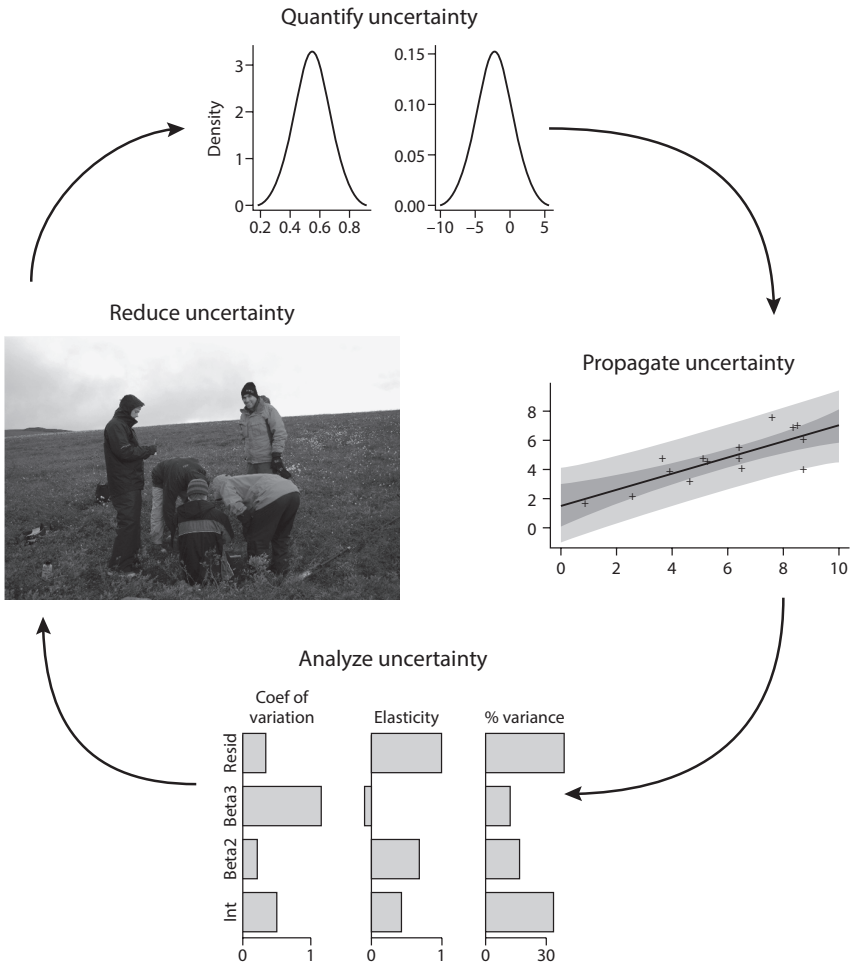

FIGURE 1.2. Model-data loop.

Throughout this book I will talk about the idea of the model-data loop (figure 1.2). This refers to the idea of iteratively using data to constrain models and then using models to determine what new data is most needed. The model-data loop applies regardless of the complexity of the model, and indeed should play a central role in any decision to increase model complexity. As an example, consider a simple linear model, $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3$, with one dependent variable, Y, and two independent variables, X_2 and X_3. The model-data loop might begin by *quantifying the uncertainty* in the intercept (β_1), the two slopes (β_2 and β_3), and the residual error using regression. Alternatively, it might require more sophisticated techniques to characterize the uncertainty (chapter 6) if there are additional complexities such as non-Normal error, non-constant variance, missing data, or multiple sources of error (for example, observation versus process, errors in the X's). The next step would be to *propagate the uncertainty* in the model into the forecast (chapter 11). For a regression this may be something as simple as constructing confidence and predictive intervals. Next we would *analyze the uncertainty* to discern which processes are controlling our predictive uncertainty.

Figure 1.2 illustrates a case where the intercept and residuals dominate model uncertainty, contributing 42% and 33% respectively to the predictive variance, while the first and second slope terms contribute 15% and 10%. Finally, based on the uncertainty analysis, we could target additional field measurements. In this case we might choose to simply increase the overall sample size, which will better constrain all the regression parameters. Furthermore, since X_2 explains more of the variance than X_3, sampling may be stratified to better capture variability in X_2.

The idea of using models to inform field research has been discussed in many contexts (Walker et al. 2014) but rarely occurs in practice, and when it does it is often in a qualitative way, such as identifying processes that we don't understand (Medlyn et al. 2015). Approaches to uncertainty analysis discussed here aim to make this a more quantitative endeavor (LeBauer et al. 2013; Dietze et al. 2014). While model-driven data collection is no substitute for hypothesis testing, the two very often go hand in hand. We almost always have open research questions about the processes identified as driving forecast uncertainty. More pragmatically, we are likely to have unanswered hypotheses in many areas of our research, but the model-data loop allows us to focus first on those that will give us the most return on investment when allocating scarce resources.

1.4 WHY BAYES?

Throughout this book many of the concepts and tools presented will be developed from a Bayesian statistical perspective. At this stage I want to quickly discuss the rationale for this choice and defer a more detailed description of Bayes and its forecasting applications to later chapters (chapters 5, 6, 8, 9, 13, and 14). This book assumes no prior experience with Bayesian approaches—in particular, chapter 5 aims to provide a solid common ground for readers who may not have had a prior exposure to Bayesian concepts and methods. For those with previous experience with Bayes, chapter 5 will serve as a refresher or could be skipped. That said, this is a book on forecasting, not Bayesian statistics, and those with no prior exposure to Bayes will probably find chapter 5 to be accelerated and may want to consult an introductory textbook for additional background (Clark 2007; Hobbs and Hooten 2015).

While it is definitely possible to do ecological forecasting from a classical (frequentist) statistical perspective, the Bayesian approach has a number of advantages that are even more valuable in forecasting than they are in standard data analysis (Ellison 2004; Clark 2005). First, it allows us to *treat all the terms in a forecast as probability distributions*. Treating quantities we are interested in as probabilities makes it easier to quantify uncertainties, to partition these uncertainties into different sources of error and process variability (chapter 6), and to propagate them into forecasts (chapter 11) and decision support (chapter 17). Second, as discussed earlier, the *ability to update predictions as new data becomes available* is a critical aspect of forecasting. In forecasting, we leverage the inherently iterative nature of Bayes' theorem (equation 1.1) as a means of updating forecasts. Specifically, Bayes' theorem states that the updated forecast (also known as the posterior probability) for any quantity, θ, is proportional to the likelihood of the most recently observed data, y, times our previous forecast (also known as the prior probability):

$$\underbrace{P(\theta|y)}_{posterior} \propto \underbrace{P(y|\theta)}_{likelihood} \underbrace{P(\theta)}_{prior} \tag{1.1}$$

As new data becomes available, the posterior from the previous forecast becomes the prior for the next. Thus, in forecasting, the priors embody the information provided by previous observations and forecasts. Finally, from a pragmatic perspective, Bayesian numerical methods tend to be flexible and robust, allowing us to deal with the *complexity of real-world data* and to build, fit, and forecast with relatively complex models.

1.5 MODELS AS SCAFFOLDS

Traditional approaches to modeling have focused primarily on forward modeling—taking a set of observed inputs and running them through a model to generate a set of outputs (figure 1.3A). The other approach to modeling commonly employed is inverse modeling—starting from a set of observations that correspond to model outputs and trying to infer the most likely inputs (figure 1.3B). In this book I will rely on both of these approaches, but I will also present a third modeling paradigm, which I call "models as scaffolds" (figure 1.3C), which refers to using models and data together to constrain estimates of different ecosystem variables (Dietze et al. 2013). One of the challenges of combining different data sources (chapter 9) is that frequently observations may come from very different spatial and temporal scales, or they may capture different but related processes, and thus cannot be directly compared to one another. In most cases ad hoc approaches, such as interpolating data to a common scale, are not the right answer to this problem, as they throw out uncertainties, misrepresent the actual number of observations, and introduce additional

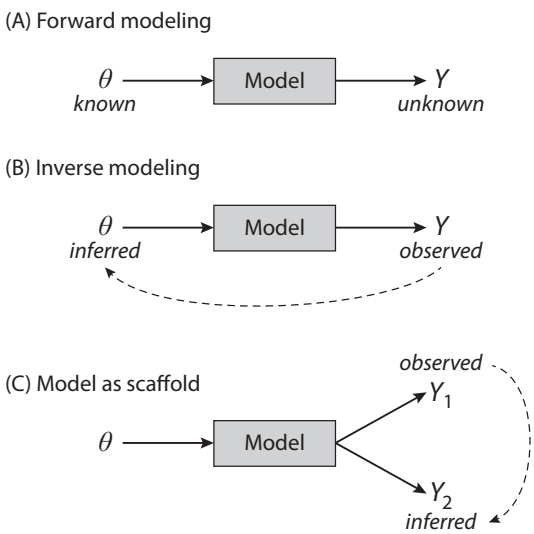

FIGURE 1.3. Modeling approaches. (A) In forward modeling the model inputs and parameters are treated as known and the model is used to predict some unknown state. (B) In inverse modeling the current state is observed (with uncertainty) and the inputs or parameters are inferred. (C) In the model as scaffold approach multiple outputs may be observed and the best estimate of the current state is inferred, both for observed (Y_1) and unobserved (Y_2) outputs, based on the covariance among model states.

assumptions that may not be consistent with our understanding of the process (for example, linearity, homogeneity, and scale independence). By contrast, it is often the case that the models we use to forecast may represent multiple spatial scales, temporal scales, and interconnected processes. Thus, while different data sets may not be directly comparable to one another, they may all be comparable to the model. In this context combining data and models is a fundamentally data-driven exercise for allowing data sets to talk to each other across different scales and processes. The model is the scaffold on which we hang different observations. Furthermore, this scaffold will leverage the understanding and current working hypotheses that are embedded in the models to describe how we believe different processes and scales are related to one another. In this sense the model serves as a covariance matrix, providing structure to the correlations between different observations. This concept of models as scaffolds will be central not only to how we approach data synthesis but also to how we update forecasts as new observations become available (chapters 13 and 14).

1.6 CASE STUDIES AND DECISION SUPPORT

Throughout this book chapters illustrating case studies from different disciplines are interspersed among chapters illustrating tools and concepts. Specifically, the case study chapters look at the status of ecological forecasting in biodiversity and endangered species (chapter 7), natural resource management (chapter 10), the carbon cycle (chapter 12), and disease ecology (chapter 15). No one discipline has solved the ecological forecasting problem, with each being stronger in some areas and weaker in others. While there are differences in the forecasting problem among ecological disciplines, there are also many overarching similarities and a lot of shared experience that is too often isolated in academic silos. These chapters aim to build bridges between these silos.

Finally, in chapter 17, I discuss how ecological forecasts can better inform decision making, touching on approaches to scenario development and quantitative decision support models. While figure 1.1 depicts decision support as an end-goal of ecological forecasting, as indeed it is the motivating factor in many real-world forecasts, in reality there are many feedback loops between forecasting and decision support, with decisions not just providing the scenarios and alternative choices to be evaluated but also defining the context, scope, and desired output variables.

Covering the details of decision support requires that we first develop a strong foundational understanding of how ecological forecasts are generated, which is why decision support is at the end of this book. However, to be able to better frame and contextualize ecological forecasts, it is also important to introduce many of the key concepts of decision support up front. First and foremost of these is that decisions are about what will happen in the future, in response to our choices, rather than about what's already happened in the past—decisions depend upon forecasts. However, decisions also depend upon the presence of choices, which are frequently evaluated as alternatives or scenarios that describe different decisions or storylines about how the external world unfolds. Because of this forecasts will be broadly separated into those that provide *predictions*, probabilistic forecasts based on current trends and conditions, versus those that provide *projections*, probabilistic forecasts driven by explicit scenarios.

Another key concept in decision support is to remember that while science provides facts and knowledge, decisions are ultimately about *values*. Furthermore, it is

the values of the community as a whole that are relevant, not just those of the scientist. While it is important to acknowledge that we all have our own values and biases, and that these inevitably affect our research, a key part of formal decision support relies on trying to compartmentalize knowledge and values. We want to objectively determine what the trade-offs are in any particular decision (for example, between different stakeholders or competing goals), so that the relevant decision maker, be that an individual, a committee, or an electorate, can decide how to balance the competing, subjective values in any trade-off. Doing so requires that ecological forecasts have clearly defined *objectives* that summarize what matters to stakeholders. Objectives should also indicate the desired direction of change (for example, increase endangered species population) for individual objectives. These objectives then need to be translated into *performance measures* that quantify the objectives in units appropriate for that objective (number of individuals, habitat area, and so on). It is the future of these performance measures (sometimes dubbed *consequences*), subject to uncertainties and alternatives, that we aim to forecast. Decision support also relies on a broad set of useful *alternatives*—any decision is only as good as the set of choices considered—while also acknowledging that a set of alternatives that is too large will overload both the forecaster and the decision maker.

Fundamentally, the goal of decision support isn't to make "optimal" decisions *on behalf of* the decision maker, but to determine the *trade-offs* among competing performance measures. The only alternatives that should be eliminated by the analyst are those that are strictly dominated (that is, lose-lose on all fronts). Beyond that, determining how stakeholders value the competing objectives in any trade-off may be outside the decision support scope, or may occur through discussion, weighting of different values, or more formal quantification of values using Utility functions (quantitative relationships between a performance measure and value).

When presenting the results of any probabilistic forecast it is important to be aware that humans do not have an innate sense of probability and rely instead on a set of *heuristics* that can be subject to a laundry list of well-known cognitive biases (Kahneman 2013). Because of this, perceptions can be sensitive to how uncertainties are presented. That said, another key concept of decision support is understanding how the *uncertainties in forecasts interact with the inherent risk tolerance or risk aversion of stakeholders*. As mentioned earlier, overestimation of uncertainty can lead to costly levels of caution, while underestimation of uncertainty can lead to risky, overconfident decision making. Even if stakeholders could perceive uncertainties perfectly, these uncertainties would have large impacts on decision making because utility functions are generally nonlinear—we don't perceive losses the same as gains. In general, utility declines as uncertainty increases, which is often dealt with by looking for strategies that are precautionary, robust to uncertainties, or adaptive. Indeed, ecological forecasting is at the core of many adaptive management approaches (Walters 1986).

To conclude, in this chapter I introduced the basic concepts of ecological forecasting and their relevance to both decision making and basic science. Forecasting was shown to depend closely on the integration of models and data, and thus requires a solid foundation in informatics and statistics. In building this foundation, I argued for taking a Bayesian approach to forecasting, as it allows forecasts to be easily updated in the light of new data. I introduced the idea of the "model-data loop," where model analyses are used to identify and prioritize measurements, and the idea of

"models as scaffolds" for allowing data of different types and on different scales to be mutually informative. I also introduced some of basic concepts of decision support and how forecasting fits into that framework. Finally, I argued the importance of accounting for uncertainty and that ecological forecasts should be made probabilistically. In the following chapter I will follow up on uncertainty and probability by walking through an example of how different sources of uncertainty affect a simple forecast, using this as a springboard to look more deeply at the problem of predictability in ecology.

1.7 KEY CONCEPTS

1. Ecological forecasting is critically important to improving our science, making it more relevant to society, and quantitatively supporting decision making.
2. Forecasting is data-intensive, and there are a number of informatics challenges to making forecasts efficient, repeatable, transparent, and defensible.
3. Forecasting requires combining data with models. Doing so requires the quantification of the uncertainties in both and the propagation of these uncertainties into forecasts.
4. Forecasting concepts apply to all models, from simple statistical models to complex computer simulations.
5. Data inform models, and models can be used to more precisely target future data as part of the model-data loop.
6. Bayesian approaches allow us to represent uncertainties probabilistically and to update forecasts easily as new data becomes available.
7. Models can serve as a scaffold for combining information on different processes, or across different scales in space and time.
8. Ecological forecasting is a key component of decision support, as all decisions are fundamentally about the future and are sensitive to forecast uncertainties. However, decisions are ultimately about human values in how we balance trade-offs among competing objectives.

1.8 HANDS-ON ACTIVITIES

https://github.com/EcoForecast/EF_Activities/blob/master/Exercise_01_RPrimer.Rmd

- Primer on the R language