Ecological Applications, 21(5), 2011, pp. 1537–1545

Introducing data–model assimilation to students of ecology

N. Thompson Hobbs[1,3] and Kiona Ogle[2,4]

[1]*Natural Resource Ecology Laboratory and Graduate Degree Program in Ecology, Colorado State University, Fort Collins, Colorado 80523 USA*
[2]*Department of Botany and Department of Statistics, University of Wyoming, Laramie, Wyoming 82071 USA*

Abstract. Quantitative training for students of ecology has traditionally emphasized two sets of topics: mathematical modeling and statistical analysis. Until recently, these topics were taught separately, modeling courses emphasizing mathematical techniques for symbolic analysis and statistics courses emphasizing procedures for analyzing data. We advocate the merger of these traditions in ecological education by outlining a curriculum for an introductory course in data–model assimilation. This course replaces the procedural emphasis of traditional introductory material in statistics with an emphasis on principles needed to develop hierarchical models of ecological systems, fusing models of data with models of ecological processes. We sketch nine elements of such a course: (1) models as routes to insight, (2) uncertainty, (3) basic probability theory, (4) hierarchical models, (5) data simulation, (6) likelihood and Bayes, (7) computational methods, (8) research design, and (9) problem solving. The outcome of teaching these combined elements can be the fundamental understanding and quantitative confidence needed by students to create revealing analyses for a broad array of research problems.

Key words: *data assimilation; data–model assimilation; ecological modeling; ecological theory; ecology curriculum; hierarchical modeling; mathematical ecology; pedagogy; statistical ecology.*

Introduction

Throughout the history of ecology, models have played a fundamentally important role in the development of insight, providing statements about nature with an economy that only mathematics can provide. Despite the longstanding importance of models to ecological understanding, there remains a cultural divide between modelers and empiricists. This divide is laid plain at virtually any national meeting of ecologists, where there are special sessions on modeling, theory, or quantitative ecology that are often set apart from the rest of the agenda. Recognition of the need to bridge this divide has motivated calls for a new emphasis on the application of quantitative methods in all areas of ecology (Green et al. 2005, Hastings et al. 2005, Jones et al. 2006) and in training students to use such methods (National Research Council 2003, Robeva and Laubenbacher 2009).

These cultural differences are easily seen in textbooks on quantitative methods in ecology. There are basically two genres that, until recently, remained astonishingly dissimilar. In the first genre, texts on modeling (e.g., Edelstein-Keshet 1988, Mangel 2006, Otto and Day 2007) teach readers how to develop equations symbolizing the operation of ecological processes and how to gain mathematical insight from them. In these texts, data are unapologetically put in the back seat while symbolic analyses, using well-established methods like analysis of local stability, come to the fore. Apologies are not needed for the absence of data because the overriding objective of these methods is to provide general insight, and as soon as data are applied to an equation, it immediately becomes specific to the particular system represented by the data (Levins 1966). In our view, the material taught in these texts is valuable to students of ecology because it requires one to think long and hard about the way ecological processes work. Symbolizing a process mathematically requires clarity of thought that surpasses the thinking required to describe the process in words. The weakness here is that important truths ultimately depend on observations. Good models can guide these, but cannot replace them.

The second genre includes texts on "statistics for ecologists" (e.g., Hairston 1989, McGarigal et al. 2000, Scheiner and Gurevitch 2001, Gotelli and Ellison 2004), the purpose of which is to teach methods for gaining insight from data, particularly data produced by designed experiments. In this genre, models are eclipsed by observations and procedures for describing relationships among them. Mathematical or process models take

Manuscript received 28 August 2009; revised 25 February 2010; accepted 11 march 2010; final version received 7 April 2010. Corresponding Editor: D. S. Schimel. For reprints of this Invited Feature, see footnote 1, p. 1427.
[3] E-mail: nthobbs@nrel.colostate.edu
[4] Present address: School of Life Sciences, Arizona State University, Tempe, Arizona 85287 USA.

the back seat, becoming almost an afterthought if they are considered at all. The arrival of a vast set of software for statistical analysis has made it possible for ecologists to analyze their data by knowing a set of instructions or a sequence of mouse clicks without any deep understanding of the statistical models and assumptions underlying the analyses they perform. In our view, the value of this material is the necessity to think about observations of nature. The weakness is that the models used for analysis often fail to thoughtfully describe the operation of processes and may be opaque to the ecologists using them.

The tables of contents of these two types of texts have almost nothing in common. However, despite their differences, the theoretical and empirical traditions in ecology have at least one feature strongly shared between them: both depend in a way that is truly fundamental on simplification. The theoretical tradition depends on models of low dimension, those with only a few state variables, because they are the only ones amendable to analytical methods and purely symbolic results. Nonlinear models with more than a few state variables quickly become intractable, requiring numerical simulation as the only feasible route to insight. The empirical tradition requires simplicity for a different reason. The unambiguous insight that results from the manipulative experiment, which has remained the gold standard of inquiry for decades, depends on observing responses of ecological systems to a few, carefully chosen influences. These investigations are feasible only at fine scales of time and space. Only a few interactions can be studied, and heterogeneity in space and time is usually minimized by clever sampling schemes or experimental designs.

In both traditions, empirical and theoretical, the clarity of insight that comes from a simple system is offset by loss of fidelity to the natural world. Both traditions are poorly suited to understanding phenomena where responses are determined by composites of forces and multiple interactions, where heterogeneity is the object of study rather than a nuisance to be designed away, where system behavior is sensitive to initial conditions, and where forecasting is an important objective of our science. In short, agronomic style experiments and analytically tractable models are often not up to the task of understanding complex systems.

Data–model assimilation offers an alternative to the traditional approaches to inquiry in ecology, an alternative that is particularly valuable for dealing with complex phenomena. To define what we mean by data–model assimilation, we start with the premise that phenomena of interest can be described by reference to some set of *states* that vary over space and/or time. For example, the state of a population can be described by the number of individuals it contains and their sex and ages; the state of a community can be described by the abundances of species in a food web and the strength of interactions among them; the state of the carbon cycle can be described by the size of different soil, microbial, and vegetation pools, and so on. In this paper we will use the term data–model assimilation in its broadest sense to describe a suite of methods for combining models of ecological processes with data to estimate states of interest, and in particular to understand how and why the states change over time or space. Although data–model assimilation methods are emerging in the ecological literature (see reviews of Ogle and Barber 2008, Cressie et al. 2009, Wang et al. 2009) training in these methods appears to be lagging behind their application.

The key features of data–model assimilation resemble the scientific method as it is taught at the most elementary level. We begin with a model of a process: population growth, community dynamics, or ecosystem function. This model represents our understanding of the way that process works, aggregating insight from first principles, individual experiments, samples, and physical laws. It represents a precisely stated hypothesis about the processes that control the behavior of states of interest. We use the model to predict the behavior of some state of interest. We compare the prediction with observations and use the comparison to update and improve the model and its future estimates. We may evaluate the model against competing, alternative abstractions of the process of interest. Data–model assimilation is the procedure of informing the process models with data to obtain estimates of (1) states, (2) model parameters, and (3) attendant uncertainties. Properly executed, these methods allow forecasting: predictions of the future values of states accompanied by confidence envelopes.

The purpose of this paper is to offer ideas for training ecologists to use quantitative methods in a new, more informative way by providing them with a foundation for applying data–model assimilation techniques to ecological problems. Our intent is to reinforce important trends in quantitative education for ecologists, most notably the emergence of superb texts that bring together the traditions described above (e.g., Hilborn and Mangel 1997, Clark 2007, McCarthy 2007, Bolker 2008, Royle and Dorazio 2008, King and Gimenez 2009, Link and Barker 2010). To do so, we will outline what we think are the essential features of a curriculum that captures the best of the empirical and theoretical traditions in ecology: the mathematical description of processes that has characterized theory and the emphasis on data by the empiricist. To put some bounds on this effort, we limit ourselves to material that we think could be taught in a semester to graduate students with unexceptional quantitative backgrounds, i.e., to motivated students who have completed at least two semesters in calculus and one in statistics. The most important goal of such a course is to provide a foundation supporting self-teaching through students' careers. A critical part of this goal is what we call quantitative confidence: the belief that given sufficient

time and access to resources, including colleagues in mathematics and statistics, students can compose the appropriate analytical approach to a great diversity of problems in ecological research.

We do not pretend that ours is the only way to organize this material. We wish to be provocative rather than prescriptive. Moreover, in addition to outlining topics, we will occasionally venture briefly into pedagogy by describing what has worked for us in teaching a particular topic. The purpose of this paper, then, is to foster discussion of how we should be training ecologists in the quantitative tools they will need to solve the pressing research problems in coming decades.

Why is New Training Needed Now?

There are at least three reasons why a new approach to quantitative training for ecologists is needed: to expand the diversity of questions that ecologists seek to answer, to exploit a fundamental change in the availability of observations, and to foster ecological forecasting in a world where forecasts are badly needed to solve environmental problems. In our experience, many graduate students in ecology emerge from obligatory classes in traditional statistical methods knowing a host of procedures without gaining much understanding about how models and data work together to provide insight. In the absence of this understanding, research design becomes an exercise resembling taxonomy—dichotomously matching the correct analysis procedure to the data that would be or (worse) have been collected (see, for example, the inside cover of Sokal and Rohlf [1995]). In our view, this process has a confining effect on the questions that ecologists are willing to ask (Hobbs et al. 2006); if the only tool that you have in your locker is analysis of variance, then all the world looks like a plot. It is the nature of science that questions can go unanswered simply because researchers believe they lack the tools to attack them. We suggest that providing students with a highly general and flexible approach to gaining insight could meaningfully expand the type of questions that ecologists are willing to ask and are able to address.

This expansion is one of the motivations for the construction of unprecedented observing systems, for example, the emerging National Ecological Observatory Network (NEON). These networks promise to fundamentally change the relationships among models, observations, experiments and insight in ecology, moving the field from one that tends to be data-poor to one that is data rich (Luo et al. 2011). In this new environment, students will need the ability to nimbly combine diverse sets of observations, results of experiments, and historical data in a statistically coherent framework. This ability will enable them to choose analyses based on the problem and resources at hand rather than choosing the problem to fit a familiar type of analysis (Little 2006, Ogle 2009). This sort of flexibility requires a first principles understanding of how we gain insight from models and data.

The final motivation for thinking about how we train students in quantitative methods is the emerging need to offer forecasts that inform decisions made by environmental managers and policy makers. Ecological models have been sharply criticized for their inaccurate predictions (Pilkey and Pilkey-Jarvis 2007), but it can be argued that these failures result not because predictions fail to match experience, but rather because modelers have neglected to estimate uncertainty in a reliable way. The ecologists of the future need to know how to go about developing forecasts, predictions made honest by a rigorous assessment of uncertainty.

Elements of a Curriculum

In the following sections, we outline what we believe are essential parts of an introductory, one-semester course in data–model assimilation for ecologists. We do not consider the list of topics to be exhaustive nor do we try to provide deep coverage of each topic. We note, however, that the topics can be treated in greater depth in a two-semester sequence. We speak from our experience in teaching this material (Appendix), admitting that our experience is neither comprehensive nor universal. We will have succeeded in our purpose here if these topics stimulate useful conversations among those who train ecologists in methods of analysis of models and data.

Element 0: Preparatory training and prerequisites.—There are fundamental differences of opinion among thoughtful educators about how to best communicate sequences of knowledge, and, hence, the need to establish a firm series of prerequisites for courses. These contrasting views are evident in the different systems of graduate education in Europe and the United States: in many European programs, students are expected to learn what they need without a specific course sequence; in the United States, coursework tends to be prescribed. We will not try to resolve these views in this paper. We believe that a solid quantitative foundation including two or three semesters of calculus as well as coursework in linear algebra, differential equations, and mathematical statistics is a tremendous asset for any student of ecology. Understanding the concepts we outline below requires an understanding of derivatives and integrals, some familiarity with differential equations and matrix methods. That said, we also believe that what you are able to learn is often more important than what you know, and that oftentimes students (including the authors) are most efficiently guided in their learning by the problems that they confront. We have seen students with minimal mathematical training spread the books out in their rooms and learn the math they need to master the material that we outline below. So, while valuing undergraduate mathematical training, we also believe in the value of self-teaching by motivated students and,

as a result, do not advocate an inflexible set of prerequisites.

Element 1: Model building as a route to insight.—The historical emphasis on procedural training in statistics has allowed ecological researchers to conduct analyses of data without carefully considering the role of models. We think it is critical for students to understand that virtually all insight in ecology depends on the analysis of some type of model and that the choice of models is perhaps the most important choice we make when we design research. The reason that models are central is simple: we gain insight by determining how well or how poorly our statements about nature are matched by observations. This determination requires predictions, and predictions require models. It is critical that students understand that all models are abstractions of nature and that one of the most creative parts of science is developing them. Because abstractions, by definition, do not include all of the detail of nature, the researcher must decide what to keep and what to throw away. Data ultimately arbitrate the wisdom of that decision.

It follows that a key goal of a course on data–model assimilation (which may also be referred to as data–model integration, or model-data fusion) should be to give students a strong "modeling intuition." That is, as a starting point for research design and analysis, students should be able to develop conceptual models of their ecological systems. They should be able to translate these conceptual models into mathematical equations, justify the functional forms used to specify the mathematical model, and provide ecologically meaningful interpretations of the model's parameters.

We have found that this intuition can be usefully developed by a series of examples from a range of ecological fields, starting with static models and moving to dynamic ones. It can be exceedingly instructive to take students through the algebraic process of deriving the classical, deterministic models in ecology, for example Holling's disc equation, the Lotka-Voltera model of predators and prey, the susceptible–infected–recovered (SIR) model of disease transmission, the Levins meta-population model, Tillman's model of resource competition, and so on. As soon as dynamic models become involved, the basics of rates and rate equations should be taught, particularly the relationship between rates in continuous and discrete time and the relationship between rates and probabilities.

All of these examples can be used to emphasize that, by design, models do not represent all of the influences that shape an ecological process, but rather include only those influences that the researcher wishes to investigate and, by choice, includes in the model. This means that researchers must be able to estimate how those "other" influences, those not included in the model, create uncertainty in the model output or predictions. This prepares the students for a treatment of uncertainty and stochasticity.

Element 2: Uncertainty.—Why are models always "wrong?" What creates uncertainty in statements about the operation of ecological systems? How can we quantify uncertainty? Traditional, procedural training has not prepared students to deal with the multiple sources of uncertainty that arise in ecological analysis (Brewer and Gross 2003). We suggest that as soon as students have some ability to construct deterministic models and think about them as abstractions of nature, they then be encouraged to think about the relationship between models and data. In particular, they should think about why is it that the predictions of models of ecological processes never perfectly match observations of those processes. The first reason is that models, because they are abstractions of processes, will not represent all of the influences that determine the behavior of a state of interest, but that observations *will* include all of those influences. Even though we have chosen to omit influences from our model, we can include their effects by treating them stochastically, by acknowledging they will create uncertainty in model predictions that we must quantify. This estimate of uncertainty is what we call process variance.

A second reason that models fail to perfectly match observations is the fact that we usually observe things imperfectly. There are very few quantities of interest in ecology that can be directly and precisely observed. We cannot observe directly the process of photosynthesis, but instead take measurements of changes in concentrations of CO_2 and water inside a cuvette to infer leaf-level fluxes. We cannot know the biomass of plants over large landscapes, but instead take measurements of reflectance of plant canopies from instruments on satellites and relate these to biomass. We almost never count all of the animals in a population but instead count a subset and somehow account for those that remained uncounted. A related problem in taking observations is that we never observe everything we would like to observe—that is, all times, locations, individuals, or driving variables of interest, and hence we must somehow learn about the behavior of "everything" based on a sample. Again, we treat the uncertainties that arise in the observation process stochastically by estimating observation error, which we define as the difference between the true state of interest and our observations of the state.

The final source of uncertainty comes from variation among locations or individuals that is not accounted for by our deterministic model. This variation may arise because of differences in genetics or experience among individuals, or from variation among locations or time periods, that we do not account for. We can acknowledge that such variation exists, even if we cannot (or choose not) to measure or model it explicitly. This acknowledgement requires us to estimate random effects.

Element 3: Basic probability.—Our experience tells us the concepts of process variance, observation error, and random effects are quickly grasped by students, but they are ill equipped to do anything useful with these

concepts. Putting these ideas to use requires a basic understanding of probability theory and the statistical distributions that are used to represent uncertainty. This material serves as the foundation for the elements that follow: hierarchical models, data simulation, likelihood and Bayesian methods, and computational tools. There are a few distributions that have broad application in ecology, notable among them are the Poisson, negative binomial, binomial and multinomial for discrete variables and the normal, lognormal, exponential, beta, and gamma for continuous ones. The properties of these distributions are treated in many fine texts (Hilborn and Mangel 1997, Casella and Berger 2002, Clark 2007, McCarthy 2007, Bolker 2008), and we will not repeat that treatment here. However, we will emphasize a point in teaching distribution theory that we find is often overlooked. Students are usually somewhat familiar with the normal distribution, which has the unusual property that its shape parameters are the same as its first moment and second central moment, the mean and the variance. Other than the Poisson, which is a somewhat special case, no other distributions have this property. Because virtually all texts will give equations for moments in terms of shape parameters, students can usually find moments relatively easily. However the converse is not true—given moments, few students can discover the appropriate shape parameters, even though this is algebraically straightforward (i.e., via "moment matching"). Students need to be able to move deftly back and forth between the moments of distributions and their shape parameters.

In addition to understanding the basic statistical distributions and the kinds of variables that give rise to them, students need familiarity with fundamental tenants of probability theory because they are essential for constructing a full statistical model that marries data and ecological processes. Such data–model integration can often be facilitated by hierarchical or conditional modeling procedures, described in more detail below. That is, simple rules relating joint, marginal, and conditional probabilities are called upon in constructing such models. Moreover, if one takes a Bayesian route to data–model assimilation, then exposure to these probability rules illuminates the basic foundation of the Bayesian framework. Although essential to data–model integration, hierarchical modeling, and Bayesian statistics, such fundamental probability rules are rarely, if ever, introduced in the introductory statistics or applied methods courses that ecology students are advised to take. The basic statistics curriculum would greatly benefit ecologists by increased emphasis on statistical modeling that introduces students to the essentials of probability theory and statistical distribution theory.

Element 4: Hierarchical models.—We believe that virtually all ecological problems can be usefully cast hierarchically (Royle and Dorazio 2008, Cressie et al. 2009). This is because we often confront the following general problem, regardless of the particular topic of our research. We are interested in the operation of a process, for example, population growth, primary production, community assembly, or disease transmission. Rarely is it possible to observe this process directly as there will always be states and parameters that we cannot measure, hereafter described as *latent*. However, we can observe data that are related to the latent state. It is hard to imagine an important problem in ecology for which this is not true. As a result, ecological problems can be usefully dissected into two parts: a process model that describes the behavior of the true, latent state of nature, and a data model that relates what we are able to observe to the latent state. This can be represented more formally as

$$y \sim f(\theta_d, \sigma_d, z)$$

$$z \sim g(\theta_p, \sigma_p) \quad (1)$$

where $f(\cdot)$ is a function relating the data, y, to the latent state, z, and $g(\cdot)$ is a function that describes how the state behaves. Thus, $f(\cdot)$ is a model of the data (with parameters θ_d), which is often termed the "likelihood"; $g(\cdot)$ is a model of the process (with parameters θ_p), which may be referred to as the stochastic process model. The σ_d quantifies uncertainty in the data while the σ_p quantifies variation in the process that is not accounted for by the process model. This general framework can accommodate an enormous range of ecological problems: static and dynamic processes, observational and manipulative studies, simple and complex models, single sources or many sources of data. The data model and the process model may be simple or complex, they may be associated with many parameters or few, several states or a single state. Contextual examples of hierarchical models (Eq. 1) from a range of subdisciplines of ecology can be found in Cressie et al. (2009), Calder et al. (2003), Ogle and Barber (2008), and Ogle et al. (2009).

Students of ecology rarely receive formal training in how to construct a hierarchical model or how to deal with fixed, random, mixed, or crossed effects. In the above example (Eq. 1), the θ's may describe fixed effects of interest, while the σ_p often reflects uncertainty introduced by random effects. Our experience is that current statistics curricula for ecologists requires that students take a series of prerequisite courses, many of which may not be very useful, before being exposed to these topics. This is an inefficient use of time, and the prerequisite courses often hamper the students' abilities to develop the flexible, creative, and intuitive thinking that is required for the types of hierarchical modeling that we see as being critical to the enterprise of data–model assimilation in ecology (also see element 0, above). We believe that ecologists should be trained to view problems using hierarchical models as a starting point. The ordering of element 4 (hierarchical models) and element 6 (likelihood and Bayes) can of course be

reversed, but we have found it is possible to teach the concepts of hierarchical modeling before introducing formal estimation procedures. These concepts can lead logically to data simulation.

Element 5: Data simulation.—If you truly understand your hypothesis about a process, then you should be able to simulate data that would arise if the hypothesis were true. Moreover, we have found in our own work that data simulation is a great way to assure that we understand our hypotheses. We suggest that data simulation is a logical way for students to assemble elements 1–4 in a way that gives strong intuition for the relationships among data, deterministic models (read hypotheses), and multiple sources of uncertainty. Data simulation forces students to think about how data emerge from a process and what distributions are appropriate for representing uncertainty in the process and in the observations it generates. It can be particularly revealing for students to simulate time series of data, observing first hand how process variance propagates over time while observation error does not. In a relatively short time, students can be simulating data sets including parameter uncertainty, process variance, observation error, and random effects. Data simulation is not simply a pedagogical trick (although we have found it is a good one), it also turns out to be an exceedingly useful skill for planning observational and experimental studies, for estimating sample sizes, and for testing and evaluating analytical approaches. Moreover, once students are comfortable simulating data, it is much easier for them to understand how parameters are estimated using likelihood and Bayesian methods.

Element 6: Likelihood and Bayes.—The first five elements provide a foundation for teaching maximum-likelihood and Bayesian approaches to data–model assimilation. Data simulation is particularly important in this regard because we can think of it as the inverse of likelihood. In data simulation, we know the parameters and the sources of uncertainty, which allows us to generate the data. In likelihood, we know the data, which allows us to estimate the model parameters and uncertainties. This relationship helps students understand a concept that, in our experience, they do not find intuitive: understanding likelihood begins with understanding the probability (or probability density) that we would observe a data point conditional on the assumption that it was produced by the processes represented in the model. This is far easier to comprehend if you have already seen, using data simulation, how the model can produce the data.

We advocate exposing students to a variety of problems in maximum-likelihood estimation, including some that they learned in traditional statistics courses, like regression or analysis of variance. In all of these problems, students should be required to choose a likelihood function, justify it, and use an optimization routine to discover parameter values. We like to begin with a simple example where the students can plot the likelihood profile and find the maximum analytically by applying differential calculus. However, real-life problems rarely allow analytical solutions, thus students should be exposed to numerical methods and to the difficulties that can arise when there are, for example, multiple, local solutions. Construction of confidence intervals based on likelihood profiles should be taught as a means for evaluating uncertainty in parameter estimates obtained via maximum likelihood.

A useful link between likelihood and Bayesian methods is to illustrate that the likelihood framework can be complimented by prior information (Hilborn and Mangel 1997, Pawitan 2001), leading to the Bayesian framework that gives the conditional distribution of the unknown quantities given the data, using

$$P(\theta \mid y) \propto P(y \mid \theta)P(\theta). \quad (2)$$

In Eq. 2, $P(y \mid \theta)$ is the likelihood, that is, the probability of the data (y) conditional on the parameters (θ). When placed in the context of Eq. 1, $P(y \mid \theta)$ is essentially given by the product of $f(\theta_d,\sigma_d,z)$ and $g(\theta_p,\sigma_p)$. Maximum-likelihood methods can be used to estimate θ based solely on $P(y \mid \theta)$, where θ represents all parameters of interest (e.g., such as θ_d, θ_p, σ_d, and σ_p, and potentially z). However, we can include results from, for example, previous studies in the prior distribution for θ, $P(\theta)$. When we incorporate prior information, the right-hand side of Eq. 2 is, of course, the numerator of Bayes law, which allows a clear transition from likelihood to Bayesian methods.

The goal of the Bayesian approach is to obtain the posterior distribution of the unknown parameters or quantities, conditional on the observed data, which is represented by $P(\theta \mid y)$. We have found that developing the classical Bayesian foundation from first principles of probability theory can be appreciated by students with minimal quantitative preparation. Woodworth (2004) offers superbly accessible material providing intuition for how Bayesian methods arrive at the probability of the model conditional on the data. The theory behind likelihood and Bayesian approaches to hierarchical models should be introduced with examples illustrating the flexibility provided by conditional (or hierarchical) modeling (e.g., Ogle and Barber 2008). Students exposed to the elements presented here should be able to make an informed decision about whether to choose a likelihood or Bayesian approach. This choice will likely depend upon computational requirements and the availability of prior information.

Given the "best" model, both likelihood and Bayesian methods provide rigorous estimates of latent states, parameters, and uncertainties. However, most often, it is unclear whether one model offers a better approximation of a generating process than another model. Model selection using likelihood and information theoretics has become a very popular framework for analysis in ecology (reviewed by Johnson and Omland 2004) as a

result of the highly influential works of Burnham and Anderson (1998, 2002). However, applying model selection procedures to hierarchical models presents complications for information theoretic methods (Royle and Dorazio 2008). Moreover, the embrace of information theoretics by ecologists is not matched by widespread agreement among statisticians about the best way to evaluate evidence in data for competing models (Link and Barker 2006, Royle and Dorazio 2008). We encourage teaching a range of approaches, including likelihood ratios, information theoretics, predictive loss, and Bayes factors, while being honest with students that none is a perfect solution and that statisticians have not reached consensus on which method is best.

Element 7: Computational methods.—The arrival of fast computers and computationally intensive methods, notably Markov chain Monte Carlo (MCMC), are responsible for the unusually rapid progress in assimilating data with ecological models seen during the last decade. Although application of MCMC methods for estimating quantities of interest in hierarchical models was originally dominated by Bayesian approaches, recent developments allow likelihood methods to achieve many of the same results without specification of prior distributions on states and parameters (Lele et al. 2007). Students should have a basic familiarity with these computational procedures. They should be able to solve simple problems by programming Metropolis or Gibbs samplers and should be familiar with data cloning (Lele et al. 2007). However, our view is that once this basic understanding is achieved, students can usefully exploit software (i.e., WinBUGS, OpenBugs, JAGS) that perform MCMC sampling given specification of all model components (e.g., data distributions, process model equations, priors, and associated probability distributions). By using MCMC front-end software rather than writing their own simulation code, students can focus on model building rather than computational procedures. This allows more time to gain experience with a broad range of problems. Clearly, there is a balance to be struck here: writing MCMC code requires mastery of requisite computational techniques, while using front-end software may increase the number of different problems that students can tackle.

Element 8: The importance of good design.—Even the most sophisticated and modern techniques cannot overcome poorly planned research. Students need to understand the basic principles of research design: randomization, replication, and stratification. Our experience suggests that these principles have far greater meaning to students when they have been exposed to the concepts of uncertainty and hierarchical modeling that we have described.

Element 9: Solving problems.—We have reserved the most important element for last. In our view, skills required to integrate models of ecological processes with data are best gained by active work on problems of appropriate difficulty: problems that are too easy leave students bored, but problems that are too hard leave them discouraged. Of course, as students mature in their skills, they need problems of increasing difficulty, but regardless of how difficult they are, we think the problems that students need share some features in common. The best problems will challenge students to compose models: models of processes and models of the way the processes give rise to data. Students should understand how to use random effects to represent, for example, variation among individuals and locations that is not accounted for directly in the process model. The best problems will involve all of the steps required to estimate states, parameters, and uncertainties. The best problems will give students a great deal of leeway in forming a solution. The ultimate freedom, of course, is to allow students to work independently on projects with their own data and/or the data of their classmates. Students should be able to apply the tools and conceptual procedures learned earlier toward the development and specification of an appropriate process model (or models) and subsequently integrate relevant data via statistical modeling procedures. The current language of choice for solving most of these problems is R (R Development Core Team 2007), and training students in R has benefits for collaboration across disciplines, particularly for collaboration with statisticians.

Discussion

Manipulative, agronomic style experiments have a long history in environmental science, providing many sturdy insights. The inferential framework for these studies is mature and the choice of analysis for a given experimental design is unambiguous. For a long time, manipulative experiments were the gold standard of inquiry, enjoying almost universal sanction by the peer review community. The training needed to support such inquiry is offered at virtually every research university and remains a requirement for many graduate programs in ecology.

Although the designed, manipulative experiment will no doubt remain one of the fundamental tools of ecological inquiry, there is a growing realization that such studies cannot address the full range of problems in ecological research. There is a growing realization that tidy experimental designs often fail to capture the decidedly untidy behavior of ecological systems, particularly the behavior of phenomena that operate at large scales of time and space, that involve multiple interacting or composite forces, and that are inherently nonlinear in their dynamics. Because complexities of nature are not easily compressed into agronomic designs, there is not a single, uniform route to analysis of observations of these systems.

We have made the case that the ecologists of the future need to be grounded in principles of mathematics and statistics that allow them to create models for analysis that are appropriate for a diverse range of

ecological problems, including those that involve complex interactions operating at large scales. Often the data available to solve these problems come from different sources, collected at vastly different spatial scales and frequencies. These data may have been collected to serve different purposes, not all of which are directly motivated by the research question at hand. Often there are data that are missing from specific times or locations. There will almost always be quantities of interest that cannot be observed directly. What this means is that each research problem has a unique set of analytical challenges and that students simply cannot be prepared to attack these problems by learning a fixed set of procedures.

We have advocated a different approach for the training of ecologists, an approach that emphasizes principles at the intersection of ecology, mathematics, and statistics. Training based on principles rather than procedures provides the flexibility that will allow the ecologists of the future to discover solutions to a broader range of problems, each of which may be quite different from the others. Often these solutions will emerge from conversations between quantitatively literate ecologists and ecologically literate mathematicians and statisticians.

Although we urge a principle-based approach, we also admit that it is hard to solve real-world problems with principles alone. Students must be trained in methods for putting those principles to use. In so doing, there is some danger that we simply substitute one set of procedures for another. This is why we believe that the ultimate goal of the training of an introductory course in methods of data–model assimilation must be the development of quantitative confidence. We must give students the self-assurance that, confronted with a new problem, they can master the contemporary methods required to solve it, or at the very least can master enough of the relevant material to collaborate effectively with statisticians and mathematicians. These methods will evolve over the course of their careers (Hobbs et al. 2006), and their graduate training must prepare them to keep pace with this evolution.

Given that experiments will remain an essential tool, does this mean that students should take traditional methods courses as well as the principles based course we have outlined? There is, of course, a limit to the amount of coursework that graduate students can be reasonably expected to take and it remains an open question whether principle-based courses should replace, rather than complement, procedural ones. A student who understands maximum-likelihood and Bayesian approaches to gaining insight from models will be able to master traditional procedures (analysis of variance, regression, etc.) relatively easily, and probably can self-teach specific applications of these procedures as the need arises, allowing the problems they confront to guide their learning (Hilborn and Mangel 1997). For example, Hobbs once had a student who, equipped with an understanding of binomial likelihoods, was able to see his way to "reinvent" logistic regression from first principles. However, the converse is clearly not true: trained in procedures, students realistically cannot find their way to develop likelihood or Bayesian based analyses.

We have found that one of the most important benefits of principles-based quantitative training is enhancing our ability to collaborate with mathematicians and statisticians. For example, an ecologist who understands the fundamentals of distribution theory, rules of probability, basics of MCMC algorithms, and knowledge of software commonly used by statisticians is well equipped to have productive conversations with his or her statistical colleagues about challenging problems. The same cannot be said for someone trained in procedures alone. These kinds of collaborations are likely to be required to solve the most challenging problems in biology in the future (Schwenk et al. 2009).

Until now, we have focused on graduate training, but we believe that the same elements can be applied to training ecologists who have long since finished their formal education. The dramatic acceleration in development of quantitative tools has left many researchers lagging behind the field, unfamiliar with methods that are key to contemporary insight. There is a great need to provide opportunities for learning for this group of scholars. For many, accessible papers and texts are all that is required for self teaching. However, we also see an unmet need for workshops, short courses, and online training directed at ecological researchers (e.g., Appendix). Moreover, we also believe that the principles we advocate could usefully extend into undergraduate teaching in biology. Many of the ideas we have offered are included in at least one text aimed at undergraduates (Woodworth 2004).

In conclusion, we were motivated to write this paper because our experience has shown us that new approaches to the analysis of ecological data can provide insights that heretofore would have been impossible to obtain. Moreover, we find these approaches provide an intellectually more satisfying way to do research. We advocate these principles and methods as a foundation for training graduate students in gaining insight from data and models.

Acknowledgments

The ideas in this paper developed in part from many helpful conversations with colleagues, including Jennifer Hoeting, Jim Clark, Ray Hilborn, Maria Uriarte, Saran Twombly, Paul Duffy, Yiqi Luo, and the many bright students who have challenged Hobbs and Ogle to be clear in their teaching. We acknowledge support from National Science Foundation (NSF) Awards EF-0914489 and EPS-0447681. The work reported here was supported in part by the NSF while N. T. Hobbs was serving as a rotating Program Director in the Division of Environmental Biology. Any opinions, findings, conclusions, or recommendations are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Literature Cited

Bolker, B. 2008. Ecological models and data in R. Princeton University Press, Princeton, New Jersey, USA.

Brewer, C. A., and L. J. Gross. 2003. Training ecologists to think with uncertainty in mind. Ecology 84:1412–1414.

Burnham, K. P., and D. R. Anderson. 1998. Model selection and inference: a practical information-theoretic approach. Springer-Verlag, New York, New York, USA.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multi-model inference: a practical information-theoretic approach. Springer-Verlag, New York, New York, USA.

Calder, C., M. Lavine, P. Muller, and J. S. Clark. 2003. Incorporating multiple sources of stochasticity into dynamic population models. Ecology 84:1395–1402.

Casella, G., and R. L. Berger. (2002). Statistical inference. Second edition. Duxbury, Pacific Grove, California, USA.

Clark, J. M. 2007. Models for ecological data. Princeton University Press, Princeton, New Jersey, USA.

Cressie, N., C. A. Calder, J. S. Clark, J. M. V. Hoef, and C. K. Wikle. 2009. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Ecological Applications 19:553–570.

Edelstein-Keshet, L. 1988. Mathematical models in biology. McGraw-Hill, New York, New York, USA.

Gotelli, N. J., and A. M. Ellison. 2004. A primer of ecological statistics. Sinauer, Sunderland, Massachusetts, USA.

Green, J. L., A. Hastings, P. Arzberger, F. J. Ayala, K. L. Cottingham, K. Cuddington, F. Davis, J. A. Dunne, M. J. Fortin, L. Gerber, and M. Neubert. 2005. Complexity in ecology and conservation: mathematical, statistical, and computational challenges. BioScience 55:501–510.

Hairston, N. G. 1989. Ecological experiments: purpose, design, and execution. Cambridge University Press, Cambridge, UK.

Hastings, A., P. Arzberger, B. Bolker, S. Collins, A. R. Ives, N. A. Johnson, and M. A. Palmer. 2005. Quantitative bioscience for the 21st century. BioScience 55:511–517.

Hilborn, R., and M. Mangel. 1997. The ecological detective: confronting models with data. Princeton University Press, Princeton, New Jersey, USA.

Hobbs, N. T., S. Twombly, and D. S. Schimel. 2006. Deepening ecological insights using contemporary statistics. Ecological Applications 16:3–4.

Johnson, J. B., and K. S. Omland. 2004. Model selection in ecology and evolution. Trends in Ecology and Evolution 19:101–108.

Jones, M. B., M. P. Schildhauer, O. J. Reichman, and S. Bowers. 2006. The new bioinformatics: integrating ecological data from the gene to the biosphere. Annual Review of Ecology Evolution and Systematics 37:519–544.

King, R., and O. Gimenez. 2009. Bayesian analysis for population ecology. Cambridge University Press, Cambridge, UK.

Lele, S. R., B. Dennis, and F. Lutscher. 2007. Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. Ecology Letters 10:551–563.

Levins, R. 1966. The strategy of model building in population biology. American Scientist 54:421–431.

Link, W. A., and R. J. Barker. 2006. Model weights and the foundations of multimodel inference. Ecology 87:2626–2635.

Link, W. A., and R. J. Barker. 2010. Bayesian inference with ecological applications. Academic Press, San Diego, California, USA.

Little, R. J. 2006. Calibrated Bayes: a Bayes/frequentist roadmap. American Statistician 60:213–223.

Luo, Y., K. Ogle, C. Tucker, S. Fei, C. Gao, S. LaDeau, J. S. Clark, and D. Schimel. 2011. Ecological forecasting and data assimilation in a data-rich era. Ecological Applications 21:1429–1442.

Mangel, M. 2006. The theoretical biologists toolbox: quantitaive methods for ecology and evolutionary biology. Cambridge University Press, Cambridge, UK.

McCarthy, M. A. 2007. Bayesian methods for ecology. Cambridge University Press, Cambridge, UK.

McGarigal, K., S. Cushman, and S. Stafford. 2000. Multivariate statistics for wildlife and ecology research. Springer-Verlag, New York, New York, USA.

National Research Council. 2003. BIO2010: transforming undergraduate education for future research biologists. National Academies Press, Washington, D.C., USA.

Ogle, K. 2009. Hierarchical Bayesian statistics: merging experimental and modeling approaches in ecology. Ecological Applications 19:577–581.

Ogle, K., and J. J. Barber. 2008. Bayesian data–model integration in plant physiological and ecosystem ecology. Progress in Botany 69:281–311.

Ogle, K., J. J. Barber, C. Willson, and B. Thompson. 2009. Hierarchical statistical modeling of xylem vulnerability to cavitation. New Phytologist 182:541–554.

Otto, S. P., and T. Day. 2007. A biologist's guide to mathematical modeling in ecology and evolution. Princeton University Press, Princeton, New Jersey, USA.

Pawitan, Y. 2001. In all likelihood: statistical modeling and inference using likelihood. Oxford Scientific Publications, Oxford, UK.

Pilkey, O. H., and L. Pilkey-Jarvis. 2007. Useless arithmetic: why environmental scientists can't predict the future. Columbia University Press, New York, New York, USA.

R Development Core Team. 2007. R: a language and environment for statistical computing. R Project for Statistical Computing, Vienna, Austria. ⟨www.r-project.org⟩

Robeva, R., and R. Laubenbacher. 2009. Mathematical biology education: beyond calculus. Science 325:542–543.

Royle, J. A., and R. M. Dorazio. 2008. Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations, and communities. Academic Press, London, UK.

Scheiner, S. M., and J. Gurevitch, editors. 2001. Design and analysis of ecological experiments. Oxford University Press, Oxford, UK.

Schwenk, K., D. K. Padilla, G. S. Bakken, and R. J. Full. 2009. Grand challenges in organismal biology. Integrative and Comparative Biology 49:7–14.

Sokal, R. R., and F. J. Rohlf. 1995. Biometry: the principles and practices of statistics in biological research. W. H. Freeman and Co., New York, New York, USA.

Wang, Y. P., C. M. Trudinger, and I. G. Enting. 2009. A review of applications of model-data fusion to studies of terrestrial carbon fluxes at different scales. Agricultural and Forest Meteorology 149:1829–1842.

Woodworth, G. G. 2004. Biostatistics: a Bayesian introduction. Wiley, Hoboken, New Jersey, USA.

APPENDIX

Schedule for a one-semester course at Colorado State University (NR 575: Systems Ecology) introducing concepts of model–data assimilation to graduate students in ecology (*Ecological Archives* A021-071-A1).