How to fit Bayesian models and influence people *or* How to do Bayesian model fitting in ecology *or* The best way to be a Bayesian in ecology today

EM Wolkovich and WD Pearse and and TJ Davies and M Betancourt?

August 12, 2023

Abstract

Improvements in algorithms and computational speed have heralded a new a era of Bayesian model fitting. Models are easier to fit, faster to run, and more flexible to fit, making them ideal for many of the complexities of ecological sampling designs. This has opened up Bayesian approaches to a new world of users, but how best to start using and implementing Bayesian approaches in ecology is still somewhere between old rules and methods and the new school where Bayesian approaches also include a specific workflow of not just how to do stats, but how to do science. Here we review a simple form of a Bayesian workflow that integrates mechanistic and statistical models with a computational toolkit that we argue could accelerate ecological science. [Maybe along the way we highlight common practices in ecology that are stupid as all tomorrow—as our approach makes clear—and outline best practices for jumping into the wonderful surf of the happy Bayesian ocean.]

1 Main text

Recent years have seen the explosion of Bayesian models across fields (CITES), including ecology. This change comes in part from increased computational power, but more from new algorithms (e.g., Hamiltonian Monte Carlo) that have made models faster and arguably easier to fit. Capitalizing on these advances R packages such as brms have seen much wider use than the previous generation of packages that streamlined fitting a set of pre-determined models using Bayesian approaches (MCMCglmm. Bayesian approaches, long heralded as more powerful, flexible and especially adept at capturing the multiple levels of variance in ecological data, seem poised to help ecology as a discipline advance in leaps and bounds.

Ecology itself seems poised to leap out of the jungle and land into the role society is asking of it as growing global challenges demand more predictive models of how communities and ecosystems will shift with climate change, habitat degradation and all that jazz. Ecology is now three decades on after the start of the synthesis movement, which required more advanced statistical and computational approaches of ecology to test fundamental theory across systems (CITES). Early concerns have given way to a widespread appreciation of the value of these new approaches, alongside a resounding fervor over 'natural history.' At the same time, the growing anthropogenic challenges have increased the need for predictive models and forecasts from ecologists.

The result is that the average ecologist today has a diversified skillset compared to that of decades ago. Ecology has long been a field with deep statistical training, but in many ways the modern ecologist is now also expected to be computational—able to handle large datasets, produce repeatable workflows and help translate models into forecasts for planning. Many ecologists now bridge field and computational methods. As such, the rise of Bayesian approaches is especially timely for ecology.

Bayesian approaches are in no way new to ecology, as they have long been used in certain subdisciplines related to estimating population sizes of things people want to eat or manage. For example, wildlife biology widely uses mark-recapture data and associated models to estimate population sizes; these methods almost always use Hidden Markov models (HMM), which can rarely be fit without Bayesian methods (CITES). Similarly, fisheries biology has tended towards more complex models—such as state-space models—to estimate fish stocks, which are similarly easiest to fit with Bayesian approaches (CITES). For decades, Bayesian training in ecology has focused on these aims, but as data and methods change, so has how ecologists are using Bayesian models.

Today, as large-scale ecological data becomes available for more diverse systems and for questions addressing other aims, more ecologists are using Bayesian models in new ways. Yet, at the same time, training in statistics and in Bayesian modeling in particular may struggle to keep pace. Bayesian approaches provide a pathway to powerful models that can transform how we understand our systems, but they can also lead to pitfalls most ecologists are not trained to notice or deal with. These pitfalls can be avoided by approaching Bayesian analyses through specific workflows (CITES), which themselves are built on a process of how to do not just stats, but science. Here we describe a broadly generalizable workflow for Bayesian analysis and show how it can revolutionize training in ecology by integrating more model building and model understanding. Once you start doing this workflow, your scientific life will never be the same.

1.1 Box by WPD with 3 paragraph explanation of Bayes

Someone with the initials WPD will hopefully write this.

1.2 A brief overview of the benefits—and pitfalls—of Bayesian models

Bayesian models have many benefits, but an often-mentioned one is that 'you can fit any model you want.' While this is not entirely true (CITES), compared to the models ecologists can fit in popular modeling packages (lme4) Bayesian modeling options can feel limitless. As long as you can write the out the likelihood of your desired model and assign priors to all parameters, you can generally 'fit' any model. This includes non-linear ones, non-Gaussian families (Poisson, beta or combinations thereof such as hurdle models), hierarchical designs and any combination of these, as well as 'joint' models where parameters estimated in one equation appear in another, thus carrying through estimated uncertainty. Such flexibility is incredibly powerful in ecology where data are often influenced by complex spatial or temporal patterns, non-linear processes are widespread and common data types are non-Gaussian (counts, percent cover etc.).

Fitting a bespoke model to your data and question frees scientists to estimate useful parameters—effectively, a way to get the numbers we often really want but don't have access to. Thus, instead of reporting a treatment's p-value and accompanying F statistic and degrees of freedom, models can be designed to estimate and report effects per °C of warming or at what level non-linearities due to high temperature begin—always with estimated uncertainty. While replication crises in other fields, driven in part by a overly zealous focus on p-values (CITES), and the rise of metaanalyses in ecology (CITES) have led to a somewhat greater focus on ‘effect sizes’ in ecology (often used to refer to very specific unitless statistics, such as Cohen’s d), bespoke models take this to a new level. Researchers can easily estimate comparable effect sizes from z-scored data, alongside estimates in meaningful natural units, such as per °C of warming or per hectare of habitat lost.

But this valuable flexibility is also one of the greatest pitfalls of Bayesian models You can fit almost whatever you want, but critical parts of your model might be almost entirely unimpacted by your data. And in ecological model fitting, we’re most often interested in parameter estimates strongly informed by our data.

This ‘pitfall’ of Bayesian is not new, nor unique to ecology—though the complexity of ecological data and processes may make it especially pernicious in ecology—and decades of statistical research has aimed to develop best practices when using Bayesian models to avoid this. These best practices generally center around a specific workflow; a variety of which can be found in exquisite detail elsewhere (CITES). We do not aim to repeat those here, but instead to provide you with a highly simplified but powerful workflow we believe when applied to Bayesian modeling in ecology could greatly accelerate progress.

1.3 Our simple Bayesian workflow

We outline a basic Bayesian workflow below that includes what we consider the major steps for Bayesian model fitting. Many steps should be familiar to statistical ecologists, but are often overlooked, whereas other steps may appear particular to Bayesian (prior predictive checks), but are actually useful for anyone—using Bayesian models or not—to challenge their models of how the world works, and learn from them. As we focus on a simplified list of major steps, many of the smaller but still critical steps are omitted. For example, visualization is required at every step—especially 2 and 4; while we do not discuss this explicitly we refer to relevant publications.

We assume a user of Stan, a relatively new probabalistic programming language, that interfaces with R, Python, Julia (and more) to write bespoke Bayesian models and underpins the R packages brms and rstanarm, which fit a suite of specific (pre-defined) models. We focus on Stan as its MCMC algorithm (a variant of Hamiltonian Monte Carlo, HMC) is fast and produces specific output to warn of model fit issues (i.e., divergent transitions) in a way other MCMC algorithms do not (Metropolis or GIBBS), but the basic workflow should apply to diverse implementations of Bayesian modeling.

Step 1: Get your model running.
This obvious step assumes a suite of crucial work already done to define your model or interest. You must understand your question, your subdiscipline, and your data enough to formulate a useful model that you want to fit. This step thus assumes you arrive with a wealth of scientific expertise and a well enough thought out model.

Now, you need to write up the model and check you wrote it up correctly. Whether writing it out in Stan, where you need to be able to write out the full likelihood and set all your own priors, or using a package that writes much of the model for you (rstanarm), you need the code and a way to verify the code is correct—test data (aka ‘fake data,’ or ‘simulated data’, etc.). To build test data you need to understand your model well enough to generate data from known parameters, then run that data through your model and confirm it returns the parameters.

This very basic model checking step is uncommon for many ecologists, but critical in our view. If you can simulate data from your model, then you can powerfully—and easily—answer questions related to statistical power, what effect sizes are reasonable, and, most likely, have new insights into how your model suggests the world works. 'All models are wrong; some models are useful,' becomes much clearer when you have the power to generate data from your model under any parameter set you want. Once your Stan output returns the parameters you expect from your test data, you can move onto interrogating your priors.

Step 2: Check your priors.
Priors are the source of much discussion and focus by Bayesian and non-Bayesian researchers alike. Often treated as the big bad wolf of Bayesian, or the unseen hand producing the model fits you get, according to some. More accurately they are half of the equation that gives you your model posterior; and they are *designed to me that way.* Bayesian philosophically is built around the idea that you have prior knowledge you trust and want to compare to new data—showing up through your likelihood.

How much they influence your model fit is in many ways up to, and your data. Depending on your model and the data, the likelihood (influenced by your data) can easily overwhelm your priors. On the other hand, as you can fit a Bayesian model with 100 parameters and only 5 datapoints, and then your priors likely matter a lot. Most dangerously, failing to think about them, check them and understand your model enough, can mean they may matter more than you know.

Luckily you can use the code you wrote to build test data and scale it up for prior predictive checks. In these, you explore a distribution of potentially reasonable values for your priors, then see how they influence your resulting output. How exactly to do this depends on your question, model and aims, but many guides can help you think through this (CITES).

In our experience, prior predictive checks can quickly disabuse newbies to Bayesian from their discomfort about priors. For example, you may not think you have a prior on how sunlight affects plant growth, until you realize your priors suggest plant growth of meters per day for one individual plant. Prior predictive checks in this way serve both as a check on the priors your using, and to further explore the model of the world tow hich you're planning to fit your data.

Step 3: Run your model on your empirical data.
This is the step many ecologists skip straight to, ourselves included. It's easy to see the appeal—fitting our new data to the model can feel like the moment when we'll learn something new. But, at least in our experience, this is not always the case. When we rush to this step, that first model we fit is often followed by another, and another—perhaps because one does not converge, or the results of another does not make immediate sense. After a while of this process, it can easily feel like we're not sure what we learned, if anything.

In contrast, by approaching the model through Steps 1-2, it's often much easier to quickly see through the results of the model fit. And also easier to plan next steps.

Step 4: Retrodictive predictive checks
Once you have your posterior based on your model and new empirical data, it's time to interrogate it. Steps 1-2 have set you up well for this, as you have a sense of what different parameter estimates do to the model. Now you have the parameter estimates from your posterior you can simulate new data from them and see how that new world looks to you—called retrodictive predictive checks (or posterior predictive checks). Exactly how to do these are—again—up dependent on your question, model and aims (but again, lots written on this, CITES). In contrast

to prior predictive checks, however, this step is built into some software. If you use rstanarm, then the package shinystan will automatically give you a set of retrodictive predictive checks, including histograms of simulated data, as well as examine the mean and variance of multiple simulated datasets. Often here you may find big differences from your empirical data and can start to generate hypotheses for why.

With those hypotheses in hand, you may very well want to tweak your model—which all part of the workflow. At that point you return to Step 1, tweak your code, and repeat the process. In this way, fitting multiple models is encouraged, but in a far more structured and careful way than traditional ecological model fitting in our experience. And there are big benefits to it.

This process more fully integrates mathematical modeling into statistical modeling. To complete Step 1, you have to understand the underlying math of your model enough to simulate data from it. This can be challenging at first (in our courses many students cannot recall how to simulate y data from a simple linear regression), but is immensely beneficial to understanding your model. Indeed, we have found the greatest insights come not from the step we all know best—fitting the model with empirical data—but from every other step in this workflow. Developing simulated data to test the model, running prior and retrodictive checks all dive you deep into understanding your statistical model, which suddenly you may find yourself thinking through much more mechanistically. In our experience this process has quickly translated into insights for our biological systems, and all changed how we approach statistical models.

1.4 How this workflow changed our science

As we have used this workflow, how we approach our statistical models has changed. These changes have generally been similar for each of us. We suspect they are not unique to us, our study systems or questions. Instead, we think they represent common approaches to statistical modeling in ecology that could help the field progress, much as we believe they have helped our science.

Understanding nonidentifiability

Identifiability and—more common in our experience nonidentifiability—refers to whether whether all parameters in a model can be uniquely identified. Models can be nonidentifiable in several ways, including when mathematically some parameters cannot be uniquely defined. Statistically, nonidentifiability often is an outcome of the empirical data combined with the model. In this definition nonidentifiability occurs when the data do not contain “sufficient information for unique estimation of a given parameter or set of parameters in a particular model” (CITES).

Nonidentifiability can come up in many ways in ecology—and be hard to see, especially if you rushing through model fitting. But if you have to write out your model and simulated data, you may suddenly realize lots of places for nonidentifiability to live. For example, when species do not occur across most sites, a model including separate parameters for site and species is often nonidentifiable, but there’s no warning in packages to tell you this; you have to learn it for yourself and test it out. We have become far better experts at nonidentifiability based on this workflow—and we have adjusted how we collect data and interpret results extensively because of it.

Simplify
Once we noticed how pervasive nonidentifiability and weak identifiability (where parameter estimates are highly uncertain because the model and data together are nearly identifiable) were

we started simplifying our models. Whereas previously when we had data that qualitatively appeared complexly nested, crossed, split or twisted, we would have tried to fit all of these intricacies (on the intercept) we are now more slow to add these to our models. By both understanding these terms better (including the many different ways each can be modeled), and understanding them better depending on the data and model in each unique context, we work slowly though what to include, and when it is most critical.

While before using this workflow, some of us would start with complex models, then simplify models until they converged. Often these were hierarchical models with many levels, built based more on a dogma of non-independence (and, somewhat relatedly, correct degrees of freedom) than an understanding of the model. We also often fit a suite of interactions: multiple two-way interactions and the occasional three-way interaction were common fare. But in simulating data, and fitting models to real, messy, imbalanced data using the workflow we came to see how much we were asking of our data and models together. Fitting a two-way interaction with half the effect size of a main effect takes a 16X sample size (CITE). This is sobering. It's more sobering when you see it played out again and again through this workflow.

We now simplify based on a more careful reckoning. Usually this starting model is not simple, it often includes grouping factors that may be difficult to fit, but that we see as absolutely critical to the question, model and data at hand. We build up from there, often grouping posteriors by some factor (site or biome) before we add it to our model. However, the ending model is often not as complex as we may have fit if we did not follow all the steps. And—through them, especially Step 4—we can feel confident our model is capturing important variation.

Understanding this variation includes re-approaching how we plot our model and data. While many of us are good at plotting raw data, and plotting our main model parameters (and plotting both these pieces together), the world in between is rarely taught. And it's in this in-between world that we have found a deep understanding of our models, what they're doing and—relatedly—what our results show. Many packages offer model predictions—based on a full or near-full model, but being able to decompose model predictions into varying components can give powerful insights. For example, model predictions can be plotted with and without major grouping factors such as 'block,' 'plot', or 'species,' and suddenly how predictable a model is (or not) based on other factors like treatment or time becomes clear.

Looking at parameters in our models
Before this workflow, not all of us commonly discussed the values that parameters in our model took—things like the slope and intercept (two common model parameters) were sometimes reported, but we did not know them as well as we knew whether the p-value for the slope was < 0.05. This changes quickly when you need to build simulated data for a phenological event (a day within the calendar year so, ideally between 1 and 366) and suddenly find it's >1000 based on the quick parameters you put into your code.

This focus on the value of parameters scales up through the workflow and across projects. Having a better sense of parameter values across different biological context, model parameterizations. and time periods gives us a better sense of how the biological world works, including what's reasonable, possible or wildly unrealistic.

1.5 How this workflow reshapes ecological training

This four-step workflow is a simplified version of the current best practices for Bayesian model fitting, but many of the skills required are not part of traditional ecological training. Writing out the math behind most statistical models enough to complete Step 1 bleeds into the skillset usually reserved for those working on theory, where coding and simulating from a model are common tasks. In contrast field, lab and otherwise empirical-data based ecologists often fit models they could simulate data from. This dichotomy seems short-sighted in our current era, of bigger, messier data and greater computational methods poised to handle such messy data—if scientists are trained in how to build the right models for their questions and data. We believe training ecologists with a skillset where they quickly use this workflow could advance ecological science rapidly.

This workflow also requires a shift in how we approach Bayesian methods—and training in Bayesian models—in ecology. Already, uptake of new Bayesian R packages highlight that Bayesian methods are no longer the purview of only wildlife and fisheries biologists, and these changes come alongside advances in Bayesian workflows, algorithms, and visualiziations (CITES), that ecology must adapt its training for. While this is an active area, we highlight three major changes.

(1) Prior 'beliefs' are changing. Best practices for determining priors is an active area of statistical research (CITES), and training should reflect current best practices. These include that 'non-informative priors' are a misnomer, as they are often informative (CITES), and priors can easily be 'weakly informative' and thus a strong focus on the dangers of priors in training can be overkill (see supp or box where WDP includes part of his rant?).

Further, training often include a strong focus on priors, including conjugate priors (because of the importance of conjugate priors in closed form solutions for particular posteriors). Modern algorithms, such as HMC, do not require conjugate priors, which are now antiquated (and thus sometimes referred to as 'the crystal deodorant of Bayesian statistics'). Prior predictive checks provide a far more powerful way to understand how priors work within a particular model, and are far more useful than rules about which priors should be fit in certain cases or memorizing which priors are conjugate.

(2) 'Random effects' are not just random. Hierarchical models contain grouping factors, sometimes referred to as 'random effects,' such as species or individual. This term, however, is misleading, imprecise and thus no longer recommended (CITES). In ecology, it also carries with a heavy weight of older 'rules' of what is 'random' versus 'fixed,' including that 'random effects are things you don't care about, like block.' After a couple retrodictive posterior checks (Step 4), you might feel differently, as hierarchical effects are (by definition) drawn from an underlying distribution—meaning you can predict outside of the specific set you sampled, meaning you can predict for a new species or individual, whereas you cannot do so for many 'fixed effects.'

(3) P-values are easily misleading ('I am arbitrary but my story is often told ... '), and there's no easy fix for that. The replication crisis, rampant in other fields, is based in part on an overly hopeful belief that p-values will separate the signal from the noise, with one easy number. In reality, small sample sizes, lack of routine reporting of interpretable effect sizes, fitting of many models without adequate explanation, poor data and code reporting habits all increase the chance of finding 'significance' at a level of ≤ 0.05 (CITES). This reality means a similar crisis is likely lurking in ecology, especially given small sample sizes alongside a tendency to fit complicated models with multiple interactions. The answer to this, however, is not Bayesian approaches, which (as we touch on above), bring their own ways to sift exciting results from a pile of chafe.

The answer is a workflow for careful model building, model fitting and model interrogation—as we outline here. Try it, and see if it transforms your life they way it transformed ours!

Take home messages (maybe)

1. You should not fit a model you cannot simulate
2. Fit simpler models
3. Know your nonidentifiability

2 Figures

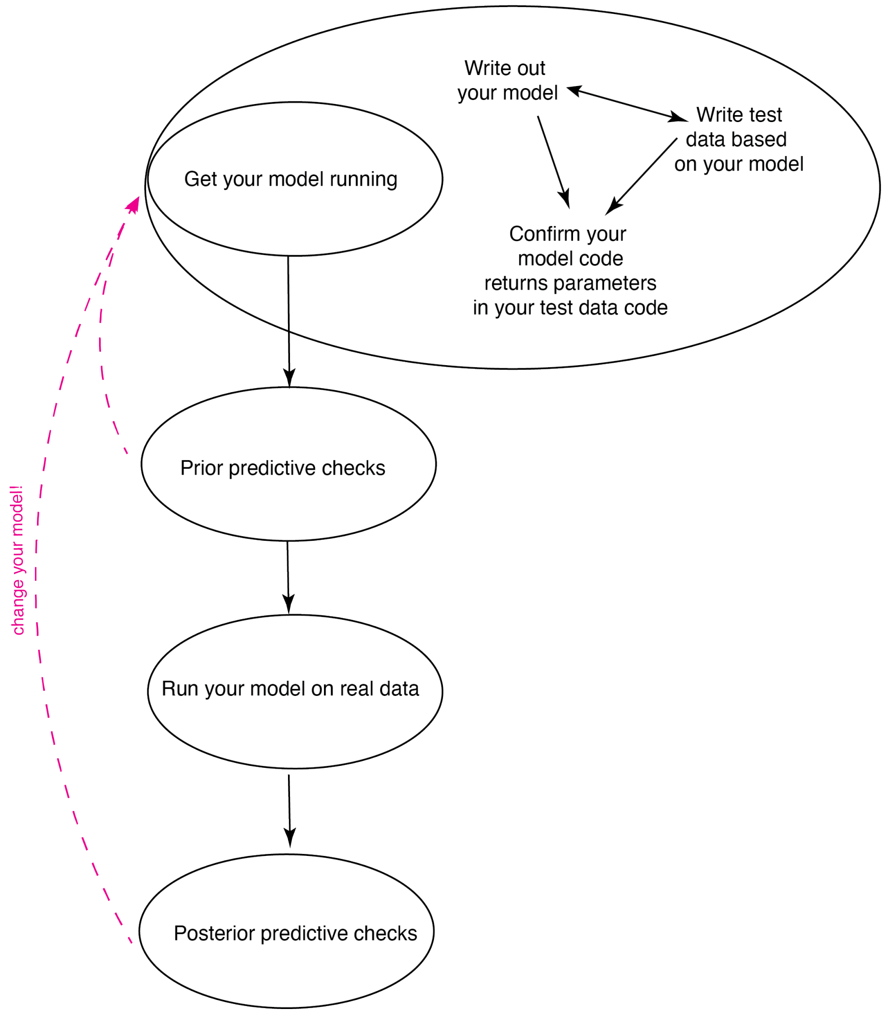
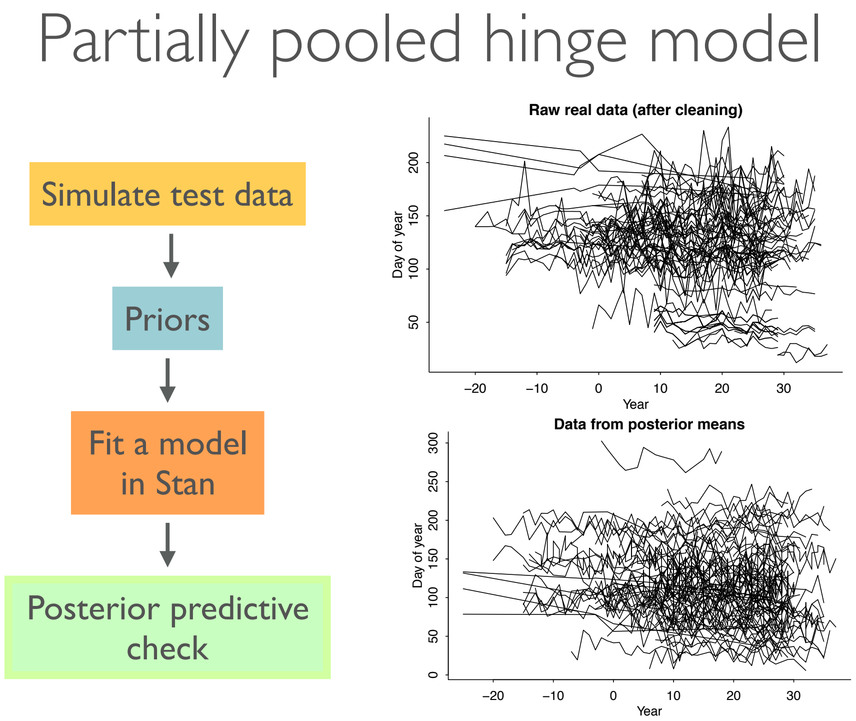


Figure 1: Simple workflow.



Partially pooled hinge model

Simulate test data

Priors

Fit a model in Stan

Posterior predictive check

Figure 2: Reminder to Lizzie of a retrodictive check we could include.