A three-step Bayesian workflow for improving ecological science

EM Wolkovich, TJ Davies, WD Pearse & M Betancourt

January 15, 2024

Abstract

Improvements in modeling languages, algorithms, and computational speed have heralded a new era of Bayesian model fitting. Models are more flexible, easier to fit, faster to run, making them ideal for capturing many of the complexities of sophisticated sampling designs and bigger, messier datasets. This has opened up Bayesian approaches to a new world of users. Yet many ecologists do not currently take advantage of these new possibilities, and instead rely on the traditional approaches, work-arounds, and diagnostics that sufficed when other options were lacking. For many the transition to Bayesian methods can be overwhelming, with analysts suddenly having to confront for example modeling assumptions and computational accuracy that have been hidden behind software defaults and largely taken for granted. In this paper we show how Bayesian methods, and even many frequentist methods, can organized into systematic workflows that drastically ease the development of analyses compatible with ecological domain expertise.

Here we describe a broadly generalizable workflow for Bayesian analysis centered on simulating data from models, and show how it can revolutionize training in ecology. Building on the increasingly computational toolkit of many ecologists, this approach represents a ground shift of not just how to fit Bayesian models, but how we should approach model fitting to advance our science. By integrating model building and testing more fully with ecological theory, concepts and understanding this simulation-based workflow helps fit models that are more robust and well-suited to provide new ecological insights. This in turn can help us refine where to put resources for better estimates, better models, and better forecasts. While we outline this workflow using Bayesian methods, the general approach could be used by anyone fitting models to data.

Main text

Recent years have seen an explosion of Bayesian models across scientific fields (van de Schoot *et al.*, 2021; Schad *et al.*, 2021; Grinsztajn *et al.*, 2021), including ecology. This change comes in part from increased computational power, but mostly from new algorithms (e.g. Hamiltonian Monte Carlo, Hoffman & Gelman, 2014; Betancourt, 2019) that have made fitting and implementing Bayesian models faster, more robust and—in many ways—easier (Carpenter *et al.*, 2017). Capitalizing on these advances R packages, such as brms (CITE), that streamline fitting a set of pre-determined models have seen growing use. Bayesian approaches, long heralded as more powerful, flexible and especially adept at capturing the multiple levels of variance in ecological data, are positioned to advance ecology as a discipline.

At the same time, ecology itself appears well positioned to apply a century of theory and empirical insights to address the global challenges society is facing. Growing anthropogenic pressures

have increased the need for predictive models from ecologists, as policy-makers demand forecasts of how communities and ecosystems will shift with climate change, habitat degradation, and other anthropogenic forces to balance conflicting societal needs and desires. The solutions ecologists can offer are rarely simple, and communicating the inherent risks and trade-offs in decision-making would benefit from forecasts that include uncertainty, and highlight the variability of ecological systems that is often as, or more important, than means or other summaries.

To meet these demands, ecology has developed ever larger datasets to test fundamental theory across systems (Hampton *et al.*, 2013), but bigger data is messier data that ecologists are not usually trained to handle. Such data generally requires a model of both the underlying biological processes and how the measurements were made. While some fields have long used these types of models (generally in fields focused on inferring population sizes of things people want to eat or manage, Muthukumarana *et al.*, 2008; Zheng *et al.*, 2007; Trijoulet *et al.*, 2018; Strinella *et al.*, 2020), most have not. Thus ecologists have tried to adapt what they were trained in—traditional statistical methods (e.g., F and t tests) and a strong focus on null hypothesis testing. They have done this often fitting multi-way interaction terms (rather than fitting mechanistic models informed by biological understanding), using random effects to correct for group level factors (rather than explicit models of data-collection biases), or comparing across a large suite of models (because they cannot fit a single model that accounts for all the idiosyncracies and biases).

But these approaches do not actually align with ecology's aims today. Beyond the reality that most traditional methods are fragile when used beyond the cleaner, simpler experiments these methods assume (e.g. spatial, temporal and phylogenetic correlations often violate these methods independence assumptions), multi-way interaction terms can make main effects hard to interpret and require much larger sample sizes to estimate reliably, null hypotheses tend to be rejected even when the underlying effect sizes are negligible (Gelman & Hill, 2009; Muff *et al.*, 2022), and many model comparison approaches prefer models whose inferences match the idiosyncratic fluctuations in the ecological data to which they are fit, but don't generalize to other observations.

Bayesian approaches provide a pathway to powerful models that can transform how we understand our systems as large-scale ecological data increasingly becomes available for more diverse systems and new questions. Ecologists recognize this potential and are increasingly using Bayesian approaches, but are often inadequately prepared to notice or manage the pitfalls of such models. Many of these pitfalls can be avoided by approaching analyses through specific workflows (Betancourt, 2020; Grinsztajn *et al.*, 2021; van de Schoot *et al.*, 2021), which themselves are built on a process of how to do not just statistics, but how to do science (**?**).

These approaches move away from a focus on null hypothesis testing, towards estimating effect sizes, using models calibrated (see Table 1) and better understood through simulating data at multiple steps—using a number of skills often reserved in ecology more for 'theorists.' Given the societal demands on ecologists today, we argue this divide is antiquated—especially as the average modern ecologist is computational, we believe they also need the skills to build and understand their own models. Breaking the training divide between theoretical and empirical ecologists, we argue, would yield greatly increased flexibility in what models can be fit and improved insights into the underlying systems.

Here we provide a highly simplified—but powerful—workflow for modeling in ecology, that builds on new insights from statistics (Betancourt, 2020; van de Schoot *et al.*, 2021)—and suggests a new way to train ecologists for the skills they need today. By integrating model building more

fully with ecological theory, concepts and understanding—and vice versa—this approach can fit models that are more robust and better-suited to providing new ecological insights and robust predictions than traditional approaches.

A short guide to statistical workflows and Bayesian approaches

Statistical analyses are designed for inference—to learn about some process or effect or behavior from data (see Table 1). Robust analyses, however, rely on our inferences being consistent with the underlying truth more often than not. Quantifying this consistency is calibration—analyzing how often a parameter estimate is close to the true value—a critical part of using models for inference. A major problem with traditional (frequentist) approaches in ecology today is their inferences are unpredictable when their foundational assumptions fail, but ecologists are not usually trained in how to recognize or deal with this. The workflow we recommend avoids these pitfalls. It provides an organized sequence of steps to explicitly build and challenge a model through simulation. We believe this is easier and more straightforward to do with a Bayesian approach and thus outline the workflow assuming a Bayesian approach, but we suggest a similar approach could, and should, be extended to other statistical inference methods.

To better understand our workflow, we provide a very brief overview of some of the fundamentals of Bayesian methods that is inherently incomplete and, by design, not very technical. This section can be skipped for those who feel already well-versed, and can be augmented for those who are new to Bayesian approaches (for example, MacElreath, 2016; Gelman *et al.*, 2014; Gelman & Hill, 2020).

A brief review of statistical inference using Bayesian approaches

Probability is often defined as "the long-term frequency with which something happens." We would expect, for example that if we tossed a coin 100 times we would see roughly 50 heads. In this case we would say that the probability of tossing a coin and getting a 'head' is $\frac{50}{100}$, equivalent to 50% or $\frac{1}{2}$. At the same time we wouldn't be very surprised if we observed 49 or even 55 heads, although we would be surprised if we saw 99.

This definition of probability—which is the *frequentist* definition—is useful in many situations, but it has a few disadvantages. First, frequentist definitions aren't very helpful when dealing with unexpected situations. Frequentist probabilities are grounded in repeatable observations, and so understanding these repeatable frequencies is of limited use when trying to make predictions for changing or entirely novel systems. Second, most common frequentist approaches rely on specific assumptions. Using frequentist statistics requires trying to match a model that we have (often just in our heads) of some ecological system to a frequentist method that mostly closely matches the assumptions of our biological model. Given the complexity of ecological data and our uncertainty about the underlying model, frequentist approaches can be especially challenging in ecology. It would be nice to have a statistical way to build models that can propagate our (un)certainties about what we do—or especially, don't about how ecological systems work.

Luckily there is an alternative definition of probability—one that allows us to incorporate so-called *belief*, quantify general uncertainty, and—through the posterior distribution—can yield useful uncertainties about our inference: Bayesian probability. The Bayesian definition of probability is:

$$probability = \frac{likelihood \cdot prior}{normaliser} \tag{1}$$

Where *probability* is the Bayesian probability, *likelihood* is exactly the same as a frequentist likelihood, `prior` is your best-guess of the probability (your 'belief prior' to collecting data), and `normaliser` is a mathematical constant that makes sure our probability cannot go above 100% or below 0% (statisticians are lazy, and will not 'give 110%'). This mathematical constant [technically it is the probability of our data; $p(data)$] is a nuisance term that is extremely challenging to estimate (sometimes it is impossible!) and held back the practical application of Bayesian statistics for many years. Nowadays, with increases in computer power, we can use numerical methods such as 'Markov Chain Monte Carlo' (MCMC) methods to avoid having to estimate its value precisely. Such methods are iterative algorithms that involve chance at each step (that the process proceeds in iterations or steps, and only the last iteration affects the next, is why this is a 'Markov Chain'), tweaking and changing parameters within your model until a distribution of parameters consistent with the data are found. This distribution is called the 'posterior' distribution to distinguish how it is our view of probability *after* the data (via the likelihood) and the prior have been considered.

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 (Gelman *et al.*, 2014; Beaumont, 2019), Bayesian modeling options can feel limitless when compared to the models ecologists can fit in popular modeling packages (e.g. lme4). As long as you can write out the likelihood of your desired model (and sometimes even if you can't; Sunnåker *et al.*, 2013) and assign priors to all parameters, you can generally 'fit' the model. This includes non-linear ones, non-Gaussian families (e.g. 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, propagating 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 (e.g. counts, percent cover, etc.).

Fitting a bespoke model to data also yields numbers we often really want but don't have access to in other approaches. We perform experiments because we want to know how a treatment affects something of interest—how much slower do birds fly when wearing backpacks, or how much faster do plants grow with more warmth—but we often become more focused on whether the treatment was 'significant' or not. We lose track of whether birds flew 5% or 50% slower, and how consistent the effect was across individuals. But models can be designed to estimate and report effects per mg of backpack weight, or per °C of warming—always with estimated uncertainty. While replication crises in other fields, driven in part by a overly zealous focus on p-values (Halsey *et al.*, 2015; Loken & Gelman, 2017), and the rise of meta-analyses in ecology (Hampton *et al.*, 2013) 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, that can be difficult to connect to useful biological values), bespoke models take this to a new level. Researchers can easily estimate comparable effect sizes from z-scored data (in units of standard deviation), alongside estimates in meaningful natural units, such as per °C of warming or per hectare of habitat lost.

Further, using Bayesian methods makes it easy to report estimates of probabilities about coefficients (*e.g.*, "with anthropogenic warming, we are 80% certain the flowers open earlier") rather than null models (*e.g.*, "we are less than 5% certain these data came from a model where flowers aren't responding to anthropogenic warming"). Relatedly, frequentist tests are based on accepted level of Type I error rates—a particular α_{crit} (often 5%), which is often confused with statistical power (β, the probability that we would detect a true effect): when we define signifi-

cance as a less than 5% chance that some data came from a null hypothesis, it does not mean we have a 95% chance of detecting a true effect (indeed statistical power is often extremely low in ecological studies, Jennions & Møller, 2003). Bayesian statistics, by allowing us to compare the probabilities that *models* are correct, and not the probabilities that *data* are taken from a give model, allows us to test what we have always cared about: the biological processes themselves.

This valuable flexibility is often mentioned as 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. In ecological model fitting, we're currently most often interested in parameter estimates strongly informed by our data, making this problem sound especially dangerous. In reality, however, this problem is not related to modeling, but to experimental design—and a flawed experimental design leading to low power for your model is much easier to see through this workflow compared to using traditional null hypothesis testing methods. Further, such models are not as common as some may suggest. Perhaps more dangerous is fitting mis-specified models, where the model is not doing what we think, either due to coding errors or poor understanding of the model. Our workflow will help you avoid that.

Priors are another major source of concern for those new to Bayesian approaches. Often treated as the big bad wolf of Bayesian, or the unseen hand producing the model fits you get, according to some. They show up as half of the equation that gives you your model posterior, and philosophically, Bayesian was built around the idea that you have prior knowledge that you trust and want to compare to new data (with your new data showing up through your likelihood). In reality, few Bayesian analyses in ecology are approached this way.

How much priors influence your model fit is up to your model and your data. Depending on those two parts, the likelihood (influenced by your data) can easily overwhelm your priors (Fig. 2). Indeed, most work on the dangers of priors and 'prior misspecification' focuses on cases where you have very little data for the model you're trying to fit. If you try to fit a Bayesian model with 100 parameters and only 5 datapoints, then your priors will likely matter a lot. Priors can also matter when you have fewer parameters and lots of data but the data are not informative for the model; for example, if you fit a model including parameters for estimating high temperature responses, but lack data at those temperatures. Priors, however, can only matter more than you know when you fail to think through and check them.

These 'pitfalls' of Bayesian are not new, not necessarily specific to Bayesian methods (Low-Décarie *et al.*, 2014), 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. Building on this research, below we describe a generalizable workflow for Bayesian analysis and outline how it can revolutionize training in ecology.

A basic Bayesian workflow

We outline a workflow below that includes what we consider the major steps for Bayesian model fitting (Fig. 1). The workflow includes model calibration (Steps 1-2), inference (Step 3) and then model development (Step 4). Many of these steps will be familiar to statistical ecologists, but are often overlooked, whereas other steps may appear particular to Bayesian (e.g. prior predictive checks), but are actually useful for anyone—using Bayesian models or not—to challenge their models of how the world works. Parts of this workflow could be dropped, or expanded as

workflows in themselves, given other aims (see Supplement: Which workflow?).

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 also jump in, and somewhat over, one of the biggest steps.

We present the workflow assuming users have a model already in hand, which—if you have collected data—is true. Your model may be only verbal or conceptual; however, for this workflow, you'll need to convert such models into mathematical versions. This can be challenging at first, as ecologists often learn only a simplified version of this step (often focused on identifying which distribution—normal, binomial, zero-inflated Poisson, etc.—their response variable most closely resembles). Yet this challenge becomes easier with practice, and if we build an ecological community where this is de rigueur. This step is also easier when approached before you collect your data. After data collection, it becomes far more tempting to focus on the particular details in the data and not the latent processes and biological models from which the data were generating.

This workflow is particularly useful for models built before you've collected your data and can more fully motivate your model by your expected experimental design, which you can then refine it after you have the data. The best models usually include both the underlying biological model being studied and a model of the data generating process (e.g., gaps between sampling dates, biases in sampling etc.). We cannot emphasize enough that this aspect of our workflow is a valuable feature, not a troublesome bug: the scientist who makes explicit and confronts the assumptions of their model(s) is, all else being equal, the better scientist. Major advances are not made by guessing what data will look like and hoping two hypotheses are distinguishable once the data are collected, and this workflow ensures that the researcher is forced to confront, consider, and overcome as many of their assumptions as possible from the start.

Getting to the point where the Bayesian workflow is part of data design and collection, however, requires starting somewhere—with some model in hand. A suite of resources for 'generative' or 'narratively generative' modeling can help (MacElreath, 2016; Betancourt, 2021b), along with two points. First, you must start somewhere, so know that you can and will improve on this skill. Second, as you start, ask lots of questions—and push yourself on your answers—about what you expect and what's reasonable biologically from your model. For example, instead of simply identifying which distribution your response variable looks most similar to, ask yourself what generates that distribution and what you think its mean, minimum and maximum are. Do you expect data below zero? Up to infinity? If not, why not? You'll be generating your model—including its priors—as you do this. Effective model building is about efficient brainstorming and this is a critical part of the process (see Supplement).

Our workflow is explained mostly program-agnostically. Though at times 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 (Carpenter *et al.*, 2017). 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 (e.g. Metropolis-Hastings or Gibbs), but the basic workflow should apply to diverse implementations of Bayesian modeling, and can be extended to other approaches (frequentist, resampling etc.).

Step 1: Check your model code.
To start the workflow, you need to write up your model in a particular modeling language and check it. As with all code, just because it runs, does not mean it does what you think it does. 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 (e.g. rstanarm), you need a way to verify the code is correct: test data.

Test data (aka 'simulated data', or 'fake data,' etc.), and the skills required to build it, are central to this workflow. With 'test data' you simulate data from your model in such a way that you can use the resulting data to test your model is correct. This means that to build test data you need to understand your model well enough to generate data from known parameters; you then run that data through your model and confirm it returns the parameters (i.e., you fix values for your model parameters, then test how well your model recovers them, see the Supplement for an example). While there's no guarantee that inferences will always recover the parameter values you set, extreme disagreement is usually a good indicator that something is amiss. At the same time these simulation studies can help us understand how often our model might lead to the correct inference (see Fig. 2), and can be easily adapted into formal tests of power. As you do this, you will also be calibrating your model—seeing how close it estimates parameters you set and under what conditions.

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 before looking at any real data. '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 and sample size you want. Conversely, if you cannot complete this step, you'll struggle to understand if the model fit well, and struggle further to meaningfully interpret the model output, making this apparently simple programmatic step actually encapsulate a far deeper understanding of your model. If (and if not you need to revisit your code) your Stan output returns the parameters you expect from your test data, you can move onto interrogating your priors.

Step 2: Check your priors.

Assigning priors generally forces you to think about your model with regards to your study system, and interrogate what's probable, possible or actually unreasonable. While many packages (e.g., brms, rstanarm) will automatically set default priors, assigning them yourself (which you have to do if you write your own code) can quickly disabuse users of their prejudices. For example, you may not think you have a prior on how sunlight affects plant growth, until you realize your 'agnostic prior' actually allows plants to grow hundreds of meters per day.

You can take this a step up with prior predictive checks, which serve both as a check on the priors you're using, and to further explore the model of the world to which you're planning to fit your data. In prior predictive checks, you coefficients from your prior distribution and then explore how your model performs under those draws. Seeing how this influences your resulting output is critical: it reveals the extent to which your model can capture variation you know to be in your data, and gives insight into whether your model is capable of distinguishing among competing hypotheses. How exactly to do this depends on your question, model and aims, but many guides can help you think through this (Betancourt, 2021a; Wesner & Pomeranz, 2021; Winter & Depaoli, 2023).

By examining the consequences of your prior model you may suddenly realize that prior values that previously seemed reasonable lead to heavily unrealistic results when embedded in your full model, with all the other priors on parameters. You may find you have set up a model where certain effect sizes mean birds fly backwards when given heavy-enough backpacks. This means both that you may want to adjust your priors, but also gives important insight into the practical consequences of certain parameter configurations. Thus, if you see them in your actual model output you'll be more likely to realize your model is problematic (Step 4 will also help with this).

Step 3: Run your model on your empirical data.
The next step is to run the model—you've now validated, test-run and have ready to go—on your exciting new empirical data. Check diagnostics so you know it's running well and adjust until it is; this includes a suite of convergence and efficiency metrics ($\hat{R}$, ESS, lack of divergent transitions etc.) that are well-discussed elsewhere (and not our focus here, see instead Betancourt, 2020; Gelman *et al.*, 2020; van de Schoot *et al.*, 2021; Gabry *et al.*, 2019)

This is the step many ecologists skip straight to, ourselves included. It's easy to see the appeal: this is the inference step and is where you get the answer! 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 do not make immediate sense. After a while of this process, it can easily feel like we're not sure what we learned, if anything. And we can get distracted from what we are actually most interested in—the inference into biology. 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. We also highlight that, more often than not, a model that doesn't converge or seems to suggest coefficients that are completely at odds with the data, results from a model that was mis-specified (Step 1) or could never capture real-world variation anyway (Step 2).

Step 4: Posterior retrodictive checks
Once you have your posterior based on your model and new empirical data, it's time to remember that it's wrong ('all models are wrong...') and ask how useful it is. This step is where you define what a model needs to be useful and then check if it achieves that goal. Just because your model gave you estimates doesn't mean the model is adequate for the data or your particular aim, and—depending on why you fit the model—you may especially want to make sure it is reasonably predictive. You can do some of this through diagnostics, such as R^2, which compares compares point predictions to the observed data, but with a posterior you can compare an entire distribution of predictions to the observed data. This is where simulating from your model can be especially insightful. It will not only indicate that the model isn't adequately fitting the data but also can suggest what the problems might be. Steps 1-2 have set you up well for this, as you have a sense of what different parameter estimates do to the model, and test data provide a sense of how it works on data similar to yours.

Now that 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 posterior retrodictive checks (or posterior predictive checks, Fig. 3-4). Exactly how to do these are—again—dependent on your question, model and aims (but there is lots written on this, Held *et al.*, 2010; Gelman *et al.*, 2000; Conn *et al.*, 2018). In contrast to prior predictive checks, however, this step is built into some software. If you use rstanarm, then the package bayesplot will automatically give you a set of posterior retrodictive checks, including comparisons of the mean and variance of simulated datasets with those from your empirical data.

Often here you may find big differences from your empirical data, and can start to generate hypotheses for why. For example, you may find patterns that suggest missing grouping factors (e.g., site or biome) through visualization, and by grouping posteriors by that factor, or you may quickly realize your model predicts impossible numbers for your biological reality because of the distribution.

Feedbacks & workflows
With those hypotheses in hand, you may very well want to tweak your model—this is 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. This approach is different than the quest for a minimum adequate model or one 'best' fit. Feedbacks in this workflow are focused far more on what is biologically reasonable, and understanding the utility—and limits—of inference from your data for your model. 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 (e.g. recalling how to simulate y data for a simple linear regression is not straightforward when you never do it), but is immensely beneficial to forcing you to understand your model and its consequences. 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 changed how we approach statistical models.

This workflow represents just one of many possible workflows. It organizes a simplified set of steps for model calibration, inference and development—all implemented by simulating data from the model in various ways. In practice parts of this workflow could be dropped, or expanded as workflows in themselves, given other aims (see Supplement: Which workflow?) and in some cases you may skip steps or need to expand or adjust them.

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 advance, much as we believe they have helped our science advance.

Understanding nonidentifiability
Identifiability refers to when all parameters in a model can be uniquely identified with infinite data. More common in our experience is nonidentifiability. Models can be nonidentifiable in several ways, including when mathematically some parameters cannot be uniquely defined. Statistically, a kin of nonidentifiability—degeneracy—is often an outcome of the empirical data combined with the model. Degeneracy occurs when the data do not contain enough information to estimate one or more more parameters uniquely (Gelman & Hill, 2009).

Nonidentifiability and degeneracy can come up in many ways in ecology—and be hard to see, especially if you rush through model fitting. But if you have to write out your model and simulate

data, you may suddenly realize lots of places for nonidentifiability and degenercies to live. For example, when species do not occur across most sites, a model including separate parameters for site and species is often degenerate, but there's no warning in packages to tell you this. Thus you may never realize what experimental designs and sampling regimes are fundamentally too imprecise for your questions of interest. We have become far better at noticing nonidentifiability and degenercies based on this workflow—and we have adjusted how we collect data and interpret results because of it.

Know your limits
Once we noticed how pervasive nonidentifiability and degenercies were we started simplifying our models. Whereas previously when we had data that qualitatively appeared complexly nested, crossed, split or twisted, we would have initially 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 now work more carefully through what to include.

Before using this workflow, some of us would start with complex models, then simplify them until they converged in our given software package. Often these were hierarchical models with many levels—for example, including every column of site, plot, transect and quadrat in our dataframe, without stopping to check how well sampled they were, or what degeneracies they might introduce. These models we built were based more fear of missing an important interaction term or on a dogma of non-independence (and, somewhat relatedly, correct degrees of freedom) than an understanding of the model and the system's ecology. Our approach was driven far more by a weak set of statistical assumptions constrained by our software package, rather than by our ecological understanding or ultimate aims.

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, compared to fitting the main effects alone (the main effects then average over the interactions, see Gelman & Hill, 2020, for more details). This is sobering. It's more sobering when you see it played out again and again through this workflow.

We now both add complexity and simplify based on a more careful reckoning. Usually our 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 still may add and consider additional grouping factors and interactions, but we do so with a careful idea of how stable the model given the data likely is with them, and we rarely fit complex three-way interactions or similar—unless we have carefully designed the model and data collection for that aim. The ending model is often not as complex as we may have fit if we did not follow all the steps.

Understanding our model inference 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,' 'site', or 'species,' and suddenly how predictable a model is (or not) based on other factors like 'treatment' or 'time' becomes clear.

Looking at parameters, not p-values
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 example, for a phenological event (an observation of a biological event on a day within the calendar year so, ideally between 1 and 365, or 366 in a leap year) and suddenly find it's >1000 based on the parameters you put into your code.

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

How this workflow intersects with ecological training

This four-step workflow is a simplified version of the current best practices for Bayesian model fitting (Betancourt, 2020; van de Schoot *et al.*, 2021), 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 not 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. Given the increasingly computational toolkit of the modern ecologist, training in how to simulate data from models may be a smaller leap than decades ago.

A reasonably competent coder could easily simulate data under a complex model that they might not have the mathematical expertise to solve analytically, if doing so was part of their training and the workflows they regularly use. While simulation methods may appear foreign initially, they are usually much easier to implement than the analytical derivations of traditional methods so often seen in textbooks and classes. If this seems unlikely, consider whether most advisors would expect that their students could, if given time and support, simulate the data and model associated with a linear regression. We suggest a much lesser proportion of advisors would expect their students to be able to derive, from first principles and using linear algebra, the estimators required to fit such a linear model. Perhaps this separation stems from the fact that simulation approaches allow interactive learning, build intuition, and stress exploring a model in its relevant—ecological—context. Ecologist are much better at thinking about ecological problems than statistical ones, and grounding our approach in ecology will likely bring the best out of our statistical modelling.

Simulating data rapidly clarifies underlying assumptions. While training in frequentist methods often includes memorizing assumptions for a particular test, or steps specifically designed to test assumptions (e.g., normal quantile plots), this workflow requires no such training. Instead it requires only the skills to identify whatever the assumptions have been encoded in their models. As such it moves away from some modeling paradigms in ecology, which focus on fewer underlying assumptions, to building models were the assumptions are transparent and motivated by the specific domain expertise available in each application.

This workflow thus 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 a few, and these changes come alongside advances in Bayesian workflows, algorithms, and visualiziations (e.g. Betancourt, 2020; van de Schoot *et al.*, 2021; Gabry *et al.*, 2019), 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 (Gelman *et al.*, 2014; Gelman & Hill, 2020; Betancourt, 2021a), and training should reflect current best practices. These include that 'non-informative priors' are a misnomer, as they are often informative (Lemoine, 2019), and priors can easily be 'weakly informative.' Thus a strong focus on the dangers of priors in training can be overkill, and verges on scaremongering.

Current training often includes a very strong focus on mathematically-convenient 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. Prior predictive checks provide a far more powerful way to understand how priors work within a particular model, and are more useful than rules about which priors should be fit in certain cases or memorizing which priors are conjugate (Betancourt, 2021a).

(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 (Gelman & Hill, 2009). 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' (for example the 'block' effect from a randomized block design). After posterior retrodictive 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 (for example, you can predict for a new species or individual), whereas you cannot do so for most categorical 'fixed' effects.

(3) P-values, and null hypothesis testing in general, are easily misleading, 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 (Halsey *et al.*, 2015; Loken & Gelman, 2017). 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 to make p-values smaller (Halsey *et al.*, 2015; Colquhoun, 2017), nor is it Bayesian approaches, which (as we touch on above), bring their own ways to sift 'exciting results' from what is actually a pile of chafe.

The answer are workflows designed for careful model building, model fitting and model interrogation informed by ecological theory and understanding—including the one we outline here. This workflow depends strongly on simulating data—for testing your model (Step 1), checking your priors (Step 2) and understanding your model results (Step 4)—an area we actively undertrain in ecology. It makes clear how relevant and important simulation is, but the relevance of simulation extends well beyond Bayesian model fitting. As larger datasets and machine learning increase their utility, and prevalence, methods to test our understanding, interrogate our models and develop new models, will depend strongly on simulation, potentially transforming ecology, as it is transforming science in general (Flynn *et al.*, 2022; Kuntz & Wilson, 2022;

Oren *et al.*, 2017). The workflow we outline here thus applies to many other statistical realms, including machine learning, resampling and traditional null-hypothesis testing using frequentist approaches.

Acknowledgements: Comments from F. Baumgarten and D. Loughnan improved this manuscript.

1 References

Beaumont, M.A. (2019) Approximate bayesian computation. *ANNUAL REVIEW OF STATISTICS AND ITS APPLICATION, VOL 6* **6**, 379–403.

Betancourt, M. (2019) The convergence of markov chain monte carlo methods: From the metropolis method to hamiltonian monte carlo. *ANNALEN DER PHYSIK* **531**.

Betancourt, M. (2020) Towards a principled bayesian workflow. `https://betanalpha.github.io/assets/case_studies/principled_bayesian_workflow.html`.

Betancourt, M. (2021a) Prior modeling. https://betanalpha.github.io/assets/case$_s$*tudies/prior*$_m$*odeling.htmll*.

Betancourt, M. (2021b) (what's the probabilistic story) modeling glory? `https://betanalpha.github.io/assets/case_studies/generative_modeling.html`.

Burnham, K.P. & Anderson, D.R. (2004) Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research* **33**, 261–304.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M.A., Guo, J., Li, P. & Allen, R. (2017) Stan: A probabilistic programming language. *Journal of Statistical Software* **76**, 10.18637/jss.v076.i01.

Colquhoun, D. (2017) The reproducibility of research and the misinterpretation of p-values. *Royal Society Open Science* **4**.

Conn, P.B., Johnson, D.S., Williams, P.J., Melin, S.R. & Hooten, M.B. (2018) A guide to bayesian model checking for ecologists. *Ecological Monographs* **88**, 526–542.

Flynn, K.J., Torres, R., Irigoien, X. & Blackford, J.C. (2022) Plankton digital twins-a new research tool. *JOURNAL OF PLANKTON RESEARCH* **44**, 805–813.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M. & Gelman, A. (2019) Visualization in bayesian workflow. *Journal of the Royal Statistical Society Series a-Statistics in Society* **182**, 389–402.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. & Rubin, D.B. (2014) *Bayesian Data Analysis*. CRC Press, New York, 3rd edn.

Gelman, A., Goegebeur, Y., Tuerlinckx, F. & Van Mechelen, I. (2000) Diagnostic checks for discrete data regression models using posterior predictive simulations. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES C-APPLIED STATISTICS* **49**, 247–268.

Gelman, A. & Hill, J. (2009) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, New York.

Gelman, A. & Hill, J. (2020) *Regression and Other Stories*. Cambridge University Press.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C.C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.C. & Modrák, M. (2020) Bayesian workflow. arXiv.

Grinsztajn, L., Semenova, E., Margossian, C.C. & Riou, J. (2021) Bayesian workflow for disease transmission modeling in stan. *Statistics in Medicine* **40**, 6209–6234.

Halsey, L.G., Curran-Everett, D., Vowler, S.L. & Drummond, G.B. (2015) The fickle p value generates irreproducible results. *Nature Methods* **12**, 179–185.

Hampton, S.E., Strasser, C.A., Tewksbury, J.J., Gram, W.K., Budden, A.E., Batcheller, A.L., Duke, C.S. & Porter, J.H. (2013) Big data and the future of ecology. *Frontiers in Ecology and the Environment* **11**, 156–162.

Held, L., Schroedle, B. & Rue, H. (2010) Posterior and cross-validatory predictive checks: A comparison of mcmc and inla. *STATISTICAL MODELLING AND REGRESSION STRUCTURES: FESTSCHRIFT IN HONOUR OF LUDWIG FAHRMEIR* (eds. T. Kneib & G. Tutz), pp. 91–110.

Hoffman, M.D. & Gelman, A. (2014) The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *JOURNAL OF MACHINE LEARNING RESEARCH* **15**, 1593–1623.

Hurlbert, S.H. (1984) Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* **54**, 187–211.

Jennions, M.D. & Møller, A.P. (2003) A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral Ecology* **14**, 438–445.

Kuntz, D. & Wilson, A.K. (2022) Machine learning, artificial intelligence, and chemistry: how smart algorithms are reshaping simulation and the laboratory. *PURE AND APPLIED CHEMISTRY* **94**, 1019–1054.

Lemoine, N.P. (2019) Moving beyond noninformative priors: why and how to choose weakly informative priors in bayesian analyses. *Oikos* **128**, 912–928.

Loken, E. & Gelman, A. (2017) Measurement error and the replication crisis. *Science* **355**, 584–585.

Low-Décarie, E., Chivers, C. & Granados, M. (2014) Rising complexity and falling explanatory power in ecology. *Frontiers in Ecology and the Environment* **12**, 412–418.

MacElreath, R. (2016) *Statistical Rethinking*, vol. 469 pp. CRC Press, New York.

Muff, S., Nilsen, E.B., O'Hara, R.B. & Nater, C.R. (2022) Rewriting results sections in the language of evidence. *Trends in ecology & evolution* **37**, 203–210.

Muthukumarana, S., Schwarz, C.J. & Swartz, T.B. (2008) Bayesian analysis of mark-recapture data with travel time-dependent survival probabilities. *CANADIAN JOURNAL OF STATISTICS-REVUE CANADIENNE DE STATISTIQUE* **36**, 5–21.

Oren, T., Turnitsa, C., Mittal, S. & Diallo, S.Y. (2017) Simulation-based learning and education. *GUIDE TO SIMULATION-BASED DISCIPLINES: ADVANCING OUR COMPUTATIONAL FUTURE* (eds. S. Mittal, U. Durak & T. Oren), Simulation Foundations Methods and Applications, pp. 293–314.

Schad, D.J., Betancourt, M. & Vasishth, S. (2021) Toward a principled bayesian workflow in cognitive science. *PSYCHOLOGICAL METHODS* **26**, 103–126.

Strinella, E., Scridel, D., Brambilla, M., Schano, C. & Korner-Nievergelt, F. (2020) Potential sex-dependent effects of weather on apparent survival of a high-elevation specialist. *Scientific Reports* **10**, 8386.

Sunnåker, M., Busetto, A.G., Numminen, E., Corander, J., Foll, M. & Dessimoz, C. (2013) Approximate bayesian computation. *PLoS computational biology* **9**, e1002803.

Trijoulet, V., Holmes, S.J. & Cook, R.M. (2018) Grey seal predation mortality on three depleted stocks in the west of scotland: What are the implications for stock assessments? *Canadian Journal of Fisheries and Aquatic Sciences* **75**, 723–732.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Maertens, K., Tadesse, M.C., Vannucci, M., Gelman, A., Veen, D., Willemsen, J. & Yau, C. (2021) Bayesian statistics and modelling. *Nature Reviews Methods Primers* **1**.

Wesner, J.S. & Pomeranz, J.P.F. (2021) Choosing priors in bayesian ecological models by simulating from the prior predictive distribution. *ECOSPHERE* **12**.

Winter, S.D.D. & Depaoli, S. (2023) Illustrating the value of prior predictive checking for bayesian structural equation modeling. *STRUCTURAL EQUATION MODELING-A MULTIDISCIPLINARY JOURNAL* .

Zheng, C., Ovaskainen, O., Saastamoinen, M. & Hanski, I. (2007) Age-dependent survival analyzed with bayesian models of mark-recapture data. *ECOLOGY* **88**, 1970–1976.

2 Glossary

Table 1: A set of major terms used and simplified definitions, organized alphabetically.

Term	*Definition*
calibration	analyzing how often an estimate is close to the true value over an ensemble of hypothetical observations; this requires knowing the true value, which we don't for our data—but we can calibrate models we plan to fit to our data (*Steps 1-2*) so we understand the models better, including their limits given data similar to ours. We emphasize simulations to calibrate model behaviors consistent with our ecological systems and understanding (e.g., working within a limited set of parameter ranges through prior predictive checks). In contrast to this approach, frequentist method are calibrated against all possible behaviors, which is not only impractical in most circumstances it's also irrelevant given that the most extreme behaviors are unlikely to manifest in reality.
degeneracy	complex uncertainties that come from a mix of sources, including, non-identified models and cases where the data cannot well inform model parameters. When the data are not informing the parameters that we care about, this highlights a measurement issue. Identifying these problems in simulation studies, can highlight when we need a better experimental design (e.g., sampling for more overlapping species across sites, or changing what we measure, etc.).
non-identifiability	when all parameters in a model cannot be uniquely identified with infinite data
posterior	product of the likelihood and prior
prior	an distribution of reasonable values for a parameter based on fundamental biological and ecological understanding, previous research, or other sources
statistical model	Approximations of observed phenomena via a mathematical equation, with parameters estimated from data using some fitting method. In this article, we often simplify to 'model.' See also the Supplement: What's a model?
workflow	a set of steps to achieve a goal, with those steps designed to help organize the process, and ideally make it more systematic

3 Figures (need work)

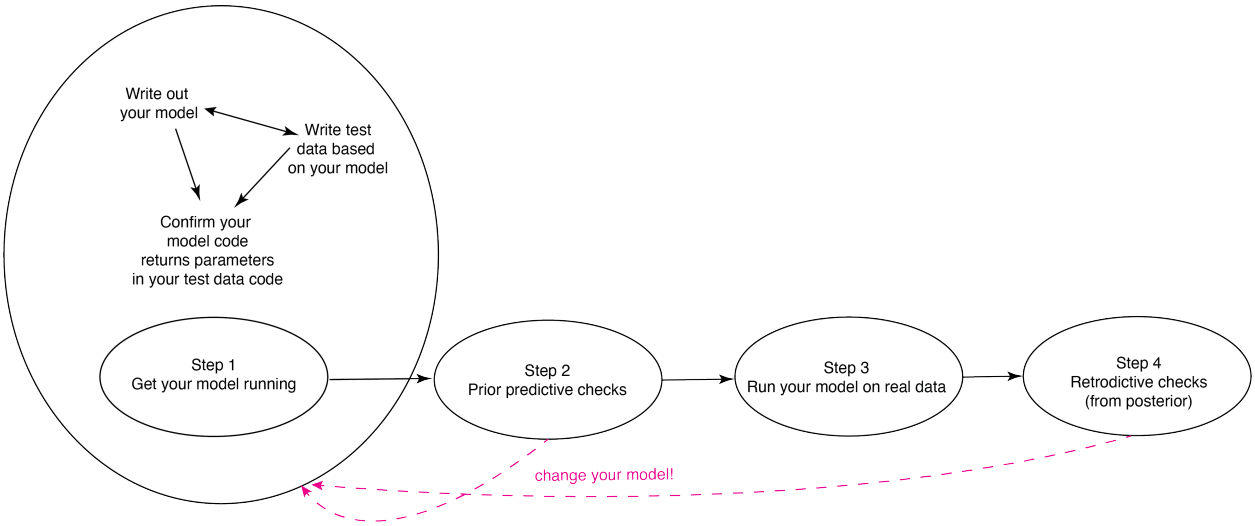

Figure 1: A very basic workflow for Bayesian model fitting includes four major steps with potential feedbacks (pink dashed arrows) and begins with testing your model through simulated (test) data.

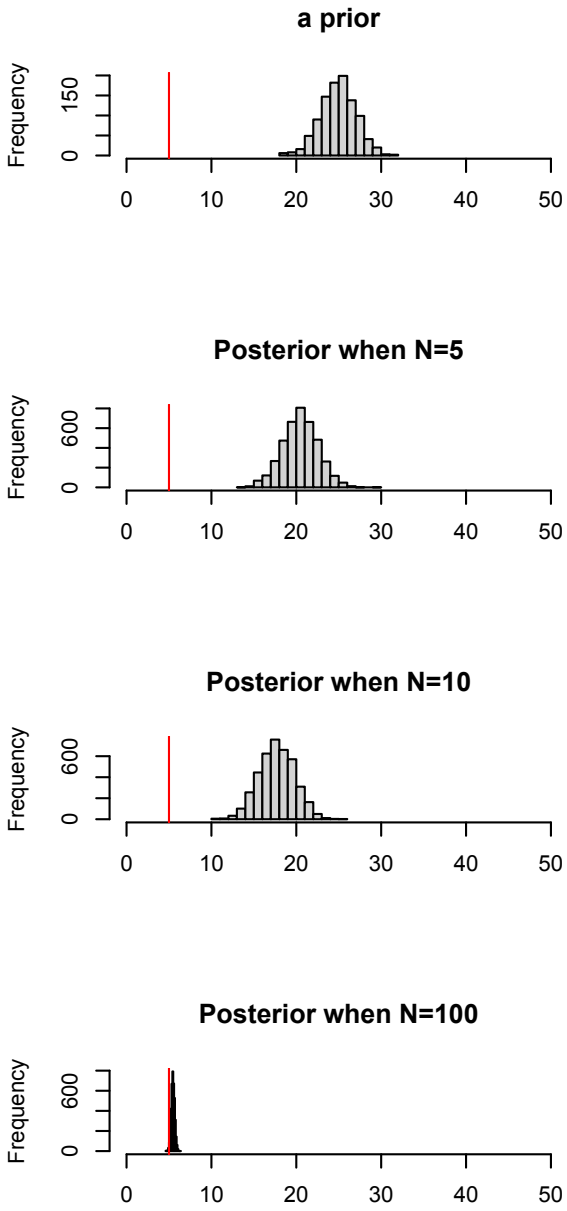

Figure 2: A simple example of how we can use simulated data to understand calibration issues in a simple mis-specified model example. Here we know the true model underlying the data is $y = \alpha + \text{normal}(0, \sigma)$ where α is 5 (shown as red vertical line) and σ is 2. The model, however, is mis-specified by a prior for α of $\text{normal}(25, 2)$ (in our experience, it is quite rare to have a prior informed by ecological knowledge be so far off, but this is an example). How mis-calibrated the model will be depends on the data: we show examples with a sample size (N) of 5, 10 and 100.

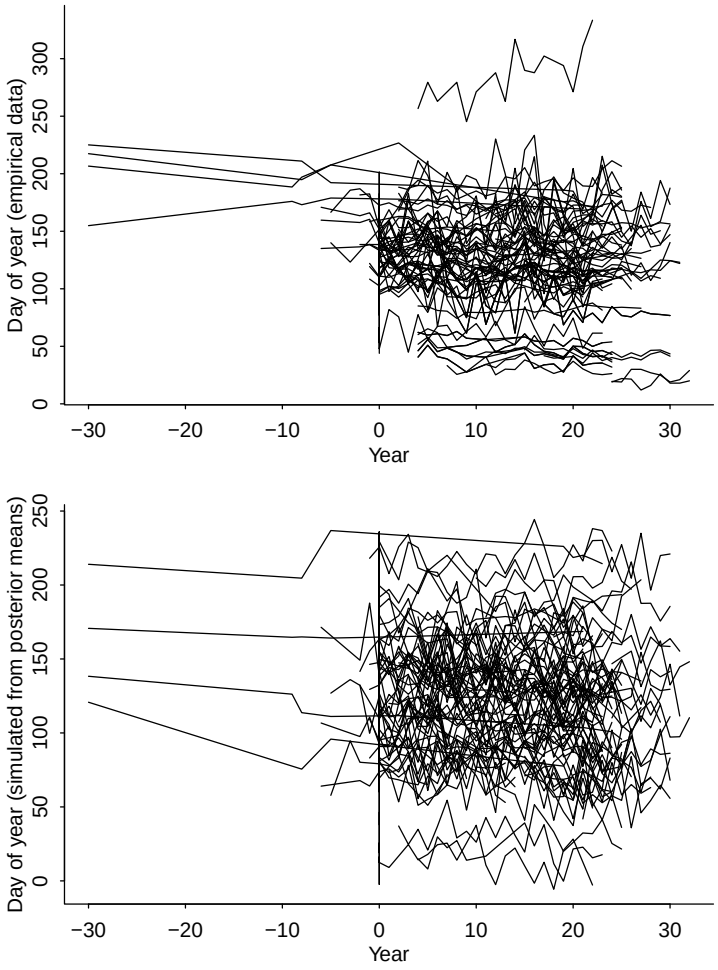

Figure 3: Example of a single retrodictive check from time-series data of phenological events over time. The raw data (top) looks similar to one simulated dataset (bottom), based on existing species number, their respective x data, and simulating from the parameters for each species. More predictive checks based on repeated simulations from the posterior, however, suggested issues in the model (see Fig. 4). See 'An example workflow' in the Supplement for more details.

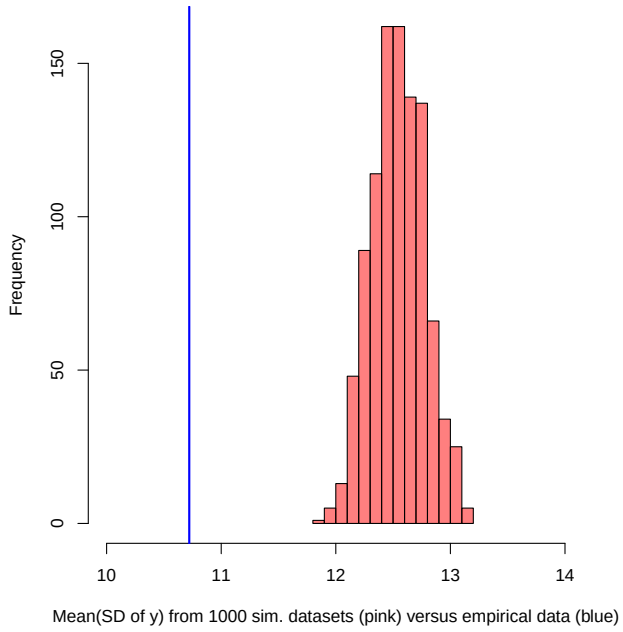

Figure 4: Example of a retrodictive check, averaging across 1000 simulations (pink histogram), for a model of time-series data of phenological events over time, showed the model consistently over-predicted variance compared to the empirical data (blue line SD). See 'An example workflow' in the Supplement for more details.