A four-step simulation-based workflow for ecological analysis and science

EM Wolkovich[1*], T Jonathan Davies[1,2], William D Pearse[3,4] & Michael Betancourt[5]

November 9, 2025

[1] Forest and Conservation Sciences, University of British Columbia, Vancouver, BC V6T 1Z4, Canada

[2] Botany, University of British Columbia, Vancouver, BC V6T 1Z4, Canada

[3] Department of Life Sciences, Imperial College London, Ascot SL5 7PY, United Kingdom

[4] Alan Turing Institute, British Library, 96 Euston Road, London NW1 2DB, United Kingdom

[5] Symplectomorphic, LLC, New York, NY 10026, USA

[*] mailto: e.wolkovich@ubc.ca

Keywords: big data; scientific workflow; data simulation; prediction; null hypothesis testing, machine learning

Abstract

Ecology is a discipline that has faced increasing challenges as the disconnect between its scientific and statistical methods has become more obvious. Growing demands for useful forecasts in an era of intensifying global change requires models that can capture the variability and underlying uncertainty of ecological systems and related data. Yet many ecologists are not trained in current methods to build the flexible robust models needed to address this challenge. Thus, there is often a reliance on a limited set of pre-defined models combined with null hypothesis testing or a temptation to adopt new approaches without fully understanding their limitations. The result is poor models that lead to incorrect predictions, alongside concerns of a looming replication crisis. Here we show how new advances in workflows can lead to better models and enhance training in ecology. Building on the increasingly computational toolkit of many ecologists, this approach leverages simulation to integrate model building and testing of empirical data more fully with ecological theory. We argue this approach can fit models that are more robust and better-suited to providing new ecological insights and improved predictions, and may provide a blueprint for other fields similarly challenged by complex systems, growing datasets and limited training in how to best approach them.

Examples & Data: We provide an example of the full workflow with complete code available at https://github.com/lizzieinvancouver/bayesianflowsexample. The data used for the example is provided and full metadata on it is available via the Knowledge Network for Biocomplexity: `doi:10.5063/F12J69B2`. All other examples use simulated data.

Introduction

Ecology is a discipline that has followed a similar trajectory to other biological and social science fields in its growing aims, study design and data collection. Historically, ecological research commonly reported results from experiments or observations of one site using frequentist test statistics associated with null hypothesis testing. Mechanistic models were typically the domain of theorists, who studied ordinary differential equations (ODE, for example of Lotka Volterra predator-prey dynamics or the relationship between stability and complexity) and rarely compared their models explicitly to empirical data (see Caswell, 1988). But in recent decades growing demands have pushed ecological science to work across systems, contribute to policy and provide models to forecast the outcomes of growing anthropogenic pressures (Hák et al., 2016; Lindenmayer and Likens, 2010). These new scales and aims have led some ecologists to try to adapt what they were trained in (e.g., null hypothesis testing using F and t tests) to increasingly larger scales, more complex datasets, and to fads in specific approaches (Grace et al., 2025).

Yet many commonly used statistical approaches do not align with these new demands. 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 independence assumptions), they will usually fail to produce robust, reproducible results. For example, an overly zealous focus on p-values has led to a replication crises in several fields, where results derived from studies with small sample sizes seem most likely the outcome of noisy data combined with a search for statistical significance through many models (effectively a garden of forking paths, Halsey et al., 2015; Loken and Gelman, 2017). Some model selection approaches, including new machine learning methods, try to avoid this by comparing across models, but may not generalize to provide useful forecasts. This is especially true when forecasts have to adapt to changes in the underlying biology (Boettiger, 2022). This leaves ecology in a predicament shared across other fields—concerns of a looming replication crisis (Filazzola and Cahill Jr, 2021; Fraser et al., 2020) and overly confident forecasts with the potential to erode public trust in science when found not to be accurate (Leroux, 2019; Boettiger, 2022).

Many researchers recognize these issues and have turned to methods better designed for forecasting from complex systems and messy data. Machine learning methods, which benefit from large datasets—often with many predictors—fitted to test and training datasets (Breiman, 2001), have revolutionized image classification in ecology and remote sensing, but are increasingly used to forecast ecological processes (e.g., Zwart et al., 2023). These methods often build complex, opaque ('black-box') models (Cox, 2001; Efron, 2001; Shmueli, 2010), thus providing opaque inference into ecological processes. Bayesian inference encourages the fitting of bespoke mechanistic models with interpretable parameters and can thus handle many of the complexities of ecological data (Hobbs and Hilborn, 2006). Long used in select applications in ecology (Muthukumarana et al., 2008; Zheng et al., 2007; Trijoulet et al., 2018; Strinella et al., 2020), Bayesian inference has recently increased in use (Anderson et al., 2021; Pichler and Hartig, 2023), alongside new algorithms (e.g. Hamiltonian Monte Carlo, Hoffman and Gelman, 2014; Betancourt, 2019) that have made fitting and implementing models faster, more robust and—in many ways—easier (Carpenter et al., 2017)

Regardless of the approach, fitting larger and more complex models has not dramatically improved forecasts, nor made research more reproducible. Instead, these new approaches have highlighted a fundamental training disconnect that applies from simple to complex models: treating scientific and statistical methods as separate.

Merging scientific and statistical training is possible 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 (Box, 1976). While these approaches are slowly gaining traction in other fields (e.g., Esfahani et al., 2021; Schad et al., 2021; Bouman et al., 2024), they are not widely used in ecology and related fields today. Such 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. We argue that potential benefits include not only a better understanding of models fit to empirical data, but also a better understanding of system dynamics by requiring explicit consideration of the generative processes underlying observations (Grace et al., 2025).

Here we outline a simplified—but powerful—workflow that builds on new insights from statistics (Betancourt, 2020; Gelman et al., 2020; van de Schoot et al., 2021) and the increasingly computational nature of ecology today. Our aim is to provide an approachable rubric for those new to fitting complex models or simply those interested in re-considering their current workflow (and is not intended to be a comprehensive overview; see 'Next steps' in the Supplement). Because of this aim and to maximize interpretability, we illustrate our workflow using examples of simple models, and suggest additional resources as users build more complex models. Our examples include several statistical inference methods, though we focus on implementing the workflow through a Bayesian statistical framework (with an example shown in R and Stan), because this framework allows integrating bespoke model building more fully with ecological theory and understanding. We suggest that adopting this workflow approach can help fit models that are more robust and well-suited to provide new ecological insights—allowing us to refine where to put resources for better estimates, better models, and better forecasts.

A four-step workflow

Our workflow outlines what we consider the major steps for building bespoke models (Fig. 1). Several of these steps will be familiar to statistical ecologists, but are often overlooked, whereas other steps may appear particular to certain methods (e.g. prior predictive checks in Bayesian analyses), but are actually useful for anyone—using Bayesian models or not—to challenge their models of how the world works. We find that it is easiest to illustrate and describe this workflow using a Bayesian framework (see *A brief review of statistical inference using Bayesian approaches* in the Supplement), but we argue this workflow can be adapted to other approaches (Fig. 2). Parts of this workflow could be expanded as workflows in themselves, given other aims (see Supplement: *Which workflow?*).

Step 1: Develop your model(s)

We start the workflow with what can feel like the biggest step—build a model (or potentially, models) based on your aims (Hilborn and Mangel, 2013). By developing a model designed for your biological question, data and aims, your statistical workflow naturally becomes a scientific workflow. You will more clearly see the assumptions and mechanisms in your model, which is especially valuable given how often our intuition of how models 'work' is wrong (Kokko, 2005). You likely already have a model, though it may be only verbal or conceptual. For this workflow, however, you'll need to convert such models into mathematical versions (Servedio et al., 2014).

Though it can feel challenging at first, this step is best approached before you collect any data. A suite of resources for 'generative' or 'narratively generative' modeling can help (McElreath, 2016; Betancourt, 2021*b*). 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. As you do this, you'll be generating your model—including its priors. Priors are important for Bayesian analysis, but the basic idea of them—coming up with a distribution of reasonable values for parameters in your model (see Table 1)—is useful to all analyses (for an example, see discussion of a heuristic model in the case study for Fig. 2b). Assigning priors generally forces you to think about your model with regard to your study system, and interrogate what's probable, possible or actually unreasonable—and can quickly disabuse users of prejudices regarding priors. 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.

Step 2: Check your model on simulated data

Once you have your model and its priors jotted down, you need to formalise it in your preferred modeling language and check it. As with all code: just because it runs, does not mean it does what you think it does. The worst errors often still permit code to run.

Test data (aka 'simulated data', or 'fake data,' etc.), and the skills required to generate 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 if your model code is correct (i.e., you fix values for your model parameters, then test how well your model recovers them, see the Supplement for several examples). This is more straightforward when your statistical model is the same as your generative model, but the basic idea can be adapted to other approaches (see Fig. 2b). While there's no guarantee that inferences will always recover the parameter values you set, even when using the correct model, extreme disagreement is often an indicator that something is amiss in the implementation of the model. At the same time these simulation studies can help understand how often a model might lead to the correct inference (see Figs. 2 and S1). As you do this, you will also be calibrating your model—seeing how accurately and precisely it estimates parameters 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 (see Fig. 2a and related Supplemental examples), 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. Thus, this apparently simple programmatic task actually encapsulates a far deeper understanding of your model. '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.

You can learn only so much, however, from data simulated from a particular parameter set. Simulation studies across multiple parameter sets allow you to investigate how robust your inferential performance might be. Prior predictive checks (Betancourt, 2021*a*; Winter and Depaoli, 2023) use the Bayesian prior model to set the scope of such simulations, but the basic idea of prior predictive checks can used in any analysis. For these, you draw values from your prior distribution and then explore how your model performs. Seeing how this influences your resulting model output reveals the extent to which your model can capture known variation in your data, and gives insight into whether your model is capable of distinguishing among competing hypotheses. If adopting a Bayesian approach, it can also serve as a check on the priors you're using (addressing one of the common concerns of those inexperienced with Bayesian models).

Step 3: Run your model on your empirical data

The next step is to run the model—you've now evaluated, 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. Which diagnostics to use depends on your exact fitting approach, with many approaches having a suite of metrics that are are well-discussed elsewhere (for Bayesian methods, this includes a suite of convergence and efficiency metrics 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 where you might gain new ecological insights. Fitting new data to the model can feel like the moment when you'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. And with the excitement of getting a model to run we can get distracted from what we are actually most interested in—the inference into our ecological system.

Following this workflow can make this step much more satisfying. Here the benefits of the workflow may become especially apparent: using it with Bayesian inference you likely have estimates in useful units with uncertainty you can understand. You can use this information to draw new conclusions, design new experiments and more—but this is also a point to stop and check your model.

Step 4: Check your model on data simulated from your empirical model output

Once you have your parameter estimates based on your model and new empirical data, it's time to remember that your model is wrong (as all models are) and ask how useful it is. You can do some of this through common model-fit diagnostics, such as R^2, which compares point predictions to the observed data. With a Bayesian posterior (see Table 1), however, you have an added benefit in that 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 when the model isn't adequately fitting the data but also can suggest what the problems might be. Using the parameter estimates from your fitted model to simulate new data (Held et al., 2010; Gelman et al., 2000; Conn et al., 2018) lets you see how that new world compares to the observed data. This is most easily done in a Bayesian framework—called posterior retrodictive checks or posterior predictive checks (Fig. 3)—where your posterior captures your uncertainty in a useful way, but can be done with estimates of your parameters and their uncertainty from other inferential frameworks. Exactly how to do this effectively, however, requires care for any particular framework. Tailoring these checks to the research question and model makes this step most likely to pick up model mis-specification and provide useful insight for improvement (e.g., Mimno et al., 2015).

Often here you may find large 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 visual posterior retrodictive checks, or you may quickly realize your model predicts impossible numbers for your biological reality. You may begin to see inadequacies in your model, or even potentially your data. This is one of the main benefits of the workflow: models don't fail silently, they fail with a wealth of context that helps to generate new models and experiments.

Feedbacks & workflows

A key feature of this workflow is that it can be iterated. If you find that you want to tweak your model then you return to the beginning, adjust your model, and repeat the rest of the workflow (Fig. 1). In this way, fitting multiple models is encouraged, but this is distinct from 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.

How this workflow changed our science

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 (Step 2). For example, when modeling phenological events (observations of biological events on numbered days within the calendar year: 1-366 days when including leap years) it is not uncommon to find seemingly-reasonable models generating predictions of events on non-existent calendar days beyond 366.

A closer inspection of our parameters also taught us a lot about identifiability and nonidentifiability, when all parameters in a model can—or cannot—be uniquely identified with infinite data, and a statistical kin: degeneracy (see Table 1). Degeneracy concerns the kinds of complex uncertainties that can arise from finite data sets (Gelman and Hill, 2009), and something we have often found in Steps 2-3 of our workflow. Nonidentifiability and degeneracy can insert themselves in many ways in ecology, and may lead us to believe we understand our system when we do not. These were issues we never thought about before using this workflow, but since then we have realized (especially in steps 1-2) lots of places for nonidentifiability and degeneracies to live—and we have adjusted how we collect data and interpret results because of it. For example, we have found fitting both site and species in a model with highly imbalanced data or trying to estimate interaction terms with low sample sizes (for more details see Gelman and Hill, 2020) leads to degenerate models, while spatial autocorrelation in environmental data can often lead to issues of nonidentifiability, but there's often no warning in common statistical packages to tell us of these problems (see Fig. 2c).

How this workflow intersects with ecological training

This four-step workflow is a simplified version of the current best practices for 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 to complete Steps 1-2 leans on 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 easily simulate data from. This dichotomy seems short-sighted in our current era of bigger, messier data and a greater diversity of methods available to handle such data. The increasingly computational toolkit of the modern ecologist makes it easier to bridge the gap between statistical models used in ecology and the field's underlying mechanistic theories.

We argue training in simulating data as part of an organized workflow could speed progress in ecology and is possible given the increasingly computational abilities of many ecologists. A reasonably competent coder could easily simulate data under a complex model that they might not have the mathematical expertise to solve analytically (e.g., solving for an equilibrium in an ODE)—if doing so was part of their training and the workflows they regularly use. While training in frequentist methods often includes memorizing assumptions for a particular test, or steps specifically designed to test particular 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 your models. It thus moves away from some modeling paradigms in ecology, which focus on fewer underlying assumptions (e.g. random forests, non-parametric), to building models were the assumptions are transparent and motivated by the specific domain expertise of researchers.

In Bayesian inference advances in developing workflows have come alongside improved algorithms, visualizations (e.g. Betancourt, 2020; van de Schoot et al., 2021; Gabry et al., 2019), perspectives on priors (Gelman et al., 2014; Gelman and Hill, 2020; Betancourt, 2021*a*) and hierarchical approaches that could also improve training. For example, new work shows that prior predictive checks provide a more powerful and intuitive way to understand how priors work within a particular Bayesian model (Betancourt, 2021*a*), compared to past approaches. Similarly, traditional ecological training in hierarchical models still often refers to grouping factors (such as species or individual) as 'random effects,' which is misleading, imprecise and thus no longer recommended (Gelman and Hill, 2009). In ecology, it also carries with it many 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). Training in retrodictive checks (Step 4) may reshape these views, as hierarchical effects are (by definition) drawn from an underlying distribution—meaning they can predict outside of the specific set sampled (for example, to predict for a new species or individual), whereas the same is not true for most categorical 'fixed' effects.

How this workflow extends to other fields

These new best practices have gained traction at the same time that ecology, alongside many other fields, has recognized that *p*-values, and null hypothesis testing in general, are easily misleading (Gelman and Geurts, 2017; Ferraro and Shukla, 2020; Filazzola and Cahill Jr, 2021; Fraser et al., 2020). Small sample sizes alongside a tendency to fit complicated models with multiple interactions makes ecological research particularly vulnerable to these problems (Gelman, 2015). Adding to this, a lack of routine reporting of interpretable effect sizes, fitting of many models without adequate explanation (or reporting), and poor data and code recording habits all increase the chance of finding 'significance' at a level of ≤ 0.05 (Halsey et al., 2015; Loken and Gelman, 2017).

The answer to these problems is not to make *p*-values smaller (Halsey et al., 2015; Colquhoun, 2017), nor is it Bayesian, machine learning or 'new' causal inference approaches, despite assertions to the contrary, which echo previous promised revolutions through the introduction of new methods (e.g., Mitchell, 1992; Burnham and Anderson, 2004; Byrnes and Dee, 2025). Ecology, like many fields, has increasingly adopted machine learning methods in hopes they will help them fit better models, but they can easily lead to poor models that do not match the underlying realities of the system (Efron, 2020; Pichler and Hartig, 2023). Similarly ecology readily took up path analysis, multi-model comparison with AIC, and a suite of other approaches, that promised better inference, but ultimately led to many papers reporting poor models, and resulting policy recommendations based on such models (Petraitis et al., 1996; Leroux, 2019). This fad approach to statistics is not unique to ecology, but the cure for it is also not yet another new statistical method.

We argue that the answer is training in workflows designed for careful model building, model fitting and model interrogation informed by underlying theory and understanding of the system being modeled (Betancourt, 2020; Gelman et al., 2020; van de Schoot et al., 2021)—including the one we outline here. Our workflow depends strongly on simulating data—for testing your model (Step 2), and understanding your model results (Step 4)—an area we actively under-train in many research fields that depend on increasingly complex statistical methods. Simulation approaches encourage interactive learning, build intuition, and stress exploring a model in its relevant context. Ecologists, similar to researchers in any domain-specific field, are much better at thinking about domain-specific scientific problems than statistical ones. Grounding statistical approaches in theory and domain knowledge will likely bring the best out of statistical modeling. While this idea is not new, we argue the need for it is especially high, as the line between estimation and prediction becomes more blurred (Shmueli, 2010). At the same time, however, computation is increasingly part of a researcher's toolkit, lowering the barriers for those wishing to adopt this workflow and improve their statistical inference.

Acknowledgements: Comments from F. Baumgarten, B. Bolker, D. Loughnan, N. Pates, V. Rudolf, J. Socolar and V. Van der Meersch improved this manuscript and J. Ngo improved Figure 1. Thanks to V. Van der Meersch also for helping with Markdown formatting.

References

Anderson, S. C., P. R. Elsen, B. B. Hughes, R. K. Tonietto, M. C. Bletz, D. A. Gill, M. A. Holgerson, S. E. Kuebbing, C. McDonough MacKenzie, M. H. Meek, et al. 2021. Trends in ecology and conservation over eight decades. Frontiers in Ecology and the Environment 19:274–282.

Betancourt, M. 2019. The Convergence of Markov Chain Monte Carlo Methods: From the Metropolis Method to Hamiltonian Monte Carlo. Annalen der physik 531.

———. 2020. Towards A Principled Bayesian Workflow. https://betanalpha.github.io/assets/case$_studies/principled_bayesian_workflow.html.$

———. 2021a. Prior modeling. https://betanalpha.github.io/assets/case$_studies/prior_modeling.htmll.$

———. 2021b. (what's the probabilistic story) modeling glory? https://betanalpha.github.io/assets/case$_studies/generative_modeling.html.$

Boettiger, C. 2022. The forecast trap. Ecology Letters 25:1655–1664.

Bouman, J. A., A. Hauser, S. L. Grimm, M. Wohlfender, S. Bhatt, E. Semenova, A. Gelman, C. L. Althaus, and J. Riou. 2024. Bayesian workflow for time-varying transmission in stratified compartmental infectious disease transmission models. PLoS computational biology 20:e1011575.

Box, G. E. 1976. Science and statistics. Journal of the American Statistical Association pages 791–799.

Breiman, L. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science 16:199–231.

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

Byrnes, J. E., and L. E. Dee. 2025. Causal inference with observational data and unobserved confounding variables. Ecology Letters 28:e70023.

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

Caswell, H. 1988. Theory and models in ecology: a different perspective. Ecological Modelling 43:33–44.

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

Conn, P. B., D. S. Johnson, P. J. Williams, S. R. Melin, and M. B. Hooten. 2018. A guide to Bayesian model checking for ecologists. Ecological Monographs 88:526–542.

Cox, D. R. 2001. Comment on statistical modeling: The two cultures. Statistical Science 16:216–218.

Efron, B. 2001. Comment on statistical modeling: The two cultures. Statistical Science 16:218–219.

———. 2020. Prediction, estimation, and attribution. International Statistical Review 88:S28–S59.

Esfahani, A. A., M. Betancourt, Z. Bogorad, S. Böser, N. Buzinsky, R. Cervantes, C. Claessens, L. De Viveiros, M. Fertl, J. Formaggio, et al. 2021. Bayesian analysis of a future β decay experiment's sensitivity to neutrino mass scale and ordering. Physical Review C 103:065501.

Ferraro, P. J., and P. Shukla. 2020. Feature—is a replicability crisis on the horizon for environmental and resource economics? Review of Environmental Economics and Policy .

Filazzola, A., and J. F. Cahill Jr. 2021. Replication in field ecology: Identifying challenges and proposing solutions. Methods in Ecology and Evolution 12:1780–1792.

Fraser, H., A. Barnett, T. H. Parker, and F. Fidler. 2020. The role of replication studies in ecology. Ecology and Evolution 10:5197–5207.

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

Gelman, A. 2015. The connection between varying treatment effects and the crisis of unreplicable research: A bayesian perspective.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2014. Bayesian Data Analysis. 3rd ed. CRC Press, New York.

Gelman, A., and H. M. Geurts. 2017. The statistical crisis in science: How is it relevant to clinical neuropsychology? The Clinical Neuropsychologist 31:1000–1014.

Gelman, A., Y. Goegebeur, F. Tuerlinckx, and I. Van Mechelen. 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., and J. Hill. 2009. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge, New York.

———. 2020. Regression and Other Stories. Cambridge University Press.

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

Grace, J. B., N. Huntington-Klein, E. W. Schweiger, M. Martinez, M. J. Osland, L. C. Feher, G. R. Guntenspergen, and K. M. Thorne. 2025. Causal effects versus causal mechanisms: Two traditions with different requirements and contributions towards causal understanding. Ecology letters 28:e70029.

Grinsztajn, L., E. Semenova, C. C. Margossian, and J. Riou. 2021. Bayesian workflow for disease transmission modeling in Stan. Statistics in Medicine 40:6209–6234.

Hák, T., S. Janoušková, and B. Moldan. 2016. Sustainable development goals: A need for relevant indicators. Ecological indicators 60:565–573.

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

Held, L., B. Schroedle, and H. Rue. 2010. Posterior and Cross-validatory Predictive Checks: A Comparison of MCMC and INLA. Pages 91–110 *in* T. Kneib and G. Tutz, eds. Statistical Modelling and Regression Structures: Festschrift in Honour of Ludwig Fahrmeir.

Hilborn, R., and M. Mangel. 2013. The ecological detective: confronting models with data (MPB-28). Princeton University Press.

Hobbs, N. T., and R. Hilborn. 2006. Alternatives to statistical hypothesis testing in ecology: a guide to self teaching. Ecological Applications 16:5–19.

Hoffman, M. D., and A. Gelman. 2014. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research 15:1593–1623.

Kokko, H. 2005. Useful ways of being wrong. Journal of evolutionary biology 18:1155–1157.

Leroux, S. J. 2019. On the prevalence of uninformative parameters in statistical models applying model selection in applied ecology. PloS one 14:e0206711.

Lindenmayer, D. B., and G. E. Likens. 2010. The science and application of ecological monitoring. Biological conservation 143:1317–1328.

Loken, E., and A. Gelman. 2017. Measurement error and the replication crisis. Science 355:584–585.

McElreath, R. 2016. Statistical Rethinking, vol. 469 pp. CRC Press, New York.

Mimno, D., D. M. Blei, and B. E. Engelhardt. 2015. Posterior predictive checks to quantify lack-of-fit in admixture models of latent population structure. Proceedings of the National Academy of Sciences 112:E3441–E3450.

Mitchell, R. J. 1992. Testing evolutionary and ecological hypotheses using path analysis and structural equation modelling. Functional Ecology pages 123–129.

Muthukumarana, S., C. J. Schwarz, and T. B. Swartz. 2008. Bayesian analysis of mark-recapture data with travel time-dependent survival probabilities. Canadian Journal of Statistics 36:5–21.

Petraitis, P., A. Dunham, and P. Niewiarowski. 1996. Inferring multiple causality: the limitations of path analysis. Functional ecology pages 421–431.

Pichler, M., and F. Hartig. 2023. Machine learning and deep learning—a review for ecologists. Methods in Ecology and Evolution 14:994–1016.

Schad, D. J., M. Betancourt, and S. Vasishth. 2021. Toward a principled Bayesian workflow in cognitive science. Psychological Methods 26:103–126.

Servedio, M. R., Y. Brandvain, S. Dhole, C. L. Fitzpatrick, E. E. Goldberg, C. A. Stern, J. Van Cleve, and D. J. Yeh. 2014. Not just a theory—the utility of mathematical models in evolutionary biology. PLoS biology 12:e1002017.

Shmueli, G. 2010. To explain or to predict? Statistical science pages 289–310.

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

Trijoulet, V., S. J. Holmes, and R. M. Cook. 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., S. Depaoli, R. King, B. Kramer, K. Maertens, M. C. Tadesse, M. Vannucci, A. Gelman, D. Veen, J. Willemsen, and C. Yau. 2021. Bayesian statistics and modelling. Nature Reviews Methods Primers 1.

Winter, S. D. D., and S. Depaoli. 2023. Illustrating the Value of Prior Predictive Checking for Bayesian Structural Equation Modeling. Structural Equation Modeling-a multidisciplinary journal .

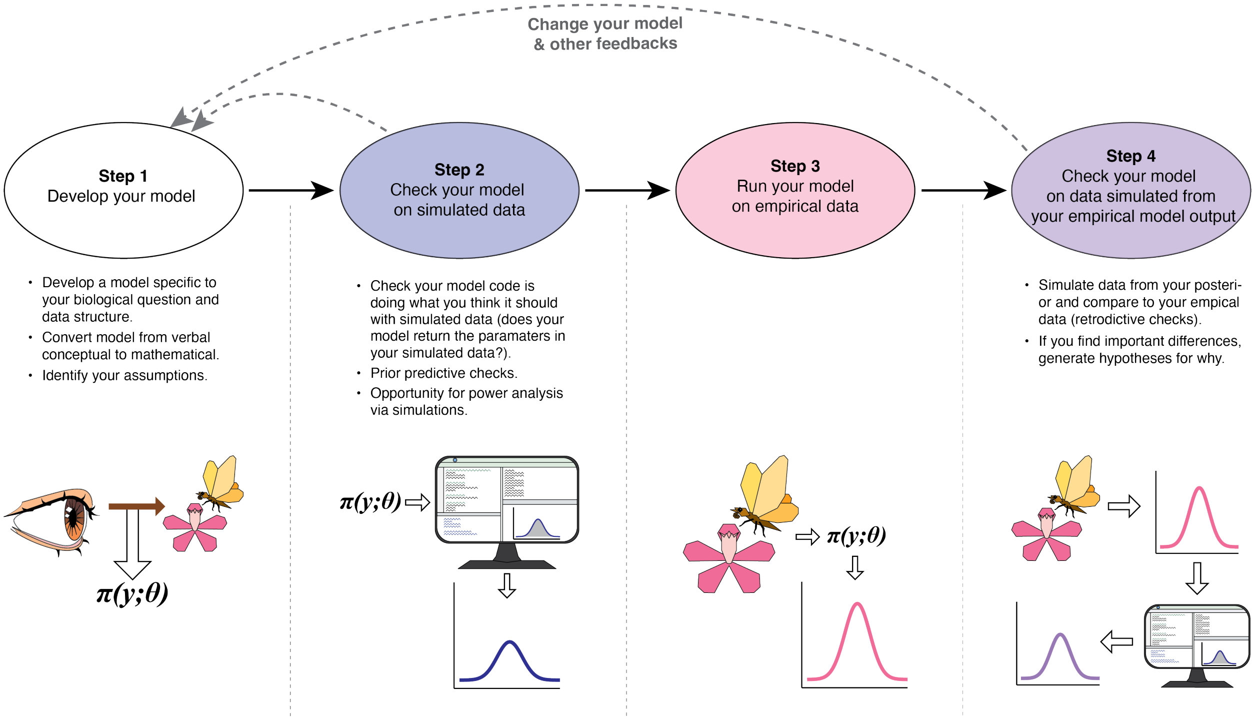
Zheng, C., O. Ovaskainen, M. Saastamoinen, and I. Hanski. 2007. Age-dependent survival analyzed with Bayesian models of mark-recapture data. Ecology 88:1970–1976.

Zwart, J. A., S. K. Oliver, W. D. Watkins, J. M. Sadler, A. P. Appling, H. R. Corson-Dosch, X. Jia, V. Kumar, and J. S. Read. 2023. Near-term forecasts of stream temperature using deep learning and data assimilation in support of management decisions. JAWRA Journal of the American Water Resources Association 59:317–337.

Tables & Figures

Table 1: Glossary: We provide below simplified definitions of the major terms we use (many of these terms, such as calibration, may be used differently depending on the particular literature).

Term	*Definition*
calibration	analyzing how often an estimate is close to the true value over an ensemble of hypothetical observations. An exact calibration would require simulating from the true data generating process which is impossible in practice. We can, however, calibrate to data simulated from the configurations of models we plan use 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 for complicated models 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
prior	an distribution of reasonable values for a parameter based on fundamental biological and ecological understanding, previous research, or other sources
statistical model	Mathematical approximations of the true data generating process labeled with numerical parameters. Evaluating a statistical model on observed data gives a likelihood function that quantifies how compatible different parameters are with the observed data, and hence can be used to 'fit' the best parameters. In this article, we often simplify to 'model.' See also the Supplement: What's a model?
posterior	product of the likelihood and prior; that is, a probability distribution that quantifies how compatible different model parameters are with both the observed data and the domain expertise encoded in the prior 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



- Develop a model specific to your biological question and data structure.
- Convert model from verbal conceptual to mathematical.
- Identify your assumptions.

- Check your model code is doing what you think it should with simulated data (does your model return the paramaters in your simulated data?).
- Prior predictive checks.
- Opportunity for power analysis via simulations.

- Simulate data from your posterior and compare to your empical data (retrodictive checks).
- If you find important differences, generate hypotheses for why.

Figure 1: The four-step iterative workflow we outline can help design models for specific ecological questions, data and aims—which makes this a statistical workflow that can naturally become a scientific workflow. It makes the step that many of us focus on—running your model on your empirical data (Step 3)—far more straightforward and insightful by using simulations both before (Step 2) and after (Step 4) it to better understand the model and data together.

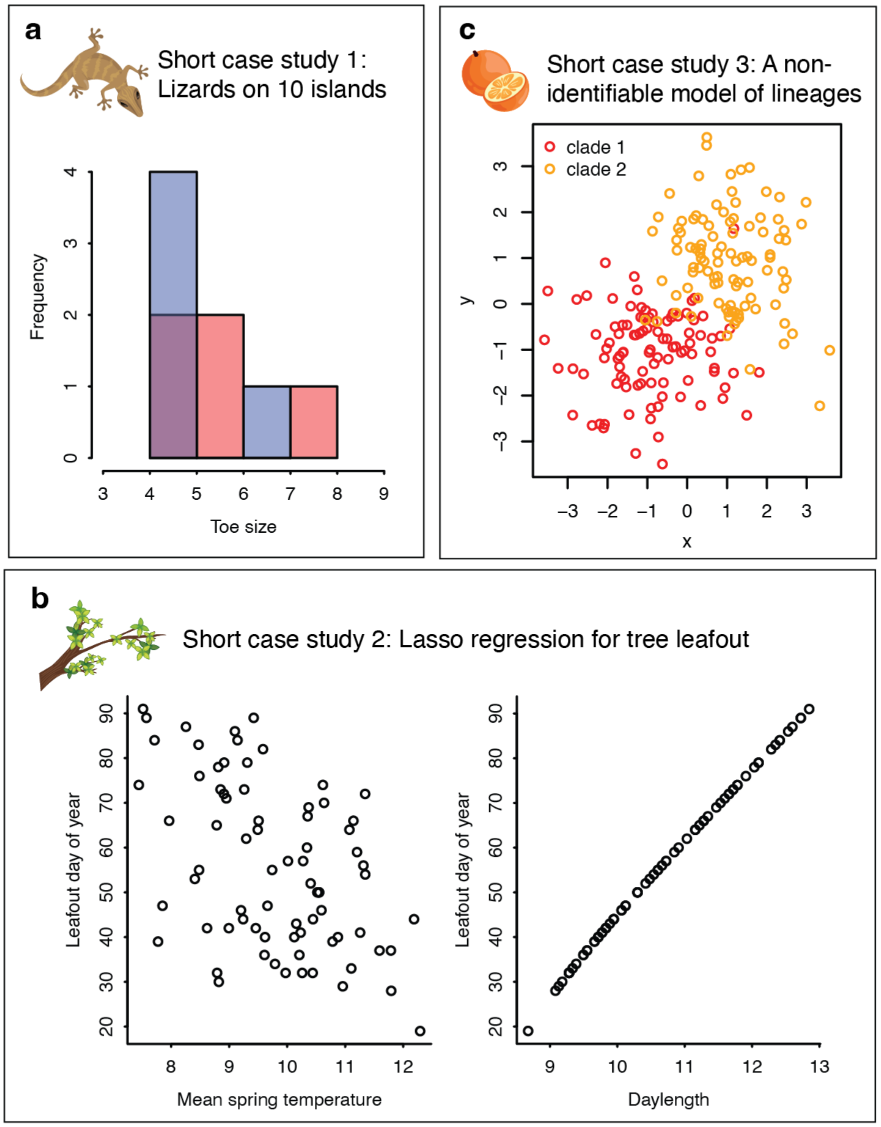


Figure 2: We provide three very simple examples of the first steps of this workflow as Supplements (in PDF from R Markdown files). One example (a) uses ordinary least squares regression considering a natural experiment on lizards on tropical islands, and simulating two different possible sample sizes. The next example (b) uses lasso regression to examine how environmental variables may predict tree leafout. The third example (c) shows several examples of non-identifiability in regression models. See supplements: ‘Is a sample size of five stormy islands enough?’; ‘Identifying predictors of tree leafout’ and ‘Three non-identifiable models, two of which are vital in biology.’

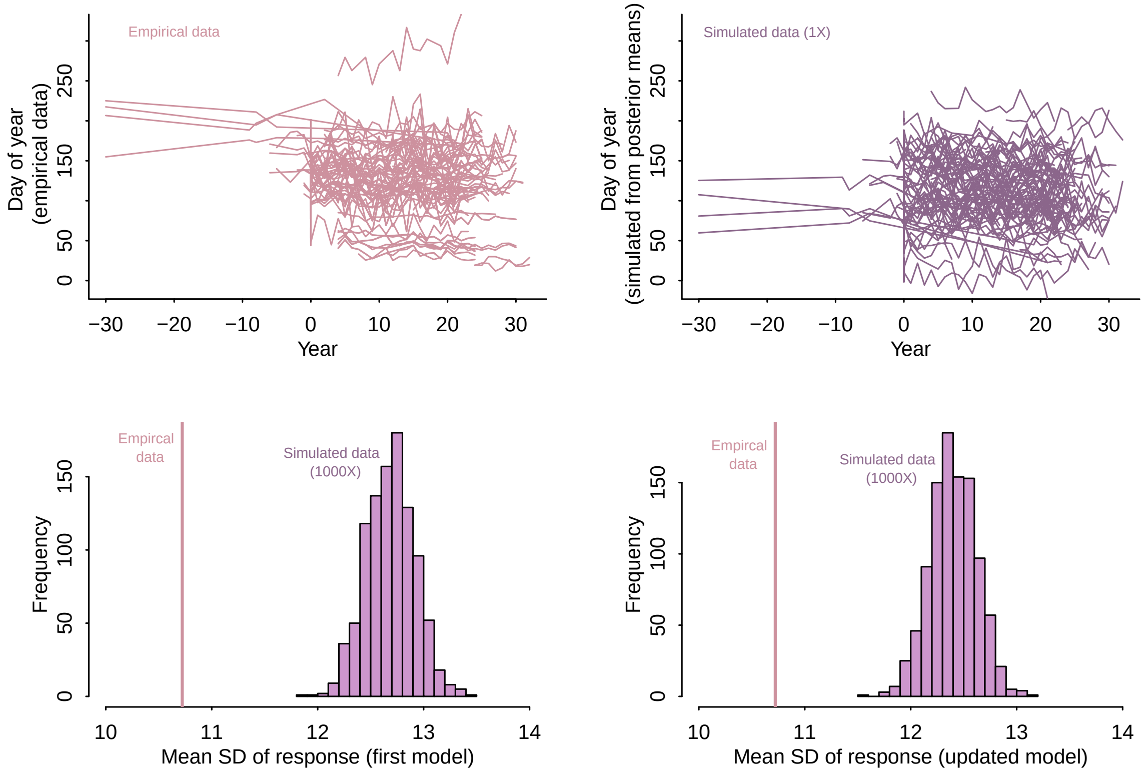


Figure 3: Example of retrodictive checks (Step 4) and feedbacks from time-series data of phenological events over time. The empirical data (top left, pink) looks similar to one simulated dataset (top right, purple), based on existing species number, their respective x data, and simulating from the parameters for each species, but the spread of the simulated data seems possibly higher. Repeating this retrodictive (or posterior predictive) check 1000 times, and taking the standard deviation (SD) of each simulated dataset, then looking at the resulting histogram confirms this (lower left in purple, empirical data SD in pink). This leads to an updated model, where the same retrodictive check looks slightly closer to the empirical data (lower right), but clearly still could be improved as part of additional feedbacks. These examples are shown in full in ‘Steps 1-4 in a Bayesian framework’ in the Supplement and available at: https://github.com/lizzieinvancouver/bayesianflowsexample.