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

EM Wolkovich, TJ Davies, WD Pearse & M Betancourt

November 1, 2025

Next steps

Our aim is to provide an approachable succinct description of a basic adaptable workflow for those new to fitting complex models. Thus anyone wanting to implement it would likely benefit from more reading, training and/or community support. We highlight some resources in the main text, but more are being produced often and our suggestions are in no way intended to be comprehensive, so we recommend checking for additional resources. For those interested in this workflow and looking for somewhere to start, we suggest:

– A number of books and online resources provide approachable introductions to fitting Bayesian models that include a similar workflow to the one we outline here. These include: *Statistical Rethinking* (MacElreath, 2016), M. Betancourt's organized writing covering probability theory to modeling and inference `https://betanalpha.github.io/writing/`, *Regression and Other Stories* (Gelman & Hill, 2020). Many of these have related resources including additional examples, videos of courses taught using the materials and more (see, for example: `https://github.com/rmcelreath/stat_rethinking_2023` and `https://avehtari.github.io/ROS-Examples/`).

– Gelman *et al.* (2014) provides a comprehensive review of Bayesian inference, including many modeling approaches.

– A number of communities support learning these approaches, but tracking them down can take some inquiries depending on your specific interests and aims. We suggest attending conferences and visiting online support sites (e.g., `https://discourse.mc-stan.org/` or `https://www.r-inla.org/discussion-group`) and asking about how people connect. For example, ask about Slack or Discord servers or check for Meetup groups.

A brief review of statistical inference using Bayesian approaches

To fit bespoke models we usually apply Bayes' theorem, which generates a *posterior* distribution from a combination of a *likelihood* and a *prior* distribution (an initial uncertainty estimate derived from basic ecological knowledge), and using iterative algorithms (e.g. MCMC, Markov Chain Monte Carlo) that provide samples that can be used to extract information from the posterior distribution. To better understand our workflow for building bespoke models, we provide a very brief overview here 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).

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 probabilities are grounded in the assumption of repeated events. 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. Second, this definition precludes the use of probability in quantifying inferences, so frequentist approaches are generally limited to point estimators and interval estimators. The behavior of point estimators can be very sensitive to the model assumptions, and even small violations of the model assumptions can make the point estimator behavior difficult to quantify. Given the complexity of ecological data and our uncertainty about the underlying model, frequentist approaches can be especially challenging in ecology.

Bayesian methods use probability more generally, allowing it to be used to quantify variation it the data but also uncertainty in our inferences about how ecological systems work. *Probability* is used to quantify uncertainty: the higher the probability of a certain interval of values the less uncertainty we have that the true behavior falls within that interval. Assuming we also have some estimate of our initial uncertainty—usually from knowledge of ecology—to inform a prior distribution (termed below *prior*), then we can apply Bayes' theorem

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

to update that prior into a posterior distribution that accounts for the information added by a likelihood (*likelihood* is the same as a frequentist likelihood). Here, `normaliser` is a mathematical constant that makes sure our probability cannot go above 100% or below 0%. This mathematical constant is a nuisance term that is extremely challenging to calculate (sometimes it is impossible!) and held back the practical use of Bayesian statistics for almost a century because that normaliser could rarely be analytically worked out. But one of the major advantages of Bayesian methods is that the solution to this problem—numerical simulations based on Markov Chain Monte Carlo

(MCMC)—provide a huge additional advantage. Now that no analytical solution needs be found, any model that can be written out mathematically can be fit to data, giving the scientist more freedom of model structure.

Which workflow?

Formally, all a 'workflow' does is organize various steps together in a systematic fashion, but there are many different workflows depending on what the aim is, which will determine which steps a workflow should include. For example a workflow aimed at calibration could look like an expanded version of our Step 1, where all the steps focus on investigating the assumptions encoded in a given model using simulated data. Or a workflow aimed at inference could expand Step 3, to focus on constructing a posterior, then investigating its model adequacy via several criteria. An inferential workflow can also be extended into a model development workflow. If the model adequacy criteria inform not only that something is inadequate about the current model assumptions but what is inadequate (ideally this happens some in Step 4) then one can use those hints to iterative improve the modeling assumptions. We present in the main text a very simplified model development workflow that combines calibration, inference and some model development, but it is not necessarily appropriate for everyone, depending on their aims.

An example workflow

Our workflow in the main text is explained mostly program-agnostically. Though at times we assume a user of Stan, a relatively new probabilistic programming language, that interfaces with R, Python, Julia (and more) to write bespoke Bayesian 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.).

For our example of the full workflow we use Stan and R, presented in Markdown (see files). As our goal is to introduce people to this approach, we present a simple—but real-world—example to walk through the basic steps. Like all models, it is a work in progress, and could easily be adjusted, expanded and likely improved (all data and code will be fully and freely available on Zenodo before publication). For now, you can download the workflow here: `https://github.com/lizzieinvancouver/bayesianflows/blob/main/examples/synchrony/example.html`.

References

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.

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

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

Figures

With 5 data points
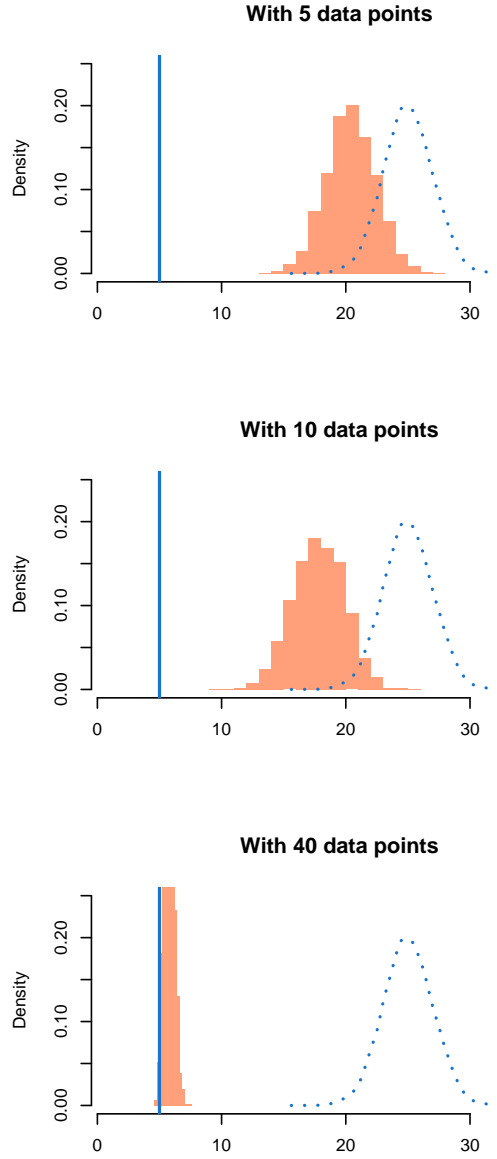
With 10 data points

With 40 data points

Figure S1: A simple example of how to use simulated data to understand calibration issues in a mis-specified model. Here we know the true model underlying the data is $y = \alpha + \text{normal}(0, \sigma)$ where α is 5 (shown as blue vertical line) and σ is 2. The model, however, is mis-specified by a prior for α of $\text{normal}(25, 2)$ (dashed blue line), resulting in a posterior (salmon-colored histogram) not centered on the true value. 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 40 data points. In practice these studies would allow us to determine how much data we would need to be robust to suspect prior models.