Closing the gap between statistical and scientific workflows for improved forecasts in ecology

Victor Van der Meersch, J. Regetz, T. J. Davies* & EM Wolkovich

March 24, 2025

* Says he is happy to help and give friendly review, but not sure he will reach level of co-author.

Deadline: 1 May 2025 (was 1 April 2025)

For: Scientific and Statistical Workflow theme issue for *Phil Trans A* as an *Opinion*

Abstract

Increasing biodiversity loss and climate change have led to greater demands for useful ecological models and forecasts. Relevant datasets to meet these demands have also increased in size and complexity, including in their geographical, temporal and phylogenetic scales. While new research often suggests that accounting for these complexities variously increases, removes or otherwise alters major trends, I argue that the fundamental approach to model fitting in ecology makes it impossible to evaluate and compare models. These problems stem in part from continuing gaps between statistical workflows – where the data processing and model development are often addressed separately from the ecological question and aim – and scientific workflows, where all steps are integrated. Yet, as ecologists become increasingly computational, and new tools make it easier to share data, the opportunity to close this gap has never been greater. I outline how increased data simulation at multiple steps in the scientific workflow could revolutionize our understanding of ecological systems, yielding new insights. Combining these changes with more open model and data sharing – and developing new efforts to race the same data – could be transformative for ecological forecasting.

Goal: Increase awareness of how we can merge statistical and scientific workflows in ecology (especially forecasting) and what we would get out of it.

1. Problem
   (a) Climate change, biodiversity crisis etc. has made it critical to understand trends to date and be able to forecast future trends (for policy)
   (b) Current workflows in ecology are not up to the task
      i. For trends: lots of different stuff reported for seemingly same data/question, makes people outside ecology wonder if we can even well document them, let alone understand them enough to suggest policy

ii. For forecasting: also high divergence between models, criticism against process-based approaches (supposedly the most robust?) for lack of transparency (in the model building and calibration) and increasing complexity, with many parameters not supported by data

(c) Here we outline the problem and provide a solution!

2. What are current workflows and where are they limiting us?

(a) For trends ...

i. easy to find different trends through small model tweaks to analyses and/or different data

ii. For example, right now many different papers report different biodiversity trends (LPI example?)

iii. New workflow would make ecologists understand uncertainty in their model data/combo (and perhaps not see/publish results as so divergent?)

(b) For forecasting ... (somehow jump to our focus on process-based models PBMs here? Something like, forecasting is big and there are diverse methods! Near-term iterative, correlative niche models, but PBMs are often considered the gold standard ... mention machine learning?)

i. as many models as researchers working on process-based models

ii. + accumulation of successive layers in the development of models = significant challenge to scientific transparency, reproducibility, interpretability
models often draw inspiration from each other, which is good (way to do science), but not always explicit... (some issues: arbitrarily established parameter value in one model then transmitted to multiple models)

iii. focus of researchers: always integrate new mechanisms, new parameters, to increase "realism"... they intuitively "feel" what kinds of adjustments is needed... but opaque from an external perspective ("black box" of model building and calibration)

iv. models rarely fitted as a whole, dozens of parameters without explicitly quantifying parameter uncertainty, and often neglect to propagate this uncertainty

v. simulations of models themselves became a subject of study to disentangle all the processes modelled and understand model sensitivity

3. Better workflows to the rescue!

(a) General overview of new workflows

i. Step 0: Research Qs and hypotheses (with a mechanism) lets you ...

ii. Step 1: Build a model!

iii. Step 2: Simulate data (and priors or something like priors that forces you to put numbers on stuff)

iv. Step 3: Design experiments/data collection (maybe you go back to Step 1 here)

v. Step 4: Simulate data from actual design/collection

vi. Step 5: Fit the model to empirical data

vii. Step 6: Retrodictive checks (feed back to 0 and 1)

(b) New vision of each workflow

4. Conclusion: world is better

Current workflows

1.

How much of forecasting do we cover?

1. PBMs
2. Near-term iterative
3. SDMs
4. Machine learning

Miscellaneous notes/points without a home

1. Ecologists need to race the same data to make progress for trends and for forecasting (point to make at end of paper maybe? And what is the workflow for this?) ... though LPI is used a lot, perhaps it is sign that ecology is ready to start racing the same data, but then we need 'analysis-ready' data so we're not all slightly differently cleaning the data etc..
2. Machine learning threatens utility of PBMs
3. We need more uncertainty propagation for trends and forecasting (uncertainty is esp. ignored in PBMs)
4. This workflow should lead to less model comparison (AIC, stepwise)
5. This workflow works for machine learning!
6. PBM: workflow should require estimating all parameters together; data simulation should reduce parameter number and highlight non-biol results
7. Current trend workflow: Should be research question focused?

Miscellaneous notes/points from thinking over 22-23 March weekend ...

1. Workflows emphasize there is no easy fix to better science or better stats
2. Maybe do a retrodictive check example with LPI? Including showing how even the Freckleton work could do more?
3. What's the pathway from model comparison of many covariates to something else? Sometimes it seems like it's just prediction shoved into a mechanistic study, but if the goal is mechanism,

there needs to be more work that either models these covariates together in a useful way (sort of approaching process-models!) or gets down to the fewer extremely relevant ones.

4. When do we need to open up the black box? Something about we need better training for what science is and our goals; machine learning is not often helping with advancing *science*

5. Scott Collins point that forecasting is not science (he claims a bunch of economists, weather folks etc. came to an ecological forecasting meeting and tried to explain that forecasting is an outcome of science, but it's not something you do science ON)