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 28, 2025

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

1. Intro
    - Ecology super challenged to predict stuff for decision making (kind of a whole new world of relevance?)
      → Example of stuff to predict: populations, policy-relevant questions...
    - General way to do this so far... (bifurcated?)
        - long term data (GBIF, Biotime...) for estimation of trends
        - PBM, SDMs for forecast
    - Gap, problem: None of this is going well
        - debates over trends, not only on the significance but also the direction!
        - predictive modeling trapped in overly complex models, hard to get scientific insights? (same as GCMs?)
    - Here we introduce a universal workflow to adress this (say more)
2. Overview
    - General scientific method we all learn stresses: RQ → study design → collect data → build model → answer
    - Divergences from this are common **(Figure 1)**
        - bad science (data lead to question)
        - important question for which we cannot get data quickly, we have to use existing data: e.g. global datasets for trends, many various small datasets for process-based (experimental...)
        - complexity makes the ideal workflow hard/impossible
    - the workflow we present here works across these realities by
        - stressing the need to think about model before study design
        - advancing data simulation

- More details on the workflow: walk through the different steps **(Figure 2)**, and emphasizes: not rigid structure, integration of everything, iterate the process of science, explicit effort to recognize the uncertainties,
  forecast is just a step: jointly model the new circumstance along with the original data
  - spend more time in critical quadrat, post-model pre-data
  - feedbacks
  - uncertainty

3. How to address current issues (two case studies, **Figure 3**)
   (a) Trends!
   - Outline current problem: different answers from different analysis (and slightly different datasets?)
     → It is a problem because we can't make decisions on +/- trends debate... (and it degrades trust in science?)
   - two big missing parts are data simulation steps:
     - retrodictive check: would highlight missing pieces (speed up current process, because without this step each slightly different model is a paper... whereas each big iteration of the workflow—including multiple feedbacks—should be a paper)
     - simulated data: you would know that you have low data sooner =¿ the debate is actually not jus models but really limited data

   (b) Forecasts!
   - Outline current problem:
     - black box: intrication between model build and data fitting (calibration), everything is mixed
     - complexity trap [mention uncertainty], counterproductive, not always better performance
     - developping a model has become the goal, whereas it should be a way to answer a research question!
   - We need the workflow to open the black box! Simulating data would allow to add a necessary step between model building and data fitting, which would highlight strong degeneracies
   - a clear framework to support (or not) additional complexity and new parameters
   - Workflow would also force you to clearly express a research question, define a limited context in which the model should apply

   (c) Step back
   - we need more data, and better question (relate this to both previous study cases)
   - where can we best reduce uncertainties through new scientific insights?
   - machine learning! If we change nothing, what's the point of not doing ML? ML > process-based without question, and ML > trends without mechanisms! Where

theory is lacking, or where we are far from mechanistic understanding, you might as well do ML!

- (ML has a way to collect and interpret large datasets...)

4. Wrap up: how to make it happen?

- usual issues: publication pressure, low standards (especially for models)
  "*we allow far more hand-waving in the presentation of modeling results than we do for experimental data*" (J. Aber, 1997)
- growing concerns, leading to increase in reproducibility and data sharing practices [mention uncertainty]
- need a little more here...
- better training! on BOTH estimation AND prediction
  - estimation: being aware of what is a parameter, and mention uncertainty propagation
  - prediction: should be a natural outcome, not a finality
- ML (short-time forecast), benchmarking models are probably useful but should not be the core of our scientific practice, not the spirit!
  $\rightarrow$ moving ecology in the right direction!