Full Title: Resisting the Lure of Complex Models As Early Career Ecologists

Short Title*: *Resisting Complex Models As ECRs

A viewpoint manuscript in preparation for submission to *Ecology Letters*.

Andrew Wood[1*]

[1] Department of Biology, University of Oxford, 11a Mansfield Road, Oxford, OX1 3SZ, U.K

* Correspondence

E-mail: wood_and@hotmail.com

Keywords: Complexity, modelling, AI, Machine Learning, hyper-dimensionality, remote sensing, high-throughput, viewpoint

Article Type: Viewpoint

Number of words in abstract: 151

Number of words in text: 1925

Number of figures, tables, and text boxes: 0

Number of citated works: 19

Abstract

Scientists have not always had freely accessible high-quality and high-resolution datasets relevant to their study systems. Today, early career researchers routinely confront a deluge of data that is relevant to their research questions. Early-career scientists face the combined challenges of using accessible yet powerful models, under high publication pressure, and with mixed guidance from scientists trained under an earlier era. There exists a temptation to reach for black-box analytical approaches to offer guidance through this wilderness of data. New complex models consisting of artificial intelligence and machine learning tools are poised to be co-opted by large numbers of early career researchers due to their modelling strength and easy, out-of-the-box usage. Just because we can use these new tools, does not mean we always should. I argue we should reconsider the role of complexity in the construction of our ecological models when we test ideas of our understanding of the natural world.

Resisting the Lure of Complexity As Ecologists and Early Career Ecologists

Early career ecologists seek to understand the biological world. We conduct scientific research to elucidate the underlying patterns and mechanisms of nature. Yet more and more often, I feel myself predisposed toward models of complexity commensurate with the phenomena themselves[1]. In the ecological world of infinite interactions and rapidly expanding datasets, I feel a strong temptation to follow the current trend of complexifying my models[2]. This "complexification" phenomenon partly comes from the lure of gathering easy research outcomes from measuring everything possible and throwing it all into a pre-made analytical framework or machine learning toolbox. We now have on our laptops the computational power to accommodate the full assemblage of nature, and increasingly accessible packages and tools to summarise it in reviewer- and publisher-friendly ways. The other enticement of overly complex analysis is the ability to find a nice little predictive tranche for every nanometre of variance within a dataset. Here the analysis becomes an incomprehensible mess, but at least the summary statistics look good.

We are all aware of the academic pressure on early-career researchers (ECRs) to produce publications. Some of us may also be aware that discussion on the issue of kitchen-sink model parameterisations and ecological model building is not a new phenomenon. Most famously, Levins (1966) remarked upon the impossibility of achieving generality, realism, and precision simultaneously[3]. But since the 1960s, the quantity of data collected has increased dramatically. This rapid increase in data collection has led to increased model complexity, a trend which has been highlighted multiple times over the last couple of years[1,2,4]. This trend means I am not the first ECR with a parameter problem. However, I do contend that this complexity conundrum is not necessarily a good thing for ecology or ecologists, and I suggest that this article serves as a reminder for all ECRs, and more specifically for ecologists, to watch the complexity of the models we use.

I acknowledge that for ECRs it is hard to avoid the benefits of pre-made black-box modelling tools. Such tools increase research speed and consequently mean that more thesis chapters can be completed; more papers can be sent for publication; and more grants applications can be written. As outcome-driven researchers with a lust for H-indices that that we must flaunt with the discretion of a peacock, this all seems very appealing. But we should all resist the urge for three simple reasons.

1. **Complex models are nuanced in their outcomes**

Nurse (2021) suggested in his article last year that "students will be better motivated and will feel more inspired if they are taught that biology has ideas"[1]. As a PhD student I feel that this is true. As students working on knowledge-based ideas, PhD students are driven by the curiosity and satisfaction and fulfilment that building accurate ecological models can deliver. But this is balanced against the highly pressurised world of academic research that traffics in a specific currency that has less to do with understanding and more to do with production. The lure of the black-box modelling system is the ability to skip over hard choices comparing competing mechanisms and deliver a simple outcome prediction, without cutting statistical corners. Black-box modelling approaches save time and answer research questions without months of agonising over fiddly mechanisms. One such example can be found in the field of phylogenetics. Done correctly, phylogenetics offers the chance to explore our past and offer insight into ecological history, biogeography, and inter-taxa and intra-population relations. Just as important, phylogenetics is essential as a blocking variable for any comparative study to account for the large trait variance due to shared ancestry. But phylogenetics is also a large field with active debate about the rate of mutation, speciation, and extinction[5,6], and intricate nuances in the ways sampling and calibration are accounted for in models[7,8]. Combining these two caveats, they suggest that simply putting data into the model to create a phylogenetic tree is not enough, it has to be a well-tuned model too. The second statement also suggests that if the tree building is not understood from the basics upwards, incorrect conclusions are easily possible. It is also possible that without understanding the mechanics of the model, it is highly likely that we will not understand why we are incorrect. George Box's famous quote of "all models are wrong but some are useful" is hard to apply to black-box models systems in ecology. If we do not know why our models are wrong, we cannot also know if they are useful. In short, just because the ecological mechanisms can be skipped in models, does not mean they should be. Karl Popper said, "science may be described as the art of systematic simplification." And systematic thinking begins with simpler models.

2. Excessive complexity in models begets unintelligibility in understanding

Excessive complexity in models begets unintelligibility in outcome; hinders understanding of processes; and obscures causality in effect. Accordingly, and despite how it may sometimes feel, being an early-career ecologist is more than just being a paper mule churning out predictions from pre-made, all-singing, all-dancing, cookie-cutter models. As ecologists, our time could be better off spent developing simpler, strategic models which are targeted in their aim and that capture the necessary information to make the inferences we desire. They are more flexible to use making them more applicable, and provide understanding across broader ranges of natural phenomena. Going back to our purpose as ecologists, science is supposed to be about developing understanding to solve problems. It has been pointed out that the paradox of enrichment[9] contributed to understanding about self-regulation in ecological systems[10], and Lotka[11]-Volterra[12] models demonstrate population dynamics, even if their simplicity is controversial[13]. Neither models are correct or found in nature, but both examples provide organising principals that allow us to understand the natural world. Simple models are integral to a fundamental understanding of nature that runs through laws or broadly consistent patterns and mechanisms in nature. In the 1970s, Bob May shook the foundations of ecology by demonstrating how stable equilibria, point cycles, and chaos can emerge from a simple, two-parameter discrete logistic model – perhaps the most elegant model the field has ever known[14,15]. The subtle but profound insight extending from May's work was that the immense complexity of natural patterns need not commensurately complex mechanism. Today we appear to have lost this wisdom.

A second outcome from ecological models are the underlying principals which emerge allow ecologists to inform decision-makers of processes occurring on both modified and wild landscapes. Without the simplicity of trophic cascade models and ideas, the reintroduction of wolves and predators would never have occurred[16]. Having worked on the Yellowstone system, the dynamics are complex but with simple models the ideas can be communicated. Without this simplicity, therefore, the whole reintroduction movement would have been more difficult.

Accordingly, we should choose models which we can inform with processes, making the outcomes more communicable when external stakeholders seek to interrogate the scientific processes they are being informed by. As ECRs, we have to remember that unintelligibility due to complexity means a lack of comprehension of the biology in the middle of the black-box. Lacking understanding provides an issue for science. In our own research we should try to understand the mechanisms, even if we don't explain all the variance through it.

3. **Big excitement and big futures**

Aside from this philosophical quandary, as ECRs, much of our time is spent developing skills which will enable us to solve future problems. To use techniques which skip any understanding is akin to finding the answers to the homework questions in the back of the textbook. It is very useful for now, and you can always go and check them again, but it doesn't allow you to develop the skills for sitting the exam. Or in this case, the detailed understanding of your chosen system. I currently work with big data sets to provide insight into local ecological processes, trying to infer what is happening to individual plants based on high-level datasets. As in much of ecology, the models that I build contain a portion of data bias, implicit assumptions and models built on models. Without deep understanding of how the models were designed, the robustness of both my model and my researcher integrity is compromised. Or put another way, by developing the techniques to pull apart the relationships between variables one develops into a more thorough, patient, and resilient ecological researcher. Inquisition becomes your passion, and you never know where the discoveries will lead you. Such personal development is the true value of mechanism over outcomes. Anything could be around the corner, and the little findings may spark curiosities which last a lifetime. Big ideas, rather than just big data, are what make ecology exciting, and so we should embrace this fully in our work!

Being careful when reconciling black-boxes with ecological modelling

Complex models will and should become part of the scientific method. But they shouldn't stop thought. The I in AI cannot replace human intelligence, rather it should assist in helping drive understanding. Recently, work has been published that demonstrates through bark beetles that deep learning is useful for the understanding of ecology[17]. It very nicely shows the utility of black-box systems for prediction, but in doing so it fails to demonstrate any sort of system-level understanding. Spurious correlations in such data may well abound, and due to the analytical techniques, we would be none the wiser for it. In the light of high-intensity data collection, from drones and satellites [18] to genetic sequences and museum specimens[19], we must be careful where the limit between automated assistance with data categorisation, annotation, and identification ends and the beginning of scientific excitement, inference, and analysis starts. As ECRs, we are primely placed to drive the field forward with fresh ideas. We can drop the complexity fetish and encourage simpler models. We can choose, through our own work, to defend the fundamentals of good science and evidence-based decision making.

When undertaking this good research, choosing to take a more mechanistic approach does not necessarily rule out any sort of tool or technique. Complex methods are not off the table because one chooses to do investigative science based upon underlying organisational principals. We are better off considering our ideas *a priori* rather than *post-hoc*. In doing so, we can stay true to the ethos of science. When I plan my work, I generally think about two questions: 1) Can I predict the outcome of this work based upon known knowledge? 2) Can I explain exactly the route by which these factors are having an effect on my outcomes? If the answer to both is yes, then great, I proceed onwards with the work with whichever technique is required to suit the task at hand. But if your answer is unknown to either then an overly complex approach may not be helpful. I find it as hard as anyone to admit not knowing the answer. But it is also unreasonable to expect that, especially as ECRs, we always do. Having studied a subject for 5 years is nothing compared to studying it for 40. Rather than resorting to not developing the science, we should develop the knowledge together with senior scientists such that we can develop understanding of the whole system.

To strive for simplicity in modelling is to push for useful, applicable, knowledge for comprehension and prediction. As early career ecologists we should be aiming for simplicity that so that we can, in short, push science forwards to develop the understanding of our natural world.

Bibliography

1. Nurse, P. Biology must generate ideas as well as data. *Nature* **597**, 305 (2021).
2. Saltelli, A. *et al.* Five ways to ensure that models serve society: a manifesto. *Nature 2021 582:7813* **582**, 482–484 (2020).
3. Levins, R. The Strategy of Model Building in Population Biology. *American Scientific* **54**, 421–431 (1966).
4. Smaldino, P. Better methods can't make up for mediocre theory. *Nature* **575**, 9 (2019).
5. Wells, T. *et al.* Species as a Heuristic: Reconciling Theory and Practice. *Syst Biol* **71**, 1233–1243 (2022).
6. Sanmartín, I. & Meseguer, A. S. Extinction in phylogenetics and biogeography: From timetrees to patterns of biotic assemblage. *Front Genet* **7**, 35 (2016).

7. Carruthers, T., Sanderson, M. J., Scotland, R. W. & Alfaro, M. The Implications of Lineage-Specific Rates for Divergence Time Estimation. *Syst Biol* **69**, 660–670 (2020).
8. Duchêne, S., Lanfear, R. & Ho, S. Y. W. The impact of calibration and clock-model choice on molecular estimates of divergence times. *Mol Phylogenet Evol* **78**, 277–289 (2014).
9. Whitnack, G. C. & Martens, H. H. Paradox of Enrichment: Destabilization of Exploitation Ecosystems in Ecological Time. *Science (1979)* **171**, 385–387 (1971).
10. van Nes, E. H. & Scheffer, M. A strategy to improve the contribution of complex simulation models to ecological theory. *Ecol Modell* **185**, 153–164 (2005).
11. Lotka, A. J. Contribution to the Theory of Periodic Reactions. *J Phys Chem* **14**, 271–274 (1910).
12. Volterra, V. Variations and Fluctuations of the Number of Individuals in Animal Species living together. *ICES Journal of Marine Science* (1931).
13. Hall, C. A. S. An assessment of several of the historically most influential theoretical models used in ecology and of the data provided in their support. *Ecol Modell* **43**, 5–31 (1988).
14. May, R. M. & Oster, G. F. Bifurcations and Dynamic Complexity in Simple Ecological Models. *https://doi.org/10.1086/283092* **110**, 573–599 (1976).
15. May, R. M. Simple mathematical models with very complicated dynamics. *Nature 1976 261:5560* **261**, 459–467 (1976).
16. Ripple, W. J. & Beschta, R. L. Trophic cascades in Yellowstone: The first 15years after wolf reintroduction. *Biol Conserv* **145**, 205–213 (2012).
17. Rammer, W. & Seidl, R. Harnessing Deep Learning in Ecology: An Example Predicting Bark Beetle Outbreaks. *Front Plant Sci* **10**, 1327 (2019).
18. Tuia, D. *et al.* Perspectives in machine learning for wildlife conservation. *Nature Communications 2022 13:1* **13**, 1–15 (2022).
19. Hansen, O. L. P. *et al.* Species-level image classification with convolutional neural network enables insect identification from habitus images. *Ecol Evol* **10**, 737–747 (2020).