

# How to do good science\*

## INTRODUCTION

There is a remarkable paucity of ecological literature on how to do ecology well. This seems to be somewhat true for all of science. Of the philosophers of science with whom I am familiar – Popper, Kuhn, Lakatos and Feyerabend – only Popper provided explicit direction for how to do science well (Popper, 1959). And Popper's advice was difficult to apply and irrelevant to a large number of research questions that most people consider science. Around the same time that Popper was presenting his ideas, Platt published – in *Science* – his recommendation for testing competing hypotheses (Platt, 1964). So, there was some discussion about how to do science well but not in a comprehensive way.

There seems to be an aversion among scientists to being prescriptive in print. Most recently, Fox (2025) asserted that one of the great strengths of ecology is our diversity of approaches. However, it

seems indisputable to me that even a diversity of approaches must share a common ontology that is the foundation of 'good' science. But I have never read a description of what those shared core values are – what the common DNA is.

That's what will follow. My prescription for doing 'good' science. This may be seen as hubris or arrogance, but I am baffled by the thought that practitioners of science should not have opinions about best practices for doing science. Butchers, bakers and candle stickmakers are all willing to weigh in on how to practice their trade well. Yet scientists remain largely silent about what makes a good scientist.

I am not expecting what follows to be definitive, exhaustive or even particularly insightful. But I do hope it opens a long and fruitful discussion about what it means to do science well. Here are the twelve principles of good science.

## TWELVE PRINCIPLES OF GOOD SCIENCE

**PRINCIPLE #1 OF GOOD SCIENCE: Do Science.**

**PRINCIPLE #2 OF GOOD SCIENCE: Prediction is the only way to demonstrate understanding and precise, quantitative understanding can only be demonstrated with precise quantitative predictions.**

**PRINCIPLE #3 OF GOOD SCIENCE: Predictions are made by models and all of human understanding is contained in our models of the natural world.**

**PRINCIPLE #4 OF GOOD SCIENCE: Complete models consist of causal mechanisms (and their interactions), functional forms of relationships between drivers and responses and strength of relationships between drivers and responses.**

**PRINCIPLE #5 OF GOOD SCIENCE: Perfect understanding implies zero or very small prediction errors.**

**PRINCIPLE #6 OF GOOD SCIENCE: The amount of understanding/knowledge in a model can be measured by its predictive ability.**

**PRINCIPLE #7 OF GOOD SCIENCE: Distinguish between causal and correlational relationships.**

**PRINCIPLE #8 OF GOOD SCIENCE: Model building is better than hypothesis testing.**

**PRINCIPLE #9 OF GOOD SCIENCE: Know the current best model.**

**PRINCIPLE #10 OF GOOD SCIENCE: Define the scale and scope of your models.**

**PRINCIPLE #11 OF GOOD SCIENCE: Models must be transferable in time. It's nice if they are transferable in space and across taxonomic groups.**

**PRINCIPLE #12 OF GOOD SCIENCE: Know exactly what part of model building your research is addressing.**

## PRINCIPLE #1 OF GOOD SCIENCE: Do Science.

To do good science, you need to know what science is. The distinction between science and non-science has become a part of the public conversation as debates about “follow the science” and “doing your own research” have escalated in the wake of the covid-19 epidemic. Despite the heat associated with those

discussions, people who insist we must ‘follow the science’ can rarely provide a coherent definition of science. Perhaps more concerning is that I know few scientists who can provide a coherent definition of science.

Let’s begin with what science is not.

### MYTH # 1: SCIENCE IS KNOWLEDGE ACQUIRED THROUGH THE SCIENTIFIC METHOD

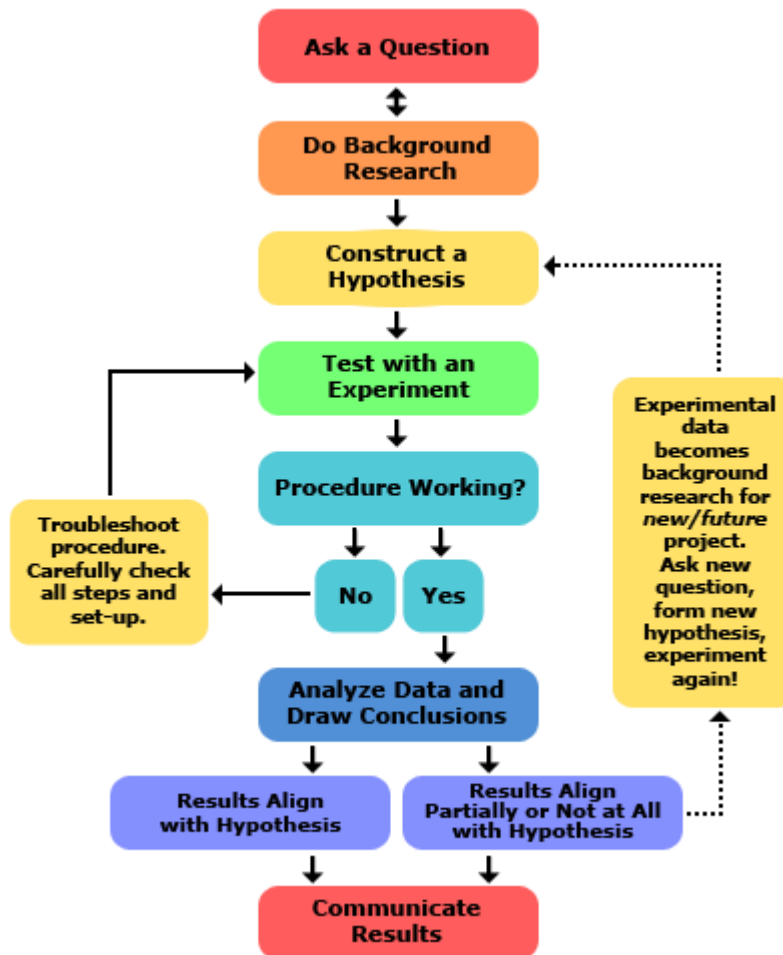


Figure 1: The Scientific Method

The Scientific Method is a useful heuristic, but it is far too restrictive. It rests on a foundation of hypothesis-testing. But a large part of what we consider science is purely descriptive – the scientists who mapped the human genome never tested a hypothesis; development of the CRISPR – Cas 9 technology never tested a hypothesis; the detection of gravitational waves never tested a hypothesis. Every practicing scientist understands that few scientists follow the scientific method when they do their work. And none of us think that's a problem, let alone disqualifying.

The critical piece of the scientific method is...using observations to answer a question. And this is one of the three pillars of doing science.

#### **MYTH #2: SCIENCE IS DISTINGUISHED FROM NONSCIENCE BY ITS FALSIFIABILITY**

Karl Popper's most significant contribution to the practice of science was 'solving' the demarcation problem – identifying the characteristics of science that separated it from nonscience. Popper concluded that "*Statements, or systems of statements, therefore convey information about the empirical world only... if they can be subjected to tests—testable, and falsifiable.*" (Popper, 1959)

However, this definition also rests on the foundation of hypothesis testing while so much of what we regard as science

(1) doesn't test hypotheses and (2) hypothesis testing may not be the best way to understand the world.

"What is science?" is not an easy question. However, if you do not have a coherent definition of science, it's time to stop making accusations about what is or isn't 'pseudoscience'. And you certainly are in no position to adjudicate what is 'good' science.

So, if the scientific method and falsifiability don't define science – what does?

#### **MY DEFINITION OF SCIENCE**

My definition is very inclusive and includes a lot of research that many people – including many scientists – would not accept as scientific. There are three components necessary for somebody to be doing science. Those three pillars are...

1. They must believe that there is an objective reality. That is, they must believe that there is a 'real' world that exists independently of human perception or human existence. That is, they must be scientific realists.
2. They must be solely committed to discovering the truth of the 'real' world. Any other motivation for doing research other than discovering what is true about the natural world is a potential impediment to doing science. Scientific research filtered

through any motivation but truth-seeking runs the risk of being non-science.

3. They discover the truth of the natural world using empirical observations. Evidence for what the natural world is and how it works comes only from observations – not from dreams, or celestial voices or other sources that cannot be measured and documented.

That's it. These characteristics are both necessary and sufficient for doing science. By definition, these same characteristics are necessary for doing good science. But they are not sufficient. A child trying to understand life in their backyard pond by observing, is doing science. But it almost certainly falls far short of good science.

These three characteristics may seem simple and even obvious. But I believe they are high enough barriers that a significant portion of 'accredited' science (i.e. published in academic journals and/or carried out in university, government or corporate laboratories) doesn't meet my definition of science. Let's take them one by one.

*Belief in objective reality:* It's not clear to me what the rationale for doing science is if you aren't a scientific realist. If the world is 'in the eye of the beholder' then each of us doing scientific research is discovering things about a world that exists only for ourselves. What you're discovering isn't relevant to what I'm discovering because we live in different

worlds. This may seem obvious. However, this belief in an objective reality implies that the natural world is identical for all of us – western cultures, indigenous cultures, cultures of all religions, men, women, people of all life experiences. It implies that there is a single truth about the natural world. When it comes to understanding the natural world, we cannot talk about 'my truth' or 'your truth'. There is only 'the' truth. Difficult as it might be to uncover. Our perception of the world may differ, but the true nature of the world is identical for all of us.

*Sole commitment to discovering the truth:* I am deeply suspicious of any scientist who responds to this with a dismissive wave and a curt "Of Course". Because this is hard. Organised science has built its entire infrastructure around perverse incentives. To be successful in science - and by extension to have the opportunity to do science – one has to publish a lot and publish in recognized journals. Getting hired, getting tenure, getting promoted, and getting funding and recognition from our peers, require publications in recognized journals. The more publications and the more prestigious the journals, the better.

The most extreme consequence of these perverse incentives is scientific fraud. Ecology has had prominent examples of alleged data fabrication – Jonathan Pruitt's "research" on the social behavior of spiders (Treleaven, 2024) and Danielle Dixson's findings on the effects of carbon dioxide on fish

behavior (Enserink, 2022), but this is likely the tip of the iceberg. Jonathan Pruitt fabricated data so clumsily that it would have been difficult not to detect. I suspect that there are ecologists fabricating data more skillfully than Pruitt, who will be able to continue doing so indefinitely. However, my belief (and hope) is that outright fraud is rare. By contrast, I believe that p-hacking, HARKing and selective reporting of results are absolutely commonplace in ecology and many other disciplines.

P-hacking involves a set of analytical practices that result in findings that are statistically significant (usually  $p < 0.05$ ). Those practices include (i) applying several statistical tests and using the one that provides statistical significance, (ii) breaking the full data set into smaller subsets and reporting the results on the subsets that have statistically significant results, (iii) including or excluding covariates until you have statistical significance and (iv) removing outliers. Scientists do this because statistically significant findings are usually easier to publish than non-significant findings.

Harking is “Hypothesizing After the Results are Known”. This occurs when a scientist stumbles on an unexpected finding in their analyses and pretends that the unexpected result was the test of a hypothesis they had before they collected the data. Research suggests this is a common practice in science (Kerr, 1998).

Selective reporting involves carrying out multiple studies/experiments but only

reporting the ones that result in statistically significant results.

Any researcher who has used any of these practices has abandoned their absolute commitment to the truth. But when there is great pressure to publish, it is easy to convince yourself that these practices are bringing you closer to the truth. They are not.

I and - I suspect - most ecologists I know have used some variation of these practices at some point in their careers.

Perhaps even subtler is – the framing of research questions and subsequent results. Scientists may frame research questions and results to imply that their research has more real-world relevance than it actually does or to imply that the research addresses a fundamental question in their discipline when it is only tangentially related (Peters, 1991). They do this to increase the ‘profile’ of their research.

Further, regardless of the research question, scientists often have a preferred outcome for reasons other than publishability. I’ll give a personal example. My lab worked on a whole-system experimental manipulation that tested the impacts of glyphosate (the active ingredient in RoundUp) on pond communities. Most of us were conservation biologists and we were sure we would find that glyphosate had negative impacts on amphibians and aquatic invertebrates. We didn’t find that. There ended up being a lot of discussion about how to frame our

results, so that we weren't giving a free pass to Monsanto and herbicide users. We shouldn't have had that conversation.

The question that every researcher should ask themselves is...

*"Imagine the result you would least like to find in the project you are working on – the finding that would be most reprehensible to you personally, that you think would do the most harm to society, that would contradict your own previous research most definitively ...would you publish it without caveat? If the answer is No – choose another project."*

This is what it means to be absolutely committed to the truth. This is what it means to do science.

*Truth through empirical observations:*

This may be the easiest of the three. We discover the truth through observations. Answers do not come to us in our dreams. There is no higher power that can provide answers that are unavailable to us through observation. Our 'gut' – our 'intuition' – cannot provide answers. Our 'gut' may guide research questions and methodological approaches and hypothesis development. but ultimately only the data – only the observations – provide answers.

So, that's science. That's it.

But good science requires much more.

*The objective of science:* It will be impossible to define good science without identifying the objective of science. This is true of most endeavors – what it means to be a 'good' driver depends on whether your goal is to arrive quickly or safely. I believe there is a general consensus among scientists, if not unanimity, that the objective of science is to discover what the natural world is made up of and how it works. That is, to discover what's true about the natural world. And that will be the foundation for all that follows – that the sole objective of science is to find out what's true about the natural world. If you disagree, then what follows will be of no interest or relevance to you.

There are a couple of ways that I represent this idea. First, as a wandering line from ignorance to the truth. The direction is – on average – towards the truth but there are often steps backwards. In this representation - truth is a point we are moving towards. The second representation is of scientists carrying information – some of which is true - to fill the space of everything we don't know...truth is a space that needs to be filled with knowledge and understanding. Both capture something of how I think about this – but the second probably captures my thoughts more closely than the first (thanks to Cody Carlyle – a graduate student in a Philosophy of Science session – for this idea)

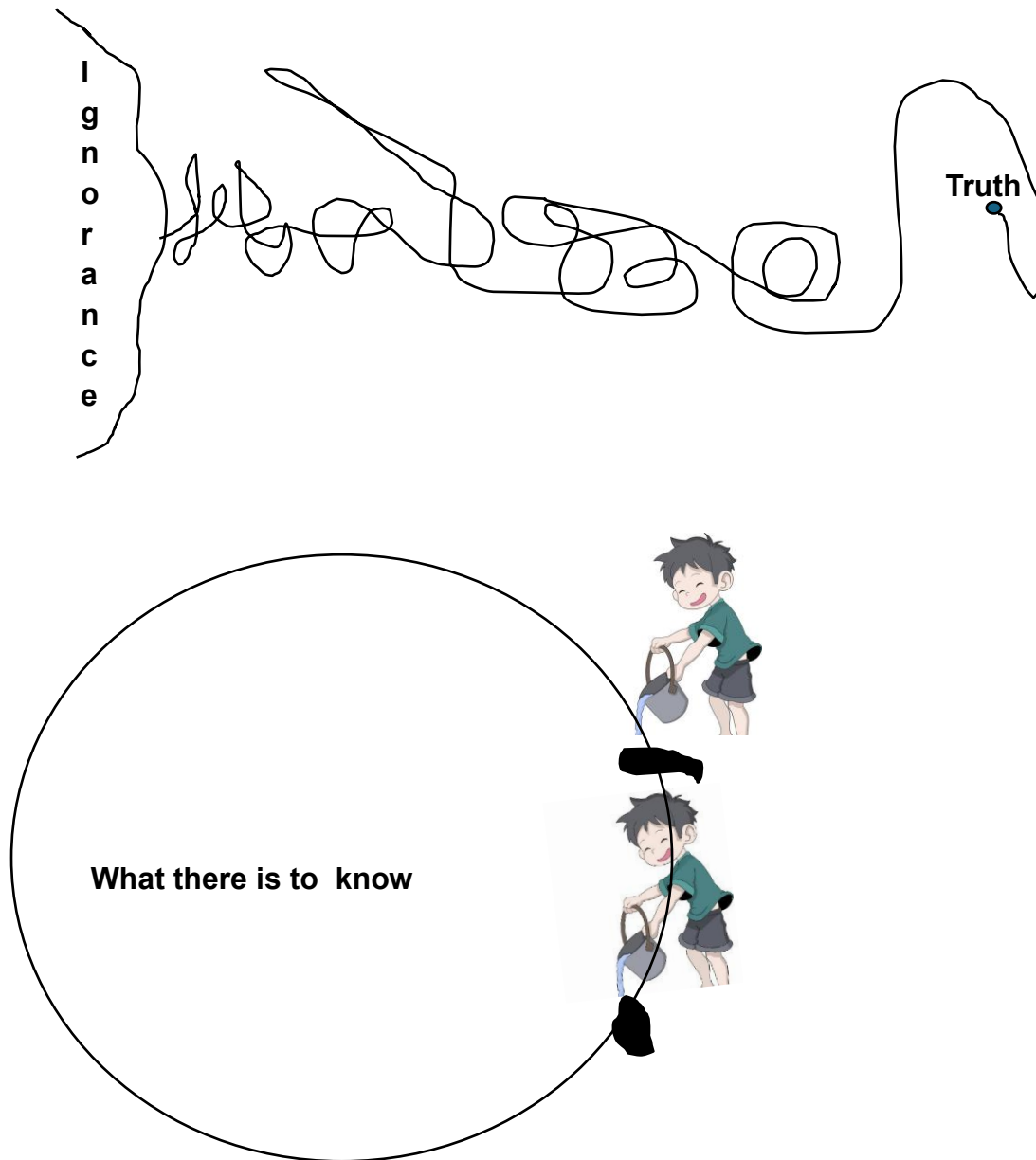


Figure 2: Schematics representing scientific progress (A) as a path from ignorance to truth and (B) filling up what we don't know with what we are learning.

So, the objective of science is to discover the truth of the natural world. We can think of that as getting closer to the truth or gathering more knowledge and understanding. So, we need an explicit definition of scientific knowledge.

*Description versus understanding:* There are two broad categories of scientific knowledge – descriptive knowledge and understanding. Hempel and Oppenheim (1948) identify the 'what' and even more importantly, the 'why' as the chief

objectives of scientific inquiry.

Descriptive knowledge answers the 'what' question – what is it? What is bird species richness on Krakatoa?; What is the melting point of mercury? What is the CO<sub>2</sub> concentration in the earth's atmosphere? Understanding provides answers to 'Why' questions. Why are there 54 bird species on Krakatoa? Why is the melting point of mercury -38.83 °C? Why is the current CO<sub>2</sub> concentration in the earth's atmosphere 416 ppm? Descriptive knowledge describes the characteristics of real objects and understanding addresses the mechanisms that cause variation in those characteristics

Descriptive knowledge – the What? Where? And When? questions are the easy part of science. Understanding – the Why? and How? are hard. Here I focus on understanding – the hard stuff.

Now that I've defined science and its objective, I can define good science – good science increases our understanding of the natural world. One scientific approach is better than another to the extent that – all other things being equal - it increases our understanding of the natural world faster.

**PRINCIPLE #2 OF GOOD SCIENCE:**  
*Prediction is the only way to demonstrate understanding and precise, quantitative understanding can only be demonstrated with precise quantitative predictions.*

Prediction has been a cornerstone of science since nearly its inception. In fact, one of the key steps in the scientific method is identifying predictions that flow deductively from the hypothesis you are proposing. Further, Popper made clear that 'risky' predictions – i.e. predictions that were easily falsifiable – were better than 'safe' predictions. And the 'riskiest' predictions are precise, quantitative predictions.

Perhaps the most famous prediction in science – Einstein's prediction about light deflection as it passed the sun - was quantitative, precise and very risky. The prediction implied by Einstein's General Theory of Relativity was that light should be deflected by 1.74 to 1.75 arcseconds as it passed near the sun, while Newtonian theory asserted that light would be affected only by gravity and would bend half that much. And measurements provided evidence for The General Theory of Relativity.

Spoiler alert: Both of these predictions emerge from mathematical representations of how the world works.

In ecology – and I suspect many other disciplines – we rarely test hypotheses with explicit predictions. And when we do make predictions, they are never

precise and quantitative. And they are rarely tested a second time on new data that weren't used to reach the initial conclusions. That is, we rarely replicate studies in ecology.

Replication has long been considered a critical component of science. And replication implies prediction. That is, we predict that whatever happened in the first study will also happen in the second, third and all subsequent studies. Replication is especially useful if the original study allows us to make precise, quantitative predictions for all subsequent studies. However, the proposal instructions for most granting agencies almost always include novelty' as a criterion. Replication is, by definition, not novel.

So, we give lip service to replication. We believe it is essential but aren't willing to fund it. And beyond funding barriers, one of the major barriers to replication is the fear that the 'very interesting' thing we found the first time isn't true.

So, while prediction is touted as critical for doing good science it has become a very low priority in ecology and other scientific disciplines.

Scientists routinely make claims to understanding the natural world – both in our published works and in public statements. Those claims need to be supported by evidence. There is undeniably near unanimous agreement among scientists that prediction is AN important piece of demonstrating

understanding – even if relatively little in how scientists do their research reflects this agreement. Where I differ from many scientists is that I claim that prediction is THE ONLY way to demonstrate understanding.

I can't show this deductively. I make the claim because I can't think of another way to demonstrate understanding. And nobody has provided me with another way. It's not that people have suggested alternatives and I've found them lacking – nobody has ever offered an alternative. And I haven't read an alternative in the literature. But the

challenge remains open. Show me another way to demonstrate understanding than prediction. So, for now...

Prediction demonstrates understanding.

There is no other way but prediction to demonstrate understanding (Houlahan et al., 2017). And if we can't demonstrate our understanding of the natural world, why should we or anybody believe we have understanding?

And where do predictions come from?

Our models of the natural world.

**PRINCIPLE #3 OF GOOD SCIENCE:**  
*Predictions are made by models and all of human understanding is contained in our models of the natural world.*

Models come in many forms including verbal, mechanical, simulation, decision trees or mathematical and they contain all of our understanding about how the natural world works. We think first of models as mathematical, but many of our ecological models are not mathematical. For example, most ecologists, if asked to describe what we understand about the relationship between species richness and the size of the area a set of species occupies, would say something like “big areas contain more species than small areas”. This is a verbal model of the species-area relationship. Similar heuristics would be used to describe the relationship between species richness and productivity and the effects of population density on population growth rates. Is this an indictment of ecology as a discipline? That’s hard to say. Perhaps given the complexity of the world and our ability to measure ecological variables, heuristics are the best ecologists can do. But it is notable that there is not a single ecological phenomenon for which ecologists have a general mathematical model that includes parameter estimates.

So, why the emphasis on mathematical models? Because to completely explain a particular natural phenomenon (say,

bird species richness on Krakatoa) a model must identify three things about the phenomenon – (1) ‘drivers’ (i.e. causal mechanisms) and the interactions among them, (2) the functional relationships between the causal mechanisms and the phenomenon of interest (e.g. linear or nonlinear) and (3) the size of the effect for each causal mechanism (e.g. the slope of a linear relationship). Combining these three sets of information into a single model is done most easily with mathematics. Species-area relationships are a simple example of how causal mechanisms, functional relationships and effect sizes can be integrated into a single mathematical model -  $SR = Area^{0.25}$ , where SR is species richness and Area is the size of the patch. In this mathematical model variability is caused by a single variable, size of patch, there is a nonlinear power relationship between the size of the patch and species richness and species richness increases as the  $\frac{1}{4}$  root of patch size. But these kinds of general mathematical models don’t exist in ecology. That is, while ecologists may use mathematical models in very specific places and/or contexts we have no general models that apply over large spatial and temporal scales. We are much more likely to use verbal heuristics. Because that is what hypothesis testing lends itself to.

Hypotheses are statements about how the world works. Hypotheses are not

models, but they often – perhaps always  
- imply a model of the world. In ecology,  
hypotheses are almost always verbal

rather than mathematical. This places  
limitations on the predictive ability of  
hypotheses. More on this later.

**PRINCIPLE #4 OF GOOD SCIENCE:**  
*Complete models consist of causal mechanisms (and their interactions), functional form of relationships between drivers and responses and strength of relationships between drivers and responses.*

Models, whether verbal, logical or mathematical, include a phenomenon of interest - the pattern or process that we are trying to understand (more formally known as the explanandum – the thing to be explained). In a mathematical model, this is the dependent variable or the response variable. But, even in a verbal model or a logical model like a decision tree, the concept of a response or dependent variable works. Models also include the causal mechanism(s) of the phenomenon of interest - the factors that cause changes in the phenomenon of interest (more formally known as the explanans – the explanation for the phenomenon of interest). In a mathematical model we would refer to them as independent or explanatory variables. But knowing that factor A is a causal mechanism for a phenomenon of interest is not enough – to completely understand a process or pattern we must know the functional form that links the causal mechanism to the phenomenon of interest and the strength of the relationship.

**Phenomenon of interest:** Any 'real' property of the natural world (i.e. object, pattern or process) is a suitable subject

for scientific study and a candidate to be a 'phenomenon of interest'.

A phenomenon of interest can be binary, categorical, integer or continuous. The prediction will be (1) one of two states (e.g. dead/alive), (2) one of several states (e.g. eye colour), (3) one from a sample space of integer values (e.g. number of individuals), and (4) one from a sample space of continuous values (e.g. body size (kg)), respectively. When only two states are possible the phenomenon of interest is usually binary. However, there are contexts where, what appears to be a binary target, is actually categorical. For example, a specific individual can only be present or absent from a particular site, but they also may be present and undetected, so there are actually three categories, present and detected, absent, present but undetected. In theory, integers and continuous values can range from  $-\infty$  to  $\infty$  but phenomena of interest may have 'hard' floors and/or ceilings. For example, the number of fish in Lake Erie, the number of bird species in Canada and the mean weight of fish in Lake Michigan have 'hard' floors of zero. Further, estimates of proportion or probability are continuous but have a hard floor of 0 and a hard ceiling of 1.

A common mistake in ecology is to model an integer or continuous phenomenon of interest as categorical (e.g. large, medium and small body size) or binary (tall or short) – there is

rarely a good rationale for doing this. When we convert integer or continuous phenomena of interest to categories, we lose information because we treat each value in a category as being equal when, in fact, there is variation among values, and the variation contains information about the phenomenon of interest.

**Causal mechanisms (“drivers”):** A driver is any factor where a change in the state or magnitude of that factor causes a change in the state or magnitude of the phenomenon of interest. Drivers, like phenomena of interest, can also be binary, categorical, integer or continuous. All four types of drivers can be included in models for the same phenomenon of interest.

Traditional approaches to model-building often extol the benefits of Occam’s Razor or the parsimony principle. That is, that our objective should be to identify some subset of important drivers. There may be practical reasons for beginning with models that include a subset of drivers and there may be utilitarian reasons for using models that don’t include all drivers. But there is never a good reason for choosing to be ignorant of some drivers. While some drivers will be more important than others, “complete understanding” requires identifying all drivers.

Generally, the relative effect of a driver decides its importance. That is, if the same change in both driver A and driver B causes a larger change in the

phenomenon of interest for driver A than driver B, then driver A is more important. However, in specific contexts a driver with a smaller relative effect may be more important because, in that specific context, the variability of the driver with the smaller relative effect is much greater than the variability of the driver with the larger relative effect. For example, in general, mean annual temperature may have a larger effect on mean annual primary productivity than mean annual precipitation but if mean annual temperature varies little and mean annual precipitation has large fluctuations then annual precipitation will be the more important driver of mean annual primary productivity.

Because many ecological phenomena have complex dynamics, interactions among drivers are likely going to be common. An interaction between two drivers simply implies that the strength of the effect of driver A changes as the magnitude of driver B changes. I have chosen to mention interactions in the ‘drivers’ section because interaction terms in a model can be thought of as an additional driver. For example, in a model with two drivers, driver A and driver B, the interaction between driver A and driver B can be thought of as a third driver. So,  $Target = A + B + A*B$ .

**Functional relationships:** The most basic distinction among functional relationships is linear versus nonlinear. And again, because ecological phenomena often have complex dynamics, nonlinear relationships may

be common. Linear relationships imply that the strength of the relationship between the driver and response variables remains constant across the entire range of the independent variable. A nonlinear relationship implies that the strength of the relationship between driver and response is greater over some ranges of the driver than others.

The most basic distinction among linear relationships is negative versus positive relationships. By contrast, nonlinear relationships can take many forms. One basic distinction is monotonic versus nonmonotonic. Monotonic relationships imply that the strength of the relationship between driver and target varies with the value of the driver, but the direction of the effect (i.e. negative versus positive) doesn't. Non-monotonic relationships imply that over some range of driver values, the relationship is positive and over others it is negative.

We may observe many different nonlinear functional forms in ecology including exponential, power, sigmoidal, parabolic, wave or saturation curves.

In ecology, we often approximate nonlinear relationships with linear functions by (1) ignoring subtle nonlinearities and assuming the non-linearities are unimportant or (2) by transforming variables. There may be contexts where it is practical to ignore nonlinearities but, just as it's never preferable to be ignorant of 'true' drivers it's never preferable to be ignorant of 'true' functional relationships.

Transforming the values of drivers and/or responses can allow us to estimate parameters using linear analytical techniques. For example, it is common to estimate the intercept and slope parameters for the area-species richness relationship using linear analytical techniques by log-transforming area and species richness values. However, despite estimating the parameters using linear analytical techniques the log area – log species richness relationship implies a power function. It is linear in the parameters but not in the variables.  $\text{Log SR} = \text{Log } c + z * \text{Log Area}$  is deterministically equivalent to  $\text{SR} = c * \text{Area}^z$ . But because parameters are being estimated with error these models are rarely, if ever, perfectly equivalent, because the first model assumes additive errors and the second assumes multiplicative errors.

Getting the functional forms right may not be as critical as getting the drivers right but it is an important step in understanding how the natural world works.

**Strength of Relationships – parameter estimation:** Estimates of the strength of relationships can take many forms – the slope of the relationship between a driver and response variable in a linear model, or the difference in means between two groups when the driver exists in one of two states, or estimates of  $c$  and  $z$  in a power equation  $Y = c * X^z$ . So, any information that moves beyond

stating 'X has an effect on Y' to 'When X changes by x then Y will change by y'.

Data scientists generally believe that they have the correct model if they have identified the correct causal mechanisms and the correct functional relationships. That is, a model can be 'right' even if the estimates of the strength of the relationships are poor and the model makes bad predictions. There is a good reason for this – because poor parameter estimation is a function of sample size and measurement error and parameter estimates can always be improved by collecting more samples or measuring more accurately. On the other hand, if the model is missing causal mechanisms, including variables that are not causally linked to the phenomenon of interest or gets the functional relationships wrong, no amount of sampling can correct the problem. So, getting it wrong about the strength of effects is fundamentally different than getting it wrong on causal mechanisms and functional relationships. Getting it wrong on causal mechanisms and functional relationships results in errors due to bias and getting it wrong on parameter

estimates leads to errors caused by variance.

However, knowing the strength of relationships between causal mechanisms and response variables is a critical part of understanding and essential to acquiring complete understanding of any particular phenomenon. So, I treat a model as 'wrong' if it (1) includes factors that are not causal mechanisms, (2) is missing causal mechanisms, (3) uses incorrect functional relationships and/or (4) gets the strength of effects wrong. However, that doesn't mean that getting the causal mechanisms wrong, the functional relationships wrong and the strength of effects wrong have equal consequences for understanding – some 'mistakes' are worse than others.

So, if our understanding of the ecological world is contained in our models then it is contained in our causal mechanisms, functional relationships and estimates of the strength of relationships. We can use this to measure scientific progress.

**PRINCIPLE #5 OF GOOD SCIENCE:**  
*Perfect understanding implies zero or very small prediction errors.*

There are two philosophical camps on this – that true stochasticity exists or it does not. The first camp believes that there is true irreducible error, a non-zero amount of variance in any response variable that cannot be explained. The second camp believes that unexplained variance in a response variable is just the sum of all the things we don't understand about a phenomenon. Heisenberg's uncertainty principle implies that true stochasticity exists for physical particles at the quantum level. However, it's not clear that this fundamental stochasticity, if it exists, would emerge at the cellular, individual, population or community levels. Or, if it emerges above quantum scales, how

large that stochasticity would be. Thus, it's not unreasonable to assume that perfect understanding leads to perfect or near-perfect predictions. My proposed approach to measuring scientific progress using predictive ability, to some extent, relies on the assumption that the theoretical floor on prediction error is zero or very close to it.

However, knowing the lower limit to prediction error isn't absolutely necessary for measuring scientific progress. It's only necessary for knowing how close we are getting to the truth. It's not necessary to know your destination to know how far you've driven from your starting point. Thus, reductions in prediction error imply scientific progress even without knowing what perfect understanding is.

***PRINCIPLE #6 OF GOOD SCIENCE:  
The amount of  
understanding/knowledge in a model  
can be measured by its predictive  
ability.***

Scientific progress can be defined as our increase in knowledge or understanding. In recent decades, science has been measured by output. That is, the quantitative measurement of the products resulting from scientific activities. Publications are the clearest example of scientific output and the trend in number of publications and elements associated with publications such as citations, has and is being used as a measure of scientific progress. However, the objective of science is not to publish more nor to cite published papers more – the objective is to increase our knowledge.

There have been several attempts to use Shannon entropy as the foundation of measuring understanding (Fanelli 2019), but none have been adopted by practicing scientists. The fact that we have no quantitative measures of understanding or increases in understanding implies that measuring understanding has eluded scientists. Thus, science has a clear objective, but we can't estimate how well science is meeting its objective.

But there is a way to measure scientific progress - prediction – predictive ability (Figure 3). Predictive ability is without doubt an imperfect and imprecise measure of understanding but is adequate while we wait for something better.

**Applying predictive ability as a measure of understanding:** The distance between zero and perfect understanding (i.e. the truth) is the distance science has to travel. However, we don't, a priori, know what the truth is and it's difficult to conceptualize what it means to have zero understanding. But – if point B represents the prediction error we would make if we had perfect understanding and point A represents the prediction error we would make with zero understanding, we can plot these in 2-

dimensional space and quantify scientific progress using predictive ability (Figure 3). That is, if we move halfway from Point A to Point B, we will have 50% understanding of the response variable representing the phenomenon of interest. The core assumption here is that the relationship between predictive ability and understanding is linear, which may not be true. However, let's begin the project of discovering the relationship between predictive ability and understanding.



Figure 3: A two-dimensional space for representing scientific progress. A is predictive ability with zero understanding and B is predictive ability with perfect understanding.

So, we have to know two things to use predictive ability to measure scientific progress – (1) how large our prediction error would be if we knew nothing about a particular phenomenon and (2) how large our prediction error would be if we had perfect understanding of a particular phenomenon. Perfect understanding implies that we know every causal mechanism, the correct functional form describing the relationship between each causal mechanism and the

response variable and the size of the effect that each causal mechanism has on the response variable. We will rarely if ever achieve perfect understanding (i.e. zero prediction error) but the distance we move from the prediction error with no understanding towards zero prediction error is a quantitative measure of scientific progress about a particular phenomenon/response variable (Figure 4).

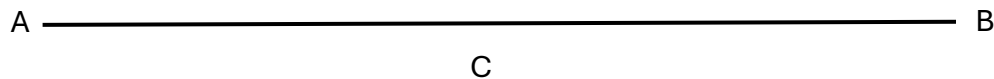


Figure 4: A two-dimensional space for representing scientific progress. A is predictive ability with zero understanding, B is predictive ability with perfect understanding and C is predictive ability with our current 'best' model.

*Predictive ability with zero*

*understanding:* The concept of 'zero understanding' overlaps with the concept of null models and there has been fierce debate in ecology for the last several decades about the definition and utility of null models. Gotelli and Graves (1996) say,

*"A null model is a pattern generating model...designed to produce a pattern that would be expected in the absence of a particular ecological mechanism."*

This definition doesn't imply zero understanding, rather it implies zero understanding about the causal mechanisms that lead to variation in the phenomenon of interest. And zero understanding of a particular phenomenon versus zero understanding of the causal mechanisms of a phenomenon are very different ideas. So, there is clearly the potential for disagreement among ecologists about what zero understanding means and how it gets converted to prediction error. An example may better illustrate this point.

Let's imagine that we want to predict bird species richness in a 100-hectare forest fragment in southern New Brunswick. What does it mean to make

a prediction with zero understanding? Does it mean that the number of species could be any integer between  $-\infty$  and  $+\infty$ ? No. Because even with zero understanding of bird species richness we know that the actual number of bird species cannot be less than zero. So, zero understanding could mean that any value between 0 and  $\infty$  would be equally probable. However, even without knowing anything about the mechanisms that drive bird species richness, we can place tight constraints on the range of possible values. There are approximately 18000 bird species in the world. A prediction made with zero understanding would simply choose a random integer between 0 and 18000. But we can place even tighter constraints without incorporating any mechanisms causing variability in bird species richness. For example, the total number of breeding bird species in New Brunswick is ~350, so we could constrain bird species richness in a fragment to be between 0 and 350. Further, researchers may have a list of many forest fragments in New Brunswick for which bird species richness is known and the predicted value for the unknown forest fragment could be selected by choosing a value

at random from the set of all forest fragments where bird species richness is known. So, there are several options for 'predictions based on zero understanding of causal mechanisms.

Certainly, the choices described above contain different amounts of knowledge – from knowing absolutely nothing about bird species richness on earth, to knowing the total number of bird species on earth, to knowing the total number of breeding bird species in New Brunswick, to having a sample of the distribution of bird species richness' in New Brunswick forest fragments. But what all of these choices have in common is that there is no explicit causal mechanism. Any null model must exclude a causal mechanism that can be used to predict bird species richness. Beyond that, the choice of null model is a technical one – analogous to the way that temperature scales differently if we use degrees Kelvin, Celsius or Fahrenheit. Different

nulls simply set the prediction error associated with 'zero understanding' at different points. Thus, the selection of a 'null' model will be somewhat arbitrary – ideally, researchers will agree on a null model, but it isn't necessary.

However, as our discipline matures the point from which to assess progress in ecology about some phenomena would not be from 'scratch'. That is, zero understanding wouldn't be our starting point. Instead, we would estimate the scientific progress of a new ecological model relative to our 'best' understanding before the new model was developed. This would simply involve comparing the prediction error associated with the old 'best' model relative to the prediction error associated with the 'new' model. If the 'new' model makes better predictions than the old 'best' model, then our understanding has increased and there has been scientific progress.

**PRINCIPLE #7 OF GOOD SCIENCE:**  
*Distinguish between causal and correlational relationships.*

Predictive ability cannot address the issue of causality, because predictive ability arises from strong correlations between the phenomenon of interest and putative causal mechanisms. But – correlation does not confirm causality. Thus, models can make good predictions without containing understanding...if the 'causal' mechanisms aren't causally related to the phenomenon of interest. There are four reasons that two variables - A and B - can be correlated. First, because A has a direct effect on B, second, because B has a direct effect on A, third, because A has a direct effect on C and C has a direct effect on B. and fourth, because C has a direct effect on both A and B. Only, the first and third reasons imply a causal relationship of A on B and only the first implies a direct causal relationship.

Ecological models intended to contain our understanding of the world must only contain causal mechanisms.

There are contexts where all we want are good predictions – we don't care

whether the model captures true understanding – but understanding requires that we can differentiate between true causal relationships and spurious or confounding relationships (i.e. reasons 2 and 4 for correlations between A and B). And predictive ability alone cannot assess whether relationships are causal or spurious. Historically, controlled experiments have been seen as the ideal, if not only, way to assess causality. This, in part, explains why, even in ecology, where we are trying to explain the distribution and abundance of living organisms in nature, there has been a great deal of emphasis on controlled experiments. However, in recent decades there have been tremendous advances in our ability to infer causality using observational data obtained in uncontrolled settings (Pearl et al. 2016). But whether our understanding of causal links comes from controlled experiments or 'causal' analyses of non-experimental observations, it is critical that any claims to understanding based on the predictive ability of a model must also provide evidence for causal links between putative drivers and response variables.

**PRINCIPLE #8 OF GOOD SCIENCE:**  
*Model building is better than hypothesis testing.*

The scientific method generally is seen as the cornerstone of scientific knowledge. The 'Method' leaves room for answering both 'what' and 'why' questions. The inductive steps of the scientific method often include research that results in descriptive knowledge, and the descriptions provoke 'why' questions leading to plausible answers (i.e. hypotheses). The emphasis in ecology over the last several decades has been on the hypothesis-testing portion of the scientific method. In particular, many ecologists found Karl Popper's argument that science is distinguished from non-science by the falsifiability attribute, compelling. Popper proposed that "scientific" hypotheses were falsifiable. Falsifiable hypotheses identified observations that could disprove them, while unscientific hypotheses had no observations that could be interpreted as inconsistent with the hypothesis.

Shortly after Popper's views found a foothold in ecology, Platt followed with his classic 1964 paper 'Strong Inference', which made the argument that the most efficient path to scientific progress was devising crucial experiments to test competing hypotheses. So, two of the most influential commentators of the last 75 years insisted that hypothesis testing must be the fundamental building block

of scientific progress. But hypothesis-testing has profound limitations.

The primary limitation of testing hypotheses is that models contain all of our understanding of the natural world and hypotheses are almost always verbal models. Hypotheses rarely allow for precise, quantitative predictions. Further, hypothesis-testing is binary – that is, the data either are evidence for or against a hypothesis. This approach to acquiring understanding may be effective when scientists are investigating simple phenomena – phenomena with one or a few causal drivers that have linear and non-interactive effects. However, phenomena that have complex dynamics – many interactive drivers and nonlinear effects - the number of plausible hypotheses and the number of crucial experiments that are required becomes enormous and perhaps, for some phenomena, infinitely large.

When we place hypothesis testing at the center of scientific discovery we constrain scientific research to a series of binary decisions about vague non-quantitative models. The result is a set of loosely connected scraps of understanding that have to be pieced together into a coherent model for any particular phenomenon. Meta-analytic techniques have been used to attempt to integrate scientific research but have been generally unsuccessful at providing coherent models of natural phenomena.

Hypothesis testing can lead to predictive models in two ways – (1) by choosing between/among very precise hypotheses (i.e. models) that describe causal mechanisms, functional relations and strength of relationships or (2) by testing very general hypotheses, usually specifying a single causal mechanism with no mention of functional forms or strength of relationships. The first approach is never used in ecology. Precise quantitative hypotheses are rare

in ecology – we usually test general, qualitative hypotheses using a null-hypothesis significance testing (NHST) framework. The NHST framework does not lend itself to building predictive models.

An iterative model-building-and-refining approach makes more sense than comparing a near-infinite set of tenuously connected hypotheses with the objective of creating coherent theory.

**PRINCIPLE #9 OF GOOD SCIENCE:**  
*Identify the current best model.*

Moving from testing hypotheses to a model-building approach does place a heavy burden on scientists for preparation before beginning research and for communicating after research is completed. Because researchers must know the current best model for the phenomenon they are working on. And if researchers believe they have improved the model after they have completed their research, they have to make the case that there is a new 'best' model.

That scientists should know the current 'best' model for any phenomenon they are studying may seem obvious to non-scientists. In my discipline – ecology – we do not have a 'best' model for a single phenomenon of interest. Not even at an explicitly defined spatial location or over a specific time period. Pick an ecological phenomenon – population fluctuations, species richness, productivity – and I defy anybody to identify the current 'best' model capturing our understanding of any of these phenomena. In fact, I defy anybody to identify a model (not necessarily the 'best' model) regularly used by ecologists that can make precise, quantitative predictions at any spatial, temporal or taxonomic scale. And remember – I'm not asking for accurate predictions...just precise and quantitative predictions.

To make precise quantitative predictions a model must have a measurable

dependent variable, at least one measurable independent variable, explicitly proscribed functional relationships and estimates for all parameters. To my knowledge, such a model does not exist in ecology. Not for species richness. Not for population size. Not for productivity.

One explanation might be that general models do not exist in the natural world. In fact, Tony Ives famously (by ecologist's standards for fame) asked the audience at his 2013 MacArthur lecture "Should ecology be about general laws". And to be sure they thought about it, he left the stage. Only 1/3 of the audience agreed that ecology should be about discovering general laws. To be frank, I'm not sure why the audience's opinion matters – the question of whether there are general laws to be found in ecology is an empirical not a philosophical one. Either they exist or they don't.

But even if general laws don't exist in ecology, all of our understanding still lies in our models. It's just that those models may have to be system-specific. But I am unaware of any consensus among ecologists on a single system-specific model that can be used to make precise, quantitative predictions.

The result is that ecologists – as far as I can tell – have almost no shared understanding of how the world works. Are populations density dependent? Who knows? Does productivity affect species richness? Maybe. But maybe it's species richness that affects

productivity. Or maybe there's a feedback loop. Blah, blah, blah.

If we were to shift from testing hypotheses to model-building, introductions to research papers would have two functions – first, to provide the authors' case for what the current 'best' model was and second, to identify how the proposed research might improve that model. That's it. By the end of the introduction, readers would be explicitly

told what the current 'best' model was and whether the authors were going to add a new causal variable, remove an old causal variable, replace the current function relationship(s) with something better or provide improved parameter estimates.

The reason we don't do this now is because it's hard. But I'm not sure how we can argue that it wouldn't be beneficial.

**PRINCIPLE #10 OF GOOD SCIENCE:**  
*Define the scale and scope of your models.*

Scale and scope are related but independent concepts. The scale of a model is defined by the grain and the extent of the sampling used to collect the data that the model was trained and tested on. The scope of a model is the transferability (i.e. the generalizability) of the model (Figure 5).

I focus here on spatial, temporal and taxonomic scale though there are others. Spatial grain is the area of the sampling unit, and the spatial extent is the total area that was sampled. For example, if we sampled plant species richness in one-meter square quadrats over a 100-hectare old field, the grain would be 1 meter squared, and the extent would be 100 hectares. Temporal grain is the time between samples and temporal extent is the total time period over which sampling occurred. For example, if we sampled plant species richness every month for ten years, the temporal grain would be one month, and the temporal extent would be ten years. Taxonomic grain is the finest level of taxonomic classification that individuals are assigned to, and taxonomic extent is level of classification that is used to decide which individuals will be sampled. For example, if plant species are identified to subspecies and we sample all angiosperms, the taxonomic grain is subspecies and taxonomic extent is Phylum.

The grain of a model is particularly important because the true model for a particular phenomenon at small spatial, temporal and/or taxonomic scales may be different than for that same phenomenon at large spatial, temporal and/or taxonomic scales. For example, the model that captures annual species richness at 1 meter squared may be very different than the model that captures decadal genera richness at 1 square kilometer. We must explicitly identify the spatial, temporal and taxonomic scales of our model.

Scope defines the spatial, temporal and taxonomic transferability of a model. Transferability is defined by the way that prediction error changes with distance from where the model was trained and tested. Spatial transferability is the ability of a model to make good predictions in different places. For example, the spatial scope of the plant species richness model is 1 kilometer if it can make good predictions in an old field up to one kilometer away from the old field the model was trained and tested on. Temporal transferability is the ability of a model to make good predictions at different times. For example, the temporal scope of the plant species richness model is one year if it can make good predictions up to one year away from when the model was trained and tested. The taxonomic scope of the plant species richness model is to the kingdom if it can make good predictions up to all communities in the kingdom Plantae.

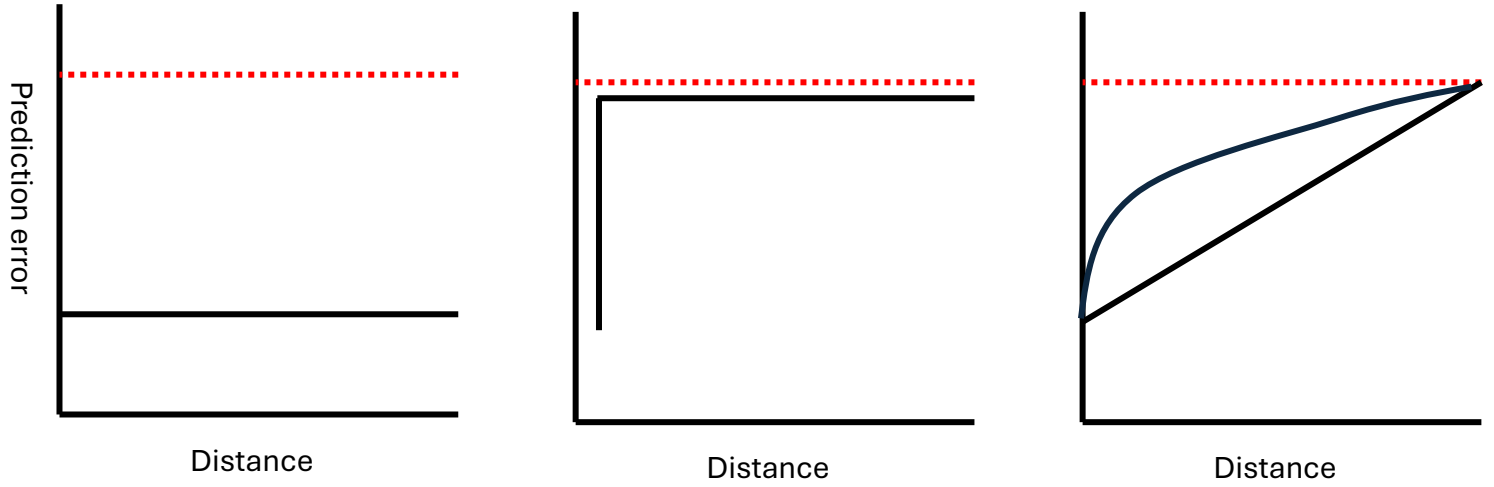


Figure 5: (A) complete transferability; (B) Zero transferability; (C) transferability degrades with distance (linear or nonlinear). The red broken line is prediction error with zero understanding, and the black solid lines are prediction error with the model.

**PRINCIPLE #11 OF GOOD SCIENCE:**  
*Models must be transferable in time.*  
*It's nice if they are transferable in*  
*space and across taxonomic groups.*

The scope of a model defines its generalizability. If a model isn't transferable in time – at least, to a limited extent – it's impossible to know if it no longer holds information or it never did.

A simple thought experiment illustrates the necessity of temporal transferability. Imagine a model that perfectly captures an ecological phenomenon and makes perfect the training set, but *one second later* no longer contains information about the phenomenon and makes predictions that are no better than a random guess. There would never be an opportunity to demonstrate that the model contained knowledge of the

natural world because the information contained in the model would be so ephemeral.

So, for a model to be useful it needs to be temporally transferable, but more than that – to be able to demonstrate that it ever contained knowledge it must be temporally transferable. The same isn't true for spatial and taxonomic transferability. We would like a model to be able to make good predictions to other places than where it was trained and tested, but if it isn't, it simply means that the spatial scope of the model is small. Similarly, if a model only makes good predictions for a particular subspecies, it simply means its taxonomic scope is small, but we have knowledge and understanding of the particular subspecies.

**PRINCIPLE #12 OF GOOD SCIENCE: *Know exactly what part of model building your research will be addressing.***

If I can make two assertions –

1. The objective of science is to increase our knowledge and understanding of the natural world.
2. All of our understanding of the natural world is contained in our models of natural phenomena.

Then I can conclude that we should not do any research that will not directly or indirectly lead to comparing a ‘new’ model to an ‘old’ model. This implies that we can identify some kinds of research as better than others.

Fox (2025) argues that a diversity of research approaches is a key strength of my discipline – ecology. The elegant metaphor is to the importance of living diversity for ecological function in the natural world. However, the defining characteristics of life are diversity AND unity. The remarkable diversity of the life we see around us is built on a genetic foundation that is structurally identical across all living organisms. Similarly, the diversity of scientific research approaches needs a unifying foundation that will guide scientists into research approaches that are most likely to lead to knowledge and understanding. The two assertions at the beginning of this section provide the basic foundation - the remaining ten

principles of good science construct a framework for guiding scientific research. And they imply specific avenues for scientific research. Below are the kinds of research that are designed to meet the objective of science. The list might not be exhaustive...but it's close.

Types of scientific research

Potentially leading to models for new phenomena

1. New and/or better descriptions of the world. These descriptions of the world are what provoke questions about why or how the patterns we observe come to be.

Potentially leading to direct improvement of models

2. Effects of new variable(s) or new interactions among variables.
3. Effects of new or improved functional relationship(s)
4. New or improved parameter estimate(s)

Indirect improvements of models

5. Improving measurement accuracy and precision
6. Development of new statistical techniques that bridge the gap between data and models

7. Estimates of spatial, temporal and/or taxonomic scope of a model
8. Theoretical models that identify potential new variables/functional relationships/interactions to test
9. Tests of causality
10. Linking models into a larger theoretical network of knowledge.

If your planned research doesn't fall into one of these categories, you should ask yourself – "How will my research potentially improve our understanding (i.e. models) of the natural world?"

These dozen PRINCIPLES are more philosophical than technical. Let's move to the technical characteristics of any research project or program that will be most likely to result in improved models.

## TECHNICAL CHARACTERISTICS OF GOOD SCIENCE

Some of what I assert in the preceding section may be seen as provocative because there are places where it contradicts what many scientists would consider to be good science (e.g. stop testing hypotheses).

By contrast most of my claims about the technical characteristics of good science are 'motherhood' statements – they will seem obvious to most practicing scientists. However, every year for a decade or so, I led a couple of graduate student seminars on the philosophy of science, and I always began the second seminar by asking "What are the characteristics of good science?". (The question before the first seminar was "What separates science from non-science?"). I was invariably surprised that nobody ever mentioned sample size or measurement error or representative sampling. The suggestions were always much higher level – 'unbiased', 'transparent', 'objective' – characteristics that were difficult to define and operationalize. I don't know if these technical characteristics were ignored because they are so obvious or because in the context of a 'philosophy of science' class the students' attention was elsewhere. Or are we losing track of how critical these technical components of good science are?

If science is the unwavering commitment to understanding the natural world through empirical observations, then the available

understanding in any study is buried inside of the observations (i.e. data). Further, the amount of understanding that can be squeezed from the observations depends on the quantity and quality of the observations.

While the previous section identifies what I believe should be the foundational cornerstones that support the diversity of scientific approaches, I also hold that probably the greatest barriers to understanding the world in my discipline – ecology - are small sample sizes, measurement error and unrepresentative sampling. I am convinced that adopting a modelling and prediction approach will help us to confront these technical and methodological concerns. But large sample sizes, small measurement error and representative sampling would almost certainly lead to greater leaps forward in ecological understanding than shifting from hypothesis testing to model building.

The solution to many of the technical concerns is additional resources. Meanwhile, increased awareness of and commitment to these technical characteristics can mitigate some of the problems. Even if the increased attention only gets us to the place where early-career scientists identify large sample sizes, small measurement error and representative sampling as characteristics of good science.

## DATA COLLECTION

### *Realism of system*

It's generally accepted that the gold standard for data is observations collected under controlled experimental conditions in the places and at the spatial and temporal scale to which you want to draw inferences. However, these kinds of data are difficult and expensive to collect. We have a few examples in ecology including the Experimental Lakes Area in northwestern Ontario and the Hubbard Brook Experimental forest in New Hampshire.

The usual tradeoff in ecology is between small-scale experimental studies and large-scale observational studies. The first sacrifices realism for the ability to make strong causal inferences and the second sacrifices strong causal inference for realism. Sacrificing realism is a greater problem than sacrificing causal inference.

Ultimately, we need models that make accurate predictions to natural systems. Thus, model building, refinement and most importantly validation must be done on observational data collected on natural systems.

However, small-scale experiments can play an important role in confirming causal links and identifying potential causal variables and functional relationships to be validated in natural systems. But anybody doing small-scale artificial experiments must explicitly discuss how their research is

relevant to improving models of natural systems.

My lab did whole-system and microcosm experiments on the impacts of glyphosate on wood frog larvae and the small-scale experiments told us almost nothing about what we found in the whole-system experiments.

### *Small measurement error*

There is little doubt that one of the major constraints on model-building and refinement is measurement error. Measurement error sets the ceiling on model predictive ability. Even if we had the true model for a particular phenomenon, measurement error would prevent perfect predictions because the observations of both dependent and independent variables would be inaccurate. And the prediction errors will be a positive function of the size of the measurement error. One fundamental phenomenon studied in ecology is population size and how it changes over space and time. However, population monitoring at large scales in natural systems is labour-intensive and expensive. And even when effort and money is invested it is very difficult to validate population estimates because validation requires comparing estimated values with known population sizes...and we are rarely able to provide known population sizes at large spatial scales in natural systems. Published ecological research

rarely estimates or reports measurement error.

Relatively little research has been done in ecology to estimate the measurement error on population estimates. The same issues arise with other fundamental ecological phenomena such as species richness or population/community biomass. More effort and money should be invested in estimating measurement error and developing sampling methods that reduce measurement error.

### Small sampling errors 1 and 2

Sampling error is more complex than it appears at first glance (at least, it was for me). And though I've never seen this mentioned in the literature, I've identified two kinds of sampling error.

First, is the traditional concept of sampling error – the error that occurs when we sub-sample a population to estimate a summary statistic (e.g. mean or standard deviation). This error occurs because the set of subsamples that you take is rarely – if ever – perfectly representative of the population. For example, we can sample and weigh twenty lake trout and use their mean weight as an estimate of the average weight of lake trout in the lake. But the estimated mean weight is unlikely to be precisely accurate because the twenty sampled fish are unlikely to be perfectly representative of the entire population of lake trout. But of course, the more subsamples you take, the closer the

parameter estimate will be to the true parameter value.

However, the concept of sampling error doesn't transfer perfectly to data collection for the purpose of building models. Because when we are using data to build models we are not trying to find the 'average' model for the phenomenon we're studying. Rather we are trying to discover the true model for the phenomenon of interest. And we believe that the true model operates homogeneously across our study area and study period. For example, if lake primary productivity (as measured by chlorophyll a levels) is a simple function of total phosphorus and we sample chlorophyll a and total phosphorus in 100 lakes, sampling error at the level of the 100 lakes sampled isn't relevant. Because we are not trying to estimate the 'average' true model across all lakes using a subsample of 100 lakes. We believe the same model is operating in each lake and thus any subsample that is without measurement error will capture the true model.

However, sampling error is a concern for the subsamples that are used in each of the 100 lakes to estimate mean chlorophyll a and mean total phosphorus. Because sampling error will lead to inaccurate estimates of mean chlorophyll a and mean total phosphorus, sampling error will reduce our ability to identify the true causal variable(s), the functional relationship(s) and make precise accurate parameter estimates. So, sampling error isn't a

concern for capturing the true underlying model, but it is a concern for estimating the values of inputs to and outputs of the models. And errors on the estimates of inputs and outputs will make it more difficult to discover the true, underlying model.

However, there is a second kind of sampling error that I have rarely, if ever, seen discussed in ecology. This error occurs if you don't sample the entire range of values found for the causal variables. This leads to errors in the model because if samples don't cover the entire range of values for causal variables it may be difficult to identify the true nonlinear functional relationships or identify and estimate the strength of interactions among causal variables. For example, if mean total phosphorus ranges from 1 ug/L to 100 ug/L and the relationship between mean chlorophyll a and mean total phosphorus is linear up to 70 ug/L and reaches an asymptote at values above 70 ug/L but our sample of lakes only covers a range from 1 to 50 ug/L we will not discover the asymptote.

The solution to the first sort of sampling error is increased sample size, but increased sample size will not necessarily solve the problems caused by the second kind of sampling error.

### *Large sample size*

Of course, sample size and sampling/measurement error are inextricably linked. Large sample sizes minimize the impacts of sampling and

measurement error and increase the precision of model parameter estimates. This isn't news to anybody doing science, but it's not always obvious to me that people worry about sample size enough.

Sample sizes are usually provided in published ecology research papers...but it's rarely discussed critically unless the authors did a power analysis. In my discipline, most scientists simply accept the sample size imposed on them by effort and financial constraints without assessing whether the sample size they use will allow them to answer their question. I have rarely, if ever, seen explicit discussions in ecological research of the consequences of sample size for particular papers. My intuition is that 50 – 100 samples of one or two causal variables is better than 10 to 20 samples on five or ten causal variables. But all I have is intuition.

### *Representative sampling*

When we subsample a population to make inferences about the population, we assume that our subsamples are representative of the population. If the assumption is true, then our only concerns are measurement and sampling error. If the assumption isn't true, it means we expect to make useful inferences to one population by sampling from a different population.

My intuition is that it's rare that the sampled population is completely independent of the population to which

we want to make inferences. But I also believe that there is rarely complete overlap between sampled populations and the populations to which we want to make inferences.

But perhaps the largest problem is that I have to rely on intuition because the question of how representative our sampling is is almost never addressed.

For example, ecologists readily acknowledge the difficulty of identifying precisely the spatial boundaries of populations. However, ecological papers studying populations rarely discuss the issue of the specific spatial boundaries of their specific research site. My explanation for this lack of discussion is that when there is a research problem without a solution you ignore the problem.

Good science requires that we make an explicit case that our samples are representative. This is rarely done in my discipline, but when ecologists use stratified sampling designs it is implicit evidence, they were concerned about representative sampling and there is often explicit evidence of their concern for representative sampling when they describe their stratified sampling design.

### *Small differences between measured variables and conceptual variables*

There is often a difference between the ecological phenomena we make inferences to and the variables we measure. The same is true for the causal mechanisms and the measured variables. For example, ecologists

estimate trophic position using stable isotope ratios and gene expression using fluorescence intensity, variables that are quite distant from the phenomena we are studying. By contrast, species richness is usually measured by counting the number of species in the study area – the measured variable is identical to the phenomenon of interest.

We almost always use indirect measures of phenomena of interest when direct measures are difficult or impossible to collect. These indirect measures are often very clever but usually require one or more assumptions to be true to be good proxies for the phenomena of interest. Indirect measures are rarely, if ever, as good as direct measures.

Scientific research should include explicit discussion of the link between what we measure and what we make inferences about.

### *Small difference between scale of sampling and scale of inference*

This is relatively rarely a problem for observational studies – generally ecologists sample at the spatial, temporal and taxonomic scales to which they want to make inferences.

However, in manipulative experiments logistics often constrain ecologists to temporal and/or spatial scales that are smaller than those to which we want to make inferences.

### *Sample few variables*

Our recent research shows that as we increase the number of variables in the model selection process the probability of choosing the 'best' model declines dramatically. For small-scale experimental studies there is rarely a good rationale for measuring many variables other than the manipulated variable(s). However, for observational studies, I understand the temptation to measure a large number of variables in the field. The rationale is usually that – for efficiency's sake – we sample everything we can, so we don't have to use labour and financial resources on multiple trips. But collecting many 'potential' causal variables and including them in the model selection process reduces our ability to find the 'best' model. Particularly, when the list of sampled variables includes some that have little or no effect on the phenomenon of interest.

Of course, the appropriate number of independent variables to sample will depend on sample size. Unfortunately, little research has been done to estimate the optimal sample size: sampled variables ratio. Peduzzi et al, (1996) is usually identified as the source of the 10:1 rule of thumb that there should be 10 observations for every parameter estimate in a logistic regression. This lack of research is illustrated by the fact that ecological research rarely, if ever, discusses the ratio of observations to parameter

estimates. And I routinely see examples of 5 – 10 parameter estimates with 30-50 observations. When we increase the number of putative causal variables included during model selection without increasing sample size the result will be poor parameter estimation and model selection.

### *Sample so that transferability can be assessed.*

We can assess transferability if data are assigned to training and test sets systematically rather than randomly. Training and test sets can be assigned systematically if individual data points are labelled in time, space or taxonomically. By contrast, if all sampled data are labelled as if they come from a single point in time, a single place or a single taxonomic group, it is impossible to assess temporal, spatial or taxonomic transferability.

This implies (1) providing temporal, spatial and/or taxonomic 'coordinates' for each data point and (2) variation in temporal, spatial or taxonomic coordinates such that data can be assigned to training and test sets based on the coordinates.

Ideally, we would identify what kind of transferability we wanted to assess before collecting our data and then design the sampling protocol accordingly. For example, if we wanted to assess whether a model was transferable across years, we would sample data in two or more years, use

one of the years as a training set and the remaining year(s) for testing. Similarly, if we want to assess whether a model was transferable in space, we would sample two or more sites that were separated in space and use one of the sites as the training set and the remaining site(s) for testing.

Note that I've focused throughout on temporal, spatial and taxonomic transferability, but there are many other dimensions along which models could be assessed for transferability. Examples include life history characteristics such as life stage or sex. It would be useful to know if models for adults made good predictions for juveniles or models for females made good predictions for males. Or environmental context such as habitat type or altitude. It would be useful to know if models built in forests made good predictions to grasslands or if upland models made good predictions to lowlands.

### *Regression rather than ANOVA sampling design*

"ANOVA"-style experimental designs - where the causal variables are assigned to categories even though the variables are continuous - are commonly used in

ecology. We should never group continuous variables...and most variables are continuous.

The problem with ANOVA-style designs is that they aren't designed to estimate parameters or to identify functional relationships. They are designed to test hypotheses. If we agree that our objective is to find the 'true' model underlying any phenomenon of interest, then we should collect data so we can identify functional relationships and estimate parameters. This means – for both observational and experimental studies – we should be sampling along the plausible range of values that would be seen in nature for each potential causal variable. Ideally, with a large enough sample size to detect non-linear relationships and interactions.

### *Assess causality*

The gold standard for assessing causality is manipulative experiments. However, in recent decades statistical analyses including structural equation modelling techniques have been designed that allow us to assess causality. Ultimately, we have to demonstrate that predictor variables have causal links to the phenomena of interest.

## DATA ANALYSIS

### *Have a training and test set*

There should always be a training set to which the data are fit and a test set which is used to validate candidate models. I don't recommend k-fold cross validation for breaking data into training and test sets. For two reasons.

First, k-fold cross validation tests a series of models against test sets that are randomly chosen. In the most extreme example – leave-one-out cross validation – n different models are trained on n-1 data points and tested against the remaining data point. Thus, there is no 'best' model because each of the n models has slightly different parameter estimates. If we choose parameters based on the weighted average of all parameter estimates we end up with parameter estimates that are approximately the same as what the parameter estimates would have been if we fit the entire dataset (I think. I haven't checked, but I think this is true). This does not provide an independent test of the predictive ability of models. It is really a test for outliers.

Second, temporal transferability is necessary for a model to be useful and spatial and taxonomic transferability are useful. So, collect data that allow transferability to be assessed and then use that design to estimate predictive ability.

### *Provide rationale for how data is split into training and test sets.*

There is substantial research available to guide selection of training and test sets. Ultimately, the ideal is to archive a consensus 'test set' for each phenomenon of interest, which would not be used to train models. These archived 'gold standards' would be used to provide a consensus on scientific progress for all scientists in a particular discipline.

However, this is a long-range objective and until then we will have to provide a clear rationale for the way we divide data.

### *Standardize data for training and test set where appropriate*

Temporal, spatial and taxonomic transferability implies applying models to places, times and groups of organisms that were not used to build the models. Because ecological models are incomplete – there are no ecological models that include all causal variables, exclude all non-causal variables, have all functional relationships correct and have precise accurate parameter estimates – when we transfer models there may be fundamental differences between the training context and the testing context that are not captured by the models. This will lead to poor predictive ability. However, standardizing the dependent and

independent variables so that predictions are in standard deviations rather than absolute values can provide improved predictive ability.

In addition, when transferability is improved by standardizing it implies that there are factors operating at a scale that create fundamental differences between the place/time/taxonomic group used in the training set and those used in the test set. Identifying those factors will lead to fundamental improvements in ecological models

### *Analytical techniques are interpretable*

The objective of science is to understand how the natural world works. This implies that the analytical techniques we use should create models that we can interpret. That is, models with (i) an explicitly defined response variable, (ii) explicitly defined causal variables, (iii) explicitly defined interactions among causal variables, (iv) explicitly defined functional relationships, and (v) precise parameter estimates.

Two types of analytical approaches that are not easily interpretable are (i) many machine learning approaches including even relatively simple ones like random forests, gradient boosted regression and neural nets, and (ii) ordination techniques like principal components analysis, correspondence analysis, canonical correspondence analysis, and redundancy analysis.

Machine learning models have been very successful in making predictions in some areas. However, the underlying causal variables, the underlying functional relationships, the interactions among causal variables and the strength of relationships between causal variables and the phenomenon of interest are very difficult to extract from the analyses. We have a model that can be used to make good predictions, but we have little comprehension of what the model is. That is, the model understands the world, but we don't understand the model. This is especially true of deep learning models, which often use dozens to hundreds of layers of nodes, with thousands of nodes per layer resulting in trillions of parameter estimates. The most important reason for simplifying models might be interpretability.

Ordination techniques rarely, if ever, provide a precise phenomenon or phenomena of interest, because the 'dependent variable' is usually a matrix of community characteristics that are unique to the training set. Unconstrained techniques such as principal component analysis, nonmetric multidimensional analyses and correspondence analysis are primarily visualization techniques that don't identify causal variables. Constrained techniques including redundancy analysis and canonical correspondence analysis do estimate how variation in community characteristics are related to environmental gradients, but do not identify functional relationships or

parameter estimates so it's not clear how they can be used to make predictions to new data.

There may be contexts where machine learning techniques and ordination techniques are useful exploratory techniques, but, in general, I recommend avoiding techniques that result in uninterpretable models or results that don't allow predictions.

*Analytical techniques that are able to identify causal variables, flexible enough to capture nonlinearities and interactions and able to estimate parameters.*

Suitable analytical techniques include generalized linear models and generalized linear mixed models where ad hoc techniques for model selection like p-value rejection thresholds or Akaike's Information criteria are used for model selection. In addition, Bayesian approaches can use informative priors to help with variable selection. Further there are analytical techniques such as regularized regression methods (e.g. LASSO), generalized additive models and multivariate adaptive regression

splines that integrate model selection, identifying functional relationships and estimating parameters. All of these techniques explicitly identify potential causal variables, identify functional relationships, interactions among variables and precisely estimate parameters, which allows the development of models that make precise quantitative predictions.

### *Assess causality*

While I believe that predictive ability is necessary to demonstrate understanding, I don't believe predictive ability alone is sufficient. Understanding requires that we confirm that relationships are causal. Historically, controlled experiments have been seen as the only route to demonstrating causality. And of course, controlled experiments at relevant temporal, spatial and taxonomic scales are still an excellent approach to demonstrating causality. However, over the last several decades analytical techniques such as structural equation modelling, regression discontinuity design and instrumental variables approaches.

## **MODEL CONSTRUCTION, ASSESSMENT AND IMPROVEMENT**

### ***Identify and estimate predictive ability of a/the Null model***

See Principle #6: predictive ability with zero understanding. If there is a general consensus in your discipline on an appropriate null model for your phenomenon of interest – use it. If not, identify an appropriate null model. Use the null model to make precise, quantitative predictions and estimate the prediction error (i.e. predictive ability) of the null model.

### ***Estimate predictive ability of old model to test data (raw and relative predictive ability)***

If there is a general consensus on the 'current' best model use that model to make precise, quantitative predictions for the test set. If there is no general consensus on the current best model, review the literature and identify the 'current' best model. Use it to make predictions to the test set.

### ***Estimate predictive ability of new model to test data (raw and relative predictive ability)***

Use the model built on the training set to make predictions to the test set and estimate the prediction error of the new model.

### ***Estimate increased understanding of new model as- (Mean\_Prediction\_Error\_Old – Mean\_Prediction\_Error\_New)/Mean\_Prediction\_Error\_Null***

If the new model leads to improved predictive ability, I conclude that the research has led to improved understanding of the phenomenon of interest. Further, conclude that the new model is the 'current' best model. If the new model doesn't lead to improved predictive ability, then conclude the old model is still the 'current' best model

## **SUMMARY**

***Science is the absolute commitment to describing and understanding the natural world using observations.***

All of the information we use to improve our descriptions and understanding of the natural world is found in the observations we collect. We extract the

information from our observations using analytical techniques. We use the information we've extracted from the observations to build models. We estimate the amount of understanding in our models by making predictions to new observations.

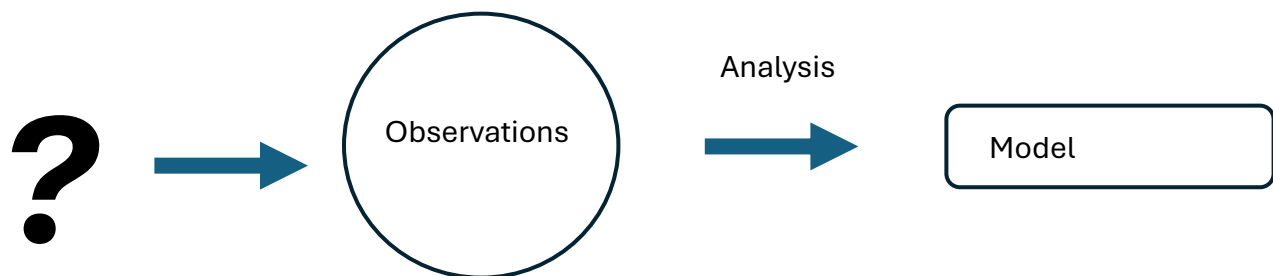


Figure 6: Scientific Method

The twelve principles of good science provide a general unifying vision for doing good science.

But ultimately increased understanding will come from (1) making observations that contain a large amount of 'true' information and little misleading 'information' (i. e. data collection), (2) extracting as much of the information in the observations as possible (i.e. data analysis) and (3) converting the extracted information into comprehensible, testable models (i.e. model construction, assessment and improvement).

Thus, the technical characteristics of good science that lead to high

information content in the observations, analyses that extract as much 'juice' as possible from the observations and result in precise quantitative models that assessed through predictive ability, are perhaps even more critical to scientific progress than the philosophical foundation provided by the twelve principles.

So, how could a research project or program be designed? Here's one defensible approach.

Step 1: Select an ecological phenomenon that you want to understand better.

Step 2: Identify the current 'best' model for the phenomenon of interest by reviewing the ecology literature.

Step 3: Identify gaps in the current 'best' model and target a particular gap.

Step 4: Collect data

Step 5: Analyse data

Step 6: Construct new model

Step 7: Compare predictive ability of new model and current 'best' model.

Step 8: If new model makes better predictions than current 'best' model, replace current model with new model

But of course, there are a large variety of research projects and programs that can indirectly serve the objectives of improving the predictive ability of ecological models.

## REFERENCES

Popper, K. 1959. *The Logic of Scientific Discovery*. Abingdon-on-Thames: Routledge.

Platt, J. R. 1964. Strong Inference. *Science* 347 – 353.

Fox, J. 2025. *The Diversity of Ecologists*. The University of Chicago Press. Chicago, USA.

Treleaven, S. 2024. A rock-star researcher spun a web of lies – and nearly got away with it. *The Walrus*.

Enserink, M. 2022. Star marine ecologist committed misconduct, university says. *Science*. Vol. 377, 699-700.

Kerr, N. L. 1998. HARKing: Hypothesizing after the results are known. *Personality and Social Science Review*. Vol. 2, 196-217.

Peters, R. H. 1991. *A Critique for Ecology*. Cambridge University Press, Cambridge, UK.

Houlahan et al. 2017. The Priority of Prediction in Ecology. *Oikos*. Vol. 126, 1-7.

Fanelli, D. 2019. A theory and methodology to quantify knowledge. *Royal Society Open Science*. Vol. 6, 181055.

Gotelli, N. J. and G. R. Graves. 1996. *Null Models in Ecology*. Smithsonian Institution Press, Washington, D.C.

Pearl, J. M. Glymour and N. P. Jewell. 2016. *Causal Inference in Statistics: A Primer*. John Wiley & Sons, Ltd. USA.

Peduzzi et al. 1996. A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology Vol. 49, 1373 – 1379.

## APPENDIX 1: GOOD SCIENCE CHECKLIST

The laws of good science and the technical characteristics of good science allow us to develop a 20-point checklist of 'good science'.

- 1. Do Science.**
- 2. Make precise quantitative predictions.**
- 3. Create models rather than test hypotheses.**
- 4. Create models that include causal variables, functional relationships and parameter estimates.**
- 5. Estimate prediction error of models relative to a null.**
- 6. Test for causality.**
- 7. Identify the current best model.**
- 8. Define the scale and scope of your models.**
- 9. Identify what part of model building your research will be addressing.**
- 10. Test transferability of model.**
- 11. Realistic**
- 12. Large sample size**
- 13. Small measurement error**
- 14. Representative sampling**
- 15. Small difference between measured and conceptual variables**
- 16. Small difference between scale of sampling and scale of inference**
- 17. Sample limited number of independent variables**
- 18. Estimate transferability**
- 19. Regression style experimental design**
- 20. Estimate amount of understanding in models**

\*I should emphasize that good science as I've defined it has one objective – to get closer to the truth of the natural world. Doing good science does not mean using science to do good. This implies that getting closer to the truth could do harm to humankind. Similarly,

good science – as I've defined it – says nothing about the ethics of the scientific research. That is, good science could be done in a way that is completely unethical and immoral. So the science we do must first be governed by ethical and moral principles before assessing whether the science is good or not.