OXFORD

Prediction in ecology and evolution

Andrew P. Hendry (iD)

Andrew Hendry (andrew.hendry@mcgill.ca) is an evolutionary ecologist at McGill University, Redpath Museum and Department of Biology, in Montréal, Québec, Canada.

Abstract

Prediction is frequently asserted to be the *sine qua non* of science, but prediction means different things to different people in different contexts. I organize and explain this diversity by addressing five questions. What does it mean to predict something? To answer this question, I describe concepts of prediction as prophecy, diagnosis, history, repeatability, and fate. What are we trying to predict? Here, I describe how predictions vary along several axes: general to specific, qualitative to quantitative, relative to absolute, point to range, and continuous to discontinuous. Where do predictions come from? In this case, I focus on deductive versus inductive reasoning. How do we test predictions? The answer here is not straightforward and I discuss various approaches and difficulties. How good are predictions? Not surprisingly, it depends on what is being predicted and how we judge success. Importantly, I do not espouse a "best" way to approach prediction but, rather, I outline its diverse manifestations so as to help organize practical thinking on the topic.

Keywords: forecasting, hindcasting, projection, accuracy, precision

In Madagascar, there must be moths with probosces capable of extension to a length of between 10 and 11 inches [25–28 cm]!

—Darwin (1862)

By the year 2000, the United Kingdom will be simply a small group of impoverished islands, inhabited by some 70 million hungry people.

—Ehrlich (1968)

Prediction is frequently touted as a primary goal of science, including in the fields of ecology and evolutionary biology (Peters 1991, Mouquet et al. 2015, Urban et al. 2016, Lässig et al. 2017, Maris et al. 2017, Burford Reiskind et al. 2021, Gompert et al. 2022). In particular, the concept of prediction is invoked as an aspirational or practical goal in numerous proposals and papers. For instance, an ISI Web of Science search (31 March 2022) for *prediction or predict* as a topic yielded 5379 papers in just *Ecology, Evolution,* and *The American Naturalist.* However, even a cursory examination of the many papers mentioning some form of prediction indicates that invocation of the term is highly variable in intent and meaning. I will provide many referenced examples of these diverse uses as my arguments unfold, but, for now, a few unreferenced signposts will serve to set the stage. One major arena for prediction can be found in the numerous and varied attempts to predict what the world will look like under a future of global change. Another arena for prediction is the use of surveys and statistics to infer whether a species is present (e.g., invasive species) or absent (e.g., endangered species) or to infer its abundance (e.g., for management of fisheries or hunting). In evolutionary biology, a common context for prediction is the use of heritability estimates to predict the evolutionary responses of organisms to natural or artificial selection.

A lot of excellent, rewarding, and important science does not so overtly invoke prediction as its goal. Instead, such basic science or curiosity-driven science has the implicit goal of simply understanding the way the world works. Even in this case, however, various forms of prediction are often involved. For instance, statistical testing in data-based science typically evaluates empirical support for formal predictions as a way of distinguishing between alternative hypotheses about how the world works. Similarly, mathematical theories are based on assumptions that come from expectations (a weaker form of prediction) of how the world works. Furthermore, past empirical outcomes of basic science or mathematical theories often form the basis for predictions in future studies. For instance, understanding the heritability of a trait is interesting in and of itself, but it can also be used to predict how that trait will change under a given form of selection. In short, although prediction is not synonymous with science, it does permeate science in a diversity of manifestations.

My goal in the present article is to examine and organize the various meanings of prediction and how they are used in ecology and evolution biology. I work toward this goal by attempting to answer five questions.

What does it mean to predict something? I will evaluate various concepts of prediction and organize them into five basic categories, three of which involve explicit prediction of either the future (prophecy), present (diagnosis), or past (history), and two of which involve implicit prediction in the sense of either repeatability or fate.

What are we trying to predict? Another way of phrasing this question might be *What are predictions good for?* I will explain the nature of prediction along five axes: general versus specific, qualitative versus quantitative, relative versus absolute, point versus range, and continuous versus discontinuous.

Where do predictions come from? I will explain the role of induction and deduction in diverse manifestations that permeate ecology and evolutionary biology.

How do we test predictions? Predictions can be made in the absence of subsequent testing, but testing them supposedly makes for "good" science. However, many predictions are extremely difficult to test, both conceptually and practically.

How good are predictions? To answer this question, I will explain how judging the success of a prediction depends heavily on what is being predicted and how we judge success.

These five questions are not independent in practice. For instance, the origins of a prediction (Where do predictions come from?) will influence tests of that prediction (How do we test predictions?), which could influence the success of that prediction (How good are predictions?). However, I have decided not to merge these questions because the correspondence is not one to one. For instance, a given origin for a prediction can lend itself to several types of tests that can have varying degrees of effectiveness, depending on the goals. I feel it is important to outline the various options for each question separately because they can be mixed and matched in various ways.

Before starting to answer the questions, I need to clearly frame my goals and intended audience while also making clear what I am not attempting to do. For starters, my article is not a philosophical argument along the lines of whether "good" predictions are possible, whether "good" science should require predictions, or what types of predictions or tests are "best." Similarly, I do not review how various philosophers of science have defined prediction, although I do provide citations that can be used as an entrée to that rich literature. Instead, my article is a pragmatic attempt to organize and operationalize the way prediction is currently being used in ecology and evolution, and so my intended audience is an inclusive mix of everyday practicing biologists focused on empirical questions. By my pitching this big tent, readers will on occasion think *Oh, that isn't the right way to do things* or *That isn't what I consider to be prediction.* That is fine, because my approach is more along the lines of "People are already doing this thing, and they are calling it *prediction*, so let's try to see how it fits in to the bigger picture."

Given this big-tent motivation, my writing often provides lists of examples that are signposts to diverse topics and studies in a particular category. These examples are far too numerous for each to have a detailed explanation and robust citation list, although I certainly provide both at important junctures. Instead, the idea is for diverse readers to perhaps spot their own research areas within these lists, which can help to orient their own thinking within the broader organization framework. Along the same lines, I cannot provide explanations for some of the listed methodologies (e.g., futures analysis, adaptive management, cross-validation), although I do provide citations for readers interested in more details. For all of these reasons, my article is quite different from previous discussions of prediction in ecology and evolution.

Definitions and explanations for how I use various terms are provided in table 1. Importantly, in this table, I am explicit about what is included in each term and therefore (by exclusion) what is not.

What does it mean to predict something?

When we say we are making a prediction in ecology or evolution, what, precisely, do we mean? It won't surprise readers when I say that *prediction* means different things to different people and, therefore, its definition will need to be hierarchical, starting broad enough to encompass the breadth of actual uses in our disciplines and then parsed into a set of more precise definitions for various forms of usage. At the broadest level, the key idea uniting all of the meanings and uses is that prediction about the state (of a system) or outcome (of a process or experiment) must be made before that state or outcome is known by the predictor. By contrast, after-the-

fact (i.e., not *prediction* but *postdiction*) interpretations of a result are fine, but they must be noted as such and not masqueraded as predictions that were tested in the study. Witness Gould and Lewontin's (1979) famous critique of post-hoc just-so stories in evolutionary biology. The reality is that just-so stories can be used to generate predictions for testing in further research, but they are not predictions tested within the study that spun the story. (Of course, another criticism in that famous paper was that many authors assumed but never tested adaptation itself—as opposed to, e.g., drift.)

I need to make a further distinction among the five concepts of prediction outlined below. The first three assume an explicit statement about an unknown future (prophecy), present (diagnosis), or past (history), whereas the last two can be considered implicit (need not be formally stated) predictions based on previous outcomes (repeatability) or presumed inevitabilities (fate). The two sets of categories are not mutually exclusive; for example, a prediction can involve an element of prophecy (about the future) that is derived from insights about repeatability (past outcomes in similar situations) or fate (a sense of inevitability).

Prediction as prophecy

"If I have eschewed the word *prophet*; I do not wish to attribute to myself such a lofty title at the present time."

—Nostradamus

The word *prophecy* can imply the utterance of a prophet or some other divine insight, but common use of the term simply means "a prediction" (Oxford Languages) or "a prediction of something to come" (Merriam-Webster). In this more general sense, prediction as prophecy might conform best to conventional interpretation and public perception. The idea is that we can know (from existing information) or guess enough about a given system to say what will happen in the future. Prophecy is the form of prediction that one finds in weather and election forecasts—albeit in the form of a probability (more about this later). In ecology and evolution, prediction as prophecy is particularly common in global change projections, with countless studies predicting range shifts or extinctions or phenological changes or shifts in community composition or any number of other future patterns (Urban et al. 2022).

Such a priori predictions of the future can serve two potential functions. (I repeat the common phrase *a priori prediction* but, of course, the *pre-* of *prediction* has the same meaning as *a priori*.) First, when faced with uncertain futures, we might wish to know which are most likely so that we can prepare for them—such as in scenario analysis or futures analysis (Duinker and Greig 2007, Urban et al. 2016). Examples include predicting invasive species (Gallien et al. 2010), anticipating forest pathogen outbreaks (Hudgins et al. 2017), and projecting sustainable fish harvests (Hyun et al. 2005). Some studies then further evaluate how particular decisions in the present might influence the probability of different outcomes in the future, such as by generating a series of projections on the basis of possible actions by management or government or societies. Importantly, any failure of prediction as prophecy is typically unknown until it is too late to do anything about it, and, therefore, the whole predictive process needs to be repeated for the next relevant time step or location. This iterative process sometimes plays out in nearly real time, such as in continuously updating adaptive management strategies (Allan and Stankey 2009).

Another motivation for prediction as prophecy is to formalize scientific honesty as idealized in the scientific method. That is,

Table 1. Definitions and explanations of terms as they are used in the present article.

Term	Definition and explanation
Hypothesis	An expectation about how a system works that can be used to generate a testable prediction that must be true if that hypothesis is correct.
Prediction	A formal assertion about a state or outcome before that state or outcome is known. For instance, one can predict the state of the world in the past, present, or future—as long as that state isn't known in advance of the prediction. Or one can predict what will happen in the future given a particular manipulation in the present. Note that the making of a prediction does not mean that the prediction will be tested—merely that it could be—or that the outcome will be apparent without the need for testing.
Expectation	Similar to a prediction, but weaker, implying higher uncertainty and less "surprise" should the expected outcome not obtain.
Question	A statement of what one would like to know about a system (past, present, or future). A question can be used to generate a prediction, or it can be used in lieu of a prediction. That is, answering a question does not require making a prediction about what the answer to that question should be.
Predictable	A state or outcome for which a prediction would be precise and accurate assuming adequate knowledge of the system. For something to be predictable, a prediction doesn't necessarily have to be made, nor—if a prediction was made—that it was accurate (leaving aside how accuracy would be assessed). The term "predictability" then would invite a quantitative assessment of just how predictable something is—or would be.
Prophecy (prediction as …)	When an explicit (that is, formally stated) prediction is about the future state of system. Note that the term *prophecy* does not require (and, in this definition, does not involve) invocation of the divine or utterance of a prophet.
Diagnosis (prediction as …)	When an explicit (that is, formally stated) prediction is about the current state of system.
History (prediction as…)	When an explicit (that is, formally stated) prediction is about the past state of system.
Repeatability (prediction as …)	When a similar set of conditions generates a similar outcome, it can be considered an implicit (need not be formally stated) prediction. A phenomenon that is repeatable would also be predictable if we knew enough about the important driving factors. The related term "repeatable" less clearly invites a quantitative assessment of just how repeatable something is or is expected to be.
Fate (prediction as …)	When logic or a wealth of evidence suggests that an event or outcome (past, present, or future) is inevitable—or nearly so, it can be considered an implicit (need not be formally stated) prediction. This term is not intended in the present article (nor in common usage) to imply absolutely surety—merely a sense of near inevitability. For instance, all indicators point toward a particular outcome and so, barring some completely unforeseen turn of events, that outcome seems almost predetermined.
Parallel or convergent (in the sense of evolution)	A specific case of "repeatable" that emphasizes the extent to which evolution generates similar outcomes under similar conditions from similar (parallel) or different (convergent) starting points.
Forecasting	Use of data from the past or present to make a prediction about the future.
Hindcasting	Use of data from the present or part of the past to make a prediction about another part of the past.
Uncertainty	The level of confidence in a prediction or in the outcome of a test of a prediction. Examples of quantitative measures of uncertainty can include probabilities of various alternative outcomes, predictive power (e.g., r^2 or likelihood) of various models, or confidence intervals for parameter estimates. Various types of uncertainty are outlined in Milner-Gulland and Shea (2017).

Note: These definitions variously overlap with (and deviate from) those stated by other authors—who often do not agree with each other. It would not be profitable to detail these various associations; rather, the definitions listed in the table are intended for interpretation within the context of the present article. Importantly, many of the definitions are stated such that they imply a yes or no answer (e.g., something is or is not predictable or repeatable)—and, in some cases, such an answer is sufficient. Most of the time, however, a more quantitative (e.g., how predictable and how repeatable) answer will be more appropriate (see the text).

scientists generate a hypothesis, make a prediction on the basis of that hypothesis, and then test that prediction with new data. If the data fail to support the prediction, then we are told the original hypothesis should be rejected and that we should generate a new hypothesis. One manifestation of this approach appears in frequentist statistics that infer support for a hypothesis only when an analysis leads the researcher to reject the null hypothesis of no effect at a chosen critical value (usually $p < .05$). A different manifestation appears in the Popperian approach of falsifying hypotheses, where the favored idea is used to generate a prediction. If a "severe" test fails to support that prediction, then we are told to abandon the hypothesis and seek a new one (Popper 1959, Chitty 1996). Of course, slavish adherence to focusing only on the statistical testing of such a priori predictions can cause a scientist to miss important but unexpected patterns in the data (Yanai and Lercher 2020). Furthermore, prediction as prophecy can be a temptation to dishonesty, such as hypothesizing after the results are known

(remaking the hypothesis to fit the data), the file-drawer problem (not publishing results that contradict the hypothesis), observer bias, confirmation bias (deemphasizing or not publishing contradictory results), and falsifying data to fit the original hypothesis. These and other issues associated with the statistical testing of predictions have been discussed at length from diverse perspectives (Jennions and Møller 2002a, Rosenberg 2005, Kardish et al. 2015, Parker et al. 2016, Fraser et al. 2018). Suggested solutions to these issues include preregistration of the experimental hypotheses, design, and analysis plan—along with the increasing use of "blind" observers (Parker et al. 2016).

Prediction as diagnosis

"I'm Al Gore, and I used to be the next president of the United States of America"

—Al Gore

Some readers might feel that prediction must be about the future; however, in common practice, this criterion is not required. For instance, many ecological and evolutionary studies predict a pattern that should be evident in nature if we assume a particular hypothesized mechanism is important; then they test that prediction using existing data from the present or past. Of course, it remains true that a prediction must be about an outcome (past, present, or future) that is not already known. In this sense, prediction as diagnosis is an attempt to infer the current state of a system, as opposed to some future state (as in prophecy above), which would be more akin to prognosis. In public perception, prediction as diagnosis might appear as answers to questions such as *Has a particular politician won an election?* or *Does a patient have cancer or COVID 19?* In ecology and evolutionary biology, it might appear as answers to questions such as *Is a given species present (e.g., an invasive species) or absent (e.g., an endangered species) from a location?, Is the productivity of a particular forest or a particular lake limited by phosphorous or by nitrogen?,* or *Are two groups of organisms the same species or are they different species?*

Prediction as diagnosis is often needed because the real (true) state of a system might be unknowable (or at least unmeasurable), and so we must employ proxies or indicators to suggest the true state (Noss 1990, Stephens et al. 2015). Examples of such proxies in ecology and evolution are numerous. We can use various indices to infer ecosystem health, such as the IBI (the Index of Biotic Integrity; Beck and Hatch 2009). We can employ standardized surveys to infer species presence or population abundance, such as direct observation along transects, indirect observation via camera traps, or very indirect (but often more inclusive) quantification via environmental DNA. In evolutionary biology, we can use genetic markers or breeding experiments to assess whether two groups are separate species (Coyne and Orr 2004), and we can measure heritability or additive genetic variance to estimate evolutionary potential (Hansen et al. 2011). Furthermore, the knowledge that organismal traits are phylogenetically correlated (i.e., species relatedness predicts trait similarity) is often used to infer the expected trait value of species in a phylogeny that has not been measured for that trait (e.g., Swenson 2013).

As these examples highlight, prediction as diagnosis sometimes emerges from attempts to detect or quantify the item of interest itself, such as actually counting individuals of a focal species, whereas, in other cases, it is based on indirect indicators, such as remote sensing of spectral profiles to infer vegetation types (Xie et al. 2008). For prediction in the case of a yes or no diagnosis, such as whether a given species is present or absent, we need to be concerned with false positives (the entity is absent even though our proxy infers that it is present) and false negatives (the entity is present even though our proxy infers that it is absent; Fielding and Bell 1997). To minimize such errors, we need to employ specific sampling designs optimized to detect rare species or phenomena, such as "montane unicorns" (Hurlbert 1990). We also need to be aware of factors influencing detectability, as well as various other biases. For instance, a focal species might inherently avoid our capture methods (e.g., "trap shyness"; Wegge et al. 2004), or we might survey and test in only convenient locations (Koenig et al. 1992) or through easily measured but imperfect proxies (e.g., water conductivity to predict salinity).

Prediction as history

"Under the appropriate set of conditions, Alexander predicted, evolution ought to produce a eusocial vertebrate, even though

eusociality in the naked mole rat (or any other vertebrate) was unknown at the time."

—Braude (1997)

Studies of the past might seem even less likely to involve prediction than studies of the present. To the public, at least, it might seem that archaeologists and paleontologists simply dig things up and try to understand them, that historians read old texts and listen to oral histories to puzzle out a narrative, and that geneticists sequence genomes and blindly use statistical outcomes to infer evolutionary history. But even these seemingly exploratory efforts are often guided by predictions of one form or other. Some of these predictions relate to when and where to look for something (Verhagen and Whitley 2011). For instance, paleontologists use expectations from geological ages and rock types to guide where to search for fossils that might capture evolutionary transitions. Other predictions of the past appear in typical hypothesis-testing form. That is, researchers have examined existing information to generate predictions about the past that have then been tested in further studies (e.g., birds evolved from dinosaurs, Neanderthals mated with *Homo sapiens*, whale populations went through genetic bottlenecks because of hunting, and the Earth was struck by a massive meteor at the end-Cretaceous extinction).

Predictions about the past (and present) are especially susceptible to retesting when new data or methods become available. For instance, the development of radioactive isotopes for dating rocks or carbon has completely overturned many ideas about the past, as has the development of DNA sequencing, especially of ancient genomes. Similarly, new statistical tools often lead to the reevaluation of existing data sets. As one example, the use of climate envelope models to predict the past distributions of species is subject to new or improved predictions as models are refined (Nogués-Bravo 2009). As another example (the full saga is explained in Hunt et al. 2008), contemporary populations of stickleback fish show trait changes that clearly reflect rapid adaptation to changing environments. The expectation, then, would be that we could infer the action of natural selection for fossil time series of stickleback showing the same trait change while experiencing the same environmental change, yet standard statistical methods developed with that goal in mind could not reject the null hypothesis of randomness. Improved statistical methods resolved the issue by providing greater support for the action of natural selection than for the random alternative (Hunt et al. 2008).

Prediction as repeatability, consistency, or reproducibility

"Insanity is doing the same thing over and over again and expecting different results."

—Anonymous

If a scientist was to conduct the same experiment multiple times, how often would the same outcome emerge? The *multiple times* could, in this instance, mean repeated measurements of the same experimental unit (i.e., measurement error), different replicates of a given treatment within an experiment (e.g., replicate vials, enclosures, ponds, or plots), repetition of an entire experiment at two different times (e.g., different years) in the same laboratory, the use of different software packages to analyze the same data (Shafer et al. 2017), or repetition of an experiment across different laboratories. Such replicates—especially of the last type—are considered to be the basis for reproducibility in science, which

has been asserted as the reason for detailed methods sections and archived code and data (Parker et al. 2016). Although some authors (e.g., Cassey and Blackburn 2006) distinguish *repeatability* (the experiment can be conducted on the basis of the information given) from *reproducibility* (the same result is obtained when an experiment is repeated), I will use *repeatability* for the latter intent because doing so is common in the literature (see below).

Moving beyond formal experiments, repeatability can be evaluated for strictly observational studies of organisms and environments. If, for instance, we are interested in the among-population correlation between environmental variables and trait values (MacColl 2011), we can ask *How consistent is the correlation across space or time?* That is, we can estimate and compare environment–trait correlations for the same set of populations in different years (Gotanda et al. 2013); different sets of populations, such as on different continents (Gilchrist et al. 2004); or different taxonomic groups, such as multiple species in the same locations (Sanderson et al. 2021). Temporal repeatability in these cases can address questions about stability, stasis, equilibria, resistance, and resilience (Pimm 1984). Spatial and taxonomic repeatability can inform questions about evolutionary convergence or parallelism, such as when similar traits or genotypes evolve independently in similar environments (Lässig et al. 2017, Oke et al. 2017, Bolnick et al. 2018, Heckley et al. 2022). Furthermore, these different levels of repeatability can be merged to answer question about the spatial or taxonomic consistency of temporal trends in response to environmental change (Parmesan and Yohe 2003, Oke et al. 2020). Examples include assertions of "globally coherent signatures of climate change" (Parmesan and Yohe 2003) or various biogeographic "rules" or "laws," such as Cope's rule, Dollo's law, Bergmann's rule, and Rapoport's rule (Ashton 2001). Figure 1 shows each of these cases in idealized form and explains some of them.

Prediction as repeatability also figures heavily in studies of trait variation at the individual level. That is, if one measures the same trait in the same individuals at multiple times, one can consider how individually consistent are those trait values—especially in relation to other individuals (Nakagawa and Schielzeth 2010). Such individual repeatabilities are sometimes asserted to set an upper limit on the heritability of a trait (Dohm 2002); that is, if variation among individuals is not repeatable through time, then it can't be heritable. The consideration of individual repeatability has taken on particular importance in the study of behavior. Specifically, if a particular behavior is repeatable through time, then perhaps it reflects something about the *personality* or *behavioral type* of animals (Bell et al. 2009). Furthermore, if individual behavior is also repeatable across contexts (e.g., in the presence of mates versus predators), then perhaps it can be considered a *behavioral syndrome* (Sih et al. 2004).

A further manifestation of prediction as repeatabilty appears in the context of heterogeneity in meta-analysis. Meta-analyses typically seek to estimate a consistent effect shared across multiple experiments and systems. Therefore, heterogeneity among studies points to at least some nonrepeatability of an effect across experiments or contexts or taxa (Senior et al. 2016). Various moderators are then added to the analysis in an attempt to explain some of the heterogeneity and thereby improve consistency of effect size estimate within a given context or taxa.

Prediction in the sense of repeatability as just described (and in the sense of fate as described below) might seem more about predictability than about prediction per se. However, I included the concepts in the present article because they give a strong sense of implicit prediction. That is, a highly repeatable pattern can lead
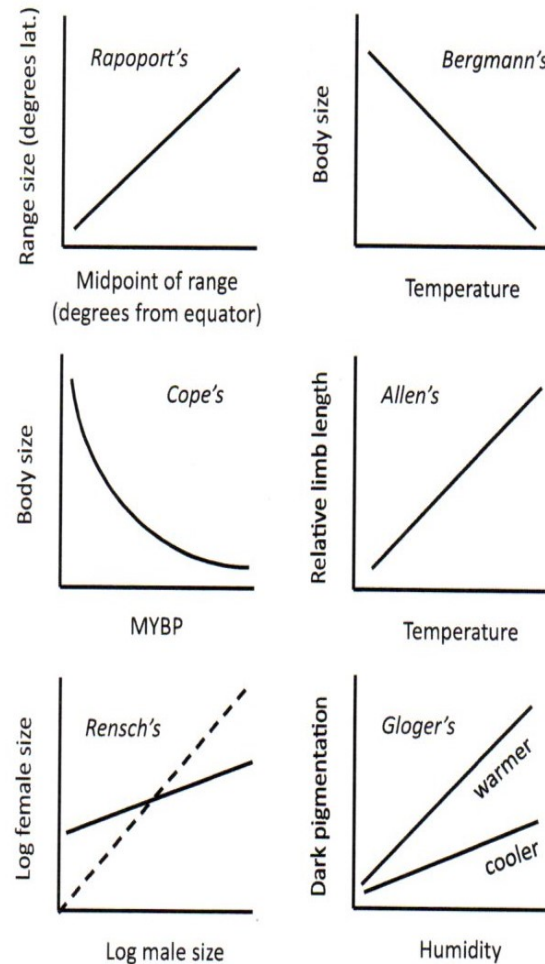

Figure 1. Six suggested "laws" or "rules" in ecology and evolution. Rapoport's rule states that organisms found at higher latitudes have greater latitudinal ranges. Bergmann's rule states that organisms found in colder temperatures have larger body sizes. MYBP is Million Years Before Present. Cope's rule states that body sizes increase in animal lineages through time. Allen's rule states that organisms found in colder temperatures have shorter appendages (limbs) relative to body size. Rensch's rule states that, within a lineage, sexual size dimorphism increases with increasing body size when males are the larger sex but decreases with increasing body size when females are the large sex (the dashed line is the 1:1 line where males and females are the same size). Gloger's rule states that dark pigmentation is more common in more humid and warmer environments. The statement of these "rules" does not imply that they are necessarily very predictive; in fact, many empirical exceptions have been reported for each.

to an implict expectation of an outcome without needing to state it as a formal prediction—as, I would argue, is common in many empirical papers. Of course, a repeatable pattern can certainly be used to generate an explicit prediction about the past, present, or future.

Prediction as fate, inevitability, or destiny

"I do not like the idea that I am not in control of my life."

—Neo in *The Matrix*

"Winter is coming."

—House Stark in *Game of Thrones*

Prediction as fate is the idea that a particular outcome is inevitable, or nearly so, under the expected range of conditions. Some scientists (and some reviewers of this article) argue that fate is not prediction in the usual sense. Take chaos as an example; the outcome can be inevitable, but it is also unpredictable in the usual a priori sense, because it is extremely sensitive to initial

conditions (May 1976, Rego-Costa et al. 2018). However, if we had perfect information, we would, in fact, have been able to make an accurate prediction of the outcome. Therefore, the relevance of including the concept in the present article emerges when *prediction* is considered broadly enough to include related concepts such as *predictability* and *repeatability*. In short, an inevitable outcome is extremely predictable, regardless of whether we make an explicit prediction of that outcome. Of course, as we will see later, such supposed fates are—in fact—rarely if ever inevitable. Therefore, my use of the term is intended to imply a sense of near inevitability; that is, all the indicators or signs point toward a particular outcome. As an example, the continuation of global warming in the years to come is inevitable.

In evolution, an important promoter of prediction as fate can be strong feedback loops (Crespi 2004). Negative feedback loops can generate evolutionary stability, such as in evolutionary cycles driven by coupled predator–prey dynamics (Blasius et al. 2020) or negative frequency dependence (Nosil et al. 2018). Positive feedback loops can generate runaway evolutionary outcomes between species, such as so-called arms races driven by predator–prey or host–parasite interactions (Dawkins and Krebs 1979) or even within species, such as runaway sexual selection (Pomiankowski and Iwasa 1998) or evolutionary environmental degradation (Hadfield et al. 2011). Other examples of suggested inevitable evolutionary fates include specialization as an evolutionary dead end (Futuyma and Moreno 1988) and taxon cycles on islands (Ricklefs and Bermingham 2002).

Ecology also can invoke a strong sense of prediction as fate, whether stemming from internal dynamics or external drivers. As in evolution, these fates are often envisioned as near-inevitable changes, such as ecological succession (Walker et al. 2010) or community assembly rules (Belyea and Lancaster 1999). Ecological fate also can arise from processes generating stability, such as in ecosystem resistance or resilience (Gunderson 2000, Folke et al. 2004). Of particular interest has been the potential inevitability of shifting between stabilizing and destabilizing outcomes, such as tipping points between alternative stable stages (Beisner et al. 2003, Scheffer et al. 2012). Short of the tipping point, systems predictably resist change despite shifts in external drivers, but past that tipping point, the system inevitably changes to a new state and can stay there even after reversal of changes in the original external drivers (i.e., hysteresis).

Prediction as fate is highly dependent on scale. In some cases, an outcome is inevitable on a short time scale: Yes, winter is indeed coming. In other cases, an outcome is inevitable only on a long timescale: All life on Earth will disappear when our sun explodes. Apart from such extreme inevitabilities, the invocation of prediction as fate does not mean that such fate is, in fact, truly inevitable—only that it has an extremely high probability barring some completely unforeseen turn of events. To return to the above example, a continuation of global warming in the years to come is inevitable—unless, of course, a massive volcanic eruption or meteor impact occurs and fills the atmosphere with ash or dust.

What are we trying to predict?

"This trend is of serious concern because it projects the global collapse of all taxa currently fished by the mid-21st century (based on the extrapolation of regression in figure 3a to 100% in the year 2048)."

—Worm and colleagues (2006)

"We predict, on the basis of mid-range climate-warming scenarios for 2050, that 15%–37% of species in our sample of regions and taxa will be 'committed to extinction'"

—Thomas and colleagues (2004)

Whatever form prediction takes—whether prophecy, repeatability, fate, or diagnosis—the specific form of the prediction can vary along several axes, which I outline as five contrasts: general versus specific, qualitative versus quantitative, absolute versus relative, point versus range, and continuous versus discontinuous. Predictions in ecology and evolution often mix and match these various categories in a diversity of ways (Burford Reiskind et al. 2021) and other ways of organizing the goals of prediction have been suggested (e.g., Maris et al. 2017).

General versus specific

General predictions extend beyond a specific context, such as a particular replicate or experiment or population or location. As such, general predictions transcend nuances and idiosyncrasies to instead reveal some underlying universal truth. This striving for "general" or "universal" laws of ecology and evolution is sometimes argued to stem from physics envy (Lawton 1999, Murray 2000, Penny 2005). General predictions also permeate extreme versions of Popperian falsifiability, in which the failure of a given experiment to confirm a prediction should cause the rejection of the general hypothesis that made that prediction (Chitty 1996). By contrast, specific predictions embrace the ubiquity of context dependence, in which an experimental or observational outcome will depend on the evolutionary context, the current environmental conditions, and the venue where the experiment or observations are conducted (Skelly and Kiesecker 2001, Elliott-Graves 2019). Figure 2 provides concrete examples of various manifestations of such context dependence in ecology and evolution. Specific predictions are not expected to uncover some universal truth; rather, they are intended to apply to a particular set of conditions. In reality, predictions always have some degree of generality and some degree of specificity, and I am simplifying the contrast for the purposes of illustration.

Even specific predictions are intended to transcend random stochastic noise to reveal some form of *predictable* context dependence, what Elliott-Graves (2019) called *causal heterogeneity* (Elliott-Graves 2019) and I call *predictable contingency*—with the intentional oxymoron focusing attention on the problem. For instance, two "replicates" of the same experimental treatment might give different outcomes because they differ in some other (unplanned or unmeasured) deterministic factor that influences the response variable. For instance, perhaps plants in two supposed replicates of the same nutrient level differ in growth because those replicates differ in temperature or light. One might imagine that random effects structure in a statistical model can circumvent this problem; however, this approach factors out some of the otherwise unexplained variation, giving an illusion of high predictability, as opposed to generating improved predictions by understanding the causes of variation. With such an understanding, however, studies can measure uncontrolled variables for inclusion as statistical covariates, or they can manipulate those other causal factors in crossed experimental designs. In plant ecology, for example, studies often examine a given factor of interest (e.g., diversity levels) under different light or nutrient levels (e.g., Johnson and Agrawal 2005). In evolutionary biology, studies often measure genetic effects in different environments, revealing context dependence in
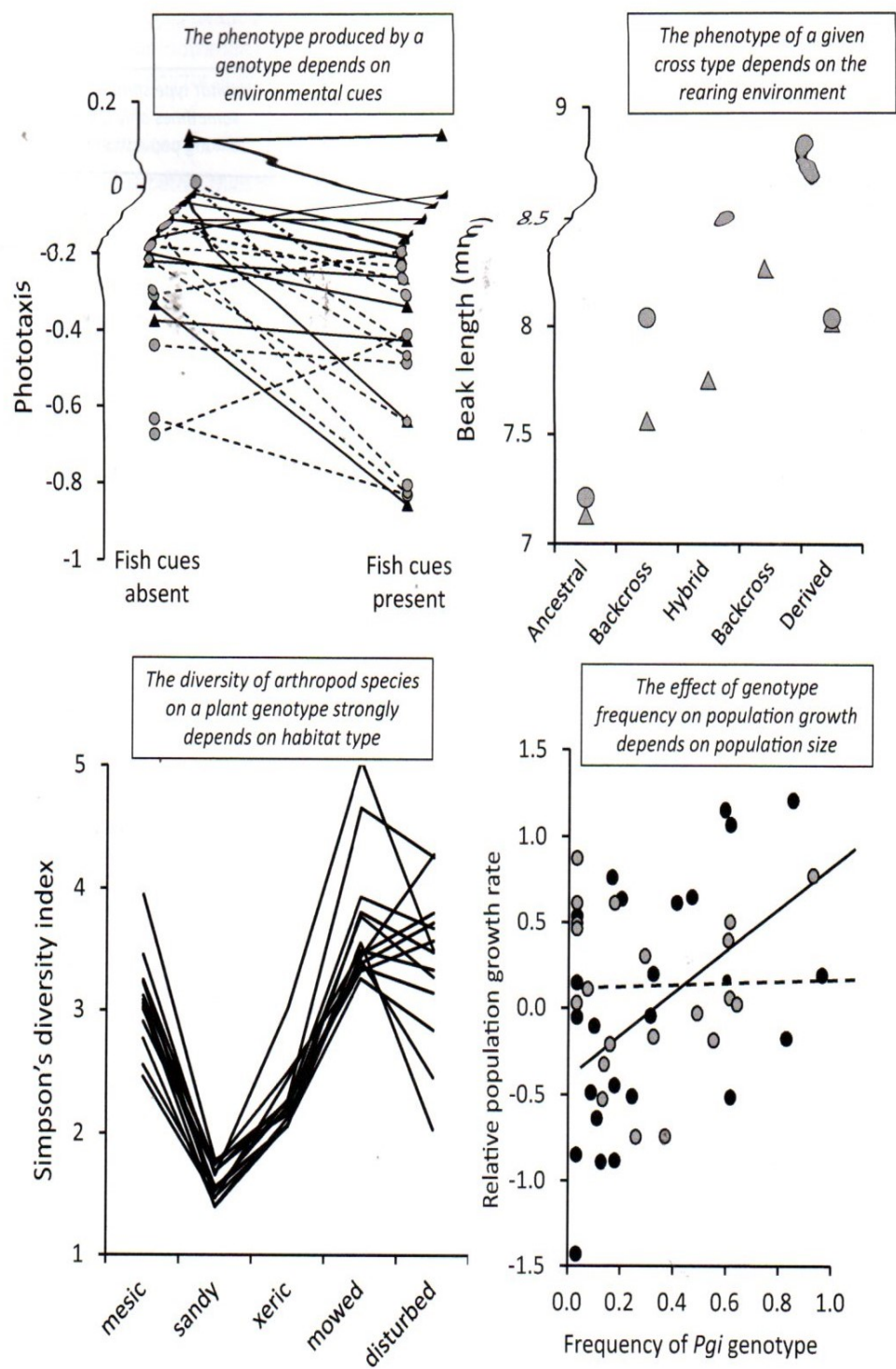
Figure 2. Examples of context dependence in ecology and evolution. The upper-left panel shows that the response to light (phototaxis score) of *Daphnia* clones (linked by lines) depends heavily on whether fish cues are present during testing (the two columns) and whether the clones came from time periods when fish predation was absent (the triangles and solid lines) or present (the circles and dashed lines; Stoks et al. 2016). The upper-right panel shows how the beak length of soapberry bugs (*Jadera heamatoloma*) in crosses between two ecotypes of the bugs (columns) is shaped by the plant on which they are reared (different symbols; Carroll 2007). The lower-left panel shows how the diversity of arthropods on primrose (*Oenothera biennis*) genotypes (lines) depends on the habitat in which the plants are grown (the columns; Johnson and Agrawal 2005). The lower-right panel shows how the effect of genotype frequency on population growth in Glanville fritillary butterfly (*Melita cinxia*) metapopulation patches depends on the size of habitat patches (different symbols and lines; Hanski and Saccheri 2006). Source: These panels are adapted, with modification, from Hendry (2017).

the form of interactions between genotype and environment (see the examples in figure 2).

Making and testing general predictions is therefore complicated by several realities. First, ecological and evolutionary patterns tend to be multicausal, such that an experiment failing to support a specific mechanism cannot be used to falsify that mechanism generally, because the role of that mechanism might depend on other measured or unmeasured factors (Hendry 2019). Such multicausality also makes it hard to apply Platt's (1964)

vision of strong inference to ecology and evolution (Hilborn and Stearns 1982, Quinn and Dunham 1983, Hilborn and Mangel 1997). Second, effects documented under one set of conditions can change when novel nonanalog conditions arise, as has been argued for the current Anthropocene (Fitzpatrick and Hargrove 2009). The hope then is to generate predictions that are positioned in the optimal place along the continuum of general to specific. Aiding this goal, the main effect of a factor of interest in a meta-analysis can be used to obtain a general prediction, whereas the

heterogeneity of estimates can be used to test more specific predictions by adding various appropriate moderators (Senior et al. 2016, Heckley et al. 2022).

Qualitative versus quantitative

Quantitative predictions are about a specific number or set of numbers: as examples, temperatures are increasing at a certain rate, two lineages split apart a certain number of years ago, the number of returning salmon is ten thousand, and so on—with or without various measures of uncertainty. Qualitative predictions, however, are about categories: as examples, temperatures are increasing more rapidly in one place than in another, two lineages split apart before two other lineages split, the number of salmon is greater this year than last, and so on. Many qualitative predictions invite a yes-or-no or a this-or-that answer: evolution is or is not parallel, an invasive species is or is not present, a behavior is or is not repeatable, flowering time is or is not changing, species A will outcompete species B, and so on. Such qualitative predictions tend to be reinforced by the frequentist approach to statistics, where we infer that something is happening if we can reject the null hypothesis at some critical value of the test statistic corresponding to (usually) $p < .05$. Qualitative predictions also lend themselves to model comparisons, where one model is accepted as "best," and that model is the only model considered thereafter (Stephens et al. 2007).

Quantitative predictions, which are more often framed as questions, tend to invite *how much*, *how many*, or *how strong* answers (Jennions and Møller 2002a, Peek et al. 2003, Hendry 2017). For instance, we might ask how much of the variation in phenotypes or genotypes or alleles is explained by parallel evolution (Oke et al. 2017, Bolnick et al. 2018, Heckley et al. 2022), what percentage of a fish community is made up of invasive species, how repeatable a behavior is (Bell et al. 2009), or how much has flowering time has advanced over the past century (Parmesan and Yohe 2003). Figure 3 shows quantitative representations of phenomena usually pitched as qualitative predictions: parallel evolution and the repeatability of behavior. Quantitative questions such as these are reinforced by estimates of effect size, with confidence intervals around those effect sizes (Halsey 2019), as well as relative weights in model comparisons (Stephens et al. 2007).

Most of the time in ecology and evolution, quantitative questions will be more useful than qualitative predictions, because it is more valuable to know how strong an effect is than simply if an effect is present. For example, if you were talking to a person who lived on the coast and you told them that sea level was going to rise, their response would surely be "Okay, by how much?" They probably wouldn't care much if the rise was going to be 0.01 millimeter per year, but they might get concerned if was 1 centimeter per year. If a manager was told fish population size was declining, they would surely ask, "By how much?" If a biologist was assessing evolutionary potential, it wouldn't be enough to know that fitness variation was heritable; it would matter just how heritable it was (Hendry et al. 2018).

Absolute versus relative

Absolute predictions tend to be specific values or ranges of values, although they also can be qualitative yes or no predictions (as above). The quotes at the top of this section provide some infamous examples. Relative predictions, by contrast, assert the order (but not the precise value) of differences: one value is bigger or smaller than another value, or one effect is more or less important than another effect (Dietze 2018). Relative predictions are especially useful when absolute predictions are tenuous. For ex-
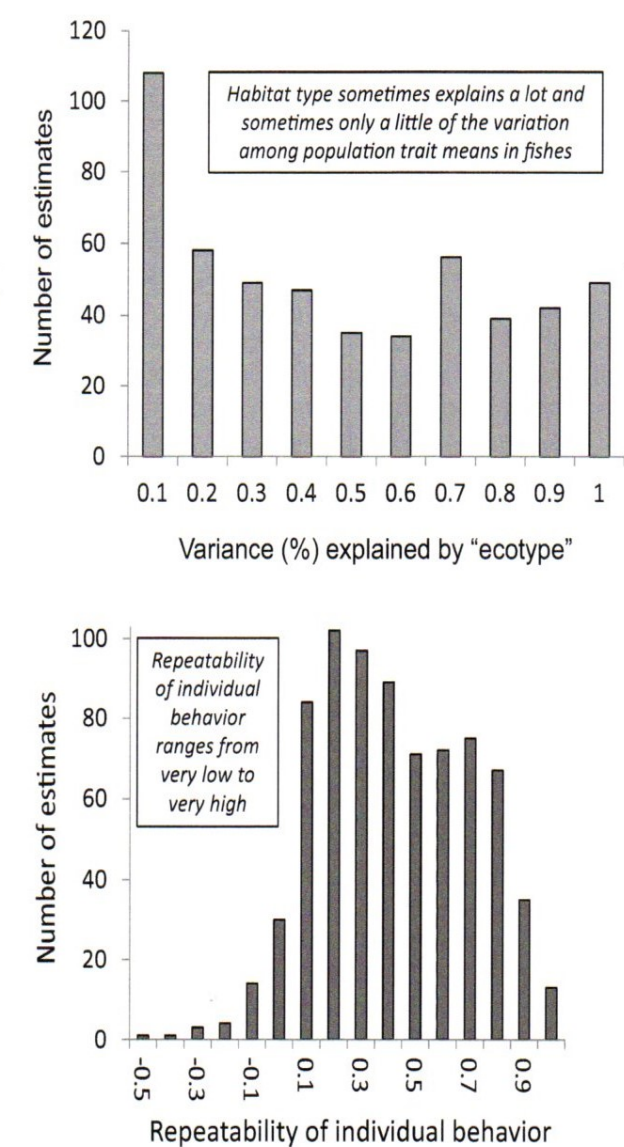

Figure 3. Illustrations of the high variation in repeatability. The top panel shows the number of estimates from studies of fishes where habitat (i.e., ecotype) explains a given percentage of the variation (r^2) among population means. The data indicate that parallel evolution varies from weak (low r^2) to strong (high r^2). The bottom panel shows the repeatability (r) of animal behavior. The estimates reveal that repeatability varies from very low to very high. Source: Both panels are reproduced, with modification, from Hendry (2017).

ample, we might not know with much precision or accuracy the extinction risk of two species, but we might know with high confidence that extinction risk is higher for one species than the other. Or we might not have enough information to generate a good absolute prediction of the rate of evolution of two trait types (e.g., life history versus morphology), but we might be very confident that one will evolve faster than the other. Other examples of relative predictions might be that species richness is higher in the tropics than elsewhere, body sizes increase with latitude (i.e., Bergmann's rule), and fitness at a given site will be higher for organisms from that site than for organisms from elsewhere (i.e., local adaptation).

Relative predictions encourage comparative assessment based the relative weights of different statistical models (Wagenmakers and Farrell 2004), sensitivity analysis of different model parameters (Cariboni et al. 2007), and elasticity analyses of population parameters (Benton and Grant 1999). For example, we might wish to determine the best environmental predictor of ecological or evolutionary change (van de Pol et al. 2016) or the best metric of evolutionary potential (Hansen et al. 2011) or the best model for

evolutionary change in phenotypes (Hunt 2007). Of course, such distinctions benefit from a clear indication via effect sizes of just how good the "best" proxy or predictor or model really is, because even the best might not be very good.

Although the above examples of relative prediction tend to qualitative (e.g., one model is better than another), relative predictions also can be quantitative, such as one model explaining twice as much variance as another. And even if the initial prediction is qualitative, its assessment often should be quantitative. That is, a researcher might predict that one ecological mechanism is stronger than another and then answer that prediction by saying just how much stronger.

Point versus range

Point predictions emphasis an average effect or discrete outcome, such as whether or not an invasive species is present or at what rate sea levels are rising or whether or not a species will go extinct. Range predictions emphasize a diversity of possible outcomes, expressed either as probabilities (e.g., 50% chance of extinction in the next 100 years) or as a confidence interval (e.g., flowering will advance 1–3 days per decade), or various alternative scenarios (e.g., different emissions scenarios for climate change).

Point predictions are more appealing in some respects. For instance, Popper (1959) argued that probabilistic predictions are unfalsifiable and therefore not scientific (Quinn and Dunham 1983), in which case, it is hard to convince some prognosticator that they were "wrong." After predicting that Donald Trump had only a 28.6% chance of winning the 2016 US Presidential election, Nate Silver (fivethirtyeight.com) still claimed success in that he gave Trump a higher chance of winning than most other statistical analysts did. Moreover, point predictions can be more useful than range predictions. For instance, knowing the actual best estimate of the heritability of a trait is more useful than simply saying a trait is heritable or not. However, range predictions can help to avoid unrealistic expectations (e.g., biodiversity targets; Hiers et al. 2016), can encourage preparation for uncertainty (Milner-Gulland and Shea 2017), and can more readily accommodate multicausality and context dependence.

Continuous versus discontinuous

Continuous predictions are typically made about variation along a continuous function (e.g., linear, logarithmic, exponential), whereas discontinuous predictions tend to be about transitions on or between functions (e.g., breakpoints, inflections, attractors). A common example of continuous prediction is the extrapolation of some past trend to some future year or of some geographical pattern past the bounds of the study area. Examples include predictions of the number of species committed to extinction by midcentury (Thomas et al. 2004) or the number of fish in the ocean by 2050 (Worm et al. 2006). By contrast, discontinuous prediction asks when a dramatic shift might take place, such as a tipping point that leads to a rapid and perhaps not easily reversible shift between alternative states (Scheffer et al. 2012). One form of discontinuous prediction asks where feedback loops switch from stabilizing to destabilizing, popularized in metaphors such as the straw that breaks the camel's back or how many rivets can be removed from a plane before it crashes (Ehrlich and Ehrlich 1981). Other examples of discontinuous prediction include the maximum sustainable rate of evolution (Kopp and Matuszewski 2014), critical transitions in speciation (Nosil et al. 2017), nutrient levels that shift lakes between alternative stable states (Scheffer et al. 1993), densities below which an extinction vortex kicks in (Fagan and Holmes 2006), and various planetary boundaries (Rockström

et al. 2009). These discontinuous tipping point predictions can be hard to test because the current state of the system might not reflect the inevitable fate that is coming, such as in the case of extinction debt (Tilman et al. 1994).

Where do predictions come from?

> "So you know what's going to happen?"
> "No, it was a feeling. But I can guess."
>
> —V for Vendetta

One might imagine that most predictions are the natural product of evidence, data, experience, and intuition—in various mixtures. That is, a researcher or prognosticator integrates previous information and interpretation to make an educated projection or guess as to the past, present, or future of a system—or what will happen in a new set of observations or experiments. Such predictions vary along a continuum from evidence based (existing data, models, theory) to gut feelings as implied by the quote above, and it is assumed we should place more confidence in the former. All predictions, however, surely contain an element of both data and intuition; for example, models have assumptions that require uncertain choices by the designer. On the other extreme, what might seem to be purely gut feeling predictions rarely occur in a vacuum of evidence but are instead based on varying degrees of experience. As such, gut feelings about a complex problem can sometimes incorporate a broader set of information and experience than can a formal mathematical or statistical model. We should obviously strive for evidence-based predictions, but the gut should not be ignored, and we shouldn't pretend it is completely absent from model-based predictions. The key, instead, is to make sure that readers and users are aware of where on the continuum from data to gut each part of a prediction falls.

Overall, philosophers often divide the main approaches to prediction into deductive versus inductive reasoning (Mentis 1988, Murray 2001, Dodds 2009). Again, actual predictions rarely fall cleanly into either one or the other of these two categories, as will be explained later. Deductive reasoning—often considered to be based on "first principles"—takes place when logical connections are used to infer what must be true, such as "all A have property B, C is a member of A, therefore C has property B." Despite what some authors (e.g., Peters 1991) have argued, deductive predictions are common in evolutionary biology. Adaptation by natural selection, for instance, can be considered a logical outcome of heritable variation that influences differential survival or reproduction—in essence, a syllogism (Endler 1986). However, as in all prediction, the outcome assumes that all else is equal. When all else is *not* equal (e.g., populations are small and subject to drift), adaptation might not follow even when the component parts are present. And of course, empirical estimates of heritability and selection could be wrong. As other examples, various biomechanical, physical, or frequency-dependent considerations can be used to define a realm of possible solutions (an envelope of possibility) or an optimum solution or evolutionary stable states, which are then later tested in the natural world. These predictions often emerge through mathematical exploration, most obviously through optimality models or game theoretic approaches (Maynard Smith 1982, Parker and Smith 1990), but also via other optimization algorithms such as evolutionary algorithms or machine learning (Simon 2013). Examples of predicted possible solutions that are then tested using real organisms include the diversity of shell shapes in an adaptive radiation of snails (Stone 1996) and pareto fronts generated by morphological constraints

evolutionary change in phenotypes (Hunt 2007). Of course, such distinctions benefit from a clear indication via effect sizes of just how good the "best" proxy or predictor or model really is, because even the best might not be very good.

Although the above examples of relative prediction tend to qualitative (e.g., one model is better than another), relative predictions also can be quantitative, such as one model explaining twice as much variance as another. And even if the initial prediction is qualitative, its assessment often should be quantitative. That is, a researcher might predict that one ecological mechanism is stronger than another and then answer that prediction by saying just how much stronger.

Point versus range

Point predictions emphasis an average effect or discrete outcome, such as whether or not an invasive species is present or at what rate sea levels are rising or whether or not a species will go extinct. Range predictions emphasize a diversity of possible outcomes, expressed either as probabilities (e.g., 50% chance of extinction in the next 100 years) or as a confidence interval (e.g., flowering will advance 1–3 days per decade), or various alternative scenarios (e.g., different emissions scenarios for climate change).

Point predictions are more appealing in some respects. For instance, Popper (1959) argued that probabilistic predictions are unfalsifiable and therefore not scientific (Quinn and Dunham 1983), in which case, it is hard to convince some prognosticator that they were "wrong." After predicting that Donald Trump had only a 28.6% chance of winning the 2016 US Presidential election, Nate Silver (fivethirtyeight.com) still claimed success in that he gave Trump a higher chance of winning than most other statistical analysts did. Moreover, point predictions can be more useful than range predictions. For instance, knowing the actual best estimate of the heritability of a trait is more useful than simply saying a trait is heritable or not. However, range predictions can help to avoid unrealistic expectations (e.g., biodiversity targets; Hiers et al. 2016), can encourage preparation for uncertainty (Milner-Gulland and Shea 2017), and can more readily accommodate multicausality and context dependence.

Continuous versus discontinuous

Continuous predictions are typically made about variation along a continuous function (e.g., linear, logarithmic, exponential), whereas discontinuous predictions tend to be about transitions on or between functions (e.g., breakpoints, inflections, attractors). A common example of continuous prediction is the extrapolation of some past trend to some future year or of some geographical pattern past the bounds of the study area. Examples include predictions of the number of species committed to extinction by midcentury (Thomas et al. 2004) or the number of fish in the ocean by 2050 (Worm et al. 2006). By contrast, discontinuous prediction asks when a dramatic shift might take place, such as a tipping point that leads to a rapid and perhaps not easily reversible shift between alternative states (Scheffer et al. 2012). One form of discontinuous prediction asks where feedback loops switch from stabilizing to destabilizing, popularized in metaphors such as the straw that breaks the camel's back or how many rivets can be removed from a plane before it crashes (Ehrlich and Ehrlich 1981). Other examples of discontinuous prediction include the maximum sustainable rate of evolution (Kopp and Matuszewski 2014), critical transitions in speciation (Nosil et al. 2017), nutrient levels that shift lakes between alternative stable states (Scheffer et al. 1993), densities below which an extinction vortex kicks in (Fagan and Holmes 2006), and various planetary boundaries (Rockström

et al. 2009). These discontinuous tipping point predictions can be hard to test because the current state of the system might not reflect the inevitable fate that is coming, such as in the case of extinction debt (Tilman et al. 1994).

Where do predictions come from?

"So you know what's going to happen?"
"No, it was a feeling. But I can guess."

—*V for Vendetta*

One might imagine that most predictions are the natural product of evidence, data, experience, and intuition—in various mixtures. That is, a researcher or prognosticator integrates previous information and interpretation to make an educated projection or guess as to the past, present, or future of a system—or what will happen in a new set of observations or experiments. Such predictions vary along a continuum from evidence based (existing data, models, theory) to gut feelings as implied by the quote above, and it is assumed we should place more confidence in the former. All predictions, however, surely contain an element of both data and intuition; for example, models have assumptions that require uncertain choices by the designer. On the other extreme, what might seem to be purely gut feeling predictions rarely occur in a vacuum of evidence but are instead based on varying degrees of experience. As such, gut feelings about a complex problem can sometimes incorporate a broader set of information and experience than can a formal mathematical or statistical model. We should obviously strive for evidence-based predictions, but the gut should not be ignored, and we shouldn't pretend it is completely absent from model-based predictions. The key, instead, is to make sure that readers and users are aware of where on the continuum from data to gut each part of a prediction falls.

Overall, philosophers often divide the main approaches to prediction into deductive versus inductive reasoning (Mentis 1988, Murray 2001, Dodds 2009). Again, actual predictions rarely fall cleanly into either one or the other of these two categories, as will be explained later. Deductive reasoning—often considered to be based on "first principles"—takes place when logical connections are used to infer what must be true, such as "all A have property B, C is a member of A, therefore C has property B." Despite what some authors (e.g., Peters 1991) have argued, deductive predictions are common in evolutionary biology. Adaptation by natural selection, for instance, can be considered a logical outcome of heritable variation that influences differential survival or reproduction—in essence, a syllogism (Endler 1986). However, as in all prediction, the outcome assumes that all else is equal. When all else is *not* equal (e.g., populations are small and subject to drift), adaptation might not follow even when the component parts are present. And of course, empirical estimates of heritability and selection could be wrong. As other examples, various biomechanical, physical, or frequency-dependent considerations can be used to define a realm of possible solutions (an envelope of possibility) or an optimum solution or evolutionary stable states, which are then later tested in the natural world. These predictions often emerge through mathematical exploration, most obviously through optimality models or game theoretic approaches (Maynard Smith 1982, Parker and Smith 1990), but also via other optimization algorithms such as evolutionary algorithms or machine learning (Simon 2013). Examples of predicted possible solutions that are then tested using real organisms include the diversity of shell shapes in an adaptive radiation of snails (Stone 1996) and pareto fronts generated by morphological constraints

(Shoval et al. 2012). Examples of predicted optima often arise in life history theory (e.g., age at maturity given mortality schedules), reproductive allocation (e.g., egg size given embryo mortality), and foraging theory (e.g., ideal free distributions during foraging). Deductive predictions are also found in ecology (Wilson 2003), where they underly—and can be used to test—formal theory such as island biogeography (MacArthur and Wilson 1967). Other potential examples include various aspects of community assembly and organization (e.g., niche packing, limiting similarity, food web assembly) and metabolic ecology (Brown 2004).

Inductive reasoning occurs when predictions arise from an accumulation of previous empirical observations. Perhaps the most extreme manifestation of this approach is the so-called empirical school of limnology, which criticized attempts to understand mechanism and argued for an exclusive focus on the best empirical predictors of a given phenomenon (Peters 1991, Rigler and Peters 1995). That is, if we want to know the color of water in a lake, perhaps we can simply measure total phosphorus. Similarly, if we want to know the metabolic rate of an endotherm, perhaps all we need to know is its body mass. Echoes of this type of approach can be seen in arguments for self-thinning (Westoby 1984), life history invariants (Charnov 1993), and the neutral theory of ecology (Hubbell 2001). Beyond such extremes, many other studies generate predictions about what might happen in a new study on the basis of previous related studies, sometimes formalized via Bayesian logic (Hilborn and Mangel 1997). In evolutionary biology, inductive predictions are especially clear in the study of parallel or convergent evolution: if five populations or species look similar when found in a similar environment, then so too should the next population or species we examine.

In reality, most predictions emerge from the joint consideration of both logic (deduction) and empirical precedent (induction). The theory of evolution by natural selection is a case in point. Deduction was clearly important (as was noted above), but so too were the numerous empirical observations that Darwin assembled from domesticated animals, the fossil record, and geographical distributions. Furthermore, logic is always conditioned by empirical experience; witness the famous prediction that, on the basis of the physics of flight, bumblebees shouldn't be able to fly (Magnan 1934). The empirical observation that bumblebees can, in fact, fly ("And yet it flies") lampooned the strict deductive approach and pointed to the need for empirical guidance.

In closing this section, it feels important to note that some purely empirical relationships are astounding predictive but also unhelpful, misleading, or tenuous. Examples include autocorrelation (e.g., offspring are always younger than their parents; Ellstrand 1983), extrapolation (e.g., the oceans will be empty of fish by 2048; Hilborn 2006), chance (p-hacking from data sets with lots of variables; Head et al. 2015), reverse causality (Räsänen and Hendry 2008), or various other lurking variables (Laubach et al. 2020). My point here is that predictions based on existing empirical relationships should really be backed by knowledge of causal associations.

How do we test predictions?

"If a man will stand up and assert, and repeat and reassert, that two and two do not make four, I know of nothing in the power of argument to stop him."

—Abraham Lincoln

Some predictions are not intended for testing in the usual scientific sense, such as predictions to help prepare for future events. In these cases, the prediction turns out to be correct or not (or to a given degree) after its period of usefulness, and the resulting test is valuable only in its ability to improve future predictions. Other types of predictions do require—or at least benefit from—testing, because they are hoped to reveal some underlying truth about the world or because they are intended for future application to situations where the outcome will be consequential.

Some predictions are easily tested from a conceptual standpoint, such as through the collection of data from new geographical locations or new experiments. Studies of parallel or convergent evolution are a case in point, where traits (e.g., color or life history) that previous work showed were associated with a particular environmental contrast (e.g., predation) are used to make—and then test—predictions about the same traits when other populations or species encounter the same contrast (for a review, see Heckley et al. 2022). Other predictions, however, are much harder to test, especially those for which new data cannot be collected and those made on temporal scales not amenable to timely assessment. In such cases, predictive models can be developed with a subset of the data and then tested (cross-validated) with the rest of the data (Roberts et al. 2017). For instance, part of a time series can be used to predict the rest of a time series, including hindcasting the past (using recent data to predict historical or archived data) or forecasting the present (using historical data to predict current conditions). In some cases, it is also possible to experimentally create past or future conditions in which empirical systems can be used to test a variety of predictions. A particularly powerful example is the resurrection of past genotypes for testing under present environmental conditions. Such resurrections have proven insightful for a variety of organisms with dormant stages, such as seeds for plants or resting stages for aquatic invertebrates or frozen samples of microbes (Orisini et al. 2013).

How do we test predictions about more distant futures? In one sense, such predictions will inevitably be tested because we simply have to wait for the future to arrive. That is, every prediction of how rising carbon dioxide levels will increase temperatures or raise sea levels (or cause range shifts or species extinctions or adaptive evolution) will inevitably be tested by the passage of time. Unfortunately, by the time we find out if our predictions were correct—and to what degree and in what sense—it will be too late to do anything about it. Therefore, we attempt to test predictions about the future in various indirect ways, such as the above-mentioned cross-validation (Roberts et al. 2017). Other approaches include the use of spatial patterns to test temporal predictions (substituting space for time; Pickett 1989), short-term experiments to test long-term outcomes (Wolkovich et al. 2012), and simulation models (Peck 2004). Although these convenient substitutions sometimes work well, experience teaches that spatial patterns do not always mirror temporal changes, that short-term experiments often do not scale up to long-term outcomes, and that the outcome of simulation models always depends on the assumptions that went into them.

How do we test whether a measurable proxy provides a good indicator of a (sometimes) unmeasurable reality? One example is the attempt to infer the absence of something. Has an endangered species gone extinct? Is an invasive species present? In such cases, the chosen survey method might simply have imperfect detectability, as is evidenced by the numerous instances where a species thought to be extinct has been rediscovered (Fisher and Blomberg 2010) or an invader thought to be absent has—in fact—been present (Trebitz et al. 2017). The same problem arises in
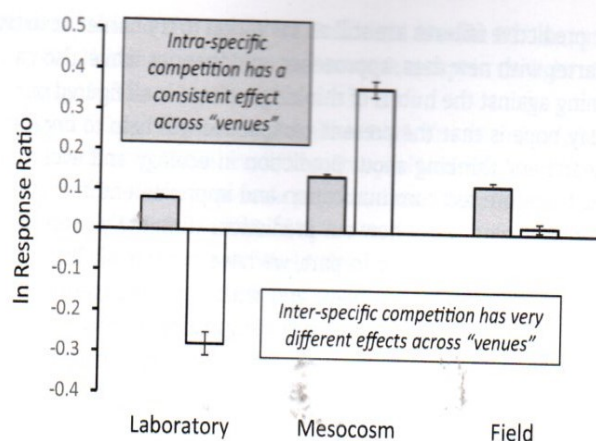
Figure 4. Illustration of the role of venue in the outcome of ecological experiments, based on tadpole growth rate experiments in the meta-analysis of Skelly and Kiesecker (2001). The bars show the average natural log response ratio when the density of intraspecific competitors was increased (the gray bars) or when an interspecific competitor was added (the open bars). The error bars represent the 95% confidence interval. The results were similar among experimental venues (laboratory, mesocosm, field) for intraspecific competition but not for interspecific competition. Source: The figure is based on data provided by David Skelly.

quantitative inference, such as estimating the abundance of a population or species. In most cases, the true number of individuals is unknowable, and so the best test for the predictability of a given proxy becomes its correlation with some other proxy. Alternatively, one can create experimental arenas with known numbers of a given species to see how close the proxy comes to the actual number present. A recent application of both methods comes in testing the predictive ability of environmental DNA (eDNA) for species presence or abundance; that is, estimates from eDNA are compared to estimates from traditional survey methods or from experimentally constituted known communities (Mahon et al. 2013).

To what extent should predictions be tested in the complex real world as opposed to controlled settings (Hendry 2019)? A classic manifestation of this debate was whether experimental tests in microcosms or mesocosms generate realistic predictions about what happens in actual lakes (Carpenter 1990). In some cases, outcomes of the different approaches correspond and are therefore particularly convincing. For example, a simple experiment adding phosphorus to only one side of a divided lake was much more effective—especially for public opinion and policy makers— at confirming the effects of phosphorous additions than were numerous previous experiments in artificial settings (Schindler 1977). In other cases, outcomes are very different between approaches. Stated another way, the venue for testing a prediction matters to the conclusion. As one example, experimental venue (lab, mesocosm, or natural ponds) has a dramatic effect on the outcome of experiments testing predictions about the effects of competition on tadpole growth (Skelly and Kiesecker 2001; for details, see figure 4). Similarly, the responses of plants to a given level of simulated warming under controlled conditions is very different from their response to the same level of actual climate warming in nature (Wolkovich et al. 2012). In evolutionary biology, estimates of additive genetic variance and heritability—and, therefore, evolutionary potential—depend critically on the environmental conditions under which the variance components are estimated. In a stunning example, Kellermann and colleagues (2009) reported zero heritability to increased desiccation in several *Drosophila* species, but van Heerwaarden and Sgró (2014) found

reasonable heritability for the same species under more realistic increases in desiccation risk.

The main point of the preceding paragraph is that context matters, such that the same experiment in different contexts can give dramatically different outcomes. As was noted earlier, my goal is not to prescribe a particular best approach to prediction and its testing but, rather, to describe and organize the various options. In closing, however, I will recommend a comparative approach—in which effect types and sizes are compared across various predictors and responses and contexts. Examples of studies comparing the same predictor across contexts include the effects of competition across experimental venues (Skelly and Kiesecker 2001), the effects of warming on plant phenology across experiment types (Wolkovich et al. 2012), and the effects of predation on evolution across various trait types (Heckley et al. 2022). Examples of studies comparing different predictors within and among contexts include different forms of competition (interspecific versus intraspecific) in Skelly and Kiesecker (2001) and different potential modifiers (lineage, predator types, time for adaptation) of responses by guppies to predation (Heckley et al. 2022). Comparative analyses such as these provide useful information on just how important context is for an outcome and also on the relative explanatory power of a given prediction.

How good are predictions?

"All models are wrong, but some are useful."

—Attributed to G. E. Box

Many examples could be provided of successful predictions in ecology and evolution. Temperatures continue to rise, as do sea levels. Flowering times and breeding times are getting earlier (Parmesan and Yohe 2003). The catches in test fisheries are correlated with the catches in actual fisheries (Hyun et al. 2005). Biocontrol agents deemed specific to an invader rarely switch to native organisms (McFadyen 1998). Independent lineages colonizing similar environments tend to evolve similar (parallel) phenotypes (Oke et al. 2019). Yet even in these obvious successes can be found plenty of ambiguities and partial failures. Climate warming predictions are so diverse that any trend is bound to match at least one of them. Sea level rise is only accurately predicted after accounting for increased water impoundment (Chao et al. 2008). The rate of advancement of flowering time under experimental warming is far slower than the actual rate of change in nature (Wolkovich et al. 2012), and some populations of animals are not showing such advancement (e.g., Lane et al. 2012). Some biocontrol agents do end up attacking native species (Louda et al. 2003). Parallel evolution is often only modest or even weak (Oke et al. 2017) and highly dependent on the level of investigation, such as fitness, phenotypes, gene networks, genes, gene expression, or alleles (Bolnick et al. 2018). Similarly, behavior is often not very repeatable, and many examples exist of evolutionary one-offs (de Queiroz 2002). Figure 3 shows empirical data that questions the classic expectations that evolution is highly parallel and that behaviors are highly repeatable.

And, of course, alongside such more or less accurate predictions can be found many abject failures, sometimes optimistically recast as surprises (Doak et al. 2008, Elliott-Graves 2019). Introductions of mysid shrimp into Flathead Lake, Montana, in the United States, had the opposite ecological effect to that expected on the basis of their previous introduction to Kootenay Lake, British Columbia, in Canada (Spencer et al. 1991).

Introductions of nonnative species often have surprising outcomes, including exceptionally high abundances and expansion to habitats not normally occupied in the native range (Urban et al. 2007). Many instances of directional selection on heritable variation do not, in fact, generate the expected adaptive responses (Merilä et al. 2001, Shaw 2019). Neutral genetic variation is a demonstrably poor proxy for adaptive genetic variation (Reed and Frankham 2001). Overall, the predictive ability of single causal factors in ecology and evolution is extremely low (Jennions and Møller 2002a), although multicausal predictive ability is more reasonable (Peek et al. 2003).

The goal of this section is not to argue that predictions are generally successful or not successful. Rather it has been to point out that most predictions tend to be partly successful and partly unsuccessful, and we need to assess this variation on a quantitative scale and light of our goals (Mouquet et al. 2015, Burford Reiskind et al. 2021). Do we care about the existence of a phenomenon (changing phenology) or the direction of a phenomenon (advancing spring phenology) or the magnitude or rate of a phenomenon (a certain number of days per decade) or the repeatability of a phenomenon (advances are seen in a certain percentage of species)? The answer will depend on the goals of the prognosticator. Overall, it is important to assert (when possible) the strength or success of a prediction on the basis of some sort of quantitative effect size (e.g., frequency of occurrence, variance explained, likelihood ratio, Cohen's d) that can be compared within and across contexts, response variables, locations, species, and studies. And such estimates should be (when possible) accompanied by a measure of uncertainty, and I don't mean a p-value but, rather, uncertainty in the effect size estimate itself, such as a confidence interval. Deviations between an a priori prediction and the actual outcome can be assessed with similar quantitative measures.

Conclusions

Prediction is not a monolithic and invariant concept, especially in ecology and evolution. Perhaps this variation should be viewed as a problem, and we should work toward a single unified way of defining and applying predictions. Or perhaps this variation is a strength. Perhaps, the basic idea of prediction is a general one, and that general idea can be leveraged into several specific approaches that are useful in different contexts. I espouse this latter viewpoint because it provides an overarching aspirational goal that reflects our ability to understand how the world works while allowing flexibility and diversity in how we achieve that goal.

If we understand the world, we should be able to say something about what will happen in a given set of circumstances, right? Well, sometimes we can, and such predictions or projections can be extremely satisfying intellectually or useful practically. Other times, however, our predictions are abjectly miserable, either because they lack precision or because they prove to be inaccurate, but why do some predictions fail in these ways? Maybe such failures point us toward other variables that will improve predictability, in which case we are making progress toward accurate and precise prediction. Or maybe such failures point to instances of the fundamental impossibility (given current data, approaches, and theory) of a robust understanding of the complexity of ecology and evolution. Perhaps that complexity would be predictable if we were able to measure all of the relevant variables with greater precision and accuracy, or perhaps it represents pure stochasticity that we must despair at ever explaining. We will never reach the ideal of Laplace's demon—an imaginary creature that knows everything about everything (Gompert et al. 2022),

but predictive failures are still an invitation to try harder or to try smarter with new data, approaches, and theories while also cautioning against the hubris of thinking we have it all figured out.

My hope is that the present perspective will help to organize researchers' thinking about prediction in ecology and evolution and, therefore, aid communication and improvement in that enterprise. Regardless of how our predictions ultimately succeed or fail, whether in whole or in part, we have a responsibility to try them—and to try in new ways and with improved clarity. With that in mind, I close with one major suggestion. I think the right approach to prediction is to start from the realization that most predictions will have elements of success and elements of failure, and we should therefore focus on more useful questions, such as *how* predictable is a given phenomenon, *what* factors contribute to variation in predictability, *how* much uncertainty exists around those predictions, and *how* can we reduce those uncertainties? Striving to answer these more realistic, more useful, and more practical quantitative questions should be a major goal of predictive science moving forward.

Acknowledgments

References cited

Allan C, Stankey GH. 2009. *Adaptive Environmental Management: A Practitioner's Guide.* Springer.

Ashton KG. 2001. Are ecological and evolutionary rules being dismissed prematurely? *Diversity and Distributions* 7: 289–295.

Beck MW, Hatch LK. 2009. A review of research on the development of lake indices of biotic integrity. *Environmental Reviews* 17: 21–44.

Beisner BE, Haydon DT, Cuddington K. 2003. Alternative stable states in ecology. *Frontiers in Ecology and Environment* 1: 376–382.

Bell AM, Hankison SJ, Laskowski KL. 2009. The repeatability of behaviour: A meta-analysis. *Animal Behaviour* 77: 771–783.

Belyea LR, Lancaster J. 1999. Assembly rules within a contingent ecology. *Oikos* 86: 402–416.

Benton TG, Grant A. 1999. Elasticity analysis as an important tool in evolutionary and population ecology. *Trends in Ecology and Evolution* 14: 467–471.

Blasius B, Rudolf L, Weithoff G, Gaedke U, Fussmann GF. 2020. Long-term cyclic persistence in an experimental predator–prey system. *Nature* 577: 226–230.

Bolnick DI, Barrett RDH, Oke KB, Rennison DJ, Stuart YE. 2018. (Non)parallel evolution. *Annual Reviews of Ecology, Evolution, and Systematics* 49: 303–330.

Braude S. 1997. The predictive power of evolutionary biology and the discovery of eusociality in the Naked mole-Rat. *Reports of the National Center for Science Education* 17: 12–15.

Brown JH. 2004. Toward a metabolic theory of ecology. *Ecology* 85: 1771–1789.