When Models Explain, When They Predict, and When They Do Neither

Claude.ai[1], ChatGPT[2], *Gemini*[3], *Alex*[4]
[1]Anthropic [2]OpenAI [3]Google [4]University of New South Wales

1. When Models Explain, When They Predict, and When They Do Neither

1.1 Introduction: Why clarity of purpose matters

The biggest source of confusion in environmental modelling is not model choice—it is unclear purpose.

Statistical models are now central to environmental science. We use them to relate climate variability to ecological change, to estimate risks such as coral bleaching or fishery collapse, and to project future conditions under climate change. Regressions, mixed models, generalised additive models, and machine-learning approaches appear routinely in the literature, often applied to similar problems.

Despite this widespread use, there is persistent confusion about what these models actually tell us. The same type of analysis may be interpreted very differently across studies. Strong, confident language is often used to describe results that rest on limited evidence. This confusion does not usually arise from poor statistical practice. It arises because many studies do not clearly state what the model is being used for. A regression relating temperature to coral bleaching, for example, might be used to:

describe patterns in historical observations,

test whether bleaching is associated with temperature anomalies,

predict bleaching risk at new locations or under future climate scenarios,

understand the mechanisms by which temperature causes bleaching.

This fourth goal—mechanistic understanding—is often the ultimate motivation, even when the methodology cannot directly support it.

These are different scientific goals. They require different standards of evidence. They support different kinds of conclusions. Yet the underlying model might be identical in all four cases.

Problems emerge when these goals are left unsaid. A model fitted to summarise past variability is discussed as though it makes reliable predictions. A model optimised for predictive accuracy is interpreted as evidence for ecological mechanisms. An exploratory analysis that screened dozens of variables is written up as though the selected predictors were specified in advance.

These practices are common, not because researchers are careless, but because the distinctions involved are rarely taught explicitly. Graduate training in statistics tends to focus on how to fit models and interpret output. It pays less attention to the prior question: what kind of answer is this model capable of providing?

The aim of this paper is to address that gap. We focus on models applied to observational data, which dominate environmental science, and we emphasise interpretation over technique. Our goal is not to argue for one class of model over another, nor to set statistical thresholds for good versus bad analyses. Instead, we aim to clarify:

what environmental models can legitimately be used for,

what they cannot tell us,

and how to align the language we use with the evidence we have.

The central argument is simple. Most problems with modelling are not technical, they are conceptual. They arise at the interpretation stage, not the analysis stage. And they can be largely avoided by asking one question before any model is fitted: "What am I trying to learn, and what kind of evidence would that require?" (Shmueli, 2010; Tredennick et al. 2021)

Table 1 Four goals, four evidence standards

Clarity about which goal is primary determines what can legitimately be concluded.

2. Core concepts

2.0.1 2.1 The same model, different questions

A single statistical model can serve very different scientific purposes.

Consider a regression relating sea surface temperature anomalies to coral bleaching severity across a network of reef sites. From this one model, we might ask:

Description: What patterns exist in the historical data? How much of the variation in bleaching can be accounted for by temperature?

Association: Is bleaching severity statistically related to temperature anomalies, after accounting for other factors?

Prediction: Can we use temperature forecasts to anticipate bleaching risk at sites or times not yet observed?

Mechanism: Does temperature cause bleaching, and if so, through what process?

These questions sound similar, but they are not the same. Description summarises what happened. Association tests whether a relationship is distinguishable from noise. Prediction asks whether the model works on new data. Mechanism asks why the relationship exists. Crucially, statistical models can address the first three directly; the fourth requires additional evidence beyond the model itself.

Table 1: Four goals, four evidence standards

Goal	What it requires	Common mistake
Describe	Model fits observed data	Claiming the patterns will hold elsewhere
Associate	Relationship survives statistical scrutiny	Claiming the association is causal
Predict	Demonstrated accuracy on new data	Inferring mechanism from predictive success
Explain mechanism	Theory + appropriate design + converging evidence	Relying on regression coefficients alone

Each question has different evidence requirements. Description requires only that the model captures patterns in the data at hand. Association requires attention to alternative explanations (could an unmeasured variable be driving both X and Y?), uncertainty in the estimated relationship, and whether the model is appropriately structured for the data. Prediction requires validation: demonstrating that the model predicts accurately when applied to observations not used to build it.

A further issue, common in environmental data, is pseudoreplication—treating non-independent observations as though they were independent. Two measurements from adjacent sites, or consecutive years at the same location, are not independent samples. If a model ignores this structure, standard errors will be too small, p-values too optimistic, and confidence in the results unjustified. Spatial and temporal autocorrelation are pervasive in environmental data; ignoring them is one of the most common sources of overconfidence in reported associations.

Problems arise when these purposes are conflated. A model built to describe historical patterns may be discussed as though it predicts the future. A model that demonstrates a statistical association may be interpreted as identifying a causal mechanism. These slippages are common, and they are the source of much confusion in the literature.

Before fitting any model, we need to ask: "What question am I trying to answer?"

The answer should determine not only how the model is built, but how the results are interpreted and described.

2.0.2 2.2 What most environmental models are actually doing

Debates about environmental modelling often focus on model complexity. Should we use a simple linear regression or a generalised additive model? A fixed-effects model or a mixed model? A classical statistical approach or machine learning? These choices matter, but they can obscure a more fundamental point: most regression-type models in environmental science are asking the same basic question.

That question is: How does the expected value of a response variable change as predictor variables change?

A multiple linear regression answers this using straight-line relationships. A generalised additive model allows those relationships to curve. A mixed model adds structure to account for non-independence—such as when the same site is measured repeatedly, or when observations from the same year share unmeasured conditions (Zuur et al., 2009). These models separate variation within groups from variation between groups. Machine-learning approaches like random forests work differently: rather than fitting one equation to all the data, they build many simple rules that apply to different subsets of observations, then combine them (Breiman, 2001).

These are real differences. They affect flexibility, performance, and the kinds of patterns a model can capture. But they do not change the fundamental nature of the question. A more complex model is not asking a deeper question—it is answering the same question with more flexible machinery.

This perspective is useful because it removes some of the mystique around advanced methods. If you understand what a multiple regression does—estimating how the average response changes with predictors—you already understand the core logic of most models used in environmental science. The rest is refinement.

One caveat: not all models fit this template. Classification models ask about probabilities of category membership. Clustering and ordination methods ask about hidden groupings or gradients in multivariate data. Time-series models with autoregressive components ask about temporal dependencies, not just predictor–response relationships. These are genuinely different questions. But for the regression-type models that dominate environmental applications, the underlying logic is shared.

2.0.3 2.3 Explaining data is not explaining nature

When environmental scientists say a model explains something, they typically mean that it accounts for variation in the observed data. A model might explain 40% of the variability in fish growth rates, or 60% of the variation in species richness across sites.

This usage is standard, but it can be misleading. Statistical explanation is not mechanistic explanation. A high R^2 is often treated as a marker of success, but it measures only how well the model fits the data at hand. A model can explain 90% of observed variation and still be wrong about mechanism—if the predictors are proxies for unmeasured processes, or if the relationships are specific to the sampled conditions. Conversely, a model with modest R^2 may correctly identify a real

but weak effect. Variance explained measures fit, not truth.

A model that explains variation tells us that certain variables are useful for summarising patterns in the data we have. It does not tell us why those patterns exist. It does not establish that the predictors are the true causes of the response. It does not rule out alternative explanations.

This distinction matters because observational data—the kind that dominates environmental science—are messy. Variables are correlated. Trends are shared across space and time. Important processes go unmeasured. A model can fit observed data well for reasons that are only indirectly related to the processes we care about.

Consider a model that relates bird abundance to vegetation cover. The model fits well and explains substantial variation. But vegetation cover is correlated with elevation, climate, land-use history, and proximity to water. The model cannot tell us whether birds respond to vegetation itself, or to something else that vegetation happens to track.

A similar issue arises in fish ecology. Suppose a model relates fish growth rates to sea surface temperature and explains substantial variation. But temperature covaries with primary productivity, prey availability, oxygen concentration, and stratification. The model cannot distinguish whether fish grow more slowly because warm water is physiologically stressful, or because warm conditions coincide with poor feeding opportunities. Both are plausible; the data do not discriminate.

This is not a flaw in the model. It is a fundamental feature of observational inference. Models summarise associations in data. Associations can arise for many reasons, only some of which reflect the causal relationships we are interested in. A helpful rule of thumb: Models explain data. Explaining nature requires additional evidence (Shmueli, 2010). That additional evidence might come from experiments, from mechanistic reasoning, from natural experiments that isolate specific processes, or from convergent findings across independent studies. The model alone does not provide it.

2.0.4 2.4 Prediction means performance on new data

Prediction is one of the most commonly invoked concepts in environmental modelling—and one of the most commonly misused.

In statistics, prediction has a specific meaning: it refers to a model's performance on data that were not used to fit the model. This might mean predicting new years, new locations, new species, or new environmental conditions. The key criterion is that the model has not seen the data before.

This definition has important consequences. A model can fit historical data extremely well and still predict poorly. If a model has too many parameters, or allows relationships to wiggle too freely, it may capture not only the underlying signal but also the noise specific to a particular dataset. When applied to new data—with different noise—it fails. This phenomenon, called overfitting, is one reason why goodness-of-fit is not the same as predictive skill.

Conversely, a model with modest explanatory power may predict reliably. If the underlying relationships are stable, and the model captures them without overfitting, it may generalise well even if it leaves much variance unexplained.

The practical implication is that claims about prediction require evidence of prediction. Statements such as X will impact Y in the future or our model predicts increased risk under climate change are claims about performance on new data. They require validation—testing the model on observations it was not trained on. Without this, such statements are speculation, regardless of how well the model fits the historical record.

This leads to an uncomfortable but important realisation. Many environmental models provide neither strong causal explanation nor reliable prediction. They may fit the data reasonably well, but they have not been tested out-of-sample. They may identify statistical associations, but these associations have not been shown to generalise.

Such models are not useless. They can summarise complex data in interpretable form, highlight plausible relationships, and generate hypotheses worth testing. But their limitations should be stated clearly. A model that describes historical patterns is not the same as a model that predicts the future—even if the fitted equation is identical. The difference lies in what has been demonstrated, not in the form of the model.

A further requirement is often overlooked: predictors must be available at the time of prediction. A model using degree heating weeks to predict bleaching is useless for forecasting if those heating weeks have not yet occurred.

2.0.5 2.5 Mechanistic understanding requires more than statistical models

The previous sections distinguished between describing patterns, testing associations, and predicting new observations. But there is a fourth goal that environmental scientists often care about most: mechanistic understanding—knowing why a system behaves as it does.

Statistical models do not directly provide this. The reason is fundamental: models relate measurements to measurements, not processes to processes.

The construct–measurement gap When we hypothesise that thermal stress causes coral bleaching, we are making a claim about theoretical constructs—abstract processes that we believe operate in nature. But we do not measure thermal stress directly. We measure sea surface temperature, or degree heating weeks, or some other proxy. Similarly, we do not measure bleaching as a physiological process; we measure colour scores, symbiont density, or mortality rates.

A statistical model relates these measurements to each other. It tells us, for instance, that higher degree heating weeks are associated with lower symbiont density. This is useful. But whether this association reflects the mechanism we hypothesise—thermal disruption of the coral–symbiont relationship—depends on questions the model cannot answer:

Does our temperature metric actually capture the thermal exposure experienced by corals?

Does symbiont density adequately represent bleaching severity?

Are there other processes, correlated with temperature, that could produce the same pattern?

These are theoretical questions, not statistical ones. A model can be statistically impeccable and still tell us little about mechanism if the measurements poorly represent the constructs, or if the constructs themselves are misconceived. Shmueli (2010) calls this the "construct–measurement gap" - explanatory power and predictive power are fundamentally different.

Variable selection for mechanistic understanding This has direct implications for how variables are chosen.

If the goal is prediction, variables can be selected based on their statistical usefulness. A variable that improves predictive accuracy earns its place, regardless of whether it has any mechanistic interpretation. Correlated proxies, composite indices, and atheoretical predictors are all acceptable if they work.

If the goal is mechanistic understanding, the logic is different. Variables should be selected before analysis, based on a conceptual model of the system. The question is not "which variables predict Y?" but "which processes do I hypothesise are operating, and how can I measure them?"

This means that mechanistic modelling often involves:

Including variables because theory demands them, even if they turn out to be statistically weak

Excluding variables that would confound interpretation, even if they improve fit

Choosing measurements that most directly represent the process of interest, not those that happen to be available or convenient

A model built this way may have lower explanatory power than one where variables are selected for fit. That is expected. The goal is not to maximise variance explained; it is to test whether a specific, pre-specified mechanism is consistent with the data.

The problem with data-driven variable selection for inference A common practice is to screen many candidate variables and retain those that show strong associations with the response. For prediction, this approach is legitimate—provided the model is subsequently validated on independent data. The validation step directly tests whether the selected variables generalise.

For inference, the same practice is far more problematic. Statistical inference assumes the model was specified before examining the data. When variables are selected because they appear significant, the resulting p-values and confidence intervals no longer mean what they claim to mean. They are too optimistic—often drastically so. A variable selected for showing a strong association will, by construction, appear to have a strong association. This circularity cannot be fixed by validation; it is a problem with the inferential framework itself.

Common workflows that produce unreliable inferences include:

Screening many predictors, then reporting p-values for the selected subset as though they were pre-specified

Using stepwise selection, then interpreting the surviving variables as ecologically important

Fitting the model, observing which variables are significant, then constructing a post-hoc hypothesis to explain the pattern

Reporting only the model that "worked" from among many attempted specifications

These practices are not inherently wrong—exploration is valuable. The problem arises when exploratory results are presented as confirmatory, or when the uncertainty introduced by the search process is ignored.

The implication is straightforward: if variables were selected based on the data, the analysis is exploratory, regardless of how it is presented. This is not a criticism—exploratory analysis is valuable. But it should be labelled as such, and its findings treated as hypotheses to be tested in future studies, not as established results.

The implication Mechanistic understanding cannot be read off model output. It requires:

A clear conceptual model of the processes thought to be operating

Measurements that credibly represent those processes

A statistical model structured to test the hypothesised relationships

Careful reasoning about alternative explanations

A regression coefficient is not a mechanism. It is a number that summarises an association between two sets of measurements. Whether that association reflects the mechanism we care about is a judgment that lies outside the model itself.

This is not a limitation of statistics. It is a recognition of what statistics does: it quantifies patterns in data. Understanding why those patterns exist requires theory, domain knowledge, and often evidence from multiple independent sources.

Box 1: The same model, three very different uses

A conceptual example from environmental science Consider a situation common in environmental research. We have repeated measurements of biological

performance—growth rates, survival, or reproductive success—from multiple sites over several years. We also have data on temperature extremes during the same period, along with biological covariates such as body size or age. We fit a statistical model relating performance to temperature while accounting for the biological structure in the data.

The model identifies relationships between temperature extremes and reduced biological performance. Several temperature metrics are statistically associated with the response.

From this single analysis, three very different scientific uses are possible. Each involves distinct questions, supports different conclusions, and requires different evidence.

A. Explaining patterns in historical data Question: Are variations in biological performance statistically associated with temperature extremes in the observed data?

What the model provides:

Some temperature metrics are associated with reduced performance after accounting for biological covariates

The model summarises patterns in the historical observations

A portion of the observed variability can be explained by the predictors

What can be concluded:

Temperature extremes are associated with variation in the response

The association persists after adjusting for other measured variables (though it might weaken or disappear if different variables were included)

What cannot be concluded:

That temperature extremes are the true cause of reduced performance

That alternative explanations have been ruled out

That the relationship will hold in other times or places

Appropriate language: "Temperature extremes were associated with reduced performance." "The model accounts for X% of the observed variation."

Inappropriate language: "Temperature drives performance decline." "Our results demonstrate the impact of heat stress."

Key message: Explaining patterns in data is not the same as explaining the processes that generated them.

B. Predicting new observations Question: Can the model accurately predict biological performance in new times, places, or conditions?

Additional requirement: The model must be evaluated on data not used for fitting—for example, withheld years, independent sites, or a separate dataset entirely.

What validation might show:

Good predictive skill within a tested domain, or

Poor generalisation despite strong historical fit

What can be concluded (if validation succeeds):

The model has demonstrated predictive skill under conditions similar to those tested

The identified relationships are robust, not artefacts of overfitting

What cannot be concluded without validation:

That the model will perform well under conditions more extreme than, or otherwise different from, those in the validation data

What cannot be concluded even with validation:

That the predictors are causal

That the model captures the true underlying mechanism

Appropriate language: "The model predicted performance with $R^2 = X$ on held-out data." "Temperature can be used to anticipate performance under similar conditions."

Inappropriate language: "Temperature controls performance." "These results confirm that warming will reduce performance."

Key message: Prediction is defined by performance on new data, not by confidence in the historical fit.

C. When the model does neither This situation is common—perhaps the most common.

A model may:

explain only modest variability in the response,

lack any formal test of out-of-sample performance,

rely on predictors that are correlated with each other and with unmeasured processes.

Such a model is not worthless. It may still be useful for:

summarising complex observations (e.g., reducing dozens of site-years to a few key relationships),

identifying variables that warrant further investigation,

generating hypotheses for future studies,

guiding the design of experiments or monitoring programmes.

But it cannot support:

strong claims about what drives the response,

confident statements about future impacts,

policy recommendations framed as predictions.

Appropriate language: "Temperature was associated with performance in the observed data, but this has not been tested on independent observations." "These patterns suggest hypotheses for further investigation."

Inappropriate language: Any claim about prediction, causation, or future impacts.

Key message: A model can be informative without being explanatory or predictive. The danger lies in claiming more than the evidence supports.

A note on language From analyses like the one described above, statements such as the following are common:

"Temperature extremes drive declines in biological performance." "Our results demonstrate the impact of marine heatwaves on ecosystem health." "These findings predict significant future losses under climate change."

What the evidence typically supports is more modest:

"Biological performance was lower during periods of temperature extremes in the observed data." "Temperature metrics were statistically associated with reduced performance after accounting for other variables." "If the observed relationships hold, future warming may be associated with reduced performance—but this has not been tested directly."

The first set of statements implies causation, mechanism, and future validity. The second set describes what was actually found. Both refer to the same analysis.

Key message: Scientific language should scale with scientific evidence. Strong claims require strong support—not strong models.

Box 2: When good fit fails to predict

Figure 2 illustrates two distinct ways that models can fail in prediction, despite fitting training data well.

Panels A–B: Overfitting

A complex model (polynomial) is fitted to training data and achieves $R^2 = 0.68$. When applied to new data from the same system, performance drops to $R^2 = 0.30$. The model captured noise specific to the training sample, not just the underlying signal. The wiggly fitted curve tracks random fluctuations that do not recur in new data.

Panel C: Confounding

A simple linear model is fitted to training data from Region A, achieving $R^2 = 0.70$. Higher temperatures are associated with slower growth. When applied to Region B, the model completely fails (R^2 0).

Why? In Region A, temperature happened to correlate with an unmeasured variable, e.g. food availability. When temperatures were high, upwelling was weak and food was scarce. The model attributed growth reduction to temperature, but food limitation was doing much of the work. In Region B, food was abundant regardless of temperature, so the relationship did not hold.

The lesson Both examples show the same outcome—predictive failure—but for different reasons. Overfitting occurs when models are too flexible for the available data. Confounding occurs when correlations in training data are local or contingent, not general.

Neither failure could be detected from the training fit alone. This is why validation on independent data is essential before claiming predictive skill.

3. Validation: what it does and does not tell us

A common response to concerns about overinterpretation is to point to validation. If a model has been tested on independent data, using a train/test split, cross-validation, or an external dataset, surely its conclusions are more secure?

This is partly true. Validation is genuinely valuable. But it is also a frequent source of misunderstanding, because it answers a different question than many researchers assume.

3.0.1 3.1 What validation actually tests

When we validate a model, we are asking: Does this relationship hold beyond the data used to fit the model?

This might involve withholding a random subset of observations, fitting to one time period and testing on another, training on some locations and predicting to others, or evaluating against a completely independent dataset.

If a model performs well under validation, we learn something important: the relationships it identifies are robust. They are not artefacts of overfitting. They are not chance correlations that happened to appear in one particular dataset.

This is a meaningful result. It tells us the model has captured something stable in the data-generating process. It increases confidence that the model is useful.

But useful for what?

Validation tells us a model can generalise—that it works on data it has not seen. It does not tell us why the model works. It does not tell us whether the predictors are causes, correlates, or proxies for something else entirely.

Key point: Validation tests robustness and generalisability, not causation.

3.0.2 3.2 Why validation does not establish causation

A model can predict accurately for reasons that have little to do with the mechanisms we care about.

Consider a model that uses sea surface temperature to predict fish recruitment. The model is validated on held-out years and performs well. What have we learned?

We have learned that temperature is a useful predictor of recruitment in the tested domain. We have not learned that temperature causes variation in recruitment.

Temperature might affect recruitment directly—through physiological effects on larvae, for instance. But it might also be correlated with other processes that drive recruitment: upwelling intensity, prey availability, predator abundance, or oceanographic transport. If any of these covary with temperature, a temperature-based model may predict well even if temperature itself is not the mechanism.

As long as these correlations persist, the model will generalise. If they break down—because of a regime shift, a change in ecosystem structure, or application to a different region—prediction may fail, even though temperature does affect recruitment in some contexts.

Validation cannot distinguish between two very different situations:

X causes Y.

X reliably co-occurs with the true cause of Y.

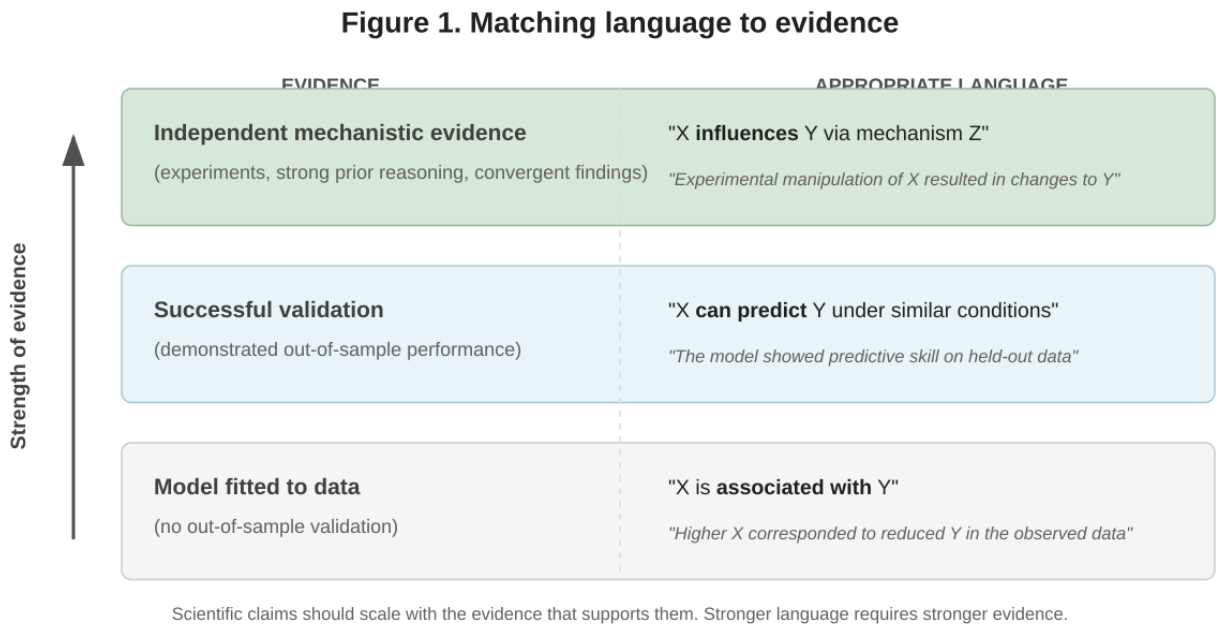

Figure 1: Good fit does not guarantee good prediction. (A) A complex model fitted to training data explains 68% of variance. (B) Applied to new data from the same system, performance degrades to $R^2 = 0.30$ because the model overfit noise in the training sample. (C) A simple model explains 70% of variance in Region A, but fails completely ($R^2 = 0.00$) in Region B because the temperature–growth relationship was confounded with unmeasured food availability. Data are simulated for illustration.

Both can produce strong predictive performance. Both can pass validation. Only one reflects the mechanism.

Key point: A model can predict well for the wrong reasons—and validation cannot tell the difference.

3.0.3 3.3 What failed validation does—and does not—imply

When a model fails to predict well on new data, the natural reaction is to conclude that the identified relationship is not real. This conclusion is often too strong.

Suppose we fit a model relating biological performance to temperature extremes using data from a particular region and time period. The model fits well. We then validate by predicting performance in a different period or region. The predictions are poor.

At least four interpretations are possible:

The original relationship was spurious. The association arose by chance or through overfitting and does not reflect a real process.

The relationship is real but context-dependent. Temperature affects performance, but only under certain conditions. The validation data came from a context where those conditions did not hold.

The model is missing key variables. Temperature is part of the story, but other unmeasured processes modulate its effect. When those processes differ between training and test data, prediction fails.

The system has changed. The relationship between temperature and performance is not stable over time due to adaptation, acclimation, or broader environmental change.

All four explanations are consistent with failed prediction. The data alone cannot tell us which is correct.

What failed validation tells us unambiguously is this: The model does not generalise reliably across the tested contexts. This means we should not use the model to make predictions in similar contexts. But it does not tell us that the underlying process is absent.

Key point: Failed prediction identifies limits to generality. It does not establish that a relationship is false.

3.0.4 3.4 Matching validation to claims

A useful discipline is to let the level of validation determine the strength of language.

Without validation: The model has been fitted but not tested out-of-sample. The appropriate language is associational: "X is associated with Y in the observed data."

With successful validation: Stronger language is warranted—but only about prediction, not causation: "X can be used to predict Y under conditions similar to those tested."

With independent mechanistic evidence: Causal language requires evidence beyond the model itself: "X influences Y via mechanism Z."

This framework does not weaken science. It makes science more precise. A study that demonstrates robust predictive skill has achieved something valuable—and should say so clearly. A study that identifies

a statistical association has also achieved something valuable—and should not overclaim.

Key point: Language should scale with evidence. Validation supports claims about usefulness, not about truth.

Scientific claims should scale with the evidence that supports them. Stronger language requires stronger evidence.

4. Inference and prediction are different tasks

The previous section focused on what validation can and cannot tell us. But there is a more fundamental issue: inference and prediction are different scientific activities, with different goals, different standards, and different measures of success.

Much confusion in environmental modelling arises not because these concepts are poorly understood in the abstract, but because they are quietly mixed within the same study. Models are built with one goal in mind and interpreted using the standards of the other.

4.0.1 4.1 Inference: evidence for associations in observed data

Inference-focused modelling asks questions such as: Is this variable associated with the response? Is the estimated effect distinguishable from zero? How large is the effect, and how uncertain is that estimate?

The emphasis is on parameter estimates and uncertainty. The focus is often on parameter estimation: not just whether temperature affects growth, but how much—what is the coefficient, and how precisely is it estimated? Prediction, by contrast, cares primarily about the response (Y); inference cares about the relationship itself (the).

Inference is typically retrospective. It is concerned with understanding patterns in observations we already have, often to evaluate competing hypotheses or quantify relationships suggested by theory.

Importantly, inference does not require out-of-sample validation to be meaningful. A well-designed experiment with proper controls supports inference without predicting new data. An observational study with careful attention to confounding and model specification can support inference about associations—even if it has never been validated externally.

Key point: Inference asks what the data tell us about relationships. It is evaluated by study design, model specification, and uncertainty quantification—not by out-of-sample prediction.

4.0.2 4.2 Prediction: usefulness for new data

Prediction-focused modelling asks different questions: How accurately can we predict new observations? Does the model generalise to new times, places, or conditions? What error should we expect when the model is used operationally?

The emphasis is on performance. We want to know whether the model works when applied beyond the data used to build it. The internal structure of the model—what the coefficients mean, which variables are important—is secondary to the question of whether predictions are accurate.

For predictive modelling, validation is not optional. It is the primary measure of success.

Key point: Prediction asks whether the model works on new data. It is evaluated by validation, not by the plausibility of the fitted relationships.

4.0.3 4.3 Why models optimised for inference often fail at prediction—and vice versa

A model can be excellent for inference and poor for prediction—or vice versa. This reflects different optimisation targets.

Inference-oriented models typically prioritise interpretability and simplicity, include only variables with clear theoretical justification, and aim for unbiased parameter estimates even at some cost to overall fit. These choices may limit predictive performance: a model that excludes correlated variables to avoid collinearity problems may predict less accurately than one that includes them.

Prediction-oriented models typically tolerate complexity if it improves accuracy, include correlated or proxy variables if they help prediction, and accept some bias in individual estimates if it reduces overall prediction error. These choices may produce models that are difficult to interpret: effects may be distributed across correlated predictors, and coefficients may lack clear mechanistic meaning.

Why biased estimates can improve prediction

This last point—accepting bias to improve prediction—may seem counterintuitive. If a coefficient is biased, how can the model predict better?

The answer lies in a fundamental tradeoff. Prediction error has two sources: bias (systematic error from model misspecification) and variance (instability in estimates due to limited data). A model that is unbiased but highly variable may, on average, predict worse than a model that is slightly biased but more stable.

Consider a concrete example. Suppose temperature and salinity both affect fish growth, but they are highly correlated in your dataset. An inference-oriented approach might exclude one variable to obtain clean, interpretable estimates of the other's effect. The resulting coefficient is unbiased—it estimates the true effect of the retained variable. But on new data, where the correlation between temperature and salinity may differ slightly, predictions may be poor because the model lacks information it could have used.

A prediction-oriented approach might retain both variables, even though their individual coefficients become unstable and hard to interpret (each coefficient is "contaminated" by the other). The model's estimate of temperature's effect is now biased—it no longer reflects temperature's true independent effect. But

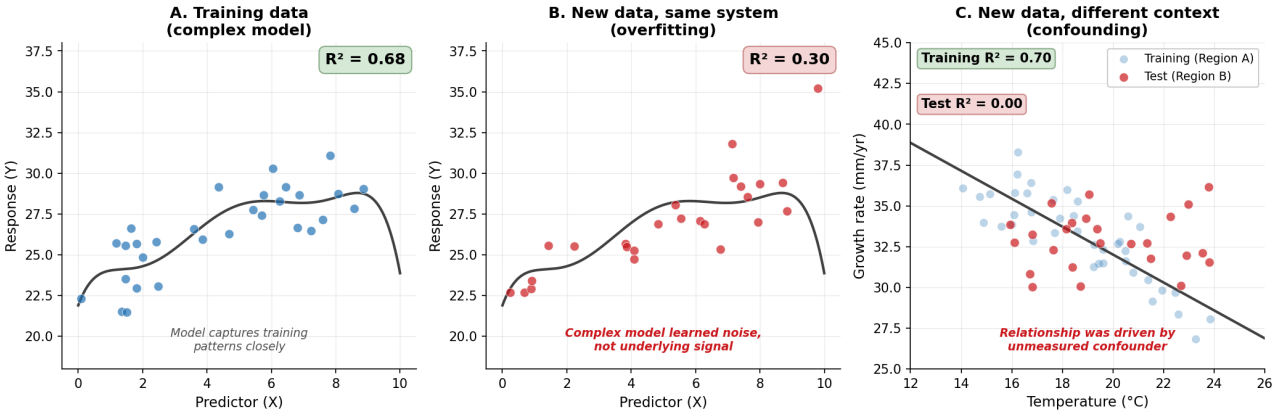

Figure 2: Systematic framework for matching language to evidence.

Table 2: Evidence and appropriate language

Evidence	Appropriate Language
Model fitted to data (no out-of-sample validation)	"X is associated with Y" / "Higher X corresponded to reduced Y in the observed data"
Successful validation (demonstrated out-of-sample performance)	"X can predict Y under similar conditions" / "The model showed predictive skill on held-out data"
Independent mechanistic evidence (experiments, strong prior reasoning, convergent findings)	"X influences Y via mechanism Z" / "Experimental manipulation of X resulted in changes to Y"

predictions of fish growth may be more accurate, because the model uses all available information about conditions that covary with growth.

This is why simpler models sometimes predict better than complex ones (they have lower variance), and why "wrong" models can outpredict "true" ones when data are limited (Breiman, 2001; Shmueli, 2010). The optimal balance between bias and variance depends on the goal. For inference, we want unbiased estimates of specific effects—bias is unacceptable. For prediction, we want accurate forecasts of the response—some bias is tolerable if it buys stability.

Neither outcome represents failure. It represents optimisation for different objectives.

Key point: Poor predictive performance does not invalidate inferential conclusions. Strong predictive performance does not validate causal interpretations.

4.0.4 4.4 The danger zone: mixing goals without saying so

The most common source of confusion in environmental modelling is not poor statistics. It is the implicit mixing of inferential and predictive goals.

Warning signs include: using predictive language (X will impact Y) without validation; drawing causal conclusions from variable importance scores in machine-learning models; reporting train/test performance but interpreting coefficients as effect sizes; treating predictive failure as evidence against a process; and presenting exploratory analysis as confirmatory.

None of these practices are fatal in isolation. The problem is the mismatch between what was done and what is claimed.

The solution is not to choose one goal and ignore the other. It is to be explicit about which goal is primary, and to ensure that interpretation matches methodology.

Key point: Most modelling problems arise at the interpretation stage. Clarity about purpose prevents over-interpretation.

5. Hypothesis-driven and data-driven modelling: a false dichotomy

Environmental models are often described as either hypothesis-driven or data-driven. These labels carry implicit judgments: hypothesis-driven work is seen as principled and rigorous; data-driven work as exploratory or atheoretical.

This framing is misleading. Most environmental modelling sits somewhere between these poles, and the distinction obscures more than it clarifies.

5.0.1 5.1 What these terms usually mean

Hypothesis-driven modelling refers to studies where predictors are selected based on prior reasoning, expected relationships are articulated before analysis, and the model is used to test whether observations are consistent with these expectations. The appeal is clear:

by specifying expectations in advance, the risk of finding spurious patterns is reduced. But hypothesis-driven does not mean that causal mechanisms are established or that confounding has been eliminated.

Data-driven modelling refers to studies where many candidate predictors are considered, variable selection is guided by statistical criteria, and models are evaluated primarily by fit or predictive accuracy. Data-driven approaches are sometimes criticised for letting the data speak for themselves, but all models embed assumptions. Data-driven methods are entirely appropriate when theory is weak, systems are complex, or the primary goal is prediction.

Key point: Hypothesis-driven does not mean rigorous. Data-driven does not mean unprincipled.

5.0.2 5.2 A continuum, not a binary

Most environmental studies fall between these extremes. A typical workflow might involve assembling predictors based on ecological reasoning (hypothesis-driven), exploring alternative variable transformations (data-driven), refining the model based on diagnostics (data-driven), and interpreting results in light of prior understanding (hypothesis-driven). This mixing is not poor practice—it is often unavoidable.

Key point: Most useful models combine hypothesis-driven and data-driven elements. This is normal and appropriate.

5.0.3 5.3 Where problems actually arise

Problems arise when exploratory analyses are presented as confirmatory, when variable selection is extensive but uncertainty is understated, when associations discovered post hoc are described using causal language, or when prediction-oriented workflows are interpreted inferentially.

A particular risk is the "garden of forking paths" (Gelman & Loken, 2014), the accumulation of small, seemingly reasonable decisions (which variables to transform, which outliers to exclude, which interactions to test) that collectively inflate the chance of finding a spurious result. Each decision may be defensible in isolation, but the combination produces a model that fits the data far better than it should. If a different analyst, making equally reasonable but different choices, would have reached different conclusions, the reported result is less robust than it appears.

These problems are about interpretation, not methodology. A data-driven study that clearly labels its findings as exploratory is doing good science. A hypothesis-driven study that overstates the strength of its prior reasoning is not.

Key point: The danger lies in over-interpretation, not in exploration itself.

6. Practical guidance

The distinctions developed in previous sections translate directly into how models are built, validated, and written about. This section offers concrete guidance for authors, reviewers, and readers of environmental modelling studies.

6.0.1 6.1 State the purpose of the model

Every modelling study should make explicit, early in the paper, what the model is being used for. This sounds obvious; in practice, it is rarely done.

A simple statement can prevent substantial confusion:

"The goal of this analysis is to describe patterns in the historical data and identify variables associated with the response."

"We aim to develop a predictive model and evaluate its performance on independent data."

"This study is exploratory; we screen candidate predictors to generate hypotheses for future investigation."

Such statements do not constrain the analysis. They constrain the claims—which is exactly the point.

Key recommendation: State whether the primary goal is description, association, prediction, or exploration. Do so before presenting results.

6.0.2 6.2 Match language to evidence

Much over-interpretation arises from a mismatch between evidence and language. A useful test: imagine you had collected different data from the same system—different years, different sites, different individuals. Would your conclusion likely still hold? If the answer is uncertain, the language should reflect that uncertainty.

For observational models without validation: Use associational language (X was associated with Y). Avoid implying prediction, causation, or generality.

For models with successful validation: Predictive language is appropriate—but only about prediction (X can predict Y under similar conditions). Validation does not license causal claims.

For claims about mechanisms: Causal language requires evidence beyond model fitting—experiments, strong mechanistic constraints, or convergent independent findings.

The framework in Figure 1 provides specific guidance. The key discipline is ensuring that the strength of language matches the strength of evidence.

Key recommendation: Strong claims require strong, aligned evidence. Let the analysis determine the language.

6.0.3 6.3 Use validation deliberately

Validation is powerful, but it is not always necessary. If the goal is prediction, validation is required. If the goal is inference, validation is useful but not required—a well-designed study with appropriate uncertainty quantification can support inferential conclusions without out-of-sample testing.

When reporting validation, be specific: What data were held out? What performance metric was used? What domain do results apply to?

Key recommendation: Use validation to support predictive claims, not causal ones. Report procedures in enough detail that readers can assess the domain of applicability.

6.0.4 6.4 Treat variable selection with humility

In environmental systems, predictors are correlated, data are limited, and the true model is unknown. A selected variable is useful for the purpose at hand, given the data available. It is not necessarily causal or mechanistically important.

Do not equate selected with causal. Acknowledge that different datasets might yield different selected variables. If multiple predictors perform similarly, say so. Consider sensitivity analyses. For detailed treatment of collinearity and variable selection in ecological contexts, see Dormann et al. (2013).

Key recommendation: Variable selection identifies what is useful, not what is true. Report selection honestly and interpret with humility.

6.0.5 6.5 Be explicit about limitations

Clear statements of limitation strengthen, rather than weaken, a study. They show that the authors understand what their analysis can and cannot support. Useful limitations to state:

"We do not test causation; the observed associations may reflect confounding."

"The model has not been validated outside the observed conditions."

"The selected predictors may act as proxies for unmeasured processes."

Key recommendation: State limitations as clearly as results. This is a sign of rigour, not weakness.

6.0.6 6.6 What good practice looks like

For a PhD thesis chapter, good practice often involves multiple models serving different purposes:

An exploratory model to identify candidate variables and generate hypotheses

A confirmatory model with pre-specified structure to test those hypotheses on independent data

A predictive model, if forecasting is a goal, validated on held-out observations and interpreted cautiously

This is not redundant—it reflects the different questions each model type can answer. The exploratory phase generates ideas; the confirmatory phase tests them; the predictive phase asks whether they are useful. Problems arise only when the boundaries blur: when exploratory results are written as confirmatory, or when predictive success is interpreted as mechanistic validation.

Explicitly labelling each component—"this analysis is exploratory," "this model tests the hypothesis that..."—protects against over-interpretation and signals methodological awareness to readers and reviewers.

6.0.7 6.7 A decision checklist

Before finalising a modelling study, work through these questions:

What question am I asking? Is the goal to describe patterns, test associations, predict new outcomes, or explore possibilities?

Does my modelling approach actually answer that question? If the goal is prediction, have I validated? If inference, have I considered confounding?

What evidence would be required to support stronger claims? What would I need to claim causation or predictive skill under novel conditions?

Does my language reflect the answers to questions 1–3? Am I using causal language without causal evidence?

If the answers are not aligned, the problem is conceptual, not statistical.

Key recommendation: Alignment between question, method, evidence, and language is the foundation of credible modelling.

7. Conclusion

Environmental models are indispensable. They help us organise complex observations, quantify relationships, identify patterns, and—when properly validated—anticipate outcomes. They are central to how environmental science generates and tests understanding.

But models are tools, not oracles. They answer the questions we pose, within the limits of the data we provide and the assumptions we embed. They do not automatically reveal mechanisms, establish causation, or guarantee predictions. These require additional evidence, careful reasoning, and honest acknowledgment of uncertainty.

Most problems in environmental modelling are not technical. They do not stem from choosing the wrong algorithm or fitting too few parameters. They stem from a mismatch between what the analysis can support and what the conclusions claim.

The remedy is not more sophisticated methods. It is clearer thinking—about purpose, about evidence, and about language.

A model built to describe patterns should not be discussed as though it predicts the future. A model validated on historical data should not be interpreted as establishing causation. An exploratory analysis should not be written up as though every finding was anticipated.

These are simple principles. They are also routinely violated.

The goal of this paper has been to make these distinctions explicit and to provide practical guidance for navigating them. The core message is simple:

Better environmental science begins with clearer questions—and with the discipline to let those questions determine what we claim.

Models are powerful when used with precision. They mislead when asked to do more than they can. The

difference lies not in the complexity of the method, but in the clarity of the thinking behind it.

Summary Box: Key messages

This box consolidates the central points developed throughout the paper.

On purpose and clarity

The biggest source of confusion in modelling is unclear purpose, not model choice.

The same model can describe patterns, test associations, predict outcomes, or explore mechanisms—these are different goals with different evidence requirements.

State the purpose of the model explicitly and early.

On explanation

Explaining variance is not explaining mechanisms.

Observational models are subject to confounding. Associations may be real without being causal.

On prediction

Prediction means performance on new data. Good historical fit does not imply predictive skill.

Claims about prediction require validation.

On validation

Validation tests robustness, not causation. A model can predict well for indirect reasons.

Failed validation shows limits to generality, not that a relationship is absent.

On inference and prediction

Inference and prediction are different tasks with different standards.

Mixing goals without acknowledgment is the most common source of over-interpretation.

On variable selection

Variable selection identifies useful predictors, not mechanisms.

Exploration is legitimate when labelled honestly; the danger lies in presenting exploratory findings as confirmatory.

On language

Language should scale with evidence:

Without validation: "X is associated with Y"

With validation: "X can predict Y under similar conditions"

With mechanistic evidence: "X influences Y via mechanism Z"

Strong claims require strong, aligned evidence.

On limitations

Stating limitations clearly is a sign of rigour, not weakness.

A final reminder

Most modelling problems arise at the interpretation stage. Aligning questions, methods, evidence, and language prevents most errors.

Glossary

This glossary defines technical terms used in this paper.

Coefficient A number estimated by a model that describes the relationship between a predictor and the response. In a linear regression, the coefficient indicates how much the expected response changes for a one-unit change in that predictor, holding other variables constant.

Collinearity The situation where two or more predictor variables are correlated with each other. When predictors are collinear, it becomes difficult to separate their individual effects on the response. Collinearity is common in environmental data, where variables such as temperature, rainfall, and season often covary.

Confounding A situation where an unmeasured variable influences both the predictor and the response, creating a spurious association between them. Confounding is a central challenge in observational studies and cannot be eliminated by statistical methods alone.

Cross-validation A method for assessing how well a model generalises to new data. The data are divided into subsets; the model is fitted on some subsets and tested on others. This provides an estimate of predictive performance that is less optimistic than simply measuring fit on the training data.

Effect size A quantitative measure of the magnitude of a relationship or difference. In regression, the coefficient is an effect size: it indicates how much the response changes per unit change in the predictor. Effect sizes are distinct from statistical significance; a relationship can be highly significant but trivially small, or non-significant but potentially important.

Exploratory analysis Analysis conducted to discover patterns, generate hypotheses, or identify candidate variables for further study. Findings from exploratory work should be interpreted cautiously and labelled as hypothesis-generating.

Confirmatory analysis Analysis designed to test hypotheses specified before examining the data. Confirmatory analysis provides stronger evidence than exploratory analysis because the risk of finding spurious patterns is reduced.

Generalised additive model (GAM) A type of regression model that allows relationships between predictors and the response to be curved rather than straight. Instead of estimating a single slope for each predictor, a GAM estimates a smooth function.

Goodness-of-fit A measure of how well a model describes the data used to fit it. Good fit on training data does not guarantee good performance on new data.

Mixed model (hierarchical model) A regression model that accounts for structure in the data, such as repeated measurements on the same individuals or observations grouped by site. Mixed models include both fixed effects (relationships of primary interest) and random effects (variation among groups).

Non-stationary A system or relationship is non-stationary if it changes over time. Non-stationarity is a challenge for prediction: a model fitted to one period may not apply to another if the underlying relationships have changed.

Out-of-sample Data that were not used to fit a model. Evaluating a model on out-of-sample data tests whether the model generalises. Out-of-sample performance is the appropriate measure of predictive

skill.

Overfitting The tendency of a model to capture noise as well as signal when fitted to a particular dataset. An overfitted model performs well on training data but poorly on new data. Validation is the primary safeguard against overfitting.

Proxy variable A variable that is not of direct interest but is correlated with something that is. In modelling, a predictor may act as a proxy for an unmeasured process—it predicts the response not because it is causally related, but because it tracks something that is.

Pseudoreplication Treating non-independent observations as though they were independent samples. Common forms include multiple measurements from the same site, consecutive time points, or spatially clustered observations. Pseudoreplication leads to underestimated standard errors and overconfident conclusions.

R^2 (variance explained) A measure of how much of the variation in the response variable is accounted for by the model. R^2 describes fit, not causation. A high R^2 does not mean the predictors are causes.

Train/test split A validation approach in which data are divided into two parts: a training set used to fit the model and a test set used to evaluate performance. Performance on the test set provides a less biased estimate of how well the model will generalise.

Uncertainty A measure of how precisely a quantity is estimated. In modelling, uncertainty is often expressed as confidence intervals or standard errors. Acknowledging uncertainty is essential for honest interpretation.

Validation The process of evaluating a model's performance on data not used for fitting. Validation tests whether identified relationships generalise beyond the training data. It is essential for predictive claims but does not establish causation.

References

Anderson, D. R. (2008). Model Based Inference in the Life Sciences: A Primer on Evidence. Springer.

Breiman, L. (2001). Statistical modeling: The two cultures. Statistical Science, 16(3), 199–231. https://doi.org/10.1214/ss/1009213726

Burnham, K. P., & Anderson, D. R. (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (2nd ed.). Springer.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., ... & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Dormann, C. F., Calabrese, J. M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., ... & Hartig, F. (2018). Model averaging in ecology: A review of Bayesian, information-theoretic, and tactical approaches for predictive inference. Ecological Monographs, 88(4), 485–504. https://doi.org/10.1002/ecm.1309

Gelman, A., & Loken, E. (2014). The statistical crisis in science. American Scientist, 102(6), 460–465.

Shmueli, G. (2010). To explain or to predict? Statistical Science, 25(3), 289–310. https://doi.org/10.1214/10-STS330

Tredennick, A. T., Hooker, G., Ellner, S. P., & Adler, P. B. (2021). A practical guide to selecting models for exploration, inference, and prediction in ecology. Ecology, 102(6), e03336. https://doi.org/10.1002/ecy.3336

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). Mixed Effects Models and Extensions in Ecology with R. Springer.