Received: 21 December 2021 | Accepted: 27 July 2022

DOI: 10.1111/2041-210X.13971

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

REVIEW

Pielou Review

Random errors are neither: On the interpretation of correlated data

Anthony R. Ives

Department of Integrative Biology, University of Wisconsin-Madison, Madison, Wisconsin, USA

Correspondence
Anthony R. Ives
Email: arives@wisc.edu

Funding information
National Aeronautics and Space Administration, Grant/Award Number: 80NSSC20K0282; National Science Foundation, Grant/Award Number: DEB-2134446

Handling Editor: Dr. Aaron Ellison

Abstract

1. Many statistical models currently used in ecology and evolution account for covariances among random errors. Here, I address five points: (i) correlated random errors unite many types of statistical models, including spatial, phylogenetic and time-series models; (ii) random errors are neither unpredictable nor mistakes; (iii) diagnostics for correlated random errors are not useful, but simulations are; (iv) model predictions can be made with random errors; and (v) can random errors be causal?
2. These five points are illustrated by applying statistical models to analyse simulated spatial, phylogenetic and time-series data. These three simulation studies are paired with three types of predictions that can be made using information from covariances among random errors: predictions for goodness-of-fit, interpolation, and forecasting.
3. In the simulation studies, models incorporating covariances among random errors improve inference about the relationship between dependent and independent variables. They also imply the existence of unmeasured variables that generate the covariances among random errors. Understanding the covariances among random errors gives information about possible processes underlying the data.
4. Random errors are caused by something. Therefore, to extract full information from data, covariances among random errors should not just be included in statistical models; they should also be studied in their own right. Data are hard won, and appropriate statistical analyses can make the most of them.

KEYWORDS
causal inference, correlated data, correlated random errors, phylogenetic correlation, prediction, spatial autocorrelation, statistical model, temporal autocorrelation

1 | INTRODUCTION

E. C. Pielou was one of the founders of modern ecology, establishing the expectation of statistical rigour when exploring data for patterns predicted by ecological theory (Simberloff et al., 2017). I bought a copy of her 'Ecological Diversity' (Pielou, 1975) when I was an undergraduate, and from the marginalia I know I read the book carefully at least twice before finishing my PhD. Pielou's mathematical background is clear in her work (Pielou, 1977); from mathematical first principles, she derived metrics that encapsulate distinguishing features of processes

that might underlie patterns in a dataset. A prime example is her famous 'Pielou diversity index' (Pielou, 1966). Building on Pielou's initial approach, the combination of deriving metrics and testing them using nonparametric means (such as permutation tests) is a pervasive tool used throughout ecology and evolution (Gotelli & Ellison, 2004).

Although this is an E. C. Pielou Review, I am going to focus on the other broad approach to statistical inference in ecology and evolution: statistical models. The feature that distinguishes statistical models from metrics is the inclusion of probability distributions; while a metric might try to summarize the data with a single number that can be used to compare among datasets, a statistical model describes both deterministic and probabilistic properties of the data that depend on how the data were generated. In other words, the statistical distributions of metrics are often unknown, whereas those for models are explicitly stated. For example, simple linear regression models include a fixed component that describes the relationship between the expected values of the dependent variable and the values of the independent variables, and random errors that capture the variance, skew and higher statistical moments of the process generating the dependent variable. By including the probabilistic properties of the data, statistical models can perform useful functions beyond those possible with metrics, such as making predictions.

When fitting statistical models, the focus is generally on the fixed component. The random errors are nuisances, rather than things of interest in their own right. Over the last several decades, ecology and evolution (along with all the social and physical sciences) have seen an explosion of statistical models to account for complex random errors that fail the iid (independent and identically distributed) test; these models explicitly account for covariances among the random errors. Perhaps the most familiar are mixed-effects models whose random effects create correlated random errors (Gelman & Hill, 2007). A good reason to use models that allow for correlated random errors is that they often guard against type I errors, that is, when a relationship between dependent and independent variables is mis-identifying as being statistically significant. As a corollary, not using models with correlated random errors can lead to rejected manuscripts. Although this might cast a negative pall on random errors that are not iid, this is a shame, because covariances among random errors reveal information about the data and can aid in making predictions.

I have centred this review around five points about random errors. These points are all based on ideas that are common in statistics and the more statistical ecological and evolutionary literature. By bringing these ideas together, I hope to champion random errors as much more than just nuisances.

(i) Correlated random errors unite many types of statistical models including spatial, phylogenetic and time-series models.

To give an example model with correlated random errors, consider the generalized least-squares (GLS) regression model

$$\begin{aligned} \mathbf{Y} &= b_0\mathbf{J} + b\mathbf{X} + \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon} &\sim N\left(\mathbf{0}, \sigma^2\mathbf{V}[\boldsymbol{\theta}]\right). \end{aligned} \tag{1}$$

For notational convenience, the n observed values of the dependent variable are given by a vertical vector $\mathbf{Y} = (Y_1, Y_2, \ldots Y_n)'$, and the independent variable is similarly given by $\mathbf{X} = (X_1, X_2, \ldots X_n)'$. The fixed component of the model contains the intercept b_0 (which is the same for every observation) multiplied by $\mathbf{J}$, the $n\times 1$ vector of 1's, and the linear relationship between Y and X as specified by the coefficient b. Although for simplicity I have only included a single independent variable, others could be added, including interactions between them. The random error term $\boldsymbol{\varepsilon}$ is given by a multivariate Gaussian (normal) distribution that has mean zero and $n\times n$ covariance matrix $\sigma^2\mathbf{V}(\boldsymbol{\theta})$ that might depend on parameters $\boldsymbol{\theta}$. If this were an ordinary least-squares (OLS) model, the matrix $\mathbf{V}(\boldsymbol{\theta})$ would be the diagonal matrix of 1's. This is the same as saying that the random errors are iid. If the diagonal elements are different, then variances differ among random errors, giving the case of heteroscedasticity. Nonzero off-diagonal elements of $\mathbf{V}(\boldsymbol{\theta})$ make the random errors correlated.

Although the term $\boldsymbol{\varepsilon}$ in Equation (1) appears as a vector of n values, it is specified in the model in terms of a multivariate Gaussian random variable having mean zero and covariance matrix $\sigma^2\mathbf{V}(\boldsymbol{\theta})$; the n values of $\boldsymbol{\varepsilon}$ are realizations of the random variable. The covariance matrix gives information about a specific value of ε_i conditional on the other values of $\boldsymbol{\varepsilon}$. Different structures given to the covariance matrix makes it possible to apply the linear model to many types of data. To make this point, I will use the model to analyse three example types of data: spatial, phylogenetic and time-series data. For spatial data, $\sigma^2\mathbf{V}(\boldsymbol{\theta})$ will contain relatively higher covariances for locations that are closer together when whatever causes the random errors is more likely to affect close locations in similar ways (Figure 1a). Similarly, random errors for phylogenetic data are likely to reflect phylogenetic relatedness, with phylogenetically related species having correlated random errors (Figure 1b). Finally, in time series, points that are closer in time are likely to have correlated residuals. Correlations among random errors can be negative as well as positive. In the specific simulations of time series I use later, the dynamics of the random errors are cyclic, and in this case the covariances in $\sigma^2\mathbf{V}(\boldsymbol{\theta})$ become negative for points separated by half a cycle period (Figure 1c). Recognizing that very different types of data differ only in their correlation structure unifies the statistical methods that can be used to analyse them (e.g. Hansen & Martins, 1996; Pinheiro & Bates, 2000). This unification is not only conceptual but also practical, because the same methods can be used to fit the data. Nonetheless, for mathematical reasons models used for some types of data employ distinct fitting methods that take advantage of the specific way in which $\mathbf{V}(\boldsymbol{\theta})$ is constructed to reduce computational burden (e.g. Bates et al., 2015; Hadfield, 2015; Harvey, 1989).

(ii) Random errors are neither unpredictable nor mistakes.

Despite their name, 'random errors' are not random in the sense of lacking any pattern, and they are not errors in the sense of making unintended and avoidable mistakes. Instead, in statistics they have a specific, technical meaning as random variables

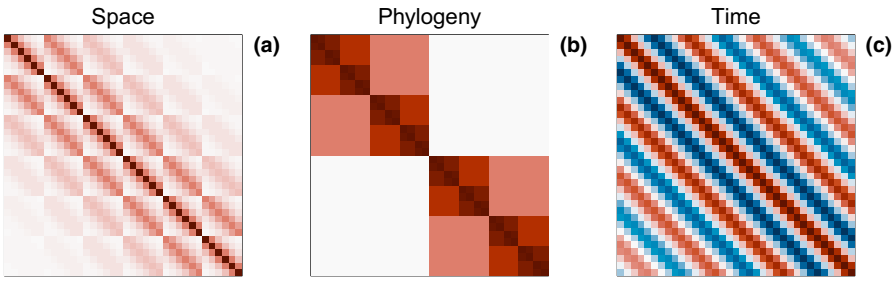

FIGURE 1 Images of covariance matrices $\mathbf{V}(\boldsymbol{\theta})$ for (a) spatial (Figure 3), (b) phylogenetic (Figure 4), and (c) temporal statistical models (Figure 6). Covariances are depicted on a blue–red scale, with dark blue, white, and dark red corresponding to −1, 0 and 1. For (a), a 6x6 map is depicted, with pixels ordered by rows (e.g. rows 1–6 correspond to the first row in the 6x6 map, rows 7–12 correspond to the second row in the map, etc.). The covariance between pixels *i* and *j* is $\exp(-d_{ij}/\rho)$. For (b), the phylogeny giving the covariances is the same as that in Figure 4b. Rows in (c) give serial autocorrelations (Equation 6); for example, the top row gives the covariance between the value of $\varepsilon(t)$ at year 1 and later years corresponding to each column. The time-series simulation model (Equation 4) gives cyclic dynamics, so the covariances cycle between positive and negative values with increasing time periods between random errors

generated from specified probability distributions. For example, for Equation (1) the distribution of the random errors is explicitly stated as multivariate Gaussian. In normal life, outside the realm of quantum physics, truly random events do not occur; even the flip of a coin and random number generators are deterministic processes. In ecological and evolutionary data, random errors are caused by something—variables—many of which will be similar to the independent variables in a model and distinguished only because they are not measured. Because these variables are unmeasured, they can only be captured by their effects in generating covariances in the random errors. Nonetheless, some of these unmeasured variables may be important for explaining the dependent variable, maybe more important than the independent variables in a model. In this review, I will simulate data using variables that are then treated as unmeasured in the statistical model fitting, showing how unmeasured variables generate correlated random errors and challenges to model fitting and interpretation.

(iii) Diagnostics for the randomness of random errors are not useful, but simulations are.

Standard statistical textbooks contain methods for testing whether random errors in regression models are iid and Gaussian. For most ecological and evolutionary datasets, these diagnostics are of little value, for three reasons. First, diagnostics might say that the random errors are not iid, but they do not indicate how severely this should reduce your faith in the results from the model; in some situations, regression models are robust to departures from iid. Second, if diagnostics identify departure from iid, which is often likely, you still have to employ alternative models to re-analyse the data, and diagnostics do not necessarily help you decide what would be a good alternative model. Third, some of the statistical problems that arise with correlated random errors cannot be diagnosed, so diagnostics can give a false sense of security. I am not saying that diagnostics in general are not useful, but diagnostics for nonrandom random errors generally are not. A better approach than diagnostics is to start with models capable of estimating covariances in random errors and then to simulate the fitted models to understand how these covariances affect inference from the model. I base the three examples in this review on simulations. Because I use the simulations to illustrate the general challenges of covariances among random errors, I did not first fit the models to real data; nonetheless, the simulations show that diagnostics are not helpful, and that simulations are useful for building understanding of the statistical consequences of correlated random errors.

(iv) Model predictions can be made with random errors.

For statistical models, prediction takes three general forms: goodness-of-fit, interpolation/extrapolation and forecasting. Goodness-of-fit involves predicting what the values of the dependent variable should be if the model were correct. These predicted (fitted) values do not assess the quality of the data, but instead assess the quality of the model to capture the essential characteristics of the data. Interpolation and extrapolation for a simple linear model (Equation 1) involve first fitting the relationship between Y and X, and then predicting the value of Y_h for a new value of X_h that does not occur within the original dataset. If the value of X_h is within the range of the observed values of X, this is interpolation, whereas if it is outside the range of X, it is extrapolation. Interpolation and extrapolation are not so different from calculating fitted values from a model, except the uncertainty of the predicted value Y_h will depend on the uncertainty in the estimates of the parameters in the model (Neter et al., 1989). Forecasting involves predicting values of Y_t at time t from values of Y_s at earlier points in time ($s < t$) and possibly X_s or X_t if these are known. Although forecasting is a form of extrapolation, it is often distinguished from extrapolation because variability (either from random errors or uncertainty in parameter estimates) will compound through time: Y_t depends on Y_{t-1} depends on Y_{t-2}, etc., and variability will expand as predictions are made further into the future (Box et al., 1994).

Prediction is most often confined to the fixed components of a model: values of Y are predicted from values of X. For example, for goodness-of-fit the procedure is to estimate the regression coefficients, $\hat{b}_0$ and $\hat{b}$, and then compute the fitted values of $\mathbf{Y}$ from $\mathbf{X}$ as $\hat{b}_0\mathbf{J} + \hat{b}\mathbf{X}$. This is not quite as simple as it sounds, because estimation

of $\hat{b}_0$ and $\hat{b}$ depends on $\mathbf{V}(\boldsymbol{\theta})$ and requires the simultaneous estimation of $\boldsymbol{\theta}$. Nonetheless, there are standard statistical tools for this, and here I will use maximum likelihood (ML) or restricted maximum likelihood (REML). The association of the independent variables with prediction is clear from other names for independent variables: 'predictor variables' and 'explanatory variables'.

It is possible, however, to make better predictions. When the random errors are correlated, having estimates of one or more random error terms (i.e. the residual errors) makes it possible to estimate the expected value of other random errors. Specifically, the expected value of ε_i conditional on the estimated values of the other random errors (and conditional on values of $\boldsymbol{\theta}$) is (Box et al., 1994; Fuller, 1996; Petersen & Pedersen, 2012 equation 353)

$$\hat{\varepsilon}_i = \mathbf{V}[i,.]\mathbf{V}[-i,-i]^{-1}\left(\mathbf{Y}[-i]-\left(\hat{b}_0\mathbf{J}+\hat{b}\mathbf{X}[-i]\right)\right), \tag{2}$$

where $\mathbf{V}[i,.]$ is row i of matrix $\mathbf{V}(\boldsymbol{\theta})$, $\mathbf{V}[-i,-i]^{-1}$ is the inverse of $\mathbf{V}(\boldsymbol{\theta})$ after row i and column i are removed, and $\mathbf{Y}[-i]$ and $\mathbf{X}[-i]$ denote $\mathbf{Y}$ and $\mathbf{X}$ with element i removed. The term $(\mathbf{Y}[-i] - (\hat{b}_0\mathbf{J} + \hat{b}\mathbf{X}[-i]))$ gives the residual errors for the observations other than i. Goldberger (1962) shows that the predicted random errors from Equation (2) are the best linear unbiased estimates (BLUE), thereby providing statistical justification for their use. If $\hat{\boldsymbol{\varepsilon}}$ denotes the vector containing values of $\hat{\varepsilon}_i$, the fitted values of the independent variable are $\hat{\mathbf{Y}} = \hat{b}_0\mathbf{J} + \hat{b}\mathbf{X} + \hat{\boldsymbol{\varepsilon}}$. This approach to predicting values of Y is similar computationally to inverse-distance weighting methods used in spatial analyses, in which smoothing is performed based on residual errors weighted by expected covariances (Cressie, 1993; Wikle et al., 2019).

I illustrate this approach for predicting the expected values of random errors from residual errors for use in goodness-of-fit, interpolation and forecasting, pairing these with the spatial, phylogenetic and time-series models, respectively. The main point is to show that covariances among random errors provide information about the data.

(v) Can random errors be causal?

The enterprise of statistics separates the goal of prediction from the goal of causal inference (Prosperi et al., 2020; Shmueli, 2010). Roughly speaking, these are associated with R^2 and p-values respectively. On the extreme prediction side, many artificial intelligence algorithms (e.g. for face recognition) focus solely on prediction, and the information from the data used by the predictive algorithms is often unknown (Athey, 2017). On the extreme causal inference side, experimental designs in which treatments are assigned randomly to replicates are used to identify independent variables (treatments) that cause a response in the dependent variable. Most data and studies lie between these extremes. Observational data dominate many areas of ecology and evolution, and they present the 'correlation is not causation' challenge. Still, it might be possible to test causal assumptions from observational data if all reasonably expected, expertly inferred and possibly confounding variables are accounted for (Bollen & Pearl, 2013; Granger, 1969; Pearl, 2010). For example, structural equation modelling analyses the relationship between Y and X by constructing latent variables that confer specific structure to the random errors, with that structure based on causal assumptions about the processes underlying the data. This intersects with the goal of my review of emphasizing that random errors are caused by something, although I have described covariances among random errors as being generated by 'unmeasured variables' to avoid calling them 'latent variables' and the association with structural equation models.

For the review, I have focused on prediction rather than causal inference. But this does bring up the following question. In the three simulation studies, I make predictions from estimates of the random errors (Equation 2). Structural equation modelling contends that, if variables and covariances among random errors in a model are properly constructed, then a causal hypothesis about the relationship between Y and X can be tested. To draw these two arguments together, is it possible to test causal hypotheses using the random errors? Some researchers seem to think so: I frequently hear evolutionary biologists say 'phylogeny caused' a pattern in their observational data, yet these phylogenetic patterns are in the random errors. I pose this question here and leave my answer for the Discussion.

2 | SIMULATION STUDIES

To illustrate these five points, I use examples of spatial, phylogenetic and time-series models, pairing each with applications of prediction: goodness-of-fit, interpolation and forecasting. I base each example around a real dataset, although I move quickly to simulated data which makes it easier to illustrate the five points; for simulations, we are omniscient about the processes generating the data. The literature on analysing data with correlated random errors is huge, and my goal is not to be comprehensive. Instead, I present issues that have puzzled me and that have likely arisen in analyses that you have performed or will in the future.

2.1 | Space: Goodness-of-fit

My first example can be introduced with the question: Is Alaska greener in the south than in the north? It is now possible to ask questions like this over large areas at high resolution thanks to remote-sensing satellites (Figure 2). The challenge, however, is that points are not independent, and therefore even though a map might contain thousands or millions of pixels, the information available is not equivalent to having thousands or millions of independent samples. To address patterns of greenness in Alaska while accounting for spatial autocorrelation, Equation (1) can be used with Y being greenness (measured by, e.g. the normalized difference vegetation index, NDVI) and X being latitude. To model spatial autocorrelation, $\sigma^2\mathbf{V}(\boldsymbol{\theta})$ could take different forms, and I will use a simple form in which $\sigma^2\mathbf{V}(\boldsymbol{\theta})$ has elements $v_{ij} = \sigma^2[\phi + (1-\phi)\exp(-d_{ij}/\rho)]$ where d_{ij} is the geographical distance between locations i and j, and $\boldsymbol{\theta} = (\rho, \phi)$ where ρ is

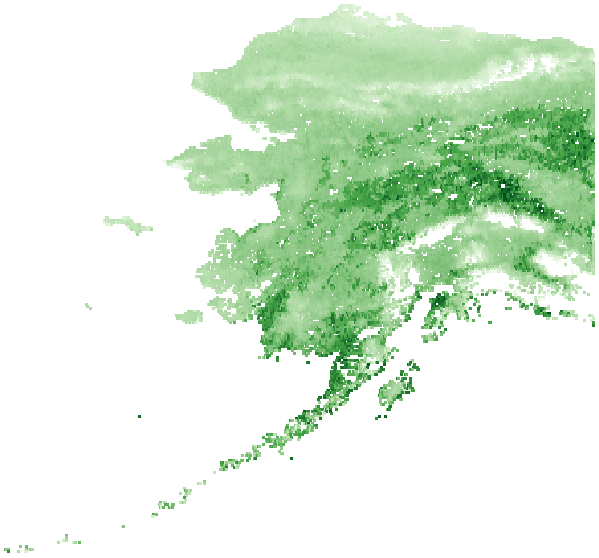

FIGURE 2 Mean annual greenness (normalized difference vegetation index, NDVI) for Alaska west of −141° longitude, 1982–2015, from the NDVI3g dataset derived from the AVHRR sensor (Ives et al., 2021). Values range from 0.09 (light green) to 17.22 (dark green), and pixels with insufficient reflectance to calculate NDVI (e.g. water bodies, permanent snow) are masked white

the 'range' parameter that scales how quickly spatial autocorrelation decreases with distance, and ϕ is the 'nugget' that gives the proportion of the variance that is local (having no spatial autocorrelation). $\sigma^2\mathbf{V}(\boldsymbol{\theta})$ can be depicted graphically to show the spatial structure of ε (Figure 1a). When applied to the Alaska data (using methods in Ives et al., 2021), this model shows a negative relationship between latitude and greenness ($b = -0.17$), although the statistical support for this relationship is not great ($p = 0.003$) considering the number of pixels (31486).

Regressions of Y on X are often conceived in terms of the effects of X on Y, with X mechanistically driving this effect, while the random errors capture a wide range of processes that might have complicated effects. In fact, interpreting the meaning of X is not necessarily easier than the meaning of ε. In the case of Alaska (Figure 2), although the measurement of latitude is clear, its biological relationship with greenness is indirect and complex. There are many variables affecting greenness that are also related to latitude. For example, annual photoperiod is mechanistically tied to latitude, and so is the angle of incidence that affects light intensity. Greenness is also affected by temperature and precipitation. While temperature and precipitation themselves show latitudinal patterns, these patterns are strongly modified by elevation, distance from oceans, etc. Furthermore, the effects of temperature and precipitation change depending on the time of year; for example, low snowfall in winter has different impacts on plants than low rainfall in summer. Even if variables such as temperature have strong effects on greenness, only variation in the variables that occur along the latitudinal gradient will be captured by the independent variable X in Equation (1). This leaves the variation in all of the variables affecting greenness that is not correlated with latitude to be captured in $\sigma^2\mathbf{V}(\boldsymbol{\theta})$. Thus, the separation between X and ε is not clean, and the real question addressed by the regression is: For all of the many variables that affect annual greenness in Alaska, how much of the variation they cause in greenness can be captured by latitude X and how much can be captured by the random errors ε?

To explore the prediction of fitted values when there is spatial autocorrelation, I simulated data on a 32×32 pixel map with the model

$$\begin{aligned} \mathbf{Y} &= b_0\mathbf{J} + b\mathbf{X} + \mathbf{U} + \boldsymbol{\gamma} \\ \mathbf{U} &\sim N\left(0, \sigma^2\mathbf{V}(\boldsymbol{\theta})\right) \\ \boldsymbol{\gamma} &\sim N\left(0, \sigma^2{}_{\gamma}\mathbf{I}\right), \end{aligned} \tag{3}$$

when there was a latitudinal gradient ($b = -1.5$). In contrast to the model fit to the data (Equation 1), the simulation model explicitly contains an unmeasured variable U that itself is spatially autocorrelated with $\sigma^2\mathbf{V}(\boldsymbol{\theta}) = \exp(-d_{ij}/r)$, while the simulated random errors γ are independent (the correlation matrix $\mathbf{I}$ is the identity matrix). Including U emphasizes that the random errors have underlying causes, but these causes are not known. In the simulations, the strength of spatial autocorrelation in U is zero, weak or strong ($r = 0$, 0.05 and 0.3, where distance is scaled so that each edge of the map has length 1). While there is no nugget for U, there is local variation caused by γ ($\sigma^2{}_{\gamma} = 0.05$). The three simulations (Figure 3) were performed with the same sequence of random numbers to generate U and γ, making results easier to compare.

When U is not spatially autocorrelated (Figure 3, top row), the effect of latitude X on greenness is visually clear, and the hypothesis that $b = 0$ is rejected ($p << 10^{-10}$). Goodness-of-fit including the estimated random errors (Equation 2) can be assessed with a prediction R^2 given by $1 - \mathrm{var}(\mathbf{Y} - \hat{\mathbf{Y}})/\mathrm{var}(\mathbf{Y})$ (this is $R^2{}_{\mathrm{pred}}$ discussed in Ives, 2019). Even though there is strong statistical support that $b < 0$, the fact that $R^2 = 0.13$ implies low goodness-of fit. With weak and strong autocorrelation in U, the statistical significance of the estimate of b drops ($p = 0.003$ and 0.086 respectively), yet R^2 increases ($R^2 = 0.58$ and 0.86), and the fitted values clearly capture the pattern in greenness (Figure 3, comparing first and second columns). The significance of b decreases because the model does not have sufficient information to adjudicate between latitude X and unmeasured U as explanations of the patterns in Y. Because in the simulation the overall patterns caused by U are stronger, statistical confidence in the effect of X is lost. This can be seen most clearly in a plot of Y versus X with strong autocorrelation ($r = 0.3$) which partially hides the linear decrease in Y with X in the simulation (red line). One could argue that the solution is to ignore spatial autocorrelation and perform the regression as OLS. The cost of this, however, is grossly inflated type I errors (false positives); if you simulate the model with $b = 0$ and $r = 0.3$, and ignore spatial autocorrelation in the regression, 85% of the simulations will reject the hypothesis that $b = 0$ ($p < 0.05$) even though it is true. Although the deterministic relationship between Y and X is hidden, however, the correlated random errors make it possible to predict values of ε_i from residuals in surrounding pixels using Equation (2), giving high R^2.

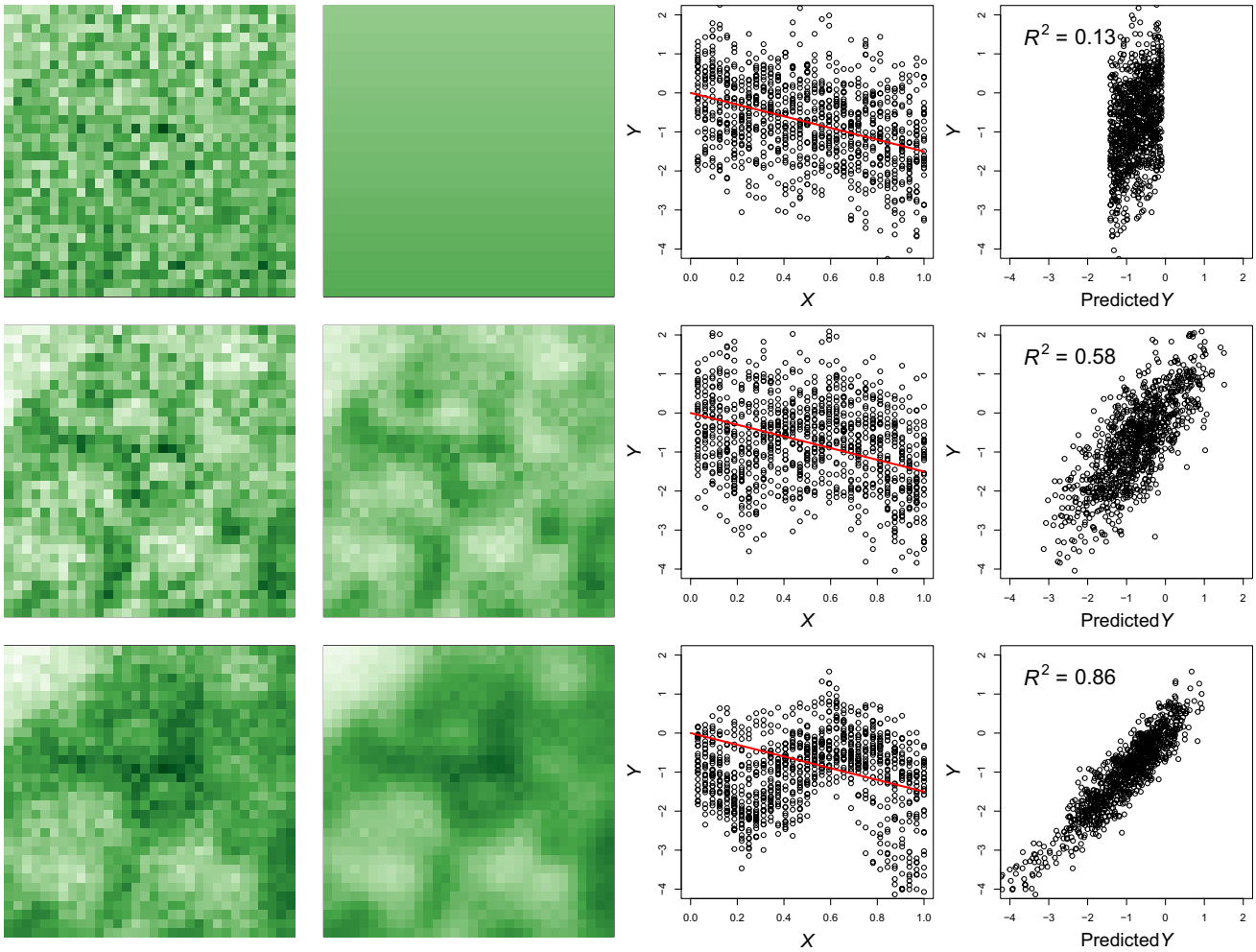

FIGURE 3 Simulation examples of a spatial process in which response variable Y depends on latitude X, an unmeasured variable U that may be spatially autocorrelated, and independent random error γ (Equation 3). Three simulations with the same sequence of random numbers were run with different strengths of spatial autocorrelation in U: none ($r = 0$), weak ($r = 0.05$) and strong ($r = 0.3$) in rows from top to bottom. Panels in the first column give the simulated data Y. Panels in the second column give the predictions of $\widehat{Y} = \widehat{b}_0 + \widehat{b}X + \widehat{\varepsilon}$ from Equation (2) that use information about the covariance among random errors. Panels in the third column give the relationship between Y and X, with the expectations that depend on the values of X (i.e. $b_0 + b_1X$) given by red lines. In an appropriate GLS analysis (Equation 1), the fitted slopes have higher p-values with increasing spatial autocorrelation ($p << 10^{-10}$, 0.003 and 0.086 respectively). Panels in the fourth column give the relationship between Y and $\widehat{Y}$ predicted using Equation (2), and the corresponding values of R^2 are 0.13, 0.58 and 0.86. Other parameter values for the simulations are $b_0 = 0$, $b = -1.5$, $\sigma^2 = 1$ and $\sigma^2_\gamma = 0.05$

Comparing predictions of Y using X to predictions of Y from the residuals in surrounding pixels might not seem fair, because values of X give additional information about the data Y—the relationship between Y and X—whereas using surrounding residuals just describes the pattern. However, in the simulation model the spatial autocorrelation is caused by U, and therefore making predictions using $\mathbf{V}(\boldsymbol{\theta})$ is just an attempt to infer the value of U in a pixel. This inference of U is not perfect; if we knew the values of U and used them in a regression model along with X to predict Y, then our R^2 would be 0.95 in all three of the simulation examples. The best we managed to do with strong autocorrelation was $R^2 = 0.86$. Still, that is not bad.

A common diagnostic used for spatial data is to compute a semivariogram for the residuals. Not surprisingly from the simulated data (Figure 3), semivariograms when U is autocorrelated ($r = 0.05$ and 0.3) show clear autocorrelation in the residual errors (Supporting Information Figure S1). This diagnostic, however, is not helpful; for example, if a semivariogram were used to select pixels sufficiently distant so that there is little remaining autocorrelation (correlation <0.05), only about 80 pixels would remain for analysing the Alaska dataset (Figure 2), and the coefficient for latitude would not be significant. Therefore, a statistical model that explicitly accounts for spatial autocorrelation is needed from the start (Cressie, 1993; Dormann et al., 2007; Legendre, 1993).

This example illustrates the limits that independent variables may have in 'explaining the data', even when good predictions can be made using the residual errors. The model can even be used (via a likelihood ratio test) to assign a p-value to the hypothesis that the random errors are not correlated ($\phi = 1$ and/or $\rho = 0$); for the

cases $r = 0$, 0.05, and 0.3, these are $p = 0.44$ ($\chi^2_2 = 1.6$), $p << 10^{-10}$ ($\chi^2_2 = 578$) and $p << 10^{-10}$ ($\chi^2_2 = 1{,}716$). Although calculating p-values is often associated with causal inference, here I am only using p-values to give evidence for the existence of correlation among random errors that depends on the distance between points. If identifying explanatory independent variables were the only goal, autocorrelation would be a nuisance. However, a broader view would include autocorrelation as part of the explanation.

2.2 | Phylogenies: Interpolation

In a classic study of social organization, Jarman (1974) sorted 75 species of African antelope according to such characteristics as feeding behaviour, diet, response to predators, body weight and habitat. While acknowledging that the causal relationships among these characteristics are impossible to know, he nonetheless describes how it is possible to explain these relationships in terms of constraints ('appropriate strategies') imposed by the needs of individuals to feed and survive the risk of predation. One relationship is between the antipredator strategies of antelope and their group size (Figure 4a). Antelope can be divided into those that freeze or hide when they fear predation and those that flee or fight. Antelope that hide occur in smaller groups than those that flee/fight. This relationship makes sense, because the antipredator strategy of hiding is harder in large groups, and fleeing and fighting are more effective if larger groups more easily evade predators (such as the 'explosive herd' flight behaviour of impala) or put up a communal defence (such as water buffalo). Nonetheless, numerous other characteristics correlate with group size, including body size and feeding mode (browsing vs. grazing). These other characteristics likely show phylogenetic correlations (or 'phylogenetic signal' sensu Blomberg et al., 2003); for example, phylogenetically related species are more likely to share similar body sizes. Therefore, when analysing the relationship between antipredator strategy and group size (Figure 4a), it is appropriate to consider the phylogenetic relationships among species (Brashares et al., 2000; Felsenstein, 1985; Garland et al., 1992; Harvey & Pagel, 1991).

A regression of group size on antipredator behaviour can be performed using Equation (1) with Y containing group size and X containing the categories 'hide' versus 'flee/fight'. The only difference between this application to phylogenetic data and that used for spatial data is the structure of the covariance $\sigma^2\mathbf{V}(\boldsymbol{\theta})$, which captures the hypothesized covariances among species depending on their degree of relatedness (Figure 1b). To model phylogenetic relatedness, sensible hypotheses for $\sigma^2\mathbf{V}(\boldsymbol{\theta})$ can be derived from phenomenological models of evolution in which species trait values diverge along a phylogenetic tree; the more distant the common ancestor on the tree, the more divergence is likely to have occurred (Felsenstein, 1985). Including a parameter $\boldsymbol{\theta}$ in $\sigma^2\mathbf{V}(\boldsymbol{\theta})$ can give a measure of phylogenetic signal in terms of the overall magnitude of covariances between species (Freckleton et al., 2002; Grafen, 1989; Hansen & Martins, 1996); for example, $\boldsymbol{\theta}$ could be Pagel's λ that equals one if the covariances among species reflect Brownian motion evolution up the phylogeny and zero if there is no phylogenetic signal (Pagel, 1997). When applied to the Jarman (1974) data (Figure 4a), there is a strong relationship between antipredator strategy and log group size ($b = 0.34$, $p = 0.0004$) and strong phylogenetic signal ($\lambda = 0.92$).

I performed simulations designed from the data of Jarman (1974) for 32 species showing a balanced branching phylogeny (Figure 4b). The simulations include an unmeasured variable U that has phylogenetic signal and an additional random error term γ that is independent among species (Equation 3). In simulation A, the antipredator strategy X is split between the two major clades, whereas in simulation B the antipredator strategy alternates between closest relatives; in both simulations, the unmeasured variable U is the same.

How well can the statistical model (Equation 1) identify a relationship between group size Y and antipredator strategy X? An analysis of the particular simulations in figure 4b gives a nonsignificant ($p = 0.15$) coefficient of antipredator behaviour on group size for simulation A, even though an OLS regression that ignores phylogeny gives a highly significant effect ($p = 10^{-8}$). This type of finding—that phylogeny 'reduces the power' of statistical tests—is common, although interpreting this as reduced power is incorrect, because the

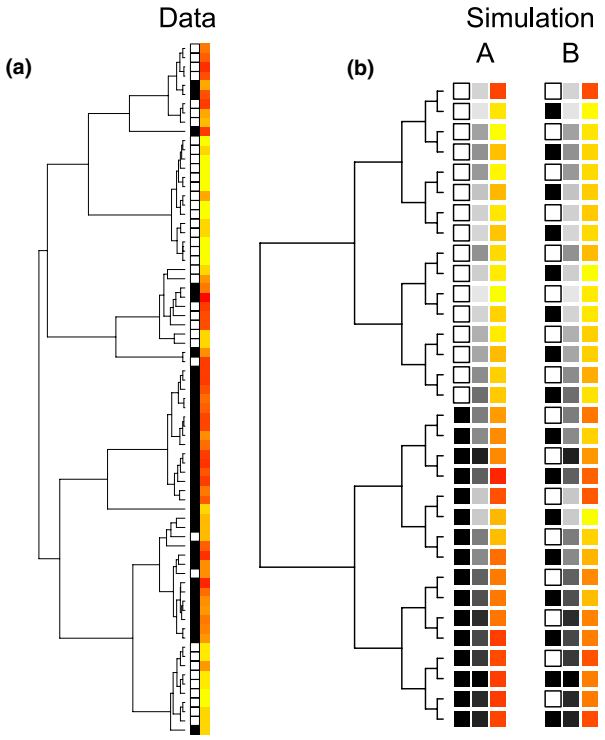

FIGURE 4 (a) Phylogenetic relationship of 75 species of African antelope from Jarman (1974), with antipredator behaviour (white = hide, black = fight/flight) and group size (1–50 corresponding to yellow - red) (data in Ives & Garland, 2010). (b) For a balanced phylogeny of 32 species, two sets of simulated values of group size (yellow - red) when antipredator behaviour is distributed according to the two major clades (simulation A) and alternating between close relatives (simulation B). Group size also depends on a third unmeasured valuable U (grey scale) that has phylogenetic signal ($\lambda = 1$)

high power implied by OLS regression never really existed. We can understand the low power to detect an effect of antipredator strategy by performing an OLS regression now including the values of U (since in the simulation we know what they are). With the inclusion of U, the effect of X is significant ($p = 0.00003$). The reason X is not significant when U was unknown is because U and X are highly correlated (−0.68). This correlation arose because X by design differed between clades and U was simulated with phylogenetic signal. Because X and U, and hence ε, were correlated, it is impossible to statistically identify the variation in Y that can properly be attributed to X. This type of issue will arise whenever X itself has phylogenetic signal as well as the random errors ε.

Simulation B shows the opposite situation in which X has anti-phylogenetic signal in which phylogenetically closely related species are less likely to share the same value of X. In this case, the coefficient for antipredator strategy on group size is highly significant ($p = 0.00006$) even though an OLS regression ignoring phylogeny gives a marginally nonsignificant effect ($p = 0.051$). In simulation B, the values of X and U are very weakly correlated (0.04), so phylogenetic signal in the residual errors of Y does not absorb the effects of variation in X.

The problems illustrated by the simulations for OLS regression—inflated type I error rates (false positives, simulation A) and the inability to identify patterns (loss of power, simulation B)—arise whenever an independent variable is correlated with the random errors. One of the basic assumptions of OLS regression is the exogeneity assumption, which states that the independent variable is uncorrelated with the random errors (Judge et al., 1985). The exogeneity assumption is violated in the phylogenetic simulations. In the simulations, we know when this is a problem, because we know the values of U underlying the covariance among random errors. This is not the case with real data, and it is only possible to diagnose this problem using a hypothesis for the covariances among ε. This hypothesis is embodied in $\sigma^2\mathbf{V}(\boldsymbol{\theta})$, so it makes sense to include $\sigma^2\mathbf{V}(\boldsymbol{\theta})$ in the analyses from the start and not worry about diagnostics.

For prediction, I will consider the task of predicting the value of log group size Y_h for a new species h with antipredator strategy X_h. I simulated data for 32 species, removed one of the species, fit the model (Equation 1) to the data from the remaining 31 species, and then predicted the value of Y_h for the species that had been removed, $\widehat{Y_h} = \widehat{b}_0\mathbf{J} + \widehat{b}X_h + \widehat{\varepsilon_h}$ (Equation 2). To assess the predictions, I used a prediction $R^2 = 1 - (Y_h - \widehat{Y_h})^2/\mathrm{var}(Y)$, where var($Y$) is the variance of the simulated log group sizes for the 31 species used to make the predictions $\widehat{Y_h}$. In application to a single dataset, there is only a single predicted value, and this measure of prediction accuracy can still be applied because it compares the squared difference between observation and prediction with the expectation of this squared difference across the remainder of the dataset.

For 1000 simulated datasets, the prediction R^2 was similar regardless of the distribution of X on the phylogeny (Table 1). This prediction R^2, however, was not as high as what could be achieved if U were known for all species. A curious result occurs when comparing predictions made using OLS regression. In this case, the prediction R^2 from X_h when X is divided between the two main clades (simulation A) is 0.33 averaged among replicates, even though the estimate of b was rarely significant (Table 1). This occurred because even though b was not significant, $\widehat{b}X_h$ nonetheless captured variation in Y that was caused by U: in the simulations, the effect of U on Y was stronger than the effect of X on Y, and because X was correlated with U, values of X predicted values of Y. In contrast, when values of X alternated between closest relatives (simulation B), the average prediction R^2 from X_h alone was 0.07, yet the estimate of b was significant in 99% of the simulations (Table 1). Thus, comparing simulations A and B, being able to detect a significant effect of an independent variable gives misleading information about the ability to make predictions from it. These conclusions hold for other measures of prediction accuracy (Supporting Information Table S1).

These simulations give two general lessons. First, in regression models with correlated random errors, the distribution of X among samples makes a difference. For OLS regression, this is not the case (Neter et al., 1989). A very common finding is that a model justifiably containing covariances in the random errors gives higher p-values ('less significance') for the regression coefficients than incorrectly applied OLS regression. Nonetheless, this is not a rule, and you can find that the covariance structure increases your power to identify significant regression coefficients. Second, if you make predictions using all of the information in a dataset—including both independent variables and $\sigma^2\mathbf{V}(\boldsymbol{\theta})$—then the accuracy of the predictions may be independent of your ability to detect significant regression coefficients. In both simulations A and B, the model had similar prediction accuracies for new values Y_h (Table 1, Table S1), and therefore the data contain the same total amount of information that can be

TABLE 1 Prediction R^2 for the values of Y_h (group size) given X_h (antipredator strategy) for one of 32 species in balanced phylogenies (Figure 4b). Simulations were performed in which the distribution of X was either divided between the two main clades (simulation A) or alternated between closest relatives (simulation B). In addition to predictions using both values of X_h and phylogenetic relationship among species (Equation 2) given by R^2, predictions were made using values of U and using OLS without information about U. Listed values of R^2 for all three types of predictions are the mean from 1000 simulations. Mean values of the estimates of phylogenetic signal (Pagel's λ) and b are also presented, along with the median p -value for the significance of b and the percentage of p -values < 0.05. See also Table S1

Distribution of X	R^2	R^2 (U known)	R^2 (OLS)	λ	b	Median $p(H_0{:}b = 0)$	Percent $p < 0.05$
A. two clades	0.61	0.81	0.33	0.70	0.64	0.29	25%
B. alternating	0.65	0.82	0.07	0.84	0.60	0.00008	99%

used to make predictions. However, this information is apportioned differently between mean and variance components of the fitted regression models. As Ted Garland wrote, "Thus, what phylogeny taketh away, phylogeny giveth back, at least with proper statistical methods" (Garland & Ives, 2000).

The key to understanding phylogenetic regression is interpreting the random error terms as being underlain by unmeasured traits that themselves have phylogenetic signal (Felsenstein, 1985). Random errors are caused by something. Even when no observable traits can be associated with the random errors, unmeasured traits can include such things as genomic architecture. Genomic architecture will show phylogenetic signal and also affect the expression of macroscopic traits, and hence may generate phylogenetic signal in the random errors.

2.3 | Time series: Forecasting

As an example time series, consider the iconic hare–lynx population cycles, a mainstay in ecology textbooks (Odum & Barrett, 1971). The remarkable hare–lynx dataset comes from the Hudson's Bay Company that recorded the purchases of hare and lynx pelts from Canada from 1845 to 1935 (Figure 5). Although this dataset is often presented as an example of a 'simple' predator–prey cycle, the real situation is more complex (Krebs et al., 2017). For this example, however, I will ignore the real complexities to allow a simple model of predator–prey interactions.

To simulate predator–prey dynamics, I used a version of the Lotka–Volterra equations in discrete time,

$$\begin{aligned} Z_1(t) &= Z_1(t)\exp\left[c_{10} - c_{11}Z_1(t-1) + c_{12}Z_2(t-1)\left(1 - hZ_1(t-1)\right) + \alpha_1(t)\right] \\ Z_2(t) &= Z_2(t)\exp\left[c_{21}Z_1(t-1)\left(1 - hZ_1(t-1)\right) - c_{22}Z_2(t-1) + \alpha_2(t)\right], \end{aligned} \tag{4}$$

where $Z_1(t)$ and $Z_2(t)$ are the abundances of prey and predator in year t, coefficients c_{ij} govern their intraspecific and interspecific interactions, h is a measure of the 'handling time' that makes the predation rate depend on prey density, and $\alpha_1(t)$ and $\alpha_2(t)$ are normal random errors with mean zero and variance σ^2 that are assumed to be independent of each other and through time. This is a nonlinear equation, and in the absence of any stochasticity ($\sigma^2 = 0$), it can give two qualitatively different dynamical patterns depending on parameter values. One pattern is stable limit cycles in which the cycles persist indefinitely, and the other is damped oscillations in which the cycle amplitude diminishes through time and population abundances

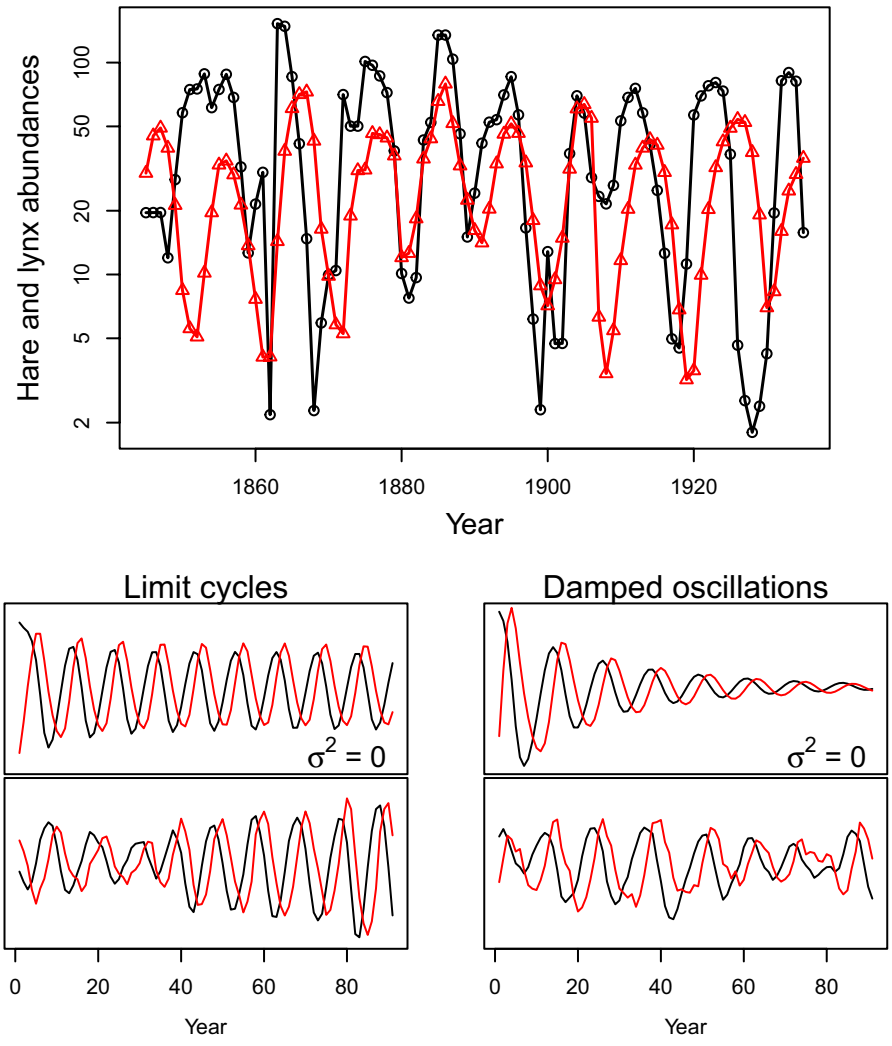

FIGURE 5 Abundances of snowshoe hare and lynx from pelts collected by the Hudson's Bay Company, 1845–1935 (Deng, 2018). The lower four panels illustrate simulations showing stable limit cycles ($c_{10} = 2$) and damped oscillations ($c_{10} = 1.5$) from different parameterizations of Equation (4). The upper panels with $\sigma^2 = 0$ have no random errors to show the deterministic dynamics, while the lower panels show the stochastic dynamics as analysed for Figure 6 and Table 2 ($\sigma^2 = 0.0025$ and 0.01 for limit cycles and damped oscillations). Other parameters are $c_{12} = -6$, $c_{11} = c_{21} = c_{22} = 1$, and $h = 0.4$

approach single points (Figure 5). In deterministic ecological theory, these two patterns are fundamentally different, yet when stochasticity is added ($\sigma^2 > 0$) the dynamics are similar to the eye (Figure 5), because the stochasticity perpetuates the damped oscillations to create quasi-cyclic dynamics.

For fitting simulated data, I am going to take two conceptual leaps. First, although the data are generated by a nonlinear simulation, I am going to use a linear model. Linear statistical models are easier to fit than nonlinear ones, and therefore the resulting forecasts might be more robust. Second, I am going to fit only data from the prey. Ecological systems are always multidimensional, involving interactions between numerous components of ecosystems. However, we never have data on everything. An important result from theory is that the dynamics caused by interacting multidimensional processes leave fingerprints on the dynamics of any one variable. Formally, the information that resides in time-lags in the dynamics of any one variable is sufficient to reconstruct the dynamics of the whole multidimensional system for both deterministic ('Taken's theorem', Takens, 1981) and stochastic systems (Stark et al., 2003).

After a little algebra, a 2D linear model that approximates Equation (4) can be recast as a 1D linear model with a time delay (Abbott et al., 2009),

$$Y(t) = b_0 + b_1 Y(t-1) + b_2 Y(t-2) + \eta(t) + d\eta(t-1), \tag{5}$$

where $Y(t)$ is the log abundance of prey. Thus, linear predator–prey interactions give rise to a 1D linear model with two time-lags in $Y(t)$ and one time-lag in $\eta(t)$ given by the moving average term $d\eta(t–1)$; Equation (5) is an autoregressive-moving average process, ARMA(2,1) (Box et al., 1994). A little more algebra turns Equation (5) into (Ives & Zhu, 2006).

$$\begin{aligned} \mathbf{Y} &= b_0\mathbf{J} + \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon} &\sim N\left(0, \sigma^2\mathbf{V}(\boldsymbol{\theta})\right). \end{aligned} \tag{6}$$

This is the same model as used for the spatial and phylogenetic examples (Equation 1), but without independent variable X. All of the dynamics are captured in $\sigma^2\mathbf{V}(\boldsymbol{\theta})$, which contains the covariances between $Y(t)$ and $Y(t–s)$ separated by s years. These covariances cycle between positive and negative values reflecting the cyclicity of the dynamics (Figure 1c).

It might seem odd that a time-series model (Equation 5) that appears to have independent variables can nonetheless be expressed in terms of a covariance matrix $\sigma^2\mathbf{V}(\boldsymbol{\theta})$ (Equation 6). This might seem less odd when considering that the apparent independent variables $Y(t–1)$ and $Y(t–2)$ are themselves determined from prior observations, and therefore $Y(t–1)$ and $Y(t–2)$ are themselves random variables.

I simulated 70 years of data using parameter values that produce either stable limit cycles or damped oscillations, and then forecast the abundance of prey for 30 years. When the dynamics show stable limit cycles, the forecasts from the linear 1D model are better than when the dynamics show damped oscillations (Figure 6a,b). This contrast is due to the uncertainty caused by the random errors $\varepsilon(t)$, rather than the uncertainty in parameter estimates: the uncertainty in the forecasts caused by only the random errors is shown in green, while the total uncertainty is shown in blue (Figure 6a,b). Stable limit cycles impose more regular structure onto the time series, and this gives more information for forecasting in contrast to the case with quasi-cycles. This underscores that even though the ability to make forecasts can be improved with better statistical models, some real-world systems will be inherently difficult to predict, simply because there is little signal generated by the processes underlying the data, as in the case of quasi-cycles.

I also statistically fit the data to the same nonlinear 2D model that produced the data (Figure 6c,d). To compare the forecasts of the two fitted models, I applied them to 1000 simulations for parameter values giving either limit cycles or damped oscillations (Table 2). The nonlinear 2D and linear 1D models had similar accuracies when the dynamics showed stable cycles. However, the nonlinear 2D model was outperformed by the linear 1D model when the dynamics showed damped oscillations. In fact, the average prediction R^2s for the nonlinear 2D model were negative for forecasts of 20 and 30 years (Table 2), implying that simply forecasting the mean prey abundance would have been more accurate. Given the different performances of linear and nonlinear models for making forecasts depending on whether the data showed limit cycles or damped oscillations, it would be useful to have diagnostics to distinguish the two. A standard diagnostic for time-series models, computing the autocorrelation among residuals, failed to show any difference between the time series (Supporting Information Figure S2), and a standard diagnostic for detecting nonlinear dynamics (Brock et al., 1996) failed to reject the hypothesis that the dynamics are linear for both simulations. Therefore, diagnostics do not help to determine whether dynamics are linear or not.

I want to highlight two lessons from this simulation study. First, even though variables are not measured, they may leave their imprint on the random errors. In the spatial and phylogenetic simulations, I used unmeasured variables to generate spatial and phylogenetic covariances. In the time-series example, even if information about predator abundances is unknown, the signature cyclicity of predator–prey interactions is still left on the prey abundances, and this creates covariances among random errors that can be used for forecasts. Second, while it is important to use a statistical model that can fit the data well, sometimes even a simple linear model can fit the data well enough to make good predictions. Here, a simple 1D linear model could outperform the 2D nonlinear model that was used to simulate the data. The surprisingly good performance of the linear 1D model occurs when the prey dynamics show damped oscillations, dynamics that linear models are capable of mimicking. Therefore, the performance of a linear statistical model fit to data generated by nonlinear processes does not so much depend on the nonlinearities underlying the processes but instead on whether the qualitative form of the dynamics (e.g. limit cycles vs. damped oscillations) can be mimicked by a linear model.

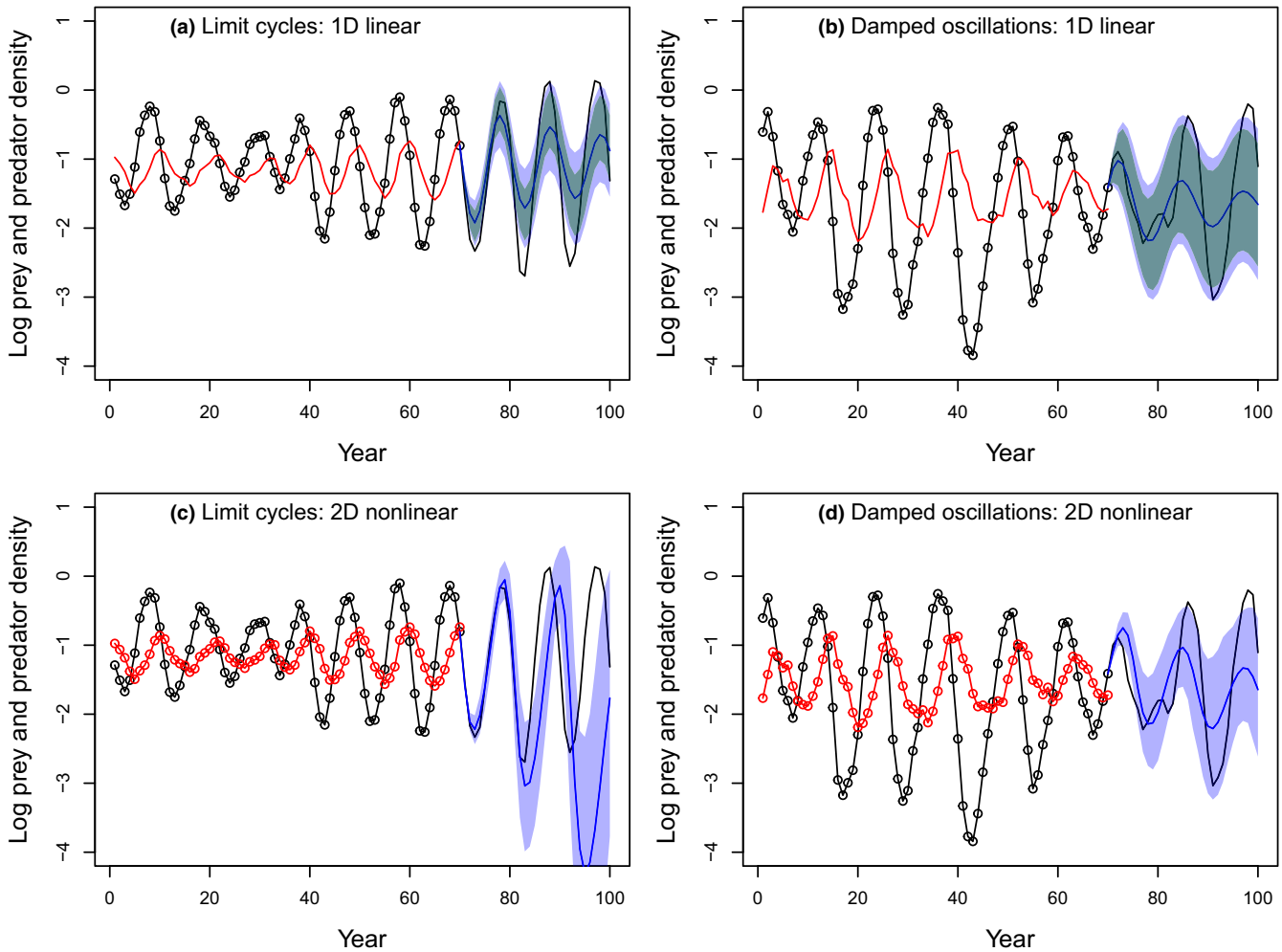

FIGURE 6 Forecasts from 70 years of data from the simulation model in Equation (4) for parameterizations showing (a, c) limit cycles and (b, d) damped oscillations. (a, b) The linear 1D model (Equation 6) was fit to only the prey data (black points), while (c, d) the nonlinear 2D model (Equation 4) was fit to both prey (black points) and predator (red points) data. Forecasts are given as solid blue lines with ±1 *SD* in the blue regions; the continuing black line gives simulated values that are being forecast. For the linear 1D model, forecasts are also given without accounting for parameter uncertainty by the green region

TABLE 2 Prediction R^2 for forecasts of prey abundance from 70 years of data simulated from a nonlinear predator–prey model. Forecasts for 5, 10, 20 and 30 years were made using a linear 1D model (Equation 6) and the nonlinear 2D model used to simulate the data (Equation 4). Values of the prediction R^2 are scaled so that when they are zero, the predictions are as accurate as using the mean prey abundance as the forecast. Values of R^2 are the averages from 1000 simulations

Dynamics	Years	R^2 (linear 1D)	R^2 (nonlinear 2D)
Limit cycles	5	0.81	0.88
	10	0.78	0.75
	20	0.47	0.46
	30	0.30	0.26
Damped oscillations	5	0.60	0.44
	10	0.50	0.12
	20	0.44	−0.08
	30	0.34	−0.13

3 | DISCUSSION

The five points that centre this review are illustrated in different ways by the three simulation studies involving spatial, phylogenetic and time-series data. The points are scattered throughout the simulation studies, so here I gather the results around the five points.

(i) Correlated random errors unite many types of statistical models, including spatial, phylogenetic and time-series models.

The three simulation studies all used the same GLS approach, with specific models differing only in the form of the covariance matrix of random errors (Figure 1). To allow this comparison among spatial, phylogenetic and time-series data, I have restricted the analyses to linear Gaussian models (Equation 1). Generalized linear models (GLMs) are designed for non-Gaussian data (McCullagh & Nelder, 1989), and they can be extended to generalized linear mixed models (GLMMs) that also allow covariances among random errors

(Gelman & Hill, 2007). The details differ between GLMMs and GLS, but just as for GLS, GLMMs can be applied to spatial, phylogenetic and time-series data. More-complex models, both frequentist and Bayesian, can similarly be applied to different types of correlated data by the careful specification of covariance matrices for the random effects. The fact that spatial, phylogenetic and time-series models differ only in how random errors covary shows the value in understanding how covariance matrices are constructed and analysed in statistical models.

(ii) Random errors are neither unpredictable nor mistakes.

If random errors are correlated, then they are likely affected by some real but unmeasured variable. In some cases, these unmeasured variables might really be nuisances, such as measurement errors that are systematically biased and generate correlated random errors. Most often, however, the unmeasured variables have real biological effects on the dependent variable. This means that the random errors contain useful information. I am not arguing that the random errors should necessarily be the focus of study. However, correlation among random errors should be carefully considered for the information it might give. For example, if you perform an analysis that includes correlated errors, the estimates for the covariance matrix should always be reported. It might also be useful to report R^2 values for models without and with correlated random errors fitted to the same data so that the information content from the independent variables can be compared to that available from unmeasured variables (e.g. Table 1). I do not want to disguise the fact that there are technical issues in defining an R^2 for models with correlated random errors (Kvalseth, 1985), but R^2s can still be useful (Ives, 2019).

(iii) Diagnostics for the randomness of random errors are not useful, but simulations are.

When there are reasons to suspect that random errors are correlated, it makes sense to skip diagnostics and start with statistical models that explicitly incorporate hypotheses about how the random errors are correlated. Diagnostics might only serve to identify the obvious (such as spatial autocorrelation in the greenness of Alaska, Figure 3, Figure S1) or fail to identify properties of the data that are important (such as nonlinearities in time series, Figure 6, Figure S2). In the GLS regression in Equation (1), $\sigma^2\mathbf{V}(\boldsymbol{\theta})$ contains parameters $\boldsymbol{\theta}$ that govern the pattern and strength of covariances among random errors, and statistical tests of $\boldsymbol{\theta}$ (e.g. whether $\phi = 1$ and/or $\rho = 0$ in the spatial model, and whether $\lambda = 0$ in the phylogenetic model) will tell whether the assumption of zero covariances among random errors can be rejected. I emphasize that $\sigma^2\mathbf{V}(\boldsymbol{\theta})$ should be treated as a hypothesis. Furthermore, there are different ways in which $\sigma^2\mathbf{V}(\boldsymbol{\theta})$ can be formulated (e.g. different spatial functions and different phylogenetic transforms), and these can be competed against each other to select the one that fits the data best (e.g. using likelihoods).

The simulations illustrate the importance of incorporating correlation among random errors. When analysing real datasets, simulations can also be used to assess the performance of the statistical model. Models with correlated random errors can have poor statistical properties, such as bias and inflated type I errors, for small datasets, where 'small' might be several hundred data points. The best way to identify these problems is to fit a model to data, simulate data from the model, and then refit the simulated data to test, for example, the type I error rates; I recommend this for any statistical model beyond OLS (Ives, 2018b).

(iv) Model predictions can be made with random errors.

I have emphasized that predictions can be made from covariances among random errors just as from independent variables. Conceptually, this is justified by recognizing that covariances among random errors are generally caused by unmeasured variables. Practically, this involves data smoothing in which the predicted value of a random error for a given data point is calculated as the weighted average of residual errors from the other data points, where the weights depend on the estimated covariances among random errors (Equation 2). Interpreting predictions from covariances among random errors as a smoothing technique de-mystifies the process. It also points to a difference between predictions from the covariances among random errors and predictions from independent variables. Suppose you fit a spatial model with no independent variables to data from Alaska (Figure 2) and the model has a high R^2. This model, however, would give no information to predict the level of greenness elsewhere in the world. In contrast, if latitude were included and had an effect on greenness in Alaska, it might be expected to have a similar effect in northern Canada or Eurasia. A fundamental difference between predictions from independent variables and predictions from covariances among random errors is that the latter are necessarily local in scope: they are based on smoothing residual errors, and therefore the predictions can only be made for points that are close enough in data space to have correlated random errors.

For many types of data and questions, I do not see the limited scope of predictions from correlated random errors as a large hindrance, because the same limitations likely apply to the independent variables as well. For example, I would be sceptical that a statistically significant effect of latitude on greenness in Alaska implies a similar relationship between latitude and greenness in northern Canada or Eurasia, because greenness depends on many variables that are captured at most crudely by latitude, if at all. Latitude does not 'cause' greenness in a direct sense, and therefore any inference about the association of greenness with latitude might be local in scope just as inference made from covariances among random errors.

(v) Can random errors be causal?

I asked this question to address two issues about interpreting random errors as being caused by something. First, it is not uncommon to hear researchers talk about random errors as causes, such as evolutionary biologists talking about patterns 'caused by phylogeny'. Similarly, Tobler's First Law of Geography states "everything

is related to everything else, but near things are more related than distant things" (Tobler, 1970). Tobler's 'Law', however, is not a causal law of nature, but is instead an empirical pattern that is commonly observed (Miller, 2004). Similarly, describing phylogeny as causal is a convenient shorthand for saying that there are many unmeasured variables that themselves could generate patterns in data that reflect phylogenetic associations. While it is important to be clear when interpreting the results of specific analyses of specific datasets about what can logically be inferred, I do not think it is necessary to be dogmatic in banning colloquial use of the word 'cause'; if a relationship between *Y* and *X* is statistically significant, I do not think we should exorcize anybody who, out of convenience, refers to 'the effect of *X* on *Y*' or even says '*X* causes *Y*'.

The second reason I have asked this question is concern that my championing of random errors as representing unmeasured variables will throw this review into the debate about statistics and causation (Bollen & Pearl, 2013). It should not. I personally think the issue of causation has much more to do with the data than the statistical methods used to analyse them. For example, scientists would probably agree that the most convincing way to show causation is to perform an experiment in which levels of *X* are randomly assigned to many replicates. Suppose I performed a highly replicated, well-controlled, randomized experiment in the field to show that natural enemies have a large negative impact on pea aphid pests in lucerne (alfalfa) fields. I conclude that natural enemies *caused* a reduction in pea aphid population growth at the field site and during the weeks I performed the experiment. However, this conclusion does not mean that, if I were to do the same experiment at a different site or at the same site in a different year, I would get the same result (Ives, 2018a). My point is that even though knowing causal relationships is useful and good justification for doing experiments, causal relationships may have limited scope and predictive power in ecology and evolution. Demonstrating a causal relationship is a far cry from demonstrating a 'law of nature'.

The point of statistics is to better understand data, and scientific inference depends on the correct coupling of data and appropriate analyses. Regardless of the specific scientific inferences you hope to draw from your data, I suspect that careful consideration of random errors will be informative. Data are hard won, and statistics should be used to make the most of them.

ACKNOWLEDGEMENTS

Thanks to my laboratory, members of the course iBio/Ent 540 (UW-Madison, 2021) and Ted Garland for valuable comments on the manuscript, and especially to Aaron Ellison for his inspirational insights. The manuscript also benefitted greatly by help from reviewers. This work was supported in part by NASA AIST program grant 80NSSC20K0282 and NSF grant DEB-2134446.

CONFLICT OF INTEREST

The author declares no conflict of interest.

DATA AVAILABILITY STATEMENT

Data for Figures 1, 4 and 6 are available from Ives et al. (2021), Ives and Garland (2010) and Deng (2018) respectively. They are also included with the computer code used for all of the analyses (Ives, 2022).

ORCID

Anthony R. Ives https://orcid.org/0000-0001-9375-9523

REFERENCES

Abbott, K. C., Ripa, J., & Ives, A. R. (2009). Environmental variation in ecological communities and inferences from single-species data. *Ecology*, *90*, 1268–1278. https://doi.org/10.1890/08-0487.1

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, *355*, 483–485. https://doi.org/10.1126/science.aal4321

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Blomberg, S. P., Garland, T., Jr., & Ives, A. R. (2003). Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. *Evolution*, *57*, 717–745. https://doi.org/10.1111/j.0014-3820.2003.tb00285.x

Bollen, K. A., & Pearl, J. (2013). Eight myths about causality and structural equation models. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 301–328). Springer Netherlands.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis: Forecasting and control* (3rd ed.). Prentice Hall.

Brashares, J. S., Garland, T., Jr., & Arcese, P. (2000). Phylogenetic analysis of coadaptation in behavior, diet, and body size in the African antelope. *Behavioral Ecology*, *11*, 452–463. https://doi.org/10.1093/beheco/11.4.452

Brock, W. A., Scheinkman, J. A., Dechert, W. D., & LeBaron, B. (1996). A test for independence based on the correlation dimension. *Econometric Reviews*, *15*, 197–235. https://doi.org/10.1080/07474939608800353

Cressie, N. A. C. (1993). *Statistics for spatial data* (revised ed.). Wiley & Sons, Inc.

Deng, B. (2018). An inverse problem: Trappers drove hares to eat lynx. *Acta Biotheoretica*, *66*, 213–242. https://doi.org/10.1007/s10441-018-9333-z

Dormann, C. F., McPherson, J. M., Araújo, M. B., Bivand, R., Bolliger, J., Carl, G., Davies, R. G., Hirzel, A., Jetz, W., Kissling, D. W., Kühn, I., Ohlemüller, R., Peres-Neto, P. R., Reineking, B., Schröder, B., Schurr, F. M., & Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography*, *30*, 609–628. https://doi.org/10.1111/j.2007.0906-7590.05171.x

Felsenstein, J. (1985). Phylogenies and the comparative method. *The American Naturalist*, *125*, 1–15. https://www.jstor.org/stable/2461605

Freckleton, R. P., Harvey, P. H., & Pagel, M. (2002). Phylogenetic analysis and comparative data: A test and review of evidence. *The American Naturalist*, *160*, 712–726. https://doi.org/10.1086/343873

Fuller, W. A. (1996). *Introduction to statistical time series* (2nd ed.). John Wiley & Sons.

Garland, T., Jr., Harvey, P. H., & Ives, A. R. (1992). Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Systematic Biology*, *41*, 18–32. https://doi.org/10.1093/sysbio/41.1.18

Garland, T., Jr., & Ives, A. R. (2000). Using the past to predict the present: Confidence intervals for regression equations in phylogenetic

comparative methods. *The American Naturalist*, *155*(3), 346–364. https://doi.org/10.1086/303327

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Goldberger, A. S. (1962). Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, *57*, 369–375. https://doi.org/10.2307/2281645

Gotelli, N. J., & Ellison, A. M. (2004). *A Primer of Ecological Statistics*. Sinauer Associates.

Grafen, A. (1989). The phylogenetic regression. *Transactions of the Royal Society of London B, Biological Sciences*, *326*, 119–157. https://doi.org/10.1098/rstb.1989.0106

Granger, C. W. J. (1969). Investigating causal relations by econometric models and Ccoss-spectral methods. *Econometrica*, *37*, 424–438. http://www.jstor.org/stable/1912791, https://doi.org/10.2307/1912791

Hadfield, J. D. (2015). Increasing the efficiency of MCMC for hierarchical phylogenetic models of categorical traits using reduced mixed models. *Methods in Ecology and Evolution*, *6*, 706–714. https://doi.org/10.1111/2041-210x.12354

Hansen, T. F., & Martins, E. P. (1996). Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. *Evolution*, *50*, 1404–1417. https://doi.org/10.1111/j.1558-5646.1996.tb03914.x

Harvey, A. C. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press.

Harvey, P. H., & Pagel, M. D. (1991). *The comparative method in evolutionary biology*. Oxford University Press.

Ives, A. R. (2018a). Informative irreproducibility and the use of experiments in ecology. *Bioscience*, *68*, 746–747. https://doi.org/10.1093/biosci/biy090

Ives, A. R. (2018b). Mixed and phylogenetic models: A conceptual introduction to correlated data. leanpub.com.

Ives, A. R. (2019). R2s for correlated data: Phylogenetic models, LMMs, and GLMMs. *Systematic Biology*, *68*, 234–251. https://doi.org/10.1093/sysbio/syy060

Ives, A. R. (2022). Code and data for Ives (2022) Random errors are neither: On the interpretation of correlated data. *figshare*. Software. https://doi.org/10.6084/m9.figshare.20407158.v1

Ives, A. R., & Garland, T. (2010). Phylogenetic logistic regression for binary dependent variables. *Systematic Biology*, *59*, 9–26. https://doi.org/10.1093/sysbio/syp074

Ives, A. R., & Zhu, J. (2006). Statistics for correlated data: Phylogenies, space, and time. *Ecological Applications*, *16*, 20–32. https://doi.org/10.1890/04-0702

Ives, A. R., Zhu, L., Wang, F., Zhu, J., Morrow, C. J., & Radeloff, V. C. (2021). Statistical inference for trends in spatiotemporal data. *Remote Sensing of Environment*, *266*, 112678. https://www.sciencedirect.com/science/article/pii/S0034425721003989

Jarman, P. J. (1974). The social organisation of antelope in relation to their ecology. *Behaviour*, *48*, 215–267. https://doi.org/10.1163/156853974X00345

Judge, G. G., Griffiths, W. E., Hill, R. C., Lutkepohl, H., & Lee, T.-C. (1985). *The theory and practice of econometrics* (2nd ed.). John Wiley and Sons.

Krebs, C. J., Boonstra, R., & Boutin, S. (2017). Using experimentation to understand the 10-year snowshoe hare cycle in the boreal forest of North America. *Journal of Animal Ecology*, *87*, 87–100. https://doi.org/10.1111/1365-2656.12720

Kvalseth, T. O. (1985). Cautionary note about R2. *American Statistician*, *39*, 279–285. https://doi.org/10.2307/2683704

Legendre, P. (1993). Spatial autocorrelation: Trouble or new paradigm? *Ecology*, *74*, 1659–1673. https://doi.org/10.2307/1939924

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman and Hall.

Miller, H. J. (2004). Tobler's First Law and spatial analysis. *Annals of the Association of American Geographers*, *94*, 284–289. http://www.jstor.org/stable/3693985

Neter, J., Wasserman, W., & Kutner, M. H. (1989). *Applied linear regression models*. Richard D. Irwin.

Odum, E. P., & Barrett, G. W. (1971). *Fundamentals of ecology*. Saunders.

Pagel, M. (1997). Inferring evolutionary processes from phylogenies. *Zoologica Scripta*, *26*, 331–348. https://doi.org/10.1111/j.1463-6409.1997.tb00423.x

Pearl, J. (2010). An introduction to causal inference. *The International Journal of Biostatistics*, *6*, 7. https://doi.org/10.2202/1557-4679.1203

Petersen, K. B., & Pedersen, M. S. (2012). The matrix cookbook. http://matrixcookbook.com

Pielou, E. C. (1966). The measurement of diversity in different types of biological collection. *Journal of Theoretical Biology*, *13*, 131–144. https://doi.org/10.1016/0022-5193(66)90013-0

Pielou, E. C. (1975). *Ecological diversity*. John Wiley and Sons.

Pielou, E. C. (1977). *Mathematical ecology* (2nd ed.). John Wiley & Sons.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. Springer.

Prosperi, M., Guo, Y., Sperrin, M., Koopman, J. S., Min, J. S., He, X., Rich, S., Wang, M., Buchan, I. E., & Bian, J. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, *2*, 369–375. https://doi.org/10.1038/s42256-020-0197-y

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

Simberloff, D., Sanders, N., & Peres-Neto, P. (2017). A Homage to EC Pielou: One of the 20th Century's Most Accomplished Scientists. In I. Stephenson (Ed.), *Methods blog, methods in ecology and evolution*. Elsevier.

Stark, J., Broomhead, D. S., Davies, M. E., & Huke, J. (2003). Delay embeddings for forced systems. II. Stochastic forcing. *Journal of Nonlinear Science*, *13*, 519–577. https://doi.org/10.1007/s00332-003-0534-4

Takens, F. (1981). Detecting strange attractors in turbulence. In D. A. Rand & L. S. Young (Eds.), *Dynamical systems and turbulence* (pp. 366–381). Springer-Verlag.

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, *46*, 234–240. https://doi.org/10.2307/143141

Wikle, C. K., Zammit-Mangion, A., & Cressie, N. (2019). *Spatio-temporal statistics with R*. Taylor and Francis Group.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ives, A. R. (2022). Random errors are neither: On the interpretation of correlated data. *Methods in Ecology and Evolution*, *00*, 1–14. https://doi.org/10.1111/2041-210X.13971