Journal of Helminthology

www.cambridge.org/jhl

Can't see the forest for the trees? Statistical considerations for disease macroecology

A. Filion[1,†] and J.-F. Doherty[2,3,†]

[1]New Zealand Department of Conservation, Dunedin, New Zealand; [2]Department of Zoology, The University of British Columbia, Vancouver, Canada and [3]Michael Smith Laboratories, Department of Biochemistry & Molecular Biology, The University of British, Columbia,Vancouver, Canada

Commentary

Cite this article: Filion A and Doherty J-F (2026). Can't see the forest for the trees? Statistical considerations for disease macroecology. *Journal of Helminthology*, **100**, e109, 1–9
https://doi.org/10.1017/S0022149X26101874

Received: 27 May 2026 UTC
Revised: 21 June 2026 UTC
Accepted: 22 June 2026 UTC

Keywords:
host–parasite networks; zoonotic spillover; pathogen distribution modelling; ecological big data; disease risk mapping

Corresponding author:
J.-F. Doherty;
Email: jeff-doherty@hotmail.com

[†]Both authors contributed equally to this paper.

Abstract

Disease macroecology relies on large, complex datasets to understand the biotic and abiotic factors shaping parasite distributions and emerging infectious disease risk. These datasets span local to global host–parasite interactions and often integrate diverse host or parasite traits across evolutionary histories. Selecting appropriate statistical approaches requires first asking key questions about the study system: How well is the system understood? How important is predictor accuracy? How much bias is present in the data? We outline three broad analytical pathways commonly used in disease macroecology (frequentist models, Bayesian approaches, and machine learning) and link them to typical data contexts, providing R coding examples using current packages. We emphasise that these pathways are intended as flexible guidelines rather than prescriptive frameworks. Although more advanced approaches have grown in popularity over the last three decades, traditional significance tests remain widely used. While these tests remain appropriate for well-controlled or small-scale experimental systems, they are often unsuitable for complex ecological datasets, where violations of assumptions, hierarchical structure, and sampling biases can lead to unstable or difficult-to-reproduce inferences. We advocate for reporting biologically meaningful predictors, effect sizes, measures of uncertainty, and transparent analytical workflows. By explicitly linking common data challenges in disease macroecology to their consequences for model choice and inference, we encourage more reproducible, transparent, and context-appropriate statistical practices for understanding and forecasting parasite distributions and emerging disease risks.

Introduction

The typical dataset produced for the ecological study of infectious diseases is inherently complex: it can include local, regional, or global patterns of host–parasite interactions across large historical or evolutionary temporal scales (Barrile et al. 2023; Filion et al. 2020; Mahon et al. 2024; Pedersen and Davies 2009). These big data can be integrated into diverse statistical approaches to predict disease risk or identify the drivers of parasite diversity and distribution in single- or multi-host systems (Buhnerkempe et al. 2015; Doherty et al. 2021; Stephens et al. 2016). Statistical approaches in disease ecology are often grouped into categories such as classical difference-finding tests (Student's *t*-test, various ANOVA tests, chi-squared test, correlation tests, etc.), regression-based models in frequentist and Bayesian frameworks (linear regressions, generalised linear models, multilevel or hierarchical models, etc.), and machine-learning methods (random forests, gradient boosting, logistic bagging, etc.). However, these distinctions can obscure the fact that many commonly used techniques, such as *t*-tests, ANOVA, and linear regression, are part of a unified modelling framework. A more useful distinction lies in their inferential goals and their capacity to accommodate key data features, including hierarchical structure, nonlinearity, and uncertainty. These differences ultimately determine their suitability for complex ecological datasets. Many of the statistical challenges we highlight, such as pseudoreplication, hierarchical structure, and the limitations of null-hypothesis significance testing, have been recognised in ecology for decades (Hurlbert 1984). However, the rapid expansion of large, heterogeneous datasets in disease macroecology introduces new combinations of these challenges, including overlapping spatial, phylogenetic, and sampling biases (Doherty et al. 2021; Filion et al. 2022). As a result, existing general statistical guidance is not always sufficient to inform model choice in these contexts.

While a number of these tests can adequately address the general questions asked in disease ecology, others are simply not suited to account for the large-scale and intricate nature of big data and unravel the complex patterns that are fundamental to host–parasite interactions (virulence, host immunity, phylogeny, etc.) (Acevedo et al. 2019; Fountain-Jones et al. 2018; Lymbery et al. 2014). Indeed, there has been a push by statisticians in recent decades for researchers to reassess how they analyse data, emphasising a more thoughtful and practical approach when testing the relationship between variables, e.g., reassessing the reliance on *p*-value–centric significance testing (Leek and Peng 2015; Nakagawa and Cuthill 2007; Sterne and Smith 2001; Wasserstein,

Schirm & Lazar 2019). These considerations prioritise the inclusion of meaningful explanatory variables and the uncertainty surrounding predictions, which can become critical when testing for patterns and projecting risks of emerging zoonotic infectious diseases, which is further complicated by the fact that most disease datasets are themselves sparse and incomplete. Difference-finding tests remain useful in controlled experimental contexts where assumptions are met and the goal is to test specific, well-defined hypotheses, but their applicability becomes more limited as ecological datasets increase in complexity. Importantly, concerns about irreproducibility are often driven not by the statistical framework itself, but by analytical practices such as selective reporting, multiple testing, and insufficient documentation of modelling decisions.

Despite the increasing importance and availability of more robust and adaptable statistical methods like Bayesian modelling and machine learning for complex ecological data (Ellison 2004; Peters et al. 2014; Wilson et al. 2024), simpler and less flexible difference-finding tests still hold a firm place in disease ecology (see Box 1). This may reflect the poor adoption of more modern approaches in the basic training of biostatistics in higher education curricula. For instance, studies in the UK and USA both found that undergraduate and graduate students misuse or even lack the knowledge to run rudimentary statistical analyses commonly used by researchers in biological or life sciences (A'Brook and Weyers 1996; National Research Council 2003; Metz 2008; Weissgerber et al. 2016). This lack of statistical training stagnates the advancement and learning of proper inference testing (Carver et al. 2016; Maurer et al. 2019). Moreover, researchers and mentors tend to stick to the same analyses throughout their careers for various reasons, which could influence the methodological approaches of their students, further impeding discovery and statistical innovations (Smaldino and McElreath 2016; Ziliak and McCloskey 2008). If this were the case, how then can we expect students and early-career researchers to adopt suitable statistical methods for the often complex datasets found in disease ecology? We emphasise that our framework does not distinguish between statistical models themselves, e.g., GLMMs, but rather between different inferential goals and analytical workflows; in many cases, frequentist and Bayesian approaches share similar model structures but differ in how inference is conducted.

Box 1.
Here, we provide a broad overview of the statistical methods used by disease ecologists since the term was popularised in the mid 1990s (Grenfell and Dobson 1995). We searched in Web of Science for all publication records ranging from 1995 to 2024 inclusively using a suite of search words representing the four most common general types of statistical tests that exist in the wider field of ecology. We found a total of 7,483 publications matching our search criteria (see Box 1 Methods in Supplementary Material for our search words), with most publications using some type of difference-finding test (2,407 publications), followed closely by frequentist modelling (2,372 publications) and Bayesian modelling (2,234 publications), with machine learning methods being the least used (470 publications). We assumed that researchers would include statistical methods at least in the abstract in approximately equal proportion across the four types. Frequentist and Bayesian modelling caught up to difference-finding tests around the year 2007, with roughly equal proportions split between the three from that year onward. The use of machine learning shows a relatively slower pace of adoption compared to other established methods, before becoming more widely accepted (Box 1, Figure 3). Worryingly, and given the increasing calls from many groups to move away from traditional *p*-value significance tests, we also find that the number of publications using more traditional difference-finding tests has remained relatively steady at around 25% of publications since 2008; this type was as commonly used as frequentist and Bayesian modelling in 2024. Encouragingly though, the increasing use of machine learning since around 2017 appears to be partly making up for a slight decrease in Bayesian modelling.

Choosing a proper statistical analysis can easily become a daunting task if we do not first take the time to ask the right questions. Before getting bogged down in data purgatory and statistical significance fishing expeditions, we suggest some key considerations when exploring data and before running actual analyses. We walk you through some of the main types of modelling approaches useful for disease ecology, and provide a basic decision tree to help choose between these different strategies based on three simple questions (Figure 1): How much do you know about the system? How much do you care about the accuracy of your predictors? How much bias in the data are you working with? Based on your answers, we suggest statistical approaches that work well for different ecological data contexts and provide examples of disease-ecology studies and R coding snippets to help readers start their own analyses. We emphasise that spatial scale is used here illustratively and that model choice should ultimately depend on data structure, research aims, and available information rather than scale per se. The following sections present three example contexts (local, regional, and global) to demonstrate how analytical strategies may vary with data characteristics, not as rigid prescriptions. Having gone through our own journeys of self-motivated education in biostatistics, we hope that this perspective will serve as a useful guide for those navigating through the complex and constantly evolving landscape of statistical inference.

Diagnosing common analytical mismatches

To move beyond general recommendations, we highlight common mismatches between data structure and analytical approach in disease macroecology. These mismatches often arise when the complexity of the data is not aligned with the assumptions or capabilities of the chosen model. First, ignoring hierarchical structure, such as repeated sampling across sites, hosts, or time, can lead to pseudoreplication, inflated Type I error rates, and overconfident parameter estimates. Second, applying overly flexible or data-intensive models to small or well-controlled datasets can result in overfitting and reduced interpretability, without improving inference. Third, using predictive models on spatially biased or unevenly sampled datasets can produce spurious correlations and limit the transferability of results across regions or time periods. Identifying these mismatches prior to analysis provides a practical framework for selecting appropriate statistical tools and improving the robustness of ecological inference.

Analytical pathways in disease ecology

Disease ecologists looking for answers to some of the most complicated scientific questions in their field have access to an expanding range of statistical tools, increasing the odds of applying a suboptimal test for the system they are investigating. Here, we provide general recommendations for parasitologists to apply to their own system, with readily executable R scripts for guidance (see Supplementary Material). These R scripts include simple instructions that are detailed enough to understand the first main steps for each analysis, including data exploration and generic modelling. We will showcase three different scenarios that represent, at least from our point of view, the main analytical pathways in disease ecology. Throughout the text, we use pathogens (viruses, bacteria, fungi, and protists) and parasites (helminths and other metazoans)

https://doi.org/10.1017/S0022149X26101874

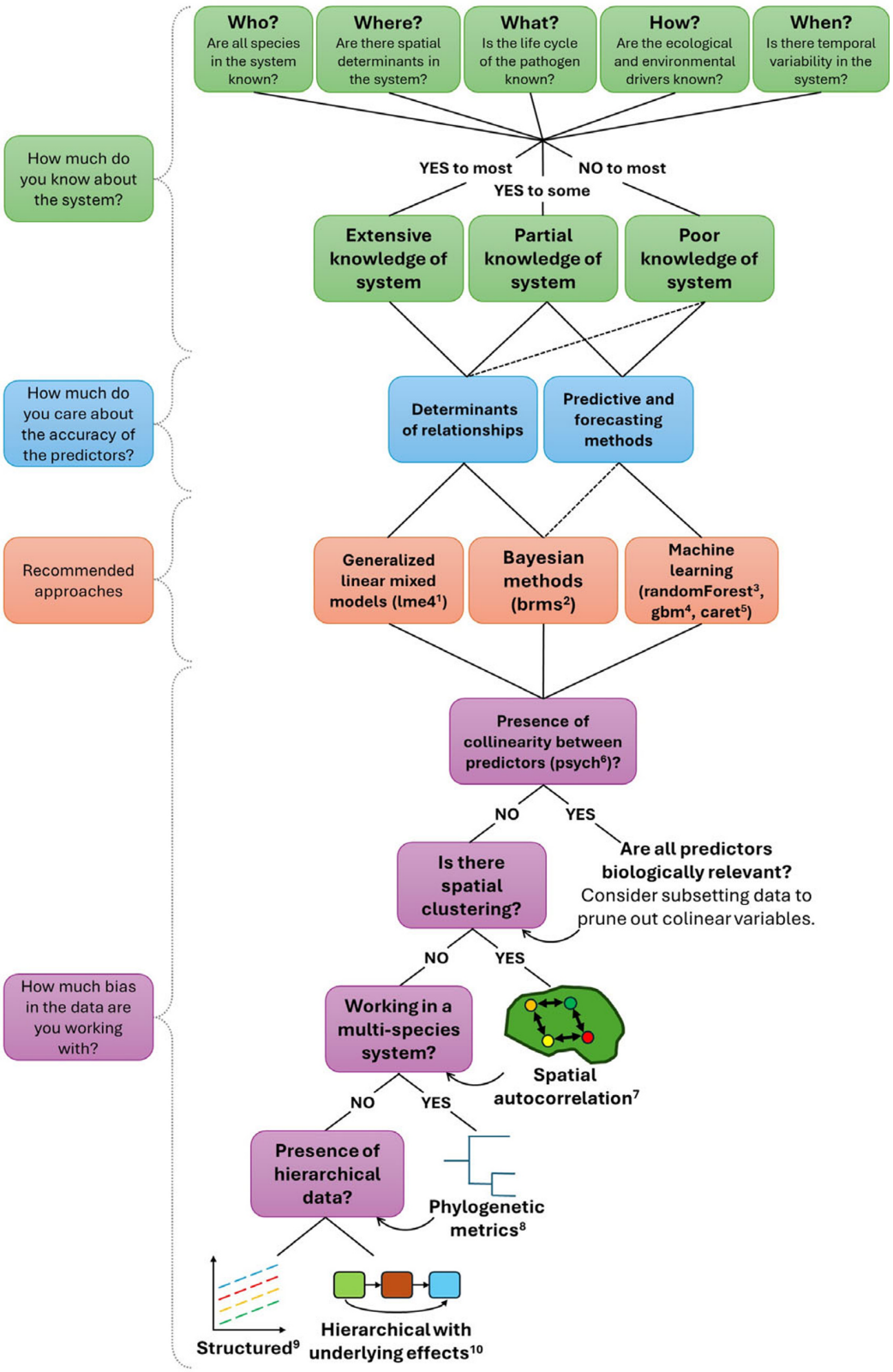

Figure 1. Decision tree based on three important questions (coded in green, blue, and purple) useful for analysing data in parasite or disease ecology. The first decision is based on five overarching questions applicable to any host–parasite system: Who? (which species are investigated); Where? (what spatial patterns are there); What? (what are the hosts and the life cycle of the pathogen); How? (what determines pathogen transmission); When? (what is the seasonal or temporal patterns of the pathogen). Full arrows represent recommended pathways of analysis, whereas dashed arrows represent atypical approaches, which may depend on the system investigated. Superscripted numbers refer to R packages or tutorials available in the resources section.

https://doi.org/10.1017/S0022149X26101874

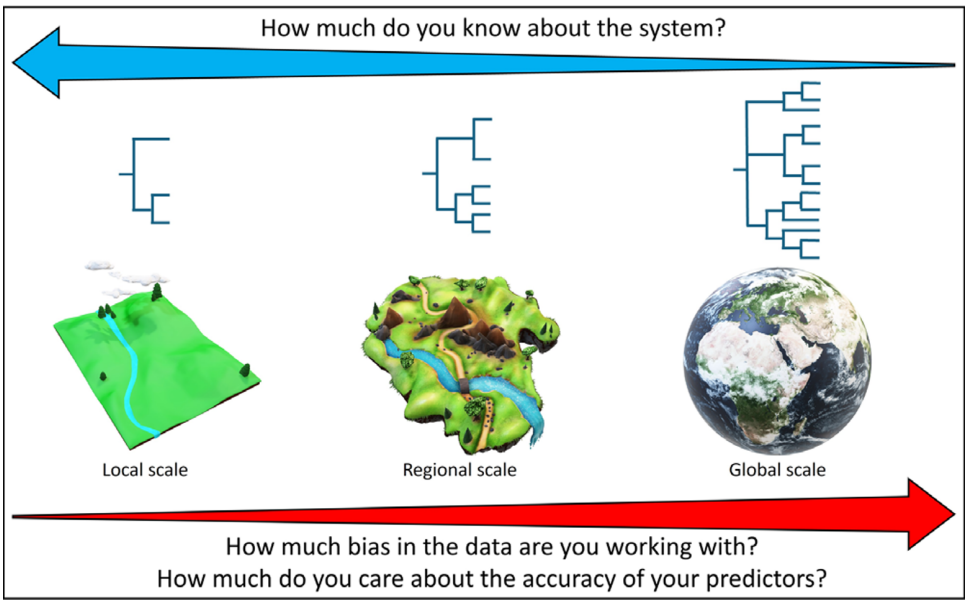

Figure 2. Scale of study system and how it generally impacts the knowledge of the system, and the inherent biases that can be found in the data. These will also impact how much researchers should care about the accuracy of the model predictors. Note that host and parasite phylogenetic patterns will also increase in scale and in complexity with an increase in geographical scale.

interchangeably, as both can cause disease, and both are studied in disease or parasite ecology. We emphasise that the following examples use spatial scale as a heuristic rather than a rule; the goal is to illustrate how analytical strategies often correspond to, but are not determined by, differences in data complexity, sample structure, and available information. Generally speaking, as we transition from a relatively small-scale local system to a large-scale global one (Fritsch et al. 2020), we naturally lose resolution and the ability to understand what drives parasite prevalence (How much do you know about the system?) (Figure 2). Because of this loss of resolution, researchers typically care more about accurate predictions of disease emergence rather than how a system naturally operates (How much do you care about the accuracy of your predictors?). This is especially true if there is an urgent need for disease risk mapping, e.g., predicting emergence of a novel infectious zoonotic disease. In turn, uncertainty typically increases and precision decreases with geographical scale, and varying sampling intensity can introduce different forms of bias, which limits the kinds of statistical tests that ought to be used (How much bias in the data are you working with?). Below, we highlight these biases and suggest common statistical approaches depending on the data context of the study system (Figure 1), while emphasising that alternative frameworks, e.g., GAMs, SEMs, or hierarchical Bayesian models, may be equally appropriate depending on data availability and research aims.

Example 1: local-scale systems

Local-scale systems refer to studies confined to a single landscape, population, or ecosystem where environmental variation and host diversity are limited. These systems are ideal for answering system-specific questions or testing ecological and evolutionary theory due to their relative simplicity and because most potential sources of bias can be explicitly measured or controlled. Applying an extremely powerful statistical method requiring complex model building or some sort of stochastic element (machine learning) will probably achieve the wrong goal (predicting instead of understanding), be time-consuming, and is akin to killing a fly with a bazooka, i.e., overkill. In most cases, whether investigating the ecological drivers of a parasite or trying to get a deeper understand of mechanisms leading to higher abundance or prevalence of a given disease-causing agent, a generalised linear mixed-effects model (GLMM) will be appropriate. This can be done with the 'glmer' function in the *lme4* package (Bates et al. 2015). These frequentist models are very useful to provide a clear understanding of the potential drivers of ecological phenomena while also allowing to account for biasing covariates and hierarchical data, such as controlling for a known pathogen covariate, e.g., body mass, and random intercepts for sampling site or host species and, when justified, random slopes for temporal or environmental covariates (Burnham and Anderson 2002; Zuur, Ieno and Elphick 2010) (see Part 1 in Supplementary Material). A critical diagnostic step at this scale is assessing whether observations are independent. When repeated measurements occur across sites, hosts, or time, failing to account for grouping structure can result in pseudoreplication and inflated statistical significance. In such cases, mixed-effects models provide a straightforward and widely accepted correction. If no such grouping exists, a standard generalised linear model (GLM) is sufficient. This type of analysis requires *a priori* decisions when building your model, i.e., which variables to include based on your knowledge of the system. Model selection using information criteria (AIC, WAIC, or cross-validation) remains essential for identifying parsimonious models and avoiding over-fitting. As these methods are relatively well known and have steadily been used for the past 30 years or so (see Box 1), we will not discuss them extensively here (for more information, see Zuur et al. 2009). Alternative frameworks such as generalised additive models (GAMs) or structural-equation models (SEMs) may be equally useful depending on the dataset and research objectives. GAMs are particularly effective when relationships between predictors and response variables are nonlinear or involve smooth gradients across

https://doi.org/10.1017/S0022149X26101874

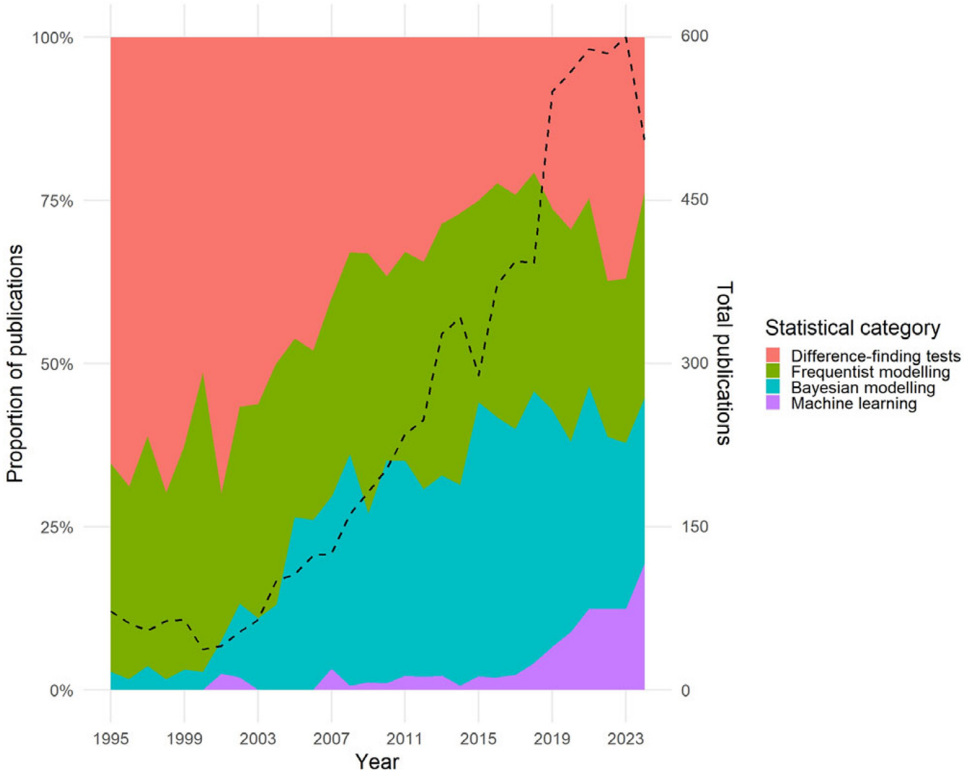

Figure 3. (Box 1): Stacked area chart showing the yearly proportion of publications of the four general statistical approaches common in disease ecology. The second y-axis represents the total number of publications per year for all types (dashed line).

space or time, while SEMs can explicitly represent causal hypotheses through linked equations and are valuable when indirect or mediating effects between ecological variables are of interest.

Example 2: regional-scale systems

Regional-scale systems encompass multiple populations or landscapes within a biogeographic region, where spatial heterogeneity and inter-community interactions become more influential. Transitioning from local to regional systems, thereby crossing landscape boundaries, naturally increases the complexity of the investigated data, but also the number of biases present in the system. For instance, increasing the spatial scale also increases the interactions among different host and parasite species; this adds an additional layer of complexity as local-scale drivers tend to dilute at large spatial scales (Cohen et al. 2016). Another key challenge at this scale is distinguishing true ecological patterns from artefacts of sampling effort. Uneven sampling across regions can create apparent spatial trends that reflect observation bias rather than biological processes. Incorporating hierarchical structure, spatial random effects, or explicitly modelling sampling effort can help mitigate these issues. Moreover, interactions between different communities or metapopulations will add a significant amount of complexity to any disease system, exponentially increasing the number of biases when trying to disentangle disease patterns and spread. For example, how can you be sure that the abundance of an investigated parasite is not driven by an increase in the biodiversity of alternative hosts present in the wider community, potentially diluting or boosting parasite host range? Even though statistical methods allow for the inclusion of such phylogenetic interactions since the early 1990s (Harvey and Pagel 1991), the adoption rate of these methods among disease ecologists remains slow (Filion et al. 2022). Other metrics, such as connectivity between landscape patches and habitat density (Correa Ayram et al. 2016), must also be considered. As such, an approach incorporating more powerful and adaptable statistics is needed. Luckily for disease ecologists, other types of models other than standard GLMs exist to overcome such issues. In this light, Bayesian methods provide key improvements, allowing researchers to consider community-wide interactions, more complex hierarchical data, host and parasite phylogenies, and more (Banner et al. 2020; Filion et al. 2022).

A principal advantage of Bayesian modelling is its capacity to incorporate prior information, allowing estimates to adjust according to existing knowledge of the system. From our point of view, the main differences between frequentist and Bayesian statistics are mostly philosophical, as Bayesian models report credible intervals instead of confidence intervals. As such, a credible interval would give the user a 95% probability that the true value of the predictors lies within this interval, as opposed to a confidence interval, which defines that in n number of repeated experiments, the confidence interval would contain the true predictor value 95% of the time (Kruschke 2010). However, these methods still require the user to have some *a priori* knowledge about the system they are investigating, as they need specific priors for each variable to run efficiently. While there are a few functions in the main R packages that offer a comprehensive Bayesian approach to most systems, e.g., 'get_prior' in the *brms* package (Bürkner 2017, 2018), the recommendations for priors are to account for the distribution family of the response variable and any random effects, and do not offer any

https://doi.org/10.1017/S0022149X26101874

consideration for the predictors. This is where having some knowledge of the system is very useful. For example, if investigating the effect of temperature on cercarial emergence in an intertidal snail or on mosquito emergence in a vector-borne disease system, you would expect that an increase in this variable would result in higher parasite abundance based on an *a priori* literature search and would therefore adjust the priors for this coefficient accordingly based on the expectation you have in your system. Without any kind of knowledge on the system, the model would default to 'flat' priors, thus offering no real advantage of switching to Bayesian from frequentist modelling. Another key improvement in these methods is that they are highly customisable, but also have a wide online community to support the improvement of novel methods[4] (see Part 2 in Supplementary Material). Like with frequentist approaches, model selection (BIC, LOO-CV, etc.) helps select the model that best balances fit with complexity. Many of the conveniences often attributed to Bayesian models, such as model flexibility or phylogenetic integration, actually stem from particular software implementations rather than from Bayesian inference itself. For instance, numerous tutorials that are constantly evolving allow users to include phylogenetic information[8] or create SEMs[11] using *brms*, facilitating the switch to these methods and improving the range of tools available to any end-user.

Example 3: global-scale systems

Finally, going from regional systems to even wider spatial scales (continental or near-global studies) tends to exponentially increase the complexity of the data and covariate structuring. For example, helminth diversity in anuran amphibians is primarily driven by the distance between sampling locations at global scales, whereas climatic effects overshadow spatial distance at regional scales (Martins et al. 2021). In vector-borne systems, avian influenza (H_5N_1) is known to be regionally driven by the presence of suitable habitats for waterfowl species like wetlands, poultry farming, and seasonal migration of waterfowl (Gilbert et al. 2006). However, when spatial scale was increased, the same authors found that regional drivers of avian influenza spread lost their significance and were overpowered by socioeconomic factors and the large migration patterns of waterfowl (Gilbert et al. 2008). Thus, a new category of biases, i.e., large-scale biases (socioeconomic factors, climate change, community stability, etc.), can become predominant at global scales, even though they may have little or no influence at local scales (de Angeli Dutra and Poulin 2024; Mora et al. 2022; Sokolow et al. 2022). Global-scale analyses can still yield strong mechanistic insights (e.g., Wilson et al. 2024, global *Toxoplasma* prevalence), demonstrating that large spatial coverage does not preclude inference about drivers. While disentangling and quantifying the relevant drivers of parasite abundance becomes more of a mission at increasing scales of geography, an advanced category of statistical tools exists to help researchers deal with this level of complex data. Having gone on our own fishing expeditions due to the large quantity of available databases at these scales (see Doherty et al. 2021), we posit that, in these systems, a careful hypothesis-driven approach should be prioritised. This is where machine learning approaches thrive, especially if one is considering more accurate forecasting of disease emergence patterns.

Most machine learning approaches tend to be robust to the inclusion of many predictor variables and are flexible when it comes to the distribution of the response variable (Elith, Leathwick and Hastie 2008). They also allow for nonlinear and threshold effects to be detected using marginal plots, a particularly useful tool when investigating large-scale data (Kuhn and Johnson 2013). Because of the stochastic nature of these approaches, model structuring is mostly determined by the data used, and thus few *a priori* decisions are generally required. As these methods are relatively recent, we include more detailed considerations below. We note that machine learning approaches are not limited to global or regional scales; they can also be effective in local contexts when dense temporal or surveillance datasets are available, e.g., seasonal influenza forecasting. Other frameworks such as hybrid Bayesian–machine-learning methods or SEMs may also serve similar predictive or explanatory roles, depending on data type and objectives. Despite their flexibility and predictive power, machine learning approaches are not universally appropriate. In disease macroecology, datasets are often spatially biased, sparse, or unevenly sampled, which can lead to overfitting and poor transferability of models across regions or time periods. Without careful validation, models may capture artefacts of sampling effort rather than true ecological processes, limiting their usefulness for inference and prediction. Despite their flexibility, machine-learning models are sensitive to data leakage, hyperparameter choices, and spatial autocorrelation. Without explicit reporting of these elements, results may vary substantially across repeated analyses, limiting reproducibility and comparability between studies.

To help curb limitations in machine learning inference, something that is particularly useful when it comes to machine learning is knowing how to create a repeatable model with your dataset. Most literature recommendations tend to suggest using about 2/3 of your database to train your model (hereafter called the training dataset), leaving about 1/3 of the data as a validating dataset (Kuhn and Johnson 2013). When choosing your training dataset, a careful and balanced approach should be considered, as it will further determine the robustness of your results. For instance, take any given disease with 1,000 records of presence/absence locations worldwide, with 950 being locations with no disease detection and 50 being locations with actual records. If the training dataset is imbalanced and includes a disproportionally high number of absence locations, the model's relevancy will decrease since the validating dataset missed a great deal of information, i.e., predicting bias towards absence data. This will subsequently lead to more problems as this newly trained model will return a very high accuracy (in this example, about 95%), because it will be able to predict most of the data 100% of the time! A good way to avoid such issues lies with careful planning and execution. First, remixing all the rows in your dataset before creating your training and validating datasets is a good way to make sure that there is no 'data clumping' (see Kuhn and Johnson 2013 for details on this method). Second, approaches that circumvent this sampling bias when creating the training dataset, such as oversampling rare instances or undersampling the majority group, will help achieve a balanced and accurate model. Third, and this is perhaps the most relevant for new machine learning users, there are built-in functions in R packages, such as the 'sampling' function in the *caret* package (Kuhn 2008), that can create a balanced sampling design between response variable categories. However, as with any approach, there are a few downsides with machine learning methods. For instance, they tend to behave erratically when presented with a suite of collinear variables, giving weight to only one predictor while systematically rejecting all others, with no repeatable patterns between model runs (Elith, Leathwick and Hastie 2008).

In addition, machine learning methods tend to be data hungry, requiring a vast amount of data to run efficiently (Domingos 2012). Another key limitation of machine learning approaches is their

https://doi.org/10.1017/S0022149X26101874

susceptibility to overfitting, particularly when predictors are numerous relative to sample size or when data are spatially biased, a common feature of global disease datasets. In such cases, models may perform well on training data but fail to generalise to new regions or time periods, limiting their usefulness for forecasting. Often, this happens when the model has too many predictors and tries to analyse the noise of the data instead. A simple workaround is to start with a simpler model with few predictors, and build up from there (Domingos 2012). We provide detailed examples of machine learning using a few of the most popular R packages for ecologists (Part 3 in Supplementary Material), such as *caret*, *gbm* (Ridgeway et al. 2024), and *randomForest* (Liaw and Wiener 2002). Ensuring model transferability, i.e., the ability to make reliable predictions beyond the training domain, should therefore be a central consideration when applying these methods, particularly in global disease risk mapping. While simple train/test splits are widely used, alternative approaches such as k-fold cross-validation or spatial blocking may provide more robust estimates of model performance, particularly in spatially structured ecological datasets where observations are not independent.

Other considerations

At this point, the reader will have noticed that we have barely mentioned anything about difference-finding tests (ANOVAs, *t*-tests, and others). This is because we caution against an overreliance on *p*-value–centric significance testing in complex ecological datasets, particularly when model assumptions are not met or when analytical decisions are not transparently reported (Leek and Peng 2015; Nakagawa and Cuthill 2007; Sterne and Smith 2001; Wasserstein, Schirm and Lazar 2019). These tests can have inflexible assumptions about the nature of the data they can analyse, which limits their applicability when dealing with hierarchical structures, random effects, phylogenetic signals, or substantial data biases. From our point of view, assessing differences between groups from *p*-values alone as opposed to calculating relevant effect sizes retains little biological value when comparing patterns across studies, something that could become critical with the increasing rate of emerging infectious diseases worldwide (Baker et al. 2022). That said, difference-finding tests remain useful in small-scale, controlled experiments or well-balanced study designs where assumptions are met, and the goal is to test specific hypotheses in controlled settings, e.g., lab-based experiments. A concise overview of the strengths, limitations, and typical use cases of each analytical approach is provided in Table 1 to help readers select appropriate methods.

We strongly recommend that researchers present the underlying code that generated their results in such a way that is fully transparent, easily understood with detailed instructions included in the script, and easily reproduced with the data provided. Like the R scripts that were provided in this paper, readers should be able to download the scripts and the data from a paper and, within minutes to hours (pending proper package installation and the type of analysis), obtain the same major results as in the paper (within the same margin of error if working with Bayesian or machine learning methods). We also encourage authors to include examples of residual and predictive checks in their code, e.g., plotting residuals vs. fitted values, visualising posterior predictions, or comparing observed vs. predicted data, to verify model fit and performance.

Within the past few years, we have also seen an exponential growth of specialised machine learning models known as large language models (LLMs) such as OpenAI's GPT-5[12] and Meta's Llama[13]. These LLMs can be useful to help write scripts for statistical analyses in R or Python, as they are capable of synthesising innumerable quantities of scripts and queries from the internet (although, see Chen et al. 2021). These LLMs are becoming useful and powerful tools to help build or improve R and Python scripts using simple prompts for the purpose of statistical tests in practically any area of research. Here, we urge caution upon those who are just beginning to learn how to write their own scripts. Regardless of your needs, we strongly encourage that you have a solid foundation in R or Python coding before copy-pasting and running codes generated solely by LLMs. In our experience, this will undoubtedly save a lot of headaches and time as these prompt-generated codes are often imperfect and can hide subtle mistakes if you are not familiar with script writing. Moreover, because hallucination rates remain disproportionately high in most current LLMs, blindly trusting an LLM-generated script without domain knowledge can easily lead to biased results. In short, just because it is convenient does not mean it is reliable. Learning basic coding from a good guide in data science (e.g., Wickham et al. 2023) will set you on the right path. The recommendations presented here are most relevant to ecological and disease-ecological studies, where hierarchical sampling and natural variability are common, rather than to tightly controlled experimental systems.

Table 1. Summary of common statistical approaches used in disease macroecology

Approach	Strengths	Limitations	Typical use cases
Difference-finding tests (e.g., t-tests, ANOVAs)	Simple implementation; widely taught; suitable for small, controlled experiments	Inflexible assumptions; poor handling of hierarchical or unbalanced data; cannot incorporate random effects or phylogeny	Preliminary comparisons, controlled laboratory or mesocosm studies, and balanced datasets with minimal hierarchical structure
Frequentist models (e.g., GLMs, GLMMs)	Transparent inference; interpretable coefficients; broad software support	Limited flexibility for complex dependencies; assumes fixed parameters	Local- to regional-scale ecological datasets with moderate complexity
Bayesian models (e.g., hierarchical, phylogenetic)	Incorporate prior knowledge; quantify full uncertainty; handle hierarchical data	Computationally intensive; require prior specification; slower to converge	Regional- to continental-scale analyses, multilevel or data-sparse systems
Machine-learning models (e.g., random forests, boosted trees, neural networks)	Handle large nonlinear datasets; strong predictive accuracy; minimal distributional assumptions	Limited interpretability; risk of overfitting; high computational demand	Predictive mapping, global-scale risk assessment, high-volume or remote-sensing data

https://doi.org/10.1017/S0022149X26101874

Future perspectives

Disease macroecology is an ever-expanding field, and with increasing amounts of ecological, environmental, and molecular data becoming available, these illustrative analytical frameworks will continue to evolve. The purpose of this perspective was to provide simple examples of how data characteristics, scale, and research objectives can guide appropriate statistical approaches without implying that a single 'best model' exists for any given context. Rather than advocating for any single class of models, our goal is to align analytical choices with data structure, inference goals, and the limitations inherent to ecological datasets. Future methodological developments should focus on flexibility, combining interpretability from classical models with the predictive capacity of machine learning, and on transparent model validation so that results remain reproducible across platforms and data sources.

We urge disease ecologists to move away from difference-finding tests and stop relying solely on *p*-values. Too often, when preparing comparative or meta-analyses, we find ourselves having to reject potentially useful studies or datapoints from our dataset because of poorly reported results. We suggest that disease ecologists report the predictive power of each model predictor, i.e., effect sizes and confidence or credible intervals (for machine learning, the relative variable influence). In addition, authors should include sufficient information on model structure, diagnostics, and data-processing steps so that analyses can be independently verified and reproduced. With the current rate of emerging infectious diseases worldwide, we urge disease ecologists to adopt a standardised statistical protocol when publishing their data and their scripts, and to keep innovating in the methods and tools they use in an open and transparent way. Continued integration of flexible analytical frameworks, combining interpretability from classical models with the predictive capacity of modern computational methods, will further strengthen the field. We strongly believe that only then will researchers begin to fully grasp the drivers and patterns of parasite occurrence to build more accurate guidelines and better forecast future emergence of disease.

Rather than reiterating general statistical principles, our goal is to contextualise them within the specific challenges of disease macroecology, where scale, bias, and data heterogeneity interact in ways that complicate standard analytical workflows. By explicitly linking these data characteristics to their statistical consequences, we provide a practical framework to support more transparent, reproducible, and context-dependent analytical decisions.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0022149X26101874.

Availability of data and material. Not applicable.

Code availability. Code is available in the online supplementary material.

Acknowledgements. We thank Patrick Stephens for his insight and for providing feedback on an early draft of the manuscript.

Author contribution. J-FD conceived the idea, and both authors designed the paper; both authors collected and analysed the data and wrote the manuscript. Both authors contributed critically to the drafts and gave final approval for publication.

Financial support. J-FD was supported by a Natural Sciences and Engineering Research Council of Canada Postdoctoral Fellowship (PDF-578318-2023).

Competing interests. Not applicable.

Ethics approval. Not applicable.

Consent to participate. Not applicable.

Consent for publication. Not applicable.

Resources.

1. https://www.guru99.com/r-random-forest-tutorial.html
2. https://uc-r.github.io/gbm_regression
3. https://cran.r-project.org/web/packages/caret/vignettes/caret.html
4. https://discourse.mc-stan.org/c/interfaces/brms/36
5. https://www.rensvandeschoot.com/tutorials/lme4/
6. https://benwhalley.github.io/just-enough-r/correlations.html
7. https://mgimond.github.io/Spatial/spatial-autocorrelation-in-r.html
8. https://cran.r-project.org/web/packages/brms/vignettes/brms_phylogenetics.html
9. https://www.highstat.com/index.php/books2?view=article&id=17&catid=18
10. https://stats.oarc.ucla.edu/r/seminars/rsem/
11. https://rpubs.com/jebyrnes/brms_bayes_sem
12. https://openai.com/index/introducing-gpt-5/
13. https://llama.meta.com/

References

A'Brook R and Weyers JDB (1996) Teaching of statistics to UK undergraduate biology students in 1995. *Journal of Biological Education* **30**, 281–288.

Acevedo MA, Dillemuth FP, Flick AJ, Faldyn MJ and Elderd BD (2019) Virulence-driven trade-offs in disease transmission: a meta-analysis. *Evolution* **73**, 636–647.

Baker RE, Mahmud AS, Miller IF, Rajeev M, Rasambainarivo F, Rice BL, Takahashi S, Tatem AJ, Wagner CE, Wang L-F, Wesolowski A and Metcalf CJE (2022) Infectious disease in an era of global change. *Nature Reviews Microbiology* **20**, 193–205.

Banner KM, Irvine KM, Rodhouse TJ and O'Hara RB (2020) The use of Bayesian priors in ecology: the good, the bad and the not great. *Methods in Ecology and Evolution* **11**, 882–889.

Barrile GM, Augustine DJ, Porensky LM, Duchardt CJ, Shoemaker KT, Hartway CR, Derner JD, Hunter EA and Davidson AD (2023) A big data–model integration approach for predicting epizootics and population recovery in a keystone species. *Ecological Applications* **33**, e2827.

Bates D, Maechler M, Bolker BM and Walker SC (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67**, 1–48.

Buhnerkempe MG, Roberts MG, Dobson AP, Heesterbeek H, Hudson PJ and Lloyd-Smith JO (2015) Eight challenges in modelling disease ecology in multi-host, multi-agent systems. *Epidemics* **10**, 26–30.

Bürkner P-C (2017) brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* **80**, 1–28.

Bürkner P-C (2018) Advanced Bayesian multilevel modeling with the R package brms. *The R Journal* **10**, 395–411.

Burnham KP and Anderson DR (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edn. New York: Springer.

Carver R, Everson M, Gabrosek J, Horton N, Lock R, Mocko M, Rossman A, Rowell GH, Velleman P, Witmer J and Wood B (2016) *Guidelines for Assessment and Instruction in Statistics Education College Report 2016*. Alexandria, VA: American Statistical Association.

Chen M, Tworek J, Jun H, Yuan Q, Henrique Ponde de Oliveira P, Kaplan J, Edwards H, Burda Y, Nicholas J, Brockman G, Ray A, Puri R, Krueger G, Petrov M, Khlaaf H, Sastry G, Mishkin P, Chan B, Gray S, Ryder N, Pavlov M, Power A, Kaiser L, Bavarian M, Winter C, Tillet P, Such FP, Cummings D, Plappert M, Chantzis F, Barnes E, Herbert-Voss A, Guss WH, Nichol A, Paino A, Tezak N, Tang J, Babuschkin I, Balaji S, Jain S, Saunders W, Hesse C, Carr AN, Leike J, Achiam J, Misra V, Morikawa E, Radford A, Knight M, Brundage M, Murati M, Mayer K, Welinder P, McGrew B, Amodei D, McCandlish S, Sutskever I and Zaremba W (2021) Evaluating large language models trained on code. Cornell University Library, arXiv.org, Ithaca.

Cohen JM, Civitello DJ, Brace AJ, Feichtinger EM, Ortega CN, Richardson JC, Sauer EL, Liu X and Rohr JR (2016) Spatial scale modulates the strength of ecological processes driving disease distributions. *Proceedings of the National Academy of Sciences* **113**, E3359–E3364.

Correa Ayram CA, Mendoza ME, Etter A and Salicrup DRP (2016) Habitat connectivity in biodiversity conservation: a review of recent studies and applications. *Progress in Physical Geography* **40**, 7–37.

National Research Council (2003) *Bio 2010: Transforming Undergraduate Education for Future Research Biologists*. Washington, DC: National Academies Press.

de Angeli Dutra D and Poulin R (2024) Network specificity decreases community stability and competition among avian haemosporidian parasites and their hosts. *Global Ecology and Biogeography* **33**, e13831.

Doherty J-F, Chai X, Cope LE, de Angeli Dutra D, Milotic M, Ni S, Park E and Filion A (2021) The rise of big data in disease ecology. *Trends in Parasitology* **37**, 1034–1037.

Domingos P (2012) *A few useful things to know about machine learning*. pp. 78–87. Association for Computing Machinery, New York, NY.

Elith J, Leathwick JR and Hastie T (2008) A working guide to boosted regression trees. *The Journal of Animal Ecology* **77**, 802–813.

Ellison AM (2004) Bayesian inference in ecology. *Ecology Letters* **7**, 509–520.

Filion A, Doherty J-F, Poulin R and Godfrey SS (2022) Building a comprehensive phylogenetic framework in disease ecology. *Trends in Parasitology* **38**, 424–427.

Filion A, Eriksson A, Jorge F, Niebuhr CN and Poulin R (2020) Large-scale disease patterns explained by climatic seasonality and host traits. *Oecologia* **194**, 723–733.

Fountain-Jones NM, Pearse WD, Escobar LE, Alba-Casals A, Carver S, Davies TJ, Kraberger S, Papeş M, Vandegrift K, Worsley-Tonks K and Craft ME (2018) Towards an eco-phylogenetic framework for infectious disease ecology. *Biological Reviews* **93**, 950–970.

Fritsch M, Lischke H, Meyer KM and Murrell D (2020) Scaling methods in ecological modelling. *Methods in Ecology and Evolution* **11**, 1368–1378.

Gilbert M, Xiao X, Domenech J, Lubroth J, Martin V and Slingenbergh J (2006) Anatidae migration in the Western Palearctic and spread of highly pathogenic avian influenza H5N1 virus. *Emerging Infectious Diseases* **12**, 1650–1656.

Gilbert M, Xiao X, Pfeiffer DU, Epprecht M, Boles S, Czarnecki C, Chaitaweesub P, Kalpravidh W, Minh PQ, Otte MJ, Martin V and Slingenbergh J (2008) Mapping H5N1 highly pathogenic avian influenza risk in Southeast Asia. *Proceedings of the National Academy of Sciences* **105**, 4769–4774.

Grenfell BT and Dobson AP (1995) *Ecology of Infectious Diseases in Natural Populations*. Cambridge: Cambridge University Press.

Harvey PH and Pagel MD (1991) *The Comparative Method in Evolutionary Biology*. Oxford: Oxford University Press.

Hurlbert SH (1984) Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* **54**, 187–211.

Kruschke JK (2010) What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences* **14**, 293–300.

Kuhn M (2008) Building predictive models in R using the caret package. *Journal of Statistical Software* **28**, 1–26.

Kuhn M and Johnson K (2013) *Applied Predictive Modeling*. New York: Springer.

Leek JT and Peng RD (2015) Statistics: p values are just the tip of the iceberg. *Nature* **520**, 612.

Liaw A and Wiener M (2002) Classification and regression by randomForest. *The R Journal* **2/3**, 18–22.

Lymbery AJ, Morine M, Kanani HG, Beatty SJ and Morgan DL (2014) Co-invaders: the effects of alien parasites on native hosts. *International Journal for Parasitology: Parasites and Wildlife* **3**, 171–177.

Mahon MB, Sack A, Aleuy OA, Barbera C, Brown E, Buelow H, Civitello DJ, Cohen JM, de Wit LA, Forstchen M, Halliday FW, Heffernan P, Knutie SA, Korotasz A, Larson JG, Rumschlag SL, Selland E, Shepack A, Vincent N and Rohr JR (2024) A meta-analysis on global change drivers and the risk of infectious disease. *Nature* **629**, 830–836.

Maurer K, Hudiburgh L, Werwinski L and Bailer J (2019) Content audit for p-value principles in introductory statistics. *The American Statistician* **73**, 385–391.

Martins PM, Poulin R and Gonçalves-Souza T (2021) Drivers of parasite β-diversity among anuran hosts depend on scale, realm and parasite group. *Philosophical Transactions of the Royal Society B* **376**, 20200367.

Metz AM (2008) Teaching statistics in biology: using inquiry-based learning to strengthen understanding of statistical analysis in biology laboratory courses. *CBE: Life Sciences Education* **7**, 317–326.

Mora C, McKenzie T, Gaw IM, Dean JM, von Hammerstein H, Knudson TA, Setter RO, Smith CZ, Webster KM, Patz JA and Franklin EC (2022) Over half of known human pathogenic diseases can be aggravated by climate change. *Nature Climate Change* **12**, 869–875.

Nakagawa S and Cuthill IC (2007) Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews* **82**, 591–605.

Pedersen AB and Davies TJ (2009) Cross-species pathogen transmission and disease emergence in primates. *EcoHealth* **6**, 496–508.

Peters DPC, Havstad KM, Cushing J, Tweedie C, Fuentes O and Villanueva-Rosales N (2014) Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology. *Ecosphere* **5**, 1–15.

Ridgeway G, Edwards D, Kriegler B, Schroedl S, Southworth H, Greenwell B, Boehmke B and Cunningham J (2024) gbm: Generalized boosted regression models. *CRAN*, 1–39.

Smaldino PE and McElreath R (2016) The natural selection of bad science. *Royal Society Open Science* **3**.

Sokolow SH, Nova N, Jones IJ, Wood CL, Lafferty KD, Garchitorena A, Hopkins SR, Lund AJ, MacDonald AJ, LeBoa C, Peel AJ, Mordecai EA, Howard ME, Buck JC, Lopez-Carr D, Barry M, Bonds MH and De Leo GA (2022) Ecological and socioeconomic factors associated with the human burden of environmentally mediated pathogens: a global analysis. *The Lancet Planetary Health* **6**, e870–e879.

Stephens PR, Altizer S, Smith KF, Alonso Aguirre A, Brown JH, Budischak SA, Byers JE, Dallas TA, Jonathan Davies T, Drake JM, Ezenwa VO, Farrell MJ, Gittleman JL, Han BA, Huang S, Hutchinson RA, Johnson P, Nunn CL, Onstad D, Park A, Vazquez-Prokopec GM, Schmidt JP and Poulin R (2016) The macroecology of infectious diseases: a new perspective on global-scale drivers of pathogen distributions and impacts. *Ecology Letters* **19**, 1159–1171.

Sterne JAC and Smith GD (2001) Sifting the evidence—what's wrong with significance tests? *BMJ* **322**, 226–231.

Wasserstein RL, Schirm AL and Lazar NA (2019) Moving to a world beyond "p < 0.05". *The American Statistician* **73**, 1–19.

Weissgerber TL, Garovic VD, Milin-Lazovic JS, Winham SJ, Obradovic Z, Trzeciakowski JP and Milic NM (2016) Reinventing biostatistics education for basic scientists. *PLoS Biology* **14**, e1002430.

Wickham H, Çetinkaya-Rundel M and Grolemund G (2023) *R for Data Science*, 2nd edn. Sebastopol, CA: O'Reilly Media.

Wilson AG, Lapen DR, Provencher JF and Wilson S (2024) The role of species ecology in predicting *Toxoplasma gondii* prevalence in wild and domesticated mammals globally. *PLoS Pathogens* **20**, e1011908.

Ziliak ST and McCloskey DN (2008) *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor: University of Michigan Press.

Zuur AF, Ieno EN and Elphick CS (2010) A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution* **1**, 3–14.

Zuur AF, Ieno EN, Walker NJ, Saveliev AA and Smith GM (2009) *Mixed Effects Models and Extensions in Ecology with R*. New York: Springer.