Check for updates

Received: 23 January 2023 | Accepted: 13 November 2023

DOI: 10.1111/2041-210X.14270

Methods in Ecology and Evolution

REVIEW

Four principles for improved statistical ecology

Gordana Popovic[1] | Tanya Jane Mason[2,3] | Szymon Marian Drobniak[4,5] | Tiago André Marques[6,7] | Joanne Potts[8] | Rocío Joo[9] | Res Altwegg[10] | Carolyn Claire Isabelle Burns[11] | Michael Andrew McCarthy[12] | Alison Johnston[13] | Shinichi Nakagawa[4] | Louise McMillan[14] | Kadambari Devarajan[15,16] | Patrick Leo Taggart[17] | Alison Wunderlich[18] | Magdalena M. Mair[19,20] | Juan Andrés Martínez-Lanfranco[21] | Malgorzata Lagisz[4] | Patrice Pottier[4]

Correspondence
Gordana Popovic
Email: g.popovic@unsw.edu.au

Funding information
Australian Research Council DECRA, Grant/Award Number: DE180100202; Australian Research Council Discovery, Grant/Award Number: DP210100812; CEAUL, funded by FCT - Fundação para a Ciência e a Tecnologia, Portugal, Grant/Award Number: UIDB/00006/2020; Christiane Nüsslein-Volhard Foundation; Computational Biodiversity Science and Services Program (Bios2-Canada); Fundação de Amparo à Pesquisa do Estado de São Paulo, Grant/Award Number: 17/16650-5; National Agency of Research and Innovation (ANII-Uruguay); National Research Foundation of South Africa, Grant/Award Number: 114696; NSW Government through its Environmental Trust, Grant/Award Number: 2018/SSC/0049; Polish National Science Centre, Grant/Award Number: UMO-2020/39/B/NZ8/01274; UNSW Scientia Doctoral scholarship

Handling Editor: Nick Isaac

Abstract

1. Increasing attention has been drawn to the misuse of statistical methods over recent years, with particular concern about the prevalence of practices such as poor experimental design, cherry picking and inadequate reporting. These failures are largely unintentional and no more common in ecology than in other scientific disciplines, with many of them easily remedied given the right guidance.
2. Originating from a discussion at the 2020 International Statistical Ecology Conference, we show how ecologists can build their research following four guiding principles for impactful statistical research practices: (1) define a focussed research question, then plan sampling and analysis to answer it; (2) develop a model that accounts for the distribution and dependence of your data; (3) emphasise effect sizes to replace statistical significance with ecological relevance; and (4) report your methods and findings in sufficient detail so that your research is valid and reproducible.
3. These principles provide a framework for experimental design and reporting that guards against unsound practices. Starting with a well-defined research question allows researchers to create an efficient study to answer it, and guards against poor research practices that lead to poor estimation of the direction, magnitude, and uncertainty of ecological relationships, and to poor replicability. Correct and appropriate statistical models give sound conclusions. Good reporting practices and a focus on ecological relevance make results impactful and replicable.
4. Illustrated with two examples—an experiment to study the impact of disturbance on upland wetlands, and an observational study on blue tit colouring—this paper explains the rationale for the selection and use of effective statistical practices and provides practical guidance for ecologists seeking to improve their use of statistical methods.

For Affiliation refer page on 277.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2024 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society.

KEYWORDS
HARKing, model assumptions, p-hacking, pre-registration, *p*-values, questionable research practices, reproducibility crisis, research waste

1 | INTRODUCTION

When reporting research findings, ecologists, like other scientists, want their results to reflect what truly happens in the system being studied and to communicate both ecological relevance and the level of support for their conclusions. For their results to hold up, researchers need to follow good research practices. Failure to follow good practices has led to low reproducibility of findings in many fields (Begley & Ellis, 2012; Camerer et al., 2018; Open Science Collaboration, 2015). Poor research practices are also common in ecology (Anderson et al., 2000; Fidler et al., 2006; Fraser et al., 2018), and their use can distort findings, waste resources, inadequately report what is happening in ecological systems and ultimately have the potential to misrepresent research to other researchers, policymakers and the public.

Poor research practices stem partly from misunderstandings or misinterpretations of statistical methods and principles of study design. Some of the most common and consequential of these practices include:

- Hypothesising after results are known (HARKing; Kerr, 1998), a practice that 51% of ecologists and evolutionary biologists report engaging in (Fraser et al., 2018).
- Not reporting non-significant results; a form of cherry picking, which 64% of ecologists admitted to doing at least once (Fraser et al., 2018).
- Hypothesis testing based on a null hypothesis that is known a priori to be false; Anderson et al., 2000 found the vast majority (95%) of *Ecology* articles they evaluated contained null hypotheses that were likely known to be false a priori.
- Misinterpreting non-significant results as evidence of 'no effect' or 'no relationship', which happens approximately 63% of the time non-significant results are reported in ecology (Fidler et al., 2006).
- Providing insufficient detail on methods and analysis. Almost 80% of ecology papers fail to provide enough detail to be computationally reproducible (with most, 73%, failing to include accompanying analysis code; Culina et al., 2020).

When presented with this report card, ecologists may justify their research practices as being harmless or unavoidable. However, cherry picking and HARKing are known to lead to overconfidence in results, by shrinking *p*-values and confidence intervals. In the context of publication bias, where 'significant' results are far more likely to be published, this inflates the rate of false positives ('significant' results when there is no actual effect in the population, see for example Forstmeier et al., 2017). Unfortunately, we do not know the true rate of false positives in ecology, as replication studies are very uncommon, with only 0.023% of all studies published to date representing a true replication (Kelly, 2019). However, we do know that a large proportion (70%) of studies in ecology support their original hypotheses (Fanelli, 2010). Among other possible explanations, this may suggest a high rate of false positives. A related issue is that 'significant' results have exaggerated effect sizes (Berner & Amrhein, 2022; Kimmel et al., 2023), so the strength of biological effects is often overstated, and even more so for underpowered studies (where sample sizes are too small for accurate inference). Estimates of exaggerated effect sizes in ecology range from 66% to 400% (Lemoine et al., 2016; Yang et al., 2023). The other side of the coin is research waste. Purgar et al. (2021) found that between 82% and 89% of research in ecology appears to be avoidably wasted due to a combination of low-quality studies, publication bias and poor study design, analysis and reporting.

In framing statistical best practice with reference to four principles, we hope to guide ecological researchers, particularly those just starting out, to present the best possible evidence for their conclusions by avoiding these pitfalls. The four principles we have identified are (Figure 1):

1. First, define a focussed research question, then plan sampling and analysis to answer it.
2. Develop a model that accounts for the distribution and dependence of your data.
3. Emphasise effect sizes to replace statistical significance with ecological relevance.
4. Report your methods and findings in sufficient detail so that your research is valid and reproducible.

These principles are listed in approximate order of impact, and later principles require the foundation provided by earlier principles to be effective. For example, reporting effect sizes will not improve a study if the model does not account for dependence. Defining focussed research questions before any data are collected can eliminate HARKing, especially when paired with registration, which has the additional benefit of eliminating cherry picking. Developing a plan for sampling helps to better answer research questions, leading to more meaningful and impactful results, and increases the likelihood population effects are found (statistical power) by promoting better sampling design. Correctly modelling data also increases power and leads to appropriate estimates of uncertainty. Emphasising effect sizes puts the focus on ecological relevance, which is the most meaningful result of ecological research. Comprehensive reporting of methods and findings allows others to successfully replicate your study, lending more weight to your findings and moving the field forward.

2041210x, 2024, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14270 by Area Sistemi Dipart & Document, Wiley Online Library on [02/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

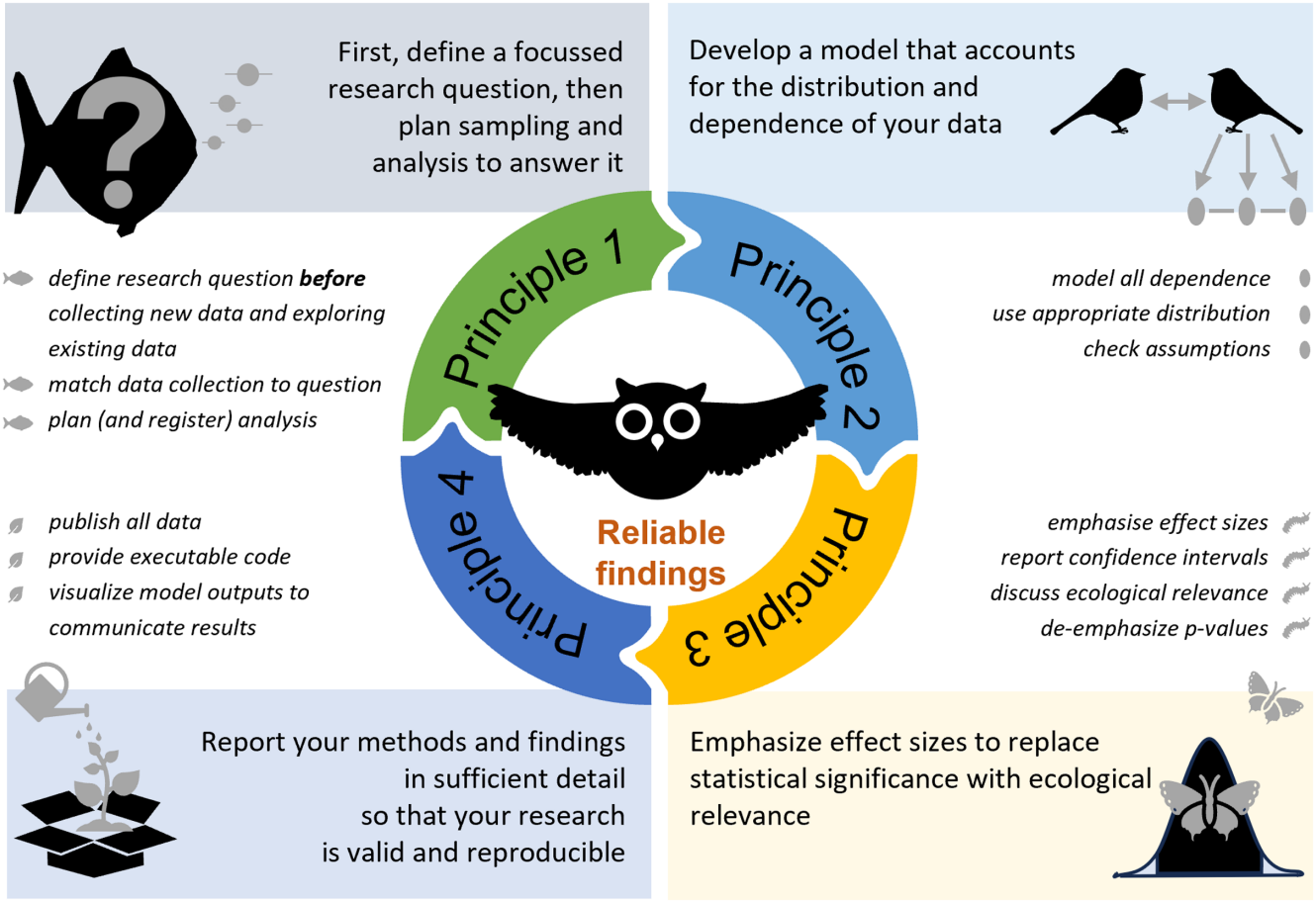

FIGURE 1 Four principles for improved statistical ecology.

These four principles arose out of conversations at the 2020 International Statistical Ecology Conference. The discussion group which conceptualised this paper included a range of ecological statisticians currently working in the field. We noticed that while excellent literature describing good practices in statistics exists, including in ecology, these tend to focus on protocols for conducting specific analyses (e.g. Steel et al., 2013; Zuur et al., 2010; Zuur & Ieno, 2016) or addressing specific problems (Nakagawa & Parker, 2015), rather than a small and digestible number of principles to follow for all inferential analyses. The principles have been written with a frequentist focus but can be easily applied to Bayesian approaches. By defining these principles, we hope to empower ecologists to pursue more robust and meaningful research and encourage collaborations in ecological statistics by helping to develop a common research *methodology*.

Within each section, we will mention useful R packages (R Core Team, 2023) for each principle. We chose R since it is an open and free software and the most used statistical software in ecology (Lai et al., 2019).

Throughout, we will demonstrate the principles with two ecological examples: the wetland experiment, which examines how disturbances affect upland wetlands (Mason et al., 2022); and the bird study, which examines the relationships between hatching date and the expression of carotenoid-based coloration in nestlings (Janas et al., 2020). A full workflow of the principles applied to these examples is available at github.com/gordy2x/principles.

1.1 | Wetland experiment

Underground mining is known to disrupt surface and groundwater flows, which may affect nearby wetland communities. The researchers wanted to examine how differing water availability affected wetland plant communities, both alone and in combination with a fire disturbance. For this study, mesocosms were collected from multiple wetlands, and were then randomised to water and fire treatments in a glasshouse. Mesocosms, for the purpose of this study, were columns of soil and plants, collected by hammering PVC pipes (diameter of 150 mm and a depth of 250 mm) to ground level and extracted with trenching shovels. They were then placed in tubs in a glasshouse, and tub water levels were manipulated to simulate different levels of groundwater availability. A fire event was simulated by sequentially applying biomass removal (clipping), heat and smoke to half of the mesocosms in each water treatment after 20 months (Mason et al., 2022, ssee Figure 2).

2041210x, 2024, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14270 by Area Sistemi Dipart & Document, Wiley Online Library on [02/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

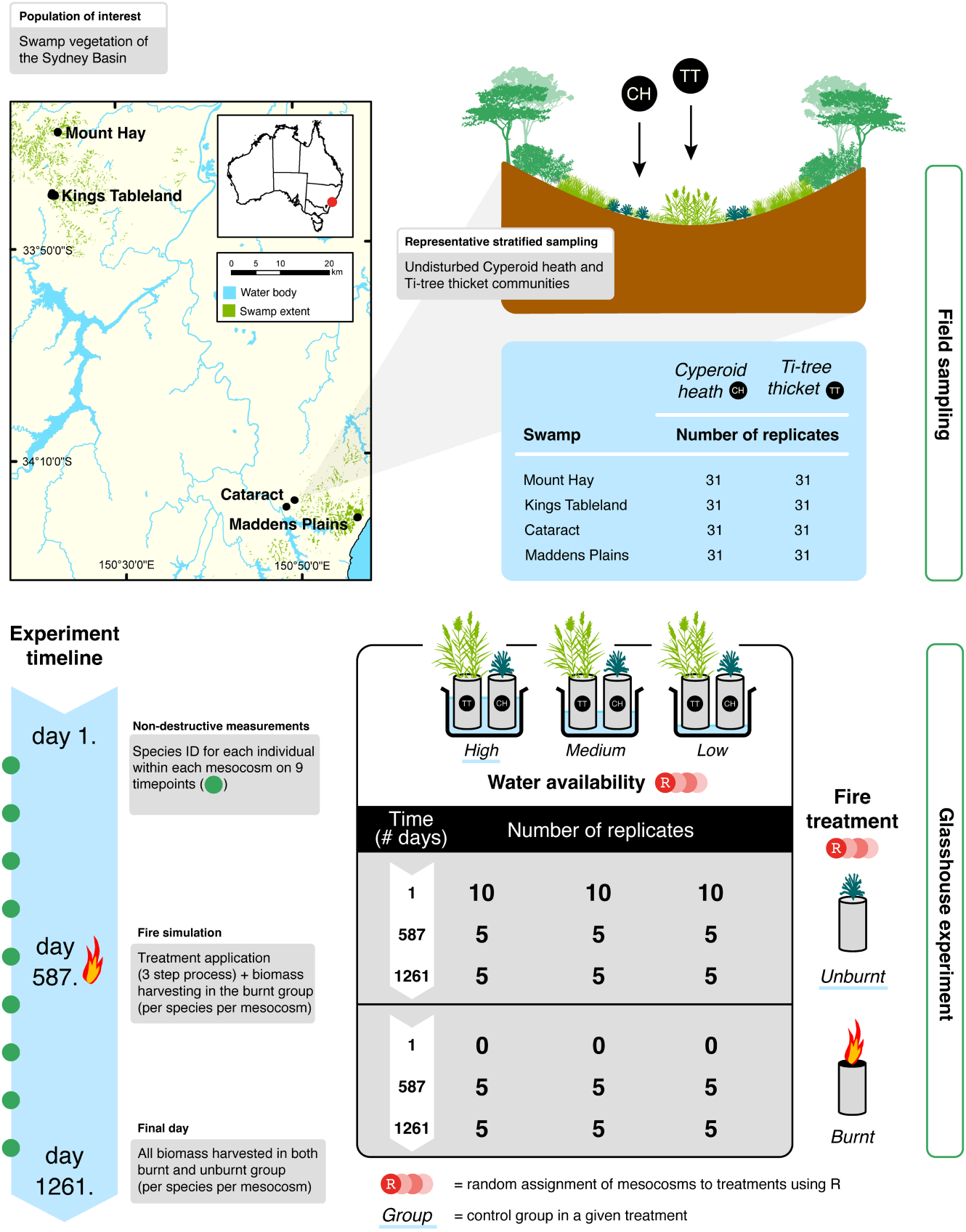

FIGURE 2 Experimental design for wetland experiment to answer 1. How does water availability affect the biomass, richness and composition of wetlands plant communities? 2. How are these changes modified by fire?

2041210x, 2024, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14270 by Area Sistemi Dipart & Document, Wiley Online Library on [02/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

1.2 | Bird study

Blue tits hatch at different points in the season and have variation in carotenoid-based coloration, which is thought to be important for reproductive success. Considering decreasing caterpillar abundance in the second half of the blue tit-hatching season, as well as the prediction that individuals of higher quality breed earlier in the season, researchers expected that the expression of nestlings' carotenoid-based coloration should be negatively correlated with hatching date, and that the effect would be more pronounced in males. Data for this study were collected during six consecutive breeding seasons, beginning in 2011. Each year, nest boxes were regularly inspected from mid-April, and the laying date, number of eggs and hatching date were recorded. The variables of interest here are those most related to blue tit signalling: UV chroma and feather brightness (Janas et al., 2020, see Figure 3).

2 | PRINCIPLE 1. FIRST, DEFINE A FOCUSSED RESEARCH QUESTION, THEN PLAN SAMPLING AND ANALYSIS TO ANSWER IT

2.1 | Define your research question

Developing a good research question is the most important part of the research process. A poorly conceived question can not only limit the usefulness of your findings but also lead to methodological problems throughout the study. A good research question is feasible, interesting, novel, ethical and relevant (the FINER criteria; Hulley, 2007). Each research question should lead to a specific statistical analysis plan, which you would ideally register (see Section 2.3). In the wetland example, the researchers considered a glasshouse experiment to be feasible and ethical, literature review confirmed it was novel and it was both interesting and relevant as planning approval for longwall mining under wetlands is a live and highly contested issue in wetland management.

The PICO framework (Haynes, 2006) is often used to frame research questions in health and medical studies, but can readily be adapted to ecological studies. For each research question, it is important to define the:

- *Population*—population of interest or the target population, usually defined by species, space and time.
- *Intervention*—the treatment that will be applied to subjects in a randomised experiment or the explanatory variable of interest in an observational study.
- *Comparison*—identifies what you plan to use as a reference or control group to compare with your treatment group.
- *Outcome*—represents what outcome(s) you plan to measure, to examine the effectiveness of your intervention or effect of the explanatory variable of interest, often called the response. In ecology, this may be abundance, richness, location, velocity, diversity, among others.

It is often helpful to try to predict what will happen in your study, as this can help clarify your research question. Rather than stating the hypothesis as mere presence of effects (e.g. 'we predict that biomass and richness will depend on water availability'), we recommend focussing on the direction and ideally the magnitude of the expected effects, which can also assist in estimating required sample sizes (Gelman & Carlin, 2014). The wetland researchers thought that mesocosms with less available water would have lower biomass and richness (by 20% or more) than mesocosms with more water availability, and that the effect of lower water would be compounded by fire. The population of interest in this example is upland swamp plant communities of the Sydney Basin, Australia. The intervention groups were combinations of water level and fire: with low, medium and high levels of water availability; and burnt and unburnt fire levels. The control group was represented by unburnt high-water mesocosms. Outcomes included biomass, richness and the presence/absence of each species (species composition). The bird researchers predicted carotenoid-based feather coloration (UV chroma, brightness) would be negatively correlated with hatching date in the population of wild blue tits on the Swedish island of Gotland.

Clearly defining a research question *before* collecting data (or exploring previously collected data) is the most important step in ensuring robust study design. While the benefits of clearly defining a research question are clear, it may be less obvious what the consequences are of refining or revising a hypothesis during or after your study. Hypothesising after the data relating to your original research question have been gathered has a distorting effect on your results, deceptively suggesting that the evidence for your post hoc hypothesis is stronger (Forstmeier et al., 2017). This is most serious when these decisions are made knowing the outcome (e.g. *p*-value) which may constitute HARKing or p-hacking. This is problematic because any changes you make to your hypothesis after collecting data are likely to reflect patterns in the sample that do not necessarily reflect patterns in your population.

What if it becomes necessary to change the research question during the study? Such situations include unexpected field challenges, or when you are analysing pre-existing data and only discover they cannot be used to answer your intended research question when you start exploring them. In case you need to change your research question, it is important to be transparent about the changes and the reasons for them. If changes are made knowing the outcomes, results should be reported as exploratory (see below). Even if you have registered your study (see Section 2.3), departures from analysis plans can be made (Hardwicke & Wagenmakers, 2023).

This prohibition against post hoc hypothesising does not prevent researchers from doing exploratory analyses to inform future research questions and hypotheses. Indeed, generating such hypotheses can be one of the best ways to define productive avenues for further research. As ecologists, we often summarise very complex data to model it simply. For example, you might combine multispecies data into a richness metric to answer your primary question, as in the wetland experiment. This complexity however is often of great interest for understanding the ecology of your study species

2041210x, 2024, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14270 by Area Sistemi Dipart & Document, Wiley Online Library on [02/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

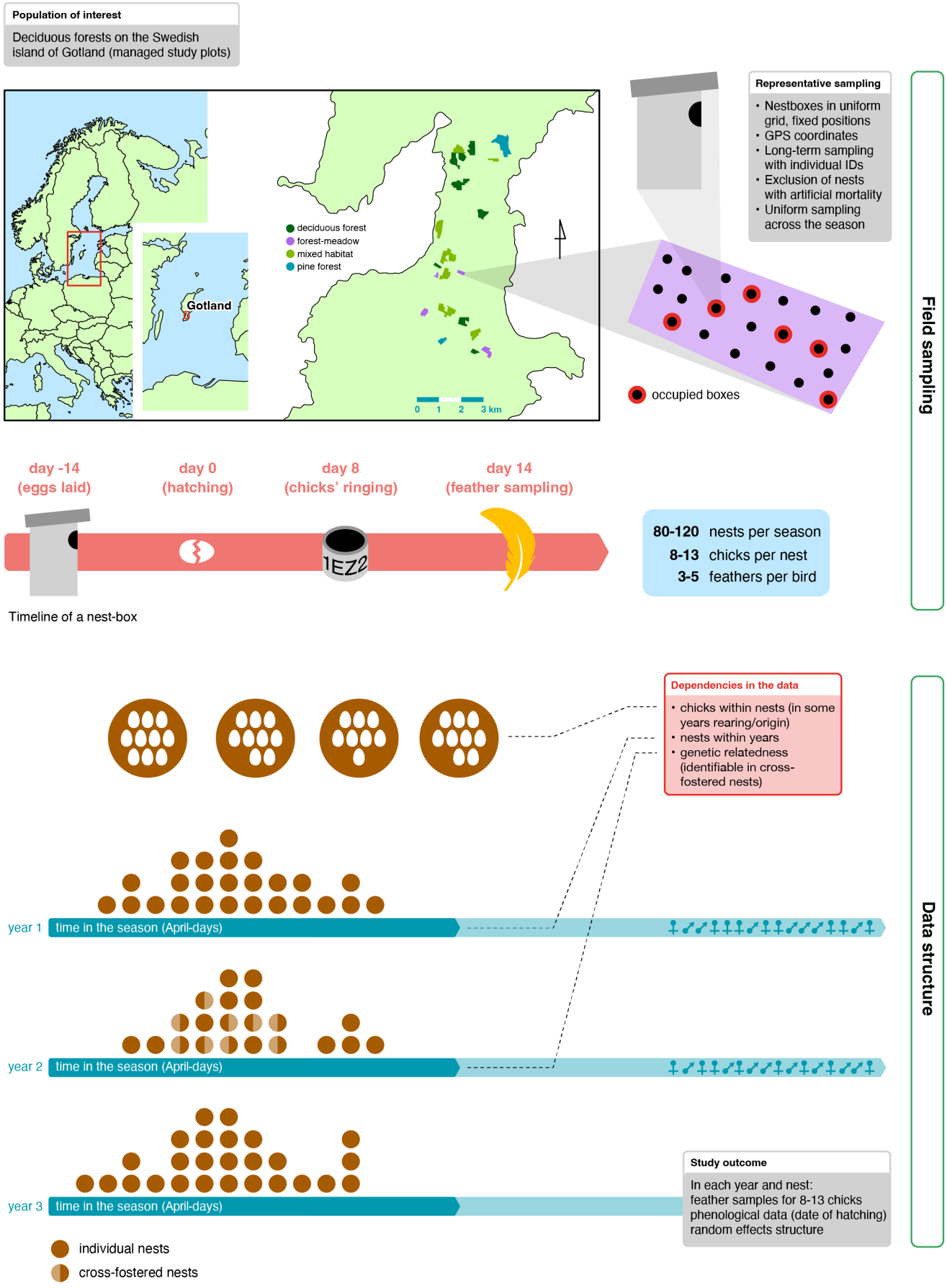

FIGURE 3 Sampling design for bird study to investigate the association between hatching date and carotenoid-based coloration in blue tits.

2041210x, 2024, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14270 by Area Sistemi Dipart & Document, Wiley Online Library on [02/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

and communities, and spending time exploring your raw data often leads to new discoveries and avenues of research. Data exploration can take the form of plotting, tabulating, model fitting and hypothesis testing (including *p*-values). Any time you explore data without a pre-specified research question in mind, you should think of it (and report it) as exploratory or as hypothesis generating (Forstmeier et al., 2017).

2.2 | Match data collection to research aims

Once research questions are outlined, the next step is to decide which data are needed to answer them. Collecting data may include sampling, and experimental manipulation and observation.

2.2.1 | Sampling

Sampling, the process of measuring characteristics (e.g. presence, leaf size and traits) of a set of sampling units (often individuals), should be done in a way that allows unbiased (and precise) estimates of relationships (parameters) of the target population.

Probability samples, where each sampling unit in the population has a known positive probability of being selected for the sample, allow unbiased inference. When the probability is equal for all sampling units, this is called a simple random sample. Stratified sampling involves dividing the population into disjoint groups (strata, e.g. vegetation type or sex) and using a simple random sample within strata. Stratified sampling is often useful when a simple random sample might result in a stratum having too few sampling units for accurate inference. Unbiased inference can be obtained from all probability samples as known biases due to unequal sampling can be modelled, for example, with offsets (to correct for known biases, e.g. sampling intensity) or covariates (to correct for imbalances associated with measured variables, e.g. site accessibility).

The complex spatial and temporal nature of ecology can make probability sampling challenging, and ecologists have in the past settled for 'convenience' samples or even haphazard data collection (Smith et al., 2017). However, this leads to unknown biases in parameter estimates and should be avoided. Helpful tools for probability sampling are the `sample()` function in base R and the `spsample()` function in the `sp` package (Pebesma & Bivand, 2005).

When working with pre-existing datasets rather than collecting new field data, it is important to consider any inherent biases, for which a risk-of-bias assessment tool (e.g. Boyd et al., 2022) may be helpful. When using data collected by others, it is a good practice to contact and collaborate with the researchers who collected the data, as they have the best understanding of their data. A probability sample can sometimes be approximated by carefully selecting the data to use in analyses (Johnston et al., 2021). Alternatively, you can reframe your results to the population for which your data can provide unbiased inference (Williams & Brown, 2019). For example, if you are tagging birds, the birds you tag will be those you can catch (usually not a probability sample), and then the inference will apply to the population of birds that are more easily caught.

2.2.2 | Association and causal relationships

Researchers are most often interested in causal relationships to answer questions about whether and how a response is promoted, caused or induced by a given set of covariates. Finding such cause-and-effect relationships might be the first step to unveil possible ecological mechanisms. Researchers may not explicitly use the word 'cause' but imply causal relationship using causal language such as 'increase', 'decrease', 'improve', 'influence' and 'affect'.

If instead you are only interested in associations, and content to use words like 'associated', 'correlated' and 'predicted', concepts of causal inference covered below are not necessary. The bird study is an example of a non-causal analysis. Researchers do not believe that hatching date caused coloration differences. They believe instead that unmeasured variables (caterpillar abundance and parental quality) cause these differences, and that these are correlated with hatching date, which would lead to a correlation between coloration and date.

Estimating causal relationships is a much harder task than many ecologists realise. The first thing to note is the adage 'correlation does not imply causation'. This means, if we find an association between two variables, we cannot simply conclude there is a causal relationship even when controlling for other variables. Instead, to be able to conclude causation, we need to work much harder. Importantly, every method for causal inference involves making some assumptions. Discussion of these assumptions, their plausibility and any deviations should form a key part of any reporting of causal conclusions.

2.2.3 | Causal conclusions from experiments

Experiments, where researchers manipulate the environment and observe the consequences, can demonstrate causation. To do this, they must have:

- Controls, to know what would have happened without the manipulation;
- Replication within treatments, to apportion observed differences to the manipulation instead of random variation; and
- Randomisation of sampling units to treatments, to avoid bias and confounding with unmeasured/uncontrolled variables.

The assumptions for inferring causality from experiments are met if the experiment is well designed and things go to plan (Kimmel et al., 2021).

The wetland example demonstrates how experiments can be done even in complex environments. Mesocosms were collected

2041210x, 2024, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14270 by Area Sistemi Dipart & Document, Wiley Online Library on [02/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

from four undisturbed upland wetland sites in the Sydney Basin. Transects were established wholly within identified vegetation communities and mesocosms were extracted at random intervals. These wetlands (two in the Blue Mountains and two in the Woronora Plateau) were considered representative of the diversity of wetlands in the Sydney Basin. Each mesocosm was then randomised to treatments, allowing the researchers to demonstrate the cause–effect relationships between water and fire treatments and plant communities. There were several treatment combinations, with the control group defined according to the research question. For example, when comparing water treatments, the controls were high-water unburnt mesocosms, as these were thought to represent the natural state of the wetland. Replication was achieved by sampling and allocating multiple mesocosms ($n = 5$) to each treatment. Sampling units (mesocosms) were allocated to treatments (e.g. low-, medium- and high-water levels) randomly. This randomisation can easily be achieved using widely available software like the `randomizr` package (Coppock, 2019), as was done for the wetland experiment, or simply by drawing numbers out of a hat.

An additional component of many well-conducted experiments is the practice of blinding, where researchers collecting measurements or analysing data do not know the treatment allocation. Blinding can remove unconscious observer bias, but is unfortunately not widely practiced in ecology, where only about 13% of eligible studies undertake blinding (Kardish et al., 2015).

2.2.4 | Causal conclusions from observational data

For ecologists, it is often not feasible or desirable to conduct controlled experiments. It may be that large amounts of observational data on the population of interest already exist, and their use may be preferable to conducting expensive experiments that produce small experimental datasets. These data may be sourced from repositories that combine data from many studies (e.g. GBIF, Movebank and eBird) or from long-term studies, which are often not driven by specific research questions, but instead aim to capture time trends and document unanticipated behaviour (ecological surprises) in habitats, populations and ecosystems. Alternatively, conducting a manipulative experiment may be unethical, for example, it may cause environmental degradation. Finally, manipulative experiments conducted in the laboratory or under artificial conditions may not be generalisable to the population of interest. In the wetland experiment, one of the hypotheses was that compositions would change over time from wetter- to drier adapted species as hydrological resources diminished. However, the glasshouse set-up meant that colonisation of terrestrial (along with wetland) plants from a regional pool was not possible and the range of available species was constrained.

To infer causation from observational (i.e. non-experimental) data, we need to know quite a bit about the system we are studying, and make stronger assumptions, some of which can and should be tested, some of which cannot.

If assumptions are met, you can control for confounding variables using regression type methods to obtain causal relationships. Before inspecting the data, start by drawing a causal diagram, also termed a directed acyclic graph (DAG; see e.g. Arif & MacNeil, 2023). A DAG ideally encodes all known and assumed causal relationships between variables in the investigated system. Once you have a DAG, you can use the so-called 'backdoor' or 'frontdoor' criteria to find the variables to control for, and those to leave out, to estimate the causal relationship of interest (see e.g. Arif & MacNeil, 2023 for a guide). Among others, it is usually recommended to control for confounders (i.e. variables influencing both explanatory variable and response) but not for colliders (i.e. variables that are influenced by both the explanatory variable and response). The `DAGitty` (Textor et al., 2017) R package and online app can help you draw a DAG. Do not use model or variable selection methods to choose variables to control for when estimating causal relationships (Arif & MacNeil, 2022; Stewart et al., 2023). Many of the assumptions you are making are encoded in your DAG, so ideally it should be included in your methods and mentioned in your discussion. For a detailed overview on the full range of causal methods, see for example Hernan and Robins (2023).

In summary, manipulative experiments make by far the least restrictive assumptions for causal inference and give very good estimates of causal effects. However, as we mentioned, experiments do have drawbacks, and it is best to complement manipulative experiments with natural experiments and longitudinal monitoring (Driscoll et al., 2010). When using observational data, it is important to think deeply about and communicate the assumptions you are making, especially if you intend using causal language to communicate your results.

2.3 | Plan analysis and consider registration

Before you collect or explore the available data, it is critical to develop a robust analysis plan. Planning your approach to analysis early can substantially reduce the complexity of the final analysis and improve the clarity of your results. An effective data analysis plan ensures that your study addresses the research question and is developed in tandem with your sampling method (see Section 3 for how to best develop a model that accounts for the characteristics of the data).

An analysis plan must include a comprehensive list of models and tests to be conducted, specifying in each case the model type (e.g. mixed effects model with Gaussian distribution), outcome/response variable (e.g. feather brightness), the principal predictor of interest (e.g. hatching date) and variables to control for (e.g. fixed effects of sex and body weight, random effects of year, nest of origin and nest of rearing). Often, we will not know which model will best fit our data, in which case we should include a description of how a good model will be chosen. The bird study plan was to determine the outcome type (e.g. normal, log transformed) and the relationship type (linear, quadratic) with reference to residual plots. When conducting

2041210x, 2024, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14270 by Area Sistemi Dipart & Document, Wiley Online Library on [02/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

multiple tests, appropriate methods for controlling for multiple testing should be implemented (see e.g. Pike, 2011) to control for false positives.

For the most robust experimental design, we recommend registering your study (we follow Rice & Moher, 2019 to prefer the term registration to pre-registration). Registration archives a detailed description of one's study *before* data collection or sometimes before data analysis (Nosek et al., 2018; Parker et al., 2019; Rice & Moher, 2019). Such description includes research aims (Section 2.1: research questions, hypotheses and predictions), study design (Section 2.2: data collection process) and a data analysis plan (Section 3.1: the statistical models to be fitted). Registration can be timestamped using public registries, such as *Open Science Framework* or *As Predicted*, and it can also be embargoed if needed. A registered report is similar to a registration in that they both commit to their study hypotheses and analysis plans prior to data collection, but it is a distinctive procedure (Chambers, 2013) where the introduction and methods are peer reviewed ahead of data collection. Currently, several ecological journals publish registered reports, including *BMC Ecology and Evolution*, *Ecology and Evolution*, and *Conservation Biology*, and less-specialised journals, such as *PLoS ONE*, *BMC Biology*, *Scientific Reports* and *Nature Communications*.

Registration can have a substantial impact on research quality. A recent study has shown that registered reports in biomedical and psychological research supported only around 40% of their original hypotheses (Allen & Mehler, 2019), relative to 80%–95% in traditional literature. The lower proportion points to a combination of registered reports precluding HARKing and p-hacking (where analyses are changed until a significant result is found), as well as increasing the chances of researchers publishing null results. No one has performed an ecological counterpart of this study yet. We anticipate registration and registered reports would bring a more reasonable ratio between positive and negative findings in ecology.

3 | PRINCIPLE 2. DEVELOP A MODEL THAT ACCOUNTS FOR THE DISTRIBUTION AND DEPENDENCE OF YOUR DATA

3.1 | Model dependence

The independence of errors is a critical assumption of almost all statistical methods. Errors are the unmodelled portion of the data after modelling dependence and impacts of covariates. The independence assumption can be met by collecting independent data (e.g. using a simple random sample) or by appropriately modelling (accounting for) any dependence in the data.

Dependence often arises from sampling design. Consider, for instance, a hierarchical (or nested) sampling design like in the bird study, where multiple hatchlings were sampled per nest. Dependence arises because hatchlings in one nest are more similar to each other than those in different nests. Such subsampling (or 'pseudo-replication' as it is often called) is a common source of dependence (Hurlbert, 1984), as is temporal dependence where, for example, the abundance of species in a wetland mesocosm is measured repeatedly. Again, these repeated measures (Gurevitch & Chester, 1986) result in observations that are more similar to each other for each mesocosm than they are between mesocosms. Spatial dependence, where sites closer in space are more similar than sites further apart, is also common. Dependence can also arise due to the complex nature of the system. For example, in multi-species surveys, dependence between species abundances may arise from common responses to unobserved environmental gradients, phylogeny or interactions between species.

Dependence is not problematic, per se; replicate observations can be extremely valuable if understood, but they must be accounted for correctly in the analysis (Steel et al., 2013). If dependence is not correctly accounted for, it will lead to incorrect inference (often underestimating uncertainty). Modelling common dependence types is straightforward in most statistical software (see Warton, 2022 for a detailed guide; and the mixed, multilevel and hierarchical models in R CRAN Task View; Bolker et al., 2022, for a comprehensive list of models). Briefly, generalised linear (mixed) models (McCulloch & Neuhaus, 2006), implemented in the `lme4` (Bates et al., 2015, p. 4) package, can model dependence due to clustering by including a random intercept for the clustering variable (`~ (1|mesocosm)`). Random intercepts are also often used for longitudinal data (repeated measurements over time), though these data often need more structured temporal dependence by for example adding random slopes (`~ (time|mesocosm)`), which allow the change over time to vary by mesocosm (Schielzeth & Forstmeier, 2009). For multivariate dependence, there are hierarchical models implemented in the `gllvm` (Niku et al., 2019, 2020) and `Hmsc` (Tikhonov, Opedal, et al., 2020; Tikhonov, Ovaskainen, et al., 2020) packages, as well as generalised estimating equations (e.g. `mvabund`; Wang et al., 2012). If the data you are modelling have several types of complex dependence, then it is less straightforward to find software to model these. You may need to use special purpose software (see, e.g. the Analysis of Ecological and Environmental Data CRAN Task View; Simpson, 2023) or adapt flexible software to your problem (e.g. RStan: Stan Development Team, 2023; INLA: Rue et al., 2009; NIMBLE: De Valpine et al., 2017; greta: Golding, 2019).

The wetland and bird examples have several layers of dependence, as is common in ecological data. In the wetland experiment, mesocosms within one swamp are dependent, repeated measurements of each mesocosm are dependent and measurements include multiple species, which have complex interdependences. The wetland researchers fitted hierarchical models using the `glmmTMB` (Brooks et al., 2017) and `lme4` packages, where, in addition to the effects of interest (vegetation, water and fire treatments), they included fixed effects for the swamp, random effects for mesocosm and a reduced rank correlation structure between species (see Figure 4). Dependence in the bird example is due to clustering (by year and nest of origin). In addition, a cross-fostering experiment conducted in the study that was unrelated to this research question induced correlation due to the nest of rearing (see Supplementary

2041210x, 2024, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14270 by Area Sistemi Dipart & Document, Wiley Online Library on [02/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

```
library(lme4); library(emmeans)
rich_mod <- lmer(log(biomass + 1) ~ swamp + veg + days * water +
                                    days * fire + time * fire +
                                    (1 | Mesocosm), data = bio_data)
emmip(rich_mod,  ~ water ~ time | fire, Cis = TRUE)
```

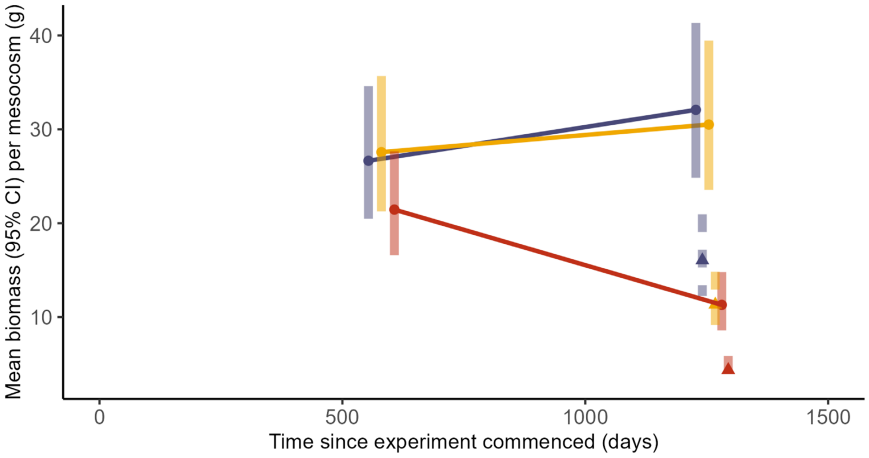

FIGURE 4 Estimated marginal means (±95% confidence intervals) for species richness over time (days since commencement) in treatment mesocosms. Treatment levels are high (▬), medium (▬) and low (▬) water availability, and unburnt (—▪—) and burnt (···▪···) fire treatment levels.

materials for details). Researchers used the `lme4` package for analysis, including year, the nest of origin and the nest of rearing as random effects.

3.2 | Check assumptions

We have already discussed the importance of modelling dependence in Section 3.1. In addition, statistical models assume the response comes from a particular probability distribution with key properties including the response type (continuous, binary, counts) and its mean–variance relationship. Examples include: the Gaussian (normal) distribution for continuous data like weight, which assumes constant variance of the residuals; binomial distribution for presence/absence with a binary response and a quadratic mean–variance relationship; Poisson or negative binomial distribution for counts, like abundance, which assumes the variance is equal to or greater than the mean. Rather than fitting the data to the model, we should aim to develop a model that accounts for the characteristics of the data. If assumptions are violated (in a consequential way, see below), then the model does not account for the characteristics of the data, and researchers should endeavour to find a better fitting model. The violation of model assumptions can bias parameter estimates, for example, by confounding location and dispersion effects (Warton et al., 2012); or underestimate uncertainty, resulting in overconfidence in results. If a better model cannot be found (one may not exist or the correct model may not be implemented in code), then sensitivity analyses should be done. Sensitivity analysis is simply the process of trying alternative models that make different assumptions to assess whether the conclusions of the analysis are robust to these changes (i.e. give consistent results).

Every model fitted must be checked and adjustments made before conclusions can be drawn. Assumptions are usually checked visually, by inspecting plots of residuals. A plot of residual versus predicted (fitted) values can show departure from linearity or the assumed mean–variance relationship. Plots of residuals over time or space would show the magnitude of temporal or spatial autocorrelation, respectively. Normal quantile plots can diagnose departures from normality (if a normal distribution is used). The effect of violating assumptions varies from model to model. For example, in a linear regression, the lack of normality is not as critical as constant variance, the violation of which may increase Type I error rates (Glass et al., 1972). As most distributions fit by ecologists are discrete (e.g. Poisson, negative-binomial, binomial), there are many choices of residuals. Quantile residuals, implemented in the `statmod` (Dunn & Smyth, 1996) and `DHARMa` (Hartig, 2020) packages, are perhaps the most suited for checking assumption violations for ecologists.

Sometimes, models require assumptions that cannot be checked. Two common scenarios are missing data, where it is necessary to make assumptions about the type of missingness (Nakagawa & Freckleton, 2008), and causal inference from observational data. Here, sensitivity analysis can help assess how sensitive results are to these assumptions.

4 | PRINCIPLE 3: EMPHASISE EFFECT SIZES TO REPLACE STATISTICAL SIGNIFICANCE WITH ECOLOGICAL RELEVANCE

4.1 | Report ecological relevance by emphasising effect sizes

Ecologists seek tools that allow them to say something about the ecological 'significance', or to use a better and less confusing term, assess the 'ecological relevance' of their findings (Martínez-Abraín, 2008).

2041210x, 2024, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14270 by Area Sistemi Dipart & Document, Wiley Online Library on [02/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Since that is hard, ecologists often settle for stating that their results are 'statistically significant'. While this is tempting, statistical significance is insufficient for the reporting of the results of ecological modelling because it provides no information on effect sizes (Nakagawa & Cuthill, 2007).

Given even a miniscule effect size in the population—say as an example, a 0.01% difference in the mean biomass between two water treatments—statistical significance could still be achieved if one could increase the sample size sufficiently. This will not make a 0.01% difference any more ecologically relevant. The effect size (here, mean difference) is the quantity of interest and must be prioritised in analysis and reporting, along with a discussion on how ecologically relevant the measured effect size is, which must be based on knowledge of the system (see Kruschke, 2018, for review on meaningful effect sizes in different fields).

Focussing on the estimation of effect sizes requires reporting and interpreting confidence intervals (or equivalents when using other statistical approaches, such as credible intervals for Bayesian analysis). The width of these intervals indicates the uncertainty associated with an estimate. If you focus on estimation rather than significance testing, you can better capture the nuances of statistical analysis and interpretation.

4.2 | De-emphasise *p*-values

The *p*-value is defined as the probability, under the assumption of no effect or no difference in the population (i.e. under the null hypothesis), of obtaining a test-statistic equal to or more extreme than what was observed in the sample. This is a mouthful, and it is no wonder it is so commonly misinterpreted. Essentially, this means *p*-values are a measure of how incompatible the data are with a specified statistical model (Amrhein & Greenland, 2022). A *p*-value is *not* the probability that the null hypothesis is true, nor is it an indicator of the size of the effect, nor the probability that the data were produced by random chance alone, among other misinterpretations (Wasserstein & Lazar, 2016).

P-values and 'significance' cut-offs have dominated scientific publishing since the early 20th century, originally because without the use of computers, widely used tables of test statistics encouraged this paradigm, though the practice was always controversial among statisticians (Kennedy-Shaffer, 2019). There have been calls for abandoning *p*-values and hypothesis testing altogether, both due to their ubiquitous incorrect interpretation and due to previously mentioned problems like cherry picking, and testing hypotheses which are known a priori to be false. More commonly, statisticians recommend we stop dichotomizing *p*-values (into significant/non-significant; McShane et al., 2019), which not only confuses statistical and ecological significance, but is also one of the causes of bad practices like *p*-hacking. Statisticians also recommend de-emphasising the importance of *p*-values (cf. Hardwicke et al., 2023). Current best practice is to report *p*-values as an addition to more important quantities like effect size and confidence intervals, which also makes it harder to misinterpret the *p*-values. As Wasserstein et al. (2019, p. 2) recommend, when interpreting results it is best to 'accept uncertainty. Be thoughtful, open, and modest'.

When modelling changes in biomass over time in the wetland example, the researchers found 'evidence ($p = 0.006$) that differences in biomass between unburnt low- and high-water mesocosms increased over time, with biomass differences between high- and low-water mesocosms more than doubling (change = 2.2; 95% CI: 1.2–4.0) between Day 587 and Day 1261 of the experiment'. The confidence intervals suggest that the biomass difference is conceivably as small as a 20% increase (1.2) or as large as a quadrupling (4.0). The former might be viewed as being of modest ecological relevance while a doubling or more might be viewed as a large effect. The researchers additionally found 'no evidence of differences in biomass changes between high- and medium-water mesocosms ($p = 0.945$; change = 1.1; 95% CI: 0.6–2.0)'. The wide 95% confidence interval (0.6–2.0) suggests that the difference between high- and medium-water mesocosms could conceivably be almost halving (0.6) or doubling (2.0) of biomass. The large *p*-value and wide confidence interval may be a result of a too small sample size or a too large variance, or it could theoretically be that there is no difference between high- and medium-water mesocosms. Since the researchers cannot tell which of these factors was the cause of the large *p*-value, it would be wrong to conclude that there is no difference between high- and medium- water levels.

5 | PRINCIPLE 4: REPORT YOUR METHODS AND FINDINGS IN SUFFICIENT DETAIL SO THAT YOUR RESEARCH IS COMPELLING AND REPRODUCIBLE

5.1 | Make it easy to reproduce your study findings

Replicability and reproducibility are important considerations when reporting your research. While definitions vary (Goodman et al., 2016), we will use the Turing Way Community (2023) definition, which says a result is *reproducible* when the same analytical steps performed on the same dataset consistently produce the same answer. A result is *replicable* when the same analysis performed on different datasets produces qualitatively similar answers. As noted previously, replication studies in ecology are rare (Kelly, 2019) and few published studies provide sufficient detail about study design, data collection and analysis for replication studies to even be attempted (Culina et al., 2020). Apart from the scientific benefits of comprehensive reporting, the successful replication of your study will lend a lot more weight to your findings, and this should be your goal. The aim is to help anyone who attempts to replicate your work by providing all the information they might need. Of relevance, collaborative networks are perhaps a good way forward for increasing replication studies in ecology, like recent examples of the Nutrient Network, US Long-Term Ecological Research network and Zostera Experimental Network (Yang et al., 2023). This is because such

2041210x, 2024, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14270 by Area Sistemi Dipart & Document, Wiley Online Library on [02/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

networks allow replications at multiple sites, which, in turn, enable us to examine the heterogeneity and generalisability of a phenomenon (cf. Ives, 2018).

If you have never included data and code with your publications, it can seem overwhelming (Gomes et al., 2022), but it is not all or none: each step towards reproducibility is worthwhile. Our advice follows that of a colleague (D. Falster, pers. comm.), who recommends you start small and work up to full reproducibility; first, writing the code so you yourself can reproduce the results after some time has passed; second, by making sure a collaborator can reproduce them; finally it is only a small extra step to provide completely reproducible code in your publications. Some good practices, in approximate order of difficulty, are the following:

- Always provide your complete data in a 'non-proprietary machine-readable format' such as .csv (not as hard scanned-in pages or PDF files);
- Keeping a research journal, noting any changes to study design or analysis, and why these were made, then reporting these in your methods and supplementary materials;
- Make sure your code is executable—that is your script file or rmarkdown/Quarto document should run from top to bottom without interruptions, exceptions and errors;
- Always supplement your papers with complete code to reproduce your results (and if using non-code based methods, provide a detailed workflow to trace your analytical steps);
- Where possible, avoid using commercial software and promote Open Source and freely available computational tools;
- Avoid using ephemeral and transient hosts to keep your code and data (e.g. personal websites, departmental web archives). Use free, publicly accessible and well-maintained repositories instead (e.g. Figshare, Dryad, Zenodo);
- When presenting code use version-control systems (such as Git) and rich documents integrating code with comments (e.g. by using Quarto, rmarkdown and R packages such as `knitr`);
- Declare packages used in the analysis and their versions (e.g. with the help of the `renv` package). If software/packages you use may change beyond being re-usable quickly consider packing your code and data into a self-contained package (e.g. using Code Ocean or Docker).

5.2 | Visualise model outputs to communicate results

As we have discussed in Section 4, results should focus on effect sizes and confidence intervals to promote a focus on ecological relevance. One of the most effective ways to do this is to plot the *model outputs*. By model outputs, we mean the estimated effects and confidence intervals produced by the model. While some packages have inbuilt functions to do this, the `emmeans` (Lenth, 2021) and `sjPlot` (Lüdecke, 2021) packages can plot model outputs from almost any commonly used R package. An alternative is to use the `predict()` function in most R packages to calculate predictions with uncertainty, then plot them manually. It is important to plot model outputs with uncertainty (i.e. confidence intervals), which is easily accomplished by using `CIs=TRUE` in `emmeans::emmip`, or `se.fit=TRUE` in many packages' `predict()` functions. The wetland and bird studies included plots created with `emmeans` (code in Figure 4), then modified with `ggplot` (Wickham, 2016); model estimates and confidence intervals for richness over time are reproduced here (Figure 4).

6 | CONCLUSIONS

This paper has proposed guidelines to enhance statistical methodology for ecological studies suitable for use by individual researchers and research teams. While outside the scope of this work, undoubtedly there is more that can be done at a systemic level to encourage the adoption of stronger experimental design and reporting practices throughout the discipline. The authors are aware of journals that are beginning to embed more robust methodological requirements into review processes. We acknowledge that it can be challenging for researchers to remain engaged with the complexities of statistical practice when deeply engaged in their specific areas of research, and it is our hope that clear guidance and critical engagement with statistical methods will help to build statistical competence and fluency to the benefit of the ecological research community.

AUTHOR CONTRIBUTIONS

Gordana Popovic conceived the idea, and co-led writing and synthesis with Carolyn Claire Isabelle Burns. Tanya Jane Mason provided the wetland example. Szymon Marian Drobniak provided the bird example. Tiago André Marques, Joanne Potts, Rocío Joo, Res Altwegg, Michael Andrew McCarthy, Alison Johnston, Shinichi Nakagawa and Louise McMillan led writing of sections (authorship order is randomised). Kadambari Devarajan, Patrick Leo Taggart, Alison Wunderlich, Magdalena M Mair, Juan Andrés Martínez-Lanfranco, Malgorzata Lagisz and Patrice Pottier contributed to multiple sections (authorship order randomised). All authors contributed critically to the drafts and gave final approval for publication.

AFFILIATIONS

[1]Stats Central, Mark Wainwright Analytical Centre, UNSW Sydney, Sydney, New South Wales, Australia; [2]Centre for Ecosystem Science, School of Biological, Earth and Environmental Sciences, UNSW Sydney, Sydney, New South Wales, Australia; [3]Science, Economics and Insights Division, NSW Department of Climate Change, Energy, the Environment and Water, Lidcombe, New South Wales, Australia; [4]Evolution and Ecology Research Centre, School of Biological, Earth and Environmental Sciences, UNSW Sydney, Sydney, New South Wales, Australia; [5]Institute of Environmental Sciences, Jagiellonian University, Krakow, Poland; [6]Centre for Research into Ecological and Environmental Modelling, The Observatory, University of St Andrews, St Andrews, Scotland; [7]Centro de Estatística e Aplicações, Departamento de Biologia Animal, Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal; [8]The Analytical

2041210x, 2024, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14270 by Area Sistemi Dipart & Document, Wiley Online Library on [02/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Edge Statistical Consulting, Blackmans Bay, Tasmania, Australia; [9]Global Fishing Watch, Washington, District of Columbia, USA; [10]Centre for Statistics in Ecology, Environment and Conservation, Department of Statistical Sciences, University of Cape Town, Rondebosch, South Africa; [11]Sydney, New South Wales, Australia; [12]School of Agriculture, Food and Ecosystem Sciences, The University of Melbourne, Parkville, Victoria, Australia; [13]Centre for Research into Ecological and Environmental Modelling, Mathematics and Statistics, University of St Andrews, St Andrews, UK; [14]School of Mathematics and Statistics, Victoria University of Wellington, Wellington, New Zealand; [15]Organismic and Evolutionary Biology Graduate Program, University of Massachusetts at Amherst, Amherst, Massachusetts, USA; [16]Department of Natural Resources Science, University of Rhode Island, Kingston, Rhode Island, USA; [17]Vertebrate Pest Research Unit, Department of Primary Industries NSW, Queanbeyan, New South Wales, Australia; [18]Institute of Biosciences, São Paulo State University, São Vicente, São Paulo, Brazil; [19]Statistical Ecotoxicology, University of Bayreuth, Bayreuth, Germany; [20]Theoretical Ecology, University of Regensburg, Regensburg, Germany and [21]Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada

ACKNOWLEDGEMENTS

The authors would like to thank everyone who participated in the vISEC2020 discussion group and subsequent discussions. We thank Glenda Wardle, Annemieke Drost, Nilanjan Chatterjee, Pedro Nicolau, Chloe Bracis, Teresa Neeman, Javier Seoane, Sarah Marley, Noa Rigoudy, Brenton Annan, Rebecca Groenewegen, Theresa O'Brien, Michelle Marraffini, Julie Vercelloni, Andrea Havron, Hayden Schilling, Amanda Hart, Christine Stawitz, Fabiana Ferracina, Andrew Edwards, Mick Wu, Gesa von Hirschheydt, Rick Camp, Alison Ketz, Julie Vercelloni, Sarah Saunders, Samantha Burke and Sarah Hasnain for their brainstorming contributions. Additionally, authors are grateful to Frederic Gosselin for pointing out an important reference for testing implausible H0, Daniel Falster for reproducibility advice, and reviewers for valuable suggestions and feedback. Bird study data came from the long-term Gotland study managed by Szymon M. Drobniak, Aneta Arct and Mariusz Cichoń. We thank Katarzyna Janas, Blandine Doligez and all field assistants collecting data over the years. We are also grateful to all landowners for their permission to access the study plots. Open access publishing facilitated by University of New South Wales, as part of the Wiley - University of New South Wales agreement via the Council of Australian University Librarians.

FUNDING INFORMATION

TJM acknowledges assistance by the NSW Government through its Environmental Trust (2018/SSC/0049) and Saving Our Species program. TAM thanks partial support by CEAUL (funded by FCT—Fundação para a Ciência e a Tecnologia, Portugal, through the project UIDB/00006/2020). RA is supported by National Research Foundation of South Africa (Grant no. 114696). AW was supported by the São Paulo Research Foundation (Grant no. #17/16650–5). MMM was supported by the Christiane Nüsslein-Volhard Foundation. JAM was supported by the National Agency of Research and Innovation (ANII-Uruguay), and Computational Biodiversity Science and Services Program (Bios2-Canada). PP was supported by a UNSW Scientia Doctoral scholarship. SMD was supported by the Australian Research Council (ARC) DECRA Fellowship (DE180100202) and the Opus grant from Polish National Science Centre (no. UMO-2020/39/B/NZ8/01274). SN and ML were funded by the ARC Discovery Project Grant (DP210100812).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest in relation to this paper.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14270.

DATA AVAILABILITY STATEMENT

All data, code and full workflows of the principles applied to the examples are available at https://github.com/gordy2x/principles (Popovic & Drobniak, 2023).

ORCID

Gordana Popovic https://orcid.org/0000-0002-1376-1058
Tanya Jane Mason https://orcid.org/0000-0002-4797-5644
Szymon Marian Drobniak https://orcid.org/0000-0001-8101-6247
Tiago André Marques https://orcid.org/0000-0002-2581-1972
Joanne Potts https://orcid.org/0000-0003-2752-5539
Rocío Joo https://orcid.org/0000-0003-0319-4210
Res Altwegg https://orcid.org/0000-0002-4083-6561
Michael Andrew McCarthy https://orcid.org/0000-0003-1039-7980
Alison Johnston https://orcid.org/0000-0001-8221-013X
Shinichi Nakagawa https://orcid.org/0000-0002-7765-5182
Louise McMillan https://orcid.org/0000-0002-0536-8563
Kadambari Devarajan https://orcid.org/0000-0002-9222-5404
Patrick Leo Taggart https://orcid.org/0000-0001-9523-0463
Alison Wunderlich https://orcid.org/0000-0001-9222-8536
Magdalena M. Mair https://orcid.org/0000-0003-0074-6067
Juan Andrés Martínez-Lanfranco https://orcid.org/0000-0002-1692-1168
Malgorzata Lagisz https://orcid.org/0000-0002-3993-6127
Patrice Pottier https://orcid.org/0000-0003-2106-6597

REFERENCES

Allen, C., & Mehler, D. M. A. (2019). Open science challenges, benefits and tips in early career and beyond. *PLoS Biology*, *17*(5), e3000246. https://doi.org/10.1371/journal.pbio.3000246

Amrhein, V., & Greenland, S. (2022). Rewriting results in the language of compatibility. *Trends in Ecology & Evolution*, *37*(7), 567–568. https://doi.org/10.1016/j.tree.2022.02.001

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *The Journal of Wildlife Management*, *64*(4), 912. https://doi.org/10.2307/3803199

Arif, S., & MacNeil, M. A. (2022). Predictive models aren't for causal inference. *Ecology Letters*, *25*(8), 1741–1745. https://doi.org/10.1111/ele.14033

Arif, S., & MacNeil, M. A. (2023). Applying the structural causal model framework for observational causal inference in ecology. *Ecological Monographs*, *93*(1), e1554. https://doi.org/10.1002/ecm.1554

2041210x, 2024, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14270 by Area Sistemi Dipart & Document, Wiley Online Library on [02/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531–533. https://doi.org/10.1038/483531a

Berner, D., & Amrhein, V. (2022). Why and how we should join the shift from significance testing to estimation. *Journal of Evolutionary Biology*, *35*(6), 777–787. https://doi.org/10.1111/jeb.14009

Bolker, B., Piaskowski, J., Tanaka, E., Alday, P., & Viechtbauer, W. (2022). *CRAN task view: Mixed, multilevel, and hierarchical models in R*. Version 2022-10-31. https://CRAN.R-project.org/view=MixedModels

Boyd, R. J., Powney, G. D., Burns, F., Danet, A., Duchenne, F., Grainger, M. J., Jarvis, S. G., Martin, G., Nilsen, E. B., Porcher, E., Stewart, G. B., Wilson, O. J., & Pescott, O. L. (2022). ROBITT: A tool for assessing the risk-of-bias in studies of temporal trends in ecology. *Methods in Ecology and Evolution*, *13*(7), 1497–1507. https://doi.org/10.1111/2041-210X.13857

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex*, *49*(3), 609–610. https://doi.org/10.1016/j.cortex.2012.12.016

Coppock, A. (2019). *randomizr: Easy-to-use tools for common forms of random assignment and sampling*. https://CRAN.R-project.org/package=randomizr

Culina, A., van den Berg, I., Evans, S., & Sánchez-Tójar, A. (2020). Low availability of code in ecology: A call for urgent action. *PLoS Biology*, *18*(7), e3000763. https://doi.org/10.1371/journal.pbio.3000763

De Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., & Bodik, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, *26*(2), 403–413. https://doi.org/10.1080/10618600.2016.1172487

Driscoll, D. A., Lindenmayer, D. B., Bennett, A. F., Bode, M., Bradstock, R. A., Cary, G. J., Clarke, M. F., Dexter, N., Fensham, R., Friend, G., Gill, M., James, S., Kay, G., Keith, D. A., MacGregor, C., Russell-Smith, J., Salt, D., Watson, J. E. M., Williams, R. J., & York, A. (2010). Fire management for biodiversity conservation: Key research questions and our capacity to answer them. *Biological Conservation*, *143*(9), 1928–1939. https://doi.org/10.1016/j.biocon.2010.05.026

Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, *5*(3), 236–244.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS ONE*, *5*(4), e10068. https://doi.org/10.1371/journal.pone.0010068

Fidler, F., Burgman, M. A., Cumming, G., Buttrose, R., & Thomason, N. (2006). Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. *Conservation Biology*, *20*(5), 1539–1544. https://doi.org/10.1111/j.1523-1739.2006.00525.x

Forstmeier, W., Wagenmakers, E., & Parker, T. H. (2017). Detecting and avoiding likely false-positive findings—A practical guide. *Biological Reviews*, *92*(4), 1941–1968. https://doi.org/10.1111/brv.12315

Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. *PLoS ONE*, *13*(7), e0200303. https://doi.org/10.1371/journal.pone.0200303

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*(3), 52.

Golding, N. (2019). greta: Simple and scalable statistical modelling in R. *Journal of Open Source Software*, *4*(40), 1601. https://doi.org/10.21105/joss.01601

Gomes, D. G. E., Pottier, P., Crystal-Ornelas, R., Hudgins, E. J., Foroughirad, V., Sánchez-Reyes, L. L., Turba, R., Martinez, P. A., Moreau, D., Bertram, M. G., Smout, C. A., & Gaynor, K. M. (2022). Why don't we share data and code? Perceived barriers and benefits to public archiving practices. *Proceedings of the Royal Society B: Biological Sciences*, *289*(1987). https://doi.org/10.1098/rspb.2022.1113

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*(341). https://doi.org/10.1126/scitranslmed.aaf5027

Gurevitch, J., & Chester, S. T. (1986). Analysis of repeated measures experiments. *Ecology*, *67*(1), 251–255. https://doi.org/10.2307/1938525

Hardwicke, T. E., Salholz-Hillel, M., Malički, M., Szűcs, D., Bendixen, T., & Ioannidis, J. P. A. (2023). Statistical guidance to authors at top-ranked journals across scientific disciplines. *The American Statistician*, *77*(3), 239–247. https://doi.org/10.1080/00031305.2022.2143897

Hardwicke, T. E., & Wagenmakers, E.-J. (2023). Reducing bias, increasing transparency and calibrating confidence with preregistration. *Nature Human Behaviour*, *7*(1), 15–26. https://doi.org/10.1038/s41562-022-01497-2

Hartig, F. (2020). *DHARMa: Residual diagnostics for hierarchical (multi-level/mixed) regression models*. https://CRAN.R-project.org/package=DHARMa

Haynes, R. B. (2006). Forming research questions. *Journal of Clinical Epidemiology*, *59*(9), 881–886. https://doi.org/10.1016/j.jclinepi.2006.06.006

Hernan, M. A., & Robins, J. M. (2023). *Causal inference: What if*. Chapman & Hall/CRC.

Hulley, S. B. (2007). *Designing clinical research*. Lippincott Williams & Wilkins.

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, *54*(2), 187–211.

Ives, A. R. (2018). Informative irreproducibility and the use of experiments in ecology. *BioScience*, *68*(10), 746–747. https://doi.org/10.1093/biosci/biy090

Janas, K., Lutyk, D., Sudyka, J., Dubiec, A., Gustafsson, L., Cichoń, M., & Drobniak, S. (2020). Carotenoid-based coloration correlates with the hatching date of blue tit *Cyanistes caeruleus* nestlings. *Ibis*, *162*(3), 645–654. https://doi.org/10.1111/ibi.12751

Johnston, A., Hochachka, W. M., Strimas-Mackey, M. E., Ruiz Gutierrez, V., Robinson, O. J., Miller, E. T., Auer, T., Kelling, S. T., & Fink, D. (2021). Analytical guidelines to increase the value of community science data: An example using eBird data to estimate species distributions. *Diversity and Distributions*, *27*(7), 1265–1277. https://doi.org/10.1111/ddi.13271

Kardish, M. R., Mueller, U. G., Amador-Vargas, S., Dietrich, E. I., Ma, R., Barrett, B., & Fang, C.-C. (2015). Blind trust in unblinded observation in ecology, evolution, and behavior. *Frontiers in Ecology and Evolution*, *3*. https://doi.org/10.3389/fevo.2015.00051

2041210x, 2024, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14270 by Area Sistemi Dipart & Document, Wiley Online Library on [02/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Kelly, C. D. (2019). Rate and success of study replication in ecology and evolution. *PeerJ*, *7*, e7654. https://doi.org/10.7717/peerj.7654

Kennedy-Shaffer, L. (2019). Before $p < 0.05$ to beyond $p < 0.05$: Using history to contextualize *p*-values and significance testing. *The American Statistician*, *73*(sup1), 82–90. https://doi.org/10.1080/00031305.2018.1537891

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kimmel, K., Avolio, M. L., & Ferraro, P. J. (2023). Empirical evidence of widespread exaggeration bias and selective reporting in ecology. *Nature Ecology & Evolution*, *7*(9), 1525–1536. https://doi.org/10.1038/s41559-023-02144-3

Kimmel, K., Dee, L. E., Avolio, M. L., & Ferraro, P. J. (2021). Causal assumptions and causal inference in ecological experiments. *Trends in Ecology & Evolution*, *36*(12), 1141–1152. https://doi.org/10.1016/j.tree.2021.08.008

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. https://doi.org/10.1177/2515245918771304

Lai, J., Lortie, C. J., Muenchen, R. A., Yang, J., & Ma, K. (2019). Evaluating the popularity of R in ecology. *Ecosphere*, *10*(1). https://doi.org/10.1002/ecs2.2567

Lemoine, N. P., Hoffman, A., Felton, A. J., Baur, L., Chaves, F., Gray, J., Yu, Q., & Smith, M. D. (2016). Underappreciated problems of low replication in ecological field studies. *Ecology*, *97*(10), 2554–2561. https://doi.org/10.1002/ecy.1506

Lenth, R. V. (2021). emmeans: Estimated marginal means, aka least-squares means. https://CRAN.R-project.org/package=emmeans

Lüdecke, D. (2021). *sjPlot: Data visualization for statistics in social sciencemmeans: Estimated marginal means, aka least-squares means* (R package version 2.8.10). https://CRAN.R-project.org/package=sjPlot

Martínez-Abraín, A. (2008). Statistical significance and biological relevance: A call for a more cautious interpretation of results in ecology. *Acta Oecologica*, *34*(1), 9–11. https://doi.org/10.1016/j.actao.2008.02.004

Mason, T., Popovic, G., McGillycuddy, M., & Keith, D. (2023). Effects of hydrological change in fire-prone wetland vegetation: An empirical simulation. *Journal of Ecology*, *111*, 1050–1062. https://doi.org/10.1111/1365-2745.14078

McCulloch, C. E., & Neuhaus, J. M. (2006). *Generalized linear mixed models*. John Wiley & Sons. https://doi.org/10.1002/9780470057339.vag009.pub2

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*(sup1), 235–245. https://doi.org/10.1080/00031305.2018.1527253

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews*, *82*(4), 591–605. https://doi.org/10.1111/j.1469-185X.2007.00027.x

Nakagawa, S., & Freckleton, R. P. (2008). Missing inaction: The dangers of ignoring missing data. *Trends in Ecology & Evolution*, *23*(11), 592–596. https://doi.org/10.1016/j.tree.2008.06.014

Nakagawa, S., & Parker, T. H. (2015). Replicating research in ecology and evolution: Feasibility, incentives, and the cost-benefit conundrum. *BMC Biology*, *13*(1), 88. https://doi.org/10.1186/s12915-015-0196-3

Niku, J., Brooks, W., Herliansyah, R., Hui, F. K. C., Taskinen, S., Warton, D. I., & van der Veen, B. (2020). *gllvm: Generalized linear latent variable models*. https://CRAN.R-project.org/package=gllvm

Niku, J., Hui, F. K. C., Taskinen, S., & Warton, D. I. (2019). gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in R. *Methods in Ecology and Evolution*, *10*(12), 2173–2182. https://doi.org/10.1111/2041-210X.13303

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Parker, T., Fraser, H., & Nakagawa, S. (2019). Making conservation science more reliable with preregistration and registered reports. *Conservation Biology*, *33*(4), 747–750. https://doi.org/10.1111/cobi.13342

Pebesma, E., & Bivand, R. S. (2005). Classes and methods for spatial data: The sp package. *R News*, *5*(2). https://cran.r-project.org/doc/Rnews/

Pike, N. (2011). Using false discovery rates for multiple comparisons in ecology and evolution: *False discovery rates for multiple comparisons*. *Methods in Ecology and Evolution*, *2*(3), 278–282. https://doi.org/10.1111/j.2041-210X.2010.00061.x

Popovic, G., & Drobniak, S. (2023). *gordy2x/principles: Paper acceptance* (v1.0.0) [computer software]. *Zenodo*, https://doi.org/10.5281/zenodo.10141145

Purgar, M., Klanjscek, T., & Culina, A. (2021). Identify, quantify, act tackling the unused potential of ecological research. *EcoEvoRxiv*, https://doi.org/10.32942/osf.io/xqshu

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rice, D. B., & Moher, D. (2019). Curtailing the use of *preregistration*: A misused term. *Perspectives on Psychological Science*, *14*(6), 1105–1108. https://doi.org/10.1177/1745691619858427

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *71*(2), 319–392. https://doi.org/10.1111/j.1467-9868.2008.00700.x

Schielzeth, H., & Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology*, *20*(2), 416–420. https://doi.org/10.1093/beheco/arn145

Simpson, G. (2023). *CRAN task view: Analysis of ecological and environmental data*. Version 2023-04-05. https://CRAN.R-project.org/view=Environmetrics

Smith, A. N. H., Anderson, M. J., & Pawley, M. D. M. (2017). Could ecologists be more random? Straightforward alternatives to haphazard spatial sampling. *Ecography*, *40*(11), 1251–1255. https://doi.org/10.1111/ecog.02821

Stan Development Team. (2023). *RStan: The R interface to Stan* (R package version 2.21.8) [Computer software]. https://mc-stan.org/

Steel, E. A., Kennedy, M. C., Cunningham, P. G., & Stanovick, J. S. (2013). Applied statistics in ecology: Common pitfalls and simple solutions. *Ecosphere*, *4*(9), art115. https://doi.org/10.1890/ES13-00160.1

Stewart, P. S., Stephens, P. A., Hill, R. A., Whittingham, M. J., & Dawson, W. (2023). Model selection in occupancy models: Inference versus prediction. *Ecology*, *104*(3), e3942. https://doi.org/10.1002/ecy.3942

Textor, J., van der Zander, B., Gilthorpe, M. S., Liśkiewicz, M., & Ellison, G. T. H. (2017). Robust causal inference using directed acyclic graphs: The R package 'dagitty'. *International Journal of Epidemiology*, dyw341. https://doi.org/10.1093/ije/dyw341

The Turing Way Community. (2023). The Turing Way: A handbook for reproducible, ethical and collaborative research. *Zenodo*, https://doi.org/10.5281/zenodo.3233853

Tikhonov, G., Opedal, Ø. H., Abrego, N., Lehikoinen, A., de Jonge, M. M. J., Oksanen, J., & Ovaskainen, O. (2020). Joint species distribution modelling with the r-package Hmsc. *Methods in Ecology and Evolution*, *11*(3), 442–447. https://doi.org/10.1111/2041-210X.13345

2041210x, 2024, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14270 by Area Sistemi Dipart & Document, Wiley Online Library on [02/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Tikhonov, G., Ovaskainen, O., Oksanen, J., de Jonge, M., Opedal, O., & Dallas, T. (2020). *Hmsc: Hierarchical model of species communities* [computer software]. R package version 3.0-9.

Wang, Y., Naumann, U., Wright, S., & Warton, D. (2012). *mvabund: Statistical methods for analysing multivariate abundance data*.

Warton, D. I. (2022). *Eco-stats: Data analysis in ecology: From t-tests to multivariate abundances*. Springer International Publishing. https://doi.org/10.1007/978-3-030-88443-7

Warton, D. I., Wright, S. T., & Wang, Y. (2012). Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution*, *3*(1), 89–101.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA Statement on *p*-values. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "$p < 0.05$". *The American Statistician*, *73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. https://ggplot2.tidyverse.org

Williams, B. K., & Brown, E. D. (2019). Sampling and analysis frameworks for inference in ecology. *Methods in Ecology and Evolution*, *10*(11), 1832–1842. https://doi.org/10.1111/2041-210X.13279

Yang, Y., Sánchez-Tójar, A., O'Dea, R. E., Noble, D. W. A., Koricheva, J., Jennions, M. D., Parker, T. H., Lagisz, M., & Nakagawa, S. (2023). Publication bias impacts on effect size, statistical power, and magnitude (type M) and sign (type S) errors in ecology and evolutionary biology. *BMC Biology*, *21*(1), 71. https://doi.org/10.1186/s12915-022-01485-y

Zuur, A. F., & Ieno, E. N. (2016). A protocol for conducting and presenting results of regression-type analyses. *Methods in Ecology and Evolution*, *7*(6), 636–645. https://doi.org/10.1111/2041-210X.12577

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems: *Data exploration*. *Methods in Ecology and Evolution*, *1*(1), 3–14. https://doi.org/10.1111/j.2041-210X.2009.00001.x

How to cite this article: Popovic, G., Mason, T. J., Drobniak, S. M., Marques, T. A., Potts, J., Joo, R., Altwegg, R., Burns, C. C. I., McCarthy, M. A., Johnston, A., Nakagawa, S., McMillan, L., Devarajan, K., Taggart, P. L., Wunderlich, A., Mair, M. M., Martínez-Lanfranco, J. A., Lagisz, M., & Pottier, P. (2024). Four principles for improved statistical ecology. *Methods in Ecology and Evolution*, *15*, 266–281. https://doi.org/10.1111/2041-210X.14270

2041210x, 2024, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14270 by Area Sistemi Dipart & Document, Wiley Online Library on [02/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License