The SORTEE Guidelines for Data and Code Quality Control in Ecology and Evolutionary Biology

Joel L. Pick[1]*, Bethany J. Allen[2], Benedicte Bachelot[3], Kevin R. Bairos-Novak[4], Jack A. Brand[5,6], Barbara Class[7], Tad Dallas[8], Pietro B. D'Amelio[9], Erola Fenollosa[10], Esteban Fernández-Juricic[11], Dylan G. E. Gomes[12], Matthew J. Grainger[13], Thomas Guillemaud[14], Christian John[15], Ruby Krasnow[16], Malgorzata Lagisz[17,18], Sebastian Lequime[19], Daniel S. Maynard[20], Shinichi Nakagawa[18], Rose E. O'Dea[21], Matthieu Paquet[22], Quentin Petitjean[23], Alfredo Sánchez-Tójar[24], Natalie E. van Dis[25,26], Laura A. B. Wilson[17,27], Edward R. Ivimey Cook[28]*

*Corresponding authors: joel.l.pick@gmail.com and e.ivimeycook@googlemail.com

[1]Institute of Ecology and Evolution, University of Edinburgh, Edinburgh UK
[2]GFZ Helmholtz Centre for Geosciences, Potsdam, Germany
[3]Oklahoma State University, OK, USA
[4]Australian Institute of Marine Science, PMB 3, Townsville MC, QLD, 4810, Australia
[5]Department of Wildlife, Fish, and Environmental Studies, Swedish University of Agricultural Sciences, Umeå 907 36, Sweden
[6]Institute of Zoology, Zoological Society of London, London NW1 4RY, UK
[7]Direction pour la Science Ouverte (DipSO), INRAE, France
[8]University of South Carolina, SC, USA
[9]Department of Biology, Reed College, Portland, Oregon, 97202, USA
[10]Department of Biology, University of Oxford, UK
[11]Purdue University, USA
[12]Marine Reserves, Oregon Department of Fish and Wildlife, Newport, OR, 97365, USA
[13]Knowledge Synthesis Department, Norwegian Institute for Nature Research (NINA), Trondheim, Norway
[14]Isa, Université Côte d'Azur, INRAE, Sophia-Antipolis, France
[15]Marine Science Institute, University of California, Santa Barbara. Santa Barbara, CA 93106 USA
[16]University of Maine, School of Marine Sciences, Orono, ME, USA

[17]Evolution & Ecology Research Centre, School of Biological, Earth & Environmental Sciences, University of New South Wales, Kensington, NSW, 2052, Australia
[18]Department of Biological Sciences, University of Alberta, Edmonton, AB T6G 2E9, Canada
[19]Cluster of Microbial Ecology, Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen, The Netherlands
[20]Department of Genetics, Evolution, and Environment, University College London, London, UK
[21]School of Agriculture, Food and Ecosystem Sciences, University of Melbourne
[22]SETE, Station d'Écologie Théorique et Expérimentale, CNRS, Moulis, France
[23]Abeilles et Environnement (UR406), Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE), Avignon, France.
[24]Department of Evolutionary Biology, Bielefeld University, Germany
[25]Organismal and Evolutionary Biology, University of Helsinki, P.O. Box 4, 00014 Helsinki, Finland
[26]Department of Animal Ecology, Netherlands Institute of Ecology (NIOO-KNAW), PO Box 50, 6700 AB Wageningen, The Netherlands
[27]School of Archaeology and Anthropology, The Australian National University, Acton ACT 2601, Australia
[28]University of East Anglia, Norwich, UK

Abstract

Open data and code are crucial to increasing transparency and reproducibility, and in building trust in scientific research. However, despite an increasing number of journals in ecology and evolutionary biology mandating for data and code to be archived alongside published articles, the amount and quality of archived data and code, and subsequent reproducibility of results, has remained worryingly low. As a result, a handful of journals have recruited dedicated data editors, whose role is to help authors increase the overall quality of archived data and code. There is, however, a general lack of consensus around what a data editor should check, how to do it, and to what level of detail, and the process is often vague and hidden from readers and authors alike. Here, with the input from multiple data editors across several journals in ecology and evolutionary biology, we establish and describe the first standardised guidelines for Data and Code Quality Control on behalf of the Society for Open, Reliable, and Transparent Ecology and Evolutionary Biology (SORTEE). We then introduce the SORTEE-led guidelines as a flexible six-stage framework that journals can implement incrementally and/or apply on a case-by-case basis, particularly when some checks (e.g., computational reproducibility) are not feasible (e.g., proprietary software). We conclude with practical advice for journals and authors, arguing that flexible adoption of these standardised guidelines will improve the consistency and transparency of the data editor process for readers, authors, data editors, and the wider scientific community.

Keywords: Data sharing, Code Sharing, Computational Reproducibility, Open Science, Data Re-use, Methodological Rigor, FAIR principles, Transparency, Data Editor

Introduction

A major focus of open science efforts in the last two decades, especially in ecology and evolutionary biology, has been open data, and more recently, open code. Open data and open code refer to the public archiving of the data and code associated with published research. The goals and societal benefits of data and code archiving have been widely discussed (e.g., Parr & Cummings 2005, Barnes 2010, Molloy 2011, Wilkinson *et al* 2016, Goldacre 2019, Gomes *et al* 2022, Ivimey-Cook *et al* 2023; see Box 1). Despite some reticence about open data and code (see Gomes *et al* 2022 for an overview of these fears), previous work has shown that data archiving is supported by the majority of academics in ecology and evolutionary biology, who perceive that the benefits outweigh any costs (Soeharjono & Roche 2021). Indeed, the two most important issues to the more than 1,000 members of the Society for Open, Reliable, and Transparent Ecology and Evolutionary Biology (SORTEE) from 2021-2025 have consistently been open data and open code (SORTEE 2026).

BOX 1: Goals of data and code archiving

From an idealistic perspective, data and code archiving has three main goals: to allow data reuse, to increase transparency, and to provide computational reproducibility (components that have been highlighted previously, e.g., Wilkinson *et al* 2016). We cover each of these in turn below.

A. Allow data reuse

The main focus of data archiving in the past has been on ensuring the potential for data reuse. Most prominently, this has included the development of the FAIR principles (Findable, Accessible, Interoperable, Reusable; Wilkinson *et al* 2016). There are several key motivations for this. First, data archiving prevents data loss. Data is typically collected using public money, and so can be seen as a public good that should not be lost when a researcher leaves their job or when a computer is lost/broken. Loss of data (and code) represents a massive source of research waste (Purgar *et al*. 2022). Second, archived data better allows for research synthesis (Culina *et al*. 2018, Hennessy *et al* 2022). Meta-analysis plays a key role in generalising results across systems, however, published

results are often not described in enough detail to be included in a meta-analysis. Provision of the data allows re-analysis to gain the information a meta-analyst needs and can substantially increase the data available for synthesis (Kim *et al* 2021). Third, methods are frequently updated, and data archiving allows for data to be re-analysed when new methods become available. Finally, data archiving allows for new questions to be asked with existing data. This is not only an efficient use of time and money, but reduces the use of animals in experiments, a central tenet of animal ethics (Janssens *et al* 2023).

To achieve the goal of allowing data reuse, it is essential that data and their accompanying metadata (see Table 1) are both present and in a form that allows for re-use. However, archived data are often incomplete or not in a state where they can be reused (Roche *et al* 2015, Roche *et al* 2022). Helping authors to ensure their archived data are FAIR should therefore be a key goal for data editors.

B. Increase transparency

Another key goal of data and code archiving is to increase transparency. Transparency involves making the research process visible, and is associated with building trust and credibility in science (Vazire 2017). While trust in scientists remains high globally, even a small minority's distrust can shape how research findings are received by the public and decision-makers (Cologna *et al* 2025), reinforcing the need for researchers to actively build and strengthen trust both with other academics and with the general public.

Analyses in ecology and evolutionary biology are becoming increasingly complex (Touchon & McCoy 2016, Feng *et al*., 2020), and there are often several ways to perform an analysis, with varying outcomes (Gould *et al* 2025). Descriptions of data filtering, processing, and analysis included within articles are not always sufficient to fully reproduce analyses (Archmiller *et al* 2020; Minocher *et al*. 2021; see Table 1). Provision of analysis code alongside a manuscript would therefore allow the analytical methods used to be directly assessed by reviewers when provided during peer review (Fernández-Juricic 2021) and by the general readership upon publication. Transparent methods allow work to be built on more easily, making science more efficient. Unlike data, the goal of code archiving in empirical articles is rarely direct code reuse (if this is the goal, then it is often more appropriate to create software packages, to allow generalisation of a method). Code archiving instead allows software code to be used for reference, adapted for new use, to allow similar methods to be applied to new datasets, or for methods to be extended.

Data and code archiving also allow mistakes to be found. Coding mistakes are easily made, and whilst many may have negligible effects on the results of an article, some will have major effects (Gihawi *et al* 2023; Mandhane 2024). The availability of the data and code that support an article make it possible for these mistakes to be found, and importantly, corrected in the future (Bolnick & Paull 2016, Delgado Manzanedo *et al* 2021). Finally, following several high profile cases of academic fraud (e.g., Viglione *et al* 2020, https://retractionwatch.com/2017/05/01/remarkable-ever-accepted-says-report-science-retract-study-fish-microplastics/, https://retractionwatch.com/2022/08/09/science-retracts-ocean-acidification-paper-more-than-a-year-after-a-report-on-allegations-in-its-own-pages/,), it has become increasingly clear that, as a community, we would benefit from a higher degree of transparency in how the results of published articles are generated. Although the provision of data and code will not stop data fabrication, publicly available data makes detecting data fabrication much easier as the data can be scrutinised and presents an additional hurdle to generating fraudulent results. As data and code are increasingly provided alongside journal articles in our field (although more so for data than code; Culina *et al* 2020; Kimmel *et al* 2023; Sánchez-Tójar *et al* 2025), not providing these resources leads readers to ask 'What are the reasons why the authors did not want to share their data and code?' (see Gomez *et al*, 2023).

To fulfil the goal of transparency, readers therefore need to *see* what has been done to obtain the reported results. This requires the presence of all data and code needed to reproduce the results presented in an article, as well as linking what was done in that article with the structure and form of both the data and code (using metadata, and appropriate code annotation).

C. Provide computational reproducibility

Perhaps the most ambitious goal of data and code archiving is computational reproducibility, which ultimately builds trust and credibility in published results (Power and Hampton 2018, Reinecke *et al* 2022). We can define computational reproducibility as '*obtaining consistent results using the same input data, computational methods, and conditions of analysis*' (National Academies of Sciences, Engineering, and Medicine 2019). Although this definition of computational reproducibility is commonly cited, the terms within it are not clearly defined. In the context of a research article, we interpret this to mean that, given the available data (i.e. input data) and code or workflow (i.e. computational methods), and using the same software versions (and hardware if appropriate) outlined in the article and metadata (i.e. conditions of analysis), we should be able to reproduce the results presented in the article.

In some cases, exact reproducibility is difficult to achieve (i.e. generating the *exact* numbers presented in an article), and often computational reproducibility is assessed with some tolerance level (e.g., Archmiller *et al* 2020; Kambouris *et al* 2024). However, practices such as having appropriate metadata that adequately describes the software and package versions, and setting of seeds (where the same pseudorandom numbers are generated each time) within code for stochastic methods, will help to achieve these goals. We should note that it is unlikely we will be able to demonstrate computational reproducibility in cases where analyses are highly computationally intensive, but solutions exist. We discuss these points in more detail in the main text.

To achieve the goal of computational reproducibility, as a minimum requirement, we need all code and data to be present. The next component of computational reproducibility is that the provided code runs without error. This requires the code to be explicit about what data files it uses and where these are located, it requires data files to be in the same directory structure and with the same names as expected by the code, and it requires the same version of all software packages used in the code to be loaded. Any code that cannot be rerun without error in a clean workspace cannot be considered computationally reproducible.

In response to this call for open data, starting in 2010, many journals in ecology and evolutionary biology began to mandate data archiving, e.g., Journal of Animal Ecology, Functional Ecology, and Heredity, to name a few (for a full list see the Joint Data Archiving Policy; https://doi.org/10.25504/FAIRsharing.z67ht2). As a result, an increasing number of journals in ecology and evolutionary biology have mandatory open data policies (estimated to be 20% out of 196 journals in 2020, 35% in 2023; Berberi and Roche 2022, Berberi and Roche 2023 and 41% in 2024; Ivimey-Cook *et al* 2025). This action has resulted in a large increase in the number of publications in ecology and evolutionary biology having open data (Vines *et al*. 2013; Culina *et al* 2020 found 79% in 14 journals that have a code archiving policy with no change from 2015/16 to 2018/19; Sanchez-Tojar *et al*. 2025 found 37% in 12 journals without code archiving policies with an increase over time; Kimmel *et al* 2023 found 78.5% in 5 journals

97 from 2018-20; Belkhir *et al*. 2025 49% in 110 journals in 2024). Compared to many other fields,
98 ecology and evolutionary biology are at the forefront of data sharing (Tedersoo *et al*. 2021).
99
100 However, despite a high proportion of ecology and evolutionary biology studies archiving data,
101 archived data are often of low quality (Roche *et al* 2015, 2022), with most datasets either
102 incomplete (some or all of the data allowing the study to be reproduced is not present) or
103 unusable (e.g., data are not machine readable, in a proprietary format, or are archived with no
104 metadata; see Table 1). Based on 362 open datasets from 2013-2019, Roche *et al* (2022)
105 calculated that 56.4% of datasets were complete, and 45.9% were reusable (out of 362), a
106 situation that has only marginally improved over the last decade (from a sample of 100 articles
107 in 2012/13, 44% were complete, and 36% reusable; Roche *et al* 2015), with only reusability
108 having statistically increased from 2013 to 2019 (Roche *et al* 2022). Several studies have
109 further sought to directly assess analytical reproducibility (defined as reproducing the published
110 results using the same data). However, these assessments rely heavily on data provided by
111 authors upon request (Archmiller *et al* 2020; Minocher *et al* 2021), as rates of archived data
112 recovery were low (11% in Minocher *et al*. 2021). Conditional on having the full dataset,
113 reproducibility was moderate (42% and 58% of articles were fully reproducible in Archmiller *et al*
114 2020 and Minocher *et al* 2021, respectively), but whether the quality of data provided directly
115 from authors for these studies differs from data that has been archived is not clear. This is
116 similar to other fields; in the *Journal of Psychological Science*, only 9 out of 25 articles were
117 reproducible (given the data) without author intervention (2014/2015; Hardwicke *et al* 2021) and
118 even when journals have mandated data archiving only 62% (85/136) of datasets were reusable
119 in the journal *Cognition* (2015/2016; Hardwicke *et al*, 2018). Overall, these results suggest that
120 most studies across fields either do not have archived data or provide a dataset that has limited
121 utility, but when full datasets are provided, reproducibility can be high. The lack of high quality
122 data impedes all the goals of data archiving (see Box 1).

The use of code for data preparation and analysis is now almost ubiquitous, particularly using the R coding language (Lai *et al* 2019, R Core Team 2022). Increasingly, journals encourage or mandate code archiving (15% in 2015 Mislan *et al* 2015; 75% in 2020 Culina *et al* 2020; 88.4% in 2024 Ivimey-Cook *et al* 2025). However, the actual rates of code archiving still remain low (2015-2016: 2.5%, 2018-2019: 7.0% in journals without a code archiving policy; 23% in 2015/16 to 30% in 2018/19 in journals that encourage or mandate, Culina *et al* 2020; 18% 2018-2020 in 5 major ecology journals, Kimmel *et al* 2023). Several recent studies have further tried to assess computational reproducibility (the ability to reproduce the results given the archived data and code; see Box 1), but have concluded that it is likely to be low in ecology and evolutionary biology. Kambouris *et al* (2024) found that out of 177 meta-analyses in ecology and evolutionary biology from 2015-17, only 26 provided both data and code. From these, only 7 studies (27%) could be exactly reproduced (with the results of 15 (58%) studies being reproduced to within 10% of the original results). Kellner *et al* (2025) found 7% of 497 articles on species distribution and abundance from 2018-2022 had code that ran. Trisovic *et al* (2022) found that out of 9,000 unique R files from the Harvard Dataverse, 74% failed to complete, which lowered to 56% when basic cleaning was applied. The lack of high quality code impedes all the goals of transparency and computational reproducibility (see Box 1). Together, the lack of functional data *and* code in public repositories limits the verifiability of published empirical research claims (Henderson *et al* 2024) and ultimately erodes trust in science.

Alongside several high profile fraud scandals (see Box 1), the lack of adherence to journal archiving policies, and the low quality of data and code archiving has led several journals to recruit data editors (https://www.amnat.org/announcements/data-and-code-announcement.html, Thrall *et al*. 2023, Barrett, 2024, Barrett & Montgomerie 2025). Data editors are responsible for screening and assessing the archived data and code of manuscripts being reviewed by a

journal, to assist authors in complying with journal mandates on data and code provision - hereafter, we refer to this process as *data and code quality control*. It is worth stressing that data editors are not acting as gatekeepers; the role of a data editor is to help authors adhere to community standards of data and code archiving. At the time of writing, we are aware of seven journals in ecology and evolutionary biology that have data editors that screen the data and code of some or all manuscripts that are published (American Naturalist, Behavioral Ecology, Ecology Letters, Ethology, Journal of Evolutionary Biology, Proceedings of the Royal Society B, and Peer Community Journal). Behavioural Ecology and Sociobiology has an editor with a related role, but only screens a small number of manuscripts at the request of other editors, and additionally provides statistical support (Fernández-Juricic *pers. comms.*).

Data and code quality control by data editors is primarily for the benefit of the authors. A large part of the data editors' role is to help authors ensure their data and code more closely adhere to the open principles that we have adopted as a research community. At the end of the process, authors will therefore have higher quality archived data and code for each publication, which has been linked to increased citation rates (Piwowar *et al* 2007, Piwowar and Chapman 2010, Christensen *et al*. 2019, Maitner *et al*. 2024) and increases the prospect of future collaboration based on using and developing archived data and code. Having well-archived and documented code also provides a clear advantage for Early Career Researchers that may pursue careers outside of academia, where a proven ability to generate reproducible code is often more of a selling point than publications (Allen and Mehler 2019, König *et al* 2025). There are many benefits to working reproducibly (Markowetz 2015), from helping with the continuation of research, to avoiding errors that could later influence results and ultimately require correction or retraction of published work. Data and code quality control further forces authors to be doubly sure that their dataset is accurate and that the code they used generates the expected results.

While authors benefit most, we believe there are also a multitude of additional benefits for journals, readers and the wider research community. For instance, it is in a journal's best interest to be the purveyor of high quality and reliable scientific research. Adopting data and code quality control signals to readers and the general community that scientific quality and transparency are priorities for the journal. By increasing transparency of analyses, data and code quality control can allow a journal to build its reputation as a reliable and trustworthy source of high-quality science. Through the provision of higher quality data and code, quality control increases the impact of both the article being evaluated, and the journals in which the manuscript is published. Furthermore, it allows other authors and researchers to extend and reuse data and code for further analyses, which can lead to an extended impact for both the original journal and author(s). Finally, ensuring the archiving of high-quality data and code facilitates rigorous post-publication evaluation of claimed results. By increasing transparency and reproducibility of analyses, data and code quality control will therefore increase the trust of published work.

We hope that an emphasis on the quality of archived data and code may additionally help to facilitate data and code review (the detailed evaluation of code; Ivimey-Cook *et al* 2023) within research groups prior to submission, creating more opportunities to actively involve co-authors in a study and resulting in a robust and healthy lab culture that promotes cohesion. Such practices can inspire early-career researchers involved in the study by promoting open science through practical experiences, while helping to strengthen trust in scientific integrity in the long term.

In this article, we outline what data editors are, discuss the costs and benefits of data and code quality control, and then provide detailed guidelines for data editors that can be used for data and code quality control in journals.

201

202 **Table 1.** Glossary of key terms

Term	Definition
Metadata	Refers to a description and information about the data and code. Typically in the form of a text file called a README. Other variations on this are possible, e.g., a codebook or a data dictionary.
Data Editor	An editorial position at a journal. The responsibility of this editor is to screen and quality control the data and code that will be publicly archived alongside manuscripts under review at the journal.
Data and Code Quality Control	The process of checking the suitability of data and code for public archiving.
Data and Code Archiving	The process of depositing data and code in a public repository.
FAIR principles	Findable, Accessible, Interoperable, and Reusable principles for Data (Wilkinson *et al* 2016) and Code (Barker *et al*. 2022). See https://www.go-fair.org/fair-principles/
Raw data	Unprocessed and unfiltered data. This would include any raw files e.g., photos, audio recordings, videos, and data sheets.
Data Filtering	The process of removing some data to create the dataset used in the analysis (e.g., removal of individuals with a certain characteristic). We refer to the resulting data as *Filtered Data.*
Data Processing	Transforming data from one form to another. Includes data that is extracted from images or videos, data that has been summarised, transformed, or is the result of calculations. We refer to the resulting data as *Processed Data.*

Repository	A framework providing long term storage for many projects, such as Zenodo, Dryad, Figshare, etc
Project	A collection of files archived for a specific manuscript (note, this is similar to what GitHub refer to as a repository)

What is Data and Code Quality Control and What is it Not?

Data and code quality control by data editors is about increasing the quality of the archived data and code, and ensuring that they meet minimum standards (e.g., the data are complete and usable, and the code is documented and runnable). The guidelines we lay out in the section below give a detailed explanation of what data and code quality control by data editors entails. Ideally, data editors would ensure that archived data and code achieve the goals laid out in Box 1, to allow data reuse, to increase transparency, and to provide computational reproducibility. The importance of these goals may vary across different groups; journals will likely focus more on transparency, whereas readers may be more concerned with data reuse and computational reproducibility. These goals require varying levels of code and data checking, and so, during the early stages of journals recruiting data editors, not all of these goals may be achievable. We also acknowledge that not everyone may agree that achieving all of these goals is the ultimate objective of data and code quality control by data editors. In practice, journals may prioritise and implement goals aligned with their editorial policy over time, while allowing the data editor to apply them flexibly on a manuscript-by-manuscript basis depending on the type of data and tools used by the authors. While the importance of each goal may differ among stakeholders, they each help improve the openness, reliability, and transparency of the scientific publication process.

Importantly, whilst data editors are responsible for checking that the archived data and code meet certain minimum standards, they are not responsible for *reviewing* data and code, and so data editors will rarely themselves detect errors or fraud. Data quality control is not about verifying the actual data (e.g., detecting data fabrication) but rather ensuring that data is available, in the appropriate format, and has the corresponding metadata to be scrutinised. The presence of a data editor at a journal will therefore not necessarily prevent fabricated data being published. We also make a clear distinction here between code *quality control* and code *review* (Ivimey-Cook *et al* 2023, Hillemann *et al* 2025). Code review is the detailed evaluation of code including assessing factors such as alignment between the code's intended or stated purpose and its actual implementation, the consistency of coding style, and the efficiency of the code, that goes beyond the task of a data editor. Code review is an important part of research (Ivimey-Cook *et al* 2023) and we encourage that research groups engage in this practice as a way to improve the quality of published research (Bavota and Russo 2015). However, data editors are not experts in every field of study, nor are they statisticians or specialists in all programming languages. Therefore, data and code quality control should not extend to assessing the suitability of analyses or code itself.

Are there Costs to Data and Code Quality Control?

Although journals adopting data and code quality control will increase the quality of data and code archiving associated with published articles, which we believe will have widespread benefits to journals, authors, readers and the wider research community (see above), we acknowledge there may be some costs to the widespread adoption of this process.

First, adopting data and code quality control may present an additional burden for a journal. This is likely to primarily impact the length of time required for peer review. To mitigate this problem,

several journals currently have data and code quality control alongside peer review (see *Suggestions for Journals* at the end of this article). For journals that currently have in-house data editors, data and code quality control does not create a per-manuscript burden to find extra reviewers.

The process of data and code quality control may add a time burden to authors (although not if authors were already adhering to many journal's existing requirements on data and code archiving). However, this time burden will reduce over time as data and code quality control becomes standard practice and making well-documented data and code becomes a natural part of a researcher's workflow. This short-term investment will also come with both short and long-term benefits as outlined above. Increasingly, data and code management skills have wide applicability and are becoming part of routine teaching at undergraduate or postgraduate level (Kohrs *et al* 2023). We acknowledge that the costs to authors will fall disproportionately on those with less access to training on open data and code practices. However, these researchers are actually those that may benefit most from the process of data and code quality control, which is designed to aid researchers adhere to data and code archiving requirements. Those with lower access to training therefore stand to gain the most from interactions with data editors and the resulting skills learned from increasing the quality of their data and code archiving.

We acknowledge that researchers with higher access to training are more likely to be those appointed as data editors. From one perspective, this can be seen as reinforcing existing hierarchies, however, it can also be argued that these more privileged individuals should take on the burden of these service roles to help increase access to this training across diverse groups.

Finally, ensuring that all data and code are archived and checked for computational reproducibility will necessitate more resources for data storage and re-running potentially computationally expensive analyses. Data storage is already reported as a leading cause of increased carbon footprint (https://direct.mit.edu/imag/article/doi/10.1162/imag_a_00043/118246/Ten-recommendations-for-reducing-the-carbon). However, we would argue that to prevent data loss, the data and code behind any study should always be responsibly archived, regardless of the process of data and code quality control. As we outline in the guidelines below, we also do not advocate for the storage of multiple instances of the data, if it is already stored in a public archive. Re-running analyses will also have an environmental impact. However, there is a limit to what data editors are reasonably expected to re-run, and so it is unlikely that highly computationally expensive analyses will routinely be repeated.

SORTEE Guidelines for Data and Code Quality Control

At the time of writing, the practice of data and code quality control is highly variable both across and within journals. To address this, the Society for Open, Reliable and Transparent Ecology and Evolutionary biology (SORTEE) started a working group with 22 data editors from 6 journals in Ecology and Evolutionary (EE) biology, comprising American Naturalist, Behavioural Ecology and Sociobiology, Ecology Letters, Journal of Evolutionary Biology, Peer Community Journal, and Proceedings of the Royal Society B. The goal of this group was to propose a set of structured guidelines for standardising the process of data and code quality control across all EE journals. As a whole, these guidelines provide a high bar for data and code quality control, and so either all or parts of these guidelines can be adopted by journals and presented to authors and readers. We note that validation of Stages 1-4 of these guidelines *by data editors*

achieves the highest level (Level 3) of the Transparency and Openness Promotion Guidelines (TOP2025; Grant *et al* 2025), for data and code.

We see several benefits of adopting a set of standardised guidelines both for authors and data editors. First, authors know what is expected of them regardless of the journal. By working towards the same standards of data and code archiving in advance of submission, authors can easily submit their manuscripts to a variety of journals, or transfer their manuscripts between them, and know that they do not need to change their submissions to meet different standards across journals. Ultimately, this reduces the burden for authors and streamlines the submission process. In addition, standardisation of this workflow will encourage authors to share data and code even when not explicitly required, help those with less experience of sharing data and code, and facilitate the community-wide adoption of open data and code as a default. Second, data editors have a standardised template for review, designed by both open science advocates and active and experienced data editors from multiple journals, to gain the balance of idealism and practicality. This can make the process both more efficient for data editors and more thorough for the community. Ultimately, standardisation will help inform decision-making for the journals. Third, readers know what checks have occurred prior to publication, which can ultimately help build additional trust in scientific reporting and open science practices, particularly when computational reproducibility has been assessed.

In an ideal scenario, data and code quality control helps achieve the goals of data reuse, transparency, and computational reproducibility (see Box 1). In reality, depending on the time, computing resources, and expertise of the reviewing data editor, only some of these will be feasible at a large scale. We propose that data and code quality control can be broken into six stages in the following coherent order to sequentially address the goals of data and code archiving outlined above (Figure 1):

- Stage 1: Data must be archived and adhere to FAIR guiding principles.
- Stage 2: Archived data corresponds with the data reported in the manuscript.
- Stage 3: Code must be archived and adhere to FAIR guiding principles.
- Stage 4: Archived code corresponds with the workflow reported in the manuscript.
- Stage 5: Archived code runs with the archived data.
- Stage 6: Results can be computationally reproduced by running the archived code.

We discuss each of these stages in more detail below and provide guidelines for how these can be assessed. Note, each component of each stage is associated with its own guideline. To help data editors perform their assessment according to these guidelines, we have created an app, that can be accessed at https://github.com/SORTEE/DC.

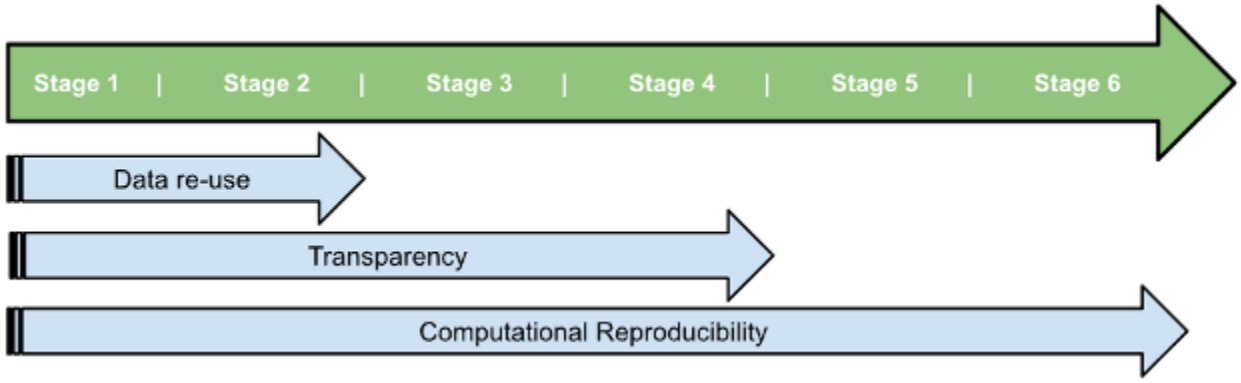

Figure 1. *A diagrammatic representation of how the three goals of data and code archiving (blue arrows) match the stages of data and code quality control (green arrow). Stages 1 and 2 (1 = Data must be archived and adhere to FAIR guiding principles and 2 = Archived data corresponds with the data reported in the manuscript) are needed to achieve the goal of data re-use. Stages 1-4 (3 = Code must be archived and adhere to FAIR guiding principles. and 4 = Archived code corresponds with the workflow presented in the manuscript) are needed to achieve the goal of transparency. Stages 1-6 are needed to achieve the goal of full computational reproducibility (5 =Archived code runs with the archived data and 6 = Results can be computationally reproduced by running the archived code). Figure created by EIC.*

Stage 1. Data must be archived and adhere to FAIR guiding principles.

For data to be open and amenable for reuse, it must adhere to FAIR guiding principles (Findable, Accessible, Interoperable and Reusable). Data must be placed in an open repository

(*Accessible)* with a permanent Digital Object Identifier (DOI, or another globally unique and persistent identifier) that is cited in the manuscript (*Findable*) with a licence that describes re-use (*Reusable*). Metadata (e.g., a README) must also be present to describe the data (*Findable, Accessible* and *Reusable*). The data itself must be in a machine-readable, non-proprietary format (*Interoperable*). We discuss each of these points in more detail below. By a data editor assessing the archived data to Stage 2 of these guidelines, journals would achieve TOP 2025 level 3 for Data Transparency (Grant *et al* 2025).

Stage 1.1. Data files are accessible and in an open repository

For data to be readily Accessible and Findable, it must be archived in a public data repository with an associated persistent DOI (or any other globally unique and persistent identifier, such as ARK (for archives, and datasets), Handle (for digital objects), Accession Numbers ('omics' data, e.g., GenBank), RRID (research resources)) that is separate from the DOI of the resulting published article. Furthermore, the data must be clearly cited in the manuscript, and listed with its DOI (or other identifier) in the reference list, so readers know where to access the underlying data. There are a multitude of different repositories to fit a variety of needs (see https://doi.org/10.5281/zenodo.10651775 for information about repositories; Harvard Longwood Medical Area Research Data Management Working Group 2023). An important feature of a repository is that it guarantees long-term storage and file immutability (i.e., they cannot be deleted or modified once published). Common general repositories that provide these include DataVerse, Dryad, Figshare, and Zenodo. There are also additional topic specific repositories, such as GenBank for depositing genetic and other biological sequences. We note that whilst GitHub is popular, it does not produce DOIs and files can be changed or deleted, and so studies that use GitHub should be linked to a data repository for archiving prior to submission (e.g., Zenodo). Similarly, the Open Science Framework (OSF) does not provide file immutability; projects and files can be changed or deleted. Data should not be provided as supplementary

material attached to the manuscript online as this does not provide a globally unique persistent identifier and may not be open and accessible; all data must be archived in a public repository (see Stage 1.3. for sensitive data).

When necessary, projects in public data repositories can be anonymised to adhere to the journal's policy of double blinding (see: https://methodsblog.com/2023/08/23/double-anonymous-peer-review-frequently-asked-questions/). Furthermore, many repositories offer embargoes, if necessary.

Guideline 1.1: Data are open and freely available (with some exceptions, see Stage 1.3.) and are located in a permanent public repository with an associated globally unique persistent identifier, that is cited in the text and reference list of the manuscript .

Stage 1.2. Data are associated with a license

Data must be associated with an appropriate license that indicates how the data can be shared and re-used. In our experience, most researchers have little knowledge of such licenses. Without a license, data cannot be legally reused under many circumstances (e.g., depending on the jurisdiction; https://choosealicense.com/no-permission/). Therefore, to avoid confusion, it is important for authors to specify a license that outlines how their data can be used, and whether or not attribution is required. There are several different licenses to choose from but typically Creative Commons licenses are used for data (see: https://chooser-beta.creativecommons.org/), with several repositories including a license by default. The most permissive license is the CC0 license, which puts the data freely into the public domain with no requirement for attribution. Some repositories assign this license by default (e.g., Figshare) or mandate the use of this license (e.g., Dryad). Another commonly used license is CC-BY 4.0,

where reusers of the data must give credit to the original author but are allowed to distribute, remix, adapt, and build upon the created material (including for commercial uses). This license is also used as a default by some repositories (e.g., Zenodo). There are several other more restrictive licenses, for example the CC-BY-NC 4.0 which prohibits commercial use. Note that these guidelines do not recommend or enforce any specific type of license.

Guideline 1.2: Data must be associated with a license.

Stage 1.3. Data files are present and complete

The simplest but most important requirement is that the data supporting the results presented in a manuscript must be complete in the archived project. Ideally, raw data and processed data should both be provided (see Table 1). The term “raw data” refers to all collected data prior to any filtering (subsetting of data based on reported exclusion criteria) or processing (extraction, transformation, summarising, aggregation, and prior to any formal calculations; see Table 1). Note that we do not count transcribing data from a written to a digital format and subsequent error-checking of data entry errors as processing or filtering.

There are several reasons why the raw data should be archived: First, to prevent data loss, which is achieved by archiving the most complete dataset possible; Second, to maximize data re-use, as only providing filtered data can exclude particular future uses; Third, the process of filtering and processing data is prone to mistakes (e.g., coding errors). Such errors are a natural and inevitable part of the research process, but being able to detect them makes the scientific process more efficient and reliable, and identifying and correcting these mistakes is only possible if the raw data are available. Finally, to increase transparency, allowing the reader a

clearer insight into the process that resulted in the final dataset used for analysis. We note that the data required to be archived also depends on the goal; computational reproducibility of the results presented in a manuscript can be achieved with processed data, whereas the goal of data reuse is dependent on raw data being archived. What constitutes raw data is often reliant on the nature of the data (see below). Ultimately, whether the archived data are most appropriate for a given manuscript will be at the discretion of the data editor and dependent on journal policy. Below we provide some guidance for specific cases.

In the simplest case, data have been collected for a stand-alone study. In this case, the raw data are simply all the collected data, and should be provided in full (for exceptions see below). If data originates from videos, images or sound files, then these are considered to be raw data. Therefore, where possible, these files should also be made available. The processed data should be provided alongside the raw data, with a description of the processing/filtering in the metadata (if this is not already described in the code files). This is particularly important if the raw data are not interoperable (e.g., outputs from proprietary software; see 1.4 below). Although most databases allow a considerable amount of data to be stored (Per project: Dryad - 300 GB, Zenodo - 50 GB, Figshare - 20 GB, Dataverse - 10GB), raw data may exceed these limits (e.g., video data that is several terabytes large). In such cases where the data are too large to be feasibly uploaded, then a representative, manageable subset of this raw data should be provided, so the extraction process can be assessed (e.g., providing several example videos).

If the data used in an analysis originates from a larger database (e.g., from a long-term study) then ideally the entire database would be considered the raw data. We can foresee many circumstances where the authors may feel this is inappropriate, for example, due to worries about the data being used without permission (Mills *et al*. 2015, but see Evans 2016 for an empirical assessment) or misused (Weissgerber *et al*. 2024). In such cases, filtered data may

be provided, alongside clear details of the filtering process that would allow the same data to be extracted at a later point (e.g., location and version of the database that the data was extracted from, how it can be accessed, the database queries used to extract the data or other similar instructions of how to generate the same subset for analysis, and any exclusion criteria). Data editors may need to assess the suitability of archived data on a case-by-case basis to ensure that the data are provided in the rawest form possible according to the journal guidelines, and that sufficient information on the generation of archived datasets is present. If the database is already open, rather than re-archiving the data (which, if large, may come with environmental costs) the authors can cite the database, include a clear description of what data were used, the data extraction procedure, and where appropriate, provide an immutable snapshot of the database if it is dynamic.

In some cases, restrictions may apply to making raw data publicly available. For instance, if the dataset contains sensitive information about medical records, identifying personal information (which may breach General Data Protection Regulations (GDPR)), or geographic locations pertaining to endangered species or fossil sites at risk of vandalism (see Chapman 2020). In many cases, data can often be obfuscated or anonymised to enable data archiving. There may also be issues with indigenous data sovereignty (for best practices on governance and stewardship of indigenous data in combination with FAIR principles see CARE principles (Collective benefit, Authority to control, Responsibility, Ethics); Carroll *et al* 2020). Processed data used in the manuscript should instead be provided alongside suitable metadata which describes the raw data in as much detail as possible, while still preserving anonymity and sovereignty. Where data cannot be provided, simulated data with the same structure and properties could also be provided, to allow for Stage 5 to be assessed. Information about how and where to make data requests should also be included in the metadata.

In all cases, the data availability statement in the manuscript should clearly outline whether the authors have archived raw and/or processed data. This section should also contain guidance on how to access and request the raw data if necessary and appropriate.

Guideline 1.3: Authors must either provide:

a) raw data, along with the processed data and/or code to prepare the data for analysis, or

b) a sample of raw data alongside processed/filtered data when full raw data upload is not possible , or

c) processed/filtered data with a detailed description of how to both obtain and process/filter the raw data.

Stage 1.4. Data files are in an interoperable format

To both facilitate review and allow reuse, data must be in a universally interoperable format, meaning that the data can be exchanged and used across different software and operating systems. File types specific to proprietary software (e.g., .sps files from the SPSS program) are not interoperable, so do not facilitate data re-use. For example, .xls files are a proprietary format, whereas .xlsx files are not, meaning they are interoperable. However, .xlsx files can contain information that is lost when importing data into statistical software (e.g., formatting). Similarly, tabular data are sometimes archived in a .RData file (or equivalent). Although this can be used with open source software (i.e., R), again, this data format restricts its use, as it requires knowledge of R to extract the data, and may be dependent on the version of R that was used to save it. Simpler text-based file formats such as .csv (comma-separated-values), .tsv (tab-separated-values) and .txt (plain text) files provide a more interoperable format, as they can be used by more software and across more systems, and so are preferable. Where possible, it

would therefore be more suitable to archive the raw data in a more interoperable format (i.e., .csv or .txt). Lastly, data should not be stored within PDF or Word documents, which can be prone to error when data are copy-pasted (e.g., for re-using) and, which cannot be readily imported into statistical programs for analysis. In some cases, there might be no option other than to provide data in a non-interoperable format (e.g., if the data was collected using proprietary software) but this should be provided alongside interoperable extracted data with a clear description of the conversion processes in the metadata including the particular software version that was used. There are continuous advances in this area; for instance, parquet files are interoperable, highly compressed, efficiently read, and highly accurate for storing large datasets.

Guideline 1.4: Data files must be provided in an interoperable format.

Stage 1.5. Data metadata present and adequate

Data files alone do not contain enough information for a user to fully understand their contents. Data files must therefore be accompanied by metadata. The most common form of this metadata is a README file, which describes and explains the content of the data and its provenance. The README should provide general information about the manuscript, e.g., the manuscript title and abstract, the authors and relevant contact information, date and location of data collection, and a list of all relevant funders. In the case of double-blind review, some sections can be left blank until acceptance (see example in Figure 2). The README should also include any relevant licence information (e.g., CC-BY, see above), and information about data derived from other sources (e.g., from other articles or online data). Finally, the metadata should contain detailed descriptions of each data file, describing its structure and what variables it contains, what units of measurement they are in, and how they link to the data described in the

manuscript, e.g., each column in a .csv should be explained and described (see example in Figure 2). This information can be provided in several ways: 1) as part of the main README file, 2) by creating additional README files to describe the data and code files (as shown in figure 2), or 3) by providing a data dictionary for each data file (e.g., a .csv file with a column for column names, and another for the description of the variable). We use the term “adequate” here to describe data-associated metadata that is sufficiently detailed so that anyone can understand the data without needing to read the resulting manuscript to understand its contents.

Guideline 1.5: Detailed metadata, including (but not limited to) a README file, must accompany the data (see example in Figure 2).

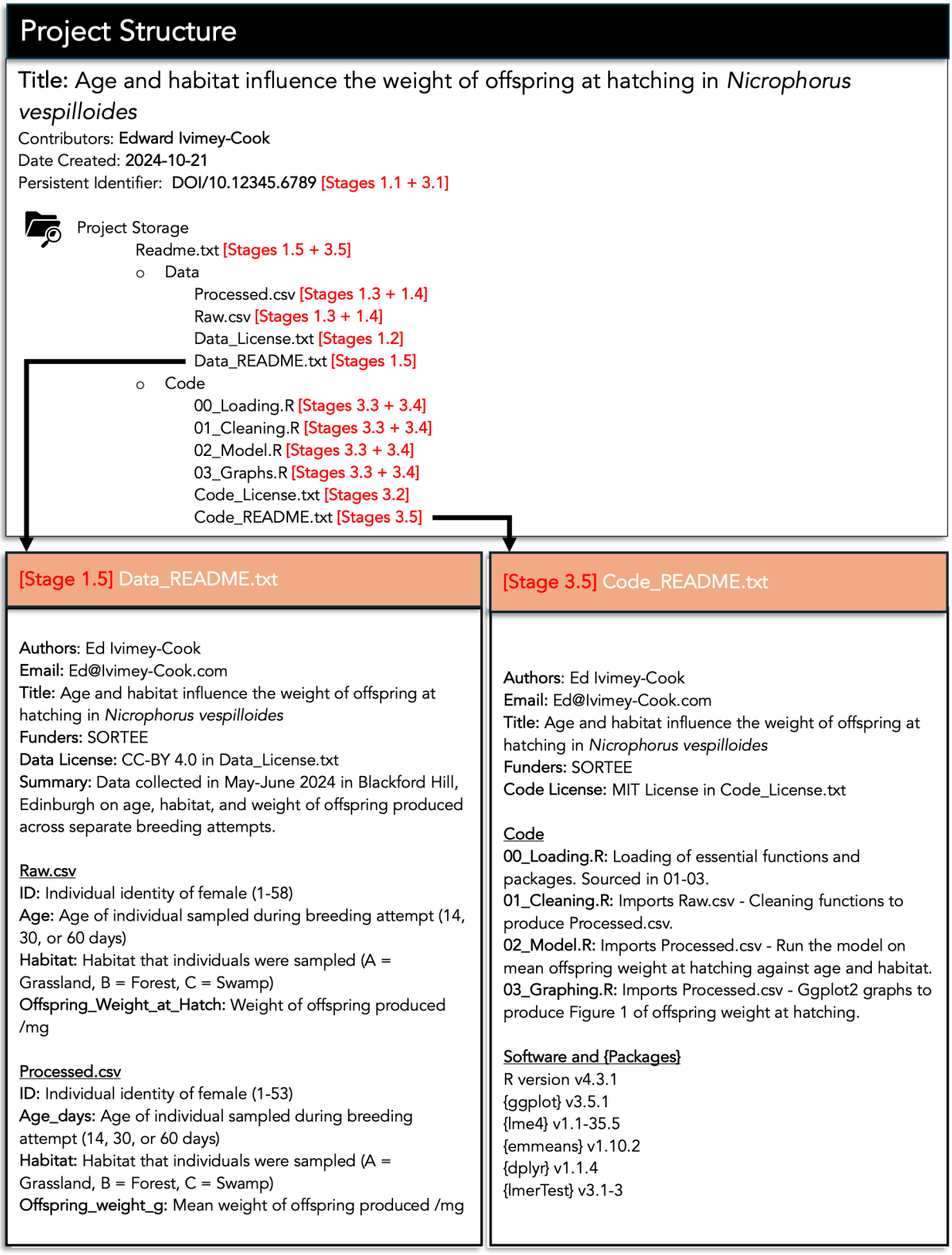

Figure 2. *Example project structure and metadata (two README files) showing how the various components adhere to the SORTEE guidelines for data and code curation. The numbers in red refer to*

the stages being addressed. The Data README should contain information on the manuscript (authors with corresponding contact details, the title of the manuscript, and any funders), the license file, along with information of the data (a brief summary of collection, and column-by-column description of the data files along with any measurement units or levels of factors). For code, the README should contain the same information as the data initially (information on the manuscript and code license) but also contain a description of each code file in the order they are meant to be used (which also clearly indicates which data file is used in each script). Lastly, the README should contain a list of all software and packages used with associated version numbers. Figure created by EIC.

Stage 2. Archived data corresponds with the data reported in the manuscript

For archived data to support a manuscript, as well as being present in a form that facilitates reuse, it must correspond with the data reported in the manuscript. For this to be assessed, the data editor needs to check that the variables and data described in the manuscript (most likely in the Methods) are present in the data files provided. For example, if the manuscript mentions that offspring weight was measured at three habitats, the data file should contain an offspring weight variable and a habitat variable (see example in Figure 3). The dimensions of the data should also correspond with those described in the manuscript; discrepancies in the size of the dataframe may suggest that some unreported data processing or filtering has taken place. A clear description of all these aspects within the text is essential; without it, the data will not correspond with the manuscript, undermining its potential for reuse, transparency and reproducibility. Some journals use AI to facilitate this process (e.g., the DataSeer.ai application: https://dataseer.ai/), which produces a report detailing the expected data that should be provided based on the description within the manuscript.

Guideline 2: The structure and contents of the archived data files must match the description in the manuscript.

Manuscript

Methods

We monitored a population of Nicrophorus vespilloides over three habitats (Grassland, Forest, and Swamp) on Blackford Hill, Edinburgh from May-June 2024. Over this time, we discovered 58 broods (23 Grass, 20 Forest and 15 Swamp), from which we took one larva and brought them back to the lab for rearing. We then monitored three breeding attempts (at 14, 30, and 60 days post-eclosion) for each of the resulting beetles. For each breeding attempt, we measured the hatching weight of offspring born to each brood (in mg). Five individuals died during the experiment and so we removed them from the dataset before analysis. In total we analysed records of 159 broods (53 females each with three broods). To test whether age and habitat affected hatching weight, we first took the mean weight of offspring at hatching per individual per age class. We then modelled hatching weight as a function or age (continuous) and habitat (factor), including a random factor of individual ID in a linear mixed model. [Stages 2 + 4]

Results

The effect of age on offspring weight was significantly negative (Age = -0.03, SE = 0.004, $p<0.001$). [Stages 5 + 6]

[Stages 2 + 4] Processed.csv

	ID	Habitat	Age_days	Offspring_Weight_g
2	1	A	14	0.22
3	1	A	30	0.33
4	1	A	60	0.21
5	2	B	14	0.28
6	2	B	30	0.24
7	2	B	60	0.22
...	...	...	...	...
160	53	C	60	0.34

[Stages 2 + 4] 02_Model.R

```
####02.1 Modelling code

#Load packages
source("00_Loading.R")

#Import processed data
Processed_data <-
read_csv("Processed.csv")

#Lmer of offspring weight with age and
habitat #with a random effect individual
ID - requires #"Processed.csv"
model1 <- lmer(Offspring_weight_g ~
                 Age_days +
                 Habitat +
                 (1|ID),
               data = Processed_data)


summary(model1)
```

[Stages 5 + 6] Terminal

```
> model1 <- lmer(Offspring_weight_g ~ Age_days + Habitat + (1|ID), data = Processed_data)
> summary(model1)
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: Offspring_weight_g ~ Age_days + Habitat + (1 | ID)
   Data: Processed_data

Fixed effects:
             Estimate   Std. Error         df   t value  Pr(>|t|)
(Intercept)   5.051133   0.196697   195.718112   25.680    < 2e-16 ***
Age_days     -0.030746   0.003729   191.946582  -8.246    2.54e-14 ***
HabitatB     -0.056885   0.090364   189.647801  -0.630     0.530
HabitatC      0.058789   0.091050   188.100452   0.646     0.519
```

Figure 3. *Matching a manuscript to archived data and code. The methods in the manuscript can be checked against the archived data (Stage 2) and code (Stage 4). In terms of the data, the same variables that are described in the manuscript need to be present in the data, and the data need to be the same*

size as referred to in the manuscript. In terms of the code, the models that are described in the manuscript need to be clearly labelled in the code. The numbers in red refer to the stages of the guidelines being addressed. Figure created by EIC.

Stage 3. Code must be archived and adhere to FAIR guiding principles

To facilitate transparency and computational reproducibility, all code used to reproduce the results should be provided alongside data. The guidelines for code archiving are broadly similar to those of data archiving outlined above, with a few subtle differences, which we outline below. By a data editor validating Stages 3 and 4 of these guidelines, journals would achieve TOP 2025 level 3 for Analytic Code Transparency (Grant *et al* 2025).

Stage 3.1. Code files are accessible and in an open repository

As with data, code files must be accessible within an open repository with an associated globally unique persistent identifier (see Stage 1.1 for suitable repositories). This can either be the same repository as the data or a separate one. This choice may depend on the chosen repository. For example, Dryad recommends only archiving data, and directs users to archive code with Zenodo, as code is not always compatible with a CC0 license, which Dryad mandates. Care must be taken when archiving code and data separately, as the two archived projects should clearly link to each other as well as to the manuscript, and authors must provide information about how to structure the data and code directories so that the code will run with the data. For example, if the code assumes that the data are in the same parent directory within a folder called ‘Data’ (see Figure 2), then the data files will have to be organised like this for the code to run. Where possible, we suggest that data and code are archived together as this will minimise issues with computational reproducibility (see Stage 5). In some cases, this may not be possible, particularly if there are large data files that are stored in a field-specific repository (e.g., genomic data stored on GenBank). In this case, we do not expect data and code to be

archived together. Instead, details of how to access and organise the data to work with the archived code need to be included in the code README. Similar to data, code should not be included in the supplementary material of a submission. Again, if necessary, the archived projects can be anonymised (see Stage 1.1 above).

Guideline 3.1: Code files are open and freely available and are located in a permanent public repository with an associated globally unique persistent identifier, preferably in the same archived project as the data files, and are cited in the text and the reference list of the manuscript.

Stage 3.2. Code is associated with a licence

As with data, any archived code must also have an associated license to enable code sharing and repurposing. It is worth noting that typically licenses used for code differ from those used with data (i.e. creative commons licenses), and there is a multitude to choose from. We suggest using permissive code licences whenever possible, for example, MIT, BSD 2-Clause, GNU, and Apache (see: https://choosealicense.com/).

Guideline 3.2: The code must be associated with a license.

Stage 3.3. Code files are present and complete

Without code alongside the data used in the analysis (and the computational environment in which the analysis took place - see 3.5), full transparency and computational reproducibility are impossible to achieve. At a minimum, the analytical code used for statistical analyses and

graphing should be present, but we recommend providing all parts of the analysis pipeline, from data filtering and processing to model analysis and graphing.

Analyses are not always done in programming languages (e.g., R). However, several analytical programs (particularly those that use Graphical User Interfaces (GUIs)) will output a script or log detailing the analysis procedure (e.g., SPSS or Minitab), which should then be archived. If this is not possible, the researcher should clearly document which menu options were selected in the GUI and in what order, with sufficient detail to enable reproducibility. Alternatively, they can provide screenshots showing all selected options during the analysis. It should be noted that these output scripts or instructions with GUI-based software are often proprietary so will still limit reproducibility (discussed further in stage 6).

Guideline 3.3: All code used for generating the results of the manuscript (including filtering, processing, graphing, and analysis) must be present either in one or more code files.

Stage 3.4. Code is in an interoperable format

For code files to be opened and usable, it is essential they are provided in an interoperable format such as a text (.txt), R (.R) or Python (.py) file. These file formats can be readily opened by text editors and other integrated development environments (such as VSCode). Code must not be provided as a .pdf or pasted into a Word document, because even if the script can be copied and pasted, this increases the risk of unintentional errors, as these programs often insert additional characters (or spacing) that can be misinterpreted by the analysis software (e.g., Python code failing to run due to improper indentation).

Guideline 3.4: Code files must be provided in an interoperable format.

Stage 3.5. Code metadata present and adequate

Code metadata must be present in two different forms, in a separate README file and also within the code file itself. As with data, a detailed README file must be provided along with the code files that describes general information about the manuscript, e.g., manuscript title and abstract, the authors and contact information, list of all relevant funders, the globally unique persistent identifier of related data (if different), and information about the code license. Additionally, the README file should also include an outline of the workflow of the code (if multiple files exist), how to use it with the data (if archived separately) and a brief description of each code file including what data they require (i.e., raw or processed data), what they do (i.e., filtering, processing, modelling etc) and what they produce (e.g., Figure X or Table X). The README should include details of the name and version of the analytical software used (e.g., R or Python) along with the names and version numbers of the loaded (not base) packages used (e.g., these can be obtained using sessionInfo() in R). This information is essential for detailing the computational environment and enabling computational reproducibility. Finally, the README should also state whether and how the authors have used large language models (LLMs) in code generation (Resnik & Hosseini 2025). We note that it is unlikely that a data editor would be able to verify this.

The second form of metadata is included within the code files, in the form of detailed code annotation and sectioning. A header at the top of the script with a title and a quick overview of what the code does can also be very helpful, especially when the whole analysis pipeline is split across multiple scripts. The code should be broken down into distinct sections with clear headings describing their purpose (e.g., loading packages, data processing, data filtering).

Annotation should then clearly describe what the code does, how to run it (if necessary, the length of time it might take, for example if the code takes multiple hours to complete), and what it produces. From the perspective of a data editor, the most important thing is that sections of the code are clearly signposted to help assess Stage 4 (see below). Therefore, line-by-line annotation, whilst important to readers and users, is not as vital as clear labelling of sections of code and their purpose for data editors. There is no expectation that the data editor will understand all the code they quality control, and it should not be the role of data editors to review, interpret, or correct the code. Similar to the data metadata (stage 1.5), the term "adequate" refers to the code metadata providing all the information necessary to understand the analysis code without reading the manuscript.

Guideline 3.5: A sufficiently detailed README must accompany the code. Code must also be broken into sections with clear annotation stating the purpose of the code with clear links to the relevant sections, figures, and tables in the manuscript.

Stage 4. Archived code corresponds with the workflow reported in the manuscript

It is crucial that all the code needed to reproduce the results of the manuscript, including any supplementary material, is archived (Figure 3). This stage should not involve the data editor critiquing the analytical techniques used or performing a formal code review (Ivimey-Cook *et al* 2023, Hillemann *et al* 2025; although annotation is required for transparency, see Stage 3.5). It should rather involve an assessment of whether the specific code is present to perform all stages of the analyses, including producing any graphs and subsequent results stated within the manuscript. At this stage, we are also not interested in whether the code reproduces the results

in the manuscript, just that the code is *present* to produce the results. As a data editor is unlikely to be an expert in all analyses, across all packages or all programming languages, clear code annotation and signposting by the author is necessary for this to be assessed. To our knowledge, there is currently no software that performs the same task for code as DataSeer does for data. However, given the rapid progress in AI, such a tool may become available soon (Cooper *et al* 2024).

Guideline 4: The structure and content of the archived code must match the description of data filtering, processing, and analysis, and the presentation of results in the manuscript .

Stage 5. Archived code runs with the archived data

This stage is a prerequisite for full computational reproducibility. The data editor must be able to run the code with the provided data and code metadata, using the described software, without errors. The metadata provided must therefore be sufficient for a reader to install appropriate programs and libraries (and their versions) required to run the code, and to understand which code files should be run and in what order. If the data editor cannot run the code with the archived data and metadata, they cannot progress to the last stage of the guidelines, and so the data editor should then ask the authors to fix the issue. Common issues include a package or module not being installed or loaded within the code, a missing code chunk, variable names in the code and data files not matching, and code referencing data files with names that do not match the archived data files. We stress that it is not the responsibility of the data editor to solve these problems and make the code run as intended, but rather the onus should be placed on the authors.

One of the most common reasons code does not run is due to the use of local or absolute file paths that do not transfer to another user's operating system. A more reproducible way of specifying file paths is to use relative file paths and there are multiple ways to do this. A common way for RStudio users is creating an RStudio project file (.Rproj), or similarly using the {here} R package (Muller and Bryan 2020) outside of RStudio (see https://docs.posit.co/ide/user/ide/get-started/). Alternatively, local file paths can be specified when R is opened, for example by opening R within a certain directory when using the terminal or using an integrated development environment (IDE) that allows users to specify a project folder (e.g., the R GUI or VSCode). Whichever method is used should be noted in the metadata. Given the multitude of methods, the use of absolute file paths or a different method of specifying relative file paths than the data editor is familiar with should not be a reason for a data editor to return the code to an author, as long as the data editor can make it run on their computer with minor changes. We class this as a minor error that should simply be noted in the final review.

Similarly, data editors should not be expected to install exact software versions in the first instance. If there are errors upon running (or results do not match - see Stage 6), then the data editor should note this in their review, and the corresponding version should be installed and code run again. Finding that the results are not reproducible with different versions can be an insightful piece of information regarding the robustness of the results. In this instance, the authors should clearly flag that their results are sensitive to the software version used, and justify the use of a particular version in the manuscript and metadata if the results substantially differ (see Stage 6).

In some cases, the data preparation or analysis may be computationally expensive and so either require specialist hardware (such as access to a high performance cluster) or take a

considerable time to run. This should be clearly indicated within the metadata, alongside a saved output. Ideally, example code should be provided that demonstrates that the code will run. For example, if a statistical model will take a long time to run, the authors can provide example code for an analysis of a subset of the data, or present a model that runs for a reduced duration. Alternatively, data editors could also simply check that the code initiates, and then terminate the run before completion. Although this does not allow computational reproducibility to be fully assessed (see Stage 6), it at least demonstrates that the code runs. Similarly, the code may come from proprietary software or use packages from proprietary software (e.g., ASReml; Butler *et al* 2017). In such cases, the data editor will not be able to run the code and so full computational reproducibility cannot be assessed. If this is the case, this should be clearly documented in the metadata. In the case that only part of the analysis requires proprietary software, the metadata should clearly indicate which parts of the code can be assessed by the data editor. As we outline below, in both the case of computationally expensive analyses and the use of proprietary software, where possible the authors must provide saved outputs from these analyses for the data editor to review. For example, outputs of large Bayesian models that can take a considerable time to run can be saved (e.g., as a .rds file) and archived. Build systems, such as the *targets* package in R or *snakemake* in Python can provide reproducibility signatures (hashes) for computationally-intensive steps along with intermediate data objects.

In some cases, the problems of using proprietary software can be overcome by ensuring that the code can be executed using non-proprietary software or by providing alternative executable formats. For example, GNU Octave can be used to run MATLAB code and MATLAB Compiler allows converting MATLAB (.m) files into standalone applications, ensuring data editors and users can run the code without owning the proprietary software. Although such alternatives can

provide computational reproducibility, authors must carefully test for compatibility and note any limitations or differences in the metadata.

Guideline 5: Code must be able to run without error using the archived data. With the exception of easy to fix file path errors, all errors should be addressed by the author.

Stage 6. Results can be computationally reproduced by running the archived code

For this final stage, the data editor should assess computational reproducibility by checking whether the results in text, tables, and graphs within the manuscript and supplementary material match those obtained by running the archived code with the archived data. This can only be assessed if the archived code runs without error (Stage 5). In most cases, we expect that exact reproducibility of the results is possible (i.e. the exact number in the manuscript should be generated by running the code), and any deviations would mean that the computational reproducibility test has failed. In some cases, authors may have used additional software to post-process figures. In these cases the data represented within the figure is still expected to be the same, but the code may not reproduce the figure exactly.

One reason that the reproduced results may slightly differ is through the use of stochastic methods that involve (pseudo) random number generation, such as Monte Carlo methods (e.g., simulations or Bayesian analysis using Monte Carlo Markov Chains (MCMC)) as these will produce a slightly different result each time they run. However, this variation can be avoided by setting a seed (e.g., using set.seed() function in R or random.seed() in Python; see Box 1) at the beginning of any code section that would be run independently, which means that the

pseudorandom number generation is the same each time the code is run, enabling the same results to be reproduced, including for analyses and figure generation (e.g., with point jittering). We note that setting seed does not always ensure computational reproducibility, for instance the use of rmvnorm() from the {MASS} R package does not create the same random numbers across different operating systems due to floating point errors. The use of different hardware may similarly lead to subtly different results.

In some circumstances, the authors may have used LLMs for data generation or analysis (for instance, extracting data from images or video recordings or extraction of data from literature). As of writing, the use of LLMs in analysis poses a significant problem for reproducibility due to the variability of generated outputs (Fukataki *et al.* 2025, Meyer *et al.* 2025, Staudinger *et al.* 2024). The use of setting seed parameters is currently only available for some LLM models (e.g., OpenAI) but even this has been suggested to not guarantee reproducible results (Vadlapati 2023, Morin & Willetts 2020). Therefore, although the code/prompts used can still be archived, the results may not be exactly reproducible. The use of LLMs in this way is similar to the use of proprietary software, in that it causes one part of the analysis pipeline to become unreproducible. This, however, should not affect the reproducibility of results that comes from further analysing the processed data generated by LLMs, which should be archived. The use of LLMs in this context may also be highly computationally expensive, further limiting reproducibility, as we discuss in Stage 5 above.

If there is no way for the data editor to generate the exact result (e.g., because the software does not allow setting a seed) then the data editor can allow a degree of tolerance for the result which should be noted in their review. Archmiller *et al*. (2020) suggest comparing the conclusion (the direction and significance of results) as well as the numbers of the original and reproduced results. In the first case, if the direction of the effect changes, or the statistical significance

changes, then this should be viewed as failing the computational reproducibility test. For results close to zero or the significance threshold, small changes in the results might change direction or significance, respectively. Hardwicke *et al*. (2021) therefore suggested using % error (i.e. (reproduced– original)/original × 100), as this is not dependent on the scale of the results, where 0-10% was classified as a minor deviation and >10% as a major deviation, and therefore, not reproducible. However, this % error method (1) still allows for a substantial deviation from the reported values, (2) would result in different tolerances for different effects within the same model, and (3) is most meaningful when effect sizes are on a ratio scale, which typically they are not. Perhaps most importantly, reproduced results should fall well within the reported uncertainty of the original result, and if they do not, this should be viewed as a failure to reproduce the results. The data editor should communicate the conditions under which computational reproducibility was assessed (e.g., the tolerance threshold) in their review. As opposed to in-text results and tables, figures cannot be exactly compared without the use of specialist software, but should be compared by eye for reproducibility.

The use of computationally expensive methods or proprietary software may mean that the data editor cannot feasibly re-run the analysis in full (see Stage 5 above). If none of the code can be run by the data editor, for example if it all takes place using proprietary software or it involves very computationally expensive analysis, then computational reproducibility cannot be assessed (both Stages 5 and 6 would fail). Clearly, this should not prevent the publication of a manuscript containing such analyses in journals where data editors assess Stages 5 and 6 of these guidelines. In this case, we would therefore recommend that it is highlighted in the manuscript that computational reproducibility could not be assessed (e.g., in the data and code availability statement, or open research sections). If it is only part of the code that cannot be run by the data editor (e.g., a computationally expensive model), then the output of this code should be provided by the author in the archived project and noted in the metadata, so that the output can

be compared to the manuscript by the data editor. Given the large variation in data workflows and different proprietary software that may be used, we encourage data editors, editors and journals to be very open to discussion, constructive and flexible in their roles to adopt, or work towards adopting these guidelines.

Guideline 6: Results reproduced by the data editor with the archived data and code must match those presented in the manuscript. A tolerance threshold can be given when there is not an exact match but the authors must state clearly in the code metadata why this mismatch might occur. If saved model outputs are instead provided, this must also be clearly stated in the metadata.

Table 2. Summary table of the guidelines for each the six Stages of the SORTEE Guidelines for Data and Code Quality Control in Ecology and Evolutionary Biology

Stage	Guidelines
1. Data must be archived and adhere to FAIR guiding principles	
1.1 Data files are accessible and in an open repository	Data are open and freely available and are located in a permanent public repository with an associated globally unique persistent identifier, that is cited in the text and reference list of the manuscript.
1.2. Data are associated with a license	Data must be associated with a license.
1.3. Data files are present and complete	Authors must either provide: a) raw data, along with the processed data and/or code to prepare the data for analysis, or b) a sample of raw data alongside processed/filtered data when full raw data upload is not possible, or c) processed/filtered data with a detailed description of how to both obtain and process/filter the raw data.
1.4. Data files are in an interoperable format	Data files must be provided in an interoperable format.
1.5. Data metadata present and adequate	Detailed metadata, including (but not limited to) a README file, must accompany the data.
2. Archived data corresponds with the data reported in the manuscript	The structure and contents of the archived data files must match the description in the manuscript.
3. Code must be archived and adhere to FAIR guiding principles	
3.1. Code files are accessible and in an open repository	Code files are open and freely available and are located in a permanent public repository with an associated globally unique persistent identifier, preferably in the same archived project as the data files, and are cited in the text and the reference list of the manuscript.
3.2. Code is associated with a licence	The code must be associated with a license.

	3.3. Code files are present and complete	All code used for generating the results of the manuscript (including filtering, processing, graphing, and analysis) must be present either in one or more code files.
	3.4. Code is in an interoperable format	Code files must be provided in an interoperable format.
	3.5. Code metadata present and adequate	A sufficiently detailed README must accompany the code. Code must also be broken into sections with clear annotation stating the purpose of the code with clear links to the relevant sections, figures, and tables in the manuscript.
4.	**Archived code corresponds with the workflow reported in the manuscript**	The structure and content of the archived code must match the description of data filtering, processing, and analysis, and the presentation of results in the manuscript.
5.	**Archived code runs with the archived data**	Code must be able to run without error using the archived data. With the exception of easy to fix file path errors, all errors should be addressed by the author.
6.	**Results can be computationally reproduced by running the archived code**	Results reproduced by the data editor with the archived data and code must match those presented in the manuscript. A tolerance threshold can be given when there is not an exact match but the authors must state clearly in the code metadata why this mismatch might occur. If saved model outputs are instead provided, this must also be clearly stated in the metadata.

863

Suggestions to Authors

Data and code quality control is becoming increasingly common across journals in ecology and evolutionary biology. Consequently, authors will have to adhere to certain guidelines for data and code sharing. Although the guidelines presented here are largely aimed at data editors, knowledge of the checks that a data editor is expected to perform will help authors understand what is needed from their data and code prior to submission. It may also increase the likelihood of authors catching their own mistakes because of the checks that will be performed. We hope that the widespread adoption of these guidelines will make the process more transparent for authors and also consistent across journals in the event of manuscript resubmission elsewhere. We acknowledge that making data and code readily accessible and reusable adds to the work load of authors (at least initially). We have several suggestions to ease this process:

Adhere to the data and code quality control guidelines from the beginning of the study

Working to make a project accessible and reproducible at the end of a study is a lot of work. We would recommend creating a clear directory structure and creating metadata (e.g., a README) at the beginning of the study, and updating the metadata as new files are added. Similarly, annotating code as authors produce it, not only with section descriptions but also with information about how they run and what output they produce, is far easier than going back and annotating code at the end of the study. Bearing reproducibility in mind while working on a project also makes it far more likely that someone else will be able to reuse the project and successfully run the code. This is even more useful if authors plan on collaborating with multiple people during the study's lifetime. Generally, there exists a multitude of benefits to working reproducibly (Markowetz 2015). We acknowledge that for many authors this may present a steep learning curve, however adherence and knowledge of these guidelines will promote learning and progression over time. Whilst these guidelines are generally aimed at data editors,

there are several resources designed for authors, which we would direct authors to, for example the British Ecological Society guides on reproducible code (Cooper & Hsing 2025) and data management (Harrison 2018) and the TADA (Transferable, Available, Documented, Annotated) guidelines (Ivimey-Cook *et al*. 2025), aimed specifically at ecologists and evolutionary biologists.

Prepare data and code according to the data and code quality control guidelines guidelines before submission to any journal

Inherently linked to the point above, if authors have not prepared data and code according to the data and code quality control guidelines from the start of the study, it is advisable to at least have data and code ready for submission. This will minimise any problems during both the submission process and the data and code quality control, and allow for easy transfer between journals. We have included a summary of the guidelines for data editors that authors may find useful in Table 2.

Perform a pre-submission code review

It is advisable for authors to send their project containing data and code to a colleague or co-author for a code review prior to submission to a journal (Ivimey-Cook *et al* 2023). This enables checking whether the code runs with the data in the project structure provided. They can then check whether there is appropriate and adequate metadata, whether data and code match the manuscript, and whether the code reproduces the results in the manuscript. Importantly, co-authors may also be more likely to spot any mistakes in the code as they are familiar with the study and data, and data editors do not check the reliability of code. This could be done within research groups, where the task of code reviewing is shared between members of the team, or as part of a larger ‘code club’. Open science organisations, such as SORTEE, have their own code clubs which are open to join. For further advice on setting up code clubs see Ivimey-Cook *et al* (2023).

Consider presenting code and associated outputs using Markdown or Quarto

Presenting everything in one self-contained document such as a Markdown or Quarto file can be very helpful for data editors and future readers or users (Buckley *et al* 2025). It allows for a clear link between the code, the data, and the resulting outputs that may need to be assessed.

Suggestions for Journals

Data and code quality control should start at submission

Currently, in many journals, data and code quality control occurs after (or close to) acceptance. We recommend that data and code are required at submission (see above for methods to anonymise data and code), and that data editors perform a light check of the data, code, and associated metadata (e.g., Stages 1 and 3) early on (before sending to review). This enables reviewers to both see and review the data and code during peer review (if they choose to), and also engages the authors in the data and code quality control process at an early stage. That way, any problems can be highlighted and addressed early in the process. Computational reproducibility checks (Stages 5 and 6) would ideally be conducted later in the process, at a point where the code (particularly related to statistical analysis) is unlikely to change because of further review, to avoid a data editor having to perform these checks multiple times. Any checks need to also be clearly communicated to those not performing the data and code quality control (i.e., editors and reviewers).

Ensure journals have data editors with a mixture of coding expertise

There exists a multitude of different languages in which to write code and analyse data. Although R is one of the most popular in Ecology (from 58-80% of studies in ecology and evolution; Lai *et al* 2019, Culina *et al*. 2020, Kambouris *et al*. 2024), code is often written in

939 other languages such as Python, Matlab, SAS, Julia, to name a few. It is therefore important
940 that a journal considers having multiple data editors with varying coding language, data type
941 and area expertise. This means that data editors can be suitably paired to each manuscript.
942
943 **Have clear guidelines on the journal website**
944 Authors will be more likely to adhere to the guidelines adopted by the journal prior to submission
945 if these are clearly displayed on the website, ideally under ‘Instructions to Authors’ sections.
946 These need to outline what stages of the guidelines the data editors check (e.g., Stages 1-4),
947 what they expect at each stage from the author, and what the authors need to state in their data
948 availability statement. Having data, code, and associated metadata already in a state ready for
949 quality control will reduce much of the work for the data editor. Some journals additionally
950 provide template README files to help authors.
951
952 **Have clear statements within manuscripts**
953 For readers to know what quality control checks have been performed and to highlight the
954 journal's endeavours to ensure the highest quality research, it should be clearly stated within the
955 data and code availability section what checks have been performed. For instance “Data and
956 code were checked from Stage 1-4 of the SORTEE guidelines for Data and Code Quality
957 Control”. This statement should also contain information if a check has not been able to be
958 performed, for instance, if the use of proprietary software, sensitive data, or computationally
959 expensive analyses were involved, impeding computational reproducibility tests.
960
961 In psychology and medicine, open science badges have previously been used to indicate
962 manuscripts that adhere to certain open science practices (e.g., open data, open code, open
963 materials, pre-registration) with the ultimate goal of encouraging authors to adopt these
964 practices. Evidence for their effectiveness in increasing data and code sharing is mixed, with

early observational studies reporting increases in data sharing after badge implementation (Kidwell *et al*., 2016), but a subsequent randomized controlled trial finding no such effect in a biomedical journal context (Rowhani-Farid *et al*., 2020). Note that the journals surveyed in these cases did not have data editors actively checking the data and code archiving.

The presence of badges has also been shown to increase the trust of researchers in published articles (Schneider *et al* 2022). Journals could choose to use such badges following data and code quality control to indicate that presence of open data (Stage 1-2) and open code (Stage 3) has been verified, and further badges could be developed for computational reproducibility (Stages 5-6).

Have clear definitions and policies of what code and data the journal requires

Ideally, all the data and code used to generate the results should be archived including both raw and processed data and all the code used to process, filter, model, and graph. However, this is at the discretion of the journal and therefore should be made explicit to the authors prior to submission. We recommend that the form of data and code is clearly described in the data availability statement of the manuscript, for instance, “Processed data and code used in modeling and graphing are archived here…”. Again, the journal requirements need to also be clearly communicated to all levels of the journal hierarchy to ensure successful implementation.

Conclusion

Here we present a standardised set of guidelines for data and code quality control for journals in ecology and evolutionary biology. As it stands, rates of data and code archiving, and importantly the quality of archived data and code, are low. By recruiting data editors, journals can positively impact the state of open data and code, and in doing so increase research transparency and

reproducibility. With the SORTEE data and code quality control guidelines, we propose steps to increase the quality and consistency of data and code quality control across journals that currently have data editors, and provide a template for journals wanting to start data and code quality control. We believe that these guidelines will have substantial benefits for journals, for authors, and for the wider scientific community.

Acknowledgments

We thank Bob Montgomerie for extensive discussion and comments on the manuscript. Thanks also to Lars Vilhuber for a useful discussion on the role of data editors in Economics. Finally, we would like to thank Ignasi Bartomeus, Noam Ross and François Keck for their feedback and the positive review process at PCI Ecology.

Funding

The authors declare that they have received no specific funding for this study

Author Contributions

Conceptualisation - JLP and EIC

Writing - Original Draft - JLP and EIC

Writing - Review & Editing - All authors

Visualisation - EIC

Project administration - JLP

Supervision - EIC

Conflict of Interest

SORTEE has been financially supported by Dryad, Figshare, the Center of Open Science (which hosts the Open Science Framework; OSF), Peer Community In, the American Society of Naturalists and the Royal Society, all of which are mentioned in the guidelines. EIC was the 2025 president of SORTEE. EIC, ML, MP, AST were on SORTEEs board of directors. JLP, KBN, CJ, SN, and EIC were members of the 2025 and 2026 SORTEE advocacy committee. JLP, BJA, KBN, JAB, BC, PDA, DG, CJ, RK, ML, SN, ROD, MP, QP, AST, NvD, EIC are SORTEE members. BB is a data editor at the American Naturalist. EIC, AST, ROD, NvD, MJG, TD, EF, PDA and QP are data editors at Ecology Letters. JAB, DG, DSM and LW are data editors at Proceedings B, as was BJA at the initiation of this project. SL is the data editor at Journal of Evolutionary Biology. EFJ is the data editor from Behavioural Ecology and Sociobiology. BC, RK, ML, MP, are data editors at PCI. TG is on the Executive board of Peer Community Journal and president of Peer Community In and BC is the editorial coordinator for PCI. They did not intervene in the evaluation process made by PCI Ecology.

References

Allen, C. & Mehler, D.M.A. 2019. Open science challenges, benefits and tips in early career and beyond. *PLOS Biol.* **17**: e3000246.

Archmiller, A.A., Johnson, A.D., Nolan, J., Edwards, M., Elliott, L.H., Ferguson, J.M., *et al.* 2020. Computational Reproducibility in The Wildlife Society's Flagship Journals. *J. Wildl. Manag.* **84**: 1012–1017.

Barker, M., Chue Hong, N.P., Katz, D.S., Lamprecht, A.-L., Martinez-Ortiz, C., Psomopoulos, F., *et al.* 2022. Introducing the FAIR Principles for research software. *Sci. Data* **9**: 622.

Barnes, N. 2010. Publish your computer code: it is good enough. *Nature* **467**: 753–753.

Barrett, L. & Montgomerie, R. 2025. A data editor for behavioral ecology. *Behav. Ecol.* **36**: araf077.

Barrett, S.C.H. 2024. Proceedings B 2024: the year in review. *Proc. R. Soc. B Biol. Sci.* **292**: 20250065.

Bavota, G. & Russo, B. 2015. Four eyes are better than two: On the impact of code reviews on software quality. In: *2015 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pp. 81–90.

Belkhir, K., Smadja, C.M., Antoine, P.-O., Scornavacca, C. & Galtier, N. 2025. An overview of open science in eco-evo research and the publisher effect. EcoEvoRxiv.

Berberi, I. & Roche, D. 2023. Living database of journal data policies in E&E. , doi: 10.17605/OSF.IO/D6SP3. OSF.

Berberi, I. & Roche, D.G. 2022. No evidence that mandatory open data policies increase error correction. *Nat. Ecol. Evol.* **6**: 1630–1633.

Bolnick, D. & Paull, J. 2016. Retraction: Morphological and dietary differences between individuals are weakly but positively correlated within a population of threespine stickleback. *Evol. Ecol. Res.* **17**: 849.

Buckley, Y.M., Bardgett, R., Gordon, R., Iler, A., Mariotte, P., Ponton, S., *et al.* 2025. Using dynamic documents to mend cracks in the reproducible research pipeline. *J. Ecol.* **113**: 270–274.

Butler, D.G., Cullis, B.R., Gilmour, A.R., Gogel, B.G. & Thompson, R. 2017. ASReml-R Reference Manual. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

Carroll, S.R., Herczog, E., Hudson, M., Russell, K. & Stall, S. 2021. Operationalizing the CARE and FAIR Principles for Indigenous data futures. *Sci. Data* **8**: 108.

Chapman, A.D. 2020. Current Best Practices for Generalizing Sensitive Species Occurrence Data.

Christensen, G., Dafoe, A., Miguel, E., Moore, D.A. & Rose, A.K. 2019. A study of the impact of data sharing on article citations using journal policies as a natural experiment. *PLOS ONE* **14**: e0225883.

Cologna, V., Mede, N.G., Berger, S., Besley, J., Brick, C., Joubert, M., *et al.* 2025. Trust in scientists and their role in society across 68 countries. *Nat. Hum. Behav.* **9**: 713–730.

Cooper, N., Clark, A.T., Lecomte, N., Qiao, H. & Ellison, A.M. 2024. Harnessing large language models for coding, teaching and inclusion to empower research in ecology and evolution. *Methods Ecol. Evol.* **15**: 1757–1763.

Cooper, N. & Hsing, P.-Y. 2025. Guide to Reproducible Code. British Ecological Society. https://doi.org/10.5281/zenodo.16421733

Culina, A., Berg, I. van den, Evans, S. & Sánchez-Tójar, A. 2020. Low availability of code in ecology: A call for urgent action. *PLOS Biol.* **18**: e3000763.

Evans, S.R. 2016. Gauging the Purported Costs of Public Data Archiving for Long-Term Population Studies. *PLOS Biol.* **14**: e1002432.

Feng, X., Qiao, H. & Enquist, B.J. 2020. Doubling demands in programming skills call for ecoinformatics education. *Front. Ecol. Environ.* **18**: 123–124.

Fernández-Juricic, E. 2021. Why sharing data and code during peer review can enhance behavioral ecology research. *Behav. Ecol. Sociobiol.* **75**: 103.

Fukataki, Y., Hayashi, W., Nishimoto, N. & Ito, Y.M. 2025. Developing artificial intelligence tools for institutional review board pre-review: A pilot study on ChatGPT's accuracy and reproducibility. *PLOS Digit Health* 4(6): e0000695. https://doi.org/10.1371/journal.pdig.0000695

Gihawi, A., Ge, Y., Lu, J., Puiu, D., Xu, A., Cooper, C.S., *et al.* 2023. Major data analysis errors invalidate cancer microbiome findings. *mBio* **14**: e01607-23.

Goldacre, B., Morton, C.E. & DeVito, N.J. 2019. Why researchers should share their analytic code. *BMJ* **367**: l6365.

1085 Gomes, D.G.E., Pottier, P., Crystal-Ornelas, R., Hudgins, E.J., Foroughirad, V., Sánchez-
1086 Reyes, L.L., *et al.* 2022. Why don't we share data and code? Perceived barriers and
1087 benefits to public archiving practices. *Proc. R. Soc. B Biol. Sci.* **289**: 20221113.

1088 Gould, E., Fraser, H.S., Parker, T.H., Nakagawa, S., Griffith, S.C., Vesk, P.A., *et al.* 2025. Same
1089 data, different analysts: variation in effect sizes due to analytical decisions in ecology and
1090 evolutionary biology. *BMC Biol.* **23**: 35.

1091 Grant, S., Corker, K., Mellor, D., Stewart, S., Cashin, A., Lagisz, M., *et al.* 2025. TOP 2025: An
1092 Update to the Transparency and Openness Promotion Guidelines. OSF.

1093 Hardwicke, T.E., Bohn, M., MacDonald, K., Hembacher, E., Nuijten, M.B., Peloquin, B.N., *et al.*
1094 2021. Analytic reproducibility in articles receiving open data badges at the journal
1095 Psychological Science: an observational study. *R. Soc. Open Sci.* **8**: 201494.

1096 Hardwicke, T.E., Mathur, M.B., MacDonald, K., Nilsonne, G., Banks, G.C., Kidwell, M.C., *et al.*
1097 2018. Data availability, reusability, and analytic reproducibility: evaluating the impact of a
1098 mandatory open data policy at the journal *Cognition*. *R. Soc. Open Sci.* **5**: 180448.

1099 Harrison, K. 2018. Guide to Data Management. British Ecological Society

1100 Harvard Longwood Medical Area Research Data Management Working Group. 2023. Harvard
1101 Biomedical Repository Matrix. Zenodo. https://doi.org/10.5281/zenodo.10651775

1102 Henderson, A.S., Hickson, R.I., Furlong, M., McBryde, E.S. & Meehan, M.T. 2024.
1103 Reproducibility of COVID-era infectious disease models. *Epidemics* **46**: 100743.

1104 Hennessy, E.A., Acabchuk, R.L., Arnold, P.A., Dunn, A.G., Foo, Y.Z., Johnson, B.T., *et al.* 2022.
1105 Ensuring Prevention Science Research is Synthesis-Ready for Immediate and Lasting
1106 Scientific Impact. *Prev. Sci.* **23**: 809–820.

1107 Hillemann, F. [freddy], Burant, J.B., Culina, A. & Vriend, S.J.G. 2025. Code review in practice: A
1108 checklist for computational reproducibility and collaborative research in ecology and
1109 evolution. *EcoEvoRxiv*.

1110 Ivimey-Cook, E.R., Pick, J.L., Bairos-Novak, K.R., Culina, A., Gould, E., Grainger, M., *et al.*
1111 2023. Implementing code review in the scientific workflow: Insights from ecology and
1112 evolutionary biology. *J. Evol. Biol.* **36**: 1347–1356.

Ivimey-Cook, E.R., Sánchez-Tójar, A., Berberi, I., Culina, A., Roche, D.G., Almeida, R.A., *et al.* 2025. From Policy to Practice: Progress towards Data- and Code-Sharing in Ecology and Evolution. *Proc. R. Soc. B Biol. Sci* **292** 20251394.

Ivimey-Cook, E.R., Culina, A., Dimri, S., Lagisz, M., Moran, N., Nakagawa, S., *et al.* 2025. TADA! Simple guidelines to improve analytical code sharing for transparency and reproducibility. EcoEvoRxiv, https://doi.org/10.32942/X2D93K

Janssens, M., Gaillard, S., Haan, J.J. de, Leeuw, W. de, Brooke, M., Burke, M., *et al.* 2023. How open science can support the 3Rs and improve animal research. *Res. Ideas Outcomes* **9**: e105198.

Kambouris, S., Wilkinson, D.P., Smith, E.T. & Fidler, F. 2024. Computationally reproducing results from meta-analyses in ecology and evolutionary biology using shared code and data. *PLOS ONE* **19**: e0300333.

Kellner, K.F., Doser, J.W. & Belant, J.L. 2025. Functional R code is rare in species distribution and abundance papers. *Ecology* **106**: e4475.

Kidwell, M.C., Lazarević, L.B., Baranski, E., Hardwicke, T.E., Piechowski, S., Falkenberg, L.-S., *et al.* 2016. Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLOS Biol.* **14**: e1002456.

Kim, B., Moran, N.P., Reinhold, K. & Sánchez-Tójar, A. 2021. Male size and reproductive performance in three species of livebearing fishes (Gambusia spp.): A systematic review and meta-analysis. *J. Anim. Ecol.* **90**: 2431–2445.

Kimmel, K., Avolio, M.L. & Ferraro, P.J. 2023. Empirical evidence of widespread exaggeration bias and selective reporting in ecology. *Nat. Ecol. Evol.* **7**: 1525–1536.

Kohrs, F.E., Auer, S., Bannach-Brown, A., Fiedler, S., Haven, T.L., Heise, V., *et al.* 2023. Eleven strategies for making reproducible research and open science training the norm at research institutions. *eLife* **12**: e89736.

König, L., Gärtner, A., Slack, H., Dhakal, S., Adetula, A., Dougherty, M., *et al.* 2025. How to bolster employability through open science. OSF.

Lai, J., Lortie, C.J., Muenchen, R.A., Yang, J. & Ma, K. 2019. Evaluating the popularity of R in ecology. *Ecosphere* **10**: e02567.

Maitner, B., Santos Andrade, P.E., Lei, L., Kass, J., Owens, H.L., Barbosa, G.C.G., *et al.* 2024. Code sharing in ecology and evolution increases citation rates but remains uncommon. *Ecol. Evol.* **14**: e70030.

Mandhane, P.J. 2024. Notice of Retraction: Hahn LM, et al. Post–COVID-19 Condition in Children. JAMA Pediatrics. 2023;177(11):1226-1228. *JAMA Pediatr.* **178**: 1085–1086.

Manzanedo, R.D., HilleRisLambers, J., Rademacher, T.T. & Pederson, N. 2021. Retraction Note: Evidence of unprecedented rise in growth synchrony from global tree ring records. *Nat. Ecol. Evol.* **5**: 1047–1047.

Markowetz, F. 2015. Five selfish reasons to work reproducibly. *Genome Biol.* **16**: 274.

Meyer, A,, Schömig, E. & Streichert, T. 2025. ChatGPT and reference intervals: a comparative analysis of repeatability in GPT-3.5 Turbo, GPT-4, and GPT-4o. *Front. Artif. Intell.* 8:1681979. https://doi.org/10.3389/frai.2025.1681979

Mills, J.A., Teplitsky, C., Arroyo, B., Charmantier, A., Becker, Peter.H., Birkhead, T.R., *et al.* 2015. Archiving Primary Data: Solutions for Long-Term Studies. *Trends Ecol. Evol.* **30**: 581–589.

Minocher, R., Atmaca, S., Bavero, C., McElreath, R. & Beheim, B. 2021. Estimating the reproducibility of social learning research published between 1955 and 2018. *R. Soc. Open Sci.* **8**: 210450.

Mislan, K. a. S., Heer, J.M. & White, E.P. 2016. Elevating The Status of Code in Ecology. *Trends Ecol. Evol.* **31**: 4–7.

Molloy, J.C. 2011. The Open Knowledge Foundation: Open Data Means Better Science. *PLOS Biol.* **9**: e1001195.

Morin, M. & Willetts, M. 2020. Non-Determinism in TensorFlow ResNets. arXiv. https://doi.org/10.1145/3673791.3698432

Müller, K. & Bryan, J. 2020. here: A Simpler Way to Find Your Files.

National Academies of Sciences, E., Affairs, P. and G., Committee on Science, E., Information, B. on R.D. and, Sciences, D. on E. and P., Statistics, C. on A. and T., *et al.* 2019. Reproducibility. In: *Reproducibility and Replicability in Science*. National Academies Press (US).

Parr, C.S. & Cummings, M.P. 2005. Data sharing in ecology and evolution. *Trends Ecol. Evol.* **20**: 362–363.

Piwowar, H.A. & Chapman, W.W. 2010. Public sharing of research datasets: A pilot study of associations. *J. Informetr.* **4**: 148–156.

Piwowar, H.A., Day, R.S. & Fridsma, D.B. 2007. Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLOS ONE* **2**: e308.

Powers, S.M. & Hampton, S.E. 2019. Open science, reproducibility, and transparency in ecology. *Ecol. Appl.* **29**: e01822.

Purgar, M., Klanjscek, T. & Culina, A. 2022. Quantifying research waste in ecology. *Nat. Ecol. Evol.* **6**: 1390–1397.

R Core Team. 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Reinecke, R., Trautmann, T., Wagener, T. & Schüler, K. 2022. The critical need to foster computational reproducibility. *Environ. Res. Lett.* **17**: 041005.

Resnik, D.B. & Hosseini, M. 2025. Disclosing artificial intelligence use in scientific research and publication: When should disclosure be mandatory, optional, or unnecessary? *Accountability in Research*, 1–13. https://doi.org/10.1080/08989621.2025.2481949

Roche, D.G., Berberi, I., Dhane, F., Lauzon, F., Soeharjono, S., Dakin, R., *et al.* 2022. Slow improvement to the archiving quality of open datasets shared by researchers in ecology and evolution. *Proc. R. Soc. B Biol. Sci.* **289**: 20212780.

Roche, D.G., Kruuk, L.E.B., Lanfear, R. & Binning, S.A. 2015. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? *PLOS Biol.* **13**: e1002295.

Rowhani-Farid, A., Aldcroft, A. & Barnett, A.G. 2020. Did awarding badges increase data sharing in BMJ Open? A randomized controlled trial. *R. Soc. Open Sci.* **7**: 191818.

Sánchez-Tójar, A., Bezine, A., Purgar, M. & Culina, A. 2025. Code-sharing policies are associated with increased reproducibility potential of ecological findings. *Peer Community J.* **5**.

Schneider, J., Rosman, T., Kelava, A. & Merk, S. 2022. Do Open-Science Badges Increase Trust in Scientists Among Undergraduates, Scientists, and the Public? *Psychol. Sci.* **33**: 1588–1604.

Soeharjono, S. & Roche, D.G. 2021. Reported Individual Costs and Benefits of Sharing Open Data among Canadian Academic Faculty in Ecology and Evolution. *BioScience* **71**: 750–756.

SORTEE. 2026. SORTEE 2025 Annual Report Accessed 2026-01-13 https://osf.io/gsw6x

Staudinger, M., Kusa. W., Piroi, F., Lipani A. & Hanbury, A. 2024 A Reproducibility and Generalizability Study of Large Language Models for Query Generation. *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region.* 186 - 196 https://doi.org/10.1145/3673791.3698432

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., *et al.* 2021. Data sharing practices and data availability upon request differ across scientific disciplines. *Sci. Data* **8**: 192.

Thrall, P.H., Chase, J., Drake, J., Espuno, N., Hello, S., Ezenwa, V., *et al.* 2023. From raw data to publication: Introducing data editing at Ecology Letters. *Ecol. Lett.* **26**: 829–830.

Touchon, J.C. & McCoy, M.W. 2016. The mismatch between current statistical practice and doctoral training in ecology. *Ecosphere* **7**: e01394.

Trisovic, A., Lau, M.K., Pasquier, T. & Crosas, M. 2022. A large-scale study on research code quality and execution. *Sci. Data* **9**: 60.

Vadlapati, P. 2023. Does Seed Matter?: Investigating the Effect of Random Seeds on LLM Accuracy. *International Journal on Science and Technology*, 14, 1-5. https://doi.org/10.5281/zenodo.14288248

Vazire, S. 2017. Quality Uncertainty Erodes Trust in Science. *Collabra Psychol.* **3**: 1.

Viglione, G. 2020. 'Avalanche' of spider-paper retractions shakes behavioural-ecology community. *Nature* **578**: 199–200.

Vines, T.H., Andrew, R.L., Bock, D.G., Franklin, M.T., Gilbert, K.J., Kane, N.C., *et al.* 2013. Mandated data archiving greatly improves access to research data. *FASEB J.* **27**: 1304–1308.

Weissgerber, T.L., Gazda, M.A., Nilsonne, G., ter Riet, G., Cobey, K.D., Prieß-Buchheit, J., *et al.* 2024. Understanding the provenance and quality of methods is essential for responsible reuse of FAIR data. *Nat. Med.* **30**: 1220–1221.

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., *et al.* 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**: 160018.