Journal of Ecology

Using dynamic documents to mend cracks in the reproducible research pipeline

Journal:	*Journal of Ecology*
Manuscript ID	JEcol-2024-1172
Manuscript Type:	Editorial
Key-words:	Aquatic plant ecology, Biogeography and macroecology, Biological Flora, Determinants of plant community diversity and structure, Dispersal, Ecological omics, Ecophysiology

Journal of Ecology Editorial

Using dynamic documents to mend cracks in the reproducible research pipeline

Yvonne M. Buckley [1], Richard Bardgett [2], Rowena Gordon [3], Amy Iler [4, 5, 6], Pierre Mariotte [7], , Samantha Ponton [3], Andy Hector [8]

[1] Co-Centre for Climate + Biodiversity + Water, School of Natural Sciences, Trinity College Dublin, Dublin 2, Ireland

https://orcid.org/0000-0001-7599-3201 [2]

[2] Department of Earth and Environmental Sciences, The University of Manchester, Manchester, UK

orcid.org/0000-0002-5131-0127

[3] British Ecological Society, London, UK

[4] Chicago Botanic Garden, Negaunee Institute of Plant Conservation Science and Action, Glencoe, IL 60022, USA

[5] Rocky Mountain Biological Laboratory, Crested Butte, CO 81224, USA

[6] Northwestern University, Program in Plant Biology and Conservation, Evanston, IL 60208, USA

https://orcid.org/0000-0002-3354-7593

[7] Agroscope, Grazing Systems, Posieux, Switzerland

orcid.org/0000-0001-8570-8742

[8] Department of Biology, University of Oxford, Oxford, UK

https://orcid.org/0000-0002-1309-7716

Summary

1. There is a research reproducibility crisis, including in ecology. The research pipeline from conception to publication has many cracks, which means that it may not be possible to repeat and verify published results.
2. Reproducibility means that the results of a study can be reproduced from the original data. It is a critical step in the quality assurance of a study; indeed, the re-use and subsequent citation of methods from reproducible research can increase the impact of the work beyond the findings of the specific study.
3. Given the original data, code, and documentation, in theory, all research results could be reproduced. However, sufficient information must be available to understand and reproduce the data handling, analysis, and modelling. Information should also be accessible, enabling reproduction with reasonable effort.
4. Various open-source software options exist that allow scientists to easily annotate their scripts in a way that makes it simple to produce dynamic documents that give a more accessible account of the analysis (html, pdf and various word processor file types). Popular software options - including Jupyter notebooks, the R markdown package and

the new multi-language Quarto application - produce documents that weave together the input code and software-generated output (text, tables, and figures) with the author's explanatory text to produce a clear narrative of the analysis process.

5. Therefore, we now encourage the submission of supplementary dynamic documents to the *Journal of Ecology* to improve the reproducibility and transparency of research published in the journal. Reproducibility can be assessed prior to the submission of the work for publication, during peer review and post-publication. Authors are encouraged to provide three file types: the data, an executable dynamic document and a static reproducibility PDF file that integrates and annotates the input code with the statistical output. We provide some basic examples of dynamic documents for reproducibility.

The reproducibility crisis

Scientific studies should be reproducible, and reliable results are, by definition, replicable. Unfortunately, over the last couple of decades, researchers in several disciplines have discovered that a worrying amount of our past and current research is neither reproducible nor replicable, prompting some to declare a 'reproducibility crisis' (Ioannidis, 2005). Strictly, reproducibility is the capacity to reproduce the results of a study given the original data, code and documentation—put simply, the research was done in a way that was transparent. Replicability on the other hand is the capacity to repeat a study - collecting new data using the same methods - and (generally) come to the same conclusions (Essawy et al., 2020). However, the definitions are not rigorously applied, and the two terms are sometimes used to cover aspects of both concepts. The causes of the reproducibility crisis are complex and occur at many points of the research process from study design, through data collection and cleaning, to exploratory analysis and choice of inference method (Alston & Rick, 2021; Leek & Peng, 2015). Ecology and environmental modelling are certainly not immune to this crisis (Alston & Rick, 2021; Archmiller et al., 2020; Essawy et al., 2020). Recently, Gould et al. (2023) showed that even when given the same ecological data sets, scientists come to different conclusions due to variation in the method of analysis. This emphasises the need for data analysis decisions to be clear and transparent to other researchers trying to reproduce the work. Beyond knowing exactly what a researcher has done to prepare their data for analysis, there is also the challenge of knowing whether the statistical analyses themselves are suitable and correct for the conclusions drawn.

There are many reproducibility cracks in the research pipeline, from conception to publication and beyond. Fixing such a diverse range of problems will not be easy. However, the development of robust procedures for documenting the research process will help, and academic journals have a key role to play here. Reproducibility is a critical step in quality assurance of a study. It can be assessed prior to the submission of the work for publication, during peer review and post-publication.

Benefits of reproducible research

Journal of Ecology Editorial

Reproducible research methods have widespread benefits, including to individual researchers, the research community, peer reviewers and publishers. The rigour of recording analyses using dynamic documents usually brings benefits for the data analyst as well as their collaborators, co-authors and readers. Re-use and subsequent citation of methods may increase the impact of the work beyond the findings of the specific study. Reproducibility also has potential to improve the peer-review process. Currently, editors and reviewers often must trust that the authors' performed analyses are correct and suitable based on the limited information given in most manuscripts and supplementary information. The alternative is for a reviewer to ask the journal to request the data and code and repeat the analysis themselves. However, notwithstanding the obvious additional time spent to do this, reviewers unfamiliar with the language used to script the data handling (checking, correcting, rearranging and summarising data) and analysis have limited capacity to reproduce the results or even assess the reproducibility of the analyses given the information provided by the authors. Asking reviewers to reproduce results is therefore not a workable solution.

A major advantage of reproducible research is that it can inform your own future research. As we increasingly re-use data in different ways, and in combination with other data types, knowing the data analysis pipeline of a previous project can speed up future projects. For example, decisions taken to exclude data from the original raw data file in the derived data product can easily be reversed to extend analyses beyond the original use. Collaborative data collection efforts are increasingly common in ecology and can present significant data processing challenges as data are provided by participants in different formats and with variations on the stated protocol. The series of decisions taken to go from raw data to derived data product should be transparent to enable reversal or alternative data cleaning decisions to be implemented. Variation in results between analyses using the same original raw data is entirely possible due to different decisions at the data handling stage.

Fundamental principles of accessible reproducibility

There are two key steps to assessing the reproducibility of a piece of research:
1) Reproducibility: is sufficient information available to understand and reproduce the data handling, analysis and modelling?
2) Accessibility of reproducibility: can the analysis be reproduced with reasonable effort?

The first step can be achieved by providing data and code with a description of the processes followed, but this may not be particularly accessible. Accessibility increases as the data, code, outputs, and explanation of the analysis are better integrated, leading in turn to an increased ability to reproduce results. The more detail given here, the easier it is to reproduce the results whilst understanding the data handling decisions made.

Journal of Ecology Editorial

While provision of the data and a description of the data handling and analysis in the methods section of paper may be sufficient for reproducing relatively straightforward, commonplace analyses, the reproducibility of more complex data handling and analyses can be very difficult to assess at the peer review and post-publication stages. To facilitate easier assessment of reproducibility that can be used in the pre- and post-publication phases of a study, we recommend one recent advance in reproducible research: the ability to combine text and code to produce dynamic documents that record, demonstrate and enable reproduction of the process of data cleaning and analysis.

Dynamic documents for reproducible research

The basic format in which we publish our professional research is gratifyingly similar to the way in which we learn to write up our experiments in school, from the introduction that states the background, questions and hypotheses, through the materials and methods, to the results and conclusions. However, the final write-up is only as good as the data and analysis on which it is based. While spreadsheets are popular in the business world it is generally recommended that they are not used in the scientific process beyond data entry (with the data stored in a simple, widely used, non-proprietary file type such as a .csv or .txt file with an archived back up). In particular, data handling (or 'wrangling': checking, correcting, re-arranging, and summarising data) has often been done using a spreadsheet with no audit trail. But to be reproducible it should be done using a script, annotated with explanatory comments. Even potential errors should be left in the raw data and corrected as part of the scripted handling process. In practice, data handling and analysis are often poorly documented: from the rationale, through exploratory stages, to the formal testing and inference. A fundamental concept is that analyses should be scripted – that is conducted using a computer programme (script) - rather than an unrecorded (and therefore unrepeatable) series of mouse clicks using menu-driven software.

While scripts are an essential part of reproducible research analysis, their mixture of computer code and brief explanatory comments means they are not the most accessible documents. Various open-source software options now exist to allow scientists to easily annotate their scripts in a way that makes it simple to produce a document that gives a fully readable account of the analysis. Popular examples of such software include Jupyter Notebooks, the R Markdown package and the new multi-language Quarto application. This software produces documents that weave together the input code and the software-generated output (text, tables and figures) together with the author's explanatory text to produce an understandable narrative of the analysis process from start to finish (Fig. 1). These documents are dynamic – if the code is edited or data altered, then the changes are implemented in the next version of the output document. Options even exist for recording the versions of the software used, and for version control (tracking changes as the dynamic document changes over time) – for example, using GitHub. The resulting document can then be downloaded as an html, PDF or other word processing file type to be shared. A good dynamic document walks the editors, reviewers and readers through the whole analysis process. This addresses both the issue of the lack of clarity of the analysis and allows for any errors in data processing or poor statistical analysis to be spotted by editors or reviewers during the peer review process.

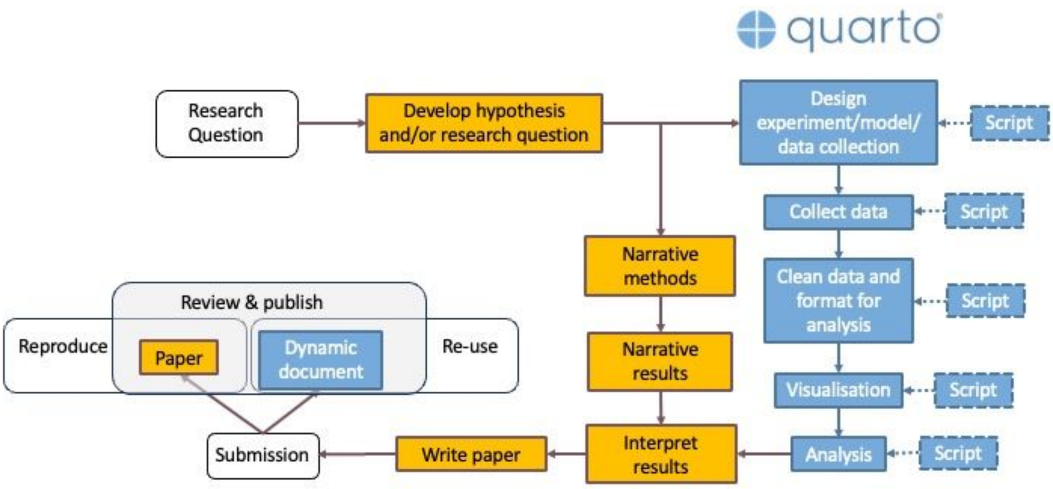

Fig. 1. A simplified linear workflow for production of a paper and associated dynamic document. Yellow boxes indicate elements of the workflow which may not require associated data handling/analysis and scripts (for example The first “Design experiment/model/data collection” step may use scripts for randomisation, to use power analysis to calculate replication etc). Blue boxes indicate elements that could be contained in a dynamic document such as Quarto. The dynamic document is submitted for review and published together with the paper and provides a resource for re-use, modification and reproduction of elements of the paper.

Reproducible Research Publishing

Making research more reproducible is no simple task because the causes of the current reproducibility crisis are complex and varied. However, the recent advances for documenting analysis workflows will help improve the situation. Science, including ecology, should be moving towards being able to repeat all analyses. Journals can help encourage these initiatives through facilitating use of reproducibility tools and the development and enforcement of appropriate submission requirements. The British Ecological Society has published two free Guides to Better Science on the subject of Reproducible Code (Cooper & Hsing, 2017) and Data Management (Harrison, 2018) which we recommend for full, worked through examples and further advice. While this could be seen as an increase in the administrative burden of submitting a paper, we encourage researchers to embrace the practice of creating dynamic documents as a critical and routine step in our process. The move to the use of dynamic documents brings notable improvements for little additional effort, and the satisfaction of knowing that we are improving the reproducibility of our research as well as that of others, ultimately improving science. Their use also eliminates many reviewer questions, and the work involved in responding to them.

Journal of Ecology Editorial

Reproducible Research Documents for Journal of Ecology

Journal of Ecology now therefore encourages the submission of supplementary dynamic documents to improve the reproducibility and transparency of research published in the journal.

Journal of Ecology already mandates the archiving of data and encourages the archiving of code used for data handling and analysis. Here we go further and encourage the provision of dynamic documentation for all stages of the data and analysis pipeline as part of the supplementary information submitted with the main manuscript (these could be produced using Jupyter Notebooks, R Markdown, Quarto or other similar options).

In terms of the journal editorial process there are two options. At submission, as part of the supplementary information, we encourage the provision of a reproducible research document in PDF form (in the software terminology the rendered or knitted document form) preferably with the executable dynamic document (e.g. the Quarto .qmd document or R markdown .rmd files). The static (pdf) and/or dynamic (executable) reproducibility documents will be available to editors and reviewers during peer review, and reviewers can tick a box to tell us when they have used them, so we can assess their utilisation. It is worth noting that any scrollable elements may be made static in the PDF format, so code should be arranged to avoid any crucial information being missed in this conversion.

After acceptance we encourage authors to archive the executable dynamic document online, together with the data (in an approved permanent repository like Dryad, Zenodo etc.). The static reproducibility document can be published with the paper to aid readers and enable post-publication reproducibility. We provide some examples (from our own work) as supplementary material to this editorial. These are not intended as ideal models but provide a range of past attempts to give a sense of the types of documents that can aid reviewers and readers and contribute to the further development of reproducible research methods.

In summary, authors submitting to *Journal of Ecology* are encouraged to provide readers of the published paper with access to three file types: (i) the data; (ii) the executable dynamic document which includes the code; and (iii) a static reproducibility PDF file that integrates and annotates the input code with the statistical output. These documents will help to fix the cracks causing the reproducibility crisis in ecology, whilst also strengthening the quality of authors' work that is published in *Journal of Ecology*.

Acknowledgements

YMB receives financial support from a Co-Centre award number 22/CC/11103. The Co-Centre award is managed by Research Ireland, Northern Ireland's Department of

Agriculture, Environment and Rural Affairs (DAERA) and UK Research and Innovation (UKRI), and supported via UK's International Science Partnerships Fund (ISPF), and the Irish Government's Shared Island initiative.

Author contributions

Yvonne Buckley and Andrew Hector conceived the ideas, and led on the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Conflict of interest

All the authors are *Journal of Ecology* editors and this represents our views.

References

Alston, J. M., and Rick, J. A. (2020). A Beginner's Guide to Conducting Reproducible Research. *Bull Ecol Soc Am* 102(2): e01801. https://doi.org/10.1002/bes2.1801.

Archmiller, A.A., Johnson, A.D., Nolan, J., Edwards, M., Elliott, L.H., Ferguson, J.M., Iannarilli, F., Vélez, J., Vitense, K., Johnson, D.H. and Fieberg, J. (2020). Computational Reproducibility in The Wildlife Society's Flagship Journals. *Journal of Wildlife Management* 84: 1012-1017. https://doi.org/10.1002/jwmg.21855.

Essawy, B.T., Goodall, J.T., Voce, D., Morsy, M.M., Sadler, J.M., Choi, Y.D., Tarboton, D.G., Malik, T. (2020) A taxonomy for reproducible and replicable research in environmental modelling. *Environmental Modelling & Software* 134. https://doi.org/10.1016/j.envsoft.2020.104753

British Ecological Society (2018). Guides to Better Science: Data Management. https://www.britishecologicalsociety.org/wp-content/uploads/2019/06/BES-Guide-Data-Management-2019.pdf

Cooper *et al.* (2017). Guides to Better Science: Reproducible Code. https://www.britishecologicalsociety.org/wp-content/uploads/2019/06/BES-Guide-Reproducible-Code-2019.pdf

Gould et al. (2023) Same data, different analysts: Variation in effect sizes due to analytical decisions. EcoEvoRxiv. https://doi.org/10.32942/X2GG62.

Ioannidis JP. (2005) Why most published research findings are false. PLoS Med. Aug;2(8):e124. https://doi.org/10.1371/journal.pmed.0020124

Leek J.T. & Peng R.D. (2015) *P* values are just the tip of the iceberg. *Nature* 520: 612.

Supplementary Materials

Journal of Ecology Editorial

These examples (from our own work; newest to oldest) are not intended as ideal models but provide a range of past attempts to give a sense of the types of documents that can aid reviewers and readers and contribute to the further development of reproducible research methods.

1)
Jackson, J. et al. (2024). Experimental drought reduces the productivity and stability of a calcareous grassland. Journal of Ecology, 00, 1–15. https://doi.org/10.1111/1365-2745.14282.
The reproducible research files are accessible via the Data Availability Statement section: "All code, output and data used in the current study are archived using the Zenodo repository: https://doi.org/ 10.5281/zenodo.8135588 (Jackson et al., 2023), which were created from the following GitHub repository: https:// github.com/jjackson-eco/raindrop_biodiversity_analysis.".

2).
Veryard, R. *et al. (2023) Positive effects of tree diversity on tropical forest restoration in a field-scale experiment*. *Science Advances* 9, eadf0938. DOI:10.1126/sciadv.adf0938.
The reproducible research documents are downloadable (600.38 KB) as: Other Supplementary Material for this manuscript includes the following: Other file **DOWNLOAD**).

3)
Jackson, J. et al. (2022). Short-range multispectral imaging is an inexpensive, fast, and accurate approach to estimate biodiversity in a temperate calcareous grassland. Ecology and Evolution, 12, e9623. https://doi.org/10.1002/ece3.9623.
The reproducible research files are accessible via the Data Availability Statement section: "The code, data, and figures for the current study can be found in the following Zenodo archive with DOI 10.5281/zenodo.7043832, which was made from the following GitHub repository https://github.com/jjackson-eco/multispectral_biodiversity.".

4)
Tuck SL, et al. (2016) *The value of biodiversity for the functioning of tropical forests: insurance effects during the first decade of the Sabah biodiversity experiment*. ***Proc. R. Soc. B*** 283: 20161451.
The reproducible research document is provided as supplementary file: rspb20161451supp5.pdf.