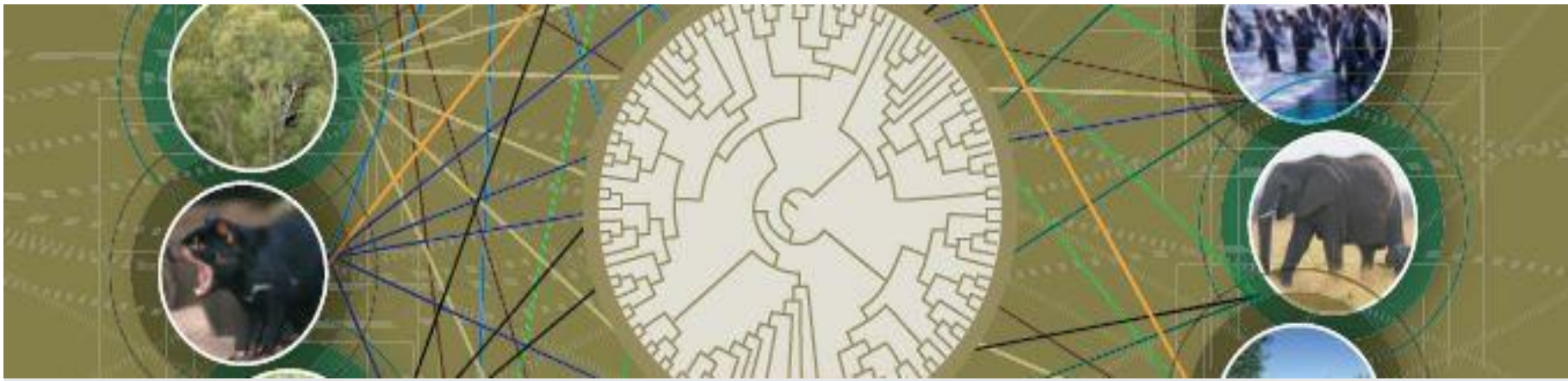
Part of SPRINGER NATURE



Data and Scientific Journals

Patrick Goymer

Chief Editor

Nature Ecology & Evolution

nature research

The big data era in biology

A different way of doing science

- Data led
- Does not fit always fit with hypothesis-testing idealised view
- Depends on openness and collaboration
- Ownership of data becomes an issue
- Requires new tools and attitudes
- Wide range of attitudes across disciplines
- Importance of meta-analysis
- Journals can innovate

But it is not without controversy

The NEW ENGLAND JOURNAL of MEDICINE

HOME | ARTICLES & MULTIMEDIA | ISSUES | SPECIALTIES & TOPICS | FOR AUTHORS | CME

EDITORIAL

Data Sharing

Dan L. Longo, M.D., and Jeffrey M. Drazen, M.D.

N Engl J Med 2016; 374:276-277 | January 21, 2016 | DOI: 10.1056/NEJMe1516564

Share:

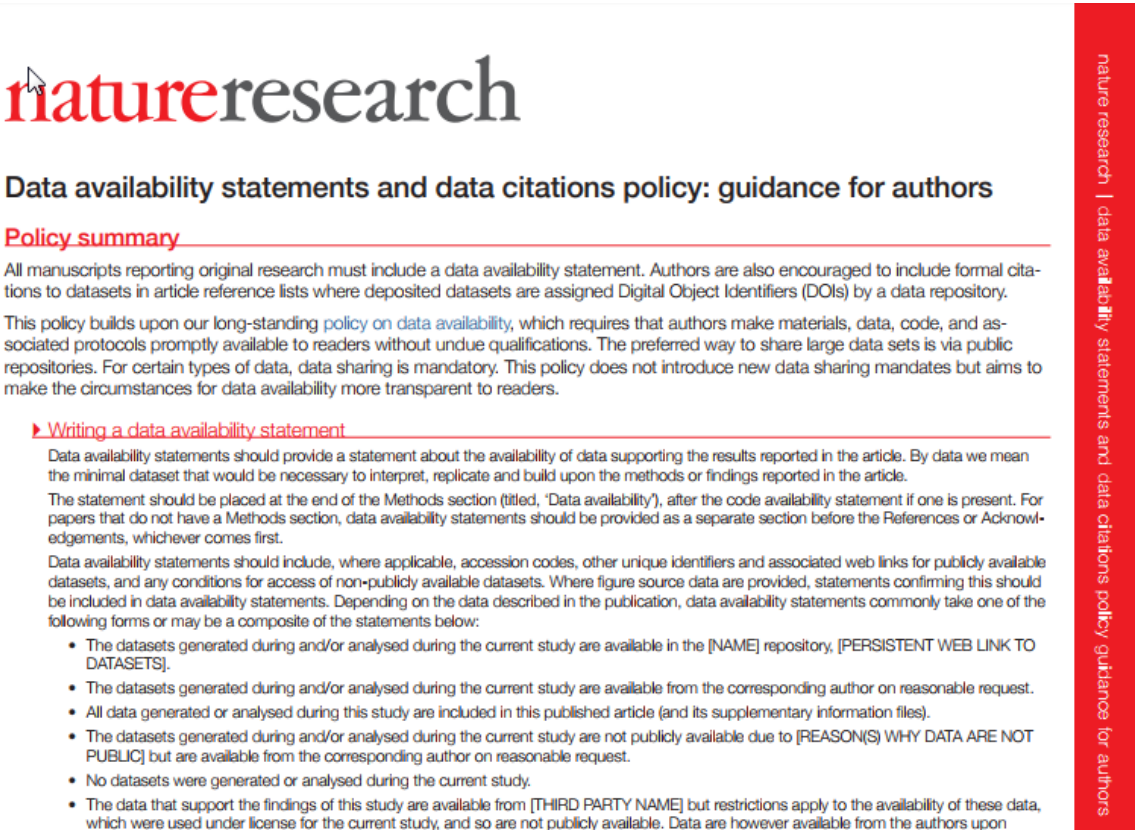
Article | References | Citing Articles (76) | Letters | Metrics

The aerial view of the concept of data sharing is beautiful. What could be better than having high-quality information carefully reexamined for the possibility that new nuggets of useful data are lying there, previously unseen? The potential for leveraging existing results for even more benefit pays appropriate increased tribute to the patients who put themselves at risk to generate the data. The moral imperative to honor their collective sacrifice is the trump card that takes this trick.

However, many of us who have actually conducted clinical research, managed clinical studies and data collection and analysis, and curated data sets have concerns about the details. The first concern is that someone not involved in the generation and collection of the data may not understand the choices made in defining the parameters. Special problems arise if data are to be combined from independent studies and considered comparable. How heterogeneous were the study populations? Were the eligibility criteria the same? Can it be assumed that the differences in study populations, data collection and analysis, and treatments, both protocol-specified and unspecified, can be ignored?

Data availability statements

- https://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf



nature research

Data availability statements and data citations policy: guidance for authors

Policy summary

All manuscripts reporting original research must include a data availability statement. Authors are also encouraged to include formal citations to datasets in article reference lists where deposited datasets are assigned Digital Object Identifiers (DOIs) by a data repository.

This policy builds upon our long-standing policy on data availability, which requires that authors make materials, data, code, and associated protocols promptly available to readers without undue qualifications. The preferred way to share large data sets is via public repositories. For certain types of data, data sharing is mandatory. This policy does not introduce new data sharing mandates but aims to make the circumstances for data availability more transparent to readers.

Writing a data availability statement

Data availability statements should provide a statement about the availability of data supporting the results reported in the article. By data we mean the minimal dataset that would be necessary to interpret, replicate and build upon the methods or findings reported in the article.

The statement should be placed at the end of the Methods section (titled, 'Data availability'), after the code availability statement if one is present. For papers that do not have a Methods section, data availability statements should be provided as a separate section before the References or Acknowledgements, whichever comes first.

Data availability statements should include, where applicable, accession codes, other unique identifiers and associated web links for publicly available datasets, and any conditions for access of non-publicly available datasets. Where figure source data are provided, statements confirming this should be included in data availability statements. Depending on the data described in the publication, data availability statements commonly take one of the following forms or may be a composite of the statements below:

- The datasets generated during and/or analysed during the current study are available in the [NAME] repository, [PERSISTENT WEB LINK TO DATASETS].
- The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.
- All data generated or analysed during this study are included in this published article (and its supplementary information files).
- The datasets generated during and/or analysed during the current study are not publicly available due to [REASON(S) WHY DATA ARE NOT PUBLIC] but are available from the corresponding author on reasonable request.
- No datasets were generated or analysed during the current study.
- The data that support the findings of this study are available from [THIRD PARTY NAME] but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon

nature research | data availability statements and data citations policy guidance for authors

Mandatory data deposition

Mandates for specific datasets

For the following types of data set, submission to a community-endorsed, public repository is mandatory. Accession numbers must be provided in the paper. Examples of appropriate public repositories are listed below.

Mandatory deposition	Suitable repositories
Deposition	PRIDE
Protein sequences	Uniprot
DNA and RNA sequences	Genbank
	DNA DataBank of Japan (DDBJ)
	EMBL Nucleotide Sequence Database (ENA)
DNA and RNA sequencing data	NCBI Trace Archive
	NCBI Sequence Read Archive (SRA)
Genetic polymorphisms	dbSNP
	dbVar
	European Variation Archive (EVA)
Linked genotype and phenotype data	dbGAP
	The European Genome-phenome Archive (EGA)
Macromolecular structure	Worldwide Protein Data Bank (wwPDB)
	Biological Magnetic Resonance Data Bank (BMRB)
	Electron Microscopy Data Bank (EMDB)
Microarray data (must be MIAME compliant)	Gene Expression Omnibus (GEO)
	ArrayExpress
Crystallographic data for small molecules	Cambridge Structural Database

Other repositories

- Scientific Data hosts a comprehensive list on their website: https://www.nature.com/sdata/policies/repositories
- Ecological data fits into few of the structured repositories, but PANGAEA is a possibility.
- Therefore Figshare and Dryad are used in majority of cases.
- Fairsharing is another source of information on databases and policies: https://fairsharing.org/

PANGAEA


PANGAEA.
Data Publisher for Earth & Environmental Science
SEARCH
SUBMIT
ABOUT
CONTACT
Submit Data
Welcome to PANGAEA® Data Publisher
Our services are generally open for archiving, publishing, and re-usage of data. The World Data Center PANGAEA is member of the ICSU World Data System.
ALL TOPICS
Search for measurement type, author name, project, taxa,...
OCEANS
(95220)
LITHOSPHERE
(37802)
BIOLOGICAL CLASSIFICATION
(28449)
ATMOSPHERE
(23822)
PALEONTOLOGY
(22458)

Data available on request

- Journals are moving away from this
- Many no longer allow it and demand deposition in a repository
- However, sometimes this deposition can have an embargo attached
- In most cases, authors are prepared to use a repository when asked to do so.
- One size does not fit all

A case study of good data practice

nature ecology & evolution

PERSPECTIVE

PUBLISHED: 23 MAY 2017 | VOLUME: 1 | ARTICLE NUMBER: 0160

Our path to better science in less time using open data science tools

Julia S. Stewart Lowndes[1]*, Benjamin D. Best[2], Courtney Scarborough[1], Jamie C. Afflerbach[1], Melanie R. Frazier[1], Casey C. O'Hara[1], Ning Jiang[1] and Benjamin S. Halpern[1,3,4]

Reproducibility has long been a tenet of science but has been challenging to achieve—we learned this the hard way when our old approaches proved inadequate to efficiently reproduce our own work. Here we describe how several free software tools have fundamentally upgraded our approach to collaborative research, making our entire workflow more transparent and streamlined. By describing specific tools and how we incrementally began using them for the Ocean Health Index project, we hope to encourage others in the scientific community to do the same—so we can all produce better science in less time.

Science, now more than ever, demands reproducibility, collaboration and effective communication to strengthen public trust and effectively inform policy. Recent high-profile difficulties in reproducing and repeating scientific studies have put the spotlight on psychology and cancer biology[1–3], but it is widely acknowledged that reproducibility challenges persist across scientific disciplines[4–6]. Environmental scientists face potentially unique challenges in achieving goals of transparency and reproducibility because they rely on vast amounts of data spanning natural, economic and social sciences that create semantic and synthesis issues exceeding those for most other disciplines[7–9]. Furthermore, proposed environmental solutions can be complex, controversial and resource intensive, increasing the need for scientists to work transparently and efficiently with data to collaborative and openly shared and communicated science—with an aim of inspiring others. Our story is only one potential path because there are many ways to upgrade scientific practices—whether collaborating only with your 'future self' or as a team—and they depend on the shared commitment of individuals, institutions and publishers[6,16,17]. We do not review the important, ongoing work regarding data management architecture and archiving[8,18], workflows[11,19–21], sharing and publishing data[22–25] and code[25–27], or how to tackle reproducibility and openness in science[28–32]. Instead, we focus on our experience, because it required changing the way we had always worked, which was extraordinarily intimidating. We give concrete examples of how we use tools and practices from data science, the discipline of turning raw data into understanding[33]. It

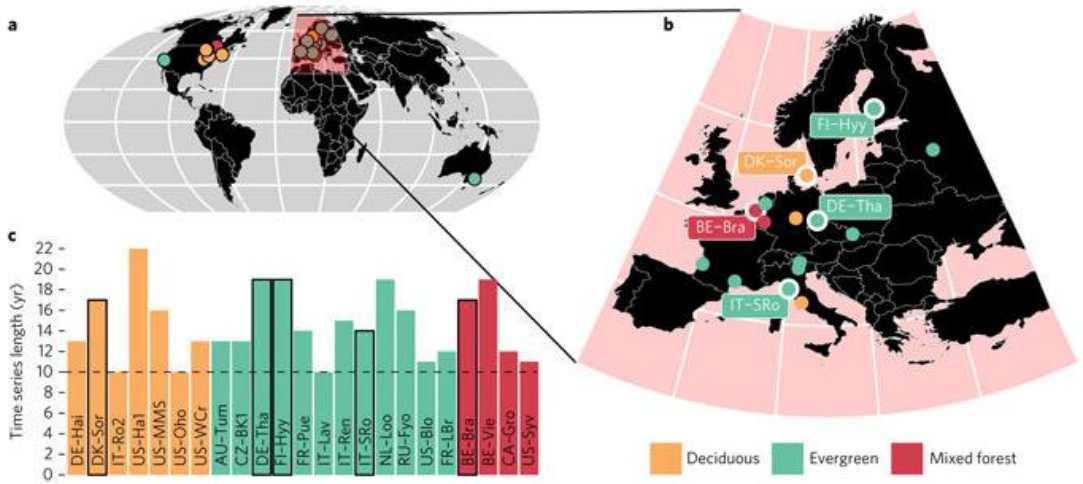
Long term ecological research

https://www.nature.com/articles/s41559-017-0306-4

nature ecology & evolution

Figure 1: Spatial distribution of the analysed forest sites.

From: Ecosystem functioning is enveloped by hydrometeorological variability



a, The 23 sites with long-term (≥10 yr) micrometeorological and NEE measurements. **b**, European sites for which TRW and AGB data are also available (white circles). **c**, Length of the analysed time series of micrometeorological and eddy covariance measurements (the five European sites with additional measurements are highlighted in black). Different colours correspond to different forest types.

Multi-site ecology

Ambio
February 2016, Volume 45, Issue 1, pp 29–41

Contributions of a global network of tree diversity experiments to sustainable forest plantations

Authors | Authors and affiliations

Kris Verheyen, Margot Vanhellemont, Harald Auge, Lander Baeten, Christopher Baraloto, Nadia Barsoum, Simon Bilodeau-Gauthier, Helge Bruelheide, Bastien Castagneyrol, Douglas Godbold, Josephine Haase, Andy Hector, Hervé Jactel, Julia Koricheva, Michel Loreau, show 13 more

Open Access | Perspective
First Online: 12 August 2015

Abstract

The area of forest plantations is increasing worldwide helping to meet timber demand and protect natural forests. However, with global change, monospecific plantations are increasingly vulnerable to abiotic and biotic disturbances. As an adaption measure we need to move to plantations that are more diverse in genotypes, species, and structure, with a design underpinned by science. TreeDivNet, a global network of tree diversity experiments, responds

Meta-analysis

Methods in Ecology and Evolution

View issue TOC
Volume 8, Issue 6
June 2017
Pages 777–784

Commentary

Will your paper be used in a meta-analysis? Make the reach of your research broader and longer lasting

Katharina Gerstner ✉, David Moreno-Mateos, Jessica Gurevitch,
Michael Beckmann, Stephan Kambach, Holly P. Jones, Ralf Seppelt

First published: 25 March 2017 Full publication history

DOI: 10.1111/2041-210X.12758 View/save citation

Cited by (CrossRef): 0 articles Check for updates Citation tools ▾

Am score 214

Funding Information

Summary

1. Ecological and evolutionary research increasingly uses quantitative synthesis of primary research studies (meta-analysis) for answering fundamental questions, informing environmental policy and summarizing results for decision makers.

Reporting checklists

nature research

Corresponding author(s):

☐ Initial submission ☐ Revised version ☐ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.

the data paper

SCIENTIFIC DATA

- A clear, peer reviewed description of data, to maximize usage
- Citable publications that give credit for reusable data

Get Credit for Sharing Your Data

Publications will be indexed and citeable.

Open-access

Articles are published by default under a Creative Commons Attribution licence (CC BY). Each publication supported by CC0 metadata.

Focused on Data Reuse

All the information others need to reuse the data; no interpretative analysis, or hypothesis testing

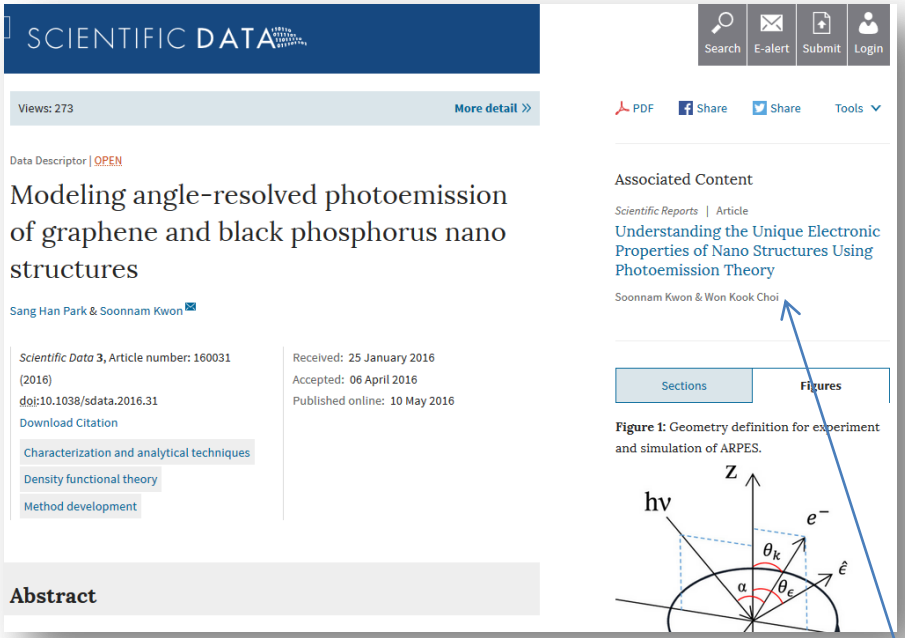
Peer-reviewed

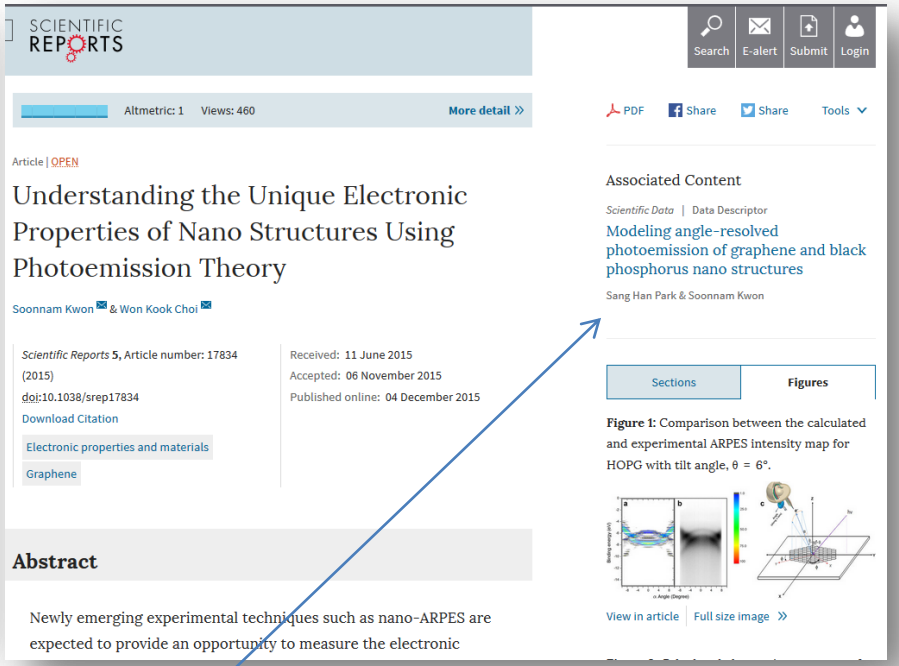
Rigorous peer-review focused on technical data quality and reuse value

Promoting Community Data Repositories

Not a new data repository; data stored in community data repositories

Complementing other journals





Bidirectional links

A Data Descriptor

SCIENTIFIC DATA

MENU SCIENTIFIC DATA

Search E-alert Submit Login

Altmetric: 32 Citations: 1 More detail »

PDF Share Share Tools

Data Descriptor | OPEN

Spatially-explicit models of global tree density

Henry B. Glick, Charlie Bettigole, Daniel S. Maynard, Kristofer R. Covey, Jeffrey R. Smith & Thomas W. Crowther

Scientific Data **3**, Article number: 160069 (2016)
doi:10.1038/sdata.2016.69
Download Citation

Received: 18 April 2016
Accepted: 01 July 2016
Published online: 16 August 2016

Ecological modelling Forestry Macroeco

Associated Content

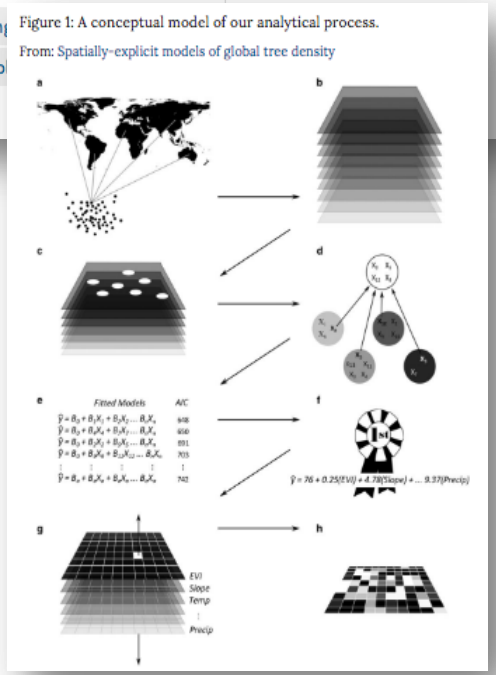
Nature | Article

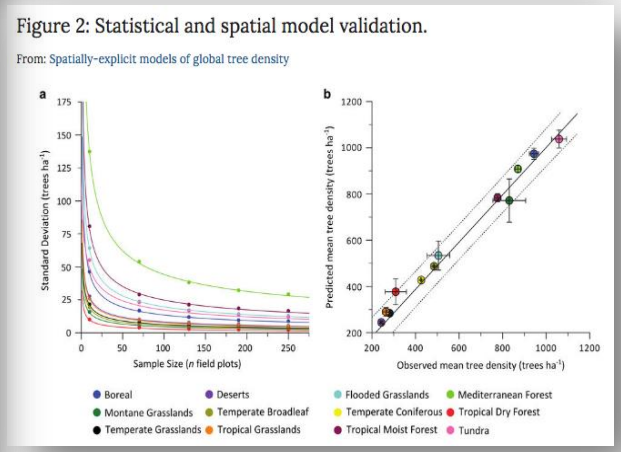
Mapping tree density at a global scale

T. W. Crowther, H. B. Glick [...] M. A. Bradford

Sections Figures References

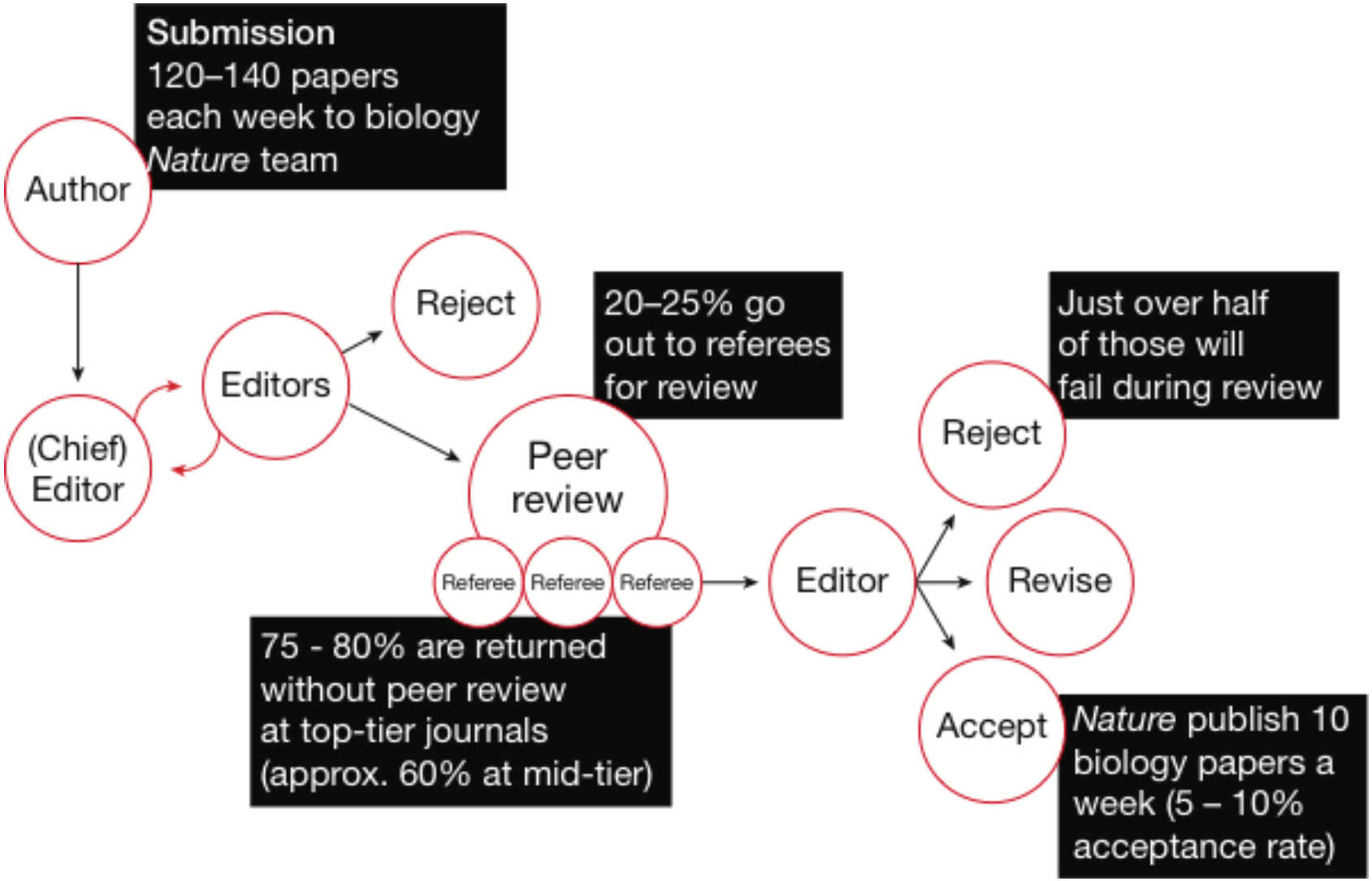
Abstract
Background & Summary
Methods
Technical Validation
Data Records





- **Expands on related paper at *Nature*, published ~ 1 yr earlier**
- The full underlying data (~4 GB) are richly described and openly available for others to use (CC BY).
- Data are hosted by figshare and tightly linked with the article.

The editorial process: statistics



What Nature editors are looking for…

- Technically solid studies
- Important to the field but of interest to a broad audience

Amongst our considerations are…

- Conceptual vs. incremental advance
- Depth of findings (mechanism, physiological relevance, generality)
- Community resource
- Importance across disciplines
- Potential to stimulate further advances

Manuscript transfer service

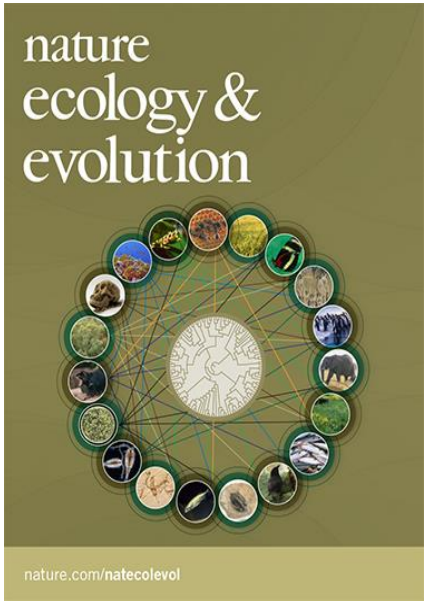
nature COMMUNICATIONS

SCIENTIFIC REPORTS

What is *Nature Ecology & Evolution?*

- *Nature Ecology & Evolution* presents cutting-edge research across the full spectrum of ecology and evolutionary biology, encompassing approaches at the molecular, organismal, population, community and ecosystem levels, as well as relevant parts of the social sciences.

- *Nature Ecology & Evolution* provides a place where all researchers and policymakers interested in all aspects of life's diversity can come together to learn about the most accomplished and significant advances in the field and to discuss topical issues.



- Ecology and evolution are key to all the global challenges – e.g. **biodiversity loss, food security, climate change, antibiotic resistance, healthy oceans, cancer.**

- No other single journal offers primary and secondary content across the whole field, from fossils to conservation, molecules to behaviour, medicine to agriculture.

Thanks for listening. Please find me during the conference for a chat and a hat. And feel free to email me with any questions: p.goymer@nature.com