IOP Publishing Environ. Res. Lett. 00 (2016) 000000

Environmental Research Letters

OPEN ACCESS

RECEIVED
18 May 2016

REVISED
29 June 2016

ACCEPTED FOR PUBLICATION
18 July 2016

PUBLISHED
DD MM 2016

LETTER

Running an open experiment: transparency and reproducibility in soil and ecosystem science

Ben Bond-Lamberty[1], **A Peyton Smith**[2] **and Vanessa Bailey**[2]

[1] Joint Global Change Research Institute, DOE Pacific Northwest National Laboratory, College Park, MD, USA
[2] Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA

E-mail: bondlamberty@pnnl.gov

Keywords: open data, soil science, reproducible research, open science

Abstract

Researchers in soil and ecosystem science, and almost every other field, are being pushed—by funders, journals, governments, and their peers—to increase transparency and reproducibility of their work. A key part of this effort is a move towards *open data* as a way to fight post-publication data loss, improve data and code quality, enable powerful meta- and cross-disciplinary analyses, and increase trust in, and the efficiency of, publicly-funded research. Many scientists however lack experience in, and may be unsure of the benefits of, making their data and fully-reproducible analyses publicly available. Here we describe a recent 'open experiment', in which we documented every aspect of a soil incubation online, making all raw data, scripts, diagnostics, final analyses, and manuscripts available in real time. We found that using tools such as version control, issue tracking, and open-source statistical software improved data integrity, accelerated our team's communication and productivity, and ensured transparency. There are many avenues to improve scientific reproducibility and data availability, of which is this only one example, and it is not an approach suited for every experiment or situation. Nonetheless, we encourage the communities in our respective fields to consider its advantages, and to lead rather than follow with respect to scientific reproducibility, transparency, and data availability.

1. Introduction

Science is becoming increasingly collaborative and data-intensive [1]; in conjunction with revolutions in Internet-based communication, this has created new research opportunities across former geographic and disciplinary barriers. At the same time, many factors are pushing scientists to increase data access and use 'best practices' in dealing with data and code [2, 3].

Scientific journals are adopting increasingly stringent data access and deposition policies, e.g. those of *Scientific Data*[3], *PLoS One*[4], and *Science*[5]. These policies generally share common assumptions and goals: maximizing access to data; encouraging deposition into structured repositories as opposed to journal supplementary information; and specifying that it is not acceptable for authors to be solely responsible for ensuring data access. Funding agencies are moving in this direction as well, with organizations such as the US National Science Foundation[6] and the UK Wellcome Trust[7], along with many others, requiring explicit data management plans, unfettered reasonable access to primary data, and use of established repositories.

Finally, growing numbers of scientists are pushing for open science and data on moral and political grounds, as well as purely scientific ones, arguing that it is not acceptable to sequester taxpayer-funded research behind private publishers' paywalls [4]. A second focus revolves around ensuring reproducibility [5] and enabling larger synthetic activities. Such analyses [6] are made possible by the assembly of large, internally consistent data sets; examples in ecology,

[3] http://nature.com/sdata/data-policies
[4] http://journals.plos.org/plosone/s/data-availability
[5] http://sciencemag.org/authors/science-editorial-policies#data-deposition
[6] http://nsf.gov/bfa/dias/policy/dmp.jsp
[7] http://wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Data-management-and-sharing/index.htm

soil science, and biogeosciences of such databases include BAAD [7], TRY [8], FLUXNET [9], and SRDB [10]. Online databases and collaborative tools have also expanded the abilities of researchers to collaborate across large distances, both improving data access and facilitating multidisciplinary research partnerships [11].

Here we briefly discuss what we see as the primary arguments for data sharing and openness, and then describe a recent 'open experiment' example.

1.1. Repeatability and reproducibility

Reproducibility of experimental results is at the heart of science and a requirement for results to be accepted as factual [12]. Too often, however, sufficient details and data are not publicly available to even *repeat* a study (i.e. perform it again in a comparable manner, while not expecting exactly the same result). For example, surveying the 2000–2014 biomedical literature, Iqbal *et al* [13] found that none of 441 randomly-chosen studies provided raw data, and only one provided full protocols. The issues of reproducibility and repeatability in ecology (and its many related fields) have been raised and debated for years [14, 15] but ecological, soil science, and global change journals differ widely in their data deposition requirements, when such requirements exist at all.

As data and code have become increasingly intertwined, the availability of the latter has become a fundamental problem as well. Specialized modeling groups have worked to improve reproducibility and archiving practices [16]. More generally, scientists in all fields are increasingly building and using software in their work, though often without strong training in this area [3]. In addition to open-source data analysis languages such as R [17], scientific workflow systems such as Kepler[8] or Taverna[9] record information about the data processing, analytical process and decisions, and statistical analysis. Providing open code does not magically produce bug-free code, mistake-free analyses, or instantly better science [18]. But it does encourage authors to invest the time upfront to clean up their code, data, and documentation when a paper is written, rather than deferring this task—often until key details have been forgotten, if not forever. This also allows for real-time peer review of both code and data.

1.2. Data loss

Vines *et al* [19] published a shocking finding, based on a survey of 516 biology articles from 2 to 22 years old: the odds of a data set being available post-publication fell by 17% each year, and the chances that the contact author's email address still worked declined by 7% per year. Similarly, Reichman *et al* [20] estimated that less than 1% of ecological data collected is made available after publication, and noted, as an example, that much current and historical data relevant to the 2010 *Deepwater Horizon* oil spill are already inaccessible or lost [20]. In global change ecology, Wolkovich *et al* [6] reported that they were able to acquire only ~10% of other researchers' raw data sets in preparation for a meta-analysis [21].

Data loss hits ecosystem, soil, and global change ecology particularly hard, as climate changes make ecological data effectively irreproducible [6]: we can never remeasure exactly the same system state. This results in a critical need for syntheses and meta-analyses [14, 15], which depend on the existence of documented protocols and (ideally raw) data. A subtler data loss issue is the 'file drawer problem' [22], where unpublished but potentially valuable data are lost forever. Scientists' use of strong and consistent data curation practices [2, 23] can mitigate the problem, but both our anecdotal experience and quantitative studies [19] suggest that in the long term, data cannot be reliably preserved by individual researchers.

This in turns means that the role of established, structured data and code repositories is critical. These provide a much-needed improvement on 'Supplementary Information' sections accompanying journal articles, which have become *de facto* repositories of data, but often have inconvenient formats (e.g. PDF), use restrictions, and uncertain long-term availability [4, 24]. Most journals neither desire nor have the necessary expe[illegible]in data storage and management [14]. In contrast, the best repositories provide easy data uploading, immediate assignment of data digital object identifiers, and long-term stability and availability. They do not, however, eliminate all risks: for example, the genomics sciences have a long history of requiring sequences to be deposited, but there are multiple repositories of varying curatorial levels, data quality, and formats [25].

1.3. Institutional and social trust

Finally, there are longer-term issues of trust to consider—in particular, the public's trust in science and science's trust in people [26]. Both of these have been weakened by public controversies over particular issues—most particularly relevant to our fields of research, climate change [27]. In addition to such issues of trust in the *correctness* of science, there is that of trust in *utility* of science: why should the public fund scientists if the latter are not producing demonstrably replicable and reusable results? Why should politicians and other stakeholders not push for greater return on public investment in scientific research, and conclude that more data openness serves this cause [4]?

These reasons have led, at least in part, to governmental efforts to have the results of federally funded scientific research made available to the public, industry, and across the scientific community [28]. Strong

[8] https://kepler-project.org
[9] http://taverna.org.uk

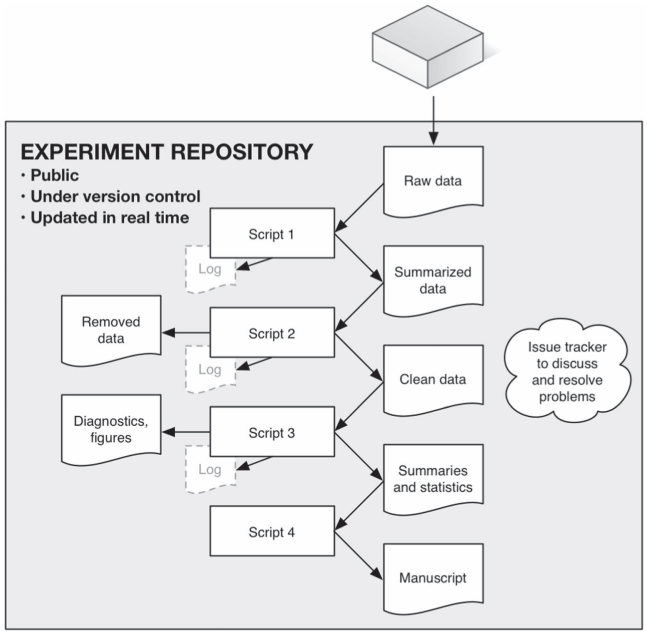

Figure 1. The 'repository' containing all code and data of a recent soil incubation experiment (29). From top to bottom: raw data from a lab analyzer is uploaded; an initial script processes data into a standardized, summarized form; a second script looks for inconsistencies and outliers in the data; a third constructs final data products, summaries, statistics, and figures; and a final one integrates these products to produce a submission-ready manuscript. (This schematic slightly simplifies the actual repository structure.) The repository is publicly available at https://github.com/bpbond/cpcrw_incubation, and throughout the experiment showed real-time updated diagnostics summarizing progress.

'open science' and 'open data' movements argue that the completeness of information provided by open science is fundamentally beneficial, complementing and perhaps replacing older systems for establishing trust within science [26], and between science and the public. These movements increasingly deny that 'the intellectual property rights of publishing (scientists)' [14] take precedence over all other factors, at least as far as publicly-funded research is concerned.

2. An open experiment: one example

In early 2015, we planned a laboratory incubation experiment to characterize the chemical and biological properties of sub-Arctic, active layer soils subjected to changes in temperature and moisture. In this experiment, we would measure greenhouse gas fluxes from soil cores over 100 days, and measure the cores' physical, chemical, and biological characteristics under temperature and moisture changes [29]. This required (i) a multidisciplinary team that was not located in one time zone; (ii) integrating a variety of different data; (iii) performing quality control and diagnostics rapidly, so if e.g. instrument problems arose we would lose only the minimum amount of time and data; (iv) tightly integrating data, statistical analyses, and manuscript results.

We used GitHub, a web-based *Git repository* hosting service, to store our data, scripts, and manuscripts. 'Git' is a popular, free, and open source version control software: it tracks all changes (when, what, by whom) made in a 'repository', a collection of folders and files. Many scientific and other users of Git use the 'GitHub' or similar web services, as they offer a wide variety useful additional functionality, particularly for teams or collaborative projects.

The design of our repository is shown in figure 1, and the repository itself can be found at https://github.com/bpbond/cpcrw_incubation. It consists of a series of scripts that feed their results from one to the next, starting with raw data and ending with final analyses, figures, and manuscripts.

This system had a number of characteristics:

- The entire data processing and analytical system is online and documented. It is written in R [17], an open-source and widely used language and environment for statistical computing and graphics.
- The version control system let us make incremental changes, work out problems, look at histories (i.e., who made what change when).
- An issue tracker let us discuss problems, reference changes in the repository, create to-do lists, and assign responsibilities.
- An informational webpage provided non-technical explanations of the experiment and broader project.
- Manuscripts were directly tied into the data system (see figure 1), with numerical results flowing directly into e.g. results sections. This meant that changes to the data (and thus statistical results, etc) propagated

automatically and consistently. Tools such as R Markdown[10] have made this process much easier to build and maintain.

- Log files provide an audit trail [15] of what analytical steps were taken and, critically, the specific versions used for each unit of software, including the main R system itself. This is critical as software changes with time, a potentially large problem in reproducing previous analyses.

We found a number of advantages to this system:

- The public nature of the repository encouraged us to flesh out documentation, use clean and clear coding, and think about the longer-term ramifications of decisions.
- Real-time diagnostics (figure 2) let the team, our DOE program managers, and other interested parties see at a glance the progress of the experiment.
- The issue tracker helped our team—which was not physically located all in one place—to communicate, discuss, and track 'issues' (i.e., problems or questions that arose).
- Investing the effort to set up a software pipeline before, not after, the experiment was performed meant that we could reliably and easily identify and diagnose problems.
- The real-time collaboration, and the public exposure of all stages inspired us to keep all activities moving quickly. In this case, it was less than nine months between sampling soil cores in the Alaskan forest and submitting two resulting manuscripts.
- Late changes to the analysis pipeline (for example, when we identified an incorrect calculation) did not result in time-consuming and error-prone pasting of new data values into our manuscript.
- Our project and experiment received more exposure and publicity than it otherwise would have.

3. Barriers and caveats

There were thus significant advantages to implementing and using an automated analytical pipeline in an open repository. It is worth considering in depth, however, the costs of such an approach, the investments required, and the barriers encountered.

- More upfront time is required in an open repository system, as the basic building blocks are assembled into an automated pipeline that runs from raw data to diagnostics to final products. While frustrating, this is generally good: automation makes later analytical stages much faster (see above), while openness provides a strong incentive to write clean, clear code from the beginning, for example, as opposed to deferring documentation.
- To our knowledge, there is no 'template' for an experimental repository such as the one used here. This results in unnecessary time spent 're-inventing wheels', in particular with respect to repository design decisions. Clear, flexible, and powerful templates will help scientists take advantage of repository-based and open science approaches.
- Effective data management in a repository system takes some programming skills. Basic programming is becoming an increasingly important part of most scientists' training, just as basic statistics has long been a critical skill. For scientists accustomed to spreadsheet-based data processing, investment in new skills and practices, or new partnerships with skilled programmers, will facilitate more open analyses.
- Version control (here, the Git software) is critical for tracking provenance, ensuring robustness, and efficiently sharing changes. It introduces significant complexity, however, particularly for any use beyond the most simple operations. In our experience, version control software remains too difficult for most scientists to use effectively[11]. We urge software developers to work closely with diverse science communities to develop more user-friendly tools.
- Git and GitHub are fundamentally designed for working with code, not data or long documents. This situation continues to improve on a technical level[12], but some operations (e.g., storing very large files; tracking small changes to columnar data; commenting on words or phrases) remain awkward or even impossible.

In summary, while tools and platforms such as Git, GitHub, RStudio, and RMarkdown have drastically improved the feasibility and accessibility of running an 'open experiment', they remain significant hurdles for many scientists. We applaud efforts to lower the technical and information hurdles to open science [30].

A number of caveats are also in order. The model we present here—a public repository updated throughout the experiment, analysis, and publication process—is far from perfect. For example, there are a number of 'best practices' of scientific computing [3] that we did not employ, in particular with respect to automation. In addition, this approach may not be

[10] http://rmarkdown.rstudio.com

[11] https://xkcd.com/1597/

[12] E.g., https://github.com/blog/1885-better-word-highlighting-in-diffs

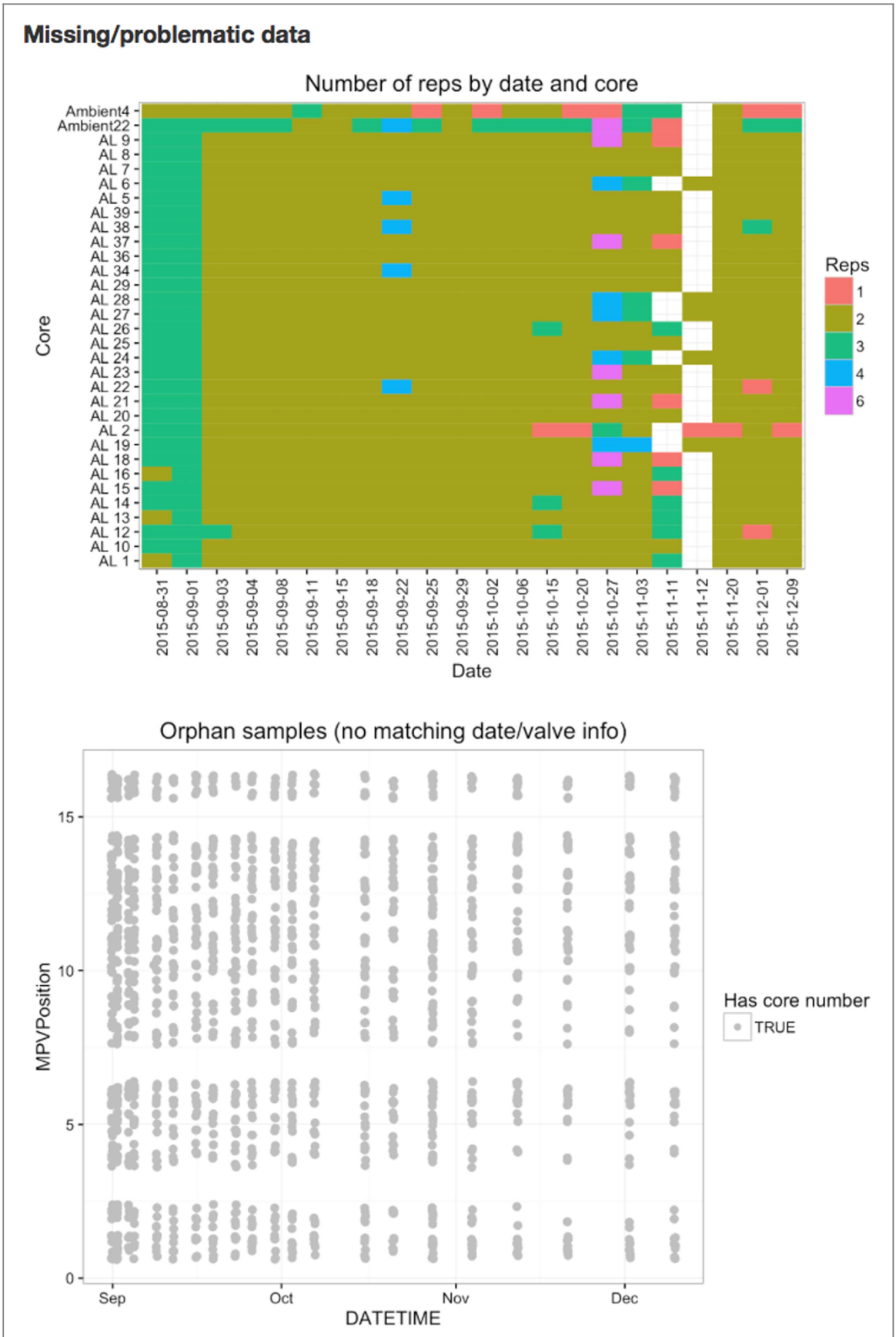

Figure 2. Two of the many diagnostic figures that were produced as the incubation experiment proceeded. Top: the number of replicate measurements made on each soil core (y), by date of measurement (x); the 'gap' at 2015-11-11/12 shows a potential problem in our data-processing pipeline, related to the instrument's clock crossing midnight. Bottom: data matches between multiplexer valve number (y) and soil core, by date of measurement (x); gray dots indicate that every valve has been correctly matched with a soil core, while a red dot would indicate a data entry problem or other error. A wide range of similar diagnostic graphics and statistics let us quickly identify problems or interesting patterns in the data. All diagnostics were publicly available in real time.

appropriate for, nor applicable to, all study types, even in our professional fields of soil, ecosystem, and global change science; it certainly is not appropriate for sensitive data, e.g. human-subject research.

More generally, by itself transparency does not guarantee repeatability and reproducibility [18], and it may raise new concerns about, for example, protecting scientists from harassment [31]. Nor will open science by itself fundamentally enhance public trust in science, as there are still many social and political challenges to overcome [26].

Scientists also frequently cite other concerns: about data sharing, being 'scooped', not receiving sufficient credit, and time constraints [6]. The benefits are not always clear to researchers who might otherwise be open to following open-data practices [24, 30]. We cannot easily dismiss all of these concerns: for example, does our community adequately credit researchers who contribute to global databases that subsequently produce high-impact papers? Such meta-analyses rely on the collection of primary data, and it is critical that field and experimental researchers' efforts are adequately valued and cited [32].

Finally, we recognize that, in general, our professional and career incentives do not yet align well with 'an open research culture' [33]. That is, scientists do not receive appropriate credit for datasets relative to publications, as hiring, promotion and tenure decisions all tend to reward publications, not datasets. The advent of 'data descriptor' articles, in journals such as *Scientific Data*, improves but does not solve this problem.

4. Conclusions

A variety of forces continue to push scientists towards more transparency in their methods, code, and data, with the goals of increasing reproducibility, enabling syntheses and meta-analyses, and improving trust in, and return from, publicly-funded science. The open experiment example we highlight here offers instrument-to-final product reproducibility and a very high level of transparency, although it is only one of a number of possible models [30]. Elements of this case study (e.g., the use of issue-tracking or version control software) might be usefully adopted in isolation, but we hope the entire experiment will be an example of individual scientists' decisions and practices [6] having a larger impact. We encourage the communities in our respective fields to consider its advantages, and to lead rather than follow with respect to scientific reproducibility, transparency, and data availability.

Acknowledgments

This research was supported by the Office of Science of the US Department of Energy as part of the Terrestrial Ecosystem Sciences Program. The Pacific Northwest National Laboratory is operated for DOE by Battelle Memorial Institute under contract DE-AC05-76RL01830.

Author contribution

BB-L designed the repository and script system described here, and wrote the manuscript with contributions from all authors.

References

[1] Adams J 2012 Collaborations: the rise of research networks *Nature* **490** 335–6
[2] Hart E *et al* 2015 Ten simple rules for digital data storage *PeerJ*
[3] Wilson G *et al* 2014 Best practices for scientific computing *PLoS Biol.* **12** e1001745
[4] Neylon C 2012 Science publishing: open access must enable open use *Nature* **492** 348–9
[5] Stodden V 2011 Trust your science? Open your data and code *Amstat News*
[6] Wolkovich E M, Regetz J and O'Connor M I 2012 Advances in global change research require open science by individual researchers *Glob. Change Biol.* **18** 2102–10
[7] Falster D S *et al* 2015 BAAD: a biomass and allometry database for woody plants *Ecology* **96** 1445
[8] Kattge J *et al* 2011 TRY—a global database of plant traits *Glob. Change Biol.* **17** 2905–35
[9] Baldocchi D D *et al* 2001 FLUXNET: a new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities *Bull. Am. Meteorol. Soc.* **82** 2415–34
[10] Bond-Lamberty B and Thomson A M 2010 A global database of soil respiration data *Biogeosciences.* **7** 1915–26
[11] Markowitz V M, Chen I-M A, Chu K, Pati A, Ivanova N N and Kyrpides N C 2015 Ten years of maintaining and expanding a microbial genome and metagenome analysis system *Trends Microbiol.* **23** 730–41
[12] Kuhn T 1962 *The Structure of Scientific Revolutions* (Chicage, IL: University of Chicago Press) 212p
[13] Iqbal S, Wallach J D, Khoury M J, Schully S D and Ioannidis J P A 2016 Reproducible research practices and transparency across the biomedical literature *PLoS Biol.* **14** e10023333
[14] Cassey P and Blackburn T M 2006 Reproducibility and repeatability in ecology *BioScience* **56** 958–9
[15] Ellison A M 2010 Repeatability and transparency in ecological research *Ecology* **91** 2536–9
[16] Thornton P E *et al* 2005 Archiving numerical models of biogeochemical dynamics *Eos* **86** 431–2
[17] R Development Core Team 2016 *R: A Language and Environment for Statistical Computing Version 3.2.3.* (Vienna, Austria: R Foundation for Statistical Computing)
[18] Easterbrook S M 2014 Open code for open science? *Nat. Geosci.* **7** 779–81
[19] Vines T H *et al* 2014 The availability of research data declines rapidly with article age *Curr. Biol.* **24** 94–7
[20] Reichman O J, Jones M B and Schildhauer M P 2011 Challenges and opportunities of open data in ecology *Science* **331** 703–5
[21] Wolkovich E M *et al* 2012 Warming experiments underpredict plant phenological responses to climate change *Nature* **485** 494–7
[22] Rosenthal R 1979 The file drawer problem and tolerance for null results *Psychological Bull.* **86** 638–41
[23] Rüegg J *et al* 2014 Completing the data life cycle: using information management in macrosystems ecology research *Front. Ecol. Environ.* **12** 24–30
[24] Molloy J C 2011 The open knowledge foundation: open data means better science *PLoS Biol.* **9** e1001195
[25] Lagesen K, Ussery D W and Wassenaar T M 2010 Genome update: the 1000th genome—a cautionary tale *Microbiology* **156** 603–8

[26] Grand A, Wilkinson C, Bultitude K and Winfield A F T 2012 Open science: a new 'trust technology'? *Sci. Commun.* **34** 679–89
[27] IPCC 2013 *Working Group I contribution to the IPCC 5th Assessment Report Climate Change 2013: The Physical Science Basis*
[28] Holdren J P 2013 Increasing Access to the Results of Federally Funded Scientific Research (Available from: https://whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf)
[29] Bond-Lamberty B, Smith A P and Bailey V L 2016 Temperature and moisture effects on greenhouse gas emissions from deep active-layer boreal soils *Biogeosciences* submitted (doi:10.5194/bg-2016-234)
[30] Hampton S E *et al* 2015 The Tao of open science for ecology *Ecosphere* **6** 1–13
[31] Lewandowsky S and Bishop D 2016 Research integrity: don't let transparency damage science *Nature* **529** 459–61
[32] Kueffer C *et al* 2011 Fame, glory and neglect in meta-analyses *Trends Ecol. Evol.* **26** 493–4
[33] Nosek B A *et al* 2015 Promoting an open research culture *Science* **348** 1422–5

Q3

QUERY FORM

JOURNAL: Environmental Research Letters

AUTHOR: B Bond-Lamberty *et al*

TITLE: Running an open experiment: transparency and reproducibility in soil and ecosystem science

ARTICLE ID: erlaa3469

The layout of this article has not yet been finalized. Therefore this proof may contain columns that are not fully balanced/matched or overlapping text in inline equations; these issues will be resolved once the final corrections have been incorporated.

We have been provided funding information for this article as below. Please confirm whether this information is correct. Biological and Environmental Research.

Page 6

Q1
Please check the details for any journal references that do not have a link as they may contain some incorrect information.

Page 6

Q2
Please provide the volume and page number or article number in reference [2].

Page 7

Q3
Please provide updated details for reference [29] if available.
