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

Ben Bond-Lamberty1, Peyton Smith2, Vanessa Bailey2

February 2, 2016

1Joint Global Change Research Institute, DOE Pacific Northwest National Laboratory, College Park, MD USA; [bondlamberty@pnnl.gov](mailto:bondlamberty@pnnl.gov)

2Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA USA

#### Abstract

Researchers in soil and ecosystem science, and almost every other field, as increasingly being pushed--by funders, journals, governments, and themselves--to ensure 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, and enable powerful meta- and cross-disciplinary analyses. Here we describe a recent "open experiment", in which we documented every aspect of a soil incubation experiment online. This involved making raw data, scripts, diagnostics, final analyses, and manuscripts publicly 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 communications, 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.

#### Introduction

Science is becomingly increasingly collaborative and data-intensive (Adams 2012). At the same time, a wide variety of factors are encouraging, at times pushing, scientists to increase data access and use "best practices" in dealing with data and code (Wilson et al. 2014, Hart et al. 2015).

Journals are adopting increasingly stringent data access and deposition policies, e.g. Scientific Data[[1]](#footnote-1), PLOS ONE[[2]](#footnote-2), and Science[[3]](#footnote-3). These policies generally share common assumptions and goals: maximizing access to scientific 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. Funders are moving in this direction as well, with organizations such as the U.S. National Science Foundation[[4]](#footnote-4) and the U.K. Wellcome Trust[[5]](#footnote-5), 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/political grounds, as well as purely scientific ones, in particular ensuring reproducibility (Stodden 2011) and enabling larger synthetic activities. Such analyses (Wolkovich et al. 2012b) are made possible by the assembly of large, internally consistent data sets. Examples in ecology, soil science, and biogeosciences include BAAD (Falster et al. 2015), TRY (Kattge et al. 2011), FLUXNET (Baldocchi et al. 2001), and SRDB (Bond-Lamberty and Thomson 2010).

We briefly discuss what we see as the primary arguments for data sharing and openness, before detailing a recent "open experiment" example.

**Reproducibility**

Reproducibility of experimental results is at the heart of science (Kuhn 1962). Too often, however, it is impossible to attempt, because sufficient details and data are not publicly available. For example, surveying the 2000-2014 biomedical literature, Iqbal et al. (2016) found that none of 441 randomly-chosen studies provided raw data, and only one provided full protocols. The issue of reproducibility in ecology (and its many related fields) has been raised and debated for years (Cassey and Blackburn 2006) but ecological, soil science, and global change journals differ widely in their data deposition requirements.

As code and data have become increasingly intertwined, the availability of the former has become a fundamental problem as well. Modeling groups have worked to improve reproducibility and archiving praactices (Thornton et al. 2005). Scientists in all fields are increasingly building and using software in their work, however, often without strong training in this area (Wilson et al. 2014). In addition to open-source data analysis languages such as R (R Development Core Team 2016), scientific workflow systems such as Kepler (<https://kepler-project.org>) or Taverna (<http://www.taverna.org.uk>) record information about the data processing, analytical process and decisions, and statistical analysis. Providing open code does not magically produce bug-free code, or mistake free analyses (Easterbrook 2014). 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, forever).

**Data loss**

Vines et al. (2014) published a shocking result, based on surveying 516 biology articles from 2 to 22 years old: the odds of a data set being available 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. (2011) 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 recent *Deepwater Horizon* oil spill are already inaccessible or lost (Reichman et al. 2011). In global change ecology, Wolkovich et al. (2012b) reported that only ~10% of raw data sets were eventually shared in their attempt to conduct a meta-analysis (Wolkovich et al. 2012a).

Data loss hits ecosystem, soil, and global change ecology particularly hard, as global to local climate changes make ecological data effectively irreproducible (Wolkovich et al. 2012b): we can never remeasure exactly the same system state. A subtler data loss issue is the "file drawer problem" (Rosenthal 1979), where unpublished but potentially valuable data are lost fovever. Scientists' use of strong and consistent data curation practices (Rüegg et al. 2014, Hart et al. 2015) can mitigate the problem, but all our anecdotal experience and quantitative studies (Vines et al. 2014) 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.

TODO - VLB: Supplementary Information sections have become repositories of data and of non-core analyses. How does our open experiment replace (improve upon?) SI? Thoughts? - journals not ready, and don't want (Cassey and Blackburn 2006).

**Institutional and social trust** (TODO)

Finally, there are longer-term issues of trust to consider with respect to reproducibility and openness.

Governmental efforts to have the results of federally funded scientific research made available to the public, industry, and across the scientific community (Holdren 2013).

* Trust in correctness of science: Potential to increase public trust in research results (e.g. climate change)
* Trust in utility of science: Why should the public fund scientists if the latter aren't producing demonstrably replicable and reusable results?

TODO: mention that of course authors have some rights. But what is being denied, increasingly, is that "the intellectual property rights of publishing ecologists" (Cassey and Blackburn 2006) take precedence over all other factors, at least as far as publicly-funded research is concerned.

#### 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. 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 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" web service, as it offers excellent performance and a wide variety useful additional functinality, 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 comprise a *pipeline*, start 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 (R Development Core Team 2016), 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).
* A issue tracker let us discuss problems, reference changes in the repository, create to-do lists, and assign responsibilities.
* Information webpage on the experiment and broader project.
* Manuscripts as well! (Including this one.) Results are linked directly into the manuscripts, so that if a result changes, it is updated automatically in the text. Tools such as R Markdown[[6]](#footnote-6) have made this process much easier to build and maintain.

Advantages:

* 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 our team, program managers, and interested other 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.

#### Conclusions

Transparency is no magic panacea ensuring reproducibility (Easterbrook 2014), and may raise genuine concerns about, for example, protecting scientists from harassment (Lewandowsky and Bishop 2016). Scientists also frequently cite other concerns about data sharing, for example being "scooped" not receiving sufficient credit, and time constraints (Wolkovich et al. 2012b). We cannot easily dismiss of these these: 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 (Kueffer et al. 2011). We also recognize that, in general, our professional and career incentives do not yet align well with "an open research culture" (Nosek et al. 2015)--that is, scientists do not receive 'proper' credit for datasets, relative to publications, even when the former may be much more broadly valueable.

The model we present here--a public repository updated throughout the experiment, analysis, and publication process--is far from perfect. First, there are a number of "best practices" of scientific computing (Wilson et al. 2014) that we did not employ: for example, there was no unit testing to ensure module correctness, nor an automated build tool automating the workflow (i.e., the individual scripts all have to be run manually). Second, it may not not be appropriate for applicable to all study types, even just in our professional fields of soil, ecosystem, and global change science. (And it certainly is not appropriate for sensitive human-subject data.) Third, Git and GitHub are designed for working with code, not data, although this situation continues to improve on a technical level[[7]](#footnote-7).

Nonetheless (TODO: it's a lot better than nothing, and offers instrument-to-final product reproducibility).

It doesn't have to be adopted wholesale: for example, the use of an issue tracker alone might be valuable for some groups and improve their science (though not reproducibility).

This is example, we hope, of the impact of individual scientists' decisions and practices (Wolkovich et al. 2012b).

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.

**References**

Adams, J. 2012. Collaborations: The rise of research networks. Nature **490**:335-336.

Baldocchi, D. D., E. Falge, L. Gu, R. J. Olson, D. Y. Hollinger, S. W. Running, P. M. Anthoni, C. Bernhofer, K. J. Davis, and R. D. Evans. 2001. FLUXNET: a new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. Bulletin of the American Meteorological Society **82**:2415-2434.

Bond-Lamberty, B., and A. M. Thomson. 2010. A global database of soil respiration data. Biogeosciences **7**:1915-1926.

Cassey, P., and T. M. Blackburn. 2006. Reproducibility and repeatability in ecology. BioScience **56**:958-959.

Easterbrook, S. M. 2014. Open code for open science? Nature Geoscience **7**:779-781.

Falster, D. S., R. A. Duursma, M. I. Ishihara, D. R. Barneche, R. G. FitzJohn, A. Vårhammar, M. Aiba, M. Ando, N. Anten, M. J. Aspinwall, J. L. Baltzer, C. Baraloto, M. Battaglia, J. J. Battles, B. Bond-Lamberty, M. van Breugel, J. Camac, Y. Claveau, L. Coll, M. Dannoura, S. Delagrange, J.-C. Domec, F. Fatemi, W. Feng, V. Gargaglione, Y. Goto, A. Hagihara, J. S. Hall, S. Hamilton, D. Harja, T. Hiura, R. Holdaway, L. S. Hutley, T. Ichie, E. J. Jokela, A. Kantola, J. W. G. Kelly, T. Kenzo, D. King, B. D. Kloeppel, T. Kohyama, A. Komiyama, J.-P. Laclau, C. H. Lusk, D. A. Maguire, G. le Maire, A. Mäkelä, L. Markesteijn, J. Marshall, K. McCulloh, I. Miyata, K. Mokany, S. Mori, R. W. Myster, M. Nagano, S. L. Naidu, Y. Nouvellon, A. P. O'Grady, K. L. O'Hara, T. Ohtsuka, N. Osada, O. O. Osunkoya, P. L. Peri, A. M. Petritan, L. Poorter, A. Portsmuth, C. Potvin, J. Ransijn, D. Reid, S. C. Ribeiro, S. D. Roberts, R. Rodríguez, A. Saldaña-Acosta, I. Santa-Regina, K. Sasa, N. G. Selaya, S. C. Sillett, F. Sterck, K. Takagi, T. Tange, H. Tanouchi, D. Tissue, T. Umehara, H. Utsugi, M. A. Vadeboncoeur, F. Valladares, P. Vanninen, J. R. Wang, E. Wenk, R. Williams, F. de Aquino Ximenes, A. Yamaba, T. Yamada, T. Yamakura, R. D. Yanai, and R. A. York. 2015. BAAD: a Biomass And Allometry Database for woody plants. Ecology **96**:1445-1445.

Hart, E., P. Barmby, D. S. LeBauer, F. Michonneau, S. Mount, T. Poisot, K. H. Woo, N. ZImmerman, and J. W. Hollister. 2015. Ten simple rules for digital data storage. PeerJ.

Holdren, J. P. 2013. Increasing Access to the Results of Federally Funded Scientific Research.

Iqbal, S., J. D. Wallach, M. J. Khoury, S. D. Schully, and J. P. A. Ioannidis. 2016. Reproducible research practices and transparency across the biomedical literature. PLoS Biology **14**:e10023333.

Kattge, J., S. Díaz, S. Lavorel, I. C. Prentice, P. Leadley, G. Bönisch, E. Garnier, M. Westoby, P. B. Reich, I. J. Wright, J. H. Cornelissen, C. Violle, S. P. Harrison, P. M. van Bodegom, M. Reichstein, B. J. Enquist, N. A. Soudzilovskaia, D. D. Ackerly, M. Anand, O. K. Atkin, M. Bahn, T. R. Baker, D. D. Baldocchi, R. Bekker, C. C. Blanco, B. Blonder, W. J. Bond, R. A. Bradstock, D. E. Bunker, F. Casanoves, J. Cavender-Bares, J. Q. Chambers, F. S. Chapin, J. Chave, D. A. Coomes, W. K. Cornwell, J. M. Craine, B. H. Dobrin, L. Duarte, W. Durka, J. J. Elser, G. Esser, M. Estiarte, W. F. Fagan, J. Fang, F. Fernández-Méndez, A. Fidelis, B. Finegan, O. Flores, H. Ford, D. C. Frank, G. T. Freschet, N. M. Fyllas, R. V. Gallagher, W. A. Green, A. G. Gutierrez, T. Hickler, S. I. Higgins, J. G. Hodgson, A. Jalili, S. Jansen, C. A. Joly, A. J. Kerkhoff, D. Kirkup, K. Kitajima, M. Kleyer, S. Klotz, J. Knops, K. Kramer, I. Kühn, H. Kurokawa, D. Laughlin, T. D. Lee, M. Leishman, F. Lens, T. Lenz, S. L. Lewis, J. Lloyd, J. Llusià, F. Louault, S. Ma, M. D. Mahecha, P. Manning, T. Massad, B. E. Medlyn, J. Messier, A. T. Moles, S. C. Müller, K. Nadrowski, S. Naeem, Ü. Niinemets, S. Nöllert, A. Nüske, R. Ogaya, J. Oleksyn, V. G. Onipchenko, Y. Onoda, J. Ordoñez, G. Overbeck, W. A. Ozinga, S. Patiño, S. Paula, J. G. Pausas, J. Peñuelas, O. L. Phillips, V. Pillar, H. Poorter, L. Poorter, P. Poschlod, A. Prinzing, R. Proulx, A. Rammig, S. Reisch, B. Reu, L. Sack, B. Salgado-Negret, J. Sardans, S. Shiodera, B. Shipley, A. Siefert, E. Sosinski, J.-F. Soussana, E. Swaine, N. G. Swenson, K. Thompson, P. E. Thornton, M. Waldram, E. Weiher, M. A. White, S. White, S. J. Wright, B. Yguel, S. Zaehle, A. E. Zanne, and C. Wirth. 2011. TRY – a global database of plant traits. Global Change Biology **17**:2905-2935.

Kueffer, C., Ü. Niinemets, R. E. Drenovsky, J. Kattge, P. Milberg, H. Poorter, P. B. Reich, C. Werner, M. Westoby, and I. J. Wright. 2011. Fame, glory and neglect in meta-analyses. Trends in Ecology and Evolution **26**:493-494.

Kuhn, T. 1962. The structure of scientific revolutions. University of Chicago Press.

Lewandowsky, S., and D. Bishop. 2016. Research integrity: Don't let transparency damage science. Nature **529**:459-461.

Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. L. Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, and T. Yarkoni. 2015. Promoting an open research culture. Science **348**:1422-1425.

R Development Core Team. 2016. R: A language and environment for statistical computing. Version 3.2.3. R Foundation for Statistical Computing, Vienna, Austria.

Reichman, O. J., M. B. Jones, and M. P. Schildhauer. 2011. Challenges and opportunities of open data in ecology. Science **331**:703-705.

Rosenthal, R. 1979. The file drawer problem and tolerance for null results. Psychological Bulletin **86**:638-641.

Rüegg, J., C. Gries, B. Bond-Lamberty, G. J. Bowen, B. S. Felzer, N. E. McIntyre, P. A. Soranno, K. L. Vanderbilt, and K. C. Weathers. 2014. Completing the data life cycle: using information management in macrosystems ecology research. Frontiers in Ecology and the Environment **12**:24-30.

Stodden, V. 2011. Trust Your Science? Open Your Data and Code. Amstat News **July**:21-22.

Thornton, P. E., R. B. Cook, B. H. Braswell, B. E. Law, W. M. Post, H. H. Shugart, B. T. Rhyne, and L. A. Hook. 2005. Archiving numerical models of biogeochemical dynamics. Eos **86**:431-432.

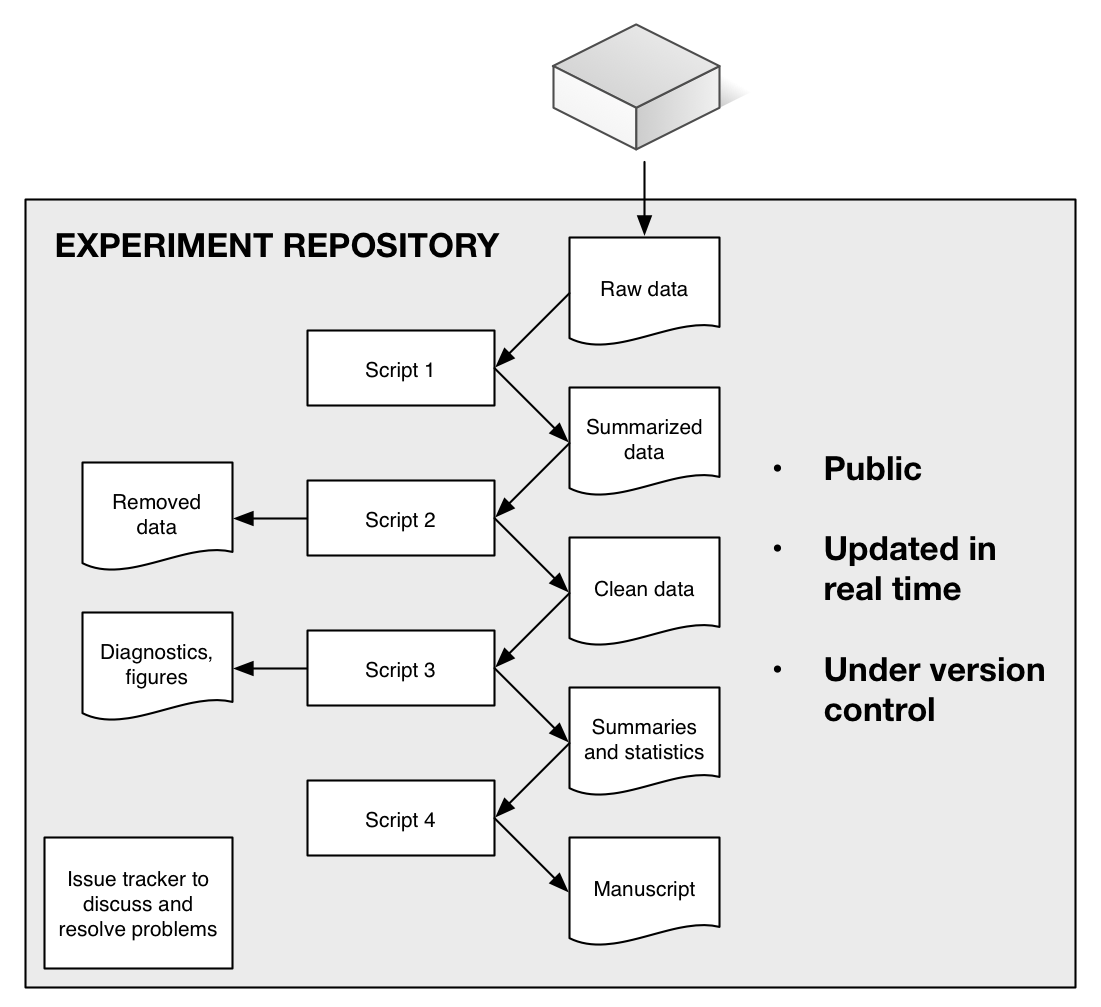
Vines, T. H., A. Y. K. Albert, R. L. Andrew, F. Débarre, D. G. Bock, M. T. Franklin, K. J. Gilbert, J.-S. Moore, S. Renaut, and D. J. Rennison. 2014. The availability of research data declines rapidly with article age. Current Biology **24**:94-97.

Wilson, G., D. A. Aruliah, C. T. Brown, N. P. Chue Hong, M. Davis, R. T. Guy, S. H. D. Haddock, K. D. Huff, I. M. Mitchell, M. D. Plumbley, B. Waugh, E. P. White, and P. Wilson. 2014. Best practices for scientific computing. PLoS Biology **12**:e1001745.

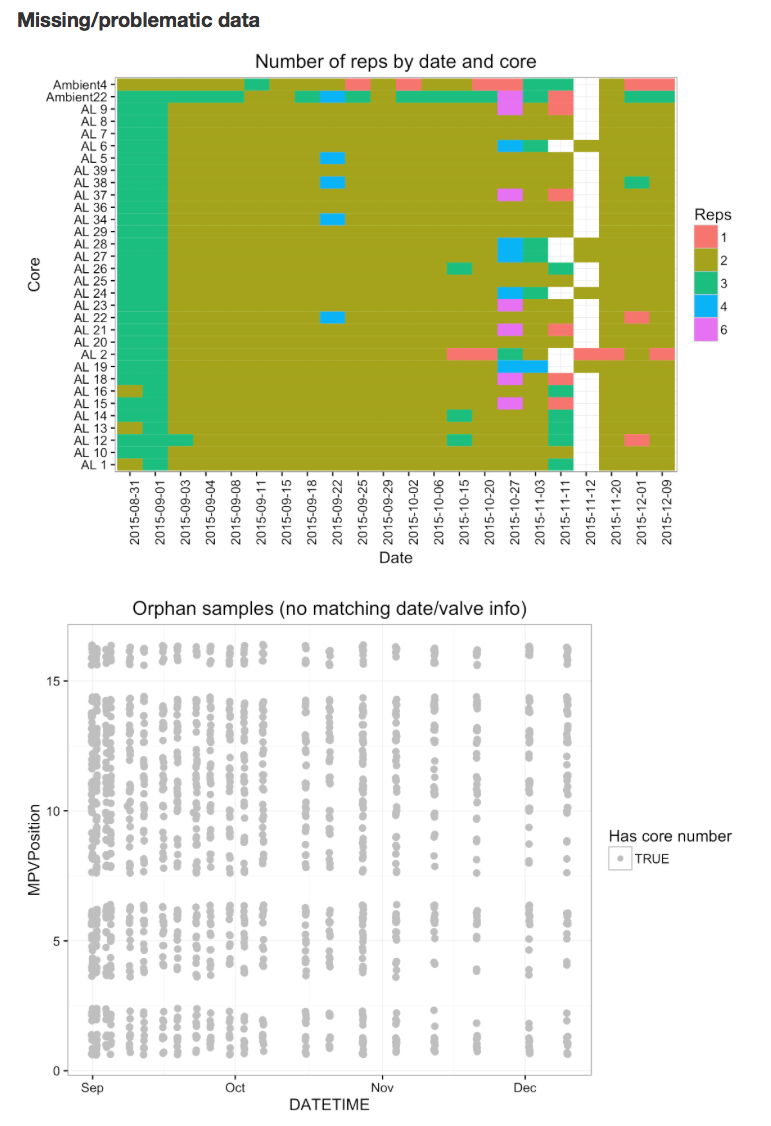
Wolkovich, E. M., B. I. Cook, J. M. Allen, T. M. Crimmins, J. L. Betancourt, S. E. Travers, S. Pau, J. Regetz, T. J. Davies, N. J. B. Kraft, T. R. Ault, K. Bolmgren, S. J. Mazer, G. J. McCabe, B. J. McGill, C. Parmesan, N. Salamin, M. D. Schwartz, and E. E. Cleland. 2012a. Warming experiments underpredict plant phenological responses to climate change. Nature **485**:494-497.

Wolkovich, E. M., J. Regetz, and M. I. O'Connor. 2012b. Advances in global change research require open science by individual researchers. Global Change Biology **18**:2102-2110.

**Figure 1.** The 'repository' containing all code and data of a recent soil incubation experiment. From top to bottom: raw data from a lab analyzer is uploaded; an initial script processes raw data into a standardized, summarized form; a second script looks for inconsistencies and outliers in the data; a third computes 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.



**Figure 2.** Two of the many diagnostic figures that were produced, in real time, as the incubation experiment proceeded. They show (top) the number of replicate measurements made on each soil core, by date of measurement, and (bottom) measurements with a data mismatch problem, indicating a data entry or other error (in the figure no errors, which would be in red, are extant). A wide range of similar diagnostic graphics (see <https://github.com/bpbond/cpcrw_incubation#current-diagnostics>) and statistics let us quickly identify problems or interesting patterns in the data.



1. <http://www.nature.com/sdata/data-policies> [↑](#footnote-ref-1)
2. <http://journals.plos.org/plosone/s/data-availability> [↑](#footnote-ref-2)
3. <http://www.sciencemag.org/authors/science-editorial-policies#data-deposition> [↑](#footnote-ref-3)
4. <http://www.nsf.gov/bfa/dias/policy/dmp.jsp> [↑](#footnote-ref-4)
5. <http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Data-management-and-sharing/index.htm> [↑](#footnote-ref-5)
6. <http://rmarkdown.rstudio.com> [↑](#footnote-ref-6)
7. <https://github.com/blog/1885-better-word-highlighting-in-diffs> [↑](#footnote-ref-7)