A minimum standard for publishing computational results in the weather and climate sciences

by Damien Irving

A procedure for publishing reproducible computational research is described that could be adopted as a minimum standard by journals in the weather and climate sciences.

The rise of computational science has led to unprecedented opportunities in the weather and climate sciences. Ever more powerful computers enable experiments that would have been considered impossible only a decade ago, while new hardware technologies allow data collection in even the most inaccessible places. To analyze the vast quantities of data now available to them, modern practitioners—most of whom are not computational experts—use an increasingly diverse set of software tools and packages. Today's weather or climate scientist is far more likely to be found debugging code written in Python, MATLAB, Interactive Data Language (IDL), NCAR Command Language (NCL), or R than to be poring over satellite images or releasing radiosondes.

This computational revolution is not unique to the weather and climate sciences and has led to something of a reproducibility crisis in published research (e.g., Peng 2011). Most papers do not make the data and code underpinning key findings available, nor do they adequately specify the software packages and libraries used to execute that code. This means it is impossible to replicate and verify most of the computational results presented in journal articles today. By extension (and perhaps even more importantly), it is also impossible for readers to interrogate the data processing methodology. If a reader cannot find out which Python library was used in regridding a particular dataset, how can they build upon that regridding method and/or apply it in their own context?

A movement within the computational science community has arisen in response to this crisis, calling for existing communication standards to be adapted to include the data and code associated with published findings (e.g., Stodden and Miguez 2014). The movement has also been active in producing best
practice recommendations to guide scientists and stakeholders (e.g., Prlić and Procter 2012; Stodden 2012; Sandve et al. 2013; Stodden and Miguez 2014), and similar calls and guidelines have appeared in numerous editorials and commentaries in recent years (e.g., Barnes 2010; Merali 2010; Ince et al. 2012). In response to this sustained campaign, there has been a modest but perceptible reaction from funding agencies and academic journals. Agencies like the National Science Foundation now require dataset disclosure and encourage software availability; however, this is not consistently enforced and compliance is largely left to the authors themselves (Stodden et al. 2013).

A recent review of journal policies found a trend toward data and code availability but, overall, the vast majority of journals have no data or code policy (Stodden et al. 2013).

Similar to many other computational disciplines, in the weather and climate sciences progress on code availability is lagging behind data availability. The societies behind the major journals (the American Meteorological Society, Royal Meteorological Society, American Geophysical Union, and European Geosciences Union) all have official data policies (e.g., Mayernik et al. 2015b); however, only two of the four indicate that code is included under their broad definition of data or metadata. Where code is included, statements regarding code availability consist of brief, vague suggestions that are not enforced by editors and reviewers. New journals such as Geoscientific Model Development have arisen for documenting work where code/software is the primary output (e.g., the development of a new climate model), but little progress has been made in documenting the computational aspects of research where code is ancillary to the main focus (i.e., where the code is not of sufficient consequence to require a standalone paper devoted to its description). Given that much of the research conducted by weather and climate scientists is based on previously documented datasets and/or models (e.g., a paper might analyze a reanalysis dataset or the output from running a well-known atmospheric model forced with anomalous sea surface temperatures), ancillary code availability (as opposed to data availability or primary code availability) is the component of the reproducibility crisis common to essentially all research today.

While it is tempting to simply decry the slow response of journals and funding agencies in the face of this crisis, the reality is that examples of reproducible weather and climate research upon which to base new communication standards have only just begun to emerge. For instance, the Max Planck Institute for Meteorology (MPI-M) recently enacted a policy (Stevens 2015b) that requires all primary data (including ancillary code) to be archived, and papers adhering to that policy are now starting to be published (e.g., Stevens 2015a). There are also a limited number of examples from other research disciplines where highly motivated computational scientists have taken a variety of different approaches to publishing reproducible results (e.g., Ketcheson and Ahmadia 2012; Crooks and Hailegiorgis 2014; Bremges et al. 2015; Schmitt et al. 2015).

To stimulate further discussion and progress in this area, this essay describes a procedure for reporting computational results that was employed in a recent Journal of Climate paper (Irving and Simmonds 2015, hereafter IS2015). Building on the aforementioned examples, the procedure was developed to be consistent with recommended computational best practices and seeks to minimize the time burden on authors. For the reasons articulated above and the fact that data availability has already been addressed in a recent BAMS essay (Mayernik et al. 2015a), the focus of the procedure is on ancillary code availability. It should provide a starting point for weather and climate scientists looking to publish reproducible research, and it is proposed that the procedure could be adopted as a minimum standard by relevant academic journals and institutions.

**A REPRODUCIBLE PAPER. Rationale.** In devising a procedure that might be adopted as a minimum communication standard, it was important to first consider why computational scientists do not currently publish their code. The literature on this topic is relatively sparse; however, a recent survey of the machine learning community found that the top reason for not sharing code was the perceived time required to prepare it for publication, followed by the prospect of dealing with questions from users (Stodden 2010). While many researchers are sympathetic to the ideals of open science and reproducible research, it appears that the practicalities seem too difficult and time consuming, particularly when the pressure to publish is so strong and unrelenting.

The bewildering array of suggested tools and best practices is one reason why an individual working in the weather and climate sciences might place reproducibility in the “too hard” basket. An appropriate solution for any given scientist no doubt exists within that collection of tools and practices, but it is heavily obscured. Consider the “regular” scientist described in the sidebar. In consulting the literature on reproducible computational research, they would
be confronted with options including data prov-

eance tracking systems like VisTrails (Freire and
Silva 2012) and PyRDM (Jacobs et al. 2014), software

environment managers like Docker and Vagrant

(Stodden and Miguez 2014), and even online services

like RunMyCode.org where your code and data can

be run by others (Stodden et al. 2012). These might

be fantastic options for small teams of software engi-

neers or experienced scientific programmers dealing

with very large workflows [e.g., the postprocessing

of thousands of Coupled Model Intercomparison

Project (CMIP) model runs], complex model simula-

tions, and/or production style code (e.g., they might

be developing a satellite retrieval algorithm that has

high reuse potential in the wider community), but a

regular scientist has neither the requisite computa-

tional experience or a research problem of sufficient

scale and complexity to necessarily require and/or

make use of such tools.

The procedure employed by IS2015 was designed

with this regular scientist in mind. It looked to mini-
mize both the time involved and the complexity of

the associated tools, while at the same time remain-
ing faithful to established best practices in scientific

computing.

Components. At first glance, the only difference be-
tween a regular journal article and that of IS2015 is
the addition of a short computation section (see “Example
computation section” for more information). That

section accompanied the traditional description of

data and methods within the article and briefly cited
the major software packages used in the research be-
fore pointing the reader to three key supplementary
items: 1) a more detailed description of the software
used, 2) a version controlled and publicly available

code repository, and 3) a collection of supplemen-
tary log files that captured the data processing steps

taken in producing each key result. The repository

was hosted at a popular code sharing website called

GitHub, while the detailed software description and

log files were hosted at Figshare (Irving 2015a), which

is a website where researchers commonly archive

the “long tail” of their research (e.g., supplementary

figures, code, and data).

Software description. There is an important differ-
ence between citing the software that was used in a
study (i.e., so that the authors get appropriate academ-
ic credit) and describing it in sufficient detail so as to
convey the precise version and computing environ-
ment (Jackson 2015). Recognizing this, IS2015 began
their computation section with a high-level descrip-
tion of the software used, which included citations
to any papers written about the software. Authors of
scientific software are increasingly publishing with
journals like the *Journal of Open Research Software*,
so it is important for users of that software to cite
those papers within their manuscripts. This high-
level description is also useful for briefly articulating
what general tasks each software item was used for (e.g., plotting, data analysis, file manipulation). Such an overview does not provide sufficient detail to recreate the computing environment used in the study, so IS2015 provided a link to a supplementary file on Figshare that documents the precise version of each software package used and the operating system upon which it was run (viz., the name, version number, release date, institution, and DOI or URL).

**Code repository.** An important best practice in scientific computing is that of modularizing code, rather than copying and pasting (Wilson et al. 2014). As such, any communication standard for code should expect/encourage authors to provide a whole library of code (i.e., a repository containing many interconnected scripts) as opposed to a single script for each key figure or result. A related best practice is the use of a version control system like Git, Subversion, or Mercurial. These systems are easily linked to an online hosting service such as GitHub or Bitbucket, which is the means by which a code repository can be made publicly available. Most of the examples of reproducible research currently available in the literature produced a pristine GitHub or Bitbucket repository for their paper (i.e., one that contains only the code that is directly relevant to that paper); however, this should not be an expectation for regular scientists who are not looking for broad-scale community uptake of their code. Not only is this a time-consuming practice that would likely involve a degree of cutting and pasting, it also goes against the workflow that version control promotes. By providing a link to their everyday repository instead, authors can quickly and easily signal to readers what code was used to produce the results in their paper. The authors may further develop that code in the future, so the reader then also has easy access to the latest version if they would like to use it in their own work and/or suggest improvements (referred to as a “pull request” in version control parlance). There are all sorts of extra bits and pieces of code in the everyday repository that is linked to by IS2015. Readers will probably never look at that code (because it is not referred to in the associated log files), but what is the harm if they do? Students and other scientists could potentially learn from it, and in the best case scenario they would inform the authors of a bug or potential improvement to the code.

In addition to providing a link to the associated Github repository, IS2015 provided (on Figshare) a static snapshot of the repository taken at the time the manuscript was accepted for publication. The motivation for doing this was to ensure that a static version of the code is available in case the associated GitHub repository is ever moved or deleted. Unlike GitHub, archiving sites such as Figshare and Zenodo issue DOIs, which function as a perpetual link to the resource and can only be obtained by agencies that commit to maintain a reliable level of preservation (Potter and Smith 2015). Since this snapshot does not provide a revision history of the code, it would not have been appropriate to provide it instead of the link to GitHub. Not all of the results presented in a paper will necessarily have been generated with the very latest version of the code, hence the need for the revision history (and even if they were generated with the latest version, it is not possible to submit a pull request to Figshare or Zenodo). Both Figshare and Zenodo provide a simple interface for importing code directly from GitHub, so the process is very straightforward.

**Example computation section**

The results in this paper were obtained using a number of different software packages. A collection of command line utilities known as the NetCDF Operators (NCO) and Climate Data Operators (CDO) were used to edit the attributes of netCDF files and to perform routine calculations on those files (e.g., the calculation of anomalies and climatologies) respectively. For more complex analysis and visualization, a Python distribution called Anaconda was used. In addition to the Numerical Python (NumPy; Van Der Walt et al. 2011) and Scientific Python (SciPy) libraries that come installed by default with Anaconda, a Python library called xray was used for reading/writing netCDF files and data analysis. Similarly, in addition to Matplotlib (the default Python plotting library; Hunter 2007), Iris and Cartopy were used to generate many of the figures.

To facilitate the reproducibility of the results presented, an accompanying Figshare repository has been created to document the computational methodology (Irving 2015a). In addition to a more detailed account (i.e., version numbers, release dates, web addresses) of the software packages discussed above, the Figshare repository contains a supplementary file for each figure in the paper, outlining the computational steps performed from initial download of the ERA-Interim data through to the final generation of the plot. A version controlled repository of the code referred to in those supplementary files can be found at https://github.com/DamienIrving/climate-analysis.
The following is the log file corresponding to Fig. SB1. It shows every command line entry executed in creating the figure from the initial download of the underlying data to the final generation of the figure. Details regarding the software and code referred to in the log file are provided in the “Example computation section.”

Tue Jun 30 08:57:37 2015: /usr/local/anaconda/bin/python /home/STUDENT/dbirving/climate-analysis/visualisation/plot_hilbert.py/mnt/meteo0/data/simmonds/dbirving/ERAInterim/data/vaERAInterim_500_hPa_030day-runmean_native.nc va /mnt/meteo0/data/simmonds/dbirving/ERAInterim/data/zw/figures/hilbert_zw_w19_vaERAInterim_500_hPa_030day-runmean_native.nc va /

Tue Jun 30 07:35:49 2015: cdo runmean,30 /mnt/meteo0/data/simmonds/dbirving/ERAInterim/data/vaERAInterim_500_hPa_030day-runmean_native.nc va /mnt/meteo0/data/simmonds/dbirving/ERAInterim/data/zw/figures/hilbert_zw_w19_vaERAInterim_500_hPa_030day-runmean_native.nc va

Wed Mar 18 17:17:14 2015: cdo mergetime vaERAInterim_500_hPa_daily-2014-06-01-to-2014-12-31_native.nc ../data/vaERAInterim_500_hPa_daily_native_origin.nc ../data/vaERAInterim_500_hPa_daily_native.nc va ERAInterim_500_hPa_daily_native.nc

Mon Nov 10 17:15:49 2014: ncat -O -va,o,c,500 hPa vaERAInterim_500 hPa_daily_native.nc

Mon Nov 10 16:31:05 2014: ncat -O -va,o,c,500 hPa vaERAInterim_500 hPa_daily_native.nc

Thu Aug 21 10:57:46 2014: cdo invertlat -sellonlatbox,0,359.9,-90,90 -daymean vaERAInterim_500 hPa_6hourly_native.nc vaERAInterim_500 hPa_daily_native.nc


Wed Aug 20 23:16:22 2014: Initial download of 6 hourly, 500 hPa meridional wind ERA-Interim data in 5 or 10 year chunks (e.g., vaERAInterim_500_hPa_6hourly-2009-2014_native.nc) from http://apps.ecmwf.int/datasets/data/interim-full-daily/

Fig. SB1. Wave envelope of the ERA-Interim, 30-day running mean, 500-hPa meridional wind for 22 May 1986 and 29 Jul 2006 at 54.75°S. See IS2015 for details.
A NEW MINIMUM STANDARD. Implications and practicalities. Before formally proposing a minimum standard for the communication of computational results, it is worth considering the implications and practicalities of adopting a standard based on the approach of IS2015. The reproducibility of published results would presumably improve, which may also lead to increased trust, interest, and citations (Piwowar et al. 2007), but what else would it mean for authors and reviewers? Would there be barriers to overcome in implementing such a standard, and what could be done to make the transition easier?
Authors. As previously mentioned, Stodden (2010) identified the time involved as the largest barrier to sharing code. There are numerous authors who suggest that the best practices adopted by IS2015 (e.g., scripting, version control) save time in the long run (e.g., Sandve et al. 2013; Wilson et al. 2014), which means that once researchers have learned and adopted these practices they may actually save time. In my experience as a Software Carpentry (Wilson 2016) instructor, many weather and climate scientists are comfortable with the idea of scripting, but very few use version control. Learning these new skills is not overly time consuming (Software Carpentry teaches them in a short two-day workshop), but on a local level it requires an individual or institution to take the lead in liaising with Software Carpentry to find volunteer instructors and to coordinate other logistics.

A good example is the Australian Meteorological and Oceanographic Society, who have hosted a Software Carpentry workshop alongside their annual conference for the past three years running. Of course, it is also possible to learn these skills by following an online tutorial (e.g., all the Software Carpentry lessons are available online), but there is an added benefit to the social aspect of a workshop. It helps to reduce the shame many scientists have about the quality of their code (i.e., they see that their peers are no “better” at coding than they are), which is an important part of achieving the required cultural shift toward an acceptance of code sharing (Barnes 2010).

The other potentially time consuming task associated with adopting a minimum standard would be dealing with requests for assistance. One suggested solution to this problem is to make it clear that authors are not obliged to support others in repeating their computations (Easterbrook 2014). This is probably the only feasible solution, but it is worth noting that even if not formally obliged, some authors may fear that refusing requests will make it look like they have something to hide.

Some researchers also face barriers relating to security and proprietary, particularly if they are using large code bases that have been developed for research and/or operations within government laboratories, national weather bureaus, and private companies (Stodden 2010). Such code bases are increasingly being made public [e.g., the Australian Bureau of Meteorology and Commonwealth Scientific and Industrial Research Organisation (CSIRO) host the code for their Climate and Weather Science Laboratory in a public GitHub repository], but any proposed minimum standard would need to allow some flexibility for researchers who are unable to make their code public for these reasons (the code archived by MPI-M is available via request only, which might be an acceptable solution in many cases). For those concerned about getting appropriate academic credit for highly novel and original code, a separate publication (e.g., with the Journal of Open Research Software) or software license (e.g., Morin et al. 2012) might also be an option.

Reviewers. If the approach of IS2015 was to be implemented as a minimum standard, it would be important to convey to reviewers that they are not expected to review the code associated with a submission; they simply have to check that it is sufficiently documented (i.e., that the code is available in an online repository and that log files have been provided for all figures and key results). Not only would it be unrealistic to have reviewers examine submitted code owing to the wide variety of software tools and programming languages out there, it would also be inconsistent with the way scientific methods have always been reviewed. For instance, in the 1980s it was common for weather and climate scientists to manually identify weather systems of interest (e.g., polar lows) from satellite imagery. The reviewers of the day were not required to go through all the satellite images and check that the author had counted correctly—they simply had to check that the criteria for polar low identification was adequately documented. This is not to say that counting errors were not made on the part of authors (as with computer code today there were surely numerous errors/bugs)—it was just not the job of the reviewer to find them. Author errors are borne out when other studies show conflicting results and/or when other authors try to replicate key results, which is a process that would be greatly enhanced by having a minimum standard for the communication of computational results. This idea of conceptualizing the peer review of code as a postpublication process is consistent with the publication system envisaged by the open evaluation movement (e.g., Kriegeskorte et al. 2012).

Proposed standards. To assist in establishing a minimum standard for the communication of computational results, it is proposed that the following text could be inserted into the author and reviewer guidelines of journals in the weather and climate sciences (institutions that have their own internal review process could also adopt these guidelines). In places, the language borrows from the guidelines recently adopted by Nature (Nature 2014). It is anticipated that a journal would provide links to
examples of well-documented computational results to help both authors and reviewers in complying with these guidelines. The journal could decide to host the supplementary materials itself (i.e., the software description, log files, and static snapshot code repository) or encourage the author to host these items at an external location that can guarantee persistent, long-term access (e.g., an institutionally supported site like MPI-M provides for its researchers or an online academic archive such as Figshare or Zenodo).

**Author guidelines.** If computer code is central to any of the paper’s major conclusions, then the following is required as a minimum standard:

1) A statement describing whether (and where) that code is available and setting out any restrictions on accessibility.

2) A high-level description of the software used to execute that code (including citations for any academic papers written to describe that software).

3) A supplementary file outlining the precise version of the software packages and operating system used. This information should be presented in the following format: name, version number, release date, institution, and DOI or URL.

4) A supplementary log file for each major result (including key figures) listing all computational steps taken from the initial download/attainment of the data to the final result (i.e., the log files describe how the code and software were used to produce the major results).

It is recommended that items 1 and 2 are included in a “computation procedures” (or similarly named) section within the manuscript itself. Any practical issues preventing code sharing will be evaluated by the editors, who reserve the right to decline a paper if important code is unavailable. While not a compulsory requirement, best practice for code sharing involves managing code with a version control system such as Git, Subversion, or Mercurial, which is then linked to a publicly accessible online repository such as GitHub or Bitbucket. In the log files a unique revision number (or hash value) can then be quoted to indicate the precise version of the code repository that was used. Authors are not expected to produce a brand new repository to accompany their paper; an “everyday” repository that also contains code not relevant to the paper is acceptable. Authors should also note that they are not obliged to support reviewers or readers in repeating their computations.

**Reviewer guidelines.** The reviewer guidelines for most journals already ask if the methodology is explained in sufficient detail so that the paper’s scientific conclusions could be tested by others. Such guidelines could simply be added to as follows: “If computer code is central to any of those conclusions, then reviewers should ensure that the authors are compliant with the minimum standards outlined in the author guidelines. It should be noted that reviewers are not obliged to assess or execute the code associated with a submission. They must simply check that it is adequately documented.”

**Discussion.** To combat the reproducibility crisis in published computational research, a simple procedure for communicating computational results has been demonstrated (see IS2015) and its rationale discussed. The procedure involves authors providing three key supplementary items: 1) a description of the software packages and operating system used, 2) a (preferably version controlled and publicly accessible) code repository, and 3) a collection of supplementary log files that capture the data processing steps taken in producing each key result. It should provide a starting point for weather and climate scientists (and perhaps computational scientists more generally) looking to publish reproducible research and could be adopted as a minimum standard by relevant academic journals.

The procedure/standard was developed to be consistent with recommended computational best practices and seeks to minimize the time burden on authors, which has been identified as the most important barrier to publishing code. In particular, best practice dictates that at a minimum weather and climate scientists should be 1) writing data analysis scripts so they can rerun their analyses, 2) using version control to manage those scripts for backup and ease of sharing/collaboration, and 3) storing the details of their analysis steps in the global history attribute of their netCDF data files (or following an equivalent process for other file formats) to ensure the complete provenance of their data. To make their published results reproducible, it follows that the minimum an author would need to do is simply make those history attributes available (via log files) along with the associated code repository and a description of the software used to execute that code. The attainment of this minimum standard would involve a slight change to the workflow of many regular weather and climate scientists (e.g., most do not use version control); however, the standard has been designed to only require skills that can be learned very quickly (e.g., at a two-day Software Carpentry workshop).
While widespread adoption of this minimum standard would be a great starting point for reproducible research, it is worth noting that as a community we should ultimately aim much higher. By way of analogy, minimum standards in the construction industry ensure that buildings will not fall over or otherwise kill their inhabitants, but if everyone only built to those minimum standards our cities would be hugely energy inefficient. The proposed minimum standard for computational research ensures that published results are reproducible (which is a big improvement on the current state of affairs), but recreating workflows from the log files and daily code repositories of even just moderately complex analyses would be a tedious and time consuming process. Once comfortable with the skills and processes required to meet the minimum standard, authors should seek to go beyond them to improve the comprehensibility of their published computational results, in the same way that builders should strive for a five-star energy rating. The precise tools and methods used in this endeavor will vary from author to author; basic analyses might only require the inclusion of informative README files that explain in plain language how to execute the code, while others might choose to prepackage their code/software for ease of installation (e.g., for inclusion in the Python Package Index), make their code available to run on an online platform like RunMyCode.org, and/or provide informative flow diagrams exported from provenance tracking systems like VisTrails. As previously mentioned, it would not be appropriate to include these many varied (and often complex) options in any minimum standards, but they represent an excellent next step for scientists who have mastered the basics and will hopefully see more uptake as the computational competency of the community improves over time.

REFERENCES


Jackson, M., 2015: How to cite and describe software. [Available online at http://software.ac.uk/so-exactly-what-software-did-you-use.]


