How to write a reproducible paper

Damien Irving
Postdoctoral Research Fellow
CSIRO Oceans and Atmosphere
Hobart, Australia

@DrClimate
https://drclimate.wordpress.com
We don’t document our software and code, so it’s impossible to replicate the computational results presented in geoscience papers today.
Wanted: A concrete example

- I set out to write a reproducible paper
  - My background
    - Software Carpentry instructor
    - CSIRO support scientist
    - Regular PyCon attendee
    - Data science blog: drclimate.wordpress.com
  - Lit review
    - Barriers to publishing code
    - Computational best practices
A Novel Approach to Diagnosing Southern Hemisphere Planetary Activity and Its Influence on Regional Climate Variability

DAMIEN IRVING AND IAN SIMMONDS
School of Earth Sciences, University of Melbourne, Parkville, Victoria, Australia

(Manuscript received 22 April 2015, in final form 30 June 2015)

ABSTRACT

Southern Hemisphere mid- to upper-tropospheric planetary wave activity is characterized by the superposition of two nearly circular, quasi-stationary waveforms: zonal wavenumber 1 (ZW1) and zonal wavenumber 3 (ZW3). Previous studies have tended to consider these waveforms in isolation and with the exception studies relating to sea ice, little is known about their impact on regional climate variability. A novel approach to quantifying the combined influence of ZW1 and ZW3, using the strength of the hemispheric teleconnection as a proxy for zonal wave activity. The methodology allows the wave envelope construct to incorporate the influence of synoptic-scale Rossby wave packets and improves on existing approaches by allowing analyses in both wave phase and amplitude. While ZW1 and ZW3 are both prominent features of the climate circulation, the defining feature of highly meridional hemispheric states is an enhancement of the ZW3. Composites of the mean surface conditions during these highly meridional, ZW3-like anomalous six-months of strong planetary wave activity reveal large sea ice anomalies over the Amundsen and Bellingshausen Seas during autumn and along much of the East Antarctic coastline throughout the year. Large precipitation anomalies in regions of significant topography (e.g., New Zealand, Patagonia, and coastal Antarctica) also exhibit anomalous warm temperatures over much of the Antarctic continent were also associated with strong wave activity. The latter has potentially important implications for the interpretation of recent warming of Antarctica and the Antarctic Peninsula.

1. Introduction

The relationship between mid- to upper-tropospheric planetary wave activity and regional climate variability in the Northern Hemisphere (NH) has received a great deal of attention in recent times, as researchers try to better understand the links between the Arctic amplification and midlatitude weather (e.g., Cohen et al. 2014; Screen and Simmonds 2014). While the meridional temperature gradient has not undergone such dramatic changes in the Southern Hemisphere (SH), this flurry of research activity has highlighted the deficits in our understanding of SH planetary wave activity and its link to surface conditions. In both hemispheres, large-scale topography and continent–ocean heating contrasts provide strong forcing for longitudinally asymmetric planetary motions. Such motions, usually referred to as planetary waves, are especially strong and tend to be vertical (Holton and Hakkinen 1980). The wave amplitude increases with height, and it is important to note their influence on the development of transient weather disturbances. For example, in the tropics, they are responsible for the transport of heat and moisture, which in turn influences surface conditions. This is especially true for the SH, where the Walker circulation is a prominent feature of the atmospheric circulation.

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 impossible only a decade ago, while new software tools and libraries allow data collection in even the most inaccessible places. To analyze the vast quantities of data now available to us, modern practitioners—most of whom are not computational experts—must use increasingly diverse sets 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 underlying 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., Stoddern and Miguel 2014). The movement has also been active in producing best
These issues are less critical in our study because the primary focus is seasonal and interannual variability (as opposed to long-term trends or low-frequency variability), but they are still important to keep in mind.

**b. ERA-Interim data**

Reanalysis projects typically provide both analysis and forecast fields for download. The analysis fields are the output of the data assimilation cycle at each time interval, which for ERA-Interim is every 6 h. They represent arguably the most accurate possible depiction of the atmospheric state for several dozen variables that are all coherent on the calculation grid. These analysis fields are then used to initialize weather forecasts for the incoming hours/days. ERA-Interim forecasts are initialized twice daily at 0000 and 1200 UTC, and forecast fields are available for 3, 6, 9 and 12 h after initialization.

In this study we analyze the daily average 500-hPa meridional wind (u), 500-hPa geopotential height (Z), surface air temperature, sea ice fraction, sea surface temperature, and mean sea level pressure, calculated as the mean of the 6-hourly analysis fields from ERA-Interim. For precipitation, the “total precipitation” forecast fields were used (i.e., the sum of the convective and large-scale precipitation, which is also provided separately). Each forecast field represents the accumulated precipitation since initialization, so the daily rainfall total was calculated as the sum of the two 12-h post initialization accumulation fields for each day. The horizontal resolution of the ERA-Interim data used here is 0.75° latitude × 0.75° longitude.

**3. Computation procedures**

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 xarray was used for reading and 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 2015). In addition to a more detailed account (i.e., version numbers, release dates, and 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.)

**4. Methods for quantifying planetary wave activity**

**a. Overview of existing methods**

In analyzing SH planetary wave activity (i.e., the ZW1 and/or ZW3 patterns), previous studies have tended to define metrics based on either a stationary pattern or Fourier decomposition. With respect to the former, Raphael (2004) defines a ZW3 index that is essentially the average 500-hPa geopotential height zonal anomaly across three key points (the annual average location of the ridges of the ZW3 pattern in the 500-hPa geopotential height field), while Yuan and Li (2008) use the principal component of the leading empirical orthogonal function (EOF) mode of the surface mid-latitude meridional wind. The stationary nature of these approaches means they cannot fully capture the subtle (approximately 15° of longitude on average) seasonal migration in the phase of the ZW3 (van Loon and Rogers 1984; Mo and White 1985) or the occurrence of patterns whose phase does not approximately coincide with the location of the three analysis points or leading EOF mode.

A number of studies have analyzed the zonal waves by using a Fourier transform to express the upper tropospheric geopotential height in the frequency domain as opposed to the spatial domain (Hobbs and Raphael 2007, 2010; Turner et al. 2013). The output of a Fourier transform can be expressed in terms of a magnitude and phase for each wavenumber (or frequency/harmonic; the terminology differs in the literature), so these studies simply analyzed the magnitude and phase information corresponding to the ZW1 and/or ZW3 pattern. While this might be considered an improvement on a grid point or EOF method in the sense that the phase is allowed to vary, a shortcoming is that the result is a constant amplitude wave over the entire longitudinal domain. The two major anticyclones associated with the ZW3 pattern (located over the western and eastern South Pacific, respectively) are known to be positively covariant with respect to their location (indicating a coordinated wave pattern) but not amplitude (Hobbs and Raphael 2010),
Computation section

- Brief overview of software packages
- Link to collection of supplementary materials
  - Code, software, processing steps
  - Host with Figshare / Zenodo

http://dx.doi.org/10.6084/m9.figshare.1385387
Code

- **Minimum standard**
  - Snapshot of code library

- **Other options**
  - Version control that library
  - Link to an external hosting service
    - Allows readers to view updates & submit pull requests

Simplest thing any author can do to ensure their results are reproducible

Alternative and/or additional things that (some) authors could do to (potentially) make the reader’s life easier
Software

- **Minimum standard**
  - Manual listing: name, version number, release date, institution and DOI or URL

<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Software packages:</strong></td>
</tr>
<tr>
<td>- Climate Data Operators. 1.5.3. October 2011. Max Plank Institut fur Meteorologie. Hamburg, Germany. <a href="https://cdo.mpimet.mpg.de">https://cdo.mpimet.mpg.de</a></td>
</tr>
</tbody>
</table>

- **Other options**
  - Conda – “conda env create dirving/jclimate”
  - Docker / Nix
Processing steps

- Minimum standard:
  - Step-by-step account, raw data to result
  - Options depend on what software you’re using
    - Command line log file for each figure / result
    - Informative README file/s
Processing steps

- Other options
  - Makefile
  - Workflow management system

VisTrails provides a graphical interface to drag and drop the various components of your workflow.
Summary

<table>
<thead>
<tr>
<th>Minimum standard</th>
<th>Other options</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Code</strong></td>
<td>Version control*</td>
</tr>
<tr>
<td>_snapshot of code library*</td>
<td>External hosting service*</td>
</tr>
<tr>
<td><strong>Software</strong></td>
<td>Conda*</td>
</tr>
<tr>
<td>name, version number, release date, institution and DOI or URL of each software package*</td>
<td>Docker / Nix</td>
</tr>
<tr>
<td><strong>Processing steps</strong></td>
<td>Makefile*</td>
</tr>
<tr>
<td>command line log file*</td>
<td>Workflow system</td>
</tr>
</tbody>
</table>

*Skills can be picked up at a 2-day Software Carpentry workshop.*
Questions?

https://drclimate.wordpress.com/2016/06/16/how-to-write-a-reproducible-paper/
