Open Software - Restricted Data: A case study.

Ivan C. Hanigan * [1] David Fisher [2] Steven McEachern [3]

[1] *National Centre for Epidemiology and Population Health (ANU)*

[2] *Information Technology Services (ANU)*

[3] *Australian Data Archives (ANU)*

[*] *corresponding author: ivan.hanigan@anu.edu.au*

Unpublished working paper\ Draft: January 24, 2016

Background: This unpublished working paper was written to accompany the material presented as a speedtalk and poster at the National Climate Change Adaptation Research Facility Conference 'Climate Adaptation knowledge and partnership', June 2013, Sydney. The poster and slideshow are both available to download from this website: http://swish-climate-impact-assessment.github.io/opensoftware-restricteddata/presentations-nccarf-2013/. The latest version of this working paper is available to download from the website: http://swish-climate-impact-assessment.github.io/opensoftware-restricteddata/report1_high_level/manuscript.pdf.

Methods: We report on a project to build tools and procedures for enhancing open and transparent analysis of restricted datasets. Some datasets such as suicide or climate change scenarios need to be accessed in a restricted way. On the other hand scientists need to make their methods, models and assumptions transparent and available for scientific debate even though the datasets may require authorisation to access.

Results: We built a secure server/client computational environment for using open software with restricted data. We demonstrate the use of this system using drought and suicide as a case study. We describe the potential use of this system in modelling climate change scenarios.

Conclusions: The project shows that restricted data and open software can be used in an appropriate way to further the progress of scientific enquiry.

---

1 Background

1.1 Open software for restricted data

Some datasets such as sensitive personal information about suicide or climate change scenarios with protected intellectual property need to be accessed in a restricted way. In the context of reproducible research methods, models and assumptions need to be made transparent and available for scientific debate even though the datasets may require authorisation to access (Peng 2011).

Restrictions around access to data have increased recently in Australia, especially to the national mortality database after the discovery of an incident in which Australian population health researcher Dr Stephen Begg was reported to have hacked into information about deaths in Australia (O'Keefe 2007). The subsequent investigations led to a wide ranging modification to the procedures for approval and provision of these data that made the access much more restricted. No new research applications were approved to access these data between 2009 and 2013.

At the same time the reproducibility crisis (Peng 2011) has emerged, reducing public confidence in statements by scientists. The true extent of the problem may turn out to be overstated (Jager & Leek 2014) however the concern should be addressed to avoid the problems that a lack of confidence in scientific publications would entail, especially in respect to evidence-based policy and expenditure of public money. Appropriate access to data and analytic software addresses this issue. We investigated available workflow tools for data management and analysis and implemented a range of these products on our server. This server has enhanced our capacity for experimentation, reviews, revisions and extensions of work in this field. We present the results of this project and report that it has streamlined access to population health and environmental data for analysis.

1.2 Motivating case study: Drought and suicide

The impact of drought on mental health is plausible, however remains a gap in epidemiological knowledge (Stanke *et al.* 2013). There is concern too that this health risk will rise under future climate change (Berry *et al.* 2008). As mentioned the number of studies that have examined the relationship between suicide and drought is limited to only a handful (Page *et al.* 2002; Nicholls *et al.* 2006; Guiney 2012; Hanigan *et al.* 2012). The motivating case for this project was to use the historical exposure-response functions to estimate future climate change impacts.

There has been substantial public interest within Australia in recent decades of the putative relationship between drought and rural mental health, including suicide. The topic has frequently been raised by the media, by rural politicians and by mental health support groups (Australian Broadcasting Commission

News 2006). There have also been media reports in India indicating substantial concerns about drought and rural suicide in that country (Sarathi Biswas 2012).

The number of studies that have examined the relationship between suicide and drought is limited. However, many papers explore links between suicide and climate variables other than drought (such as temperature) and there are two major reviews papers available of the literature on climatic influences on suicide (Deisenhammer 2003; Dixon & Kalkstein 2009). However, very few studies have investigated drought specifically.

There are several mechanisms through which unusually low rainfall, especially if exacerbated by increased soil dryness due to higher temperatures may increase the suicide rate. First, droughts increase the financial stress on farmers and farming communities (even if partially compensated by drought relief welfare payments). Such difficulty may occur in conjunction with other economic stresses, such as rising interest rates, falling commodity prices, or an unfavourable foreign exchange rate. Vins et al. (2015) provide a systematic literature review of the mental health effects of drought and explore the putative causal mechanisms.

2 Methodology

The approach we took to meet the challenge of analysing restricted suicide and climate change scenario data in a safe environment was to build a new hardware and software stack using open-source software. We based our planning on the realisation that there is a growing need of these technologies in the context of reproducible research. This requires that methods, models and assumptions need to be made transparent and available for scientific debate even though the datasets may require authorisation to access. This is true not just in health data, but also including the context of data with restrictive intellectual property and licence requirements (such as climate change scenario models).

To develop an over-all view of the issue and analyse the dimensions of the problem we spent the initial phase of the project conceptualising a rich picture of the issue, and focused on identifying risks that the project might face. Several papers that describe similar systems were reviewed (Evans & Sabel 2012; Fleming *et al.* 2014) and several recommendations from these papers were adopted in our system.

Our design responds directly to the primary threat of unintentional release of sensitive data so we decided to build a secure server/client environment for analysts to develop their software in an open way, while ensuring the safety of the datasets. Other risks we identified were in relation to the provision of the server hardware and we were able to take advantage of the Australian 'Nectar Research Cloud' (http://nectar.org.au) for virtual machines to build the servers on.

Then we defined the scope and quality of the project outcomes that we were aiming to deliver. The fact that restrictions around access to data have increased recently, coupled with arguments that appropriate access to analytic software is needed to address the reproducibility crisis meant that the scope of this project was very broad. We also explored the ambitions of our stakeholders to support their publishable outputs with open software. Given that examples of un-reproducible work has spread even to the results published in top journals (Peng 2015), the scope we decided to set for this project was for very high levels of open-ness for the evidence being presented for peer-review, along with very high levels of restriction on access to the data. Luckily however we were able to rule out the need for the extreme level of restriction such as getting Australian Defence Science and Technology Organisation (DSTO) accreditation for the security of these servers against malicious hacking attacks. Our servers just needed to be tested by the standard 'vulnerability exploits' scanner used by the Australian National University IT Department.

We also looked at the workflow system Kepler to assess it's utility for providing access to the data, but found that there were a lot of limitations at the time (Curcin & Ghanem 2008) and decided that the R environment for statistical computing and graphics would be the platform we would focus on.

During the next phase of the project we dealt with issues of the costs associated with developing the software and hosting the hardware at different locations, as well as the time needed to test and get user acceptance on the services.

2.1 System design

In this case study we utilise Virtual Machines (VMs) in the cloud. Our system requires two VMs so that the storage and processing of data can be compartmentalised, with various benefits. A high level overview of the system is shown in Figure 1. Full details including Linux commands and configuration specifications are available online at http://swish-climate-impact-assessment.github.io/opensoftware-restricteddata/.

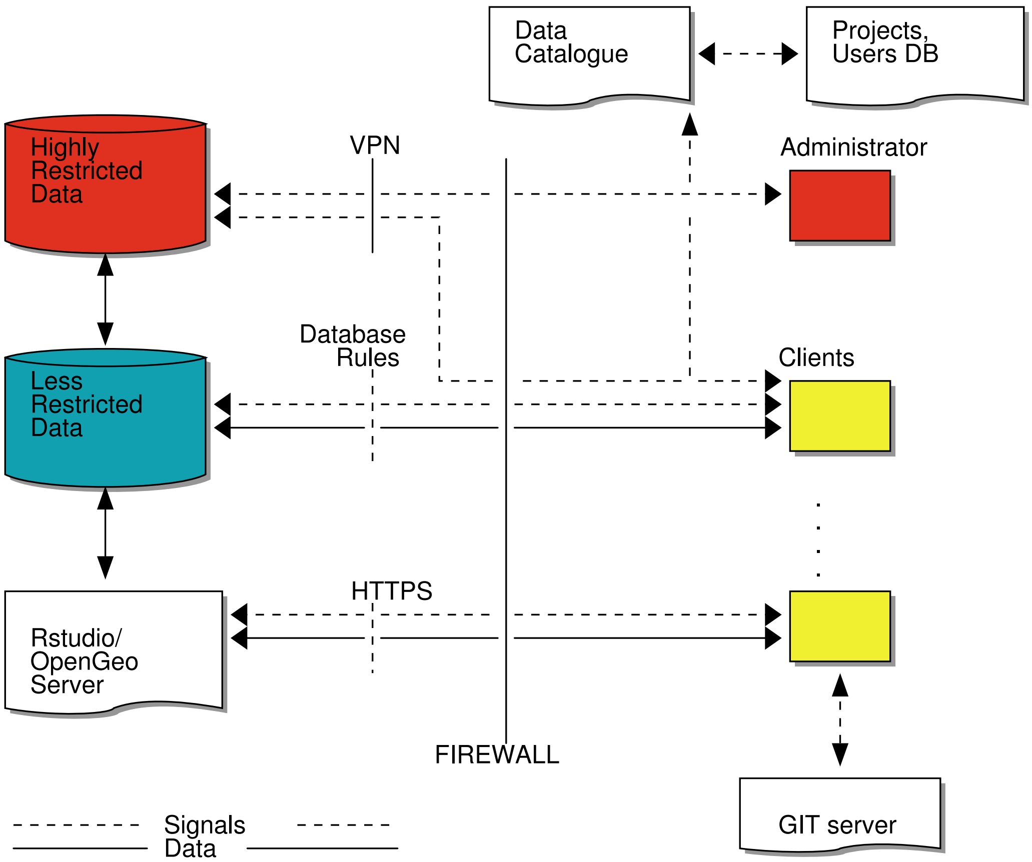

Figure 1: High Level Schematic System Design, colours indicate restrictions (red), open (blue)

2.2 Software selections

We researched a variety of systems and found the following set-up worked best for us.

Linux cluster:

- National Research Cloud www.nectar.org.au/research-cloud
- Centos 6.4 www.centos.org

Geographical Information Systems (GIS) database server:

- PostgreSQL 9.2 www.postgresql.org
- PostGIS 2.0 http://postgis.refractions.net

Statistical analysis server:

- R language for statistical computing www.r-project.org
- Rstudio server www.rstudio.com
- OpenGeo Suite http://opengeo.org

Information management:

- Projects,UsersDB Oracle XE APEX www.oracle.com
- Data Catalogue http://assda.anu.edu.au/ddiindex.html

The client side:

- Any standard web-browser
- The Kepler Project www.kepler-project.org
- pgAdmin www.pgadmin.org
- Git Version Control and GitHub www.github.com
- Emacs code editor with the starter-kit by Kieran Healy and orgmode http://kieranhealy.org/resources/emacs-starter-kit/

3 Case study

3.1 Step 1: Historical exposure-response functions

For this case study we replicated the work we had previously conducted on our personal desktop computers within the university. A key result is shown in Figure 2. That work was published already (Hanigan *et al.* 2012), however the improved IT infrastructure offered by this project allows the analysis to be re-run from a secure web-browser interface. Such improved access allowed a much broader discussion of the data, techniques and results because the researcher was able to discuss the details of the modelling with other scientists at meetings, conferences and workshops, while actually repeating the computations in real time. This is a vast improvement over the previous option of leaving the data analysis on the secure desktop computers at the university, and merely describing the computations to colleagues at the workshops.

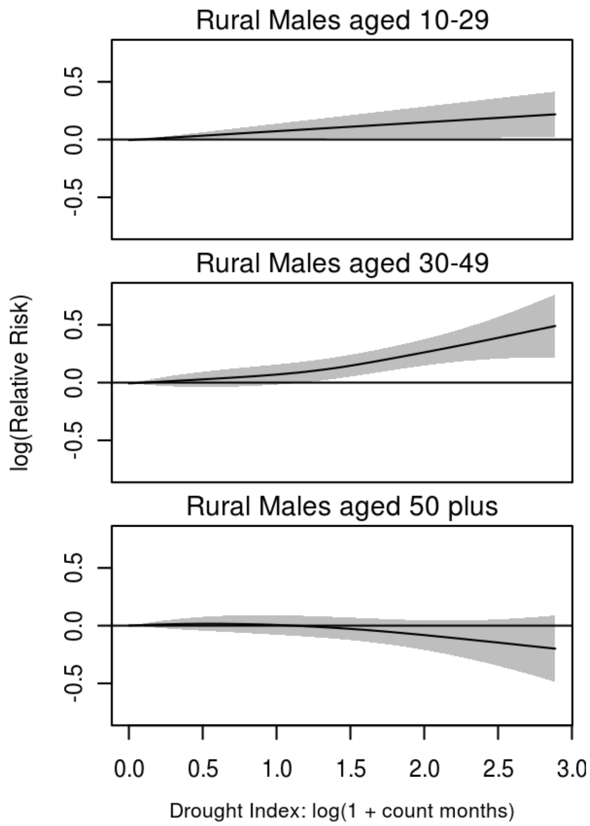

Figure 2: Drought exposure-response functions Rural Males

3.2 Step 2: Future drought scenarios and attributable fraction of suicides

Following the methods of Bambrick et al. (2008) we used the climate change scenarios provided for the Garnaut Climate Change Review (Hanigan 2013) to project estimates of future suicides under various drought conditions. The statistical method for this calculation is:

$$Y_{ijk} = \sum_{lm} (e^{(\beta_{ijk}} \times X_{lm}) - 1) \times BaselineRate_{jkl} \times Population_{jklm}$$

Where:

β_{ijk} = the ExposureVariable coefficient for zone$_i$, age$_j$ and sex$_k$

X_{lm} = Projected Future Exposure Variables

$BaselineRate_{jkl} = avgDeathsPerTime/avgPopPerTime$ in age$_j$, sex$_k$ and zone$_l$

$Population_{jklm}$ = projected populations by age$_j$, sex$_k$, zone$_l$ and time$_m$

3.3 Results

In table 1 below the two rainfall scenarios used by Berry et al. (2008) are used to demonstrate this drought impact assessment report. For full details of the data from the Garnaut Climate Change Review please refer to the original papers. For the purpose of this discussion only a brief summary is required to appreciate the relevance to this case study. The result shown below merely shows that the estimated impact of climate change can vary a lot given the input datasets and the assumptions and methods applied during analysis. The codes used to fit models, project scenarios and estimate burden of deaths are all available on the system, and can be assessed by request at the project website: http://swish-climate-impact-assessment.github.com/

Scenario	Deaths per annum	Lower 95th CI	Upper 95th percent CI
Historical (1970-2008)	4.01	2.14	6.05
A1FIR1 (Dry)	8.91	4.56	14.00
A1FIR2 (Wet)	2.93	1.50	4.47

Table 1: Estimated number of rural male suicides attributable to drought between 2000 and 2100

Below are pictorial represenations of the climate change scenarios estimated seasonal rainfall for each Statistical Division (SD) census region across the country. SDs are used because this is the geographical unit at which the suicide rates were analysed.

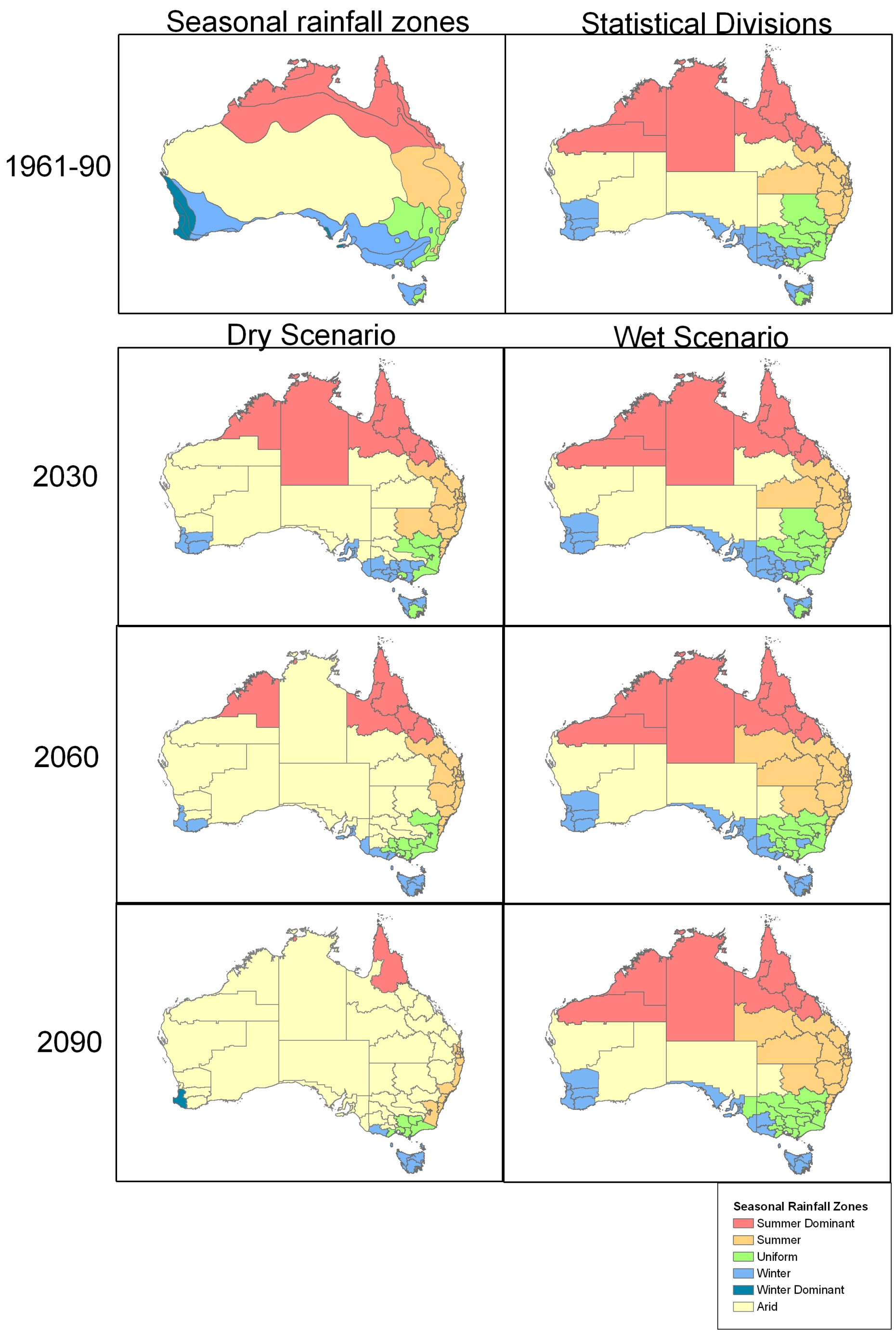

Figure 3: Seasonal rainfall under alternate future predicted scenarios (A1FI) with wet or dry assumptions

4 Discussion

4.1 Principal findings

We demonstrate the use of this system by using the climate change scenarios held on the database to project future droughts, and use the confidential suicide data that is safely stored there to estimate baseline risks and future burden of suicide attributable to the droughts. Because the server system is easy to access through a web browser these results are easily reproducible and alternate scenarios and assumptions can be tested. Because the webserver is very secure, this work can be done with confidence that the risk of unauthorised data release is minimised.

The results of our project are applicable more generally than just reproducible research. For instance in relation to issues of governance and management at the multidisciplinary institutions and multi-institutional projects it is vitally important that data sharing is enabled, while some data are still restricted because access requires approval and authorisation, or because data providers wish to be informed of re-use and identifying the profile of users of the data. It is important that data access can be made restrictive without impeding the progress of the local science agenda (collaborations, workshops, papers etc) and keeping the relevant data custodian parties informed about what is happening with the release of their datasets. The open-ness of the analytical software also has a positive effect on the value of the data infrastructure (through education and outreach activities) without risking any unethical or negligent use of these datasets.

5 Conclusions

In summary we found that the historical weather and climate change data can be attached to suicide and census data in a secure computing environment. The future drought associated deaths can then be estimated using statistical tools available in an online virtual machine. This facilitates collaboration because the data can be accessed remotely from collaborator's offices (or even workshops and conference venues) while adhering to requirements for secure data storage and access constraints.

In the case study presented it should be noted that these estimates are very uncertain, contentious and difficult to justify. However, this reinforces the argument that data management and analysis technology such as those presented are needed to enable rigorous and transparent exploration of methods and assumptions. In this way a community of like-minded scientists can challenge or verify published results.

In conclusion the system described enables data analysis in a safe environment; allows comparison of multiple climate scenarios and assumptions; is demonstrated with a climate/health impact assessment and addresses the 'reproducibility crisis'.

6 References

Australian Broadcasting Commission News. (2006). Drought lifts suicide rates: Kennett. http://www.abc.net.au/news/2006-10-13/drought-lifts-suicide-rates-kennett/1285734 [Accessed 7 Jun. 2012]

Bambrick, H.J., Dear, K.B.G., Woodruff, R.E., Hanigan, I.C. & McMichael, A.J. (2008). *Garnaut Climate Change Review of the Impacts of Climate Change on Three Health Outcomes: temperature-Related Mortality and Hospitalisations, Salmonellosis and Other Bacterial Gastroenteritis, and Population at Risk from Dengue*. Garnaut Climate Change Review, Canberra. http://www.garnautreview.org.au/CA25734E0016A131/WebObj/03-AThreehealthoutcomes/

Berry, H.L., Kelly, B.J., Hanigan, I.C., Coates, J., McMichael, A.J., Welsh, J. & Kjellström, T. (2008). *Garnaut Climate Change Review of the Rural Mental Health Impacts of Climate Change*. http://www.garnautreport.org.au/CA25734E0016A131/WebObj/03-DMentalhealth/\protect\T1\textbraceleft/\protect\T1\textdollar\protect\T1\textbracerightFile/03-DMentalhealth.pdf

Curcin, V. & Ghanem, M. (2008). Scientific workflow systems - can one size fit all? *Cairo International Biomedical Engineering Conference*, 1–9. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4786077 [Accessed 3 Nov. 2015]

Deisenhammer, E.A. (2003). Weather and suicide: The present state of knowledge on the association of meteorological factors with suicidal behaviour. *Acta Psychiatrica Scandinavica*, 108(6), 402–409.

Dixon, P. & Kalkstein, A. (2009). Climate-suicide relationships: A research problem in need of geographic methods and cross-disciplinary perspectives. *Geography Compass*, 3(6), 1–14.

Evans, B. & Sabel, C.E. (2012). Open-Source web-based Geographical Information System for health exposure assessment. *International Journal of Health Geographics*, 11(1), 2.

Fleming, L., Haines, A., Golding, B., Kessel, A., Cichowska, A., Sabel, C., Depledge, M., Sarran, C., Osborne, N., Whitmore, C., Cocksedge, N. & Bloomfield, D. (2014). Data mashups: Potential contribution to decision support on climate change and health. *International Journal of Environmental Research and Public Health*, 11(2), 1725–1746.

Guiney, R. (2012). Farming suicides during the Victorian drought: 2001-2007. *The Australian Journal of Rural Health*, 20(1), 11–15.

Hanigan, I.C. (2013). Data for the Garnaut review of climate change 2008. https://datacommons.anu.edu.au:8443/DataCommons/rest/display/anudc:5082?layout=def:display [Accessed 7 Dec. 2015]

Hanigan, I.C., Butler, C.D., Kokic, P.N. & Hutchinson, M.F. (2012). Suicide and drought in New South Wales, Australia, 1970-2007. *Proceedings of the National Academy of Sciences*, 109(35), 13950–13955.

Jager, L.R. & Leek, J.T. (2014). An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, 15(1), 1–12.

Nicholls, N., Butler, C.D. & Hanigan, I. (2006). Inter-annual rainfall variations and suicide in New South Wales, Australia, 1964-2001. *International Journal of Biometeorology*, 50(3), 139–143.

O'Keefe, B. (2007). Hackers pick up UQ cash prize. *The Australian*. http://www.theaustralian.com.au/higher-education/hackers-pick-up-uq-cash-prize/story-e6frgcjx-1111113191659 [Accessed 14 Oct. 2015]

Page, A., Morrell, S. & Taylor, R. (2002). Suicide and political regime in New South Wales and Australia during the 20th century. *Journal of Epidemiology and Community Health*, 56(10), 766–772.

Peng, R.D. (2011). Reproducible research in computational science. *Science*, 334(6060), 1226–1227.

Peng, R.D. (2015). The reproducibility crisis in science: A statistical counterattack. *Significance*, 12(3), 30–32.

Sarathi Biswas, P. (2012). Alcohol, drought lead to farmer's suicide. *Daily News and Analysis*. http://www.dnaindia.com/pune/report_alcohol-drought-lead-to-farmers-suicide_1688976 [Accessed 17 May 2012]

Stanke, C., Kerac, M., Prudhomme, C., Medlock, J. & Murray, V. (2013). Health effects of drought: A systematic review of the evidence. *PLoS Currents Disasters*, 1(1), 1–38.

Vins, H., Bell, J., Saha, S. & Hess, J. (2015). The mental health outcomes of drought: A systematic review and causal process diagram. *International Journal of Environmental Research and Public Health*, 12(10), 13251–13275.