Social evidence of a changing climate: Google Ngram data points to early climate change impact on human society

Will J. Grant[1] **and Erin Walsh**[2]

[1]*Australian National Centre for the Public Awareness of Science, Australian National University, Canberra, Australia*

[2]*Research School of Psychology, Australian National University, Canberra, Australia*

Introduction

We have long been able to point to a picture of the Earth's changing climate via a variety of physical forms of evidence, from historical instrumental data (e.g. Hansen *et al.,* 2010) to proxy indicators in glacier (e.g. Dyurgerov and Meier, 1999) and tree ring measurements (e.g. Briffa *et al.,* 2001). But can we buttress this physical data with social evidence? This article reports significant – and potentially predictive – correlations between historical discussion of a number of the predicted effects of climate change and changes in the global average temperature, prior to widespread recognition of climate change itself.

An *n*-gram approach to cultural analysis

As part of their project to index the world's stock of knowledge, Google have (as of 2011) digitised over 15 million books – around 12% of all books published. Drawing on this collection, Michel *et al.* (2011) constructed a corpus of 5195 769 digitised books (~4% of all books published), indexing 1, 2, 3, 4 and 5 word phrases (or '*n*-grams') in the texts in order to analyse our cultural record in quantitative ways. This so called 'culturomics' allows the quantitative analysis of trends in the concepts we discuss – slavery, genetic engineering, climate change – and the words we use to describe them.

Though trends in linguistic preference may be difficult to explain, it is plainly apparent that cultural salience trends (as seen in the discussion of slavery, war or earthquakes) correlate with events of the social and physical worlds. One can see, for example, major peaks in use of the word 'earthquake' in the years immediately following the 1906 San Francisco and 1931 Fuyun earthquakes (see Figure 1). Similar associations between Ngram mentions of various weather-related terms and the physical world have been demonstrated by Nicholls (2012), particularly the word 'fog', which peaked in the 1940s with the development of radar.

Yet climate change presents a far more abstract concept than the physical immediacy of an earthquake. While we can clearly detect trends in the discussion of climate change in mainstream discourse (see Discussion, below), it is worth asking two important questions of the *n*-gram data: did climate change affect society before widespread dissemination of the scientific picture, and, if so, can such proximate social evidence buttress the existing physical data? We suggest that such a cultural response to the effects of climate change, if it is detectable prior to widespread knowledge of the phenomenon itself, represents a useful proximate form of evidence of the effects of climate change and a reinforcement of the known physical evidence.

Method

We sought comparison between *n*-gram cultural data and traditional climate data. Global average temperature anomalies (from the 1951 to 1980 estimated global mean) were drawn from the NASA Goddard Institute for Space Studies' Global Land-Ocean Temperature Index (http://data.giss.nasa.gov/gistemp/tabledata_v3/GLB.Ts+dSST.txt). A comparator proximate series combining 173 temperature sensitive physical indicators was drawn from Anderson *et al.* (2013).

Cultural data were drawn from the Google Books Ngram Viewer website (http://books.google.com/ngrams/). Occurrence frequencies were downloaded for 22 one and two word *n*-grams published in the English corpus between 1880 and 2008, the period of overlap between Ngram Viewer data (1500–2008) and NASA Goddard's global temperature anomaly data (1880–2012). Data were downloaded with a smoothing of zero. As the *n*-gram data are case sensitive, simple upper and lower case forms ('Heat Wave', 'Heat wave' and 'heat wave') were downloaded and summed. Selected *n*-grams included those we believed might associate with a changing climate ('heat wave', 'unusual weather', 'unseasonal', 'drought', 'hurricane', 'cyclone', 'flood', 'flooding', 'weather', 'climate change', 'storm'), and others that clearly should not ('earthquake', 'tsunami', 'tidal wave', 'car', 'carpentry', 'computer', 'cow', 'dog', 'London', 'potato', 'tree').

The relationship between the known temperature record and Anderson *et al.*'s Paleo

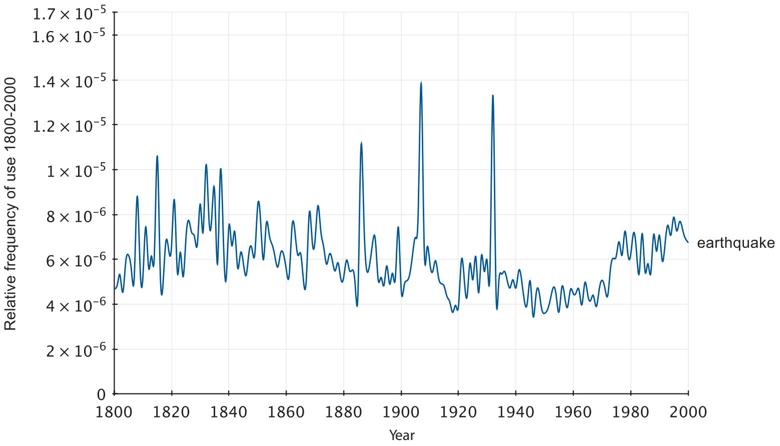

Figure 1. Annual relative frequency of use of the n-gram 'earthquake' in English language books 1800–2000.

Index (2013) was used as a comparison point for assessing correlations between word use (*n*-gram data) and temperature.

Results

Though the underlying processes of climate change have been known within the climate science community for many decades (Hulme, 2009), it would be fair to say that the scientific picture did not garner mainstream attention until the middle of the 1980s. Many have pointed to the 1988 testimony of James Hansen before a congressional hearing as being transformative (Pielke, 2010), but we can also highlight the transformation in public discourse about climate change directly in the *n*-gram data – see mentions of 'climate change' in Figures 2–5. Here one can see a significant (and probably unsurprising) lag between changes in the global average temperature and discussion of those changes. Yet a very different picture emerges when we look at discussion of some of the key weather events associated with climate change.

However, before turning to the correlations detected, it is important to address autocorrelation. Also known as perseverance, autocorrelation occurs in time series data when previous values impact upon current values (i.e. the temperature this year may be somewhat dependent on the temperature last year). If present in data, it can undermine the validity of using simple correlations and regression, and more sophisticated modelling techniques which can account for autocorrelation (such as Generalised Least Squares Regression, GLS) should be used.

For all variables, significant Durbin–Watson tests indicated the presence of autocorrelation ($p < 0.01$). This was confirmed; GLS models specifying first order autocorrelation universally fit better than those without autocorrelation specified ($-2LL\chi^2$, $p < 0.01$). Hence, autocorrelation was corrected for in analyses. In these models, incidence of the terms unseasonal, hurricane, cyclone, flood, weather, storm, car, carpentry, cow, dog and London were not significantly associated with the temperature anomaly that year. All other terms were significantly, positively associated with temperature. The pattern of results revealed within the GLS models was congruent with the conceptually simpler and more intuitively interpreted correlations, which are reported in Table 1 alongside GLS measures of model fit and significance in the interests of readability.

As Table 1 shows, very strong correlations can be seen between the mention of a number of the effects of climate change (heat waves, flooding, drought and unusual weather) and the global average temperature. The strongest correlations can be seen for 'unusual weather' (0.95, $p = 0.01$) and 'heat wave' (0.91, $p < 0.01$).

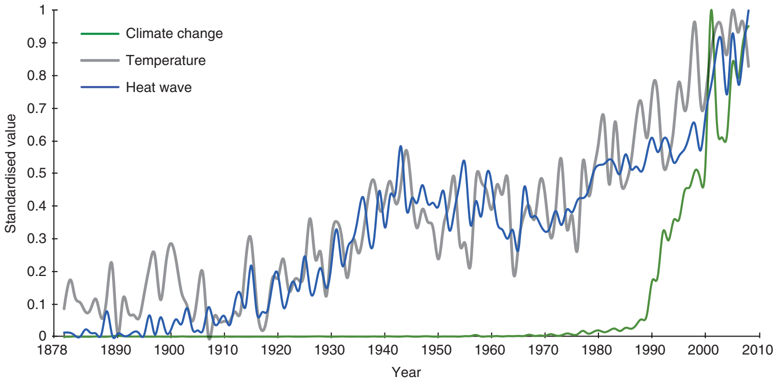

Figure 2. Variation in annual relative frequency of use of the n-grams 'heat wave' and 'climate change' in English language books, Temperature anomaly. Data scaled to show minima and maxima for each variable 1880–2008.

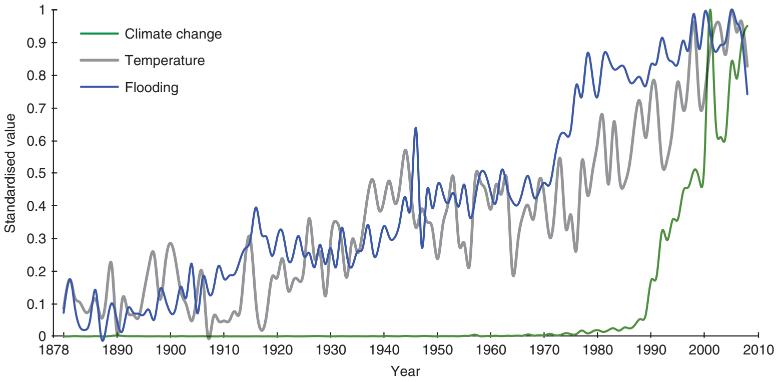

Figure 3. Variation in annual relative frequency of use of the n-grams 'flooding' and 'climate change' in English language books, Temperature anomaly. Data scaled to show minima and maxima for each variable 1880–2008.

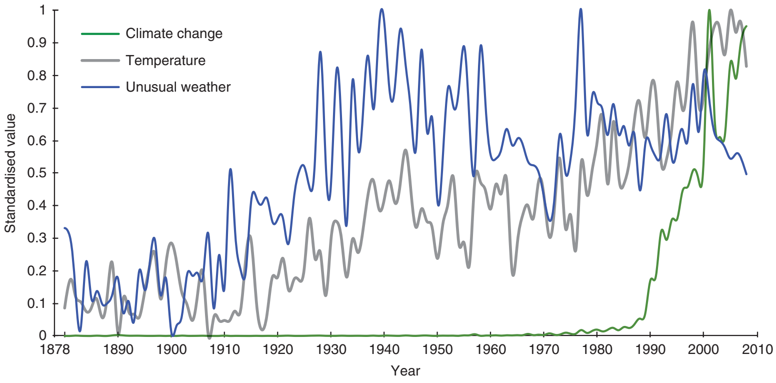

Figure 4. Variation in annual relative frequency of use of the n-grams 'unusual weather' and 'climate change' in English language books, Temperature anomaly. Data scaled to show minima and maxima for each variable 1880–2008.

These strong correlations between the temperature anomaly and discussion of climate change related terms appear dramatically in graphical form, revealing a marked difference from the long lag in the climate change mentions in Figure 2 ('heat wave'), Figure 3 ('flooding') and Figure 4 ('unusual weather'). Anderson *et al.*'s Paleo Index is included for comparison in Figure 5.

Some unexpected correlations also emerged between the temperature anomaly and *n*-grams – in particular 'computer' (0.80, $p < 0.01$) and 'tsunami/tidal wave' (0.70, $p < 0.01$). Though physical connections are not suggested, the association between usage of the word 'computer' and the changing temperature may be an artefact of modern society; the association between use of the terms tsunami/tidal wave and the temperature may be a product of the growing impact of the earth and

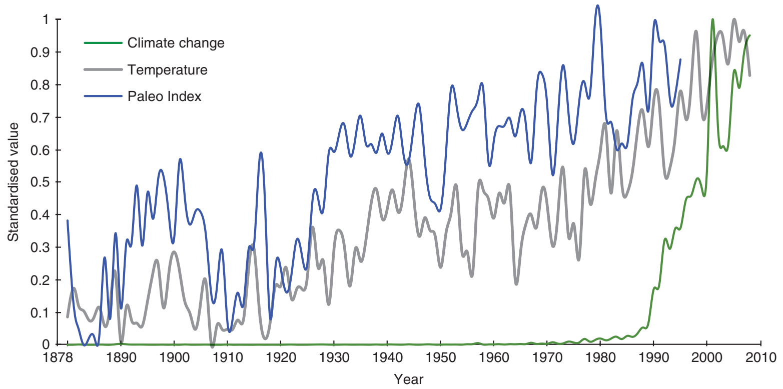

Figure 5. Variation in annual relative frequency of Anderson et al.'s (2013) Paleo Index, use of the n-gram 'climate change' in English language books, Temperature anomaly. Data scaled to show minima and maxima for each variable 1880–2008.

Table 1

Correlation and GLS model results.

	r	*GLS p*	*GLS AIC*
Paleo Index	0.80	<0.01	115
***n*-grams**			
Unusual Weather	0.95☼	0.01*	105
Heat wave	0.91☼	<0.01*	104
Flooding	0.87☼	0.05*	117
Computer	0.80	<0.01*	117
Drought	0.75	<0.01*	109
Climate change	0.74	0.02*	116
Tsunami/tidal wave	0.70	<0.01*	99
Earthquake	0.28	0.05*	122
Unseasonal	0.73	0.47	110
Hurricane	0.62	0.27	119
Car	0.79	0.75	125
Storm	–0.70	0.43	122
Weather	–0.65	0.63	123
Cow	–0.62	0.25	121
London	0.41	0.46	125
Tree	0.40	0.21	122
Flood	–0.38	0.29	121
Carpentry	0.29	0.64	117
Potato	0.20	0.52	121
Cyclone	–0.04	0.26	119
Dog	0.04	0.79	123

Note. Slopes in GLS model are not reported, as they are no longer clearly interpretable due to data centring. GLS models adjusted for autocorrelation. *indicates significance at $\alpha = 0.05$. ☼ indicates a comparatively closer relationship between *n*-gram and temperature than between Paleo Index and temperature.

ocean sciences in this period. Regardless, further investigation of all these associations – geographically and into the future as Google releases more *n*-gram data – may assist in exploring the connection.

Discussion

A correlation does not, of course, imply causation. It is impossible to rule out linguistic change or some other causative factor in driving the changes in the discussion of heat waves, flooding and unusual weather the *n*-gram data point to. Yet these strong correlations – stronger than seen in Anderson *et al.*'s (2013) Paleo Index – are indicative of a growing impact of climate change on society. These findings extend Nicholls' (2012) discussion of the use of Google *n*-gram data to examine social discussion of weather, revealing quantitative associations between social word use and the changing temperature previously unreported.

Though a rise in discussion of heat waves might be expected to associate with a heat related variable, it should be stressed that the correlation with the rise in discussion of a number of the predicted effects of climate change, such as drought and flooding, suggests a cumulative impact of climate change on society.

What is interesting is that this social evidence is unconnected with the traditional physical forms of data we have used to point to climate change. Though it might be tiresome to highlight this, such physical evidence has come under sustained political attack. As such, buttressing the physical record with social evidence presents a useful political counterargument. After all, why else would we as a society collectively decide to discuss heat waves, flooding and drought just that little bit more?

Acknowledgments

Thanks to Janie Busby Grant, Rod Lamberts, Luke Menzies, Will Steffen, Linden Ashcroft and Neville Nicholls for reading early drafts of this article.

References

Anderson DM, Mauk EM, Wahl ER *et al.* 2013. Global warming in an independent record of the past 130 years. *Geophys. Res. Lett.* **40**: 189–193. doi:10.1029/2012GL054271.

Briffa KR, Osborn TJ, Schweingruber FH *et al.* 2001. Low-frequency temperature variations from a northern tree ring density network. *J. Geophys. Res.* **106**: 2929–2941.

Dyurgerov MB, Meier MF. 1999. Twentieth century climate change: evidence from small glaciers. *Proc. Natl. Acad. Sci. USA* **97**: 1406–1411.

Hansen JR, Ruedy R, Sato M *et al.* 2010. Global surface temperature change. *Rev. Geophys.* **48**: RG4004. doi:10.1029/2010RG000345.

Hulme M. 2009. *Why We Disagree about Climate Change*. Cambridge University Press: Cambridge, UK.

Michel JB, Shen YK, Aiden AP *et al.* 2011. Quantitative analysis of culture using millions of digitized books. *Science* **331**: 176. doi:10.1126/science.1199644.

Nicholls N. 2012. Long-term changes in the usage of climate and weather words. *Weather* **67**: 171–174. doi:10.1002/wea.1924.

Pielke R. 2010. *The Climate Fix*. Basic Books: New York, NY.

Correspondence to: Will J. Grant

will.grant@anu.edu.au

doi:10.1002/wea.2504