Benford's law in the natural sciences

M. Sambridge,[1] H. Tkalčić,[1] and A. Jackson[2]

Received 23 July 2010; revised 28 September 2010; accepted 4 October 2010; published 17 November 2010.

[1] More than 100 years ago it was predicted that the distribution of first digits of real world observations would not be uniform, but instead follow a trend where measurements with lower first digit (1,2,…) occur more frequently than those with higher first digits (…,8,9). This result has long been known but regarded largely as a mathematical curiosity and received little attention in the natural sciences. Here we show that the first digit rule is likely to be a widespread phenomenon and may provide new ways to detect anomalous signals in data. We test 15 sets of modern observations drawn from the fields of physics, astronomy, geophysics, chemistry, engineering and mathematics, and show that Benford's law holds for them all. These include geophysical observables such as the length of time between geomagnetic reversals, depths of earthquakes, models of Earth's gravity, geomagnetic and seismic structure. In addition we find it also holds for other natural science observables such as the rotation frequencies of pulsars; green-house gas emissions, the masses of exoplanets as well as numbers of infectious diseases reported to the World Health Organization. The wide range of areas where it is manifested opens up new possibilities for exploitation. An illustration is given of how seismic energy from an earthquake can be detected from just the first digit distribution of displacement counts on a seismometer, i.e., without actually looking at the details of a seismogram at all. This led to the first ever detection of an earthquake using first digit information alone. **Citation:** Sambridge, M., H. Tkalčić, and A. Jackson (2010), Benford's law in the natural sciences, *Geophys. Res. Lett.*, *37*, L22301, doi:10.1029/2010GL044830.

1. Introduction

[2] The origin of Benford's law [*Benford*, 1938] goes back to the 19th century, when the astronomer Newcomb [*Newcomb*, 1881] first noticed that library books of logarithms were more thumbed in the earlier pages than the latter. He explained how this could arise if the frequency of first digits themselves were not uniform in real world observations but rather followed the rule

$$P_D = \log_{10}(1 + 1/D) \tag{1}$$

where P_D is the probability of first (non-zero) digit D occurring ($D = 1, \ldots, 9$). For example, the real numbers 123.0 and 0.016 both have $D = 1$, and the digit law suggests that numbers beginning with a 1 will occur about 30% of the time in nature, while those with a first digit of 2 will occur about 17% of the time, and so on down to first digits of 9 occurring about 4% of the time (see Table 1). This decreasing trend of probabilities with digit is shown as a histogram in Figure 1. The implications of the digit rule are significant as not only is the distribution not uniform, implying that digit frequencies are not independent, but to be true it must also hold irrespective of the units of the data as well as their source. Hence a universal property of real world measurements is implied. The result was rediscovered in 1938 by an engineer called Benford [*Benford*, 1938]. Benford also extended the law to arbitrary base, B, and to multiple digits, N. In this case (1) is unchanged except the logarithm base becomes B and D represents the corresponding N-digit integer. (With two digits there are 90 possibilities for D, i.e., $D = 10, 11, \ldots, 99$. As the number of digits increases the probability distribution in (1) tends toward uniformity.)

[3] Benford showed that 20,229 real numbers drawn from 20 sources all approximately followed the same first digit rule. These included populations of cities, financial data and American baseball league averages. Benford's results were well known in mathematical circles and despite a waning of interest his name became associated with the law. Thirty years later the same first digit distribution was noticed in numbers encountered by computers [*Knuth*, 1968]. This led to the suggestion that advanced knowledge of the digit frequency encountered by computers might be used to optimize their design, although this appears never to have been implemented. It has also been suggested that Benford's law (hereafter BL) may provide a novel way of testing realism in mathematical models of physical processes [*Hill*, 1998]. If quantities associated with those processes are known to satisfy BL then computer simulations of them should do also. More recently BL has been shown to hold in stock prices [*Ley*, 1996] and some election results (B. F. Roukema, Benford's Law anomalies in the 2009 Iranian presidential election, ArXiv:0906.2789v3, 2009).

2. Theoretical Insight

[4] Theoretical insight into the origin and reasons for BL was provided only recently [*Hill*, 1995a, 1995b, 1995c, 1998]. It was proven that BL represents the only probability distribution which is both scale and base invariant, properties which such a rule must have to be universally applicable. The scale invariance of BL means that if first digits of the variable x follow (1) then so will the first digits of the rescaled variable λx, for any value of λ. Since the Benford distribution is the only one with this property the converse is also true, i.e., if the first digits of x do not follow (1) then no rescaling will make them do so. It can also be shown that if a real valued random variable x follows a log-uniform distribution, or equivalently if its probability density $P(x) \propto 1/x$, then by simple integration its first digits will follow BL (1).

[5] A second mathematical result is that even if individual distributions of real variables do not follow BL, random

[1]Research School of Earth Sciences, Australian National University, Canberra, ACT, Australia.
[2]Institut für Geophysik, ETH Zürich, Zürich, Switzerland.

0094-8276/10/2010GL044830

Table 1. First Digit Distributions Expressed as Percentages for Various Physical Data Sets[a]

	First Digit Frequencies									Number of Values in Each Data Set	Dynamic Range of the Data (max/min)
	1	2	3	4	5	6	7	8	9		
P_D	30.1	17.6	12.49	9.69	7.92	6.69	5.80	5.12	4.58		
Geomagnetic Field	28.9	17.7	13.3	9.4	8.1	6.9	6.1	5.1	4.5	36512	10^{10}
Geomagnetic reversals	32.3	19.4	13.9	11.8	5.3	4.3	3.2	5.4	4.3	93	10^{3}
Seismic wavespeeds below SW-Pacific	30.0	17.6	13.3	9.8	7.9	6.4	5.6	4.89	4.47	423776	10^{6}
Earth's gravity	33.0	16.6	11.2	8.5	7.5	6.7	5.94	5.57	5.03	25917	10^{9}
Exoplanet mass	33.9	15.4	10.7	9.2	6.23	9.47	5.98	4.48	4.48	401	10^{5}
Pulsars rotation freq	33.9	20.7	12.7	7.6	5.3	5.0	4.94	4.67	4.88	1861	10^{4}
Fermi space telescope γ-ray source fluxes	30.3	17.9	13.0	9.9	7.6	6.96	5.23	5.23	2.72	1451	10^{5}
Earthquake depths	31.6	16.9	14.0	8.69	6.98	7.42	5.27	4.58	4.36	248915	10^{2}
S-A seismogram	28.4	15.7	12.5	9.6	8.97	7.37	6.52	6.04	4.93	24000	10^{5}
Green house gas emissions by country	29.9	17.9	11.4	7.6	9.2	8.15	5.97	4.89	4.89	184	10^{4}
Global Temp anomalies in period 1880–2008	27.7	19.4	12.7	12.1	8.9	5.4	6.61	4.32	2.81	1527	10^{2}
Fund. Phys. constants	34.0	18.4	9.2	8.28	8.58	7.36	3.37	5.21	5.52	326	10^{4}
Global Infectious disease cases	33.7	16.7	13.2	10.7	7.3	5.4	4.56	5.07	3.34	987	10^{6}
Geometric series	29.8	17.4	13.0	10.0	7.8	6.6	5.8	5.0	4.6	1000	10^{21}
Fibbonacci sequence	30.0	17.7	12.5	9.6	8.0	6.7	5.7	5.3	4.5	1000	10^{14}
Combined	30.9	17.4	13.2	9.0	7.6	6.4	5.7	4.8	5.0	10000	10^{33}

[a]The first row is the expected percentage according to Benford's law; the second row is Earth's geomagnetic field model *gufm*1 [*Jackson et al.*, 2000]; the third row is the estimated time in years between reversals of Earth's geomagnetic field for the past 84 million years [*Cande and Kent*, 1995]; the fourth row is seismic body P-wavespeeds of Earth's mantle below the SW Pacific estimated from the inversion of seismic travel times [*Gorbatov and Kennett*, 2003]; the fifth row is spherical harmonic coefficients, up to 160 degrees, of Earth's gravity field (model GGM02S) based on the analysis of 363 days of GRACE in-flight data, spread between April 4, 2002 and Dec 31, 2003 [*Tapley et al.*, 2005]; the sixth row is masses of extrasolar planets taken from the interactive ExtraSolar Planet Catalogue (URL http://www.exoplanet.eu); the seventh row is barycentric rotation frequencies of known pulsars (in Hz) from the ATNF catalogue [*Manchester et al.*, 2005]; the eight row is photon fluxes, in photons/cm^2/s, for 1451 bright objects identified by the Fermi Gamma-ray Space Telescope across the galactic in the first 11 months of operation, August 2008–July 2009 taken from the LAT 1-year point source catalog (URL http://fermi.gsfc.nasa.gov/ssc/data/access/lat/1yr_catalog/); the ninth row is earthquake depths taken from the National Earthquake Information Catalogue (with artificially assigned values at 5, 11, and 33 kms removed); the tenth row is displacement counts measured on a seismometer in Peru (station NNA) for the first 20 minutes following the first recording of the 2004 Sumatra-Andaman earthquake; the eleventh row is emissions of green house gases per country in million tons CO_2 equivalent for 2005 [*Baumert et al.*, 2010]; the twelfth row is global monthly averaged temperature anomalies from the *gistemp* database over the period 1880–2008 measured in degrees with base period 1951–1980 [*Hansen et al.*, 1994]; the thirteenth row is CODATA recommended values for fundamental physical constants [*Mohr et al.*, 2008]; the fourteenth row is total numbers of cases of 18 infectious diseases reported to the World Health Organization by 193 countries worldwide in 2007 [*World Health Organization*, 2009]; the fifteenth row is values from a geometric series ($a_o r^{n-1}$, $n = 1\ldots,10^4$) with starting point $a_o = \pi$ and factor $r = 1.05$ and the sixteenth row is terms in the Fibonacci series $F_n = F_{n-1} + F_{n-2}$, ($F_0 = 0$, $F_1 = 1$). The last row with label "Combined" is the first digit distribution of randomly selected values from all fifteen data sets (each set weighted equally).

samples from those distributions will tend to follow BL, the so called *Random samples from Random distributions* theorem [*Hill*, 1995c]. A practical application of BL that has appeared is in the detection of fraud in financial data and tax returns [*Nigrini*, 1992, 1996]. Natural finance numbers follow BL and human manipulation of such data shows up as anomalies in BL obeying statistics. We have not been able to find any applications of BL to physical phenomena, nor any recognition that the digit rule is widely applicable.

3. Empirical Evidence in the Natural Sciences

[6] Table 1 shows the first digit distributions of 15 data sets with in excess of 750, 000 real numbers with dynamic range varying over 21 orders of magnitude. Here dynamic range is defined as the absolute value of max/min excluding zeroes. The data differ in origin, number, type and physical dimension. The smallest has 93 values (the number of known reversals of Earth's geomagnetic field [*Cande and Kent*, 1995], third row of Table 1) while the largest has more than 400, 000 (seismic wavespeeds of the upper mantle beneath the Pacific [*Gorbatov and Kennett*, 2003], fourth row of Table 1). In all cases a clear trend is observed of decreasing frequency with increasing first digit, as predicted by BL. The fit of each distribution to BL predictions (first row of Table 1) is reasonable. The one exception is the mass distribution of known exoplanets (sixth row of Table 1) which has an excess of values with a first digit of 6. The 6th bin for this data set is about 9.5% whereas BL predicts it to be 6.7%. The 2.8% difference is subject to both sampling and observational error but would correspond to an excess of 11 planets being erroneously assigned a mass with first digit 6. Exoplanet masses can be difficult to estimate and in some cases only a lower bound is possible which may explain this anomaly [*Schneider*, 1999]. In the case of earthquakes (ninth row of Table 1), poorly constrained depths with assigned catalog-values produced large anomalies in bins 1, 3 and 5, corresponding to 5, 11 and 33 kms. Interestingly, once these artificial values are removed, the remainder, based on actual observations becomes consistent with BL. Overall the fit to BL's predictions seems quite striking considering that the nature of the data sets varies from direct observations of physical quantities (like photon fluxes of distant γ sources detected by the Fermi Space Telescope, eighth row of Table 1) to inferences made from indirect measurements (like estimates of the time varying spectral expansion of Earth's geomagnetic field, second row of Table 1), and from well-determined physical constants (thirteenth row of Table 1), to annually varying quantities influenced by human activity (like green house gas emissions, eleventh row of Table 1, and numbers of global diseases infections, fourteenth row of Table 1). An intriguing result is the agreement with BL of temperature anomalies over 128 years of the available record. In this case BL obeying statistics are seen in the geographical fluctuations about a globally increasing trend. Last row of Table 1 ('Combined') contains the first digit distribution of 10,000

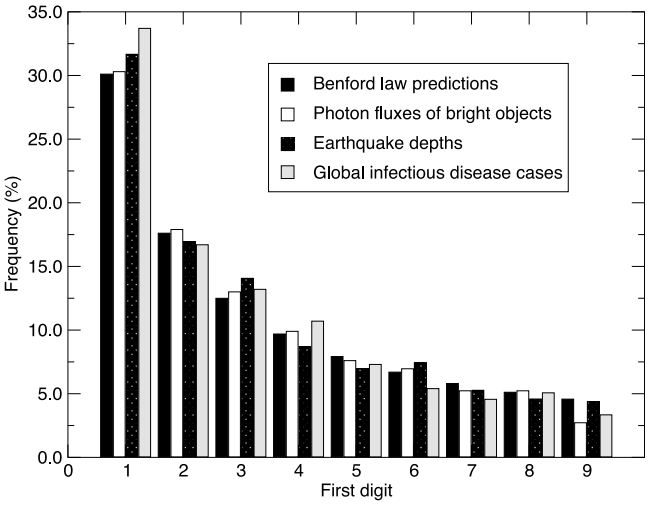

Figure 1. Benford's law predictions according to (1) for distributions of 1st digits compared to three data sets from Table 1. Columns represent eighth row of Table 1, photon fluxes for 1452 bright objects identified by the Fermi space telescope, ninth row of Table 1 depths of 248915 globally distributed earthquakes in the period 1989–2009, and fourteenth row of Table 1 987 reports of infectious disease numbers to World Health Organization in 2007. See Caption of Table 1 for full details. The 1st digit distributions from a wide variety of data sets appear to fit the predictions of the 1st digit law well.

randomly selected values from the 15 individual sets (equally weighted). Here the fit BL is even better, which is consistent with predictions of the random samples theory [*Hill*, 1998].

[7] To quantitatively assess goodness of fit we use a simple Poisson model for sampling error, i.e., where the variance of un-normalized counts in each bin is equal to the mean number of counts. For cases where observational error is small or zero (Table 1) satisfactory χ^2 values are obtained, however as the number of data increase the observed variance in each bin typically becomes larger than predicted by a Poisson model, presumably due to influence of observational errors in the data. The final combined row of Table 1, which is derived from all data sets, gives a normalized $\chi^2 = 1.17 (p = 0.31)$ which indicates an overall satisfactory fit.

4. Exploiting Benford's Law

[8] Our results suggest BL will be a natural feature of data sets with sufficient dynamic range, which raises the question of how it might be exploited. Use in a forensic mode, e.g., to detect fraud or rounding errors, is possible by simply looking for departures in the frequencies of individual digits, as in the sixth row of Table 1. (For other examples see *Nigrini and Miller* [2007] and Roukema (2009).) A more intriguing question is whether BL can be used to detect signals in contrast to background noise, e.g., in time series data. We investigated whether an earthquake could be detected by simply looking at the frequencies of first digits of ground displacement counts recorded by a seismometer. Figure 2a shows the surface displacement produced by the 2004 Boxing day Sumatra-Andaman earthquake recorded at station Nana in Peru (NNA). We compared predicted and observed distributions of first digits within a sliding 200-second window (shown as $t_2 - t_1$ in Figure 2a), for 40 minutes duration centred on the first PKP-wave arrivals from the earthquake. The sampling rate is 20 s^{-1} which gives 48000 counts in total. A goodness of fit measure to BL predictions was calculated for each window using

$$\phi = \left[1 - \left(\sum_{D=1}^{9} \frac{(n_D - nP_D)^2}{nP_D}\right)^{1/2}\right] \times 100 \qquad (2)$$

where n_D is the number of observed data with first digit D, P_D is the proportion of data expected with first digit D from (1), and n is the total number of data. In Figure 2 ϕ is plotted at the end of the sliding time window. Figure 2b shows that first digits of the noise preceding the arrival of the earthquake do not obey BL, where ϕ is below zero, but as soon as the sliding window encounters the seismic waves, at time t_2, ϕ begins to increase. The fit continues to increase steadily as more of the earthquake signal is included in the time window, which illustrates clearly that the presence of earthquakes can be detected from digit information alone, i.e., without ever seeing the details of a seismogram at all. The fact that the earthquake, rather than the noise, follows BL was contrary to our initial expectations, but is possibly explained by the much larger dynamic range of amplitudes in the former (see Figure 2). Histograms of first digits for the entire 20 minute period prior to and after the onset of the earthquake are also shown in Figure 2b (note the lack of digits 1 and 2 in the former).

[9] We also examined the same earthquake recorded at the short period station (CNB) in Canberra, Australia. Figure 2c shows the results. Time t_2 marked on Figure 2c shows the theoretical arrival time of P-waves. As with the station in Peru the fit to BL suddenly increases when the 200 second time window first encounters the Sumatra-Andaman earthquake, i.e., when the time window is in position A (Figure 2c) with its leading edge at time t_2. However rather than a gradual increase in fit, as the earthquake moves into the time window, a more complicated pattern is observed where the initial increase is followed by a decrease and eventual increase again to a peak. Upon closer inspection we noticed that a small local (Canberra) earthquake was recorded at time t_0 about 33s before waves from main event arrived. Figure 2d shows the local earthquake. This event is so small, that it only appears as an increase in high frequency content on the seismogram, while the amplitudes of digital counts remain similar. When the 200s time window reaches position B the trailing edge is at t_0 and the local event begins to exit. This is when the goodness of fit measure shows a sudden increase again indicating that the presence of the small event adversely influences the fit of the digit distribution to BL. As the local event passes out of the time window the digit distribution becomes dominated by the Sumatra-Andaman earthquake and the fit to BL improves again. It seems that the large Sumatra-Andaman earthquake obeys BL whereas the small Canberra event does not. Again a possible explanation is that the dynamic range of counts produced by the local event is too small to fit BL. Nevertheless the presence of the local event is detectable from the digit distribution, as it changes the pattern of the fit curve. After 2000 seconds the fit to BL falls away again as the amplitude decreases and the signal becomes by dominated by longer periods.

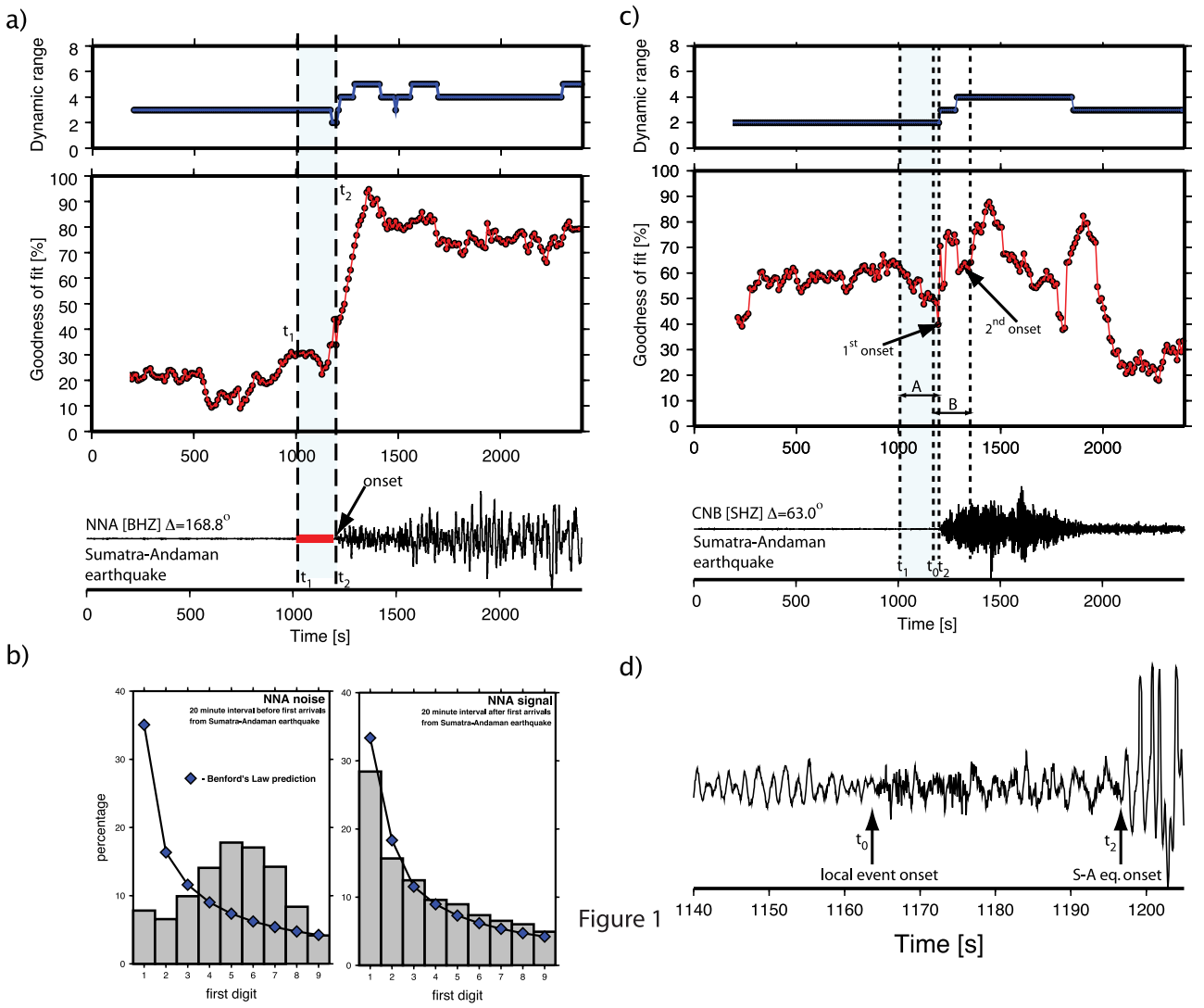

Figure 1

Figure 2. (a) (bottom) Seismogram of the Sumatra-Andaman earthquake recorded at seismic station NNA in Peru. The onset of seismic waves is marked at time t_2. Shading shows the 200-second sliding time-window in position $t_1 \rightarrow t_2$. The earthquake signal enters the moving time-window at time t_2. (middle) Goodness of fit to Benford's law (as defined in the text) as a function of time. (top) Dynamic range as a function of time. (b) Distribution of first digits for the 20-minute period (left) before time t_2 and (right) after time t_2 versus those predicted by Benford's law (blue diamonds). (c) Same as Figure 2a, for the short-period station CNB in Australia. The Sumatra-Andaman earthquake enters the time window at time t_2 (position A) and goodness of fit increases sharply. Time t_0 marks the onset of a small local event (enlarged in Figure 2d). 200 seconds after t_0 the local event begins to leave the time window (position B) which coincides with the point where goodness of fit rises sharply again, as the digit distribution becomes dominated by the major S-A earthquake. (d) The same seismogram over a shorter time period, starting at about 55 seconds before the onset of P-waves.

[10] This simple example is an illustration of how Benford's law may be exploited in seismology. Further work is required to determine whether digit information can be used to improve seismic discrimination in general. Nevertheless it suggests that digit analysis may play a role in discriminating between complex time signals that over print each other. To our knowledge this local Canberra event is the first ever earthquake detected from first digit information alone.

5. Discussion

[11] Our survey suggests that BL may hold across the sciences for data sets with sufficient dynamic range without artificial constraints, e.g., a constant value boundary condition imposed on a computational simulation, which can significantly distort the frequency table of first digits. Localized departures from BL are symptomatic of a different process overprinting the signal. As awareness of this novel phenomenon grows it seems likely that new applications will appear. One possibility is in checking the realism of computer simulations of complex physical processes, such as in the climate or oceans. Another is in the detection and elimination of rounding errors or other non-BL signals in data. We hope this work will encourage others to look at their digits more closely.

[12] **Acknowledgments.** The authors acknowledge all individuals and institutions who have made their data available for this study. Assistance and feedback was received from Pierre Arroucau, Thomas Bodin, Sue Cosetto, Ryan Lister, Peter Rickwood, Roel Snieder and two anonymous reviewers.

References

Baumert, K. A., T. Herzog, and M. Markoff (2010), Climate Analysis Indicators Tool, version 7.0, World Resour. Inst., Washington, D. C. (Available at http://cait.wri.org/cait.php?page=yearly)

Benford, F. (1938), The law of anomalous numbers, *Proc. Am Philos. Soc.*, *78*(4), 551–572.

Cande, S. C., and D. V. Kent (1995), Revised calibration of the geomagnetic polarity timescale for the Late Cretaceous and Cenozoic, *J. Geophys. Res.*, *100*, 6093–6095.

Gorbatov, A., and B. L. N. Kennett (2003), Joint bulk-sound and shear tomography for western Pacific subduction zones, *Earth Planet. Sci. Lett.*, *210*, 527–543.

Hansen, J., M. Sato, R. Ruedy, K. Lo, D. W. Lea, and M. Medina-Elizade (1994), Global temperature change, *Proc. Natl. Acad. Sci. U. S. A.*, *103*, 14,288–14,293, doi:10.1073/pnas.0606291103.

Hill, T. P. (1995a), The significant-digit phenomenon, *Am. Math. Mon.*, *102*, 322–327.

Hill, T. P. (1995b), Base-invariance implies Benford's law, *Proc. Am. Math. Soc.*, *123*, 887–895.

Hill, T. P. (1995c), A statistical derivation of the significant-digit law, *Stat. Sci.*, *10*, 354–363.

Hill, T. P. (1998), The first-digit phenomenon, *Am. Sci.*, *86*, 358–363, doi:10.1511/1998.4.358.

Jackson, A., A. R. T. Jonkers, and M. R. Walker (2000), Four centuries of geomagnetic secular variation from historical records, *Philos. Trans. R. Soc. London*, *358*, 957–990, doi:10.1098/rsta.2000.0569.

Knuth, D. E. (1968), *The Art of Computer Programming*, Addison-Wesley, Reading, Mass.

Ley, E. (1996), On the peculiar distribution of the U.S. stock indexes' digits, *Am. Stat.*, *50*, 311–314.

Manchester, R. N., G. B. Hobbs, A. Teoh, and M. Hobbs (2005), The Australia telescope national facility pulsar catalogue, *Astrophys. J.*, *129*, 1993–2006.

Mohr, P. J., B. N. Taylor, and D. B. Newell (2008), CODATA recommended values of fundamental physical constants: 2006, *Rev. Mod. Phys.*, *80*, 633–730.

Newcomb, S. (1881), Note on the frequency of use of different digits in natural numbers, *Am. J. Math.*, *4*, 39–40.

Nigrini, M. J. (1992), The detection of income evasion through an analysis of digital distributions, Ph.D. thesis, Dep. of Accounting, Univ. of Cincinnati, Cincinnati, Ohio.

Nigrini, M. J. (1996), A taxpayer compliance application of Benford's law, *J. Am. Tax Assoc.*, *18*, 72–91.

Nigrini, M. J., and S. J. Miller (2007), Benford's law applied to hydrology data—Results and relevance to other geophysical data, *Math. Geol.*, *39*, 469–490.

Schneider, J. (1999), The study of extrasolar planets: methods of detection, first discoveries and future perspectives, *C. R. Acad. Sci. Paris*, *327*, 621.

Tapley, B., et al. (2005), GGM02—An improved earth gravity field model from GRACE, *J. Geod.*, *79*, 467–478, doi:10.1007/s00190-005-0480-z.

World Health Organization (2009), *World Health Statistics 2009*, WHO Press, Geneva, Switzerland.

A. Jackson, Institut für Geophysik, ETH Zürich, Sonneggstrasse 5, CH-8093 Zürich, Switzerland.

M. Sambridge and H. Tkalčić, Research School of Earth Sciences, Australian National University, Mills Road, Bldg. 61, Canberra, ACT 0200, Australia. (malcolm.sambridge@anu.edu.au)