Newcomb-Benford Law characterization of solar wind magnetic field and geomagnetic indices

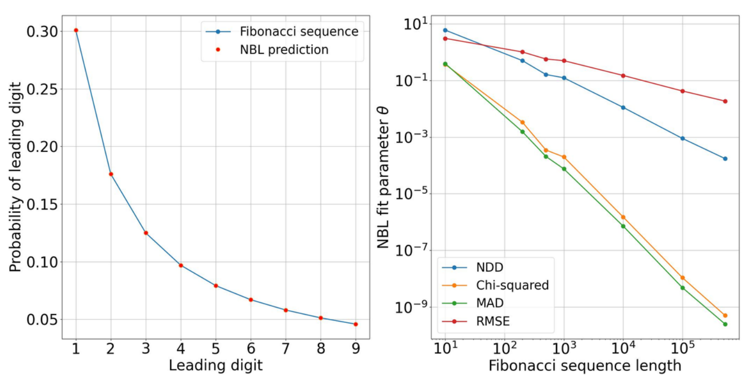
Artur M Benedito-Nunes[1], Jekaterina Gamper[1], Sandra C Chapman[1], Matthew Friel[2], and Jesper W. Gjerloev[2]

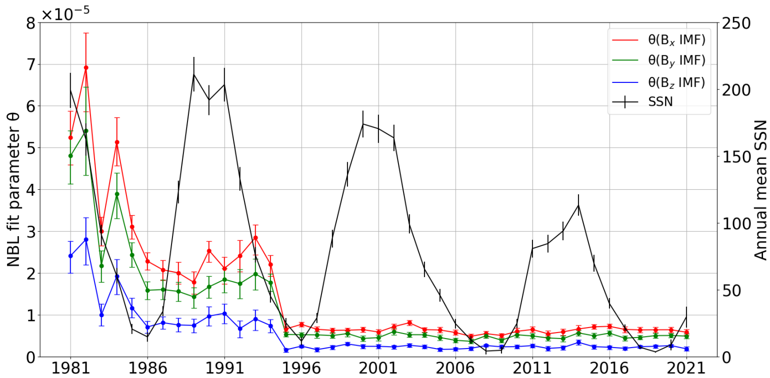
[1]University of Warwick
[2]APL-JHU

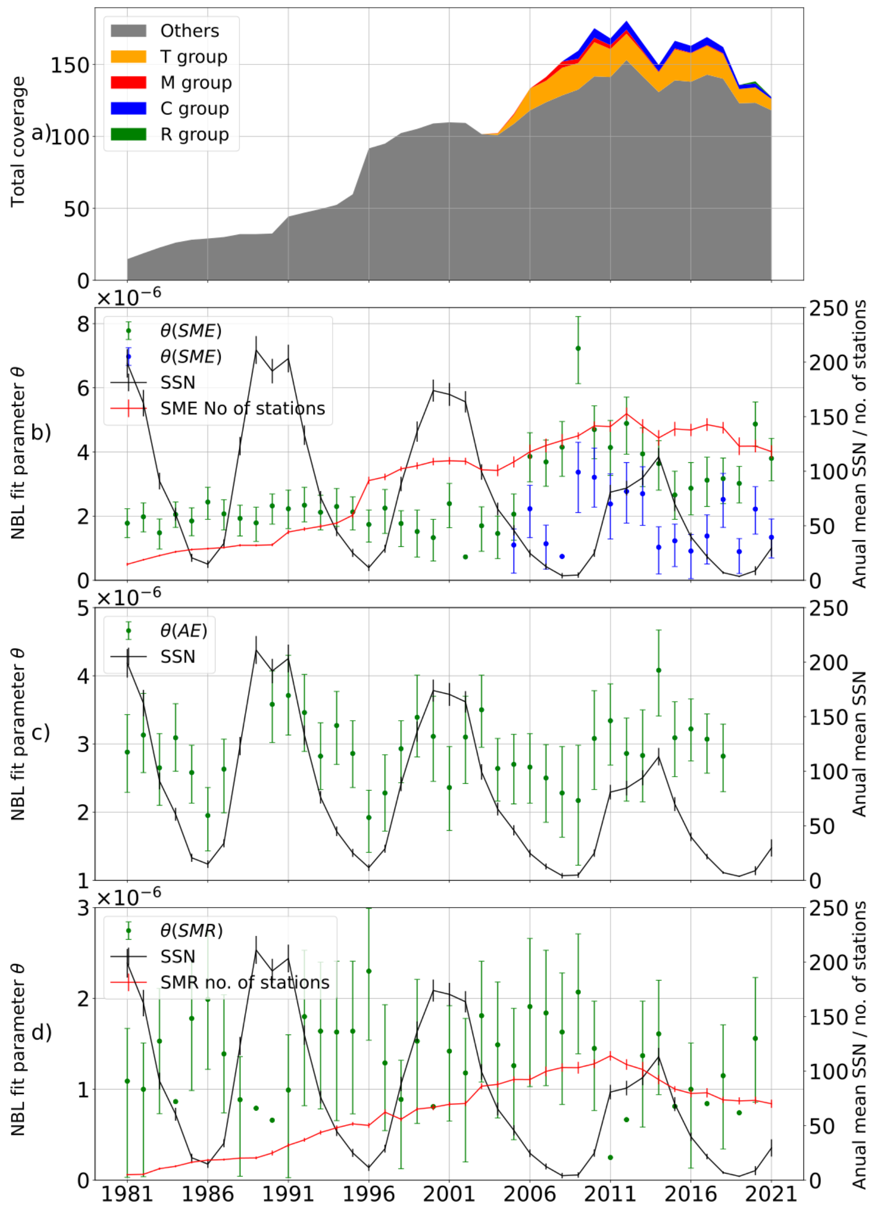
April 11, 2023

Abstract

The Newcomb-Benford Law (NBL) prescribes the probability distribution of the first digit of variables which explore a broad range under conditions including aggregation. Long-term space weather relevant observations and indices necessarily incorporate changes in the contributing number and types of observing instrumentation over time and we find that this can be detected solely by comparison with the NBL. It detects when upstream solar wind magnetic field OMNI HRO Interplanetary Magnetic Field incorporated new data from WIND and ACE after 1995. NBL comparison can detect underlying changes in geomagnetic indices AE (activity dependent background subtraction) and SME (different station types) that select individual stations showing the largest deflection, but not where station data are averaged, as in the SMR index. As composite indices becomes more widespread across the geosciences, the NBL may provide a generic data flag to indicate changes in the constituent raw data, calibration or sampling method.




×10⁻⁵
NBL fit parameter θ
Annual mean SSN
θ(B$_x$ IMF)
θ(B$_y$ IMF)
θ(B$_z$ IMF)
SSN
0
1
2
3
4
5
6
7
8
0
50
100
150
200
250
1981
1986
1991
1996
2001
2006
2011
2016
2021



a)
Total coverage
Others
T group
M group
C group
R group
b)
NBL fit parameter θ
×10⁻⁶
θ(SME)
θ(SME)
SSN
SME No of stations
Anual mean SSN / no. of stations
c)
θ(AE)
SSN
Anual mean SSN
d)
θ(SMR)
SSN
SMR no. of stations
1981
1986
1991
1996
2001
2006
2011
2016
2021

Newcomb-Benford Law characterization of solar wind magnetic field and geomagnetic indices

A.M. Benedito Nunes[1★], J. Gamper[1★], S.C. Chapman[1,2,3], M. Friel[4], J. Gjerloev[4,5]

[1]Centre for Fusion, Space And Astrophysics, Physics Department, University of Warwick, UK
[2]Department of Mathematics and Statistics, University of Tromso, Norway
[3]International Space Science Institute, Bern, Switzerland
[4]Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA
[5]Department of Physics and Technology, University of Bergen, Bergen, Norway

Key Points:

- Space weather relevant parameters and indices follow the Newcomb-Benford Law (NBL) first digit distribution to high precision
- Changes in precision to which the NBL is followed detect instrumentation changes in long-term solar wind parameters and geomagnetic indices
- NBL detects changes in indices that select extremes from constituent stations but not in indices that are multi-station averages

* The two authors contributed equally to this paper.

Corresponding author: Sandra C Chapman, `S.C.Chapman@warwick.ac.uk`

Abstract

The Newcomb-Benford Law (NBL) prescribes the probability distribution of the first digit of variables which explore a broad range under conditions including aggregation. Long-term space weather relevant observations and indices necessarily incorporate changes in the contributing number and types of observing instrumentation over time and we find that this can be detected solely by comparison with the NBL. It detects when upstream solar wind magnetic field OMNI HRO Interplanetary Magnetic Field incorporated new data from WIND and ACE after 1995. NBL comparison can detect underlying changes in geomagnetic indices AE (activity dependent background subtraction) and SME (different station types) that select individual stations showing the largest deflection, but not where station data are averaged, as in the SMR index. As composite indices becomes more widespread across the geosciences, the NBL may provide a generic data flag to indicate changes in the constituent raw data, calibration or sampling method.

Plain Language Summary

Space weather can have significant impact over a wide range of technological systems including power grids, aviation, satellites and communications. In common with studies across the geophysical sciences, space weather modelling and prediction requires long term space and ground-based parameters and indices that necessarily aggregate multiple observations which can change with time. Under certain conditions the Newcomb-Benford law (NBL) specifies the relative occurrence rates of the leading digit in a sequence of numbers arising from aggregation, that is, the number is a result of multiple operations. The NBL specifies that the leading digit, that is, the first non-zero digit in a number, is more likely to be 1 than 2, 2 than 3, and so on, down to 9 which is least likely to occur. In this first application to space weather relevant parameters, we show that how closely the NBL is followed can detect when the instrumentation providing the observations underlying these parameters and indices, or the threshold for background subtraction, has changed. In this era of 'big data', composite indices are becoming more widespread across the geosciences. The NBL may provide a generic data flag to indicate changes in the constituent raw data, calibration or sampling method.

1 Introduction

Benford's Law, also known as the Newcomb-Benford Law (NBL) [Newcomb, 1881; Benford, 1938], prescribes the probability distribution of the first digit of numbers from large sequences under conditions (see Berger & Hill [2021] and refs. therein) that can include scale and base invariance [Pietronero et al,, 2001], aggregation, and the absence of a cut-off [Nigrini, 2000]. Products of random samples from continuous distributions converge to the NBL [Hill, 1995]. The NBL gives the probability of digit d being the first digit of a standard form number in the sequence as $P(d) = \log_{10}(\frac{d+1}{d})$, so that digits $d = 1$ and 2 occur at around 30.1% and 17.61% of the time, respectively, whereas $d = 9$ occurs only 4.58% of the time. Benford [1938] demonstrated it in a wide range of domains including physical constants and physical and societal data. It has been found to apply in a broad range of observations of physical systems [Sambridge et al., 2010] and in the social [Mir, 2012; Pietronero et al,, 2001], and biological [Pröger et al., 2021] sciences. In particular, it has been proposed as a means to detect 'anomalies', that is, changes in time sequences of data, for example providing a means to detect earthquakes [Diaz et al., 2014; Sambridge et al., 2010].

In common with studies across the geophysical sciences, the study of space plasma physics and the climatology of space weather [Pulkkinen, 2007] requires long term space and ground-based observations. Magnetic field observations, both from satellites in-situ and from ground based magnetometers, are an essential component of the modelling and prediction of space weather. Geophysical data is often multipoint in character, with several hundred station observations sampling time-varying fields across the earth's surface. It is common practice across the geosciences to construct indices that capture relevant aspects of a multipoint-sampled spatial field, that is, indices based for example on the average, the variance, a threshold crossing, or an extremum across multiple station data.

An observation of a plasma parameter such as the magnetic field, either in-situ in space, or on the ground, includes various stages of processing of the raw data, involving calibration, removing offsets or background fields, coordinate rotation, and interpolation onto a common, uniform time-base. Geomagnetic indices are derived by combining data from multiple ground based magnetometer stations. The physical processes underlying these observations are also often aggregating, or multiplicative processes such as mixing and turbulence. Given sufficient dynamic range, and in the absence of a cut-off, the NBL might be expected to be followed by both solar wind parameters and geomagnetic indices, at least to some precision.

The station locations, instrumentation, calibration and processing required to derive observed parameters naturally change with time. This suggests the potential for the NBL to provide a flag that indicates that changes have occurred in the details of how long-term observations and indices are derived. In this Letter we will test this idea: that quite subtle changes in the derivation of a parameter or index can be reflected in a statistically significant change in how closely the final data product or index follows the NBL, without any information on the details of how the data product or index was derived.

We will examine how well the first digit distribution of key space weather parameters and indices follow the NBL over time. The solar wind upstream magnetic field has been observed in-situ around L1 since the 1960s [Papitashvili et al., 2020] by a succession of satellites. Comparable 1 minute data is available for the auroral AE [Davis & Sugiura, 1966], SME [Newell and Gjerloev, 2011] and ring current SMR [Newell & Gjerloev, 2012] geomagnetic indices from 1981. AE and SME have different baseline subtraction procedures but are both extremal in the sense that they are both comprised of data from the stations with the largest deflections, whereas SMR is based on a multi-station average. Since 1981, the number of stations comprising AE has not changed over time. The number of stations that comprise SME and SMR has increased by over an order of magnitude, and some more recent stations have different instrumentation. We will see that in some cases, quite small changes in the data underlying these parameters and indices can be detected simply by changes in how closely the parameter or index follows the NBL. It should be emphasised that the closeness to which a quantity follows the NBL is not an indicator of relative quality or precision per-se. Rather, it offers an indicator that there has been a change in the underlying raw observations and the process by which the final parameter or index is derived.

This Letter is organised as follows. In section 2, we describe the datasets and identify the most efficient method to estimate the fit-parameter which quantifies the precision to which the NBL is followed by a finite length sequence of data. In section 3, we estimate the fit parameter over the full records of the OMNI High Resolution Interplanetary Magnetic Field (IMF), and the AE, SME and SMR geomagnetic indices. We conclude in section 4.

2 Methods

2.1 The datasets

A series of solar wind monitors located at the L1 point upstream of the Earth have provided solar wind parameters almost continually since the 1960s. We will consider the Interplanetary Magnetic Field (IMF) for the time interval [1981 - 2021] inclusive at 1 minute resolution as extracted from NASA/GSFC's Modified (Level-3) High Resolution OMNI (HRO) data set through OMNIWeb [Papitashvili et al., 2020]. The parameters are interpolated onto a uniform timebase and mapped to the Earth's bow shock nose. The HRO 1 minute resolution IMF is derived from observations from a series of satellites, and from 1st January 1995 there was a transition from IMP8 only to IMP8, WIND and later, other satellites such as ACE. The data processing method was also modified in 1995.

Auroral indices are designed to monitor the high latitude ionospheric electrojets. The Auroral Electrojet (AE) is the difference between the Auroral Upper (AU) and the Auroral Lower (AL) indices [Davis & Sugiura, 1966]. AU and AL are derived from the 1 minute resolution GSM $\hat{e}$ field component from one of 12 high latitude ground based magnetometer stations in the northern hemisphere. The index takes the value of the data from the stations which at that instant have the largest positive (AU) and largest negative (AL) deflection. Recently, a SuperMAG [Gjerloev, 2012] analog of AE, SME, has been derived from the full set of available stations between +40 and +80 degrees in latitude [Newell and Gjerloev, 2011]. We will consider AE for the interval [1981-2018] inclusive and SME for the interval [1981-2021] inclusive.

Ring-current indices are based on averages over multiple low-latitude station observations. Our study relies on a statistical analysis, therefore rather than focus on the 1 hour time resolution DST index [Sugiura, 1964], we will consider the 1 minute resolution SuperMAG [Gjerloev, 2012] ring-current index SMR [Newell & Gjerloev, 2012]. SMR is derived from all available magnetometer stations within ± 50 degrees of latitude. Following a latitudinal correction, the GSM $\hat{n}$ displacement is first averaged over stations within four 6 hour wide local time windows to give the SMR-00, SMR-06, SMR-12 and SMR-18 local indices. These four local indices are then averaged to give SMR. We consider 1 minute SMR for the interval [1981-2021] inclusive.

Studies of the variations caused by electric currents flowing in the ionosphere and magnetosphere require a subtraction of the dominant and slowly varying Earth main field from the constituent magnetometer observations. The AE index baseline is determined from identified quietest days [Davis & Sugiura, 1966], whereas the SuperMAG indices employ an automated procedure that removes the yearly trend as well as daily variation [Gjerloev, 2012]. The number of stations comprising the AE index does not change over the interval that we will consider here. The SME and SMR indices draw upon a set of SuperMAG collated stations where there is an increase in the number, and changes to the type, of stations over time. Taken together, these index time-series provide a test-bed to see which of these changes in their construction can be detected solely by comparison with the NBL.

Our analysis will also utilize yearly mean total sunspot number (SSN) and dates of solar maxima and minima provided by SILSO.

Our analysis is a statistical comparison between the distribution of first digits of observations from intervals within these data records and the distribution predicted by the NBL. An optimal sample over which to estimate first digit distributions is 1 year, since it is (i) long enough to provide a statistically significant sample (at 1 minute resolution, 525600 data points, 527040 data points in a leap year, assuming no data gaps); (ii) is a sufficiently long time interval for the system to explore its full dynamics (quiet times, substorms and storms) and (iii) is a timescale which is short compared to the 11 year cycle of solar activity and long-term changes in how the parameters and indices are constructed. Any given year-long sample may contain data gaps, and we also exclude records that read zero; uncertainties on these variable length samples are obtained from boostrap resampling of the data as described below.

2.2 Testing for the Newcomb-Benford Law

There has been considerable debate as to the optimal fit parameter that quantifies the precision to which the numbers in a sequence of data obey the NBL [Durtschi et al., 2004; Druicá et al., 2018]. We first perform a systematic comparison of the four most commonly used estimators for the NBL goodness of fit parameter θ in order to select that which has the best performance. Using notation that the leading digit takes the value $i = 1..9$ and has a theoretical occurrence frequency T_i from the NBL and an observed occurrence frequency O_i in the data sequence, the four methods are as follows:

Normalised Distance from Observed Data (NDOD):

$$\theta_{NDOD} = \sum_{i=1}^{9} \frac{|T_i - O_i|}{T_i} \tag{1}$$

the Chi-squared test (Chi):

$$\theta_{Chi^2} = \sum_{i=1}^{9} \frac{(T_i - O_i)^2}{T_i} \tag{2}$$

the Mean Absolute Deviation (MAD):

$$\theta_{MAD} = \sum_{i=1}^{9} \frac{\|O_i - T_i\|}{N \cdot 9} \tag{3}$$

and the Root Mean Square Error (RMSE):

$$\theta_{RMSE} = \sqrt{\sum_{i=1}^{9} \frac{T_i^2 - O_i^2}{T_i^2}} \tag{4}$$

We assess the performance of these four tests by considering the Fibonacci sequence, which closely obeys the NBL [Washington, 1981]. We calculate the fit parameter θ for the leading digit of the first N values of the Fibonacci sequence, where $N = [10, 200, 500, 1000, 10000, 100000, 525600]$. This provides a lower bound for θ as a function of the length of the data record to be tested for each of the four estimators. The left panel of Figure 1 overplots the 1st digit distribution for the Fibonacci sequence of $N = 525600$ on the NBL prediction and the right panel plots the fit parameter θ obtained using the different estimators as a function of N. We can see that the MAD and Chi-squared estimators have higher sensitivity, that is, a larger dynamic range with varying N. We will use the MAD estimator here. The MAD lower bound on the fit parameter (estimated from the Fibonacci sequence) for a sequence that is the length to be tested here, that is, 1 year of minute observations or N=525600, is $\theta \approx 10^{-9}$.

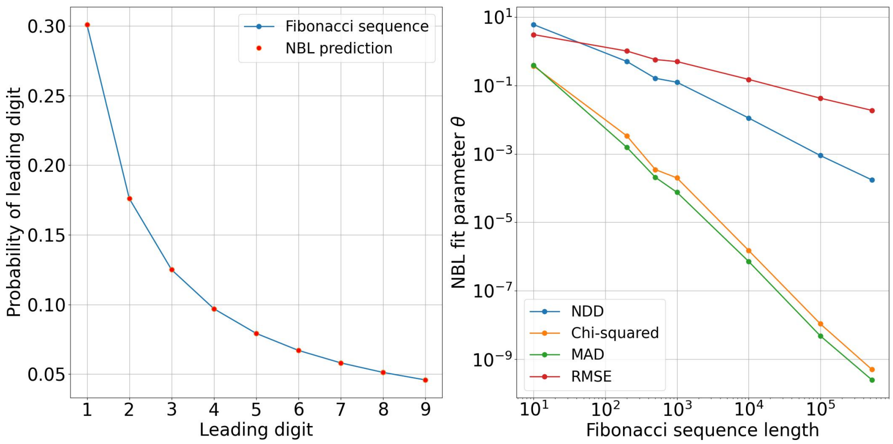


Figure 1. Left: the first digit distribution of the first 525600 numbers in the Fibonacci sequence (blue line) overplotted on the NBL distribution (red circles). Right: fit parameter of a finite Fibonacci sequence plotted as a function of length of the sequence for four estimators: NDD (blue), Chi-squared(yellow), MAD (green) and RMSE (red).

For the data analysis to follow, we will estimate 95% confidence intervals for the fit parameter θ using the stationary bootstrap [Politis & Romano, 1994] to randomly resample the data intervals. The bootstrap method estimates uncertainties by randomly resampling from the data multiple times. It provides a reliable uncertainty estimate under conditions of weak stationarity, and where the sample means form a stable distribution. The optimal length of the bootstrapping block was obtained using the method outlined in [Politis & White, 2004]. The stationary bootstrap and block length selection algorithm were implemented using the python library *arch* [Sheppard, 2021]. The Python function *arch.bootstrap.StationaryBootstrap.conf_int*, used to calculate the confidence interval, required the following inputs: seed, number of bootstrap replications, method, size, and sampling which we set to the following values, respectively: 66, 1000, "basic" (also known as empirical bootstrap), 0.95, nonparametric. We checked the validity of the bootstrap estimates by examining the distribution of the fit parameter obtained from the bootstrap re-samples. We have discarded estimates of the confidence interval where the distribution of the fit parameter for the bootstrap re-samples was not single-peaked, as well as where the confidence interval did not converge.

3 Results

3.1 Solar wind Interplantary Magnetic Field at L1

The HRO IMF dataset provides a test case to see if the NBL can detect changes in instrumentation and processing of observations for observational time series. We use the MAD estimator (eq. 3) to obtain the fit parameter for non-overlapping year-long samples of 1 minute resolution HRO IMF. Figure 2 plots the resulting fit parameters θ for IMF GSE x,y, and z components, along with 95% confidence intervals. The NBL is followed quite closely, $\theta < 10^{-4}$ across almost the entire record. However the precision to which the NBL is followed progressively improves in the first 5 years of the record then is flat until 1995, where there is a step-change improvement (lower value) in the fit parameter of over a factor of three, which significantly exceeds the 95% confidence intervals. The fit parameter is constant thereafter.

In 1995 there was change in the contributing satellites to OMNI HRO and to the processing procedure

(https://omniweb.gsfc.nasa.gov/html/HROdocum.html see also [Alterman, 2022]). Prior to 1995, the underlying observations were from IMP8 only, post 1995, they also included WIND and later ACE. The availability of WIND and ACE also resulted in fewer data gaps per year from on average, 75% of the data entries before 1995 to 8.7% after 1995; this is reflected in smaller bootstrap estimated uncertainties post 1995.

Another factor that could affect the precision to which the first digit distribution of the data follows the NBL is the dynamic range explored by the underlying observations. Increased dynamic range could improve the NBL fit precision, which might be expected to come into play during active intervals of the solar cycle. To investigate this, we overplotted on Figure 2 the yearly mean total sunspot number and we can see that the precision to which the NBL is followed is not sensitive to the overall level of activity.

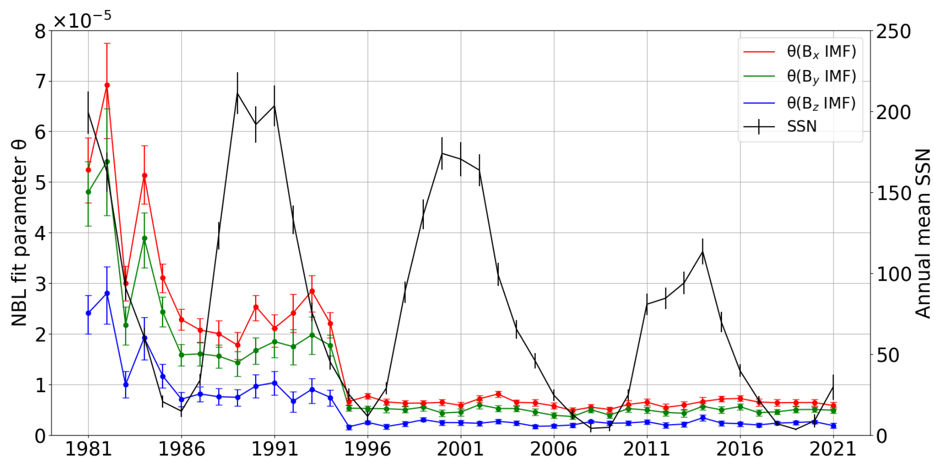


Figure 2. The MAD-estimated fit parameter θ (left ordinate) for solar wind IMF GSE $\hat{x}$ (red), $\hat{y}$ (green) and $\hat{z}$ (blue) components, estimated for 1 year non-overlapping samples, with bootstrap 95% confidence limits, are plotted versus time. Smaller fit parameter values indicate closer correspondence to the NBL first digit distribution. Yearly averages of daily sunspot number (right ordinate) is plotted (black), error bars denote the standard deviation for that year.

3.2 Geomagnetic indices

Geomagnetic indices are derived from observations from individual magnetometer stations. Before considering geomagnetic indices, we first investigated how closely the data from individual magnetometer stations follow the NBL. Some sample values for the NBL fit parameter of year-long samples of GSM magnetometer data with SuperMAG baseline subtraction are: Pebek [2014], $\hat{n}$ component: $\theta = 1.22 \times 10^{-6}$; Yellowknife [2001], $\hat{e}$ component: $\theta = 1.24 \times 10^{-7}$; Abisko [1990], $\hat{z}$ component $\theta = 2.06 \times 10^{-7}$. Given that the underlying magnetometer data follows the NBL, we would expect geomagnetic indices to follow it also to some precision. The AE and SME auroral indices are essentially comprised of data taken from the pair of ground stations that at any time observe the maximum (positive and negative) magnetic field deflections. SMR on the other hand is a multi-station average.

Estimates of the NBL fit parameter θ from non-overlapping year-long samples are plotted in Figure 3 for the SME, AE and SMR indices, with 95% confidence limits. The figure examines the effect on SME and SMR of changing station number and coverage and changes in class of magnetometer. The figure also examines the effect of different baseline removal in construction of the index by comparing SME and AE.

Panel (a) of Figure 3 plots the overall coverage provided by the ground based magnetometers collated by SuperMAG. For each year we sum over the fraction of the year that each station is taking data to obtain the total operating station-years, so that if m stations were taking data for the entire year, this would give m operating station-years. Colours discriminate a subset of stations which were introduced after 2003 which use a different class of magnetometer, these are Magstar, CARISMA, McMac, and THEMIS project operated stations coded as R, C, M and T in the SuperMAG catalog [Gjerloev, 2009, 2012]. The coverage from all other stations are indicated by grey in panel (a) of Figure 3. The SuperMAG collated stations then provide a test dataset to see if the NBL is sensitive to (i) an increase in station number but no change in class of magnetometer, as occurs before 2003 and (ii) the inclusion of a different class of magnetometer as occurs after 2003. Figure 3 panel (b) then plots the NBL parameter θ for the SME index derived from all available SuperMAG stations (green) overplotted on the NBL parameter for SME constructed excluding the R, C, M and T stations (blue), that is, just including all 'other' stations (indicated in grey in top panel). The Figure then shows that increasing the number of stations, that is, the spatial coverage, for stations of the same magnetometer class, does not change the NBL fit parameter: there is no change in the NBL fit parameter between the early record, and 1996-2002, over which period the number of magnetometers has increased by an order of magnitude. However, after 2005 there is a statistically significant divergence between the NBL for SME for the full set of stations (which now include the R, C. M and T stations), and with the R, C, M and T stations excluded.

Panel (c) of Figure 3 plots the NBL fit parameter for the AE index which is comprised of a fixed number of stations during this interval. In the first half of the AE data record there is a statistically significant correlation between the NBL fit parameter and the variation in the SSN over stronger solar cycles 22 (maximum in 1989) and 23 (maximum in 2001), it is less evident evident over weaker cycle 24. For SME, there is no statistically significant solar cycle variation in the NBL fit parameter over cycles 22 and 23. The AE index baseline is determined from identified quietest days [Davis & Sugiura, 1966], whereas the SuperMAG indices do not use the concept of quietest days, instead, an automated procedure that removes the yearly trend as well as daily variation is employed [Gjerloev, 2012]. The AE baseline will therefore track the overall level of geomagnetic activity in a different manner to SME. If the quietest days around strong solar maxima are more active than the quietest days around solar minima, then a baseline determined from those most quietest days will in turn track the yearly averaged SSN. During active years, a raised baseline would then act as a low-end cut-off which would increase the value of the NBL fit parameter. It should be emphasised that both the AE and SME records follow the NBL to high precision; changes in the NBL fit parameter are nevertheless sensitive to quite small changes in the underlying magnetometers and in the baselines used.

The NBL fit parameter for SMR is plotted in panel (d) of Figure 3, alongside the yearly averaged SSN and the total number of SuperMAG constituent stations. The SMR fit parameter is essentially constant within the bootstrap 95% confidence intervals. This suggests that the NBR fit parameter of an average over many stations is less sensitive to changes in its constituent data, in this case, the inclusion of different instrumentation post 2006.

4 Conclusions

The Newcomb-Benford Law (NBL) prescribes the probability distribution of the first digit of standard form number sequences under conditions which include aggregation (the values arise from multiple operations) scale and base invariance, and the absence of strong truncation. Long-term parameters and indices are in widespread use across the geosciences and the constituent instrumentation and construction methodology will necessarily change over time; we have investigated how the NBL can be used to flag these changes.

We explored the precision to which the NBL is followed by long-term parameters and indices that are central to the monitoring of space weather. We considered non-overlapping yearly samples of the solar wind interplanetary magnetic field (IMF) monitored at L1, and the AE, SME and SMR geomagnetic indices available at minute resolution over multiple solar cycles. Our results are as follows:

1. The OMNI (HRO) IMF, and indices AE, SME and SMR all follow the NBL to high precision (fit parameter $\theta \sim 10^{-6}$).
2. A change in the NBL fit parameter θ for the OMNI high resolution IMF parameter occurs when the data source changes from IMP8 to include data from other spacecraft such as WIND and ACE and the processing method was modified.
3. The SMR index which averages over multiple ground-based magnetometer timeseries, follows the NBL to a consistent precision across changing solar activity, a ten-fold increase in the number of stations comprising the index, and the introduction of different classes of constituent magnetometer.
4. A change in the NBL fit parameter for the SME auroral index occurs when there is a change in the class of constituent magnetometer but not when the number of the same class of stations increases.
5. Unlike the SME index, the AE index follows the NBL to a precision that tracks the relatively strong SSN variation of solar cycles 22 and 23, consistent with the latter using a baseline determined fom geomagnetically quietest days.

These results have practical implications for the design and use of long-term parameters and indices. We have examined geophysical parameters and indices which in all cases follow the NBL to high precision. Quite subtle changes in the underlying instrumentation and differences in the subtracted baseline can be detected by the NBL in long-term records of parameters (here, the IMF) and in indices that select single time-series from the set of stations (here, auroral indices). The latter may also be expected to apply to indices that select on a high threshold, again being comprised of a few timeseries selected from the set of observing stations. In all these cases, the NBL could provide a data flag that would indicate to the user that further investigation is needed in how a long-term parameter or index is utilised. Such a data flag would be informative without any detailed knowledge of how the parameter or index is constructed, important since parameters and indices are designed for widespread application as benchmarks of activity. The NBL is not sensitive to changes in the construction of indices that average or aggregate over many stations (here, ring current indices), consistent with the aggregating process driving the data records towards closer correspondence to the NBL.

We have found that how closely the NBL first digit distribution is followed is sensitive to changes in how parameters and indices are constructed. This is distinct from tracking physical changes in the system that they are designed to parameterize. The NBL fit parameter does not track the variation in activity (smoothed SSN), of the last four solar cycles in the IMF at L1, in SME or SMR. The distribution of solar wind parameters do show solar cycle variation [Tindale & Chapman, 2016] and the top few percent of the data records of both AE and SMR also track the solar cycle [Bergin et al., 2022]. Auroral indices such as AE and SME sample the ground magnetic perturbations from high-latitude current systems, the largest of which are the auroral electrojets. Auroral electrojet intensity tracks the solar cycle [Smith et al., 2017] and will have a maximum possible intensity, this is seen in auroral indices [Nakamura et al, 2015]. The electrojets are geographically localized, so that as the number of SME stations is increased, it is more likely that a station will be located in the vicinity of the maximum ground magnetic deflection. It has indeed been shown that the AE record systematically undersamples when compared to SME for later solar cycles [Bergin et al., 2020] as the number of stations comprising SME has increased. This change is not seen in the NBL fit parameter; for SME it does not change as the number of constituent stations is increased over an order of magnitude.

Acknowledgments

Artur Miguel Benedito Nunes and Jekaterina Gamper are co-first authors. We thank the University of Warwick Institute for Advanced Teaching and Learning. SCC acknowledges support from ISSI via the J. Geiss fellowship and AFOSR grant FA8655-22-1-7056 and STFC grant ST/T000252/1. We acknowledge the SuperMAG collaborators: https://supermag.jhuapl.edu/info/?page=acknowledgement

We acknowledge use of NASA/GSFC's Space Physics Data Facility's OMNIWeb service, and OMNI data and the experiment teams that acquired, processed and provided the OMNI-included data, and J.H. King and N.E. Papitashvili of NASA/GSFC for creating the OMNI data set. We acknowledge the WDC for Geomagnetism, Kyoto for the provision of the AE index data. We thank the World Data Center SILSO, Royal Observatory of Belgium, Brussels for provision of sunspot data.

Open Research

All data used in this study is freely available from the following sources (accessed on 1st October 2022).
SuperMAG [Gjerloev, 2012] indices: https://supermag.jhuapl.edu/

The AE index from the WDC for Geomagnetism, Kyoto [Nose et al, 2015] http://wdc.kugi.kyoto-u.ac.jp/wdc/Sec3.html

OMNI [Papitashvili et al., 2020] Solar wind parameters: https://omniweb.gsfc.nasa.gov/form/omni_min_def.html
OMNI HRO Documentation: https://omniweb.gsfc.nasa.gov/html/HROdocum.html

SILSO Royal Observatory of Belgium, Brussels daily total sunspot number version 2.0 from 1818: http://www.sidc.be/silso/home

The dates of solar cycle maxima and minima are as determined from the smoothed sunspot number record by SILSO: http://www.sidc.be/silso/cyclesmm

Stationary bootstrap and block length selection algorithms were implemented using the Python library of Sheppard [2021]: bashtage/arch: Release 5.3.1 (version 5.3.1). Retrieved from doi: 10.5281/zenodo.593254439

The Python function *arch.bootstrap.StationaryBootstrap.conf_int* is used to calculate the confidence interval.

References

Alterman B. L., (2022) Plasma Data Sources in the OMNI Database, Res. Notes AAS 6 135 doi:10.3847/2515-5172/ac7a2f

Berger, A., Hill, T.P. The mathematics of Benford's law: a primer, Stat Methods Appl 30, 779–795 doi:10.1007/s10260-020-00532-8

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

Bergin, A., S. C. Chapman, J. Gjerloev, (2020) AE, DST and their SuperMAG Counterparts: The Effect of Improved Spatial Resolution in Geomagnetic Indices, J. Geophys. Res., doi:10.1029/2020JA027828

Bergin, A., S. C. Chapman, N. Moloney, N. W. Watkins, (2022) Variation of geomagnetic index empirical distribution and burst statistics across successive solar cycles, J. Geophys. Res, doi:10.1029/2021JA029986

Diaz, J., Gallart, J., Ruiz, M. (2014) On the ability of the Benford's law to detect earthquakes and discriminate seismic signals. Seismological Res. Lett. 86 192-201. doi:10.1785/0220140131.

Davis, T. N., Sugiura, M. (1966). Auroral electrojet activity index AE and its universal time variations. J. Geophys. Res., 71 , 785–801 doi:10.1029/JZ071i003p00785

Druicá, E., Oancea, B., Vâlsan, C. (2018) Benford's law and the limits of digit analysis. Intl. J. Accounting Information Systems, 31 , 75–82 doi: 10.1016/j.accinf.2018.09.004

Durtschi, C., Hillison, W., Pacini, C. (2004). The effective use of Benford's law to assist in detecting fraud in accounting data. Journal of forensic accounting research, 5 , 17–34.407

Gjerloev, J. W. (2009), A Global Ground-Based Magnetometer Initiative, EOS, 90, 230-231, doi:10.1029/2009EO270002.

Gjerloev, J. W. (2012). The SuperMAG data processing technique. J. Geophys. Res., 117, 411 doi:10.1029/2012JA017683

Hill, T.P. (1995) A Statistical Derivation of the Significant-Digit Law. Statistical Science, 10, 354-363. doi: 10.1214/ss/1177009869

Mir, T. A. (2012). The law of the leading digits and the world religions. Physica A 416, doi: 10.1016/j.physa.2011.09.001

Nakamura, M., Yoneda, A., Oda, M., Tsubouchi, K. (2015). Statistical analysis of extreme auroral electrojet indices. Earth Planets Space, 67, 153 doi:10.1186/s40623-015-0321-0

Newcomb, S, (1881) Note on the frequency of use of the different digits in natural numbers. American Journal of Mathematics. 4 (1/4): 39–40. doi:10.2307/2369148.

Newell, P. T., Gjerloev, J. W. (2011). Evaluation of SuperMAG auroral electrojet indices as indicators of substorms and auroral power. J. Geophys. Res. 116 (A12) 422 doi: 10.1029/2011JA016779

Newell, P. T., Gjerloev, J. W. (2012). SuperMAG-based partial ring current indices J. Geophys. Res. 117 424 doi:10.1029/2012JA017586

Nigrini, M.J.: Digital Analysis Using Benford's Law. Global Audit Publications, Vancouver, B.C., Canada (2000)

Papitashvili, N. E., King, J. H. (2020). Omni 1-min data – [GSE solar wind, AE index, 1981-2021]. NASA Space Physics Data Facility. Retrieved from doi:10.48322/45bb-8792 (Accessed on 23-09-2022)

Pietronero, L., E. Tosatti, V. Tosatti, A. Vespignani (2001) Explaining the uneven distribution of numbers in nature: the laws of Benford and Zipf, Physica A Stat. Mech. 293, 297 doi: 10.1016/S0378-4371(00)00633-6

Politis, D. N., Romano, J. P. (1994). The stationary bootstrap. J. Am. Stat. Assoc., 89, 1303–1313 doi:10.2307/2290993

Politis, D. N., White, H. (2004). Automatic Block-Length selection for the dependent bootstrap. Econometric Reviews, 23, 53–70 doi:10.1081/ETC-120028836

Pröger, L., Griesberger, P., Hackländer, K., Brunner, N., Kühleitner, M. (2021) Benford's law for telemetry data of wildlife. Stats, 4, 943–949, doi:10.3390/stats4040055

Pulkkinen, T. Space Weather: Terrestrial Perspective. Living Rev. Sol. Phys. 4, 1 (2007) doi:10.12942/lrsp-2007-1

Sambridge, M., Tkalci̦', H., Jackson, A. (2010). Benford's law in the natural sciences. Geophys. Res. Lett., 37, 437 doi: 10.1029/2010GL044830

Sheppard, K. (2021). bashtage/arch: Release 5.3.1 (version 5.3.1). Retrieved from doi: 10.5281/zenodo.593254439

Sugiura, M. (1964). Hourly values of equatorial Dst for the IGY. Ann. Int. Geophys., 35 (9).445

Smith, A. R. A., Beggan, C. D., Macmillan, S., Whaler, K. A. (2017). Climatology of the auroral electrojets derived from the along-track gradient of magnetic field intensity measured by POGO, Magsat, CHAMP, and swarm. Space Weather, 15, 1257–1269 doi: 10.1002/2017SW001675

Tindale, E., Chapman, S. C. (2016). Solar cycle variation of the statistical distribution of the solar wind ϵ parameter and its constituent variables. Geophys. Res. Lett., 43, 5563–557 doi:10.1002/2016GL068920

Washington, L. C. (1981). Benford's Law for Fibonacci and Lucas Numbers. The Fibonacci Quarterly. 19 (2): 175–177.
World Data Center for Geomagnetism, Kyoto, M. Nose, T. Iyemori, M. Sugiura, T. Kamei (2015), Geomagnetic AE index, doi:10.17593/15031-54800

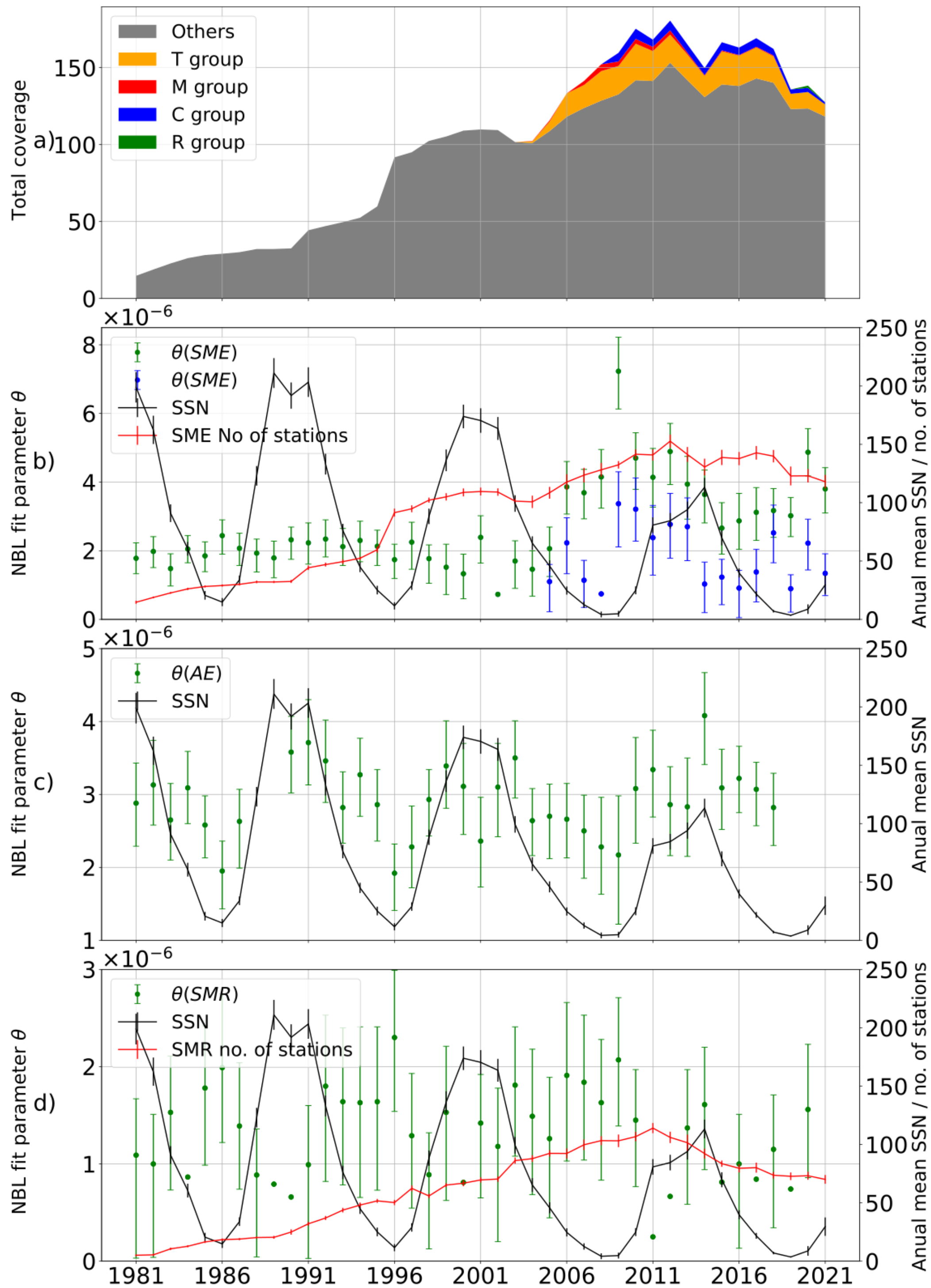
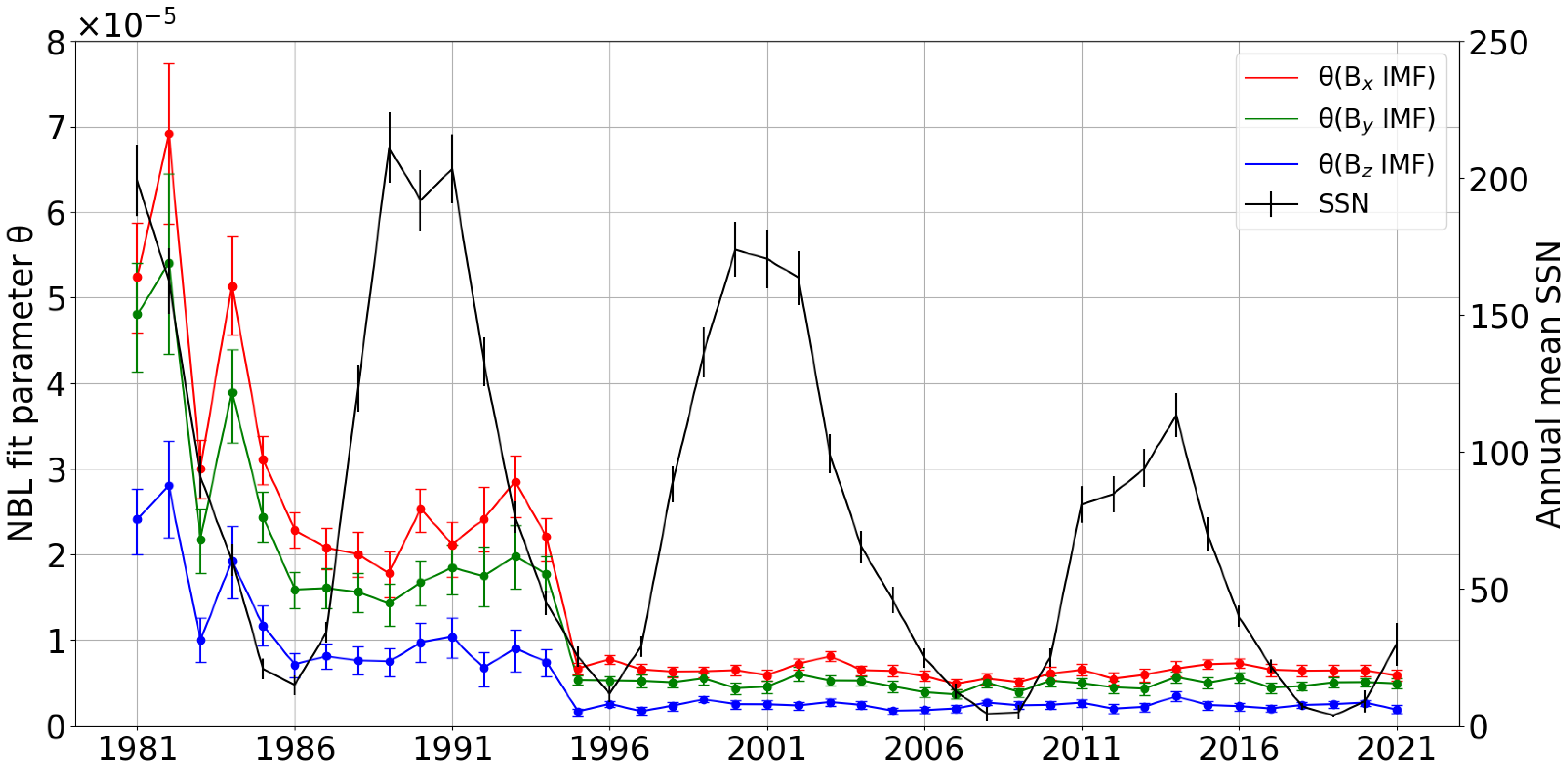


Figure 3. Panel (a): Stack plot of the coverage (total operating station-years) for different classes of SuperMAG stations. A different class of instrumentation is introduced after 2003, colours indicate specific SuperMAG station classifications. Panel (b): Left ordinate refers to the NBL fit parameter θ for non-overlapping yearly samples of the SME index. Green circles plot θ for SME derived from all stations overplotted on blue circles which plot θ for SME derived excluding R, C, M and T group stations. Panel (c) Left ordinate refers to the NBL fit parameter θ for non-overlapping yearly samples of the AE index (green circles). The fit parameter is not plotted for years 1988 and 1989 where there are significant data gaps in AE. Panel (d) Left ordinate refers to the NBL fit parameter θ for non-overlapping yearly samples of the SMR index (green circles). On panels (b-d), error bars plot bootstrap estimated 95% confidence interval uncertainties on the NBL fit parameter. The right ordinate refers to the yearly averaged SSN (black line), and in panels (b) and (d), to the annual mean number of all SuperMAG stations that operate within each year (red line).

Figure1.

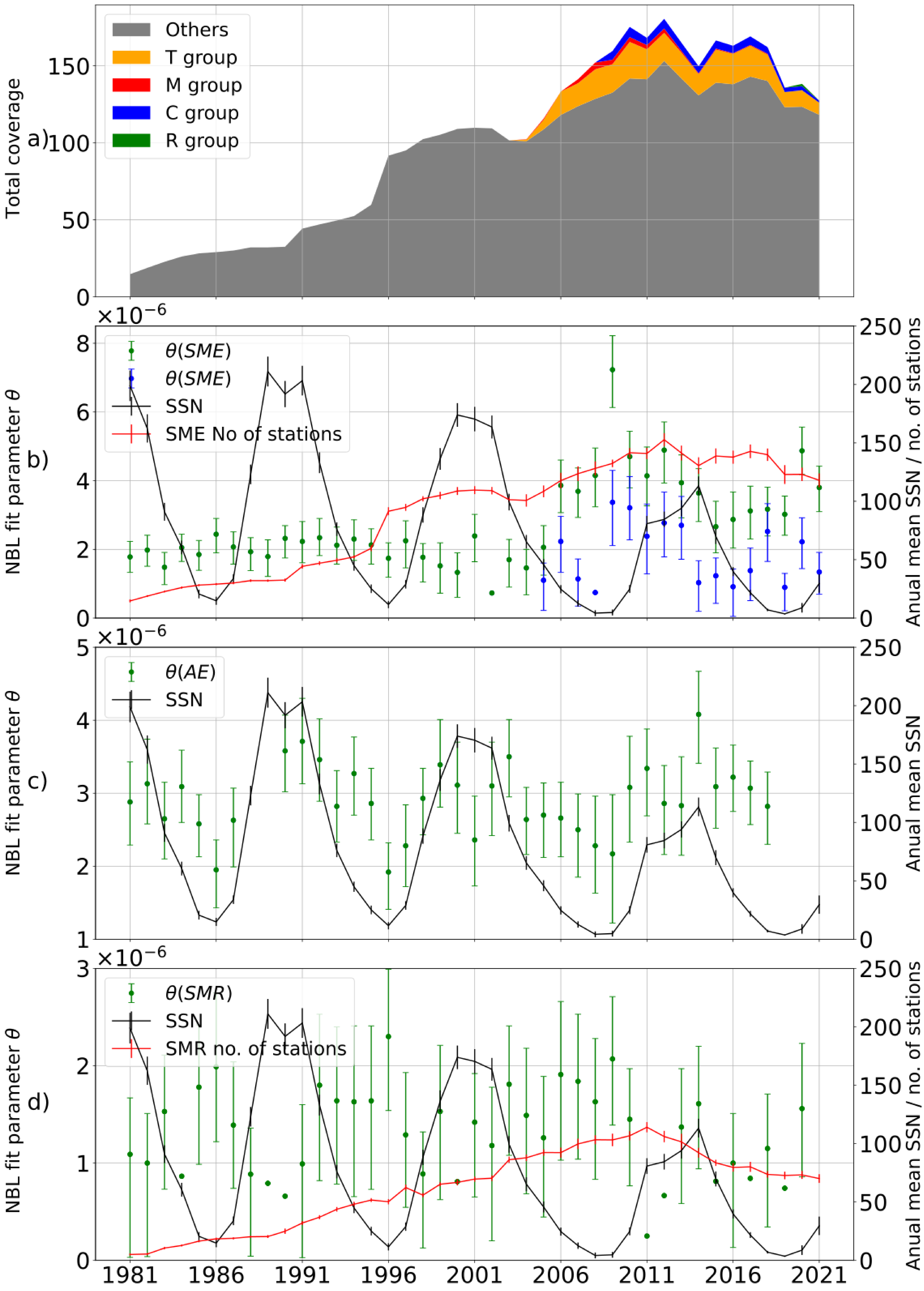
Fibonacci sequence
NBL prediction
Probability of leading digit
Leading digit
0.30
0.25
0.20
0.15
0.10
0.05
1
2
3
4
5
6
7
8
9
NBL fit parameter θ
Fibonacci sequence length
NDD
Chi-squared
MAD
RMSE
10^{1}
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
10^{1}
10^{2}
10^{3}
10^{4}
10^{5}

Figure2.



×10⁻⁵
8
7
6
5
4
3
2
1
0
NBL fit parameter θ
250
200
150
100
50
0
Annual mean SSN
1981
1986
1991
1996
2001
2006
2011
2016
2021
θ(Bx IMF)
θ(By IMF)
θ(Bz IMF)
SSN

Figure3.



Others
T group
M group
C group
R group
Total coverage
a)
0
50
100
150
×10^-6
b)
NBL fit parameter θ
θ(SME)
θ(SME)
SSN
SME No of stations
Anual mean SSN / no. of stations
0
2
4
6
8
50
100
150
200
250
×10^-6
c)
NBL fit parameter θ
θ(AE)
SSN
Anual mean SSN
1
2
3
4
5
×10^-6
d)
NBL fit parameter θ
θ(SMR)
SSN
SMR no. of stations
Anual mean SSN / no. of stations
1981
1986
1991
1996
2001
2006
2011
2016
2021