International Journal of Climatology

International Journal of Climatology | RMetS

RESEARCH ARTICLE OPEN ACCESS

Utilising Benford's Law in the Validation of Precipitation Datasets

Amee Gollop[1] | Sarah Wilson Kemsley[1,2] | Tim Osborn[1] | Manoj Joshi[1] | David Stevens[3] | Ian Harris[1]

[1]Climatic Research Unit, School of Environmental Sciences, University of East Anglia, Norwich, UK | [2]School of Geography and the Environment, University of Oxford, Oxford, UK | [3]Centre for Ocean and Atmospheric Sciences, School of Mathematics, University of East Anglia, Norwich, UK

Correspondence: Sarah Wilson Kemsley (s.wilson-kemsley@uea.ac.uk)

Received: 3 July 2025 | **Revised:** 10 November 2025 | **Accepted:** 30 November 2025

Keywords: Benford's law | data pipeline | data quality | data validation | error detection | hydroclimatology | precipitation data

ABSTRACT

The increasing magnitude and complexity of precipitation datasets necessitate robust and efficient data integrity assessment. This study systematically applies Benford's Law, a mathematical theorem describing leading digit frequencies, as a novel diagnostic tool for precipitation data in the environmental and hydroclimate sciences. We present a reproducible and robust methodology, demonstrating that global monthly precipitation consistently conforms to Benford's Law across diverse data types, including raw observations, gridded products, reanalysis and synthetic simulations. This key finding fundamentally challenges traditional assumptions regarding the influence of data origin on Benford's Law adherence, significantly broadening its applicability. Our findings underscore the importance of underlying quantitative characteristics for successful application: while regional analyses reveal that monthly precipitation data in the United Kingdom and Ireland do not conform to Benford's Law-based principles, a shift to daily temporal granularity successfully restores conformance, highlighting how temporal resolution can introduce the necessary data properties. This research uniquely positions Benford's Law as a powerful, complementary diagnostic tool capable of detecting subtle data corruptions, as demonstrated through an artificial experiment. Ultimately, this work advances the utility of Benford's Law in climate research, providing a scalable method to enhance the reliability of foundational datasets critical for climate modelling, forecasting and a wide array of hydroclimatological applications.

1 | Introduction

The rapid growth of large digital environmental datasets driven by technological advancements, underscores the necessity for robust data validation in applications such as machine learning (Huntingford et al. 2019; Kaltenborn et al. 2023), climate modelling (Brönnimann et al. 2018; Bosilovich et al. 2013) and downscaling techniques (Wilby and Wigley 1997). However, errors arising from instruments, human input or data transfer can compromise dataset integrity. This highlights a critical need for effective and scalable methods to assess dataset integrity.

The concept now known as 'Benford's Law' originated from physicist Frank Benford's (1938) observation that the early pages of logarithm tables, corresponding to numbers beginning with '1', were notably more worn than those for higher leading digits (Benford 1938). Thus, Benford's Law states that the frequency of leading digits in datasets follows a logarithmic distribution,

Abbreviations: CRU-TS, Climatic Research Unit Terrestrial Series gridded product; CRU-TSRaw, Climatic Research Unit Terrestrial Series raw ungridded input dataset; ECMWF, European Centre for Medium-Range Weather Forecasts; ERA5, Fifth-Generation ECMWF Atmospheric Reanalysis; ERA5-TS, ERA5 Terrestrial Series (land-based precipitation data); ERA5UK&I-DTS, ERA5 United Kingdom and Ireland Daily Terrestrial Series; ERA5UK&I-TS, ERA5 United Kingdom and Ireland Terrestrial Series; LD, Lead Digit; PDF, Probability Density Function; ROM, Robust Measure of Orders of Magnitude; Stochastic-TS, Stochastically Generated Terrestrial Series; UK, United Kingdom.

with lower digits appearing more often than higher ones. This phenomenon had also been independently discovered earlier by astronomer Simon Newcomb in 1881 (Newcomb 1881). Historically, Benford's Law became associated with 'natural' datasets, a perception rooted in these initial observations across varied real-world data. However, more recent theoretical advancements have refined this understanding, revealing that its adherence stems from a dataset's inherent mathematical distribution rather than its natural origin, thus expanding its applicability across diverse data sources (Cong et al. 2019; Wang and Ma 2024).

The predictable distribution of lead digits posited by Benford's Law is defined by the logarithmic formula:

$$P_d = \log_{10}\left(1 + \frac{1}{d}\right), \tag{1}$$

where P_d represents the probability of randomly selecting a number with lead digit $d \in \{1, 2, 3, \ldots, 9\}$. Many authors have provided elegant proofs for this 'lead digit law', for example, Hill (1995) and Wang and Ma (2024), and its versatility is highlighted by its many useful applications. These include: financial fraud detection (Grammatikos and Papanikolaou 2021), hydrological stream flow data assurance (Nigrini and Miller 2007), image alteration detection (Sheng et al. 2017) and assessment of corruption levels in ambient air quality data (Brown 2005). Beyond these diverse fields, Benford's Law has increasingly found utility within the natural sciences, highlighting its broad analytical power.

For example, Sambridge et al. (2010) demonstrated the prevalence of Benford's Law across the natural sciences, even finding a near Benford-like distribution in global mean temperature anomalies. They further developed a Benford's Law-based statistical anomaly detection method, which was subsequently applied by Yang and Fu (2017) to identify regime shifts in precipitation data over America and China. However, Sambridge et al. (2010) themselves cautioned that such methods might simply be detecting increased orders of magnitude. Adding a layer of complexity to the application of Benford's Law, Kossovsky (2021) highlighted a fundamental consideration: inherent correlations between leading digits may complicate an assumption of their independence in statistical analyses. Furthermore, Joannes-Boyau et al. (2015) utilised Benford's Law to assess the homogeneity and quality of tropical cyclone path length datasets, revealing improved data reliability with increased observational records. More recently, Nakamura et al. (2024) employed Benford's Law to examine synthetic tropical cyclone precipitation data for deviations from a Benford-like distribution. While these diverse studies suggest the potential of Benford's Law within climate-related research, its robust and considered applicability to precipitation datasets remains an open area for investigation.

The remainder of this paper is structured as follows. Section 2 details a reproducible and robust methodology for applying Benford's Law to precipitation datasets, rigorously incorporating its underlying assumptions. An analysis in Section 3.1 then demonstrates that terrestrial global monthly precipitation consistently conforms to Benford's Law across diverse data sources. Furthermore, a regional case study in Section 3.2 elucidates critical limitations, detailing when and why Benford's Law is not applicable, thereby guiding its appropriate use in climate sciences. Finally, to showcase its practical utility for enhancing data quality pipelines, Section 4 demonstrates the detection of hypothetical data corruption within a precipitation dataset.

2 | Developing a Robust Benford's Law Methodology for Precipitation Data

The use of global precipitation datasets, whether derived from in situ observations such as rain gauges, remotely sensed through satellite platforms or reanalysis products, is ubiquitous throughout climate science. The reliability of these datasets is thus paramount for robust climate research (Sun et al. 2018). Here, we test the adherence of a range of precipitation data to Benford's Law and, subsequently, the law's efficacy as a quality control metric. We systematically assess this adherence across a spectrum of commonly used sources of precipitation data—including raw station measurements, gridded products and stochastically generated synthetic time series—to ascertain its broad applicability to precipitation. This analysis aims to demonstrate whether Benford's Law can serve as a robust, data-type-agnostic diagnostic tool for precipitation data integrity, or if its patterns can differentiate the inherent characteristics of observational datasets from model-generated or stochastic products. To achieve this, our study incorporates a diverse range of precipitation data types, each with its own characteristics and limitations.

Weather station observations are important point-scale precipitation datasets, useful for detecting climate change signals (Alexander et al. 2006), and with high enough spatial resolution for impact assessments and water resource management (Jiang et al. 2012; Sun et al. 2018). However, raw observations can suffer from systematic errors, missing values, incomplete spatial coverage and in several cases, may not be long enough for stringent impact assessment (Wilks and Wilby 1999; Viney and Bates 2004; Daly et al. 2007; Wilby et al. 2017). Stochastic weather generators, originally developed for agricultural and hydrological modelling purposes, can simulate long sequences of continuous and stationary data, thus addressing the temporal constraints associated with raw station data (Wilks and Wilby 1999).

For applications requiring homogeneous and complete spatial coverage, interpolating raw observations provides gridded observational-based datasets with global (though typically land-only) coverage, tackling issues with spatial sparsity and more readily facilitating comparison between climate model output and observations (Asadieh and Krakauer 2015). However, it is crucial to recognise that this interpolation process can introduce biases, systematic errors and artifacts, particularly in regions with sparse observational networks, which users must carefully consider despite the considerable efforts and rigorous quality control procedures within these products (Harris et al. 2020). Reanalysis datasets are an

additional source of global (land and sea) gridded precipitation, constructed by merging observations with models and providing data at a range of temporal resolutions. Though widely used as a proxy for observational precipitation, reanalysis datasets are known to have sources of bias induced from the physical modelling and assimilation procedures, and uncertainties between reanalysis datasets themselves may arise due to differences in assimilation schemes and model physics (Sun et al. 2018).

Collectively, these diverse data types ranging from raw observations to processed and synthetic products, provide invaluable resources for climate science. Characterising these data types is important for rigorously assessing Benford's Law's utility as a (potentially) data-type-agnostic quality control tool for precipitation. The specific data sources utilised in this study are detailed in the following section.

2.1 | Data Sources

Precipitation data are gathered from: the Climatic Research Unit Terrestrial Series raw ungridded input dataset, consisting of monthly precipitation observations from weather station rain gauges (CRU-TSRaw); its final gridded product (CRU-TS; Harris et al. 2022) the fifth-generation ECMWF atmospheric reanalysis (ERA5; Hersbach et al. 2023) and a stochastically generated, synthetic dataset from a weather generator whose parameters were diagnosed from ERA5 daily precipitation fields. All gridded datasets (i.e., CRU-TS, ERA5 and stochastically generated) are interpolated to and analysed on a common $2.5^{\circ} \times 2.5^{\circ}$ latitude-longitude spatial grid to ensure consistency across results. Key summary statistics for these datasets can be found in Table S1.

The Climatic Research Unit Terrestrial Series gridded product (CRU-TS) is a widely used precipitation dataset with global land-surface coverage (excluding Antarctica) (Harris et al. 2020). In CRU-TS, precipitation is interpolated from an extensive network of weather station observations, with sparse data regions relaxed to their 1961–1990 climatology, yielding complete global coverage with no missing values. The availability of both nonquality-controlled raw station data (CRU-TSRaw) and the final gridded product (CRU-TS) is a notable asset. This unique characteristic enables a direct comparison to analyse whether quality control procedures influence leading digit frequencies. For this research, version 4.06 was accessed, covering the period from January 1901 to December 2021 as monthly totals on a latitude-longitude grid.

ERA5 monthly total precipitation and stochastically generated data are derived from ERA5 daily precipitation values (Hersbach et al. 2020). These data span the period January 2000 to December 2020 on a latitude-longitude spatial grid. To construct the synthetic precipitation time series, a two-state, first-order Markov chain-gamma model was fitted to simulate daily precipitation occurrence and amount at each grid cell (Richardson and Wright 1984). These daily simulations and ERA5 products are aggregated to monthly totals for consistency with CRU data. We use a first-order Markov chain model to condition precipitation occurrence on the previous day's precipitation status (i.e., wet or dry). This has been identified as the most globally applicable Markov-chain order for simulating precipitation occurrence (Wilson Kemsley et al. 2021). More detailed information on this process can be found in Appendix A.

For a consistent comparison across all data sources, the main analysis utilises only land-based precipitation data. This approach facilitates direct comparison between the inherently land-covering CRU data and the reanalysis and stochastically generated products. To denote datasets containing exclusively land data points, the suffix '*-TS*' (Terrestrial Series) is appended; thus, these are referred to as ERA5-TS and Stochastic-TS.

2.2 | An Algorithm for Large Datasets

To ensure the appropriate application of Benford's Law, data must adhere to two key requirements: it must span multiple orders of magnitude and its underlying distribution must be skewed (Kossovsky 2021). Here we develop a four-step methodology[1]:

1. Data points are stripped of time and space dependence, and placed into one long list (i.e., flattened). Fill values (that indicate data gaps or grid cells with no observations, such as data over the oceans in CRU-TS) are removed and zero values excluded. The resulting dataset is denoted as $X(n)$, where $n \in \mathbb{R}^{+}$.
2. A Robust Measure of Orders of Magnitude (ROM) assessment is applied to ensure data spans multiple orders of magnitude, where:

$$ROM = \log_{10}\left(\frac{P_{99}}{P_1}\right) > 2.5,$$

   P_{99} and P_1 are the 99th and first percentile values of $X(n)$, respectively, following the work of Kossovsky (2021).
3. The adjusted Fisher-Pearson standardised moment coefficient (Doane and Seward 2011) is used to assess skewness. Datasets with 'skew score' larger than 1 are considered skewed.
4. Providing the conditions in Step 2 and 3 are validated, the lead digit frequency, F_d, is calculated for $d \in \{1, 2, 3, \ \ldots\ , 9\}$. A subset S_d of $X(n)$ is constructed containing all values of $X(n)$ with lead digit d. The frequency of lead digit occurrence, F_d, can now be calculated as:

$$F_d = \frac{\text{Size}\left(S_d\right)}{\text{Size}(X(n))} \cdot 100,$$

   where Size $\left(S_d\right)$ and Size $(X(n))$ denote the number of elements in each set.

It is expected that $F_d \approx P_d$ (recalling that P_d is the theoretical Benford's Law derived lead digit frequency, specified by Equation (1)) provided that $X(n)$ is skewed and spans many orders of magnitude.

2.3 | Expected Errors

For the purposes of this work, lead digit frequencies (calculated using the methodology in Section 2.2) are considered within a maximum expected error range, E_d, for each lead digit $d \in \{1,2,3, \ldots, 9\}$. If a calculated lead digit frequency falls considerably outside of the Benford's Law determined frequency plus/minus the expected error (i.e., $P_d + E_d < F_d$ or $P_d - E_d > F_d$), then this is viewed as evidence of the dataset not adhering to Benford's Law.

Following the work of Cong et al. (2019), maximum error ranges are calculated for the terrestrial data examined in this study. To confidently apply an error range analysis, each dataset must be fitted to a probability density function (PDF). Here, all terrestrial datasets exhibit reasonable PDF fits to an exponential distribution of the form PDF $(x) = \lambda e^{-\lambda x}$ (with rate parameter λ values of 0.012, 0.025, 0.03 and 0.03 for CRU-TSRaw, CRU-TS, ERA5-TS and Stochastic-TS respectively, see Figure 1). Cong et al. (2019) demonstrates that for an exponential distribution, the total error term in Benford's Law simplifies significantly and is bound by the maximum amplitude of the periodic fluctuation around the logarithmic Laplace spectrum of the digital indicator function (Cong et al. 2019; Sornette 1998) in base 10. These maximum error bounds are explicitly calculated for the exponential distribution in Cong et al. (2019) and are reproduced here in Table 1.

3 | Assessing Benford's Law Adherence in Precipitation Data

Having established a robust methodology for assessing Benford's Law adherence in Section 2, this section now presents its empirical application to various precipitation datasets. The primary objective is to determine the extent to which diverse precipitation data, across different types and spatio-temporal scales, conform to the predicted leading digit frequencies. This analysis serves to validate Benford's Law as a potential diagnostic tool for data integrity within hydroclimatology. We begin in Section 3.1 by analysing global terrestrial precipitation from a range of sources, investigating whether data origin or type influences adherence. Subsequently, Section 3.2 provides a detailed regional case study focusing on the United Kingdom and Ireland, which critically examines the geographical and temporal conditions under which Benford's Law may or may not apply.

3.1 | Global Terrestrial Precipitation

The methodology outlined in Section 2.2 is applied to the four terrestrial datasets from Section 2.1, with error analysis as described in Section 2.3. For each of our flattened datasets, we are left with sample sizes of 9,134,011 (CRU-TSRaw), 3,920,408 (CRU-TS), 859,952 (ERA5-TS) and 845,272 (Stochastic-TS). Results are presented in Figure 1, with numerical values included for the interested reader in S1. Strikingly, across all four of these diverse datasets comprising raw observations (CRU-TSRaw; Figure 1a,b), quality-controlled gridded products (CRU-TS; Figure 1c,d), reanalysis output (ERA5-TS; Figure 1e,f) and entirely synthetic simulations (Stochastic-TS; Figure 1g,h), the calculated lead digit frequencies consistently fall within the expected error bounds of Benford's Law (see Figure 1b,d,f,h). This widespread conformance provides compelling evidence that global monthly precipitation, when sampled over a broad terrestrial domain, inherently adheres to Benford's Law, regardless of its observational or simulated origin. Crucially, this adherence is observed because the datasets meet the requisite conditions of spanning multiple orders of magnitude and exhibiting a sufficiently skewed, continuously monotonic underlying distribution (as illustrated by their histograms and PDF fits in Figure 1a,c,e,g).

This strong adherence across varied data sources strongly suggests that the precise nature of the underlying data distribution, rather than the data's classification as 'natural', 'modelled' or 'synthetic', is the determinative factor for Benford's Law conformance. This empirical evidence provides robust support for theoretical arguments, such as those presented in Wang and Ma (2024), which emphasise the mathematical properties of a dataset's distribution as the fundamental driver of Benford's Law adherence. As discussed in Section 1, many applications of Benford's Law traditionally focus on 'natural' datasets (following the seminal work of Newcomb (1881)), often implicitly assuming their inherent properties lead to adherence. The results across these four datasets directly challenge this historical assumption, significantly expanding the utility of Benford's Law to include artificial and modelled data, provided the underlying assumptions are respected.

3.2 | Regional Terrestrial Precipitation: A Case Study of the United Kingdom and Ireland

As demonstrated in Section 3.1, global monthly total precipitation data consistently adheres to the predicted lead digit frequencies of Benford's Law, irrespective of data source or type. This finding is valuable for global-scale analyses and utilised directly in Section 4. However, it is prudent to note that numerous climatic applications, such as regional forecasting, simulation and data recovery initiatives (e.g., the Rainfall Rescue project outlined in Hawkins et al. (2023)), operate at subglobal spatial scales. Therefore, it is useful to assess the applicability of Benford's Law directly to regional precipitation datasets. The two fundamental prerequisites for the valid application of Benford's Law, as detailed in Section 2.2, are that data must span across multiple orders of magnitude and possess a skewed underlying distribution. While these conditions were met by the global datasets of Section 3.1, it is not immediately evident that they would hold true at a regional scale, where climatic variability may be constrained, limiting the orders of magnitude of observed precipitation rates.

To investigate the application of Benford's Law at a subglobal scale, we focus on monthly total precipitation data from ERA5 limited to land grid points over the United Kingdom and Ireland (denoted as ERA5UK&I-TS). By decreasing the spatial extent of our analysis, the number of samples in our

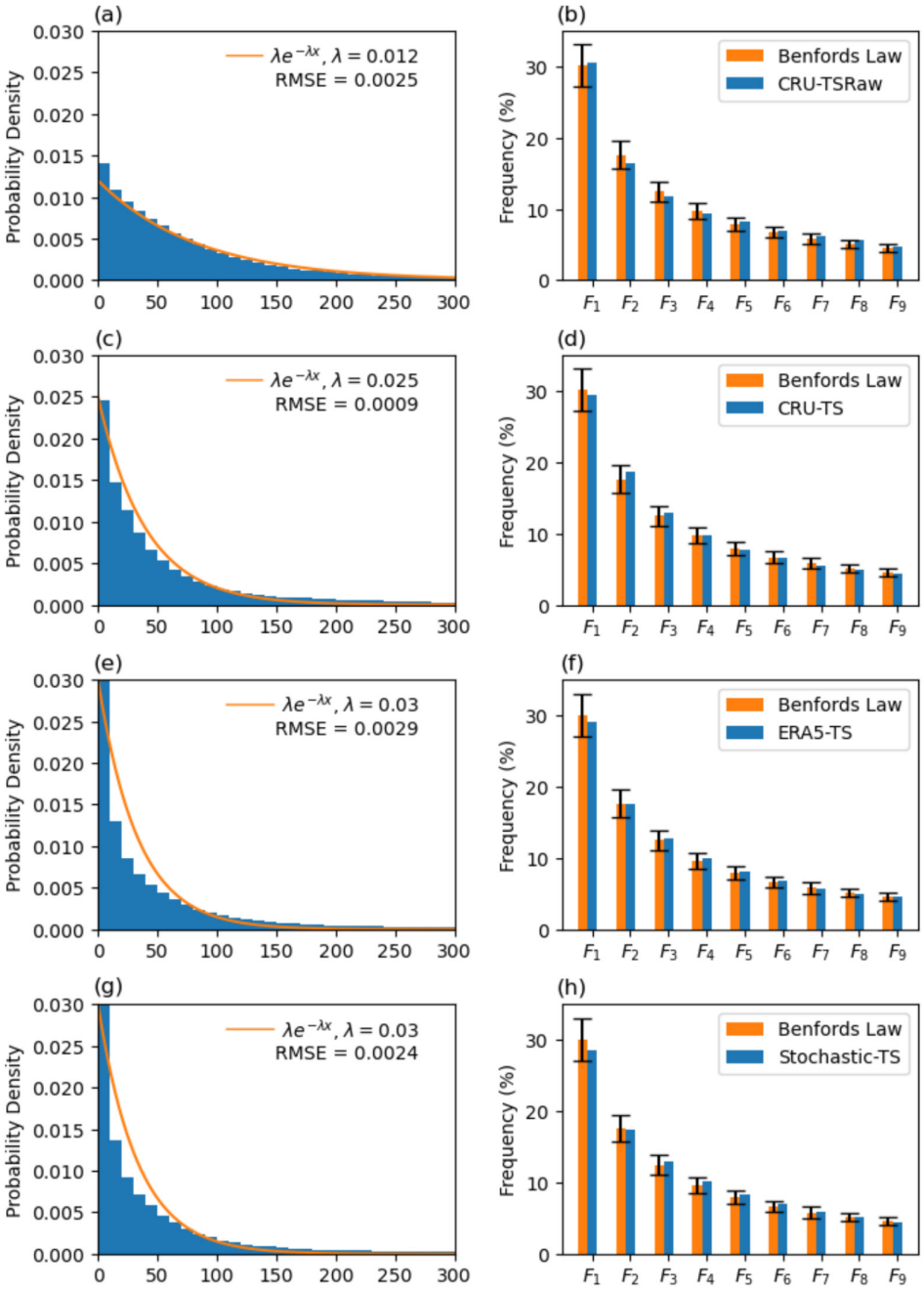

FIGURE 1 | Lead digit frequency analysis applied to raw station data (CRU-TSRaw, panels a and b), gridded observational data (CRU-TS, panels c and d), reanalysis data (ERA5-TS, panels e and f) and stochastically generated data (Stochastic-TS, panels g and h). All data are monthly totals of precipitation (mm) sampled over a global land domain. Panels a, c, e and g show histograms (blue bars) and corresponding PDF fits (orange lines) for the precipitation totals. The root mean squared errors (RMSEs) between the histograms and and PDF fits are denoted in each panel. Panels b, d, f, h show the observed lead digit frequencies (blue bars) with theoretical Benford's law predicted distributions (orange bars) and expected error intervals (black vertical lines). [Colour figure can be viewed at wileyonlinelibrary.com]

flattened dataset decreases; to increase the size (now equalling 4080), we consider ERA5UK&I-TS on a finer $1.25^{\circ} \times 1.25^{\circ}$ resolution (Hersbach et al. 2023). The United Kingdom and Ireland (hereafter UK&I) was selected for this case study due to its nonmonsoonal climate and relatively small spatial area, providing a stringent test for the orders of magnitude criterion. The methodology outlined in Section 2.2 was applied to ERA5UK&I-TS, with results presented in the top row of Figure 2. The algorithm immediately signalled nonconformance across both the ROM and skewness criteria, yielding values of 1.08 and 0.93, respectively. As illustrated in Figure 2a, the underlying distribution of ERA5UK&I-TS more closely resembles a Gaussian distribution, with characteristics that fundamentally undermine key assumptions upon which Benford's Law are predicated (Cong et al. 2019; Wang and Ma 2024), particularly the requirement for a continuously monotonic and skewed distribution.

This finding indicates that applying Benford's Law-based lead digit frequency analysis to assess the reliability of monthly total

UK&I precipitation is inappropriate and highlights the role of the algorithm's built-in validation checks as crucial safeguards. From a monthly perspective, the climatic variability within the UK&I is comparatively limited, resulting in a narrower range of precipitation values that do not span the requisite orders of magnitude. Consequently, this case study underscores the importance of thoroughly assessing the inherent characteristics of a dataset prior to the application of Benford's Law. It suggests that Benford's Law is unsuitable for geographical domains characterised by restricted climatic variability. For academic completeness, the observed lead digit frequencies for ERA5UK&I-TS are presented alongside the theoretical Benford's Law predictions in Figure 2b. This visual comparison clearly demonstrates the significant deviation from Benford's Law, precisely as would be anticipated given the dataset's underlying approximate Gaussian distribution.

Recognising the unsuitability of monthly totals for Benford's Law analysis in this regional context, we shift our investigation toward *daily* precipitation totals for the UK&I, derived from ERA5 daily precipitation totals (see Section 2.1). This new dataset, with size 121,957, denoted by ERA5UK&I-DTS (ERA5 United Kingdom and Ireland Daily Terrestrial Series), allows us to explore whether a finer temporal resolution would introduce the necessary data characteristics for Benford's Law adherence. Upon applying the methodology from Section 2.2 to ERA5UK&I-DTS, the dataset has a ROM value of 4.47 and a skew score of 2.22. Both values exceed their thresholds indicating that daily precipitation over the UK&I has strong potential for Benford's Law adherence. The histogram and fitted PDF for ERA5UK&I-DTS, presented in Figure 2c, clearly illustrate its exponential-like distribution enabling error analysis (see Section 2.3) to be applied. Figure 2d displays the calculated lead digit frequencies for ERA5UK&I-DTS which exhibit alignment with Benford's Law, all remaining within the expected error bounds for each digit. This contrast with the monthly totals underscores that while monthly aggregation in the UK&I diminishes necessary variability, daily precipitation totals retain the intrinsic properties that enable them to conform to Benford's Law. Hence, Benford's Law could be used as a quality control metric for daily ERA5 UK&I precipitation, leading to confidence that it could also be appropriate for other regional daily datasets with similar or more varied climatic behaviour.

TABLE 1 | Maximum magnitude of expected errors (percent) for each lead digit frequency. Only applicable for datasets with PDF fit $\lambda e^{-\lambda x}$ in a base 10 counting system. Values from Cong et al. (2019).

E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9
2.97	1.94	1.41	1.11	0.91	0.76	0.77	0.59	0.53

4 | Assessing Corruption Detection

Having confirmed global monthly total precipitations' adherence to Benford Law (Section 3.1), we now demonstrate a practical application of detecting data corruption via an artificial experiment utilising CRU-TS. In Section 3.2, PDF fits were

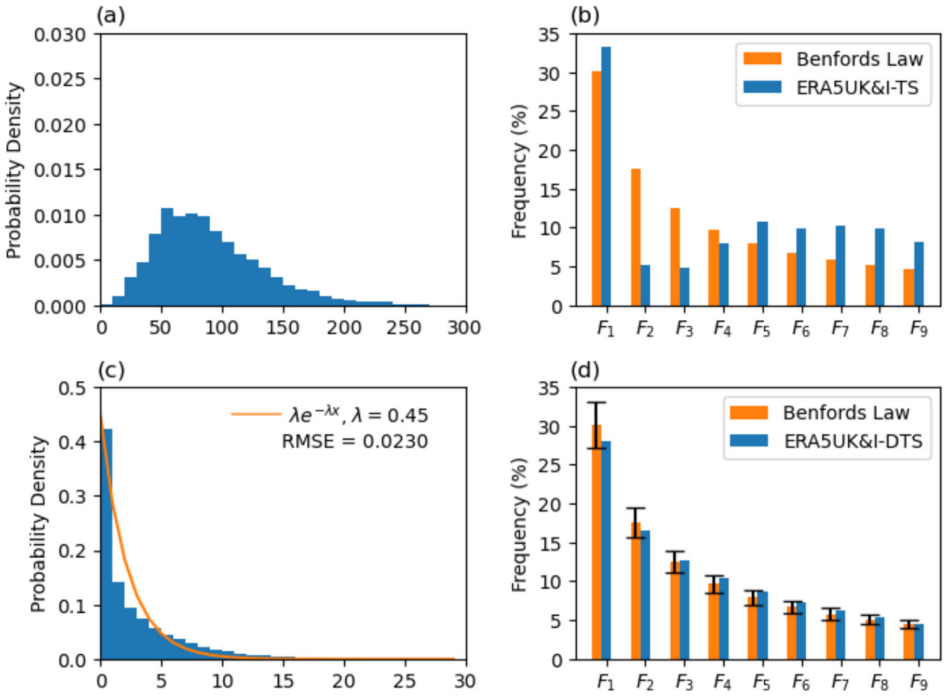

FIGURE 2 | Lead digit frequency analysis applied to only the United Kingdom & Ireland ERA5 monthly total (ERA5UK&I-TS, Row 1) and daily total (ERA5UK&I-DTS, Row 2) precipitation. Panels a and c show histograms (blue bars) for the monthly and daily precipitation, respectively, and a corresponding PDF fit for the daily precipitation only (orange line, panel c). Panels b and d show the corresponding lead digit frequencies (blue bars) for monthly (b) and daily (d) data with theoretical Benford's law predicted distributions (orange bars) and expected error intervals for the daily data (panel d, black vertical lines). [Colour figure can be viewed at wileyonlinelibrary.com]

shown to be useful in understanding why lead digit frequencies for ERA5UK&I-TS and ERA5UK&I-DTS either did or did not conform to Benford's Law. This raises the question: is Benford's Law analysis more beneficial than manually examining histograms or overall dataset characteristics? The key advantage lies in its ability to identify subtle data quality issues that may not be apparent from broad data metrics like skewness or orders of magnitude, and that can be time-consuming and subjective to detect through manual visual inspection of PDFs.

In this artificial experiment, data points within CRU-TS are randomly selected and replaced with 'corrupted' values at varying rates, first ranging from 1% to 20% of the total data entries. These corrupted values take four distinct forms:

a. Incorrect ocean fill values: Data were corrupted by placing an incorrect ocean fill value of 9999. This specific simulation was run once to evaluate a distinct type of data error; fill values of $\pm$ 9999 are commonplace in environmental datasets.

b. Mean substitution: A subset of data was corrupted by replacing selected values with the mean of the subset. This approach avoided altering the overall order of magnitude or the mean of the dataset. This specific scenario was also run once.

c. Random uniform corruption: Subsets of precipitation data were replaced with randomly selected integers from a discrete uniform distribution (bound by the original dataset's minimum and maximum). This experiment was repeated 11 times with consistent outcomes.

d. Random Gaussian corruption: Data points were replaced with random values sampled from a Gaussian distribution with the same mean (59.09 mm) and standard deviation as the original data (78.39 mm). We repeated this corruption a total of 11 times with consistent results.

Results are presented in Figure 3, where grey cells indicate data corruption detected by the algorithm (i.e., $P_d + E_d < F_d$ or $P_d - E_d > F_d$). Benford's Law proved particularly effective in detecting incorrectly filled ocean values (Figure 3a). At a 1% corruption rate, the lead digit frequency of 9 (LD9) exceeds the expected error range. At 8% corruption, the LD1 is also highlighted. Mean value corruption, mimicking 'replace with mean' approaches for infilling missing precipitation data (Hırca and Eryılmaz Türkkan 2024), was also detected at low corruption rates (Figure 3b). For example, LD5 alerted at a 2% corruption rate, followed by LD9 at 7%. By 10% corruption, over half of the lead-digit frequencies lie outside expected ranges.

Interestingly, the method showed less sensitivity to randomly introduced errors. This is likely due to random substitutions spreading corrupted values more evenly across the digits, so the lead-digit frequency distribution is not as strongly distorted. For the uniform-discrete corruption (Figure 3c)—intended to mimic random observational or satellite-retrieval errors (Ciach 2003; AghaKouchak et al. 2012; Sun et al. 2018)—corruption is only detected at rates exceeding 15%, and in these cases, the first leading digit (LD1) is the only one that falls outside the expected range. This is most likely due to the fact that one is the most common lead digit, and therefore the most sensitive to random swaps. We also find that our method is even more insensitive to Gaussian replacement, where lead-digit frequencies remained within expected bounds up to 20% corruption.

Though direct Gaussian replacement does not necessarily mimic a common source of precipitation error, we previously showed that monthly ERA5-UK&I-TS precipitation failed to meet the assumption of Benford's Law due to its near-Gaussian distribution, whereas the more skewed daily precipitation conformed (Figure 2). Because Benford's Law requires skewed data, Gaussian replacement *must* eventually be detected and thus this experiment examines a transition from a conforming (i.e., original CRU-TS) to a nonconforming dataset. We find that Gaussian replacement only breaks adherence after approximately 40% of the data is replaced. In contrast, uniform-discrete corruption (which spreads leading digits more evenly), and single-value substitutions (which impose strong bias) disrupted the Benford pattern at much lower levels. Gaussian corruption therefore alters Benford behavior only once a substantial fraction of the dataset has been replaced.

This artificial experiment highlights the utility of Benford's Law in climate data quality assessment, especially when detecting specific single-value corruptions, which are common in

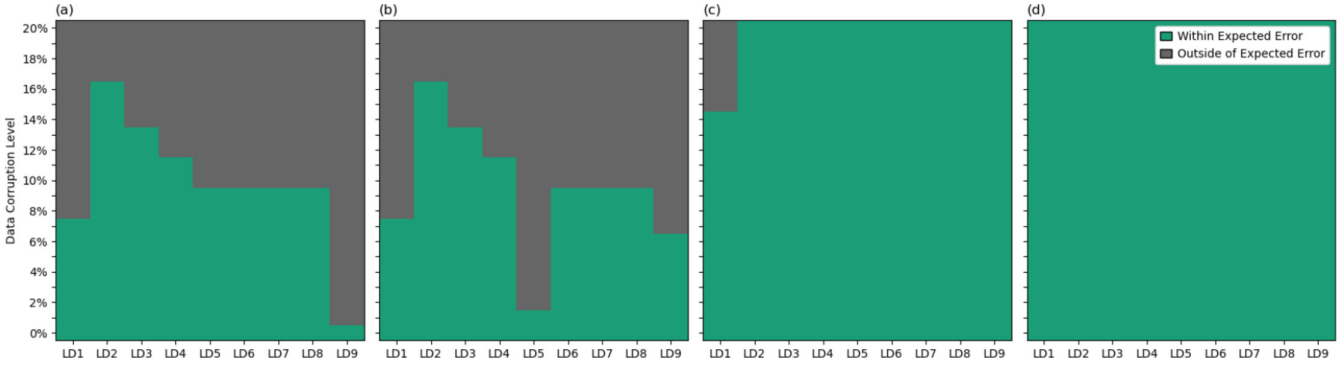

FIGURE 3 | CRU-TS corruption experiments for four different corruption types: (a) erroneous ocean fill value, (b) mean value filled, (c) random value change and (d) Gaussian corruption. The y-axis represents corruption rate as a total dataset percentage. Each column on the x-axis represents a lead digit, for example, LD1 for 'lead digit 1'. If the calculated lead digit frequency falls within/outside of the expected error range then the cell is green/grey, respectively. [Colour figure can be viewed at wileyonlinelibrary.com]

environmental datasets. We have shown that the method is less effective in detecting purely random corruption (such as uniform discrete or Gaussian) within data.

5 | Conclusions and Discussion

This study demonstrates the application of Benford's Law as a novel and potentially powerful tool for evaluating the integrity of precipitation datasets. The semi-autonomous methodology offers a practical and efficient approach to precipitation data validation and could be used to enhance data reliability. However, it is crucial to recognise that this method should be integrated as part of a comprehensive data quality pipeline, rather than used as a stand-alone metric, especially if data are vulnerable to randomised corruption. The accuracy of expected error ranges is contingent upon a thorough understanding of the parent dataset's underlying PDF, which should be carefully examined and validated prior to implementing Benford's Law.

A key outcome of this work is the compelling evidence that the data type itself does not inherently dictate the applicability or utility of Benford's Law to precipitation data. As demonstrated in Section 3.1, global monthly precipitation from diverse sources—ranging from raw station observations and quality-controlled gridded products to reanalysis output and entirely synthetic simulations—consistently conformed to Benford's Law. This widespread adherence, observed across different data generation methods, strongly supports the theoretical advancements that emphasise a dataset's inherent mathematical distribution, rather than its 'natural' or 'artificial' origin, as the primary determinant for Benford's Law conformance (Kossovsky 2021; Wang and Ma 2024). This finding expands the potential application of Benford's Law beyond its traditional scope, positioning it as a versatile tool for integrity assessment across a broader spectrum of datasets.

The UK&I-based case study (Section 3.2) provided crucial insights into the geographical and temporal granularity considerations for Benford's Law application. While monthly precipitation totals for the UK&I failed to adhere to Benford's Law, this was primarily due to a constrained range of values and a near-Gaussian distribution; an adjustment to daily precipitation totals successfully restored conformance. This outcome underscores that while regional climatic variability can indeed limit the applicability of Benford's Law at coarser temporal aggregations, moving to a finer granularity (e.g., daily totals) can reintroduce the necessary data characteristics (wider range, stronger skew) for its valid application. This highlights the importance of the initial diagnostic checks within our methodology, ensuring that the law is applied appropriately and guiding researchers in selecting suitable spatio-temporal scales for analysis.

Beyond assessing adherence, this research highlights Benford's Law's tangible impact on enhancing data quality pipelines, particularly for large and complex datasets. As discussed in Section 4, the method's ability to detect relatively low levels of corruption (at a rate of 1% and 2% for some common corruption types) positions it as a valuable complementary diagnostic tool. This is increasingly vital in an era of large data volumes, where automation and robust validation are paramount for applications ranging from climate modelling and forecasting to machine learning. The methodology developed here offers a deployable and scalable solution for routine data integrity checks in environmental sciences, contributing directly to the reliability of foundational datasets.

Looking ahead, several promising avenues for future research emerge from these findings. Expanding the analysis to global daily precipitation totals would build upon the success observed with the United Kingdom and Ireland daily data, providing a comprehensive assessment of its large-scale applicability. Further regional case studies are warranted to more fully characterise the geographical boundaries and conditions under which Benford's Law applies to precipitation. For example, the work of Yang and Fu (2017) demonstrates promise that Benford's Law could be successfully applied to Chinese and/or American precipitation datasets as they use a statistical technique closely linked to Benford's Law in detecting regime shifts over these regions. Investigating the conformance of oceanic precipitation data to Benford's Law would also be a crucial next step, given its distinct characteristics and measurement challenges compared to terrestrial precipitation. Finally, applying Benford's Law to a wider range of non-Gaussian climate data fields, potentially cloud datasets, could extend the use of these methods beyond precipitation data alone.

Author Contributions

Amee Gollop: conceptualization, investigation, methodology, validation, software, data curation, project administration, formal analysis, visualization, writing – review and editing, writing – original draft. **Sarah Wilson Kemsley:** formal analysis, project administration, data curation, validation, investigation, writing – review and editing. **Tim Osborn:** writing – review and editing, investigation, funding acquisition. **Manoj Joshi:** writing – review and editing, investigation, funding acquisition. **David Stevens:** writing – review and editing, investigation, funding acquisition. **Ian Harris:** methodology, data curation, supervision.

Acknowledgements

This research was made possible by the unwavering commitment of countless researchers, academic institutions and research centres to the principles of open science, exemplified through their provision of freely available open-access data. We specifically acknowledge the efforts behind the Coupled Model Intercomparison Project (CMIP), the European Centre for Medium-Range Weather Forecasts (ECMWF), the Climatic Research Unit (CRU) and the citizen science initiatives like the Rainfall Rescue UK project, all of which provide invaluable resources to the climate research community and were used in support of this body of work. A.G., M.J. and D.S. received funding from the UK Natural Environmental Research Council (NERC), grant number NE/W005239/1. A.G. and I.H. received funding from the UK National Centre for Atmospheric Sciences (NCAS). T.O. received funding from the UK Natural Environment Research Council (NERC), grant number NE/S015582/1. S.W.K. was funded by the UK Natural Environment Research Council (NERC), grant number NE/V012045/1. Analysis in support of this article was carried out on the High Performance Computing Cluster supported by the Research and Specialist Computing Support service at the University of East Anglia.

A.G. wishes to thank the motivational discussion and introduction to the concept of Benford's Law by Matthew J. Gollop. She also gratefully acknowledges the UK National Centre for Atmospheric Sciences

(NCAS), a NERC collaborative centre, for the provision of funding to the Climatic Research Unit, which enabled the completion of this work.

All authors would like to express their sincere gratitude to the two anonymous reviewers and to Prof. Athanassios Argiriou for their time in reviewing this manuscript, as well as for their insightful comments and constructive suggestions, which have significantly improved the quality and clarity of the paper.

Funding

This study was supported by the UK Natural Environment Research Council (NERC) (NE/W005239/1, NE/V012045/1, NE/S015582/1) and the UK National Centre for Atmospheric Science (NCAS).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are openly available in a Zenodo repository at https://doi.org/10.5281/zenodo.15731168.

Endnotes

[1]This methodology has been developed to retain generality and is thus readily applicable to many large datasets that meet the required conditions.

References

AghaKouchak, A., A. Mehran, H. Norouzi, and A. Behrangi. 2012. "Systematic and Random Error Components in Satellite Precipitation Data Sets." *Geophysical Research Letters* 39: L09406.

Alexander, L. V., X. Zhang, T. C. Peterson, et al. 2006. "Global Observed Changes in Daily Climate Extremes of Temperature and Precipitation." *Journal of Geophysical Research: Atmospheres* 111: D05109.

Asadieh, B., and N. Y. Krakauer. 2015. "Global Trends in Extreme Precipitation: Climate Models Versus Observations." *Hydrology and Earth System Sciences* 19: 877–891.

Benford, F. 1938. "The Law of Anomalous Numbers." *Proceedings of the American Philosophical Society* 78: 551–572.

Bosilovich, M. G., J. Kennedy, D. Dee, R. Allan, and A. O'Neill. 2013. "On the Rerocessing and Reanalysis of Observations for Climate." In *Climate Science for Serving Society: Research, Modeling and Prediction Priorities*, 51–71. Springer Dordrecht.

Brönnimann, S., R. Allan, C. Atkinson, et al. 2018. "Observations for Reanalyses." *Bulletin of the American Meteorological Society* 99: 1851–1866.

Brown, R. J. 2005. "Benford's Law and the Screening of Analytical Data: The Case of Pollutant Concentrations in Ambient Air." *Analyst* 130: 1280–1285.

Ciach, G. J. 2003. "Local Random Errors in Tipping-Bucket Rain Gauge Measurements."

Cong, Y., C. Li, and B.-Q. Ma. 2019. "First Digit Law From Laplace Transform." *Physics Letters A* 383: 1836–1844.

Daly, C., W. P. Gibson, G. H. Taylor, M. K. Doggett, and J. I. Smith. 2007. "Observer Bias in Daily Precipitation Measurements at United States Cooperative Network Stations." *Bulletin of the American Meteorological Society* 88: 899–912.

Doane, D. P., and L. E. Seward. 2011. "Measuring Skewness: A Forgotten Statistic?" *Journal of Statistics Education* 19: 1–18.

Grammatikos, T., and N. Papanikolaou. 2021. "Applying Benford's Law to Detect Accounting Data Manipulation in the Banking Industry." *Journal of Financial Services Research* 59: 115–142. https://doi.org/10.1007/s10693-020-00334-9.

Harris, I., P. Jones, and T. Osborn. 2022. "Cru ts4.06: Climatic Research Unit (CRU) Time-Series (TS) Version 4.06 of High-Resolution Gridded Data of Month-by-Month Variation in Climate." https://catalogue.ceda.ac.uk/uuid/e0b4e1e56c1c4460b796073a31366980/.

Harris, I., T. Osborn, P. Jones, and D. Lister. 2020. "Version 4 of the Cru Ts Monthly High-Resolution Gridded Multivariate Climate Dataset." *Scientific Data* 7.

Hawkins, E., S. Burt, M. McCarthy, et al. 2023. "Millions of Historical Monthly Rainfall Observations Taken in the UK and Ireland Rescued by Citizen Scientists." *GeoScientific Data Journal* 10: 246–261.

Hersbach, H., B. Bell, P. Berrisford, et al. 2020. "The era5 Global Reanalysis." *Quarterly Journal of the Royal Meteorological Society* 146: 1999–2049.

Hersbach, H., B. Bell, P. Berrisford, et al. 2023. "Era5 Hourly Data on Single Levels From 1940 to Present." https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels.

Hill, T. P. 1995. "A Statistical Derivation of the Significant-Digit Law." *Statistical Science* 10: 354–363.

Hırca, T., and G. Eryılmaz Türkkan. 2024. "Assessment of Different Methods for Estimation of Missing Rainfall Data." *Water Resources Management* 38: 5945–5972.

Huntingford, C., E. S. Jeffers, M. B. Bonsall, H. M. Christensen, T. Lees, and H. Yang. 2019. "Machine Learning and Artificial Intelligence to Aid Climate Change Research and Preparedness." *Environmental Research Letters* 14: 124007.

Jiang, S., L. Ren, Y. Hong, et al. 2012. "Comprehensive Evaluation of Multi-Satellite Precipitation Products With a Dense Rain Gauge Network and Optimally Merging Their Simulated Hydrological Flows Using the Bayesian Model Averaging Method." *Journal of Hydrology* 452–453: 213–225.

Joannes-Boyau, R., T. Bodin, A. Scheffers, M. Sambridge, and S. Matthias-May. 2015. "Using Benford's Law to Investigate Natural Hazard Dataset Homogeneity." *Scientific Reports* 5: 12046.

Kaltenborn, J., C. Lange, V. Ramesh, et al. 2023. "Climateset: A Large-Scale Climate Model Dataset for Machine Learning." *Advances in Neural Information Processing Systems* 36: 21757–21792.

Kossovsky, A. E. 2021. "On the Mistaken Use of the Chi-Square Test in Benford's Law." *Stat* 4: 419–453.

Nakamura, J., U. Lall, Y. Kushnir, and P. A. Harr. 2024. "A Saturated Stochastic Simulator: Synthetic Us Gulf Coast Tropical Cyclone Precipitation Fields." *Natural Hazards* 120: 1295–1318.

Newcomb, S. 1881. "Note on the Frequency of Use of Difference Digits in Natural Numbers." *American Journal of Mathematics* 4: 39–40.

Nigrini, M., and S. Miller. 2007. "Benford's Law Applied to Hydrology Data—Results and Relevance to Other Geophysical Data." *Mathematical Geology* 39: 469–490. https://doi.org/10.1007/s11004-007-9109-5.

Richardson, C., and D. Wright. 1984. "Wgen: A Model for Generating Daily Weather Variables." United States Department of Agricultural Research Service, ARS-8.

Sambridge, M., H. Tkalčić, and A. Jackson. 2010. "Benford's Law in the Natural Sciences." *Geophysical Research Letters* 37: L22301.

Sheng, G., T. Li, Q. Su, b. Chen, and Y. Tang. 2017. "Detection of Content-Aware Image Resizing Based on Benford's Law." *Soft Computing* 21: 5693–5701. https://doi.org/10.1007/s00500-016-2146-6.

Sornette, D. 1998. "Discrete-Scale Invariance and Complex Dimensions." *Physics Reports* 297: 239–270.

Sun, Q., C. Miao, Q. Duan, H. Ashouri, S. Sorooshian, and K.-L. Hsu. 2018. "A Review of Global Precipitation Data Sets: Data Sources, Estimation, and Intercomparisons." *Reviews of Geophysics* 56: 79–107.

Thom, H. C. 1958. "A Note on the Gamma Distribution." *Monthly Weather Review* 86: 117–122.

Viney, N. R., and B. C. Bates. 2004. "It Never Rains on Sunday: The Prevalence and Implications of Untagged Multi-Day Rainfall Accumulations in the Australian High Quality Data Set." *International Journal of Climatology* 24: 1171–1192.

Wang, L., and B. Q. Ma. 2024. "A Concise Proof of Benford's Law." *Fundamental Research* 4: 841–844.

Wilby, R. L., N. J. Clifford, P. De Luca, et al. 2017. "The 'Dirty Dozen' of Freshwater Science: Detecting Then Reconciling Hydrological Data Biases and Errors." *WIREs Water* 4: e1209.

Wilby, R. L., and T. M. Wigley. 1997. "Downscaling General Circulation Model Output: A Review of Methods and Limitations." *Progress in Physical Geography* 21: 530–548.

Wilks, D. S., and R. L. Wilby. 1999. "The Weather Generation Game: A Review of Stochastic Weather Models." *Progress in Physical Geography: Earth and Environment* 23: 329–357.

Wilson Kemsley, S., T. J. Osborn, S. R. Dorling, C. Wallace, and J. Parker. 2021. "Selecting Markov Chain Orders for Generating Daily Precipitation Series Across Different köppen Climate Regimes." *International Journal of Climatology* 41: 6223–6237.

Yang, L., and Z. Fu. 2017. "Out-Phased Decadal Precipitation Regime Shift in China and the United States." *Theoretical and Applied Climatology* 130: 535–544.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Appendix S1:** joc70221-sup-0001-Appendix.pdf.