Point-by-Point Response to Reviewer #1

Dear Reviewer #1,

We sincerely appreciate the time and effort you have devoted to reviewing our manuscript submitted to *Earth System Science Data* (ESSD). Your highly constructive comments and insightful suggestions have been invaluable in improving the quality, clarity, and scientific rigor of our paper.

Before addressing your specific point-by-point comments, we would like to provide a general response regarding a central theme raised across the reviews: the precise positioning, scientific significance, and intended applications of the GLBD-FED dataset, particularly in the context of existing high-quality datasets like GHCN-Daily.

General Response: Positioning and Scientific Significance of GLBD-FED

Prompted by the highly constructive feedback from the reviewers, we recognized that our original manuscript failed to adequately distinguish GLBD-FED from retrospectively homogenized climate datasets. To completely resolve this ambiguity and explicitly state the irreplaceable value of our dataset, we have completely rewritten the Introduction and added critical clarifications to the Discussion section.

We have reframed the scientific significance of GLBD-FED around two core pillars:

1. First-Hand Near-Real-Time Data vs. Homogenized Benchmarks

High-quality daily temperature datasets generally fall into two distinct tiers: homogenized benchmark datasets (e.g., GHCNd, BEST) designed for long-term decadal climate change detection, and near-real-time, first-hand datasets designed for rapid synoptic monitoring. While benchmark datasets are essential for climatology, they involve significant latency and often rely on retrospective collection. Conversely, legacy real-time datasets (like GSOD) often exhibit temporal aggregation artifacts. GLBD-FED is positioned at the foundational tier. By providing a structurally sound, temporally aligned "first estimate" directly from raw sub-daily synoptic reports, it meets the immediate need for rapid extreme weather monitoring while serving as the foundational raw material for future benchmark homogenization.

2. Strict Global Methodological Uniformity for NWP Verification (Solving the TOB Issue)

Despite the existence of high-quality regional datasets, a critical gap remains: strict temporal and methodological uniformity. National and regional benchmark datasets frequently employ diverse definitions of a "daily" period (e.g., varying local morning observation times versus midnight-to-midnight local time). This lack of standardization introduces well-documented Time of Observation Biases (TOB; Karl et al., 1986). Combining these localized datasets for global monitoring inevitably creates artificial discontinuities ("seams") across national borders.

The core scientific significance of GLBD-FED lies in its ability to eliminate these methodological borders. By applying a single, unified algorithmic framework globally, GLBD-FED enforces a universal physical 24-hour window (e.g., 0000 to 2400 UTC). This

strict temporal alignment is irreplaceable for verifying global Numerical Weather Prediction (NWP) model outputs. Modern NWP models output daily forecast summaries based on standardized UTC cycles; validating these outputs against heterogeneous regional datasets introduces severe temporal mismatch errors (Haiden et al., 2018). By aligning with WMO synoptic standards (WMO, 2017), GLBD-FED provides a seamless, time-aligned ground truth, ensuring that massive-scale synoptic weather systems are evaluated under the exact same global temporal framework.

(Note: These clarifications have been extensively integrated into the newly rewritten Introduction and Discussion sections. Detailed references, including Karl et al. (1986) and Haiden et al. (2018), have been formally added to the revised manuscript.)

Specific Responses to Reviewer #1

Reviewer #1 Comment 1: *This is an interesting paper describing a data set of potential significant value. I think it has the makings of a publishable paper but needs some work before getting to that stage. Although the authors were probably not aware of this when the dataset was being developed, it is unfortunate that both the ISD and GSOD datasets were retired in 2025 (ISD has been replaced by GHCN-Hourly, the replacement for GSOD is not yet online). As such this dataset will not be able to be updated in its current form – this is worth acknowledging, I think.*

Author's Response 1: We thank the reviewer for the encouraging assessment and the insight regarding the transition of NOAA's data services. Regarding the intended applications of our dataset, please refer to our "General Response" at the beginning of this document. Based on your feedback, we have rewritten the Introduction to emphasize GLBD-FED's role in providing a globally uniform 24-hour window that mitigates Time of Observation Bias (TOB), offering a structurally sound baseline for verifying Numerical Weather Prediction (NWP) outputs.

We agree that the retirement of the ISD archive must be transparently acknowledged. We evaluated its successor, NOAA's GHCN-Hourly (GHCNh), for near-real-time updates. However, GHCNh currently lacks the explicit sub-daily extreme reports (e.g., 12h/24h summaries) required for our temporal reconstruction algorithm. Consequently, real-time operationalization is temporarily paused. Nevertheless, the 1981-2024 GLBD-FED archive provides a temporally aligned 44-year historical baseline of first-hand data. We have explicitly acknowledged this limitation and detailed the implications of the GHCNh transition in the newly added "Dataset Positioning and Current Status" subsection within the Discussion section.

Changes in Manuscript:

> "**Dataset Positioning and Current Status**
> It is important to emphasize that GLBD-FED is fundamentally designed as a near-real-time, first-hand operational product rather than a

retrospectively homogenized benchmark dataset. Its primary utility lies in rapid, temporally accurate evaluations of regional synoptic weather events and validating numerical weather prediction models. Caution should be exercised if applying it directly to long-term decadal climate trend detection without further statistical homogenization.

Furthermore, while the GLBD-FED processing framework was built for continuous near-real-time updates, its current operationalization is constrained by upstream data source transitions. Following the recent retirement of the ISD archive, we evaluated its successor, NOAA's GHCN-Hourly (GHCNh). Our assessment revealed that the meteorological elements contained in GHCNh have been significantly reduced, lacking the specific sub-daily extreme reports necessary to robustly support our daily *Tmax* and *Tmin* reconstruction methodology. Therefore, while real-time streaming is currently paused, the completed 1981-2024 GLBD-FED archive stands as a highly valuable, temporally aligned 44-year historical first-hand baseline for the global meteorological community."

Reviewer #1 Comment 2: *The paper is lacking clear information about what the purpose of such a dataset might be. The most obvious application of a daily dataset is to support assessment of extremes – that is a link which should be given more prominence (at present there is a brief reference to extreme events research at L297-299). Also worth giving more prominence is that there is not currently, to my knowledge, a global in situ daily temperature dataset other than GSOD – the examples of daily and sub-daily products given at L55-56 are all precipitation-only (this should be stated).*

Author's Response 2: We agree that the original manuscript lacked a clear articulation of the dataset's purpose and included confusing precipitation-only references. We have restructured the Introduction accordingly.

(1) **Addressing Data Scarcity:** We removed the precipitation references and now explicitly highlight the scarcity of purely *in situ* daily temperature datasets compared to the abundance of gridded/satellite products, positioning GLBD-FED as a necessary addition to a landscape where GSOD was previously the primary baseline.

(2) **The Purpose:** We have given greater prominence to extreme weather assessment. However, to maintain scientific rigor, we framed this as supporting rapid, synoptic-scale extreme weather monitoring rather than long-term decadal climate change detection (which strictly requires retrospectively homogenized benchmarks like GHCNd).

(3) **Methodological Uniformity:** We explicitly introduced GLBD-FED's value for validating global NWP model outputs. By strictly aligning extremes to a standardized physical 24-hour window, it provides a methodologically uniform ground truth across national borders.

Changes in Manuscript:

While numerous global observational datasets exist to support climate research, there is a pronounced scarcity of daily global temperature datasets based purely on in situ station data. Historically, the development of global observational products has been exceptionally robust for precipitation, flourishing through both dense in situ gauge-based gridded analyses (e.g., GPCC; Becker et al., 2013) and multi-source merged products incorporating satellite estimates (e.g., MSWEP; Beck et al., 2019). In contrast, the available landscape for global daily temperature is much narrower. Existing prominent temperature datasets, such as GHCN-Daily (Menne et al., 2012), Berkeley Earth (BEST; Rohde et al., 2013), and HadGHCND (Caesar et al., 2006), are primarily designed either as retrospectively homogenized benchmark networks or as spatially interpolated gridded products optimized for long-term climate trend analysis. This scarcity leaves a critical gap for a purely station-based, high-fidelity daily temperature dataset that can leverage the improved global accessibility of sub-daily observations.

Furthermore, despite the existence of high-quality, long-term historical datasets, a critical gap remains in the global daily temperature data landscape: strict temporal and methodological uniformity. National and regional benchmark datasets frequently employ diverse definitions of a "daily" period—such as varying local morning observation times versus midnight-to-midnight local time—and utilize disparate temporal aggregation algorithms. This lack of standardization introduces well-documented Time of Observation Biases (TOB; Karl et al., 1986). When researchers attempt to combine these localized high-quality datasets for global monitoring, these methodological differences inevitably create artificial discontinuities ("seams") across national borders.

The core scientific significance of GLBD-FED lies in its ability to eliminate these methodological borders. By applying a single, unified algorithmic framework directly to first-hand sub-daily synoptic reports across all global regions simultaneously, GLBD-FED enforces a universal physical 24-hour window (e.g., 0000 to 2400 UTC). This strict temporal alignment is irreplaceable for advanced meteorological applications, particularly in the verification of global Numerical Weather Prediction (NWP) model outputs. Modern NWP models output daily forecast summaries based on standardized UTC cycles; validating these outputs against heterogeneous regional datasets introduces severe temporal mismatch errors (Haiden et al., 2018). By aligning with the World Meteorological Organization's standard for synoptic uniformity (WMO, 2017), GLBD-FED provides a seamless, time-aligned "first estimate" ground truth, ensuring that massive-scale synoptic weather systems are evaluated under the exact same global temporal framework.

References to be added to the bibliography

Beck, H. E., et al. (2019). MSWEP V2 global 3-hourly 0.1° precipitation: methodology and quantitative assessment. Bulletin of the American Meteorological Society, 100(3), 473-500.

Becker, A., et al. (2013). A description of the global land-surface precipitation data products of the Global Precipitation Climatology Centre with sample applications including centennial (trend) analysis from 1901–present. Earth System Science Data, 5(1), 71-99.

Caesar, J., et al. (2006). Large-scale changes in observed daily maximum and minimum temperatures: Creation and analysis of a new gridded data set. Journal of Geophysical Research: Atmospheres, 111(D5).

Haiden, T., et al. (2018). Evaluation of ECMWF forecasts, including the 2018 upgrade. ECMWF Technical Memorandum No. 831.

Karl, T. R., et al. (1986). A model to estimate the time of observation bias associated with monthly mean maximum, minimum and mean temperatures for the United States. Journal of Climate and Applied Meteorology, 25(2), 145-160.

Menne, M. J., et al. (2012). Global Historical Climatology Network-Daily (GHCN-Daily), Version 3. NOAA National Climatic Data Center.

Rohde, R., et al. (2013). A new estimate of the average Earth surface land temperature spanning 1753 to 2011. Geomics, 1.

World Meteorological Organization (WMO). (2017). Manual on the Global Observing System (WMO-No. 544). Geneva, Switzerland.

Reviewer #1 Comment 3: *Section 4.2.3 discusses reaggregation in some detail. One scenario which isn't covered here is what is done if there is a single 24-hour max/min which is reported at a time other than 0000 UTC – what is done (if anything) to convert this to a 0000-2400 equivalent?*

Author's Response 3: We thank the reviewer for highlighting this edge case. To ensure strict temporal consistency with the 0000-2400 UTC window without mathematically interpolating values, our processing pipeline employs a tiered fallback strategy when the standard reaggregation algorithm cannot be directly applied.

Changes in Manuscript:

"If the standard reaggregation algorithm fails or cannot be applied due to non-standard reporting times, a fallback strategy is implemented to ensure temporal alignment with the 0000-2400 UTC window. First, the *Tmax*/*Tmin*-24h records that cover the greatest number of hours within the target day are used as the daily values, provided they overlap for at least 12 hours. Otherwise, the absolute maximum and minimum values derived from the discrete hourly temperature observations are

employed as supplementary estimates for *Tmax* and *Tmin*, respectively, given that a high-density profile of at least 21 valid hourly observations is available for that day."

Reviewer #1 Comment 4: *The paper rather buries a key result – that many of the GSOD biases arise from the double-counting of Tmax/Tmin values reported at 0000 UTC (as discussed in section 5.3.1). How large are other biases?*

Author's Response 4: We agree that the biases arising from the 0000 UTC boundary double-counting are a key result. To explicitly quantify this, we investigated the likely duplicated daily records in GSOD from 1981 to 2024. Our analysis identified approximately 18 million cases (averaging ~1,100 occurrences per day) of likely duplicated *Tmax* records, and 11 million cases (~670 occurrences per day) of likely duplicated *Tmin* records. When compared against the strictly 24-hour physical window of GLBD-FED, these potential duplicates introduced an average bias of 0.28°C for *Tmax* and -0.85°C for *Tmin*.

Regarding "other biases": while UTC boundary double-counting drives the largest temporal aggregation errors, the remaining systematic biases primarily originate from differing sub-daily data source preferences. Specifically, GLBD-FED prioritizes WMO-standard SYNOP reports, whereas GSOD aggregates SYNOP and aviation METAR reports, introducing spatial inconsistencies (as detailed in Section 5.3.2). We have highlighted these quantitative results in Section 5.3.1.

Changes in Manuscript:

"Furthermore, we investigated the extent of likely duplicated daily *Tmax* and *Tmin* records in the GSOD dataset from 1981 to 2024 (identified as identical values derived using the same calculation method on two consecutive days). Our analysis identified approximately 18 million cases (averaging ~1,100 occurrences per day) of likely duplicated *Tmax* records and 11 million cases (~670 occurrences per day) of likely duplicated *Tmin* records. When compared against the GLBD-FED, these potential duplicates introduced an average bias of 0.28°C and -0.85°C, respectively. This indicates that the double-counting issue in legacy datasets systematically results in artificially warmer maximum temperatures and colder minimum temperatures."

Reviewer #1 Comment 5: *Comparing values from different locations is very problematic – this will not isolate the differences between the GSOD and GLBD-FED methods, as the differences deriving from the site change may be much larger than those from the method change... It would be best if possible to restrict this part of the comparison to locations where the GSOD and GLBD-FED datasets are drawing from the same site.*

Author's Response 5: We agree that comparing data from geographically distinct locations would confound the analysis. To clarify: our comparative analyses between GLBD-FED and GSOD are strictly restricted to the exact same WMO Station IDs.

The 4.5°C difference observed for Ulan Bator does not arise from comparing two distinct physical stations 25km apart. Instead, it arises from how the two datasets algorithmically handle multiple data streams (SYNOP vs. METAR) broadcasted under the **same Station ID**. For a single given Station ID, GLBD-FED explicitly prioritizes WMO-standardized SYNOP reports. In contrast, GSOD aggregates all available streams based on UTC timestamps. Because METAR sensors and SYNOP sensors at the same broad location (e.g., airport vs. observatory) can record different microclimates, GSOD's direct aggregation of these sources introduces artificial variations. Thus, the 4.5°C discrepancy strictly isolates the difference between our prioritized method and GSOD's mixed method for the exact same station entity.

Changes in Manuscript:

> "Figure 12 provides a targeted sensitivity analysis for Ulan Bator. Panel (a) displays the standard GLBD-FED daily *Tmin* calculated using our default preference for Synoptic reports, while panel (b) shows the results when the algorithm is specifically forced to utilize only METAR reports from the same station ID. The results in panel (b) nearly overlap with the GSOD data, indicating that the direct aggregation of airport-based METAR reports is a primary contributor to the significantly lower daily *Tmin* values recorded in GSOD for this identical station identifier."

Reviewer #1 Comment 6: *L50-54 – this is a very long list of datasets, but none are particularly recent – suggest being more selective and focus on most recent versions.*

Author's Response 6: We agree. This paragraph originally contained an outdated mix of precipitation and regional datasets. As part of the Introduction rewrite, we have removed this list and now focus specifically on the contemporary landscape of daily temperature datasets, limiting our examples to the most prominent flagship products relevant to our dataset's positioning: GHCN-Daily, BEST, and GSOD.
(Please refer to the comprehensive Introduction rewrite provided in the General Response section for the exact text changes).

Reviewer #1 Comment 7: *L114 – is there any indication of the geographic distribution of different observation times?*

Author's Response 7: We thank the reviewer for this suggestion. We have expanded Figure 1 to include new multi-panel spatial distribution maps illustrating the geographic data volumes for discrete hourly temperature observations, as well as 12-hour and 24-hour *Tmax*/*Tmin* summaries, across four decadal windows. These figures reveal that while hourly observations demonstrate continuous growth, the

explicitly reported extreme summaries suffer from geographic fragmentation and decadal volatility.

Changes in Manuscript:

"Figure 1-T1 visualizes the global distribution of sub-daily temperature data volumes across the 24-hour UTC cycle spanning the period from 1981 to 2024. The analysis reveals striking temporal discrepancies among different temperature parameters. For the discrete hourly temperature observations (utilized to derive Tave), the data volume is continuously distributed across all hours, characterized by a highly robust multi-peak pattern. The primary peaks align perfectly with the standard 6-hourly synoptic times (0000, 0600, 1200, and 1800 UTC), complemented by secondary peaks at the intermediate 3-hourly intervals (0300, 0900, 1500, and 2100 UTC). This dense and temporally consistent distribution provides a solid foundation for calculating highly representative daily mean temperatures.

In stark contrast, the explicitly reported extreme temperatures (Tmax and Tmin) exhibit extreme temporal concentration. The 24-hour extremes (Tmax-24h and Tmin-24h) are overwhelmingly anchored at just two specific reporting times: 0600 and 1800 UTC. Similarly, the 12-hour extremes present a highly asymmetric, diurnal-driven reporting pattern. Specifically, Tmin-12h reaches its absolute volumetric peak at 0600 UTC, capturing the nighttime cooling, whereas Tmax-12h overwhelmingly peaks at 1800 UTC, corresponding to daytime warming. These distinct structural characteristics explicitly demonstrate that while hourly observations offer continuous sub-daily coverage, explicit extreme reports are highly sparse outside of a few specific synoptic hours. This temporal fragmentation structurally mandates and quantitatively justifies the necessity of our secondary fallback strategy, which utilizes the high-frequency hourly observations to robustly reconstruct daily extremes when explicit records are absent.

The multi-panel maps in Figure 1-S1 to S5 illustrate the spatiotemporal evolution of sub-daily temperature data volumes from 1981 to 2024. The analysis reveals a striking contrast in data availability: while discrete hourly temperatures exhibit continuous and stable growth in spatial coverage and reporting frequency over the four decades, explicitly reported extremes (12-hour and 24-hour Tmax/Tmin) suffer from severe geographic fragmentation and decadal volatility, notably experiencing a pronounced global decline between 1990 and 2010. These stark spatiotemporal discrepancies visually highlight the limitations of relying exclusively on explicitly reported extremes and quantitatively justify the necessity of utilizing high-density hourly data as a secondary fallback strategy."

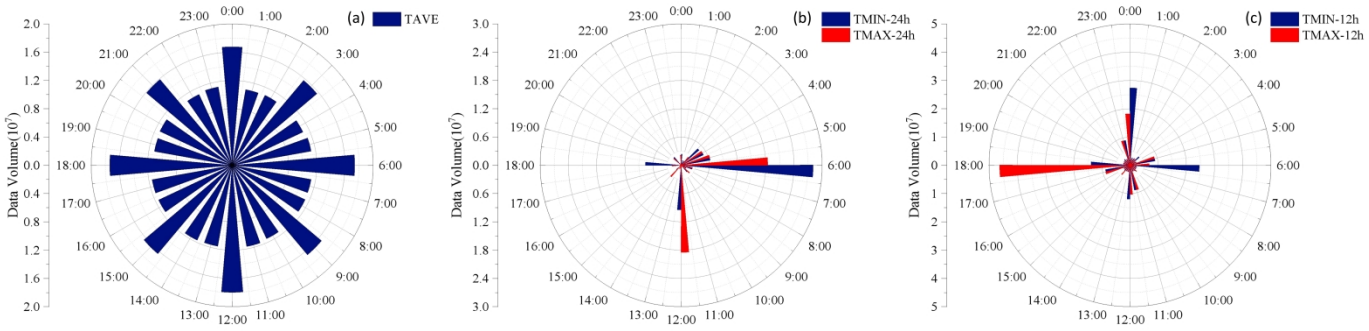

Figure 1-T1 The distribution of sub-daily temperature data amounts at each o’ clock during 1981-2024

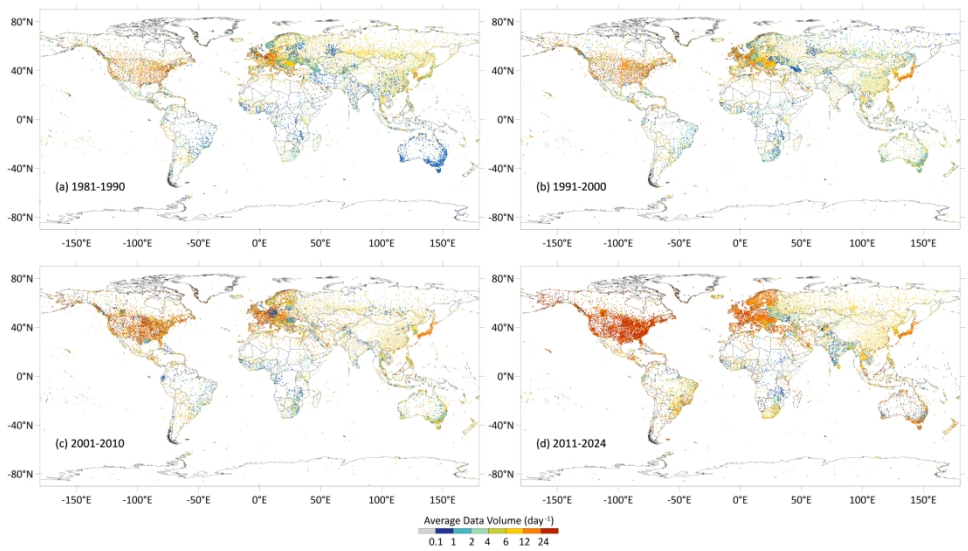

Figure 1-S1 Spatial distributions of Hourly average data volume per day for hourly Tave from ISD. Panel a,b,c,d stand for the results 1981-1990, 1991-2000, 2001-2010, 2011-2024.

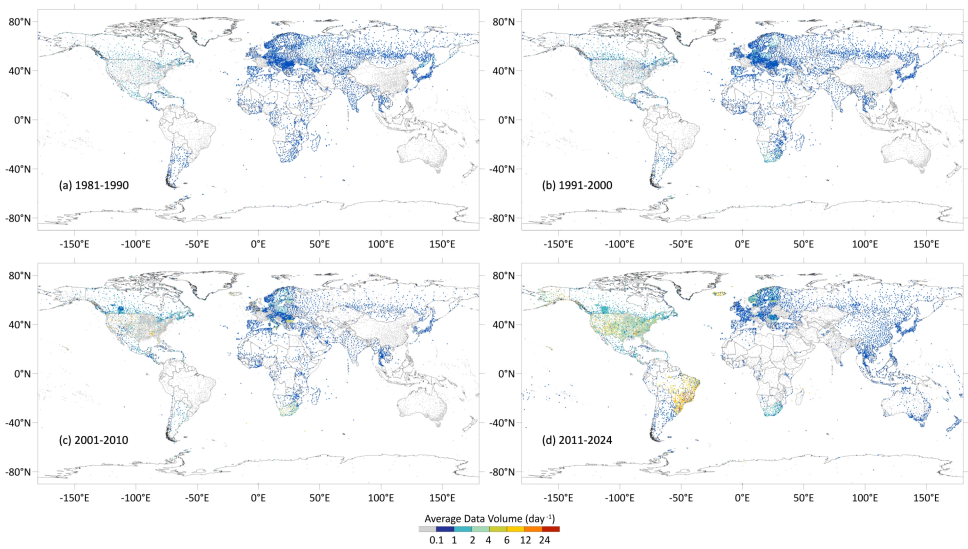

Figure 1-S2 similar to Figure 1-S1, but for Tmax-12h

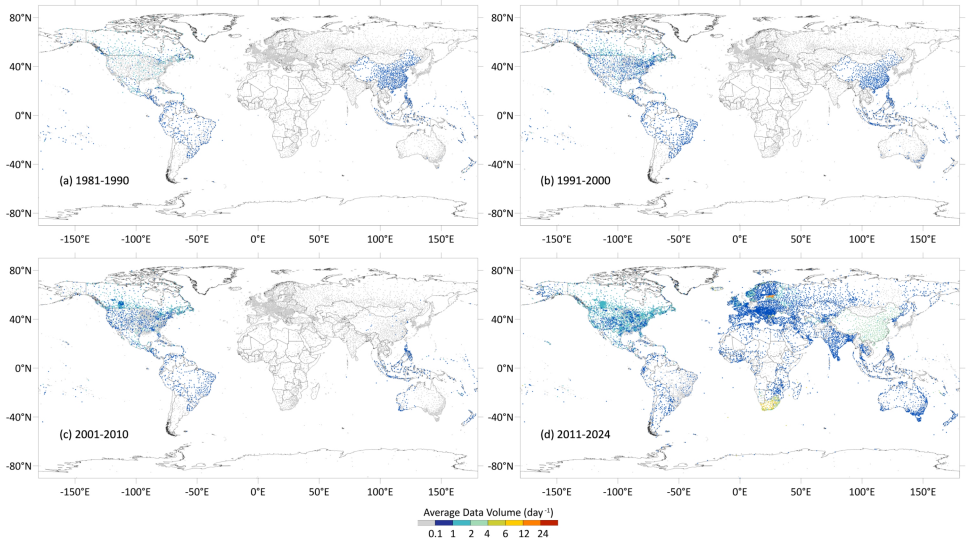

Figure 1-S3 similar to Figure 1-S1, but for Tmax-24h

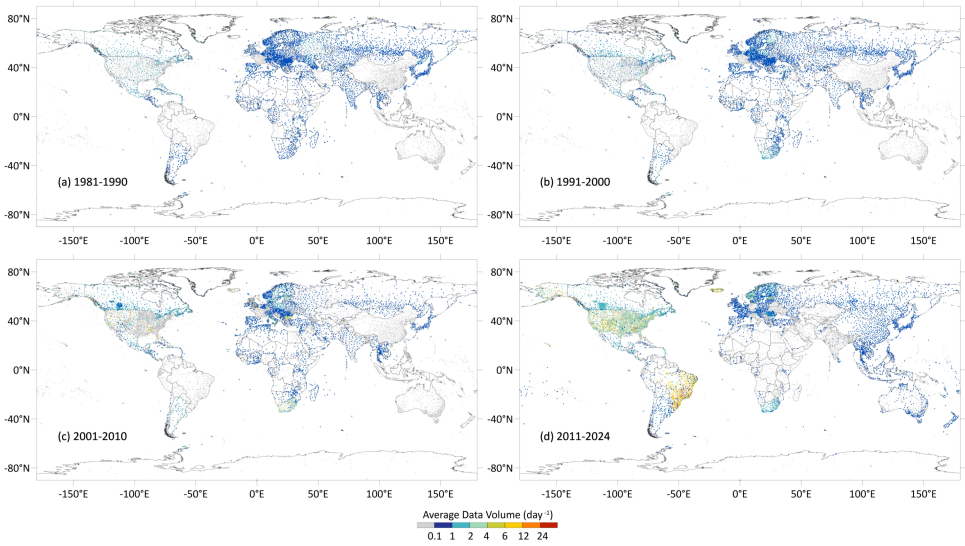

Figure 1-S4 similar to Figure 1-S1, but for Tmin-12h

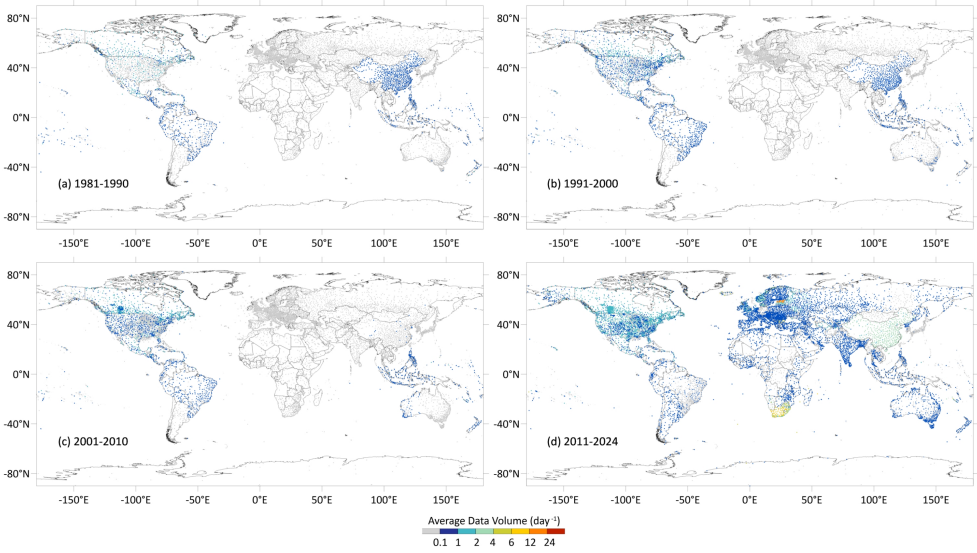

Figure 1-S5 similar to Figure 1-S1, but for Tmin-24h

Reviewer #1 Comment 8: *L129-130 – needs a cross-reference to the appendix for specifics of the quality control methods used.*

Author's Response 8: We have added the necessary cross-reference to the main text.

Changes in Manuscript:

> "All daily data underwent rigorous quality control and were assigned specific quality and date boundary codes (detailed procedures are provided in Appendix, Section 11)."

Reviewer #1 Comment 9: *Section 4.1 – this looks like a fairly labour-intensive process, how practical is it to implement this across the network?*

Author's Response 9: While the logical framework presented in Section 4.1 might appear visually intricate, its execution is entirely programmatic. The correction procedures are a set of automated conditional statements (e.g., cross-referencing reported *Tmax*/*Tmin* against the highest/lowest discrete T_{ave} within preceding windows). Once formulated, these rules were directly translated into our processing scripts. The detection and restoration process is fully automated, requires no manual intervention, and incurs minimal computational cost, making it highly scalable for global implementation.

Changes in Manuscript:

> "To ensure high-throughput scalability across the global network, these corrections were coded as automated conditional rules integrated directly into our processing pipeline. The specific programmatic procedures are as follows: ..."

Reviewer #1 Comment 10: *L184 – 'as discussed in section 3.1' – this matter is not really discussed there. There are some references which could be cited on methods for Tave calculation – see, for example, https://wires.onlinelibrary.wiley.com/doi/full/10.1002/wcc.46 and references therein.*

Author's Response 10: We thank the reviewer for recommending this valuable review paper (Trewin, 2010). First, our original intention in citing Section 3.1 was strictly to refer to the substantial differences in data volume between explicitly reported extremes and discrete hourly observations, rather than the methodologies for T_{ave} calculation. We have refined the wording to clarify this. Second, we agree that the suggested reference provides crucial support for our T_{ave} methodology and have incorporated it.

Changes in Manuscript:

> "The most analyzed variable in climate studies is mean annual or monthly temperature, historically often defined using either fixed-hour observations or the simple average of daily maximum and minimum temperatures (Trewin, 2010). However, as discussed regarding spatiotemporal data volume (Section 3.1), explicit *Tmax* and *Tmin* reports do not always appear in pairs, exhibit lower geographic continuity than hourly observations, and are highly susceptible to Time of Observation Biases (TOB). This implies that deriving the daily T_{ave} directly from the arithmetic mean of discrete, evenly distributed hourly records within the strict 0000–2400 UTC window is significantly more robust and globally uniform relative to deriving it from legacy extreme summaries."

Reviewer #1 Comment 11: *L227 – it is worth noting that some individual countries have a particularly large increase (e.g. Finland is obvious on the max for both Tmax and Tmin).*

Author's Response 11: We agree that highlighting specific countries provides a clearer understanding of the algorithm's regional impact. We have rewritten the relevant paragraph to detail these country-level enhancements.

Changes in Manuscript:

> "Figure 6 illustrates the spatial distribution of the additional data recovered by our temporal reconstruction algorithm. At the national level, several countries exhibit particularly large increases. For instance, Canada, Brazil, Finland, Denmark, Poland, Romania, Australia, New Zealand, Thailand, Indonesia, and the Philippines show notable increases in retrieved daily *Tmax* records. Similarly, regions including Canada, Brazil, Finland, Denmark, Poland, Romania, and Kazakhstan display significant growth in daily *Tmin*. Overall, the recovery of *Tmax* exhibits a wider geographic spread. This broader spatial rise is largely attributable to our algorithm successfully resolving severe baseline scarcities caused by national reporting times clashing with rigid UTC boundaries across networks in Australia and New Zealand."

Reviewer #1 Comment 12: *L245-250 – realistically with the size of the dataset this process would need to be automated – can it be confirmed in the text whether this has been done?*

Author's Response 12: We confirm that the entire quality control process is fully automated. We have revised the text to make this explicit.

Changes in Manuscript:

> "Given the massive volume of the global dataset and our objective to provide a rapid, scalable 'first estimate' baseline, the entire quality control procedure is fully automated. Throughout the processing pipeline, data quality results are evaluated via programmed conditional checks at each step and automatically flagged into three distinct categories: credible, suspicious, and erroneous."

Reviewer #1 Comment 13: *L267-270 – would this be connected with different national practices about whether or not new identifiers are assigned when a site moves?*

Author's Response 13: We agree that this fragmentation is intrinsically linked to divergent national practices regarding station identifiers upon relocation. Observation networks in the U.S. typically assign a new identifier when a station relocates, resulting in shorter data series with fewer relocation-induced inhomogeneities. In contrast, countries like China often maintain a single continuous identifier across relocations to preserve a long-term series. Because GLBD-FED is positioned as a first-hand baseline, we preserve these raw identifier transitions exactly as reported.

Changes in Manuscript:

"...Western Europe and the US demonstrate a high spatial density of daily *Tmax* data. Notably, the prevalence of shorter time series in the U.S. (indicated by green and blue dots, <20 years of records) is closely associated with specific national practices regarding station identifiers upon relocation. The U.S. observation network typically assigns a new, independent ID when a site moves (e.g., station 72200654926 [43.617°N, 96.217°W] post-2005 vs. 72200699999 [43.621°N, 96.216°W] pre-2005). While this practice yields fragmented short series within our first-hand baseline, it effectively reduces relocation-induced inhomogeneities. Conversely, as mentioned above, countries like China tend to retain a consistent ID across relocations to preserve long-term continuity, which necessitates more rigorous homogenization attention in subsequent climate applications. Brazil also displays a high spatial density of data..."

Reviewer #1 Comment 14: *Figure 7 – it looks like in the timeseries there are some individual months with a sharp drop in the number of available stations, is this correct, and if so is there any explanation for it?*

Author's Response 14: We confirm that the sharp drops in available stations in Figure 7 are real features of the data stream. These sudden declines mirror specific periods where the global acquisition of meteorological reports experienced severe disruptions. Temporary anomalies, such as telecommunication outages or server disruptions within the WMO Global Telecommunication System (GTS) and national exchange protocols, can result in significant drops in captured reports.

Changes in Manuscript:

"It should be noted that the occasional sharp drops observed in the time series are real features of the raw data stream, corresponding to specific periods of severe disruption in global data transmission caused by operational instabilities (e.g., telecommunication outages) within the WMO Global Telecommunication System (GTS) and national exchange networks."

Reviewer #1 Comment 15: *L294-297 – how is the averaging done here – is it an arithmetic mean of stations or area-weighted in some way?*

Author's Response 15: The values presented are calculated as the simple arithmetic mean across the available stations, restricted to the spatiotemporal intersection of the two datasets (i.e., using only matched records) to ensure maximum rigor.

Changes in Manuscript:

"Fig. 8 presents the multi-decadal time series (1981-2024) of global daily temperature differences between GSOD and GLBD-FED. **Calculated as the arithmetic mean across the strictly limited spatiotemporal intersection of the two datasets (i.e., only matched records present in both GSOD and GLBD-FED were included), the results demonstrate** that GSOD exhibits a warmer daily *Tmax* (around +0.3°C), a colder *Tmin* (around -0.3°C), and nearly

the same daily T_{ave} (around +0.1°C) relative to GLBD-FED throughout the entire period."

Reviewer #1 Comment 16: *L297-299 – in principle GSOD could be used for extreme events research but in practice I am not aware of any serious climate research which does so (probably because people are well aware of the limitations of the GSOD data).*

Author's Response 16: We agree with your assessment. Serious decadal climate change research generally avoids unhomogenized real-time streams like GSOD. However, datasets like GSOD are actively utilized in lower-latency applications—such as rapid extreme weather monitoring and regional operational synoptic evaluations—where heavily homogenized, high-latency datasets are not feasible. Within these near-real-time contexts, systematic TOB and double-counting biases critically misrepresent the magnitude of extreme weather events.

Changes in Manuscript:

"This means that rapid, near-real-time extreme weather assessments and operational synoptic evaluations relying on legacy datasets like GSOD would systematically misrepresent the magnitude of extreme events—overestimating warm extremes and underestimating cold extremes—compared to methodologically uniform evaluations based on GLBD-FED."

Reviewer #1 Comment 17: *L397 – 'generally reports warmer temperatures' – this isn't quite correct, what is true is that it generally reports warmer temperatures when there is a difference, but many days (presumably those where GSOD is not double-counting) have zero difference.*

Author's Response 17: We agree that the original phrasing was an overgeneralization. The overall positive bias is overwhelmingly driven by specific days where UTC-boundary double-counting occurs. To explicitly validate this dynamic, we introduced a diagnostic scatter plot analysis (provided in Figure 11) to categorize the data based on the presence of consecutive identical values. The non-repeated records predominantly fall along the 1:1 line (zero difference), while the likely duplicated records exhibit a massive positive bias of about 3.2°C.

Changes in Manuscript:

"The mean difference between the two sets of daily *Tmax* values shown in the figure is 1.3°C, indicating a mathematically higher overall daily *Tmax* in GSOD relative to GLBD-FED at this site. To thoroughly investigate the underlying cause, we categorized the data points into two groups based on the presence of consecutive identical values in GSOD (Figure 11). The black points (non-repeated values) predominantly fall along the 1:1 line, indicating zero or negligible difference on days without double-counting. However, approximately one-third of the *Tmax* daily values from GSOD at this station were likely duplicated records (red points), exhibiting a large

positive bias of about 3.2°C. Excluding these potential duplicates results in a roughly 70% reduction in the mean bias, decreasing it to 0.4°C."

Reviewer #1 Comment 18: *Figure 8 – is there any explanation for the seasonal cycle in the early years of these time series?*

Author's Response 18: We sincerely thank the reviewer for this insightful observation. The pronounced seasonal cycle in the biases prior to 1990 originates from distinct data artifacts.

To explicitly validate these underlying causes, we have provided detailed step-by-step diagnostic plots strictly for your review (see supplementary figures below), alongside the following explanation:

1. For Tmin: The seasonality is driven by the periodic fluctuation in the volume of **likely duplicated records** in GSOD associated with the 0000 UTC boundary issue. As shown in the diagnostic plots (Figure R1), there is a near-perfect synchronization between the monthly volume of these repeated records and the global bias magnitude.

2. For Tmax: Unlike Tmin, the seasonal bias in Tmax does not exhibit a strong correspondence with the volume of likely duplicated records (Figure R2), indicating a more complex diagnostic origin. We deduced the exact source of this error through the following analytical steps and determined that it is attributable to algorithmic limitations in GSOD when handling **early cold-season data in high-latitude regions** (e.g., Russia):

- **Spatial Evidence:** Analysis of the Tmax bias during Boreal Winter (DJF, Figure R4) versus Boreal Summer (JJA, Figure R3) demonstrates that the seasonal amplitude is overwhelmingly dominated by stations in Russia.
- **Mechanism (The "Winter Spike" Effect):** Case studies of specific Russian sites (Figure R5) reveal that raw hourly observations in these environments were prone to anomalous spikes (unrealistic, short-term temperature jumps).
- **Algorithmic Limitation:** GSOD's fallback strategy selected the highest available hourly observation, misclassifying these anomalous spikes as valid daily Tmax values, thus inflating the winter Tmax.
- **Quantified Outliers:** Statistical analysis confirms these hourly outliers were significantly more prevalent in winter (Figure R6 and Figure R7). In contrast, GLBD-FED implements rigorous temporal consistency checks that effectively filter out these anomalous spikes.

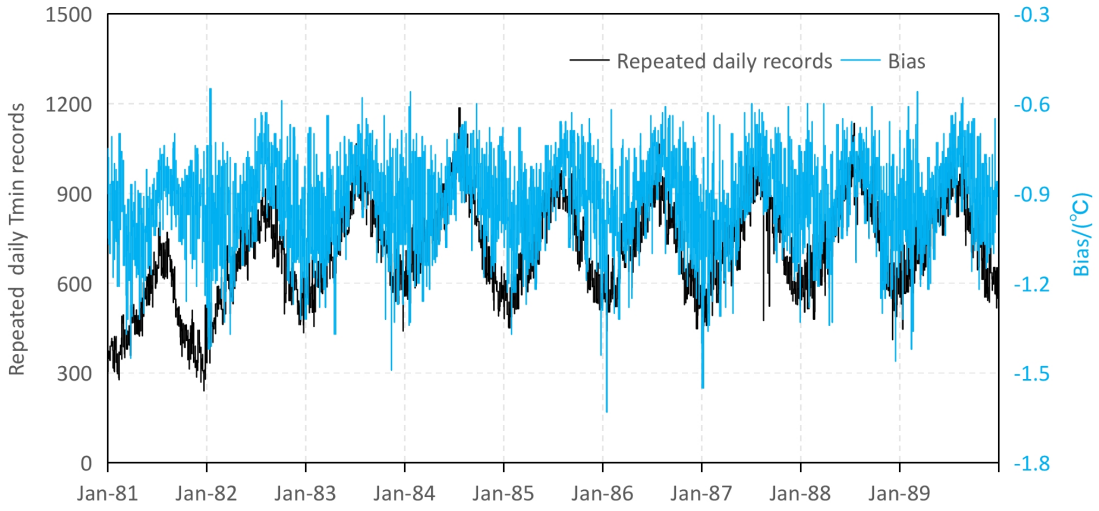

Figure R1 Repeated daily Tmin records number from GSOD (black line) and the bias between them and GLBD-FED (blue line) during 1981-1990.

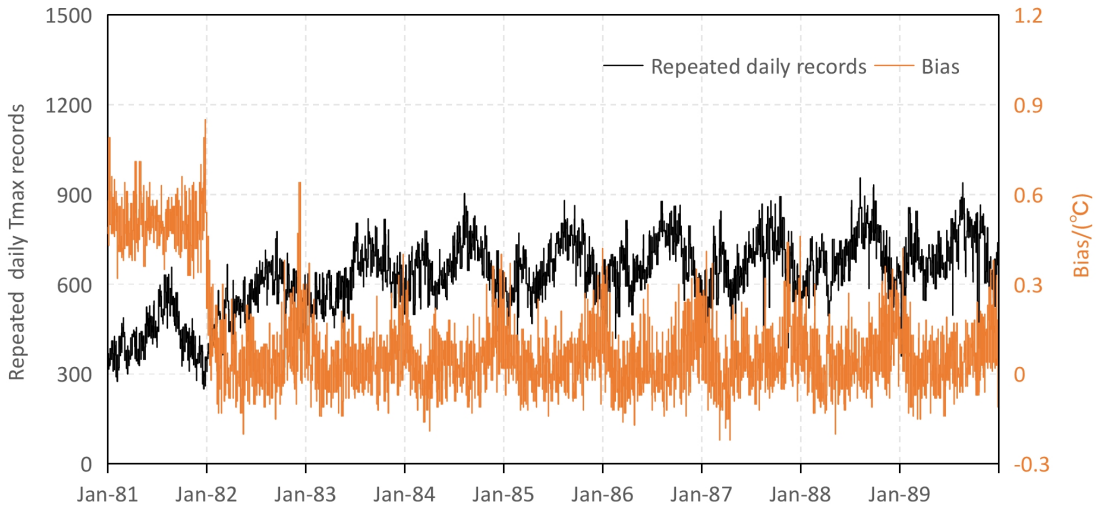

Figure R2 Similar to Figure R1 , but for daily Tmax.

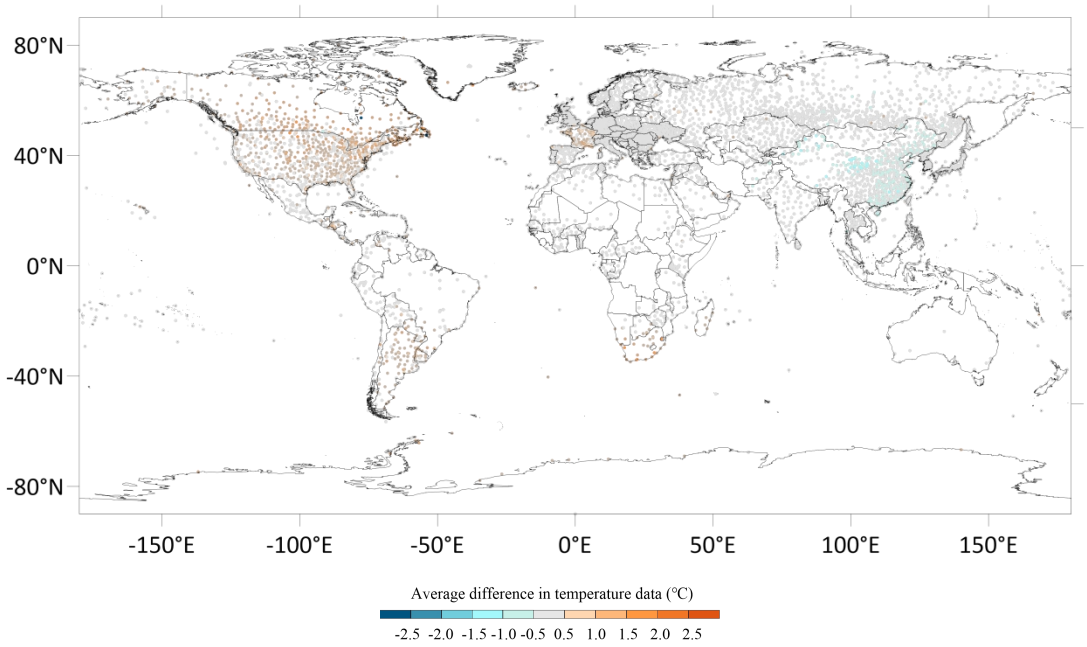

Figure R3 spatial distribution of Tmax bias between GLBD-FED and GSOD in JJA (1981-1989)

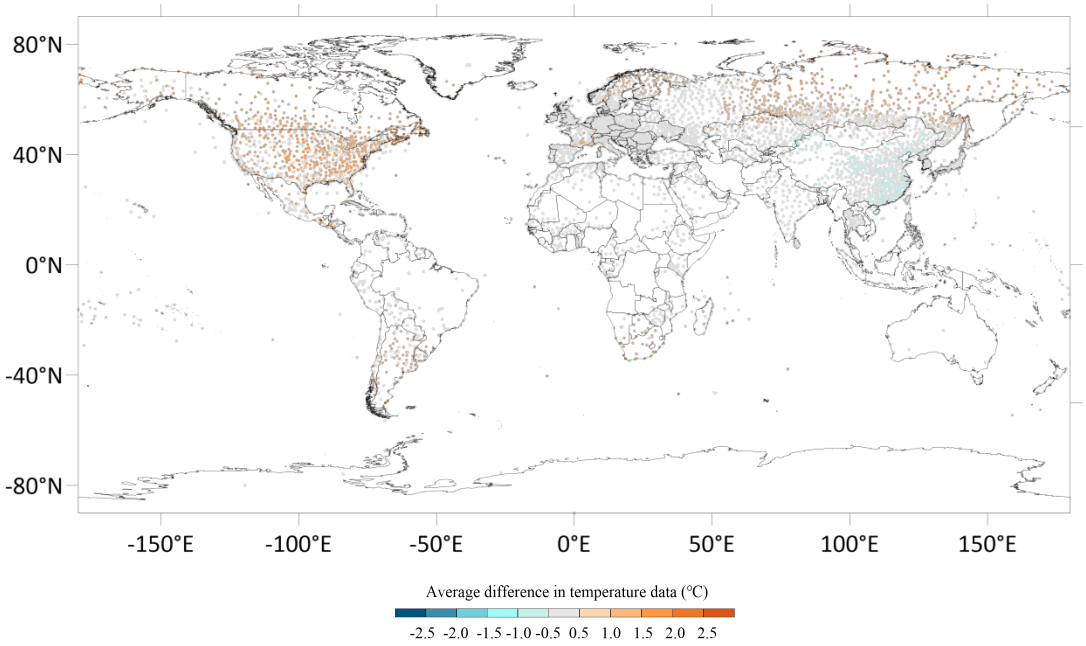

Figure R4 Similar to Figure R3 but for DJF (1981-1989)

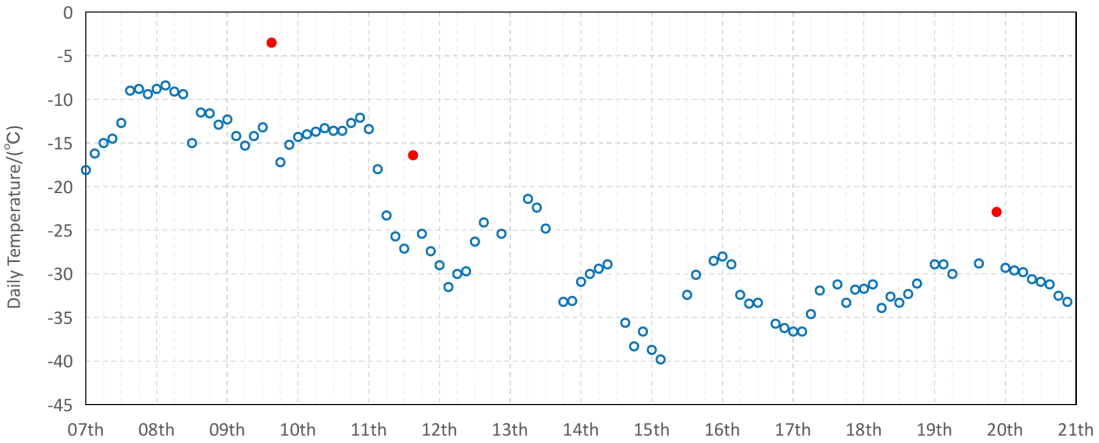

Figure R5 The hourly Tave data from the Russia site (254480-99999) during 7th to 21th Dec 1989. The cycles represent the hourly Tave and the sudden jumping ones were signed by solid red.

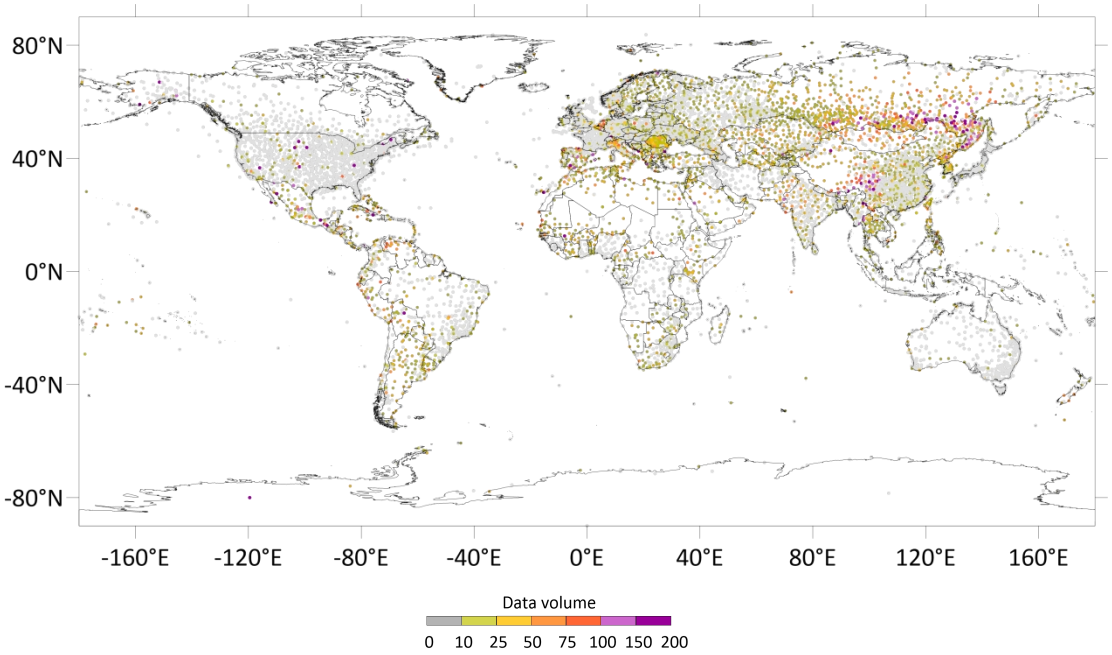

Figure R6 Spatial distribution of the outlier volume of hourly Tave at during DJF 1981-1989

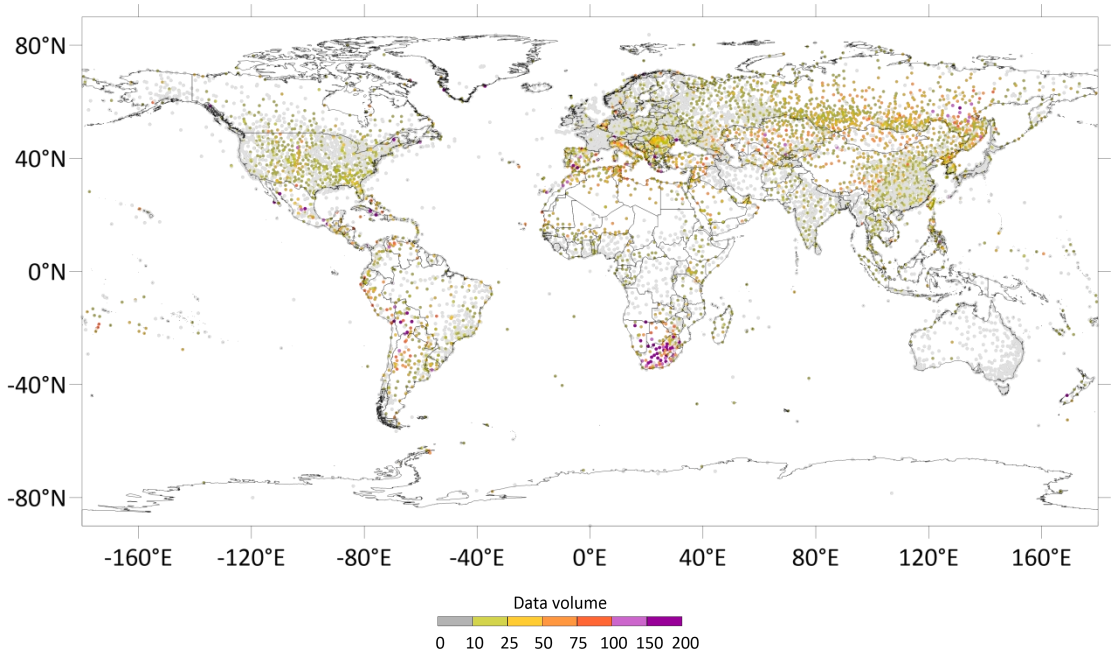

Figure R7 Similar to Figure R6, but during JJA 1981-1989.

Changes in Manuscript:

"A detailed examination of the multi-decadal time series (Figure 8) reveals two notable temporal features: a pronounced seasonal cycle in the biases prior to 1990, and a long-term downward trend in the daily average temperature (Tave) bias (decreasing from approximately +0.25°C in the 1980s to roughly +0.10°C in the recent decade). Both features are intrinsically linked to historical data quality artifacts and algorithmic characteristics.

In the early era, the seasonal cycle in the Tmax bias was largely driven by GSOD's processing of cold-season data from high-latitude regions. During boreal winters, raw observations in these environments were prone to anomalous positive spikes (unrealistic short-term jumps). GSOD's fallback extraction strategy misclassified these anomalous spikes as valid daily maximums, artificially inflating the winter Tmax and subsequently skewing the overall Tave upward. Simultaneously, the Tmin seasonality was driven by periodic variations in the volume of duplicated records within the GSOD archive.

The long-term decreasing trend in the Tave bias reflects the progressive modernization of the global observing network. As automated weather stations and enhanced transmission protocols were widely deployed, the frequency of raw sensor noise dropped significantly. Consequently, GSOD's exposure to selecting these artifacts decreased, leading to a gradual narrowing of the bias over the decades. In contrast, GLBD-FED demonstrates

higher stability throughout the 44-year period, as its rigorous temporal consistency checks successfully filtered out these anomalous spikes even during the early, noisier era."

Reviewer #1 Comment 19: *Figure 9c – it may or may not be relevant to the Australian results here that the standard times for reporting Tmin nationally in eastern Australia is 2200 or 2300 UTC (depending on season), although international synoptic reporting of this is sometimes at 0000 UTC.*

Author's Response 19: We thank the reviewer for this localized insight, which explains the regional anomalies observed in eastern Australia. This regional anomaly serves as an illustration of Time of Observation Bias (TOB). Because legacy extraction algorithms (like GSOD) depend on UTC calendar-day boundaries, this temporal misalignment leads to double-counting or shifting of minimum temperatures into the wrong calendar day.

Changes in Manuscript:

"Meanwhile, sites in GSOD with a lower daily *Tmin* (≤ -0.5°C) compared to GLBD-FED are predominantly concentrated in Australia, Russia, and Northeast Asia (panel c1), accounting for 17.4% of all evaluated sites. For eastern Australia in particular, this regional bias is highly likely attributable to structural discrepancies between national reporting practices and global transmission standards. Specifically, the standard national observation time for *Tmin* in eastern Australia is 2200 or 2300 UTC (depending on the season), whereas international synoptic reports sometimes stamp these observations at 0000 UTC. This temporal misalignment leads to Time of Observation Bias (TOB) causing double-counting or the shifting of minimum temperatures across the 0000 UTC boundary in the legacy GSOD dataset."

Reviewer #1 Comment 20: *Table 2 – the elevations are incorrect here – the general Ulan Bator metropolitan area is at approximately 1300m elevation.*

Author's Response 20: We thank the reviewer for catching this typographical error. We have updated Table 2 to reflect the correct altitude (1300m) for the Ulan Bator metropolitan area.

Changes in Manuscript:

Table 2. Sub-daily reports for Ulan Bator, Mongolia (Station ID: 442920-99999)

Reports Type	Longitude	Latitude	Altitude
Synoptic Report (FM-12)	106.867°E	47.917°N	1306m
Metar Report (FM-15)	106.767°E	47.843°N	1330m