# Point-by-Point Response to Reviewer #2

**Dear Reviewer #2,**

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 #2

**Reviewer #2 Comment 1:** *This work covers the creation of a new dataset based on NOAA's Global Summary of the Day (GSOD) dataset... However, as a data product, it is a shame that the GLBD-FED dataset (or at least, this version) will be static with the change in the source data for the GSOD from ISD to GHCNh. Most of my comments request clarifications or more information, which are detailed below. There are also some minor spacing and punctuation issues...*

**Author's Response 1:** We sincerely thank the reviewer for the positive evaluation and for recognizing the value of our work in addressing historical legacy data anomalies. Regarding your highly pertinent concern about the dataset being "static" following the retirement of the ISD archive, we completely agree that this transition must be transparently addressed. Following the retirement of ISD, our team proactively evaluated its official successor, NOAA's GHCN-Hourly (GHCNh), to seamlessly transition our near-real-time updates. However, our assessment revealed a critical limitation: GHCNh currently lacks the explicit sub-daily extreme reports (e.g., the WMO-standardized 12h/24h extreme summaries) that are strictly required to robustly drive our temporal reconstruction algorithm.

Consequently, continuous real-time operationalization is indeed temporarily paused in its current form. Nevertheless, we believe the completed 1981-2024 GLBD-FED archive stands as a highly valuable, temporally aligned 44-year historical baseline of first-hand data. To ensure transparency, we have explicitly detailed this current limitation in the newly added "Dataset Positioning and Current Status" subsection within the Discussion section. Finally, we have conducted a thorough proofreading to correct all spacing and punctuation issues.

**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 #2 Comment 2:** *Line 59: You are correct that until August 2025, the GSOD did rely on the ISD as its input data... However, with the release of the GHCNh and termination of updates to the ISD in 2025 it is important to indicate at which point you obtained the GSOD, and that the current basis may be different to the version used in your study.*

**Author's Response 2:** We thank the reviewer for highlighting this important timeline regarding the NOAA data transition. We agree that this is a critical point requiring precise clarification. We have clarified our study's data acquisition timeline and the evolving data landscape in the "Data Sources" section of the revised manuscript. We specified that the legacy GSOD dataset was accessed on 2025-03-01, reflecting its ISD-based era.

**Changes in Manuscript:**

"*(Section 3.1)* ...It is important to note that NOAA formally terminated updates to the ISD in 2025, transitioning to the new Global Historical Climatology Network hourly (GHCNh) framework. However, our evaluation indicates that GHCNh currently lacks several critical meteorological elements present in the legacy ISD—specifically the 12-hour and 24-hour temperature extremes. Given that these explicitly reported extremes are essential for our algorithm, the historical ISD remains the only viable source for reconstructing the high-fidelity 1981-2024 baseline presented in this dataset.

*(Section 3.2)* ...The legacy GSOD dataset evaluated herein was accessed on 2025-03-01, reflecting the version that relied exclusively on the ISD as its input. Readers should be aware that with the retirement of the ISD, the legacy GSOD is slated for replacement by a new product, the Surface

Summary of the Day (SSOD). Because the SSOD has not yet been officially released, a direct comparison with the true successor to GSOD is not currently possible. Furthermore, because GHCNh does not currently support our methodological requirements, real-time operationalization of GLBD-FED is temporarily paused..."

**Reviewer #2 Comment 3:** *Line 68: including references here which discuss these temporal representation issues in precipitation data could be of use to readers.*

**Author's Response 3:** We thank the reviewer for this constructive suggestion regarding temporal representation issues. During the revision process, guided by feedback across the review panel, we completely rewrote the Introduction to sharpen the manuscript's focus. As part of this comprehensive restructuring, we removed the extended discussion comparing temperature datasets to precipitation datasets, as it was deemed slightly distracting from the core narrative of temperature-specific biases. Consequently, rather than adding precipitation-specific references, we chose to rigorously expand the theoretical foundation of temporal representation challenges exclusively within the context of temperature—namely, the Time of Observation Bias (TOB). To explicitly present these updates, we have provided the relevant revised paragraph below.

**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 #2 Comment 4:** *Line 71: perhaps point forward to where these challenges are discussed.*

**Author's Response 4:** We thank the reviewer for this helpful structural suggestion. In the newly rewritten Introduction, when we mention the severe temporal aggregation artifacts inherent in legacy near-real-time datasets, we explicitly added a forward reference directing readers to the subsequent sections (e.g., Section 5.3) where these observational challenges and their impacts are quantitatively evaluated.

**Changes in Manuscript:**

"...legacy near-real-time datasets (such as GSOD) often suffer from severe temporal aggregation artifacts—such as misallocating sub-daily extremes to incorrect calendar days due to strict UTC boundary constraints (the specific mechanisms and quantitative impacts of these challenges are discussed in detail in Section 5.3)."

**Reviewer #2 Comment 5:** *Line 81: Also relevant to lines 233-35, 295-99. Somewhere it may be useful to note that to obtain "true" Tmax and Tmin values, subdaily temperature observations in the ISD (which are quasi instantaneous) even with hourly sampling will more often than not miss the Tmax and Tmin values because these do not occur at the "top" of the hour. And hence estimate of Tmaxs from the hourly data will mostly be lower, and Tmin will be higher, than the true values.*

**Author's Response 5:** We thank the reviewer for highlighting this fundamental sampling dynamic. We completely agree that discrete, quasi-instantaneous hourly observations inherently miss the "true" continuous extremes that occur between top-of-the-hour measurements. We have expanded our discussion in the revised manuscript to explicitly acknowledge this sampling bias. Interestingly, your insight into this sampling bias provides a powerful theoretical counterpoint: since deriving extremes from hourly data inherently *underestimates* Tmax, the fact that GSOD still exhibits a significant global *warm* bias relative to GLBD-FED proves that GSOD's overestimation is overwhelmingly driven by its methodological characteristics (e.g., UTC-boundary double-counting), rather than sampling limitations.

**Changes in Manuscript:**

"*(Section 4.2)* It is important to acknowledge that deriving daily extremes from discrete hourly temperature samples—used as a fallback when explicit continuous Tmax/Tmin summary reports are missing—is subject to inherent sampling biases. Because true temperature extremes rarely occur exactly at

the 'top' of the hour, estimates derived from quasi-instantaneous hourly observations tend to be slightly lower for Tmax and higher for Tmin than the true absolute values. Nevertheless, in the absence of explicit extreme reports, utilizing high-frequency hourly observations remains the only practical approach to maintain global temporal continuity.

*(Section 5.2)* Furthermore, the theoretical sampling bias of hourly data provides an important context for interpreting these differences. While the reliance on hourly data for missing records should theoretically result in an underestimation of Tmax, GSOD actually exhibits a systematically warmer daily Tmax (around +0.3°C) relative to GLBD-FED. This contrasting direction strongly suggests that the positive bias in GSOD is not a product of hourly sampling limitations, but rather an artifact of its specific aggregation methodology—namely, the UTC-boundary double-counting of records and the inclusion of anomalous data spikes, which elevate extreme values."

**Reviewer #2 Comment 6:** *Line 84: why did you choose this time range? There is a big uptick in the number of available stations in the ISD in 1973, so I'm surprised this wasn't used as the start year.*

**Author's Response 6:** We thank the reviewer for this insightful comment. While 1973 is indeed a key milestone for overall station counts in the ISD archive, our study prioritizes the functional completeness of explicit sub-daily extreme reports (Tmax/Tmin-12h/24h), which are essential for our temporal reconstruction algorithm. Preliminary analysis shows that while baseline station counts grew in 1973, the global volume of explicitly reported daily extreme records remained relatively low before experiencing a decisive surge to over 4,000 records per day in 1982. Therefore, we established 1981 as the starting point to ensure a robust and spatiotemporally consistent baseline. Extending the temporal coverage backward and forward is a core objective for our next stage of dataset development.

**Reviewer #2 Comment 7:** *Line 89-90: To get an hourly mean temperature, that would entail measurements were taken throughout the hour and then averaged. Although the ISD does contain some off-hour and sub-hourly data I think this would only be available for a subset of the stations and period of record. Please add an explanation of how you obtained the hour mean temperature from the ISD instantaneous measurements? I note from later sections that this is a 24h mean, which was not clear to me at this point, please do make it explicit here that this is a daily mean from the subdaily values*

**Author's Response 7:** You are entirely correct; the ISD provides quasi-instantaneous measurements at the top of the hour, not continuous intra-hour averages. Our original phrasing ("hourly mean temperature") was inaccurate. As you correctly deduced, the Tave parameter in our dataset refers specifically to the **daily mean**

**temperature**. We have corrected the terminology in the text to explicitly state that Tave represents the derived daily mean.

**Changes in Manuscript:**

> "The ISD-derived temperature parameters encompass five specific temporal components: **discrete sub-daily instantaneous temperatures (which are utilized to derive the daily mean temperature, denoted as Tave)**, 24-hour explicitly reported maxima and minima (Tmax-24h and Tmin-24h), and 12-hour explicitly reported maxima and minima (Tmax-12h and Tmin-12h)."

**Reviewer #2 Comment 8:** *Line 92: given your analysis starts in 1980, why do you only show results from 2011 in Figure 1?*

**Author's Response 8:** We agree that restricting Figure 1 to the post-2011 period was inadequate. We have extended the temporal coverage of Figure 1 to span the continuous 44-year period (1981-2024). Furthermore, we expanded Figure 1 to include multi-panel spatial distribution maps illustrating the geographic data volumes across four distinct decadal windows. These new figures clearly reveal that while hourly observations demonstrate continuous growth, explicit extreme records suffer from severe decadal volatility and geographic fragmentation. We explicitly utilize these comprehensive long-term distributions to quantitatively justify the necessity of our secondary fallback strategy.

**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."

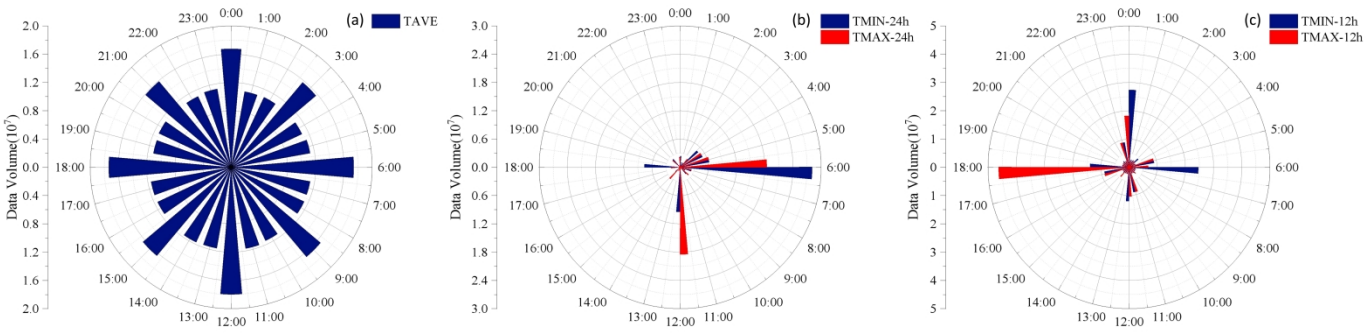



Figure 1-T1 The distribution of sub-daily temperature data amounts at each o’ clock during 1981-2024

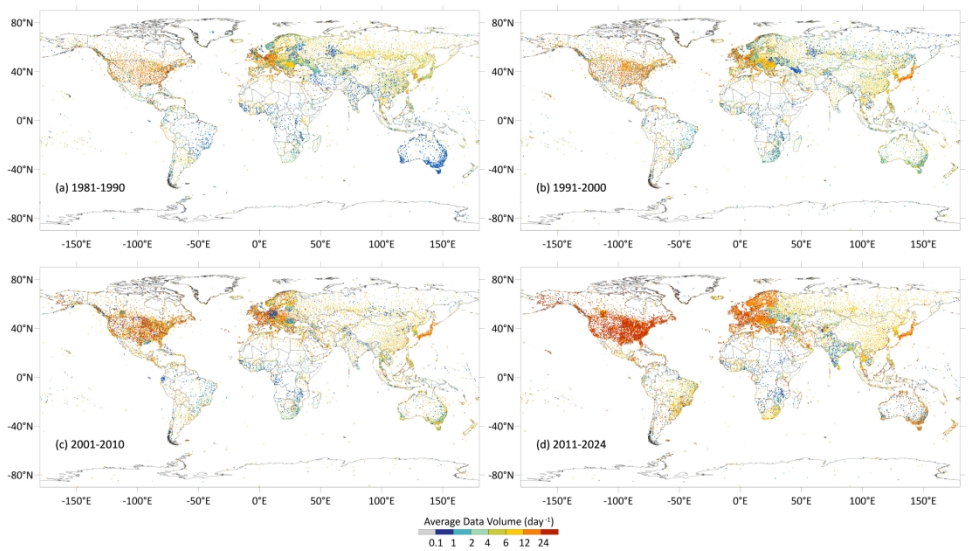



Figure 1-S1 Spatial distributions of Hourly average data volume per day for hourly Tave from ISD. Panel a,b,c,d stand for the results 1981-1990, 1991-2000, 2001-2010, 2011-2024.

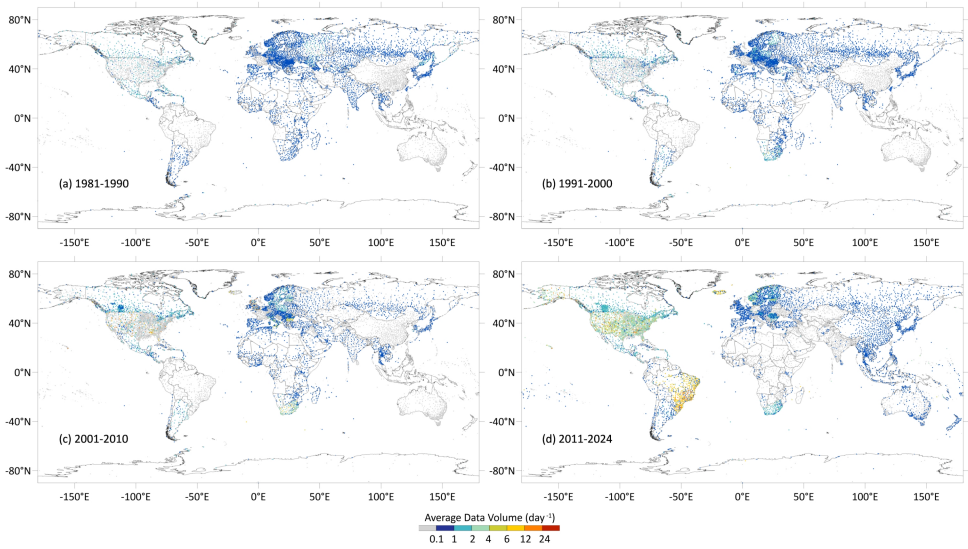



Figure 1-S2 similar to Figure 1-S1, but for Tmax-12h

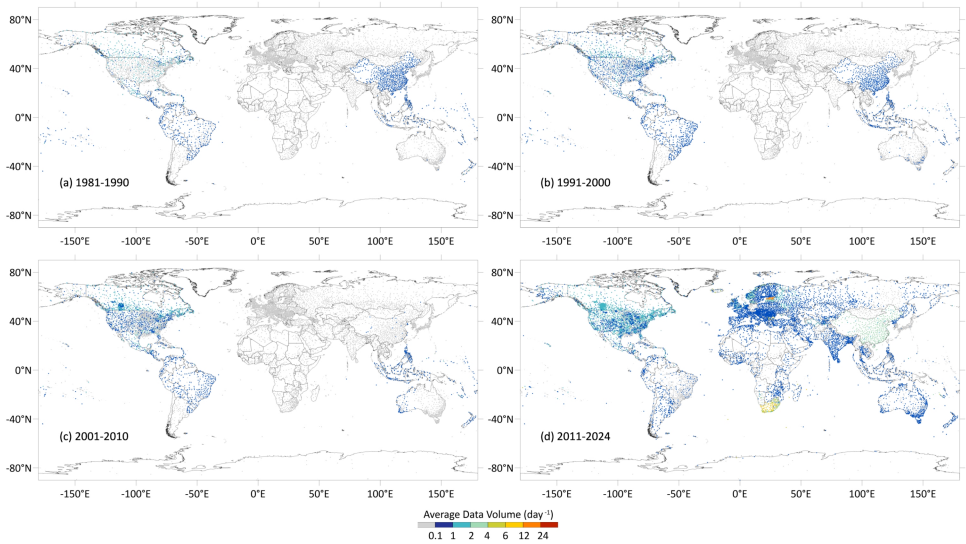



Figure 1-S3 similar to Figure 1-S1, but for Tmax-24h

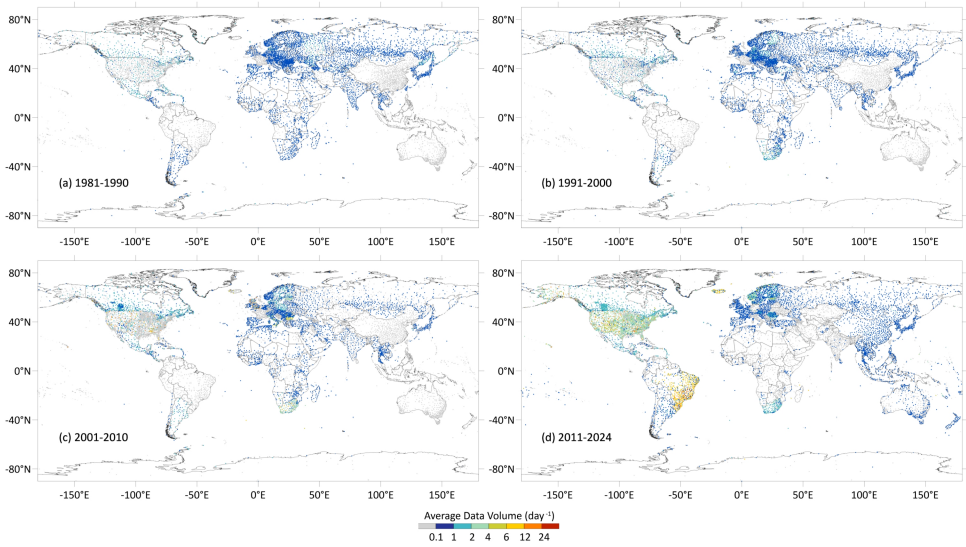



Figure 1-S4 similar to Figure 1-S1, but for Tmin-12h

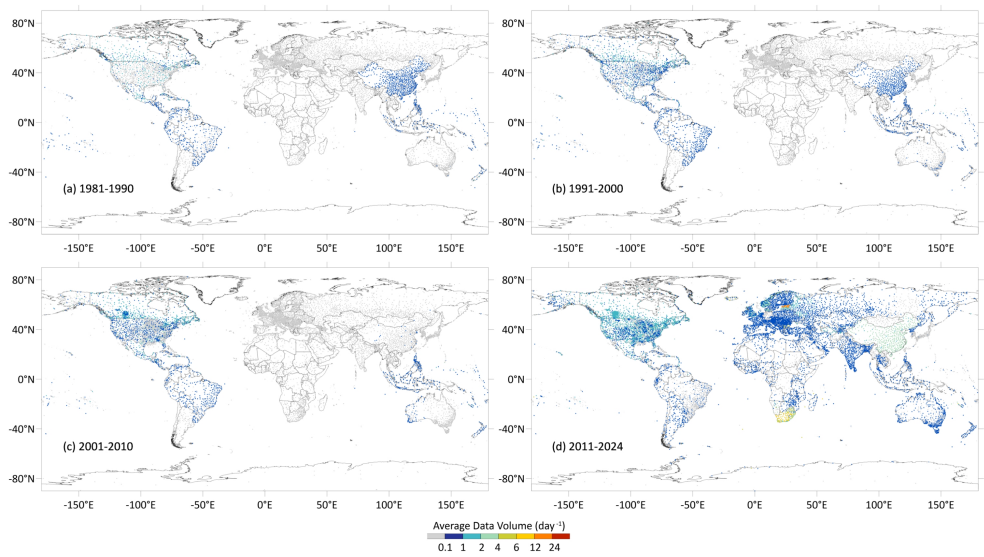



Figure 1-S5 similar to Figure 1-S1, but for Tmin-24h

**Reviewer #2 Comment 9:** *Line 94: Are you sure these are two separate 6hourly cadences, offset by 3h; or could it be a combination of 6 hourly and 3 hourly data? I also suggest noting in this paragraph any selection criteria you use to calculate the daily mean (one observation in each quarter of the day, or all available observations in each quarter....)*

**Author's Response 9:** We agree that the two "6-hourly regimes" identified are indeed a manifestation of the underlying combination of 3-hourly and 6-hourly global reporting standards. Following your suggestion, we have incorporated the specific selection criteria for calculating the daily mean into the manuscript to ensure methodological transparency.

**Changes in Manuscript:**

"Analysis of the discrete sub-daily reporting frequency reveals that the global network primarily operates on a combination of 3-hourly and 6-hourly cadences. These manifest as two dominant 6-hourly regimes offset by 3 hours: 0000/0600/1200/1800 UTC and 0300/0900/1500/2100 UTC. To ensure high-fidelity daily averaging, GLBD-FED employs an equidistant sampling criterion, requiring exactly four instantaneous observations at strict 6-hour intervals per day (e.g., $T_{hh}$, $T_{hh+6}$, $T_{hh+12}$, $T_{hh+18}$). In cases where higher-frequency sub-daily data (such as 1-hourly or 3-hourly reports) provide multiple valid combinations, we apply a sequential priority rule: the algorithm selects the first available 6-hourly sequence by sequentially scanning the starting hours from 0000 UTC (i.e., checking hh=0000 first, followed by 0100, 0200, and so on). This methodological choice ensures a standardized, reproducible approach across the global network while balancing coverage with the physical representativeness of the daily mean."

**Reviewer #2 Comment 10:** *Line 102: The plot suggests 0900 and 1200 show the peaks, please check.*

**Author's Response 10:** You are entirely correct; our original textual description of the reporting peaks was inaccurate. We have completely re-analyzed Figure 1 and rewritten the associated paragraph to correctly identify the robust multi-peak pattern: the primary peaks are aligned with 0000, 0600, 1200, and 1800 UTC, while secondary peaks occur at 0300, 0900, 1500, and 2100 UTC. *(Please see the exact revised text provided in our response to Comment 8 above).*

**Reviewer #2 Comment 11:** *Given these are max/min in 24h, we expect one value in each 24h period per station. As you show these as UTC combined with the likely observing schedule of reading max/min thermometers, how does the distribution of stations with latitude come into play in these plots?*

**Author's Response 11:** The extreme concentration of 24-hour extreme records at specific UTC hours (e.g., 0600 and 1800 UTC) is not primarily driven by the longitudinal/latitudinal geographic distribution of the stations, but rather by World Meteorological Organization (WMO) synoptic reporting protocols. We have added a brief explanatory note to the manuscript to clarify this structural artifact.

**Changes in Manuscript:**

> "It should be noted that the overwhelming concentration of explicit 24-hour extreme reports at specific hours (e.g., 0600 and 1800 UTC) is primarily an artifact of international data exchange standards, rather than a reflection of the longitudinal/geographic distribution of stations. World Meteorological Organization (WMO) synoptic reporting protocols require member states to aggregate and submit extreme records at standardized global synoptic hours (WMO, 2015). Consequently, the ISD archive structurally 'anchors' the vast majority of explicit global extreme reports to these specific UTC timestamps."

**Reviewer #2 Comment 12:** *Line 115: I presume that in panels b, c, the red and blue bars have been purposefully offset? If so, please add this to the caption?*

**Author's Response 12:** You are correct; because Tmax and Tmin are frequently reported at the exact same synoptic hours, the red and blue bars were purposefully offset slightly during plotting solely to prevent overlap and ensure visual clarity. We have added an explanatory note to the caption of Figure 1.

**Changes in Manuscript:**

> "Figure 1. The distribution of sub-daily temperature data amounts at each o'clock during 1981-2024. (Note: In panels b and c, the red and blue bars are purposefully offset slightly for visual clarity to prevent overlap, though they represent records reported at the exact same synoptic hours.)"

**Reviewer #2 Comment 13:** *Line 119: the URL doesn't resolve*

**Author's Response 13:** We apologize for this inconvenience. The original link became inactive due to NOAA's recent website restructuring. We have updated the URL in the revised manuscript to direct readers to the correct active landing page.

**Changes in Manuscript:**

> "The legacy GSOD archive used for comparison in this study was downloaded from NOAA's updated repository at **https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00516**."

**Reviewer #2 Comment 14:** *Line 136: Is the case study representative of all the mis-recorded values in the ISD? I suggest you also indicate the number of stations affected, and potentially the fraction of days for those stations (this plot could go into the Appendix).*

**Author's Response 14:** We sincerely thank the reviewer for this highly constructive suggestion. The San Antonio Oeste case study is indeed highly representative. Following your suggestion, we conducted a comprehensive global assessment and added a figure illustrating the spatial distribution of the mis-recorded Tmax/Tmin ratio at each site during 1981-2024 to the Appendix (Figure S1). As the new figure reveals, these mis-records occur in many regions globally, with notable concentrations in Western Europe, Russia, China, Canada, and South Africa. The primary cause of this widespread artifact is the legacy system erroneously classifying high-frequency reports (e.g., 1-hour interval summaries) as 24-hour extremes.

**Changes in Manuscript:**

> "The case study at San Antonio Oeste is highly representative of a systematic structural artifact within the legacy archive, where local 12-hour observation protocols conflict with global 2400 UTC formatting requirements. A comprehensive global assessment of this specific mis-recording error from 1981 to 2024 reveals that it affected numerous stations worldwide, with concentrated occurrences in Western Europe, Russia, China, Canada, and South Africa. The primary driver of this anomaly is the misclassification of high-frequency sub-daily reports (e.g., 1-hour summaries) as 24-hour extremes. The spatial distribution and frequency of the stations affected by this mis-recording artifact are detailed in the Appendix (Figure S1)."
>
> **(Added to Appendix):**
> "Figure S1 illustrates the spatial distribution of the percentage of mis-recorded Tmax and Tmin data from 1981 to 2024. These anomalies are distributed across multiple regions globally, with notable concentrations in Western Europe and South Africa. The primary cause of these errors is the

frequent misclassification of high-frequency Tmax reports (e.g., 1-hour interval summaries) as 24-hour extreme values, an artifact that is particularly prevalent across Europe."

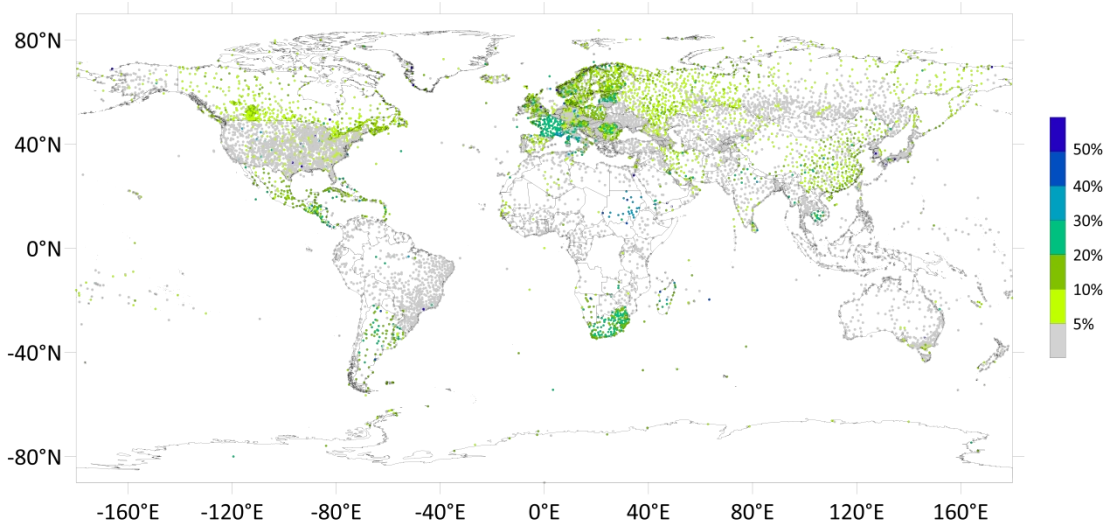



Figure S1 The spatial distribution of percentage of mis-recording Tmax and Tmin data from 1981 to 2024.

**Reviewer #2 Comment 15:** *(1) Line 166: I presume "Uplimit" is "Upper Limit" - please amend the plot if so. I'd also suggest you don't use red and green in combination given some forms of colour blindness can make these indistinguishable. (2) It also appears if some red or green lines are missing - presumably because these have identical values. I think it could help if you can ensure that both lines are visible in these cases (eg using dashed lines). (3) In panel b there are no green lines at all and so it is hard to see what has changed with the correction. (4) The red lines (12h maximums) do not align with the hourly values, being always slightly above. If this is for presentational purposes, please state this in the caption and where relevant in the text, or give an explanation for these differences.*

**Author's Response 15:** We thank the reviewer for these careful observations regarding Figure 3. We have revised the figure and comprehensively expanded its caption to address all of your points:

(1) As you correctly presumed, "Uplimit" was intended to mean "Upper Limit." Specifically, it represents the derived upper limit of temperature within the respective time window, deduced from the explicit Tmax-12h/24h reports. To ensure academic precision and prevent confusion, we have amended the figure legend and text to use precise descriptive labels (e.g., "Upper limit derived by Tmax-12h/24h"). Furthermore, we have replaced the red/green combination with a color-blind friendly palette.

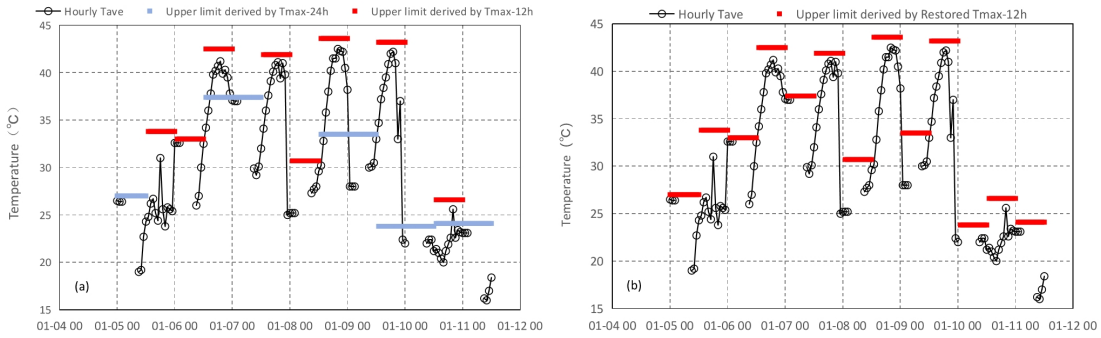
(2) The discontinuous upper limit lines are not caused by identical overlapping values, but simply indicate actual missing observations within the data stream at those specific times.

(3 & 4) The offset where the derived upper limits plot slightly above the discrete

hourly observations is a reflection of physical reality, not a visual artifact. The derived limit captures the absolute continuous peak temperature during the 12/24-hour window, which must theoretically be greater than or equal to any single discrete hourly measurement taken within that same period.

(5) The complete disappearance of the Tmax-24h records in panel (b) explicitly demonstrates our correction algorithm at work. It successfully identified these specific records as erroneously categorized and restored them to their correct physical category.

**Changes in Manuscript:**




"Figure 3. The discrete hourly observations and the derived upper limit of temperature deduced from Tmax-12h/24h during 5th Jan to 12th Jan 2023 at San Antonio Oeste, Argentina. Panels (a) and (b) represent the results from the raw and restored data, respectively. The misrecorded Tmax-24h (blue lines in panel a) are restored as Tmax-12h (red lines in panel b)"

**Reviewer #2 Comment 16:** *Line 176-81: I'm afraid I do not understand what this paragraph is trying to convey.*

**Author's Response 16:** We appreciate the reviewer pointing out this lack of clarity. We realize our use of the term "dateline" was confusing. The original paragraph was intended to describe our sliding-window fallback mechanism for calculating daily values. We have completely rewritten this paragraph and replaced the confusing terminology to make the methodology explicitly clear.

**Changes in Manuscript:**

"**4.2.1. Time window management for daily data calculation**

The standard 24-hour window for calculating daily variables (Tmax, Tave, and Tmin) is strictly defined as 0000 to 2400 UTC. However, to maximize global data retention, we implemented a sliding-window fallback mechanism. If sub-daily observations are insufficient to calculate a valid daily value within the standard UTC day, the 24-hour calculation window is systematically shifted by 1-hour increments (up to ±12 hours) until a complete 24-hour data block is found. To ensure complete transparency and traceability, the specific start and end times utilized for each adjusted daily calculation are explicitly recorded in the dataset metadata."

**Reviewer #2 Comment 17:** *Line 184-88: I think this is saying you obtain daily averages from 4 hourly averages, one in each 6h quadrant of the day? This implies that these have to be regularly spaced. Perhaps note that hh can be one of 0000, 0100....0500 rather than sequentially all of them which would give 6 values. Note that I think the end of the sequence should be 0500 rather than 0600 else there's some double counting.*

**Author's Response 17:** We thank the reviewer for identifying this notational imprecision. You are correct on both points: $hh$ represents a single specific starting hour, and the valid range must terminate at 0500 UTC to prevent the 4-point sequence from shifting into the subsequent calendar day. We have corrected the mathematical notation in the manuscript.

**Changes in Manuscript:**

> "In this context, Tave represents the average temperature, with the subscript 'daily' indicating the daily aggregation and 'hh' denoting the specific starting hour. The value of hh is strictly one of the hours from 0000 to 0500 UTC (i.e., hh $\in$ {0000, 0100, ..., 0500})."

**Reviewer #2 Comment 18:** *Line 200: what do you mean by "less apparent"? What would have happened if the Tmin-15h start time was not aligned with the Tmin-24h start time?*

**Author's Response 18:** By "less apparent," we originally intended to indicate that this 15-hour minimum was "implicitly deduced" rather than explicitly observed. We have replaced this subjective phrasing for absolute clarity. Regarding your critical question on alignment: perfect temporal boundary alignment is a strict mathematical prerequisite for this decomposition algorithm. If the bounding timestamps do not perfectly align, the interval deduction becomes mathematically invalid. We have explicitly formalized this boundary requirement in the revised text.

**Changes in Manuscript:**

> "Initially, an implicitly deduced Tmin-15h at 0000 UTC on October 12 (-1.7°C, indicated by the dashed deep green arrow) was derived from the overlapping Tmin-24h at 0900 UTC (4.4°C, light blue arrow) and Tmin-12h at 1200 UTC on October 12 (-1.7°C, deep blue arrow). It is important to note that this temporal decomposition algorithm strictly requires perfect boundary alignment between the input records; if the boundaries do not perfectly align, the deduction becomes mathematically invalid and cannot be performed."

**Reviewer #2 Comment 19:** *Line 207: I can only see one dashed green arrow (I can't tell if this is the light or dark green one). Please correct the figure/caption.*

**Author's Response 19:** We apologize for the descriptive errors in the original text and caption. We have updated the text and the corresponding figure caption to accurately identify the specific elements using clear line styles (solid vs. dashed) and distinct colors.

**Changes in Manuscript:**

"Figure 4. Application of the reaggregation algorithm for the daily Tmin value at False Pass, US (700638-99999) on October 11, 2024. The solid light and deep blue arrows represent the explicitly measured Tmin-12h and Tmin-24h records, respectively. The dashed green arrow represents the Tmin-15h implicitly deduced from these measured records."

**Reviewer #2 Comment 20:** *Line 225: There are some clear geopolitical border effects visible in Fig 6 - so you could be more specific here.*

**Author's Response 20:** We thank the reviewer for this perceptive observation. The distinct geopolitical border effects visible in Figure 6 are indeed a key feature of the spatial distribution. These sharp spatial contrasts at national boundaries reflect the reality that meteorological observation protocols, reporting frequencies, and data exchange policies are determined independently by individual National Meteorological and Hydrological Services (NMHSs). Because our temporal reconstruction algorithm salvages data by capitalizing on specific sub-daily reporting structures, its enhancement effect inherently maps onto these nation-specific administrative boundaries.

**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.

Notably, the distribution reveals pronounced geopolitical border effects. These sharp spatial contrasts at national boundaries directly reflect the differing sub-daily reporting protocols and data exchange policies enforced by individual National Meteorological and Hydrological Services (NMHSs). Because the temporal reconstruction algorithm capitalizes on specific sub-daily reporting cadences, its enhancement effect naturally maps onto these nation-specific administrative practices."

**Reviewer #2 Comment 21:** *Line 230: These figures will need to be large enough in layout so that the points can be seen. I suggest that the points are made slightly larger, and also that the colourscale used is sequential (e.g. viridis) and not a rainbow*

**Author's Response 21:** Thanks for your suggestion! We fully agree that rainbow colormaps can introduce visual artifacts and are not colorblind-friendly. As recommended, we have completely replotted Figure 6. Specifically, we have significantly enlarged the data points to improve spatial visibility and replaced the original rainbow palette with a perceptually uniform, sequential color scale to ensure optimal accessibility and accurate data interpretation.

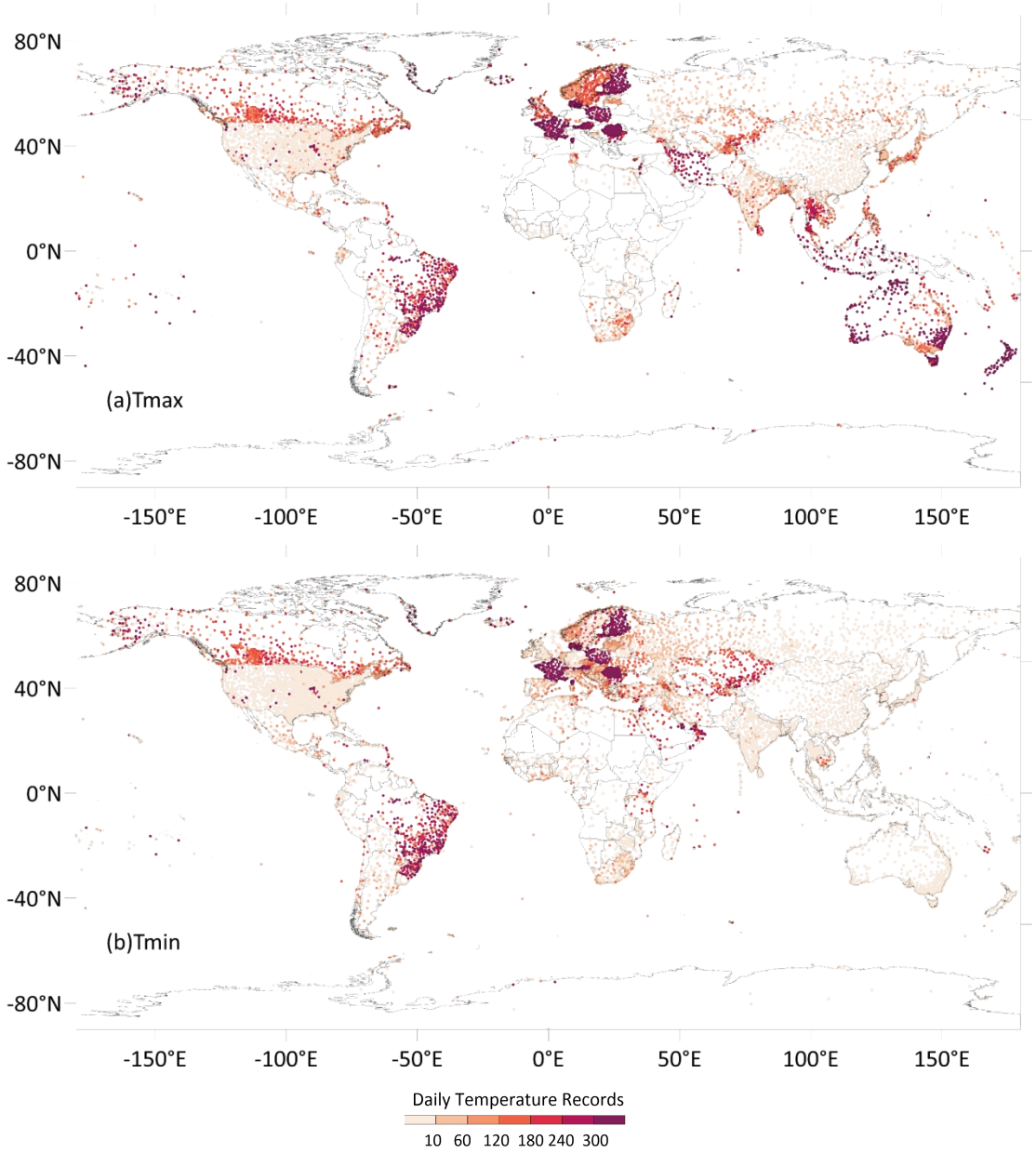



Figure 6 The spatial distribution of the increasing Tmax (panel a) and Tmin (panel b) data improved by the reaggregation algorithm in 2023. The color represents the increasing data volume.

**Reviewer #2 Comment 23:** *Line 243: How did you come up with these tests rather than any others? Can you give a simple summary of the additional temporal and spatial consistency tests here, as you have for the stuck value and inner consistency?*

**Author's Response 23:** We thank the reviewer for this question regarding our quality control methodology. The selected quality control suite (spike, stuck, inner, temporal, and spatial consistency tests) was chosen because it represents the established international standard for meteorological data validation. These specific methods strictly adhere to WMO guidelines and constitute the core automated quality assurance protocols utilized in major global benchmark datasets, such as GHCN-Daily and ISD. Following your suggestion, we have added a concise summary of the temporal and spatial consistency tests to the main text, alongside the authoritative citations that justify their selection.

**Changes in Manuscript:**

"...The daily data underwent these same tests, along with additional temporal and spatial consistency tests. Specifically, the temporal consistency test identifies values that deviate excessively from the station's long-term historical distribution, while the spatial consistency test compares a station's record with simultaneous observations from neighboring stations within a 100-km radius to detect localized anomalies. This specific suite of quality control tests was selected because it represents the established international standard, strictly adhering to WMO guidelines (WMO, 2011; 2018) and the comprehensive automated quality assurance protocols developed for major global datasets such as GHCN-Daily (Durre et al., 2010; Menne et al., 2012) and the ISD (Lott, 2004). Details of these tests for daily data are provided in the Appendix."

**References to be added to the bibliography**

Durre, I., Menne, M. J., Gleason, C. E., Houston, T. G., & Vose, R. S. (2010). Comprehensive automated quality assurance of daily surface observations. Journal of Applied Meteorology and Climatology, 49(11), 1615-1633.

Lott, N. (2004). The quality control of the Integrated Surface Database (ISD). NOAA National Climatic Data Center.

Menne, M. J., Durre, I., Vose, R. S., Gleason, C. E., & Houston, T. G. (2012). An overview of the Global Historical Climatology Network-Daily Database. Journal of Atmospheric and Oceanic Technology, 29(7), 897-910.

World Meteorological Organization (WMO). (2011). Guide to Climatological Practices (WMO-No. 100). Geneva, Switzerland.

World Meteorological Organization (WMO). (2018). Guide to Instruments and Methods of Observation (WMO-No. 8). Geneva, Switzerland.

**Reviewer #2 Comment 24:** *Line 248: Is this over the whole period of record, or per day? Please specify the time period over which the quality results are combined.*

**Author's Response 24:** We thank the reviewer for pointing out this ambiguity regarding the granularity of the quality assessment. The final quality flag is assigned on a strictly **individual data point basis** (i.e., evaluated and flagged separately for each specific daily record, such as the specific Tmax value for a given day). We have clarified this assessment granularity in the revised manuscript.

**Changes in Manuscript:**

> "The quality test results at each stage are compiled into a final assessment of data quality levels **for each individual data point (i.e., specifically for each individual daily record)**. A final quality level **for a given observation** is flagged as credible if there is no more than one suspicious test result and no erroneous test results **across all applied tests**. Conversely..."

**Reviewer #2 Comment 25:** *Line 267: Perhaps use a different example - this station moved by 450m (even with only a 20cm elevation change), but was only in the first position (72200699999) less than a single year. The GHCNh dataset which replaced the ISD may improve the linking of closely located stations in the merge process...*

**Author's Response 25:** We thank the reviewer for this insightful comment. You are entirely correct; the specific example we used was suboptimal for illustrating metadata inconsistency, and we completely agree with your assessment regarding the GHCNh dataset. The widespread fragmentation of U.S. time series shown in our plot is indeed an artifact of the legacy ISD identification system, which frequently generates new IDs for minor station relocations. As you pointed out, the GHCNh dataset effectively resolves this by linking these fragmented records under unified identifiers (e.g., USW-prefix). Rather than simply swapping the example, we now explicitly explain this legacy ISD fragmentation issue and highlight the GHCNh merging mechanism as the definitive solution.

**Changes in Manuscript:**

> "Western Europe and the US demonstrate a high spatial density of daily Tmax data. Notably, the apparent abundance of short time series in the U.S. (indicated by green and blue dots, <20 years of records) is largely an artifact of the legacy ISD station identification system. Under this older system, minor station relocations or administrative updates frequently resulted in the assignment of new station IDs, artificially fragmenting what are physically continuous observations. As recent advancements have shown, processing these records through the unified identifier framework of the newer GHCNh dataset (e.g., merging closely located fragments under a single USW-prefix station ID) effectively resolves these artificial breaks and restores the long-term continuity of the time series."

**Reviewer #2 Comment 26:** *Line 284 - b3 - there seems to be some quasi-periodic nature in the early part of the record for Tmin in the daily data volumes. Do you know why?*

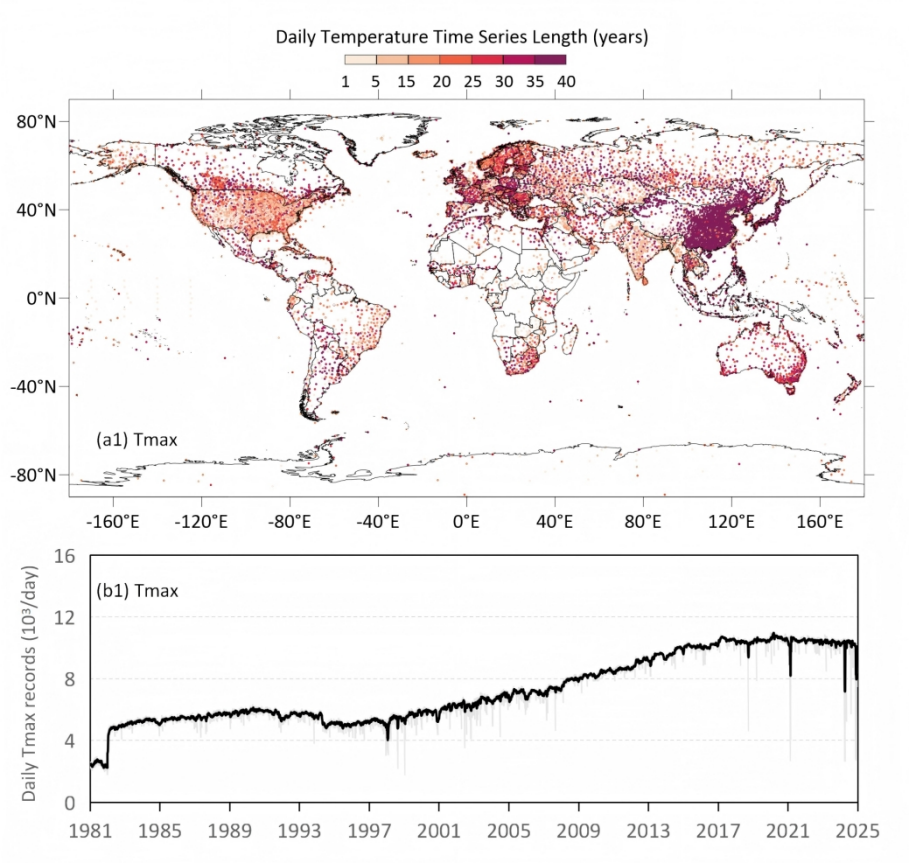
**Author's Response 26:** We thank the reviewer for this sharp observation. The quasi-periodic fluctuations (dips) in the early Tmin record are direct consequences of severe historical data scarcity in the underlying ISD archive. Our temporal reconstruction algorithm primarily relies on explicitly reported extremes (Tmin-12h/24h) to calculate daily values, utilizing high-frequency hourly observations as a secondary fallback. During those earlier years, the raw archive frequently suffered from a simultaneous lack of both components. When both data sources fail to meet the necessary computational completeness thresholds, the daily Tmin physically cannot be reconstructed, resulting in these temporary volume drops.

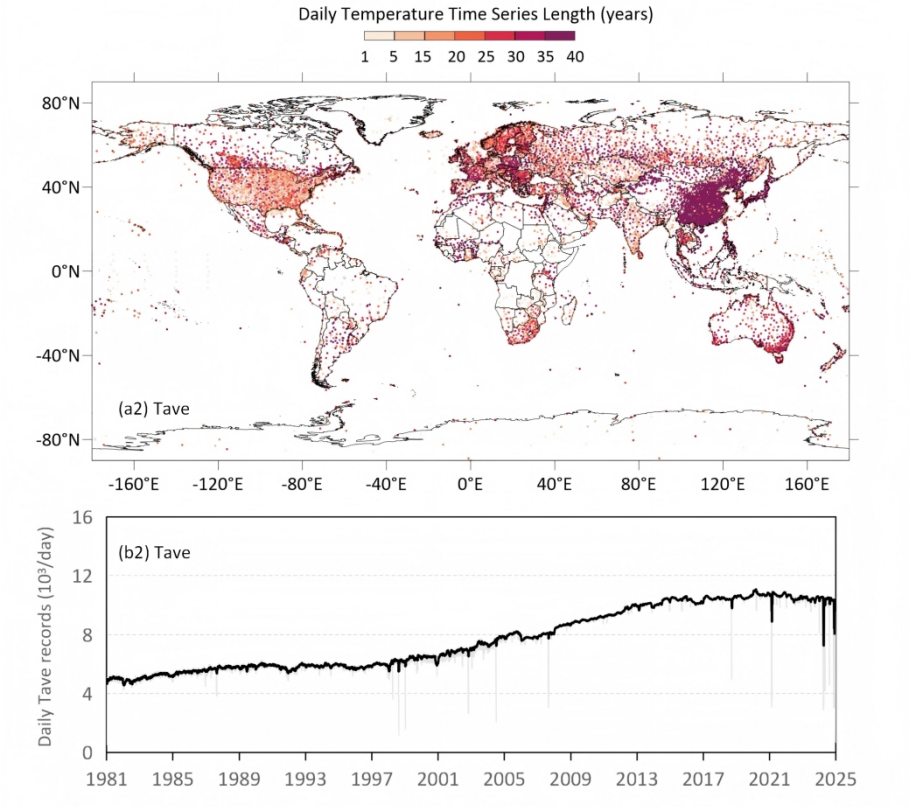
**Changes in Manuscript:**

> "Furthermore, the quasi-periodic fluctuations (dips) observed in the early Tmin record (Fig. 7, panel b3) are artifacts of historical data scarcity in the underlying ISD archive. Our reaggregation algorithm is designed to primarily utilize explicitly reported extremes (Tmin-12h/24h), relying on high-frequency hourly observations as a secondary fallback. During these earlier years, the raw archive occasionally lacked both explicit extreme reports and sufficient sub-daily observations. When neither data source meets the necessary computational thresholds, valid daily Tmin values cannot be reconstructed, leading to temporary drops in the aggregated data volume."

**Reviewer #2 Comment 27:** *Line 285: Fig7 panels a1-3. Please use a sequential colourscale here (see comments against line 230)*

**Author's Response 27:** Thanks for your suggestion. Consistent with our revisions to Figure 6 and adhering to best practices for data visualization, we have updated the color scales in Figure 7. We changed the color bar to a sequential 'rocket' colourscale to accurately and accessibly represent the daily temperature time series lengths.


Daily Temperature Time Series Length (years)
1 5 15 20 25 30 35 40
80°N
40°N
0°N
-40°N
-80°N
(a1) Tmax
-160°E -120°E -80°E -40°E 0°E 40°E 80°E 120°E 160°E
(b1) Tmax
Daily Tmax records (10³/day)
16
12
8
4
0
1981 1985 1989 1993 1997 2001 2005 2009 2013 2017 2021 2025


Daily Temperature Time Series Length (years)
1 5 15 20 25 30 35 40
80°N
40°N
0°N
-40°N
-80°N
(a2) Tave
-160°E -120°E -80°E -40°E 0°E 40°E 80°E 120°E 160°E
(b2) Tave
Daily Tave records (10³/day)
16
12
8
4
0
1981 1985 1989 1993 1997 2001 2005 2009 2013 2017 2021 2025

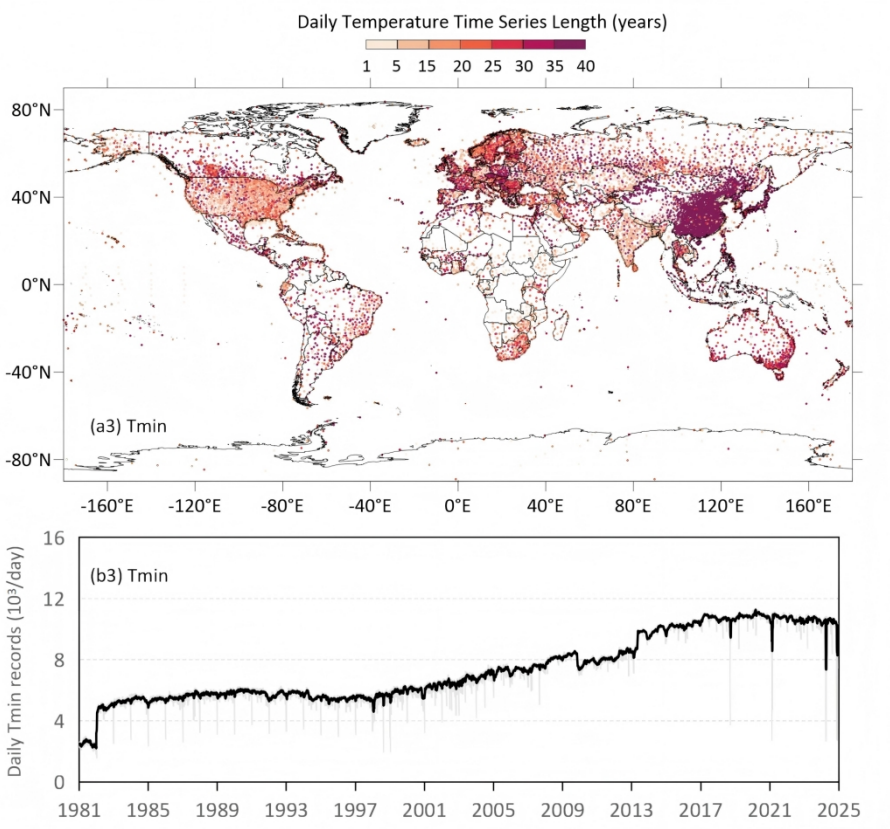



*Figure 7 The spatial distribution (panel a1, a2, and a3) and temporal changes (panel b1, b2 and b3) of global daily temperature data during 1981-2024. Panel a1 and b1 represent the Tmax, panel a2 and b2 represent the Tave, panel a3 and b3 represent the Tmin. The colorful dots in panel a represent the length of the daily data at sites; the gray and black curves stand for the daily data volume and the 15 points-smoothing result, respectively.*

**Reviewer #2 Comment 28 :**

*Line 293-99: This paragraph has a number of spelling and grammar errors.*
*Line 294: If nearly all records were included, why were some not?*

**Author's Response 28:** We sincerely apologize for the imprecise phrasing and grammatical errors in this paragraph. Regarding your question in Comment 29, our use of "nearly all" was confusing. To ensure a rigorous comparison, we strictly evaluated the **intersection** of the two datasets. The few excluded records simply lacked a spatiotemporal match in the counterpart dataset. To address both comments comprehensively, we have thoroughly rewritten this entire paragraph. We explicitly clarified the intersection logic and corrected all spelling and grammatical issues to ensure academic rigor.

**Changes in Manuscript:**

"Fig. 8 presents the multi-decadal time series (1981-2024) of global daily temperature differences between GSOD and GLBD-FED. **The comparison was strictly limited to the spatiotemporal intersection of the two datasets;**

**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 Tave (around +0.1°C) relative to GLBD-FED throughout the entire period. **This means that research utilizing** GSOD daily temperature data **would likely yield more pronounced** climate extreme events than studies based on GLBD-FED."

**Reviewer #2 Comment 29 :** *Line 302-5: I think you could be more explicit here as these two comparisons almost sound as if they are the same, as my understanding is that the GSOD selects the highest/lowest records from the hourly data within 24 hours.*

**Author's Response 29:** We thank the reviewer for pointing out this ambiguity. To clarify your understanding: GSOD does indeed employ an algorithmic hierarchy. It first extracts the maximum and minimum from explicitly reported extreme records. If these are unavailable, it then falls back to selecting the extremes from the discrete hourly temperature observations. However, the critical issue lies in its temporal attribution: a Tmax-24h reported at 0300 UTC physically represents the previous day's extreme, but GSOD directly assigns it to the current day based strictly on its reporting timestamp. In contrast, GLBD-FED strictly realigns and reaggregates these continuous extreme reports to their actual physical occurrence windows. We have comprehensively rewritten this paragraph to make this fundamental algorithmic distinction explicit. Furthermore, to quantify the magnitude of this structural bias, we added the statistical results showing that this double-counting issue alone generated approximately 18 million anomalous extreme records over the 44-year period.

**Changes in Manuscript:**

"The difference in daily data definition and extraction logic between GLBD-FED and GSOD is the primary cause of this systematic variability. GLBD-FED rigorously reconstructs true daily extremes by strictly aligning and reaggregating continuous extreme reports (Tmax/Tmin-12h/24h) to their actual physical occurrence windows. In contrast, while GSOD employs an extraction hierarchy—prioritizing explicitly reported extremes and falling back to discrete hourly observations if extremes are missing—it attributes these values to a calendar day based strictly on their UTC reporting timestamps. Consequently, in GSOD, a 24-hour extreme report that physically summarizes the previous day (e.g., a Tmax-24h reported at 0300 UTC) is erroneously treated as the daily extreme for the current day simply because of its timestamp. The profound impact of this timestamp-based misallocation is discussed in detail in Section 5.3.1.

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 #2 Comment 30:** *Line 335 (Fig 9) could the bar-plots please have some x-tick marks as the x-tick labels aren't quite aligned with the bar edges?*

**Author's Response 30:** Thanks for your suggestion! We appreciate your careful observation. We have replotted Figure 9 and added distinct x-tick marks to the bar plots in the lower panels to ensure proper alignment with the bar edges.

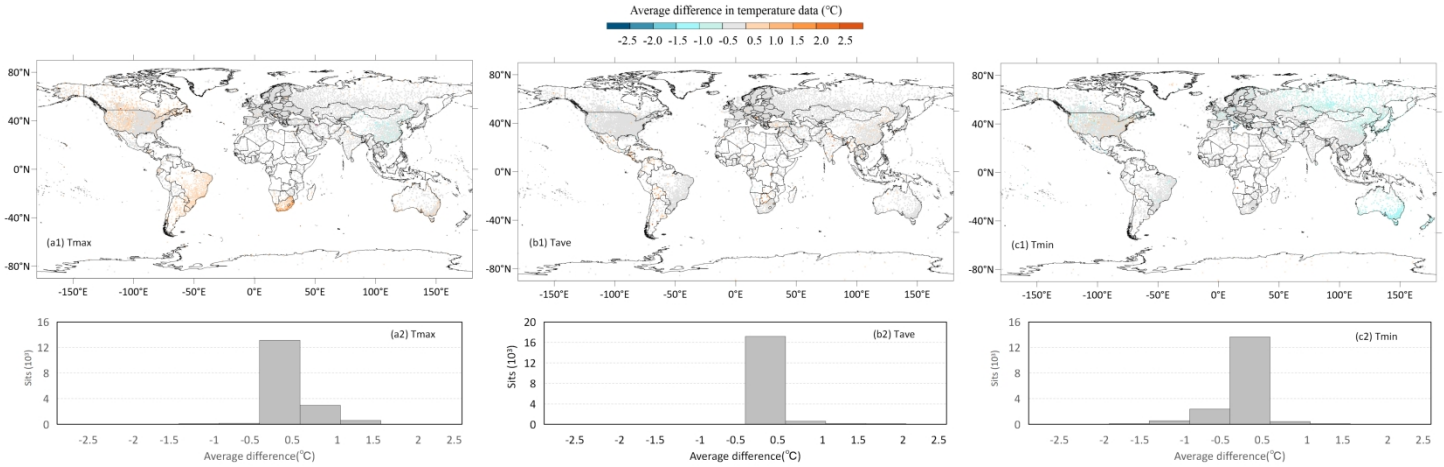



Figure 9 Spatial distribution of the difference in-situ daily temperature data between GLBD-FED and GSOD during 1981-2024 (GSOD minus GLBD-FED). Panel a1, b1, c1 show the difference in Tmax, Tave and Tmin at each site, respectively. Panel a2, b2, c2 represent the sites number distribution with diversities

**Reviewer #2 Comment 31:** *Line 311: Can you give an explanation for the periodicity for the first 15 years of the record in Tmax and Tmin?*

**Author's Response 31:** **Regarding the periodicity prior to 1990:** The "periodicity" is indeed a pronounced seasonal cycle originating from two distinct data artifacts within the GSOD processing chain. 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, 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, 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.

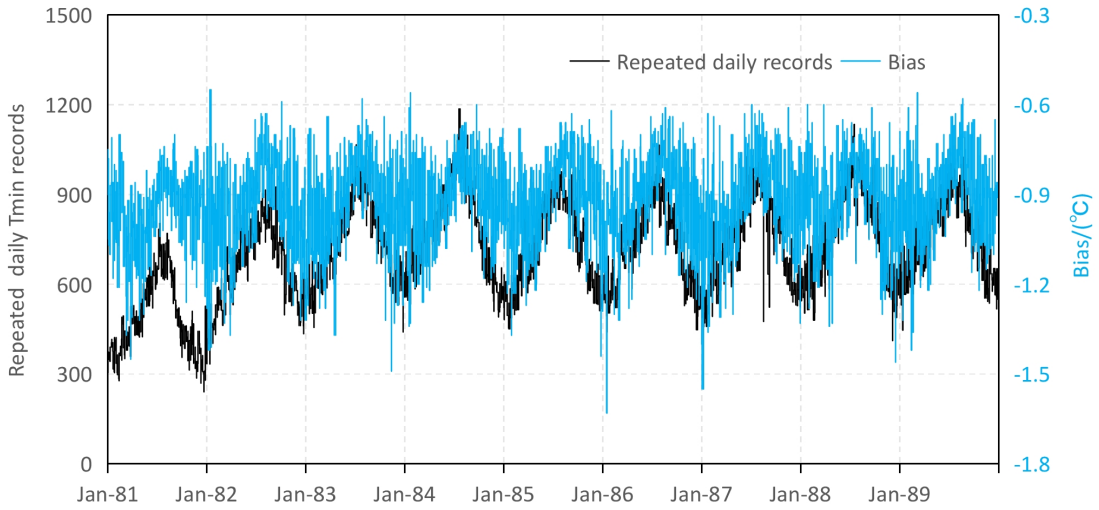



Figure R1 Repeated daily Tmin records number from GSOD (black line) and the bias between them and GLBD-FED (blue line) during 1981-1990.

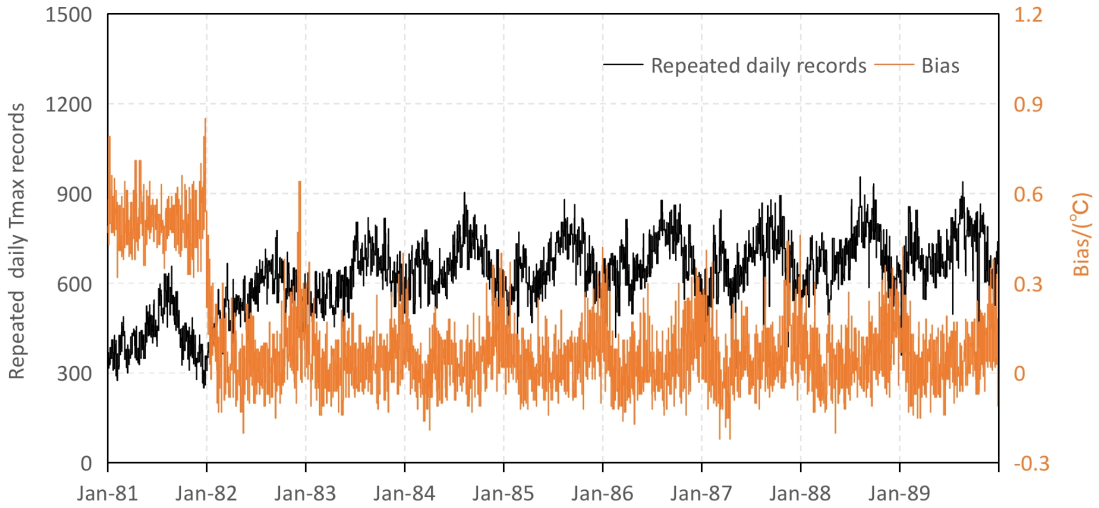



Figure R2 Similar to Figure R1 , but for daily Tmax.

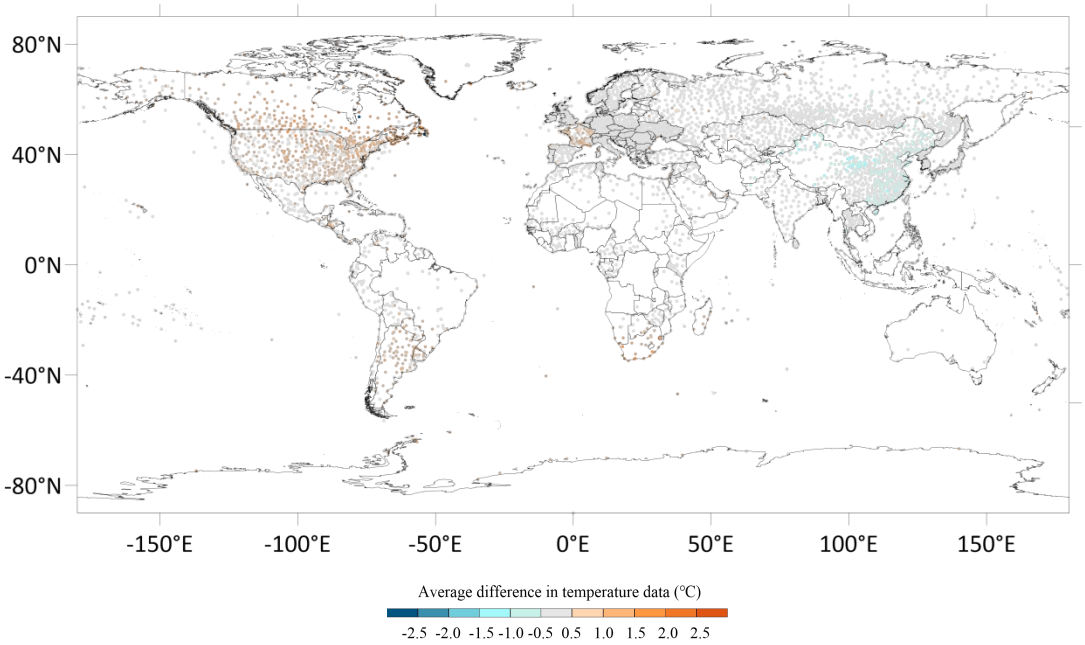



Figure R3 spatial distribution of Tmax bias between GLBD-FED and GSOD in JJA (1981-1989)

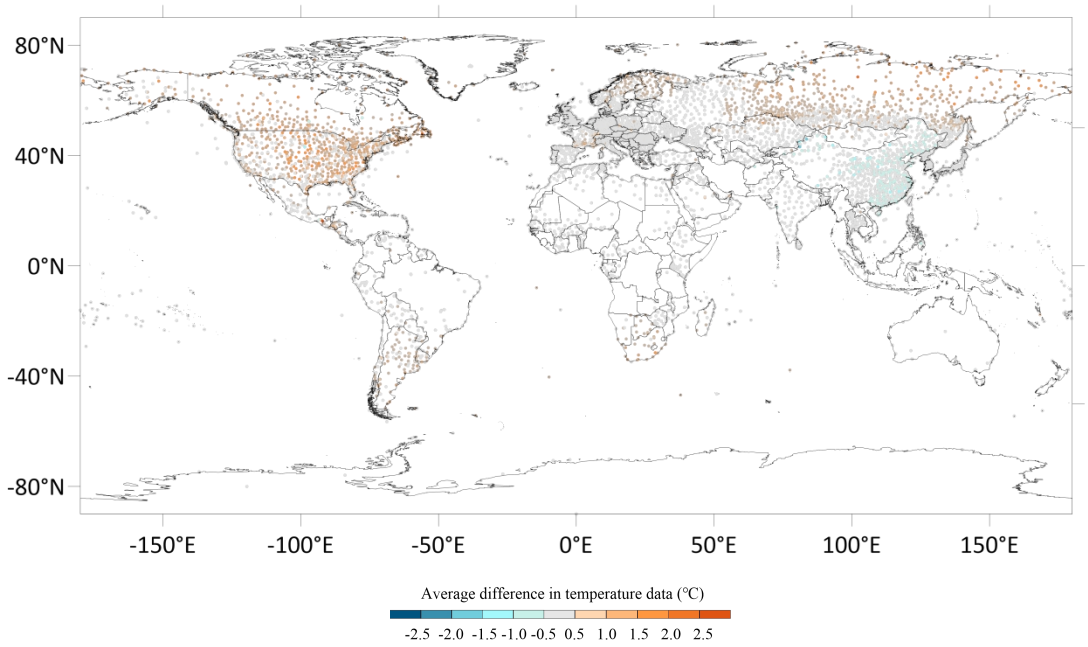



Figure R4 Similar to Figure R3 but for DJF (1981-1989)

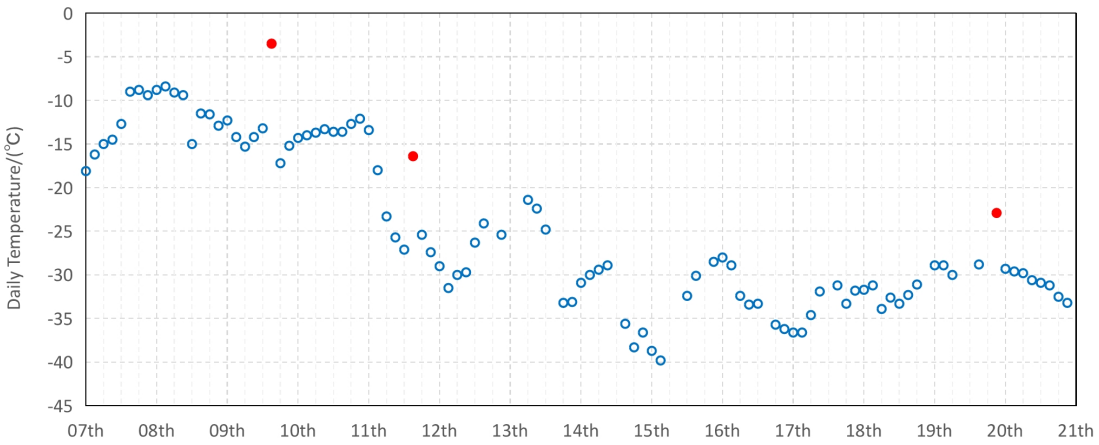



Figure R5 The hourly Tave data from the Russia site (254480-99999) during 7$^{th}$ to 21th Dec 1989. The cycles represent the hourly Tave and the sudden jumping ones were signed by solid red.

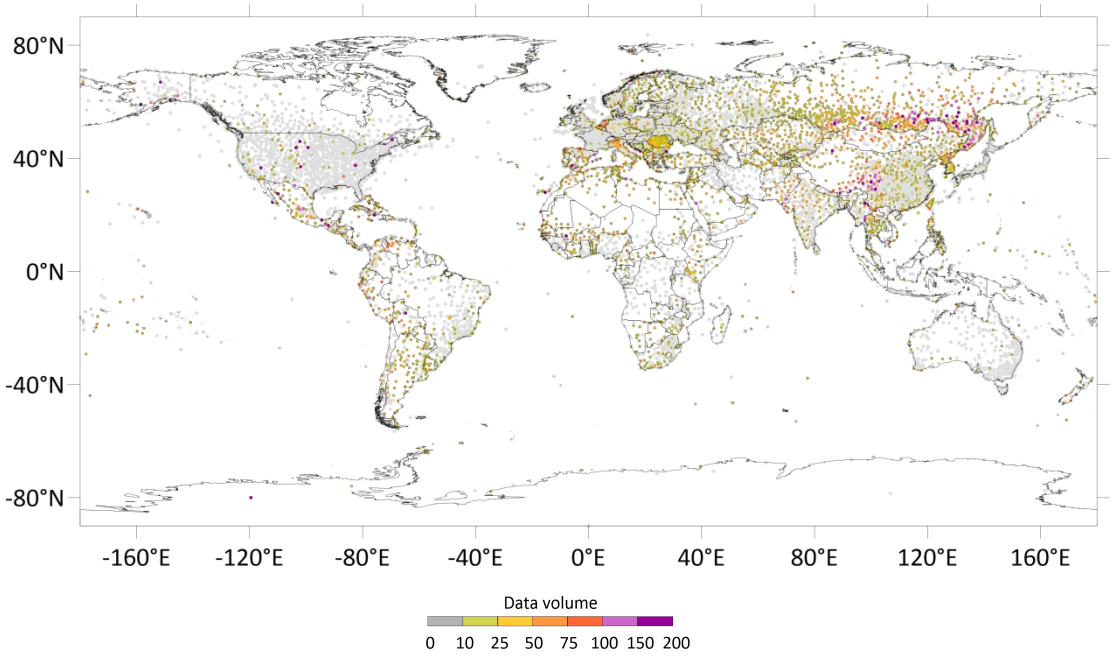



Figure R6 Spatial distribution of the outlier volume of hourly Tave at during DJF 1981-1989

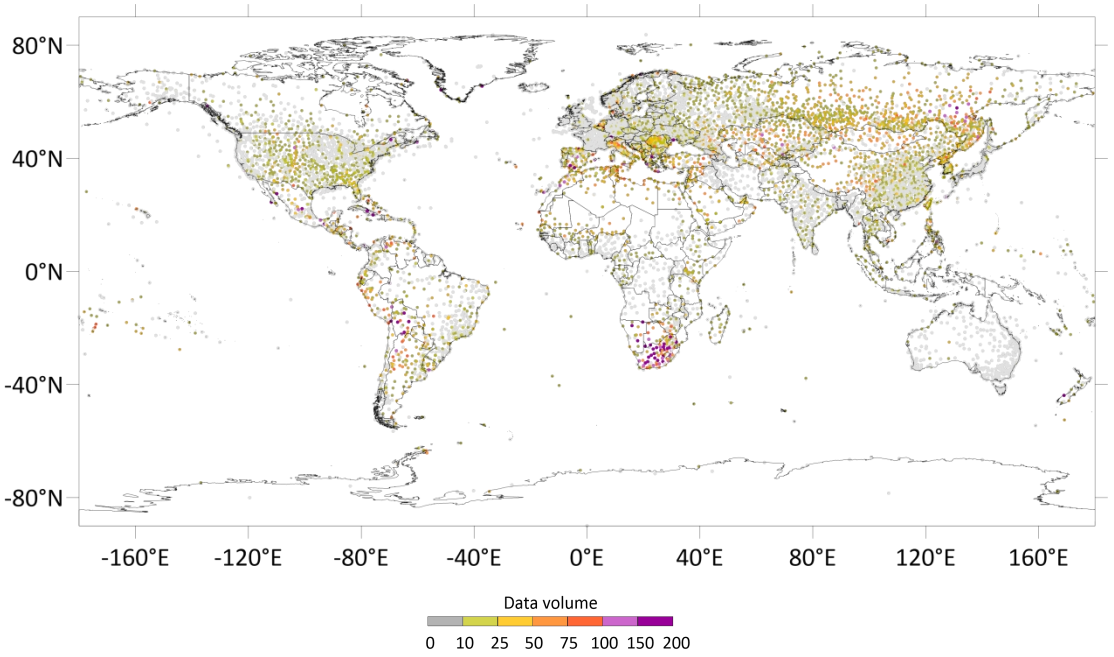



Figure R7 Similar to Figure R6, but during JJA 1981-1989.

**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 #2 Comment 32:** *Line 352: see comments against lines 302-5, and line 81, being clear here that the GSOD selects the highest/lowest from the recorded subdaily (hourly) observations I think is important.*

**Author's Response 32:** As clarified in our previous response, GSOD employs a hierarchical extraction approach rather than a mixed selection. We have objectively rewritten the text to explicitly state this methodological difference, highlighting how GSOD utilizes UTC timestamps for daily attribution.

**Changes in Manuscript:**

"First, GLBD-FED and GSOD employ different methodologies for identifying daily Tmax and Tmin values. GLBD-FED calculates daily extremes by realigning sub-daily records to their physical 24-hour occurrence windows. In contrast, GSOD derives daily extremes using a hierarchical extraction approach—first selecting explicitly reported summaries, then falling back to discrete hourly observations—but attributes them based strictly on their UTC timestamps within the calendar day. This objective methodological difference is the primary driver of the systematic discrepancies observed between the two datasets."

**Reviewer #2 Comment 33:** *Line 358: I think it would be good to include why you have selected SYNOP over METAR, and hence do the inverse of the GSOD.*

**Author's Response 33:** We appreciate the reviewer's suggestion to clarify this methodological choice. Our preference for Synoptic (SYNOP) reports over METAR is fundamentally driven by the need for global uniformity. 1) SYNOP reports are strictly

coordinated by the WMO and adhere to globally standardized protocols optimized for meteorological observation. 2) While METAR (aviation) reports serve as a primary data source in specific regions like the United States, this high density is not representative globally. 3) Because GLBD-FED is designed as a universal global product, prioritizing WMO-standardized SYNOP reports ensures much higher spatial and temporal homogeneity across different national networks.

**Changes in Manuscript:**

> "...whereas GSOD does the opposite. This prioritization in GLBD-FED is driven by the goal of maximizing global dataset uniformity. Synoptic observations are strictly coordinated by the World Meteorological Organization (WMO) and adhere to standardized global reporting protocols designed specifically for meteorological purposes. In contrast, while METAR (aviation) reports are abundant and serve as a vital data source in specific regions such as the United States, their global distribution is highly uneven. Therefore, for a universal global dataset, prioritizing WMO-standardized Synoptic reports ensures greater spatial and temporal homogeneity across different countries than relying on aviation-focused METAR data."

**Reviewer #2 Comment 34:** *Line 385: As for Figure 3/line 166, I suggest replacing "Downlimit" with something that more clearly indicates this is the daily Tmin value for these two datasets.*

**Author's Response 34:** We appreciate the reviewer's feedback, which highlights the ambiguity of our original non-standard terminology. To prevent misunderstanding and provide maximum clarity, we have replaced "Downlimit" and "Uplimit" in both the manuscript text and the relevant figures (as in response 15). Specifically, we changed them to "upper limit derived by Tmax-12h/24h" and "lower limit derived by Tmin-12h/24h" to accurately reflect the derivation source.

**Changes in Manuscript:**

> "The black and red lines represent the **lower limits derived by** GLBD-FED and GSOD, respectively..."

**Reviewer #2 Comment 35:** *Line 392: Please give the country in which this station is located at this point in the text, as you have done for the other two so far.*

**Author's Response 35:** We apologize for this omission. We have added the country name (South Africa) to the text to maintain consistency.

**Changes in Manuscript:**

> "Fig. 11 presents a comparison of daily Tmax values at the Plettenberg Bay station in South Africa (689310-99999). The abscissa and ordinate represent the daily values from the GLBD-FED and GSOD datasets, respectively..."

**Reviewer #2 Comment 36:** *Line 402: Were these likely, or potential duplicated records? Some of the red points sit on the one-to-one line so may not be erroneous duplications. There could be weather conditions which do result in identical Tmax values (even if rare)...*

**Author's Response 36:** We sincerely appreciate the reviewer's careful examination of Figure 11. You raise a very valid point: identical consecutive Tmax values can indeed physically occur due to specific stable weather conditions. Therefore, we agree that "potential duplicated records" is the more scientifically rigorous terminology, and we have updated the text accordingly. To provide a more comprehensive context, we evaluated the GLBD-FED data for this specific station, added a discussion acknowledging that the red points on the 1:1 line represent genuine stable conditions, and explained the points where GLBD-FED reports higher Tmax values (due to the physical 24-hour window capturing extremes outside the UTC boundary).

**Changes in Manuscript:**

> "The data points are further categorized into two groups based on the presence of consecutive identical values in GSOD: red points (repeated GSOD values) and black points (non-repeated GSOD values). Interestingly, approximately one-third of the Tmax daily values from GSOD this year were potential duplicated records (86 points), showing a positive mean difference of about 3.2°C relative to GLBD-FED. For context, an analysis of the GLBD-FED dataset for the same station and period reveals only 2 instances of consecutive identical values. Excluding the red points results in a roughly 70% reduction in the overall mean difference, decreasing it to 0.4°C.
>
> It is important to note the variability within this comparison. Some repeated GSOD values (red points) lie exactly on the 1:1 line of equality; these represent physically plausible weather conditions where the actual maximum temperature remained identical across consecutive days. Additionally, there are scattered points where GLBD-FED reports higher Tmax values than GSOD. These instances typically occur when GLBD-FED's physical 24-hour window captures an extreme temperature event that is otherwise split or assigned to an adjacent day under GSOD's strict UTC calendar-day boundaries. This scatter distribution visually illustrates how the distinct daily data definitions between the two datasets affect the representation of temperature records."

**Reviewer #2 Comment 37:** *Line 429: Please check the latitude and longitudes for 442920-99999. The ones presented place the station in Qingdao, Shandong Province (maybe Liuting, 548570-99999?), and the altitudes seem unlikely for Mongolia.*

**Author's Response 37:** 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 |

**Reviewer #2 Comment 38:** *Line 434: It's not clear to me from the text and caption whether these two panels show subsampled GLBD-FED data - one extracting SYNOP, the other the METAR - or whether the METAR data has been specially extracted for panel b.*

**Author's Response 38:** We appreciate the opportunity to clarify the data processing for Figure 12. Panel (a) represents the standard GLBD-FED output, which follows our default protocol of prioritizing Synoptic reports. Panel (b) is a controlled comparison where the METAR reports for the same station were specially extracted and processed using the GLBD-FED algorithm. This was done to isolate and demonstrate how the choice of sub-daily data source alone can lead to the systematic biases observed in GSOD.

**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..."

"Figure 12. Comparison of daily Tmin at Ulan Bator, Mongolia (442920-99999) between GLBD-FED and GSOD in 2024. Panel (a) compares the default GLBD-FED output (prioritizing Synoptic reports) with GSOD, while panel (b) shows a special comparison using GLBD-FED results derived exclusively from METAR reports."

**Reviewer #2 Comment 39:** *Line 439: Please give the station number in this case (as you have done with others).*

**Author's Response 39:** We apologize for this omission. We have added the station number (873280-99999) and its geographical coordinates to the text to ensure consistency.

**Changes in Manuscript:**

"The systematic discrepancies identified in Fig. 8 originate from methodological divergences between GLBD-FED and GSOD, specifically in their temporal alignment frameworks. Fig. 13(a) illustrates this mechanism through a case study at the Villa Reynolds station in Argentina

(873280-99999; 31.96°S, 65.13°W), where GSOD exhibits a systematic Tmax warm bias of +1.6°C relative to GLBD-FED..."

**Reviewer #2 Comment 40:**

*- Line 460: Please restate the averaging period at this point.*

*- Line 466: The "original" method referred to here is what was used for GSOD? If so, please make this clear.*

*- Line 468: Make it clear the increases were in the data counts, not the values.*

**Author's Response 40:** We thank the reviewer for helping us improve the precision of our concluding remarks. We have fully addressed all three points in the revised text: 1) We explicitly restated the specific 24-hour averaging period (0000–2400 UTC); 2) We clarified that the "original method" does not refer to GSOD, but rather to the conventional baseline calculation method that relies on either two consecutive 12-hour Tmax/Tmin records or a single 24-hour Tmax/Tmin record; 3) We corrected the phrasing to explicitly state that the increases refer to the data counts (volume) of valid daily records, not the temperature values.

**Changes in Manuscript:**

"1. To produce a global daily dataset representing maximum, minimum, and average temperatures over a rigorously defined 24-hour period (i.e., the standard 0000–2400 UTC day), we developed a new algorithm that decomposes sub-daily Tmax and Tmin records into finer intervals and then reaggregates them into daily extremes under physical 24-hour calculation windows. Compared to the conventional calculation method (which relied on either two consecutive Tmax/Tmin records over 12 hours or one Tmax/Tmin record over 24 hours), the new algorithm significantly increased the data counts of valid daily Tmax and Tmin records by 64% and 45%, respectively. A correction for misrecorded Tmax and Tmin records was also implemented.

2. GLBD-FED includes Tmax, Tave, and Tmin data from approximately 17,000 global sites covering the period from 1981 to 2024. The daily temperature volume of GLBD-FED increased from 3,000 records per day in the 1980s to 10,000 records per day in the 2020s. America and Asia show high spatial densities of daily temperature data, especially in recent years. In comparison to GSOD, Tmax and Tmin from GLBD-FED exhibit less extreme daily values, with slightly lower daily Tmax (approximately -0.30°C) and higher daily Tmin (approximately +0.30°C), resulting in nearly the same daily Tave (around +0.10°C). These differences are primarily attributed to the diversity in daily data definitions and the choice of data sources."

**Reviewer #2 Comment 41:** *Line 497-521: It could be useful to describe in words the behaviours that each test is looking to identify.*

**Author's Response 41:** We agree with the reviewer that adding textual descriptions improves the readability of the appendix. We have supplemented the mathematical formulas with concise textual explanations for each of the five quality control tests (including the Stuck test). These descriptions explicitly and objectively state the basic logic and the type of erroneous data each test is designed to identify.

**Changes in Manuscript:**

"**Data Quality Tests**

**Repeat test**

*This test identifies instances where identical temperature values are recorded consecutively over a specified number of days, which typically indicates sensor malfunction or reporting errors.*

$$\mathrm{qc} = \begin{Bmatrix} wrong, \sigma = 0 \\ credible, else \end{Bmatrix}$$

$$\sigma = \mathrm{std}(\mathrm{x}_{i-1}, \mathrm{x}_i, \mathrm{x}_{i+1})$$

$$\mu = \mathrm{mean}(\mathrm{x}_{i-1}, \mathrm{x}_i, \mathrm{x}_{i+1})$$

*where $\sigma$ and $\mu$ are the standard deviation and smoothing average of 3 days measurements, respectively.*

**Spike test**

*This test identifies anomalous daily temperature values that exceed predefined absolute physical boundaries or historical climatological limits for a specific station.*

$$\mathrm{qc} = \begin{Bmatrix} credible, x_{lowerlimit} \le x_i \le x_{upperlimit} \\ wrong, x_i \ge x_{upperlimit} \ \ or \ x_i \le x_{lowerlimit} \end{Bmatrix}$$

$$\mathrm{x}_{upperlimit} = \min[\max(\bar{x}) + 5℃, 80℃]$$

$$\mathrm{x}_{lowerlimit} = \max[\min(\bar{x}) - 5℃, -80℃]$$

*Where the subscript i stands for the measurement on the i-th day; $x_{upperlimit}$ /$x_{lowerlimit}$ is the upper/lower threshold value for the record and is the smaller/higher value between $max(\bar{x}) + 5℃$ and $80℃$, where $\bar{x}$ represents the subset of the historic measurements in the month (Jan, Feb,…Dec) which removes the smallest and largest 1% of values.*

**Inner consistency test**

*This test verifies the basic logical relationship among the three daily temperature variables. It ensures that the daily maximum temperature is greater than or equal to the daily average temperature, which in turn must*

*be greater than or equal to the daily minimum temperature.*

$$\mathrm{qc} = \begin{Bmatrix} credible, Tmax \geq Tave \geq Tmin, \\ wrong, Tmax \leq Tave \leq Tmin \\ doubtful, else \end{Bmatrix}$$

*Where the Tmax, Tave, Tmin stand for the daily Tmax, Tave, Tmin data, respectively*

**Temporal consistency test**

*This test detects daily records that deviate excessively from the historical statistical distribution of a given station, evaluated using the median and standard deviation.*

$$\mathrm{qc} = \begin{Bmatrix} credible, x_i \leq \mu + 2.5\sigma \\ wrong, x_i > \mu + 5\sigma \\ doubtful, \mu + 2.5\sigma < x_i \leq \mu + 5\sigma \end{Bmatrix}$$

$$\mu = \mathrm{median}(\bar{x})$$

$$\sigma = \mathrm{std}(\bar{x})$$

*Where the $\mu$ and $\sigma$ are the median value and standard deviation of $\bar{x}$, respectively.*

**Spatial consistency test**

*This test identifies records that exhibit significant discrepancies when compared to concurrent observations from neighboring stations within the same region.*

$$\mathrm{qc} = \begin{Bmatrix} credible, \delta_i \leq \min\left[\bar{\delta} + 2.5\nabla\delta, 10℃\right] \\ doubtful, \min\left[\bar{\delta} + 2.5\nabla\delta, 10℃\right] < \delta_i < \min\left[\bar{\delta} + 3.5\nabla\delta, 15℃\right] \\ wrong, \delta_i \geq \min\left[\bar{\delta} + 3.5\nabla\delta, 15℃\right] \end{Bmatrix}$$

$$\delta_i = \sqrt{\frac{\sum_{j=1}^{m} (x_j - x_i)^2}{m}}$$

$$\bar{\delta} = median(\delta_i)$$

$$\nabla\delta = \mathrm{std}(\delta_i)$$

*Where the subscript j stands for the j-th neighbouring sites (within the 100km radius around the candidate site) and m is the total number of neighbouring sites.*