Point-by-Point Response to Reviewer #3

Dear Reviewer #3,

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 #3

Reviewer #3 Comment 1: *This paper presents the development of a global in situ daily temperature dataset from 1981-2024... I am struggling to understand its purpose. For monitoring (including for extremes) and climate change applications it might not be fit for purpose and to test that you would need to assess the homogeneity of the data or do some comparisons with 'high quality' temperature data among other things. Comparing GLBD-FED with GSOD tells you about potential biases... but doesn't tell us anything really about the quality of GLBD-FED for the potential applications... Also monitoring applications will be limited as the dataset is static and it was unclear if the authors' intention was ultimately to operationalise the dataset... The authors should clarify its value and purpose and perhaps include some basic evaluation on the quality of the dataset for end user applications.*

Author's Response 1: We sincerely appreciate your thorough, highly constructive, and critical evaluation of our work. Your comments cut to the core of the dataset's identity, prompting us to critically re-evaluate and more accurately define the specific scope, intended applications, and structural limitations of GLBD-FED. We fully agree with your perspective: without rigorous statistical homogenization, a dataset should not be directly used for long-term decadal climate change detection.

To comprehensively address your concerns regarding the dataset's purpose, evaluation, and static nature, **we have completely rewritten the Introduction section and added critical clarifications to the Discussion section**. We detail our responses to your specific points below:

1. Clarifying the Dataset's Positioning: Your comment highlights a crucial distinction in climate data taxonomy. We have now explicitly clarified that the primary focus of GLBD-FED is to serve as a global, near-real-time, first-hand daily dataset. We agree that products like GHCNd are benchmark datasets designed for climate change detection. In contrast, GLBD-FED is positioned at the foundational tier. Its primary

scientific purpose is to solve the fundamental temporal aggregation problem directly from first-hand sub-daily reports. By providing a structurally sound "first estimate" without prominent algorithmic artifacts, it meets the immediate need for rapid extreme weather monitoring and serves as the foundational raw material for future benchmark homogenization.

2. The Value of Comparing with GSOD: You rightly noted that GSOD is not a homogenized benchmark product. However, it remains one of the most widely used first-hand baseline datasets globally. The purpose of our comparison is not to benchmark GLBD-FED against a gold standard, but rather to isolate and demonstrate the sheer magnitude of errors introduced *solely* by temporal aggregation artifacts in real-time operational streams. By comparing two datasets derived from the exact same source (ISD), we prove that fixing temporal aggregation removes significant artifactual extremes.

3. Addressing the "Static" Nature: As our methodology is specifically designed for near-real-time operationalization, your concern regarding the currently static nature of the dataset following the retirement of the ISD archive is highly pertinent. We have added a dedicated, transparent explanation in the Discussion section. Following the retirement of ISD, we evaluated its successor, NOAA's GHCN-Hourly (GHCNh). However, our assessment revealed a critical limitation: GHCNh lacks the necessary explicit sub-daily extreme reports (12h/24h summaries) required to robustly support our temporal reconstruction algorithm. Consequently, real-time data streaming is temporarily hindered. Nonetheless, the completed 1981-2024 GLBD-FED archive stands as a high-quality 44-year historical baseline of first-hand data.

Changes in Manuscript:

> "High-quality daily temperature datasets generally fall into two distinct operational tiers: homogenized benchmark datasets (such as GHCNd) designed for long-term climate change detection, and near-real-time, first-hand datasets designed for rapid synoptic monitoring and immediate extreme weather assessment. While benchmark datasets are essential for climatology, they often involve significant latency. Conversely, legacy near-real-time datasets (such as GSOD) often exhibit temporal aggregation artifacts—such as misallocating sub-daily extremes to incorrect calendar days due to strict UTC boundary constraints. These algorithmic characteristics introduce artificial noise that affects rapid extreme weather analysis. Therefore, the primary focus of GLBD-FED is to serve as a high-fidelity, global, near-real-time, first-hand daily dataset. By strictly resolving these sub-daily aggregation issues, GLBD-FED provides a structurally sound 'first estimate' baseline that supports reliable, large-scale extreme event monitoring and serves as temporally consistent raw material for future benchmark homogenization."
>
> "**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 #3 Comment 2: *L32: What do you mean by 'temperate performance'?*

Author's Response 2: We apologize for this confusing phrasing. By "temperate performance," we intended to describe that GLBD-FED exhibits "less extreme values" or a "more moderate diurnal temperature range" compared to GSOD. Because GLBD-FED strictly aligns observations to the actual 24-hour physical occurrence window, it effectively avoids artificially inflating Tmax or deflating Tmin (which frequently happens in GSOD when extreme temperatures from adjacent days are erroneously captured due to rigid UTC boundary splits). To eliminate any ambiguity, we have replaced "temperate performance" with "less extreme daily values" in the revised manuscript.

Changes in Manuscript:

"In comparison to GSOD, Tmax and Tmin from GLBD-FED exhibit **less extreme daily values**, with slightly lower daily Tmax (approximately -0.3°C) and higher daily Tmin (approximately +0.3°C), resulting in nearly the same daily Tave (around +0.1°C)."

Reviewer #3 Comment 3: *L49: I believe the correct acronym is 'National Meteorological and Hydrological Services (NMHSs)'*

Author's Response 3: We thank the reviewer for pointing out this oversight. You are absolutely correct. We have updated the text to use the accurate and standard acronym throughout the revised manuscript.

Changes in Manuscript:

"...National Meteorological and Hydrological Services (NMHSs)..."

Reviewer #3 Comment 4 & 5:

-L49-54. There is strange referencing here. The start of your intro is about the importance of daily in situ temperature measurements but here you reference a mixture of global, regional, daily, monthly and precipitation datasets with many of them very old references. I think you really have a chance to say here that there are really very few daily global temperature datasets based on in situ data available... This combined with improved accessibility of the data, is a major gap you are filling with GLBD-FED.

-L64: I don't know why you start to talk about precipitation here because you have not specifically mentioned it earlier and actually it would probably be best just to stick to talking about temperature datasets and gaps.

Author's Response 4 & 5: We completely agree with your observation, and we are extremely grateful for this excellent framing suggestion. You are absolutely right that our original references were overly broad and the abrupt transition to discussing precipitation was distracting. There is indeed a distinct scarcity of daily global temperature datasets based purely on *in situ* station data, especially when compared to the robust development of precipitation products.

As part of the complete rewrite of our Introduction, we have enthusiastically adopted your narrative. We entirely removed the disconnected discussion and thoroughly cleaned up the bibliography. We now explicitly highlight this major gap: while the community has access to numerous robust precipitation products (both in situ gridded analyses like GPCC and multi-source merged products like MSWEP) and heavily homogenized climate benchmarks (like GHCN-Daily, BEST, and HadGHCND), a high-quality, pure *in situ* station-based, temporally aligned daily temperature dataset has remained a critical missing piece. We have updated the text to emphasize that GLBD-FED is specifically designed to fill this gap.

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.

...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).

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.

Reviewer #3 Comment 6: *L73: OK so you're going to produce a global daily temperature dataset but it is still unclear what applications it could be used for, especially since there are known quality issues with GSOD which you are using for your comparison data.*

Author's Response 6: We fully understand this concern. As we thoroughly addressed in our response to your General Comment, comparing GLBD-FED with GSOD does not inherently prove its suitability for long-term climate change detection, precisely because GSOD itself is not a homogenized benchmark dataset.

Instead, our comparison isolates and proves the value of GLBD-FED's core innovation: fixing the temporal aggregation algorithm. Because both datasets are derived from the exact same source (ISD), comparing them demonstrates how correcting sub-daily aggregation issues eliminates the artifactual extreme values commonly found in legacy datasets. Regarding its exact applications, we have explicitly clarified in our fully rewritten Introduction and Discussion sections that GLBD-FED is designed for: 1) Evaluating historical regional synoptic weather events and validating numerical weather prediction (NWP) models, where temporally accurate, first-hand data is required. 2) Serving as a structurally sound, temporally aligned "first estimate" baseline that provides a much cleaner raw input for future statistical homogenization efforts.

Reviewer #3 Comment 7: *L85: ISD has now been superseded by the Global Historical Climatology Network hourly (GHCNh) dataset... However, this at least needs to be acknowledged somewhere as it would also affect the operationalization of GLBD-FED if that was indeed the intention.*

Author's Response 7: We completely agree with your observation and appreciate your understanding of the dataset development timeline. The transition from ISD to GHCNh is indeed a major event that directly impacts the ongoing operationalization of GLBD-FED.

As we discussed extensively in our response to your General Comment, we have actively evaluated GHCNh for potential operational updating. Unfortunately, the meteorological elements provided in GHCNh have been significantly reduced compared to ISD. Specifically, it lacks the explicit sub-daily extreme reports (e.g., the 12h/24h summaries) that are strictly required to robustly support our temporal reconstruction algorithm. Consequently, real-time operationalization is currently paused.

Reviewer #3 Comment 8: *L91: It's unclear why you chose 2011-2024 to produce Figure 1. Does the figure change if you include different time periods/the whole dataset period?*

Author's Response 8: We fully agree with your critique. To directly answer your question: yes, the spatial density and data volume do change significantly over the dataset's history, and the exclusive use of the 2011-2024 period in the original manuscript failed to capture this historical evolution.

To accurately reflect the full scope of the dataset, **we have completely redesigned Figure 1**. The revised Figure 1 no longer restricts the view to recent years. Instead, it now comprehensively displays the total data volume for the entire 1981-2024 period. Furthermore, to explicitly show how the spatial coverage changes over time, we have expanded Figure 1 to include 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 new figures clearly reveal that while hourly observations demonstrate continuous growth, the explicitly reported extreme summaries suffer from geographic fragmentation and decadal volatility. We explicitly utilize these comprehensive long-term distributions to quantitatively justify the necessity of our secondary fallback strategy.

Changes in Manuscript:

"Figure 1 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 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."

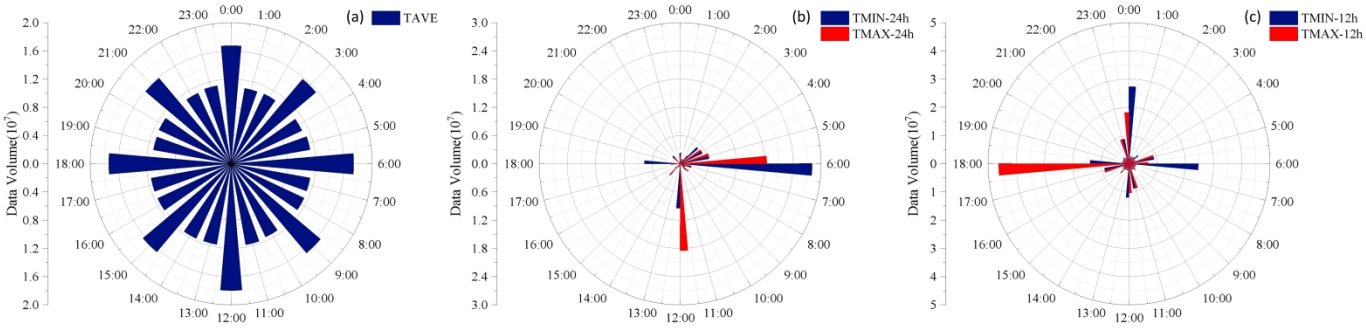
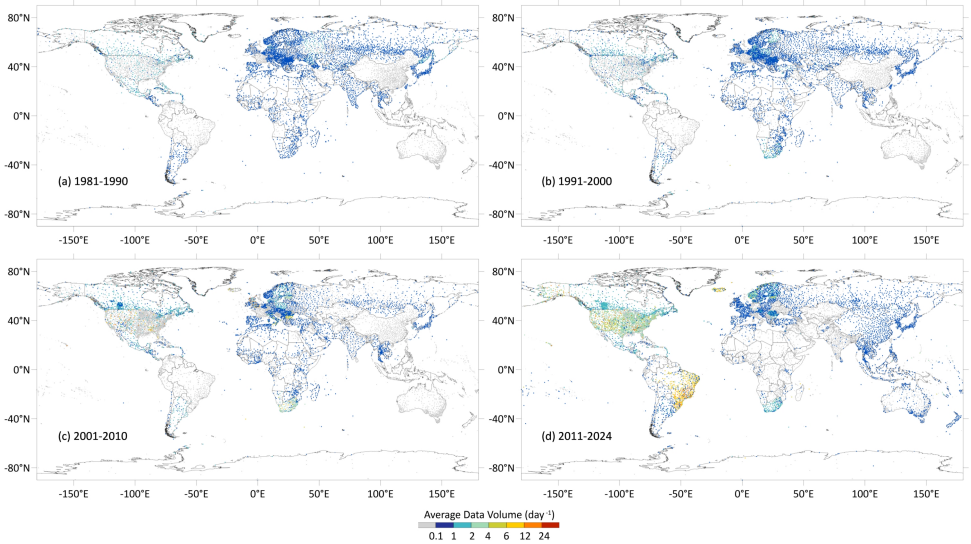


Figure 1-T1 The distribution of sub-daily temperature data amounts at each o’ clock during 1981-2024



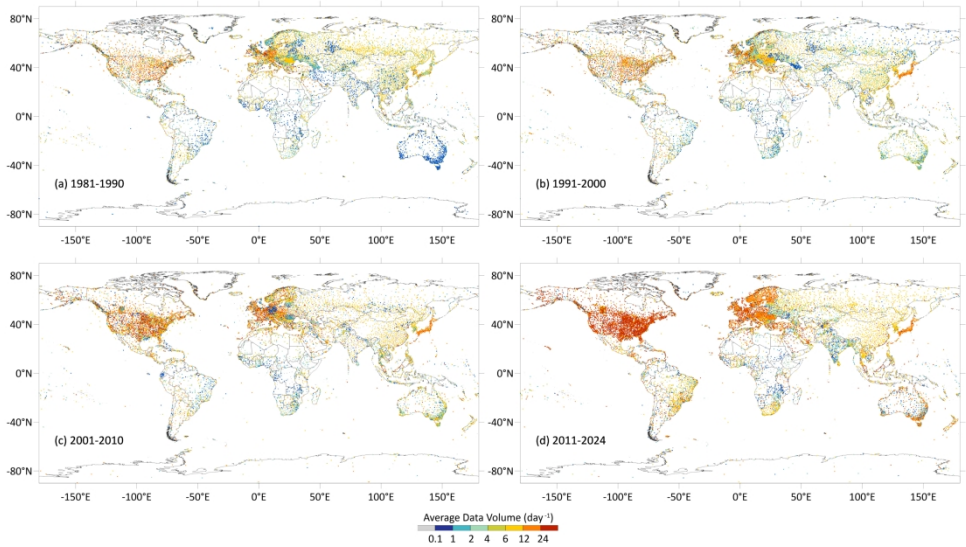


Figure 1-S1 Spatial distributions of Hourly average data volume per day for hourly Tave from ISD. Panel a,b,c,d stand for the results 1981-1990, 1991-2000, 2001-2010, 2011-2024.

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

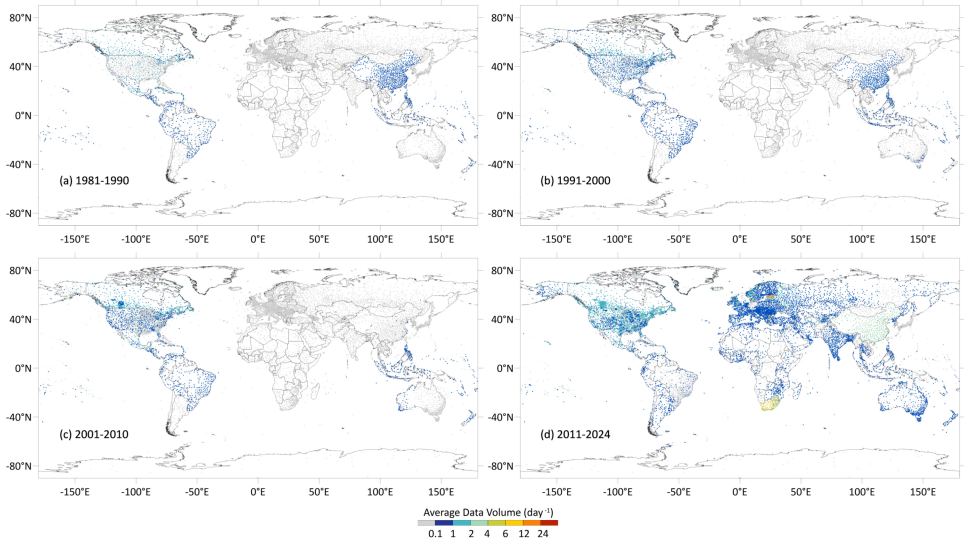


Figure 1-S3 similar to Figure 1-S1, but for Tmax-24h

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

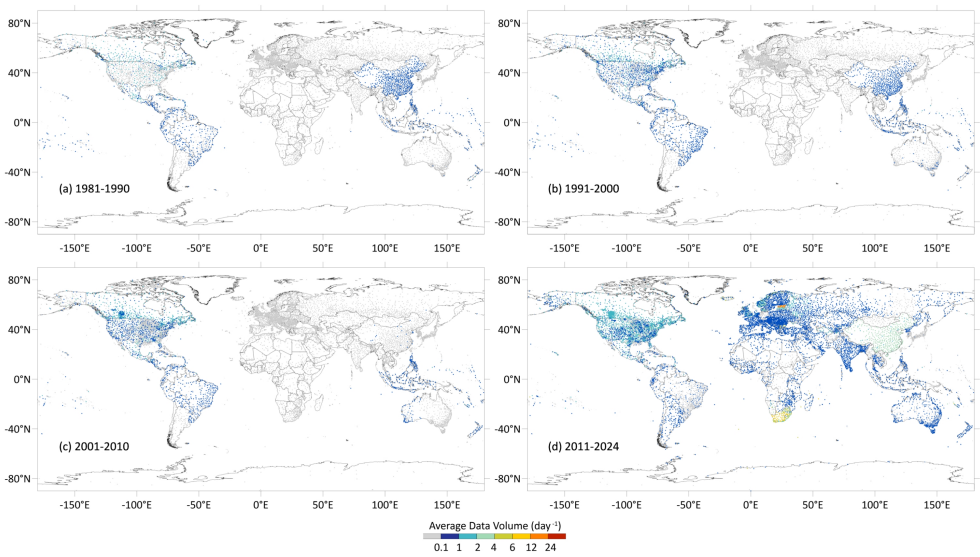


Figure 1-S5 similar to Figure 1-S1, but for Tmin-24h

Reviewer #3 Comment 9: *L110: display weaker temporal regularity – weird phrasing*

Author's Response 9: We agree with the reviewer that this phrasing is awkward and non-standard. By "weaker temporal regularity," we originally intended to describe that these sub-daily observations do not adhere to a fixed, standardized schedule, often resulting in fluctuating or uneven measurement intervals across different global stations. To correct this and eliminate ambiguity, we have replaced the phrase with the more precise and standard terminology.

Changes in Manuscript:

> "...sub-daily observations **exhibit irregular observation intervals**..."

Reviewer #3 Comment 10: *L119: The hyperlink doesn't work for me.*

Author's Response 10: We sincerely apologize for this inconvenience. The original link provided in the manuscript has indeed become inactive, likely due to NOAA's recent website restructuring. We have successfully located the new, active repository where NOAA currently hosts the historical GSOD archive data. We have updated the hyperlink in the revised manuscript to point to the correct directory.

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 #3 Comment 11: *L148: I really appreciate the effort here to actually try and correct erroneous data rather than simply flagging it. However, I assume that you flag that it has had a correction applied?*

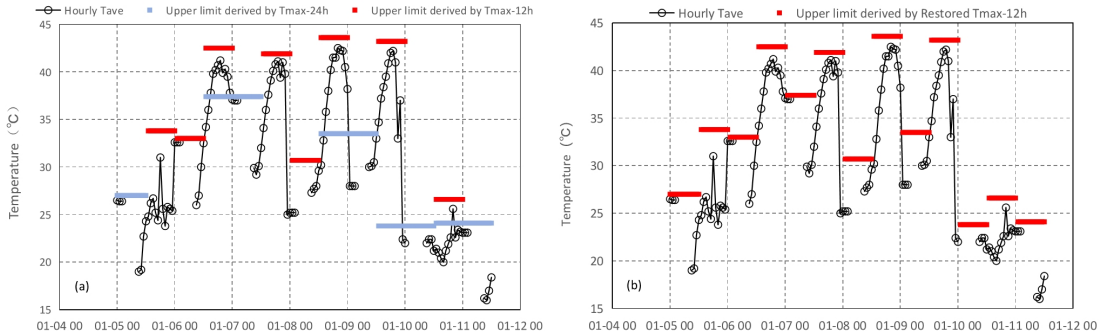
Author's Response 11: We sincerely appreciate your positive feedback on our algorithmic correction efforts. You have raised an excellent point regarding the necessity of data provenance; flagging corrected sub-daily records is indeed highly valuable for users who wish to trace the source data.

During the development of the dataset, our team carefully deliberated on this exact issue. Because GLBD-FED is fundamentally a *daily* temperature dataset, the Quality Control (QC) flags currently configured within the product are strictly designed to represent the quality and confidence level of the final *daily* aggregated values. We ultimately concluded that introducing additional flags to describe the correction history of the underlying *sub-daily (source)* data might create confusion for end-users who are primarily focused on the daily outputs. Consequently, we did not include source-correction flags in the current baseline version.

However, we fully recognize the value of your suggestion for future applications. When we upgrade GLBD-FED to its next operational version using new upstream data streams, we will implement your recommendation by introducing a dedicated metadata field or flag to explicitly indicate whether the underlying source data was algorithmically corrected, alongside the primary daily QC flags.

Reviewer #3 Comment 12: *L168: Do you mean 'upper limit of temperature' here?*

Author's Response 12: Yes, you are exactly right. We confirm the refers to the physical or climatological "upper limit of temperature" used as a threshold boundary in our data quality control procedures. We appreciate your careful reading to verify this technical detail. 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").



"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 #3 Comment 13: *L183: I couldn't work out how you deal with missing hours or whether this is a problem.*

Author's Response 13: We apologize for the lack of clarity regarding the calculation of the daily average temperature (Tave) in Section 4.2.2. You raise an essential point: irregular reporting or missing hourly observations can indeed introduce temporal sampling biases if not handled properly.

To effectively mitigate this problem, our algorithm utilizes a strict data completeness threshold before computing the daily average. Specifically, Tave is only calculated if **there are at least 4 valid hourly observations relatively evenly distributed across the 24-hour period**. If the available hourly data for a given day falls below this completeness threshold, the Tave is not computed and is simply marked as a missing value. This rigorous check ensures that all calculated daily averages are structurally representative of the entire 24-hour window, thereby resolving the risk of biases caused by excessive missing hours. We have added a brief clarification regarding this rule in the manuscript.

Changes in Manuscript:

> "To mitigate temporal sampling biases introduced by missing or irregular hourly observations, the daily average temperature (Tave) is only computed when **there are at least 4 valid hourly observations relatively evenly distributed across the 24-hour period**. Days failing to meet this data completeness threshold are recorded as missing."

Reviewer #3 Comment 14: *L225: So it does look as if you improve the data volume in America, Australia, western Europe, and southern Asia but for e.g. North America and Australia there are already high quality, long-term daily temperature records that are readily accessible so it's unclear what value additional daily data in GLBD-FED would bring. This brings me back to my general comment about the ultimate purpose of the dataset.*

Author's Response 14: We fully agree with the reviewer that regions like North America and Australia possess exceptionally high-quality, long-term homogenized national climate records (such as GHCNd in the US or ACORN-SAT in Australia). As established in our response to your General Comment, the primary goal of GLBD-FED is *not* to compete with or replace these regional benchmark datasets for localized long-term climate trend analysis.

Instead, the distinct value that GLBD-FED brings to these regions—and to the globe as a whole—is **strict global methodological uniformity**. Regional or national datasets frequently employ diverse definitions of a "daily" period (e.g., midnight-to-midnight local time versus varying morning observation times) and utilize different temporal aggregation algorithms. This lack of standardization introduces known "Time of

Observation Biases" (TOB; Karl et al., 1986). If researchers attempt to combine high-quality local datasets for global monitoring, these methodological differences introduce severe artificial discontinuities ("seams") across national borders.

By applying a single, unified algorithmic framework to raw, first-hand synoptic reports across all global regions simultaneously, GLBD-FED eliminates these methodological borders. Even in data-rich areas like North America and Australia, GLBD-FED provides a globally consistent "first estimate" that relies on the exact same physical 24-hour window as stations in Asia or Europe. This seamless global consistency is highly critical for applications such as evaluating massive-scale synoptic weather systems and strictly validating global Numerical Weather Prediction (NWP) **model outputs and forecasts**. *(As the overarching purpose and end-user applications of the dataset have been thoroughly clarified in the fully rewritten Introduction and Discussion sections, no additional text specifically targeting Line 225 was deemed necessary).*

Reviewer #3 Comment 15: *L239: The spike test seems to have a fixed threshold which may not be appropriate for regions of very low or high variability. I guess in combination with the other tests this could be accounted for but I highlight it as a possible limit to the qc checks.*

Author's Response 15: We fully agree with your highly professional assessment. Your suggestion to implement region-specific thresholds that account for varying spatial climatological variability is both scientifically rigorous and practically feasible.

As you correctly inferred, in the current version of GLBD-FED, the fixed-threshold spike test is intentionally designed as a preliminary "coarse filter." Its primary purpose is to quickly flag and eliminate grossly unphysical sensor or transmission errors (e.g., hardware short-circuits reporting physically impossible extreme values). The nuanced detection of anomalies in regions of varying natural climatological variance is currently handled by the subsequent, dynamic QC checks—specifically, the temporal consistency test (evaluating against local historical standard deviations) and the spatial consistency test (evaluating against regional neighbor observations).

However, we deeply appreciate and acknowledge your constructive advice regarding the limitations of a static threshold. While the spike test currently serves its purpose as a coarse filter, we recognize the significant value of enhancing its precision. In our subsequent work to upgrade the GLBD-FED quality control system for future operational versions, we will adopt your recommendation to explicitly incorporate spatial attributes into the spike test thresholds. This will meaningfully improve the spatial adaptability of our QC algorithms right from the initial screening layer.

Reviewer #3 Comment 16: *L293-299: I think this paragraph missed some copyediting e.g. “comparted”, “manifest”, “mean”, “remarkable”. Please check typos, grammar*

and style. In addition, why do you only compare "nearly all records" and not "all records". Which records did you not compare and why?

Author's Response 16: We sincerely apologize for the typographical errors and the confusing phrasing in this paragraph. You have raised an excellent point.

To answer your question regarding "nearly all records": our phrasing was indeed ambiguous. To ensure a scientifically rigorous, apples-to-apples comparison, we strictly evaluated only the spatiotemporal **intersection** of the two datasets. The few records that were not compared were simply those that existed in one dataset but lacked a matched counterpart (i.e., same station, same day) in the other.

To address this and correct all the grammatical and stylistic issues you rightly highlighted, we have thoroughly rewritten this entire paragraph. The revised text explicitly clarifies the intersection logic and corrects the language.

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.** It is clear 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 #3 Comment 17: *Figure 8: I note a trend in the bias for Tave through time. Do you have an explanation for this? I also note what looks like there is a seasonal cycle in the biases for all temperature aspects which you don't mention and needs to be discussed.*

Author's Response 17: We sincerely appreciate your meticulous examination of Figure 8. You have accurately identified two highly important temporal features in the bias series: the long-term downward trend in the daily average temperature (Tave) bias, and the pronounced seasonal cycles across the variables prior to 1990. We fully agree that these patterns require explicit discussion in the text, as they fundamentally stem from the same historical data quality issues and algorithmic limitations within the GSOD processing chain.

Our in-depth investigation reveals that both the early seasonal fluctuations and the long-term Tave trend (dropping from approximately +0.25°C in the 1980s to roughly +0.10°C in the recent decade) are intrinsically linked.

To explicitly validate the causes of the seasonal cycle, 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.

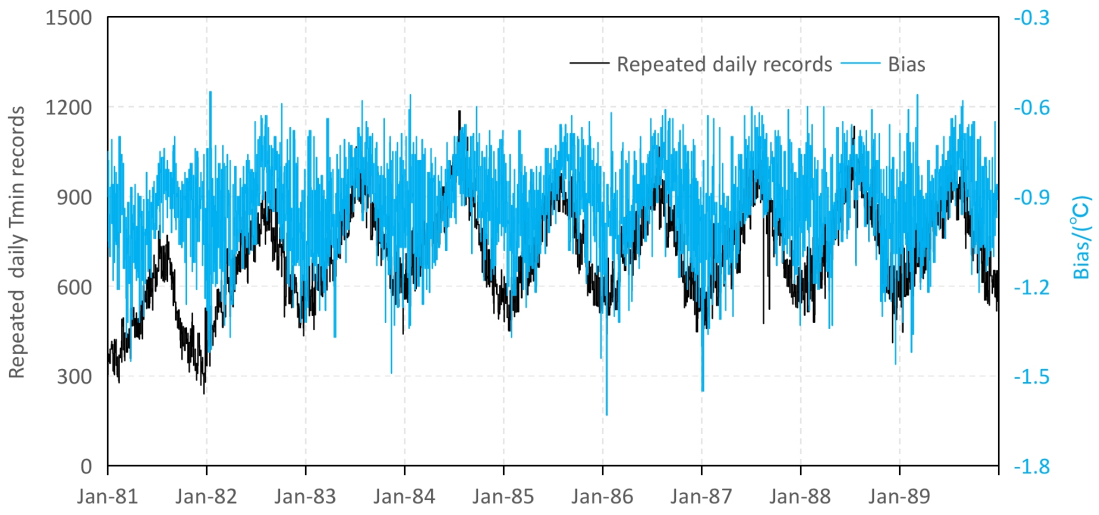


Figure R1 Repeated daily Tmin records number from GSOD (black line) and the bias

between them and GLBD-FED (blue line) during 1981-1990.

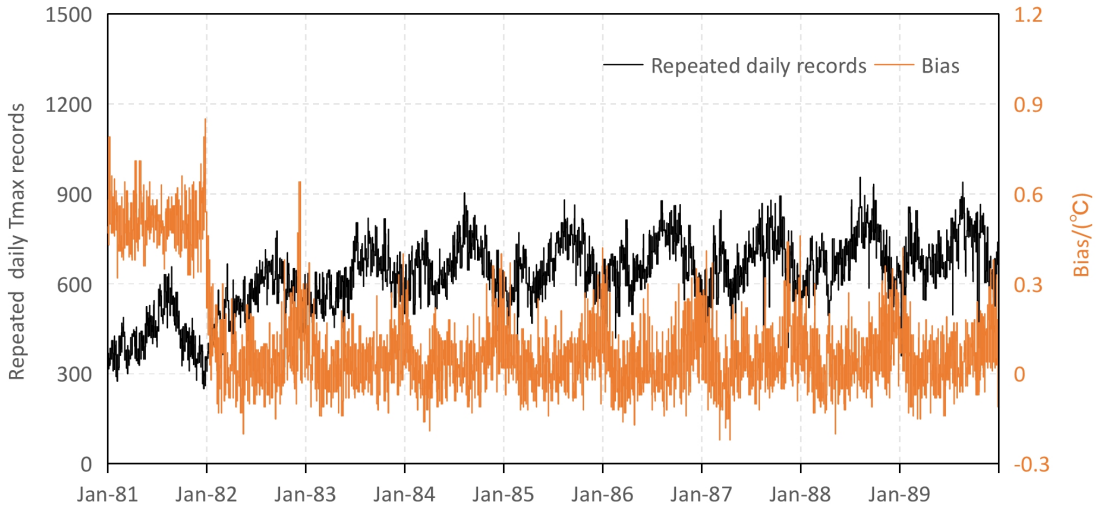


Figure R2 Similar to Figure R1 , but for daily Tmax.

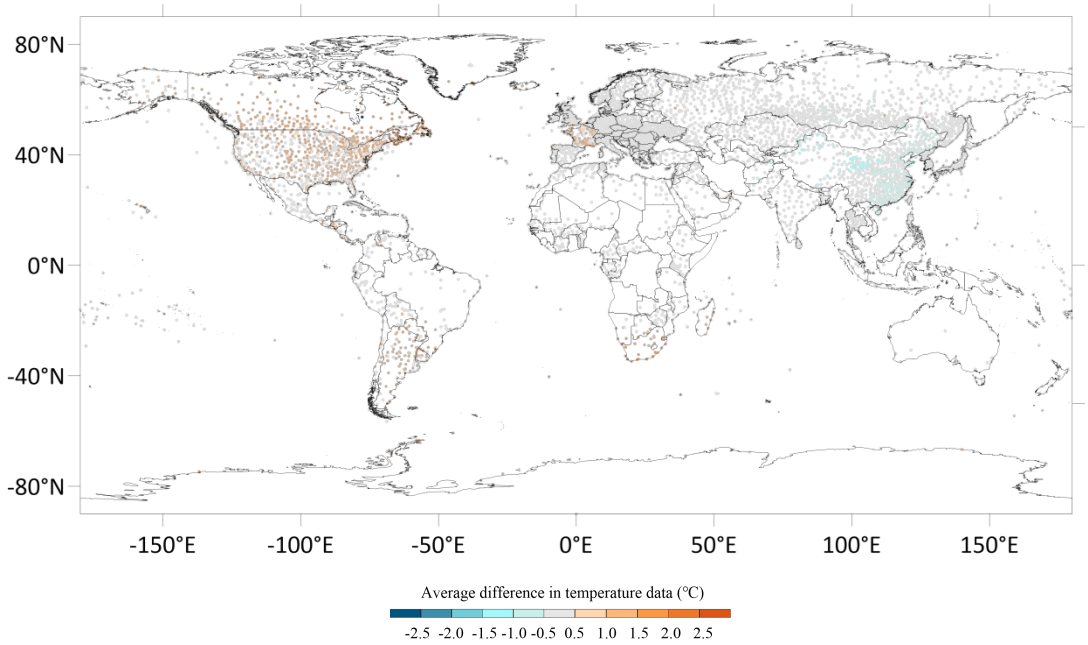


Figure R3 spatial distribution of Tmax bias between GLBD-FED and GSOD in JJA (1981-1989)

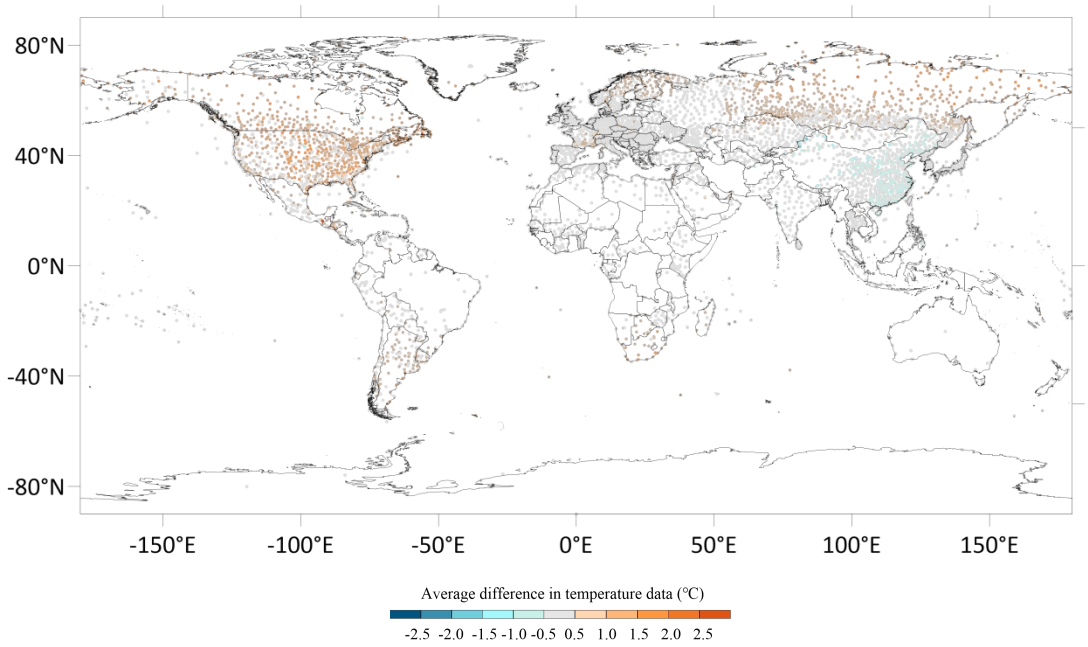


Figure R4 Similar to Figure R3 but for DJF (1981-1989)

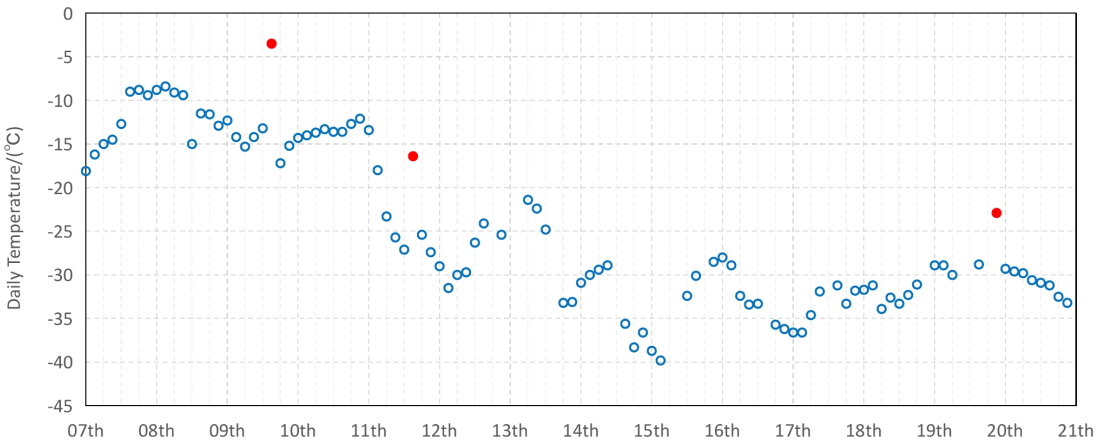


Figure R5 The hourly Tave data from the Russia site (254480-99999) during 7th to 21th Dec 1989. The cycles represent the hourly Tave and the sudden jumping ones were signed by solid red.

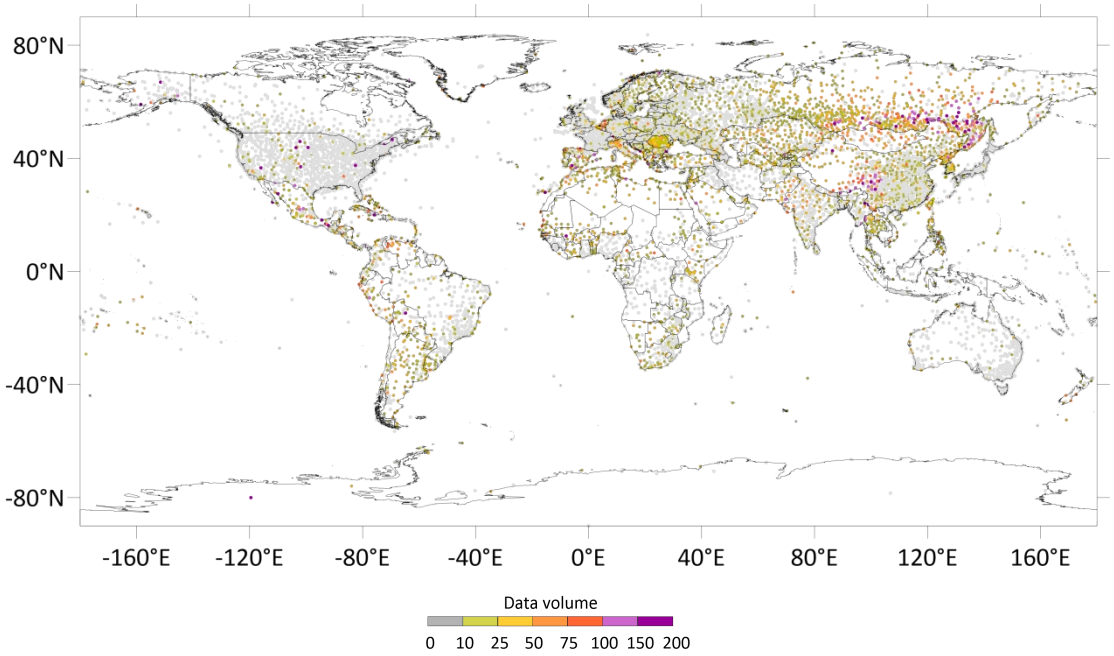


Figure R6 Spatial distribution of the outlier volume of hourly Tave at during DJF 1981-1989

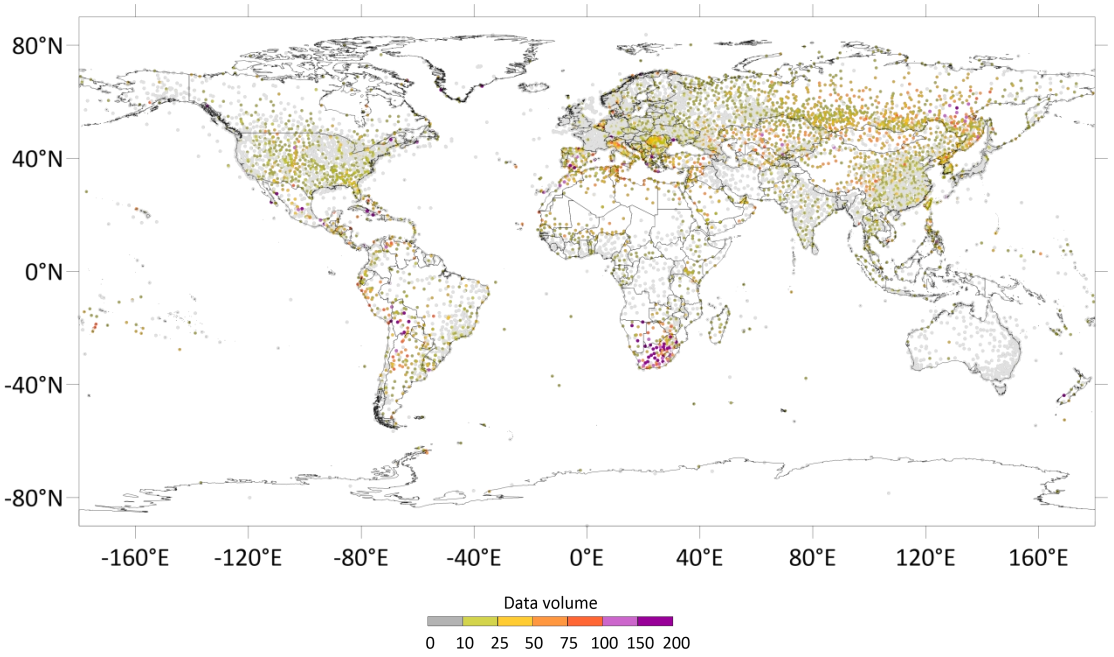


Figure R7 Similar to Figure R6, but during JJA 1981-1989.

2. The Long-Term Trend in Tave Bias:

Because GSOD overestimated Tmax in the early years due to the aforementioned sensor noise and algorithmic limitations, this directly inflated the overall daily average temperature (Tave) calculation for that era (reaching ~+0.25°C bias).

Over the past 44 years, as the global observing network progressively modernized—transitioning to automated weather stations (AWS) with improved sensor stability and enhanced WMO data transmission protocols—the frequency of raw data errors and anomalous spikes dropped significantly. Consequently, GSOD's exposure to selecting these artifacts naturally decreased. As the raw data became

cleaner, the systemic Tave bias gradually narrowed down to its current plateau of approximately +0.10°C. In contrast, GLBD-FED's algorithm remained stable throughout, as our strict temporal consistency checks successfully filtered out these anomalous spikes even in the noisy early era.

We have explicitly added this integrated explanation to the revised manuscript to address both the seasonal cycles and the long-term trend.

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."