Dear Reviewer 3,

We sincerely thank you for taking the time to review our manuscript and for providing thoughtful and constructive comments. Your feedback is greatly appreciated and has been very helpful in improving the clarity, transparency, and scientific quality of our work. In the following, we provide a detailed, point-by-point response to your suggestions. For clarity, the original comments are presented in **black**, while our responses are shown in **blue**. Sentences that are intended as revisions or additions to the manuscript are highlighted in **gold** and quotation marks, and will be formally incorporated into the revised version after the open discussion phase.

Comment 1:

Regarding the selection of only 2 out of 12 field stations from the hourly land–atmosphere interaction dataset (Ma et al., 2024) for validation after quality control, it is unclear why the other stations were excluded. Please clarify the quality control criteria and explain whether the other stations were omitted due to poor data quality or other reasons.

Response: We greatly appreciate the reviewer's astute question, which provides us with a valuable opportunity to clarify our data processing workflow and underscore the robustness of our validation approach. The situation described stemmed from an initial technical oversight that was subsequently rectified through a comprehensive data collection effort. Please allow us to explain in detail.

(1) Initial Processing Oversight

We sincerely apologize for the lack of clarity in our original manuscript. During the initial data processing phase, an error in our automated script incorrectly led us to believe that only two stations from the Ma et al. (2024) dataset (NAMORS and Arou) had successfully passed our quality control (QC) procedures and were available for use. This was an unintentional technical mistake on our part.

(2) Proactive Expansion of Validation Data

To ensure the most robust validation possible, we were not satisfied with the limited number of stations initially retained. We therefore proactively sought out and incorporated every available source of ground observation data from the region. This extensive search enabled us to integrate additional datasets, including 18 stations from the HiWATER network (Liu et al., 2018; Che et al., 2019) and 11

individual stations from other published studies (Zhang, 2018a,b; Gao, 2018; Luo, 2019; Ma, 2018; Wang and Wu, 2019; Luo and Zhu, 2020; Meng and Li, 2023). Through this expansion, our validation pool ultimately increased to a total of 31 station records from these combined sources.

(3) Discovery of Station Overlap and Final Validation Set

Upon integrating and cross-referencing all 31 records, we identified that five of the stations from our additional sources were duplicates of stations already contained within the Ma et al. (2024) dataset. In other words, a number of these stations represent the same physical locations that have been reported across different publications. For example, Arou, Yakou, Jingyangling, and Dashalong (from Ma et al., 2024) are also part of the HiWATER network; the QOMS station (cited from Ma, 2018) is included in the Ma et al. (2024) dataset; and the Maqu station (cited from Meng and Li, 2023) is likewise present in the Ma et al. (2024) dataset.

(4) Conclusion Regarding Data Quality and Selection

Therefore, to directly address the reviewer's question: the other stations from Ma et al. (2024) were not excluded due to poor data quality, but rather because of a technical error in our processing script. Some of these stations were later indirectly included through overlap with other published datasets (e.g., Arou, QOMS, Maqu). As a result, our final validation dataset comprises 31 stations from multiple independent sources, which we believe is sufficiently comprehensive to ensure the robustness and representativeness of the evaluation. We sincerely thank the reviewer again for prompting this important clarification.

We acknowledge that the original wording in Section 2.2.2 Literature-based datasets from the National Tibetan Plateau Data Center could be misleading. The phrase "(1) a publicly available dataset of hourly land–atmosphere interaction observations from 12 field stations (Ma et al., 2024), covering the period 2005–2021, from which 2 stations were selected after quality control for use as independent validation sites." may have unintentionally implied that the other 10 stations were excluded due to poor data quality, which was not the case. To avoid such ambiguity, we have revised the sentence to:

"(1) a publicly available dataset of hourly land–atmosphere interaction observations (Ma et al., 2024) , covering the period 2005–2021, from which 2 stations were used as independent validation

sites;”

Comment 2:

The quality of the figures is suboptimal. Several figures lack units, and the x- and y-axis labels are missing or unclear. For instance, Figure 2 has low color contrast, making it difficult to distinguish between different elements. Improvements in figure clarity and completeness are necessary.

Response: Thank you for pointing this out. We fully agree that the clarity and completeness of the figures are critical for readers' understanding. In the revised manuscript, we will carefully improve the figures by adding missing units, clarifying axis labels, and enhancing color contrast to make the elements more distinguishable. In addition, we will provide higher-resolution versions of all figures to further improve their clarity in the final version.

Comment 3:

The representativeness of the station data at the 1 km grid scale needs further discussion. Please elaborate on how the spatial representativeness of point stations affects the validation results, especially in regions with complex topography or sparse station coverage.

Response: We sincerely thank the reviewer for raising this profound question, which directly addresses a core challenge in the validation of gridded products. We fully agree that the spatial representativeness of point stations is a fundamental factor—particularly in regions with complex terrain or sparse station coverage—and that it must be carefully considered when interpreting validation results.

This concern has also been central to our own research considerations. Drawing on our prior experience in evaluating satellite-based precipitation products, we consistently observed that mismatches between station locations and their corresponding grid cells—especially in terms of elevation—can introduce systematic biases into validation results. Motivated by this recognition, we specifically designed Section 4.4 to investigate how such mismatches affect validation accuracy. In this section, we identified 28 stations located in high-relief regions where the elevation difference

between recorded station elevations and those derived from the 1 km DEM exceeded 50 m, and conducted a controlled experiment in which two sets of predictions were generated: one using the actual station coordinates (longitude, latitude, and elevation) and the other using the coordinates of the corresponding grid-cell centers. By comparing both sets of predictions against in-situ measurements, we were able to explicitly separate and quantify the relative contributions of model error and representativeness error arising from elevation mismatch. The results demonstrated that for variables strongly influenced by elevation — such as temperature and pressure — representativeness error constitutes a substantial component of the total validation error, with its magnitude strongly correlated with the size of the elevation difference. These findings indicate that reduced validation accuracy in high-relief areas is not primarily due to deficiencies in the reconstruction framework itself, but rather to the inherent limitations of comparing point measurements with grid-cell estimates.

Despite these limitations, ground-based stations remain the cornerstone for validating gridded products—including satellite retrievals, reanalysis, and our reconstructions—as they provide the most accurate direct measurements available. In this context, the value of our work lies not in attempting to eliminate representativeness error, but in explicitly recognizing, quantifying, and interpreting it. Section 4.4 was designed with this purpose: to enable a fairer evaluation of model performance by distinguishing error sources attributable to environmental heterogeneity from those intrinsic to the reconstruction framework itself. Looking forward, we acknowledge that technological advances—such as denser ground-based networks and emerging mobile observation platforms (e.g., drones)—may help alleviate representativeness challenges.

Comment 4:

The reconstructed sunshine duration product does not show significant advantages over existing datasets. Concerns remain regarding data consistency, likely due to instrument changes and automation upgrades in CMA sunshine duration observations over time. This issue should be addressed to ensure reliability.

Response: Thank you very much for this insightful comment. In response to this concern, and consistent with another reviewer's constructive suggestion, we have incorporated the Himawari AHI–

based daily sunshine duration (SD) dataset (Zhang et al., 2025) into our comparative analysis. This satellite-derived, high-resolution product (5 km, 2016–2023) complements the homogenized station-based SSD dataset (2°, 1961–2022) and provides an independent benchmark for recent years. The revised analysis demonstrates that the reconstructed dataset achieves accuracy comparable to SSD in long-term temporal consistency, while also performing competitively with Himawari SD in recent high-resolution comparisons. Specifically, our reconstruction yields smaller systematic bias than Himawari, while Himawari attains slightly higher correlation in daily variability. These complementary findings highlight the robustness of the reconstruction framework and its combined strengths: reduced bias relative to satellite products, temporal stability comparable to homogenized long-term datasets, and the unique provision of six decades of 1 km daily sunshine duration fields for hydrometeorological applications in topographically complex regions. The detailed revisions have been made in the following sections:

Section 2.5 Existing gridded products for comparison:

"To assess the reliability and application potential of the reconstructed meteorological variables, representative and widely used gridded datasets were selected for comparison based on their scientific relevance and availability. Specifically, for average temperature, atmospheric pressure, and relative humidity, we employed the latest version of the China Meteorological Forcing Dataset (CMFD 2.0), whose earlier versions have been extensively used in land surface, hydrological, and ecological modeling over China (He et al., 2020).

The CMFD 2.0 (He et al., 2024) provides high-resolution (0.1°), 3-hourly gridded meteorological data for the period 1951–2020, covering the land area between 70°E–140°E and 15°N–55°N. It includes near-surface temperature, surface pressure, specific humidity, wind speed, radiation, and precipitation. Compared to previous versions, CMFD 2.0 incorporates ERA5 reanalysis and station observations through updated data sources and artificial intelligence techniques, particularly for radiation and precipitation variables. It also introduces metadata on station relocations and expands the spatial coverage beyond China's borders, thereby improving temporal consistency and cross-regional applicability.

As CMFD 2.0 does not include sunshine duration, we incorporated two additional datasets for its

evaluation. This step is critical because sunshine duration reconstruction constitutes the final step in our hierarchical framework, necessitating a thorough accuracy assessment to evaluate potential uncertainty propagation. To this end, we selected two complementary benchmarks: one long-term station-based product and one recent high-resolution satellite product. 1) The sunshine duration (SSD) dataset (He, 2024) serves as the long-term, station-based benchmark. It provides a homogenized daily sunshine duration record across China from 1961 to 2022 at a 2.0° × 2.0° resolution. Developed from over 2,200 meteorological stations and corrected for non-climatic influences (e.g., station relocations and instrumental changes), it offers a reliable baseline for evaluating the temporal stability and long-term climatological consistency of our reconstruction. 2) The Himawari AHI-based daily sunshine duration (SD) dataset (Zhang et al., 2025) provides a recent, high-resolution (5 km) satellite perspective for 2016–2023. It enables a direct assessment of our product's quality during the 2016–2019 overlap period and serves as a benchmark for evaluating fine-scale spatial accuracy."

Section 4.3.2 Sunshine duration:

"To comprehensively evaluate the accuracy of the reconstructed product, two representative benchmark datasets were employed: the homogenized station-based SSD product (2°) to assess long-term temporal consistency, and the high-resolution satellite-based Himawari SD product (5 km) to examine spatial performance.

As shown in Figure 8, when compared with the SSD dataset over 1961–2019, the reconstructed product demonstrated highly consistent accuracy. The median RMSE values were identical for both products (1.48 h), and the median CC values were likewise identical (0.93). The ME differed only slightly (0.05 h for the reconstructed dataset and 0.02 h for SSD), indicating comparable bias levels. Boxplot analysis further indicated that the reconstructed product exhibited slightly narrower interquartile ranges, whereas the SSD dataset showed fewer outliers in RMSE and CC. It should be noted that although some of the 95 CMA validation stations may have been included in the SSD development, our reconstruction model excluded these stations from training, ensuring a higher degree of validation independence.

For spatial performance, the reconstructed dataset was compared with the Himawari SD dataset over the overlapping period of 2016–2019 (Figure 9). The evaluation was based on 91 stations, since three of the 95 validation stations had invalid sunshine duration values during this period and one

station was located within the SD control region. Both products showed comparable RMSE levels (1.53 h for the reconstructed dataset compared with 1.48 h for Himawari). The satellite dataset achieved a slightly higher CC (0.94 compared with 0.92), reflecting stronger agreement in daily variations, while the reconstructed dataset exhibited a smaller ME (0.08 h compared with 0.21 h), indicating reduced bias.

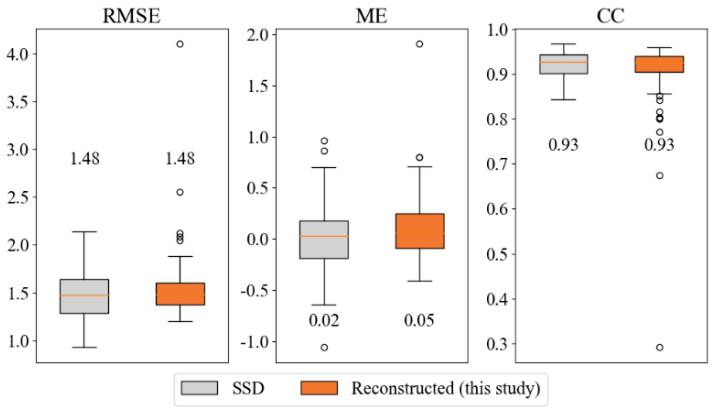

Figure 8: Boxplot comparison of RMSE, ME, and CC for sunshine duration between SSD (2.0°) and the reconstructed dataset developed in this study (1 km) from 1961 to 2019.

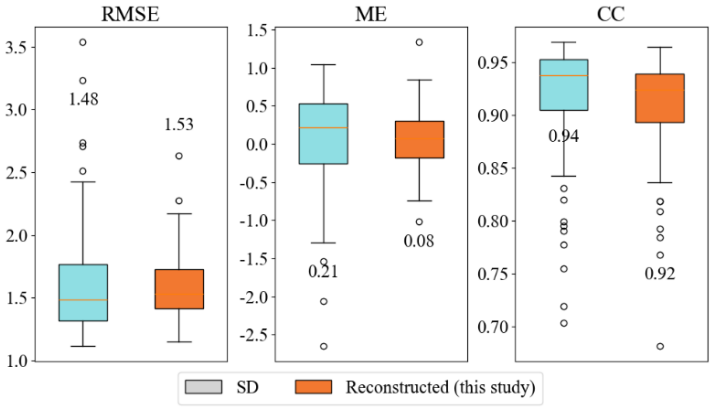

Figure 9: Boxplot comparison of RMSE, ME, and CC for sunshine duration between the Himawari AHI–based SD dataset (5 km) and the reconstructed dataset developed in this study (1 km) from 2016 to 2019.

These complementary results indicate that the reconstruction framework can achieve accuracy comparable to both a long-term homogenized station-based dataset and a high-resolution satellite-derived dataset."

In addition, we fully understand the reviewer's concern—instrument changes and automation upgrades are well-known factors affecting the homogeneity of long-term sunshine duration records. However, we would like to emphasize that the observational data used in this study are official CMA station records, which have undergone standardized quality control and are widely recognized in the

scientific community as the most reliable benchmark. Importantly, our reconstruction framework was explicitly designed to account for potential non-climatic biases arising from station relocations. Unlike conventional approaches, our model does not rely on fixed station metadata; instead, each daily observation in the training set is associated with the exact geographic information (longitude, latitude, and elevation) recorded for that specific date. As a result, if a station was relocated during 1961–2021, our model automatically treats the old and new locations as two separate geographic entities. This approach effectively avoids spurious biases introduced by site relocations and thereby improves the temporal consistency of the reconstructed product.

Comment 5:

The comparison with the CMFD product may be unfair, as CMFD utilizes a much smaller number of meteorological stations than this study. This discrepancy in input data may bias the comparison results. The authors should acknowledge this limitation and discuss its potential impact.

Response: Thank you very much for this valuable comment. We fully understand the reviewer's concern that differences in the number of input stations may affect the fairness of the comparison, and we agree that this point is indeed important.

We chose to compare our reconstruction with the CMFD product primarily because CMFD is widely used in China and has been recognized as an authoritative benchmark in the scientific community. Comparing a newly developed dataset against such a widely adopted reference is a common and necessary practice for evaluating its performance. Although CMFD is based on a relatively smaller number of stations, it achieves consistently high quality through the effective integration of reanalysis data and satellite products using advanced techniques, and has therefore become a well-recognized benchmark in this field.

We also fully acknowledge that machine learning methods depend on sufficiently large datasets, which is one of their inherent limitations. At the same time, a key advantage of such methods lies in their ability to capture the complex nonlinear relationships between meteorological variables and topographic or geographic factors. Therefore, in this study, the use of a larger set of ground-based observations is not only a necessary condition for applying the method, but also an important factor that enables the reconstructed product to better capture spatial heterogeneity and to achieve accuracy

comparable to, and in some respects even superior to, CMFD. These results highlight the potential of data-driven approaches to further improve gridded meteorological products. We will also mention this potential limitation in the discussion to ensure clarity for readers.

We sincerely appreciate the reviewer's insightful comment, which provided us with the opportunity to clarify this point and further enhance the rigor of our study.

Comment 6:

The long-term trends and homogeneity of the reconstructed dataset are not discussed. An analysis of the temporal consistency and homogeneity of the data—especially concerning non-climatic factors such as station relocations or instrument changes—would strengthen the dataset's credibility.

Response: We sincerely thank the reviewer for this important comment. We fully understand and agree that the long-term trends and homogeneity of the dataset are critical for its credibility in climate-related applications.

As this is a data paper, our central objective is to provide and evaluate a high-quality, high-accuracy reconstructed dataset that is suitable for long-term climate analyses. To ensure its reliability, we used official station observations from the China Meteorological Administration as the basis. These data have undergone standardized quality control and are widely recognized within the scientific community. Importantly, our reconstruction framework was explicitly designed to account for potential non-climatic biases caused by station relocations. Unlike conventional approaches, our model does not rely on fixed station metadata; instead, each daily observation in the training set is associated with the exact geographic information (longitude, latitude, and elevation) recorded for that specific date. As a result, if a station was relocated during 1961–2021, the model automatically treated the old and new sites as two separate geographic entities, thereby minimizing spurious biases from relocations and improving temporal consistency.

For validation, we compared the reconstructed dataset against several widely used benchmark products, including CMFD 2.0, SSD, and Himawari SD. The results demonstrate that over a period of nearly 60 years, our dataset exhibits strong consistency with these benchmark products in terms of statistical measures (e.g., correlation coefficient CC, root-mean-square error RMSE) and long-term

variability. Such cross-dataset consistency provides robust evidence of the temporal homogeneity and reliability of the reconstructed dataset.

We once again thank the reviewer for the careful review and valuable suggestions, which provided us with the opportunity to clarify and better present this dataset.