Dear Reviewer 2,

We are sincerely grateful for your careful review of our manuscript and for the constructive feedback you have provided. Your thoughtful suggestions have been very valuable for improving the overall quality and presentation of our study. In the following, we offer detailed, point-by-point responses to each of your comments. For ease of reading, the reviewer's comments are shown in **black** and our responses in **blue**. Sentences proposed as revisions or additions to the manuscript are highlighted in **gold** with quotation marks.

General comment:

This paper constructs a long-term meteorological variable dataset by decoding the nonlinear relationships between six meteorological variables and their spatial covariates. The method is innovative and the dataset is usable, but the paper needs to be revised based on the following points.

Response: We sincerely thank the reviewer for the positive overall assessment of our work, especially the recognition of the methodological innovation and the usability of the dataset. We also fully acknowledge the reviewer's suggestion that revisions are needed, and we have carefully addressed all the specific points raised. Detailed responses are provided below.

Comment 1:

Abbreviations such as "CC" in the abstract should be spelled out in full.

Response: We sincerely thank the reviewer for this valuable suggestion. In the revised manuscript, we have spelled out "CC" in full in the abstract as requested. Moreover, we have also expanded other abbreviations (e.g., "RMSE" and "ME") at their first occurrence in accordance with academic writing standards. The revised abstract can be found in our response to Comment 2.

Comment 2:

The abstract does not adequately reflect the research objectives and significance of the study and needs to be improved.

Response: We sincerely thank the reviewer for this valuable comment. We agree that the abstract should explicitly highlight the research objectives and significance. In the revised version, we have emphasized these points at the beginning of the abstract. In addition, we slightly refined the overall description of the abstract to keep it concise and balanced, thereby improving its readability. The revised abstract is shown below:

"The lack of fine-resolution and high-accuracy meteorological datasets in China has limited progress in climate, hydrological, and ecological studies. In this study, we present a 1 km daily dataset spanning 1961–2021 across China, which includes six key variables—mean, maximum, and minimum temperature, atmospheric pressure, relative humidity, and sunshine duration—to provide a reliable foundation for advancing related research and applications. The dataset was generated using a novel hierarchical reconstruction framework that leveraged daily observations from 2345 meteorological stations and incorporated topographic attributes. This approach effectively decodes the nonlinear relationships between the meteorological variables and their spatial covariates, ensuring the generation of gridded daily fields that are both high-resolution and spatially continuous. Validation against 118 independent stations confirmed the high accuracy of the dataset. For average, maximum, and minimum temperatures, the errors are minimal (median root mean square errors (RMSEs): 1.03°C, 1.19°C, 1.34°C; median mean errors (MEs): -0.09°C, -0.10°C, -0.08°C), and the consistency with in-situ data is very high (median correlation coefficients (CCs): 1.00, 0.99, 0.99). Atmospheric pressure also shows very small errors (median RMSE: 2.48 hPa; median ME: -0.02 hPa) and strong correlation (median CC: 0.98). Relative humidity exhibits relatively lower accuracy (median RMSE: 6.02%; median ME: -0.5%; median CC: 0.90), but it still exceeds standard benchmarks. Sunshine duration maintains high precision (median RMSE: 1.48 h; median ME: 0.05 h; median CC: 0.93), indicating the robustness and reliability of the dataset. Further comparison reveals that in high-altitude and topographically complex regions, the reconstructed product demonstrates higher actual accuracy than suggested by station-to-grid validation, as spatial mismatches between stations and grid cells lead to systematic underestimation. Free access to the dataset available at https://doi.org/10.11888/Atmos.tpdc.301341 or https://cstr.cn/18406.11.Atmos.tpdc.301341."

Comment 3:

The first paragraph of the introduction should provide some supporting citations.

Response: We thank the reviewer for this valuable suggestion. In the revised manuscript, we have added supporting citations in the first paragraph of the introduction to strengthen the background and provide authoritative references. Specifically, we now cite representative works demonstrating how advances in computational power and remote sensing technologies have driven hydrological modeling toward more physically based and fully distributed simulations (Lettenmaier et al., 2015; Singh, 2018), climate change research across broader scales (IPCC, 2021), as well as studies emphasizing the importance of high-resolution meteorological datasets in ungauged and topographically complex basins such as the Tibetan Plateau (Fu et al., 2020; Zhou et al., 2024). The revised paragraph as follows:

"With advances in computational power and remote sensing technologies, hydrological modeling has increasingly evolved toward fully distributed simulations (Lettenmaier et al., 2015; Singh, 2018), while climate change research continues to expand across broader spatial and temporal scales (IPCC, 2021). These developments have placed growing demands on the resolution and accuracy of basic meteorological inputs, particularly in ungauged and topographically complex basins such as the Tibetan Plateau (Fu et al., 2020; Zhou et al., 2024). High-resolution and high-quality meteorological datasets are essential for capturing fine-scale climate signals, representing land – atmosphere interactions, and supporting hydrological, ecological, and environmental assessments."

The newly added references are as follows:

Fu, Y., Ma, Y., Zhong, L., Yang, Y., Guo, X., Wang, C., Xu, X., Yang, K., Xu, X., Liu, L., Fan, G., Li, Y., and Wang, D.: Land-surface processes and summer-cloud-precipitation characteristics in the Tibetan Plateau and their effects on downstream weather: a review and perspective, National Science Review, 7, 500–515, https://doi.org/10.1093/nsr/nwz226, 2020.

IPCC: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B. (eds.), Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2391 pp., https://doi.org/10.1017/9781009157896, 2021.

Lettenmaier, D. P., Alsdorf, D., Dozier, J., Huffman, G. J., Pan, M., and Wood, E. F.: Inroads of remote sensing into hydrologic science during the WRR era, Water Resources Research, 51, 7309–7342, https://doi.org/10.1002/2015WR017616, 2015.

Singh, V. P.: Hydrologic modeling: progress and future directions, Geosci. Lett. 5, 15, https://doi.org/10.1186/s40562-018-0113-z, 2018.

Zhou, P., Tang, J., Ma, M., Ji, D., Shi, J.: High resolution Tibetan Plateau regional reanalysis 1961–present, Sci Data, 11, 444, https://doi.org/10.1038/s41597-024-03282-4, 2024.

Comment 4:

I noticed that the "Materials" section includes observation sites from different sources, but the verification sites were only selected from CMA. Have you considered selecting verification sites based on the weight of the number of sites from different sources?

Response: We sincerely thank the reviewer for raising this important point regarding the selection of validation sites. We appreciate the opportunity to clarify our methodology and believe the following explanation adequately addresses the concern.

Our validation strategy was guided by two main objectives: (1) ensuring strict internal consistency in the training dataset and utilizing as many stations as possible for training (since deep learning algorithms require sufficient data to achieve good performance), and (2) achieving independence and robustness in validation. To this end, we employed 2,345 CMA stations for model training and withheld 95 independent CMA stations for validation. The selection of these 95 stations followed the principles described in Section 2.1 , which state:

"*To support independent model validation, a total of 95 stations were selected as evaluation sites based on three principles: (1) ensuring geographical representativeness in terms of longitude, latitude, and elevation; (2) in densely monitored areas such as eastern China, a greater number of evaluation stations were retained without significantly reducing the size of the training dataset; and (3) in sparsely monitored regions such as western China (including Tibet and Xinjiang), fewer stations were assigned to the evaluation set in order to preserve sufficient data for model training.*"

However, we acknowledge that in key regions such as Taiwan and the Tibetan Plateau, CMA evaluation stations are very limited (with no stations in Taiwan). To address this, we supplemented the validation dataset with additional independent observations. Specifically, we incorporated 12 ground-based meteorological stations from the Department of Water Resources (DWR) located in the Tibetan Plateau region (see Section 2.2.1), and 8 international stations from the Global Surface Summary of Day (GSOD) dataset covering Taiwan (see Section 2.2.3). These supplementary data were added to ensure broader coverage and a more reliable evaluation in regions with sparse CMA observations.

In addition, for comparison against the China Meteorological Forcing Dataset (CMFD v2.0), we collected 31 field stations concentrated over the Tibetan Plateau from literature-based datasets archived at the National Tibetan Plateau Data Center (see Section 2.2.2). Since CMFD has assimilated or blended CMA data, and we cannot determine which specific CMA stations were used, these independent TP field stations were used as a fair and unbiased benchmark for inter-product comparison.

Through this combined strategy—using CMA-based stations for consistency and multi-source stations (DWR and GSOD) for robustness—we expanded the final validation set to 115 sites. In addition, 31 independent TP field stations from the National Tibetan Plateau Data Center were specifically reserved for inter-product comparison to ensure fairness and independence. This approach provides a more comprehensive and balanced evaluation of our dataset, especially in high-altitude and data-sparse regions. The spatial distribution of both training and validation sites is shown in Figure 1.

We are grateful for the reviewer's comment, which allowed us to better articulate the rationale and implementation of our validation design.

Comment 5:

How to consider the cumulative error caused by progressively inputted meteorological variables during the modeling process, especially in the sunshine duration model.

Response: We thank the reviewer for raising this important point. We would like to clarify that cumulative error does not occur during the training stage, since each step of model training was performed using the true location information (longitude, latitude, and elevation) and observed meteorological values from CMA stations. For example, when training the mean temperature model, longitude, latitude, and elevation were used as predictors and the observed mean temperature as the target variable; when training the maximum temperature model, the observed mean temperature together with site attributes was used as input and the observed maximum temperature as the target. This procedure was applied similarly to other variables

But potential cumulative error could only arise during the reconstruction phase, when multiple 1 km gridded meteorological fields are generated from gridded mean temperature fields reconstructed using the 1 km DEM. Nevertheless, evaluation against independent ground observations demonstrated that the reconstructed products maintained high accuracy. In particular, sunshine duration — as the final step in the progressive framework, where cumulative error would theoretically be greatest — still exhibited consistently high precision (median RMSE: 1.48 h; median ME: 0.05 h; median CC: 0.93). This demonstrates the robustness and reliability of the dataset and indicates that, although error amplification is theoretically possible, it did not significantly affect the final results.

Comment 6:

How can authors reduce errors caused by the boundaries of the study area during modeling, given that these areas have fewer observation stations?

Response: We thank the reviewer for this valuable comment. In our modeling process, we did not apply specific correction schemes to reduce potential boundary effects. However, the evaluation results demonstrate that the hierarchical deep-learning framework exhibits strong extrapolation and generalization ability. Taiwan provides a typical example of this capability: although no CMA stations were included in training phase, the model was still able to reasonably reconstruct air temperature fields in this region. Independent validation using 8 international stations from the GSOD dataset (providing average, maximum, and minimum temperature) further confirmed the accuracy of the reconstruction.

We also analyzed this issue in the manuscript (see Section 4.2). As stated: "*The spatial distribution of RMSE, ME, and CC for all six meteorological variables is further illustrated in Figures 6. Consistent with expectations, the Subtropical and Southern Temperate Zones in southeastern China (STZ-southeastern China) display the best performance across all variables, largely due to the high density of training stations in these regions. In contrast, performance metrics are relatively lower in* *the Middle Temperate, Southern Temperate, and Plateau Climate Zones of northwestern China (MSPZ–northwest China), as well as in Taiwan, where no stations were included in training. Nevertheless, model performance in these regions remains robust. Notably, despite the absence of training data in Taiwan, the MLP model accurately reconstructs air temperature in that region, suggesting strong spatial generalizability."*

Comment 7:

I noticed that Figure 4 contains a large amount of information, but the poor resolution and color quality make it difficult to see clearly. Please improve this.

Response: We sincerely thank the reviewer for this helpful comment. We will revise Figure 4 by increasing its resolution and optimizing the color scheme to enhance readability. The improved version will be provided in the final manuscript.

Comment 8:

Please indicate whether adjustments were made when the author encountered situations where the sunshine duration was less than 0 during the verification.

Response: We thank the reviewer for raising this important point. Indeed, to guarantee the physical validity of our dataset, we implemented a quality control procedure wherein any predicted sunshine duration value below 0 was reset to 0. This was necessary as the model's unconstrained regression output can generate physiologically impossible negative values near zero.

Comment 9:

The author mentioned the limitations of satellite remote sensing in estimating the meteorological variables in the "introduction". However, sunshine duration is greatly affected by cloud and aerosol parameters observed by remote sensing. When comparing sunshine duration products, please consider comparing sunshine duration datasets estimated based on remote sensing data (https://doi.org/10.5194/essd-17-1427-2025) and explain the advantages of the research method used in this study.

Response: Thank you very much for this insightful comment. Following the reviewer's 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.

To ensure logical consistency, Section 2.5 Existing gridded products for comparison in the Materials was rewritten to include the Himawari SD dataset alongside CMFD 2.0 and SSD, clearly outlining the rationale for selecting these complementary products. Furthermore, Section 4.3.2 Sunshine duration in the Results and Discussion was substantially revised to present a comprehensive comparison of our reconstructed dataset against both SSD and Himawari SD.

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 revised content of Section *2.5 Existing gridded products for comparison* is provided below:

"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."

The revised content of Section *4.3.2 Sunshine duration* is provided below:

"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.

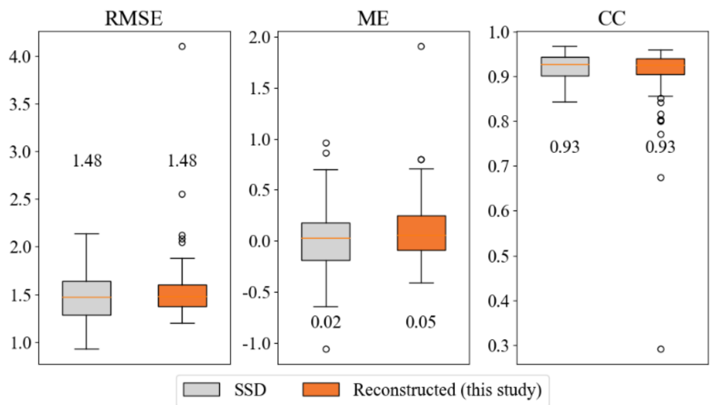

Figure 8: Boxplot comparison of RMSE, ME, and CC for sunshine duration between SSD (2.0°) and the reconstructed dataset developed in this study (1 km) from 1961 to 2019.

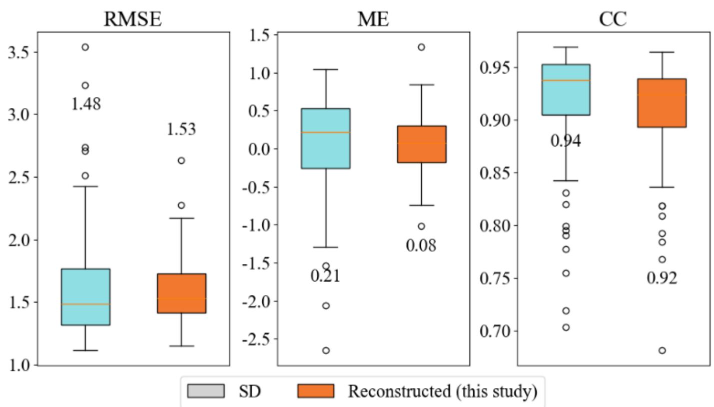

Figure 9: Boxplot comparison of RMSE, ME, and CC for sunshine duration between the Himawari AHI–based SD dataset (5 km) and the reconstructed dataset developed in this study (1 km) from 2016 to 2019.

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."

Comment 10:

Is the model independent on a daily scale? Did the authors consider modeling based on different days of year (DOYs) to enhance the model's generalization ability in the future?

Response: We thank the reviewer for this helpful question. Our reconstruction framework is indeed independent on a daily scale: for each day, the model relies exclusively on station observations and spatial covariates corresponding to that specific day, without drawing on information from preceding or subsequent days. This design ensures that daily fields are generated without temporal autocorrelation, thereby simplifying interpretation and enhancing operational applicability. Moreover, because each day is reconstructed independently, occasional data gaps on specific days do not affect the performance on other days.

In fact, we note that we also tested an alternative transformer-based approach in which temporal context from surrounding days was incorporated. This experiment, however, showed limited skill in capturing day-to-day fluctuations compared to our daily-independent model. Because ESSD primarily emphasizes the quality and validation of datasets rather than extensive methodological comparisons, we did not include this exploratory test in the main text. We acknowledge that further exploration of temporal approaches, including the reviewer's suggestion to model based on different days of the year (DOY) to capture seasonal cycles, could be valuable for future improvements in long-term high-resolution meteorological reconstructions.