Response to Reviewer#1

Manuscript number： https://doi.org/10.5194/essd-2025-717

Title： An AI-Driven Reconstruction of Global Surface Temperature with Emphasis on Refining the Antarctic Record

Comments from reviewer：

In this manuscript, the authors apply deep learning to reconstruct historical global surface temperature fields, with a specific focus on Antarctic climate dynamics. While this topic is relevant to research on climate change impacts and attribution, I have several concerns regarding the methodology, the robustness of the validation, and the demonstrated utility and uniqueness of the final products. Additionally, various statements throughout the text require further clarification. Please find my detailed comments below.

Response: We sincerely thank the reviewer for the careful evaluation of our manuscript and for the detailed and constructive comments. These suggestions have provided valuable guidance for improving our work. We have thoroughly revised the original manuscript in accordance with the reviewer's recommendations and have addressed each comment point by point. Detailed responses and corresponding revisions are provided below.

In the following, black text indicates the reviewer's comments, blue text indicates our responses, *blue italic text* indicates the revised content, and *red italic text* with strikethrough indicates deleted content, *black italic text* indicates the original content.

1. To my knowledge, the reconstruction framework based on a partial convolutional neural network (PConv) was first applied to historical climate field reconstruction by Kadow et al. (2020). The authors follow a nearly identical methodological approach. The authors must explicitly acknowledge Kadow et al. (2020) in the methodology section. This will help readers accurately contextualize the foundational literature and better understand the specific novel contributions of this work.

Response: We appreciate this suggestion and have added an explicit citation and supplementary description of Kadow et al. (2020) in the Data and Methods section. We also provide a brief overview of their key contribution to climate field reconstruction based on the partial convolutional neural network (PConv). Furthermore, we clarify the relationship and distinctions between our study and their work. Compared with the original study, our work extends the application to different sea surface temperature datasets and sea–land merging strategies. We also introduce a more systematic multi-dataset validation framework and regional evaluation approach to provide a more comprehensive assessment of the results. In addition, improvements have been made for the Antarctic region after 1961.

2. Following the previous point, the authors should at least use the latest dataset built by Kadow et al. (2020) as a benchmark to evaluate the reliability and added value of the newly proposed dataset. Also, regarding the uniqueness of the proposed dataset, it would be interesting and necessary to see how the latest Kadow et al. (2020) product performs in the Antarctic region compared to this study.

Response: We thank the reviewer for this important suggestion. We agree that incorporating the dataset developed by Kadow et al. (2020) as a benchmark is of great significance for assessing the reliability and added value of our dataset. In the revised manuscript, we have

included this dataset as one of the reference datasets and conducted corresponding comparative analyses. Specifically, we evaluate and compare the temporal trends and variability, as well as the spatial distribution characteristics, at both global and regional scales (including the Antarctic region). The relevant results and discussions have been added in the corresponding sections of the revised manuscript.

3. The evaluation and validation sections need to be strengthened. Currently, it is difficult to distinguish the unique strengths and irreplaceability of these new datasets from those of existing reference datasets solely on the basis of global/regional temperature trends. Providing a clearer demonstration of where and why this dataset outperforms existing ones would greatly improve the manuscript.

Response: We thank the reviewer for these valuable comments. We acknowledge that the validation and analysis framework in the original manuscript was relatively limited. In response, we have systematically strengthened the evaluation framework by incorporating additional analyses of the reconstructed results, including "ENSO spatial patterns and the Ocean Niño Index (ONI)" "zonal mean comparisons," and "global warming patterns." In these sections, we also conduct more comprehensive comparisons with other benchmark datasets.

Furthermore, in the original sections on "ENSO spatial patterns," "GMST and its trends," "regional land temperature time series and their trends," and "Antarctic temperature series," we have enhanced the comparisons with other datasets to make them more explicit. Through these additions, we aim to provide a more comprehensive assessment of the differences between our dataset and existing datasets. Overall, these improvements are intended to establish a more complete and targeted evaluation framework, thereby more fully addressing the reviewer's concerns.

4. I recommend moderating some of the claims regarding the capabilities of AI methods. While powerful, some statements about AI's strengths and the overall contribution of the study come across as overconfident and could be nuanced to reflect limitations of methods and datasets.

Response: We thank the reviewer for the careful and thoughtful comments. We agree that some statements in the original manuscript regarding the capabilities of the AI method may have been overly absolute and did not fully reflect the scope of applicability and inherent limitations of the method and the data. Accordingly, in the revised manuscript, we have moderated such statements, avoiding overly definitive conclusions that are not fully supported by the available evidence, and have reformulated the main conclusions in a more objective and conditional manner. Through these revisions, we aim to present a more balanced and cautious overall narrative.

5. The study provides two reconstruction products, but it is not entirely clear which one is considered superior or more reliable. A more solid comparison between the strengths and characteristics of the two products would be helpful. Furthermore, providing explicit guidance on which dataset users should select for specific use cases would add practical value to the paper.

Response: We thank the reviewer for the constructive comments. We agree that the comparison between the two reconstructed products in the original manuscript was not sufficiently systematic, which may hinder readers from assessing their relative advantages and applicable

scenarios. In the revised manuscript, we have conducted a more comprehensive and systematic comparative analysis of the two datasets, including quantitative evaluations of their performance differences across different regions and temporal scales.

Based on these analyses, we further summarize the respective strengths and limitations of the two datasets, such as their stability in data-sparse polar regions, their ability to capture the magnitude of global warming, and the consistency of long-term trends. We also provide corresponding recommendations for dataset selection to enhance their practical applicability.

Specific Comments:

L47-50: The authors mention that the "propagation of observational errors", "inconsistencies among reanalysis products", and "harsh environmental conditions" affect existing statistical reconstruction methods (PCA, EOT, DINEOF). While true, it is worth noting that AI-based methods often struggle with these exact same regional challenges. I suggest softening the critique of traditional methods here.

Response: We appreciate this suggestion and have moderated our critique of traditional statistical methods accordingly.

Changes to the manuscript: ~~*However, missing information inevitably introduces uncertainties and structural biases. The propagation of observational errors, inconsistencies among reanalysis products, and the inherent limitations of interpolation assumptions can all affect the reliability of reconstruction results (Huang et al., 2017; Morice et al., 2021). These issues are particularly pronounced in regions such as Africa, South America, and Antarctica, where sparse observations and harsh environmental conditions pose greater challenges for traditional reconstruction methods.*~~ *Nevertheless, due to the inherent incompleteness of the observational system, climate reconstruction inevitably introduces uncertainties and potential biases (Huang et al., 2017; Morice et al., 2021). These issues are particularly pronounced in regions with sparse observations or complex environmental conditions, such as Africa, South America, and the polar regions. It should be emphasized that these challenges are not specific to any single method, but rather represent common limitations shared by current climate reconstruction approaches. These methods have been widely applied at the global scale, their performance exhibits regional variability, with uncertainties becoming more pronounced in areas characterized by extremely limited observations and complex environmental conditions. Among such regions, Antarctica, owing to its unique climatic background and observational constraints, serves as a critical testbed for evaluating the effectiveness and reliability of climate reconstruction methods.*

L58: Please clarify what is specifically meant by "extreme geographical conditions" in this context.

Response: We have added a description of “extreme geographical conditions” in the revised manuscript.

Changes to the manuscript: ~~*However, due to its extreme geographical conditions, harsh climate, and logistical and communication constraints, observational data from Antarctica remain extremely scarce and temporally uneven.*~~ *However, the widespread high-elevation ice sheets and complex terrain across Antarctica pose substantial challenges to station deployment. In addition, persistently low temperatures (often below −40 °C) and strong winds hinder the stable operation of observational*

instruments, while limited transportation and communication infrastructure further complicate station maintenance and data transmission (Wang et al., 2023). To alleviate the problem of insufficient observations, ...

L98-99: Please remove the claim regarding interpretability. PConv models are generally not recognized for their physical interpretability.

Response: We have removed the description of the model's "physical interpretability."

Changes to the manuscript: ~~*By combining statistical approaches with convolutional neural networks, the methodology seeks to balance model interpretability with nonlinear representation capability. The reconstructed temperature anomaly fields are then systematically compared with existing observational and reconstructed products to assess the strengths and limitations of AI-based methods for this application.*~~ *By integrating statistical methods with convolutional neural networks and leveraging their nonlinear fitting capabilities, global and Antarctic surface temperatures are reconstructed. The results are then systematically evaluated against existing observational and reconstructed datasets to assess the strengths and limitations of this AI-based approach for this problem.*

L104-109: The introduction of the C-MST3.0 dataset is somewhat confusing, given that the reconstruction primarily relies on C-LSAT2.1. Consider revising this section to streamline the data origins.

Response: In the revised manuscript, we have reorganized and clarified the logical relationship between C-LSAT2.1 and C-MST3.0 in the "Data Sources" section.

Changes to the manuscript: *China land surface air temperature dataset (C-LSAT) is a global land surface air temperature (LSAT) product that provides both station-based and gridded data. It integrates a total of 14 data sources, including three global datasets (CRUTEM4, GHCNm, and BEST), three regional datasets, and eight national datasets. A major advancement of this dataset lies in its substantially improved station coverage across most Asian countries, particularly in China and its surrounding regions (Xu et al., 2018; Li et al., 2021; Wei et al., 2025). The dataset was assessed in the Intergovernmental Panel on Climate Change (IPCC) Sixth Assessment Report (AR6), released in August 2021, as "fully meeting the IPCC requirements," and was accordingly incorporated and utilized in the report (IPCC, 2021). C-LSAT2.1 represents its latest version (Wei et al., 2025).*

Building upon this foundation, the China global Merged Surface Temperature (C-MST3.0, Yun et al., 2019; Sun et al., 2021, 2022; Li et al., 2020, 2021) is constructed by merging the land surface air temperature dataset C-LSAT2.1 with the Extended Reconstructed Sea Surface Temperature dataset ERSSTv6 (Huang et al., 2025a, 2025b). According to the spatial extent of the reconstructed Arctic sea ice surface air temperature regions, three variants are defined: C-MST3.0-Nrec, C-MST3.0-Imin, and C-MST3.0-Imax (Li et al., accepted). ~~*The China global Merged Surface Temperature (China-MST/C-MST) dataset is an established global ST product (Yun et al., 2019; Sun et al., 2021, 2022; Li et al., 2020, 2021). The latest version, C-MST3.0, is classified into three variants (C-MST3.0-Nrec, C-MST3.0-Imin and C-MST3.0-Imax) based on the spatial coverage of reconstructed Arctic sea-ice ST (Li et al., under review). C-MST3.0 is constructed by combining ST*~~

~~*from the China land surface air temperature (LSAT) dataset C-LSAT2.1 (Wei et al., 2025) with the Extended Reconstructed Sea Surface Temperature version 6 (ERSSTv6), released by NOAA/NCEI (Huang et al., 2025a, 2025b).*~~

L136: Please elaborate on what the "influence by features at the edges of missing data" specifically entails in this context.

Response: We have added an explanation of the "influence of the edges of missing-data regions" in the corresponding paragraph of the original text, and have provided a more detailed analysis and description of this result in Section 3.2.

Changes to the manuscript： *Specifically, in Antarctic coastal transition zones where partial observations are available over the ocean but land observations are nearly absent, these grid cells in PConv can only rely on information from adjacent oceanic grids. The convolutional operation in PConv generates predictions based on the surrounding available data, which leads to a bias toward oceanic characteristics and makes it difficult to accurately represent the true local surface air temperature over land or ice sheets. In simple terms, "land-edge grid cells are biased by surrounding heterogeneous ocean grids." This effect is particularly pronounced in the merged dataset (Merge-E), which is constructed by combining with the fully ocean-covered ERSSTv6, and in regions with severe observational gaps (see Section 3.2 for a detailed discussion of GMST).*

Figure 1: This schematic appears to share strong similarities with the framework figure in Kadow et al. (2020). Please ensure appropriate citations or copyright permissions are included if it was adapted. Additionally, the numbers at the bottom of most boxes are somewhat confusing and could benefit from an explanation in the figure caption.

Response: In the revised manuscript, we have made minor modifications to the flowchart to avoid excessive similarity and have explicitly cited Kadow et al. (2020). In addition, we have added a description in the figure caption explaining the numbers at the bottom of each module in the diagram.

Changes to the manuscript:

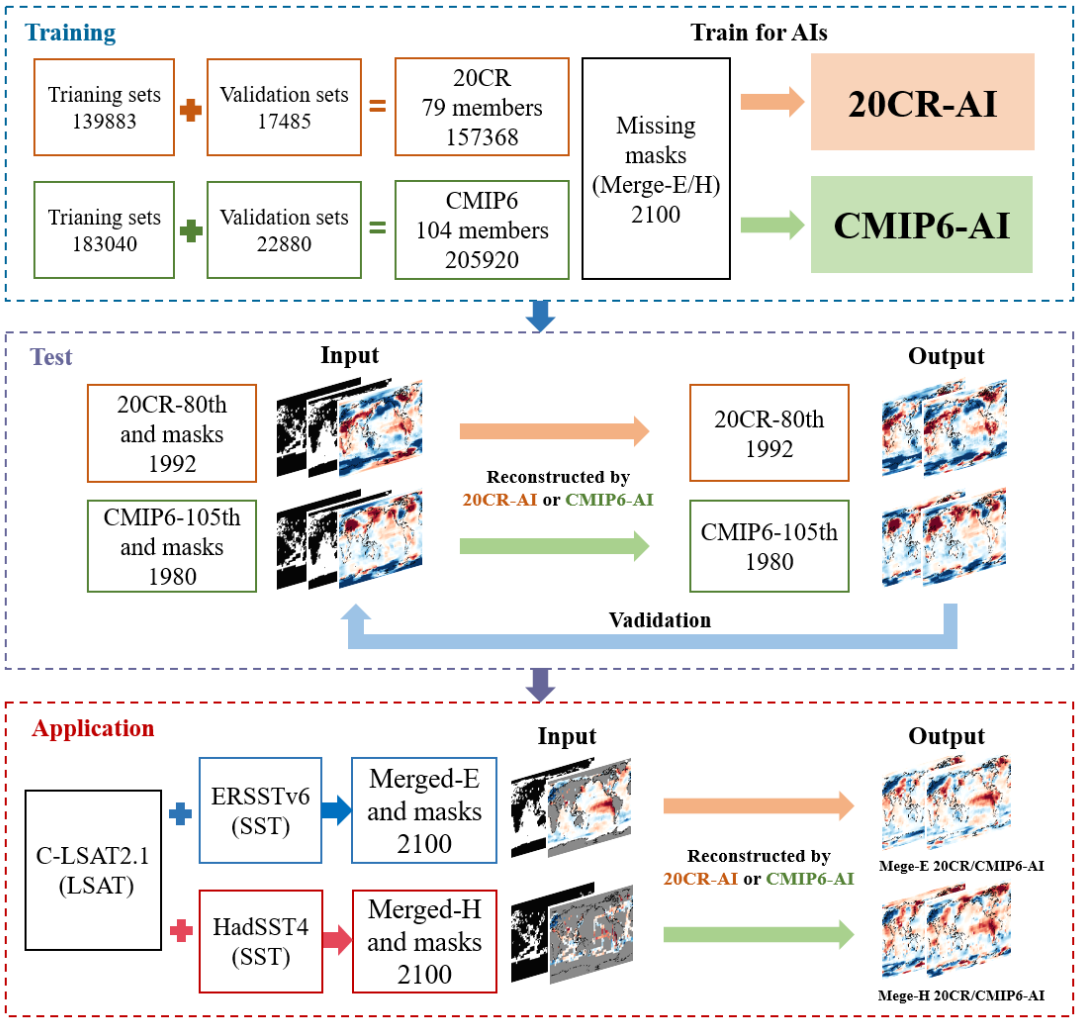

Figure 1: AI training and reconstruction (The numbers shown in the data boxes represent the sample size of the monthly anomaly fields, corresponding to the total number of months).

The reconstruction workflow is illustrated in Figure 1 (adapted from Extended Data Fig. 1 of Kadow et al., 2020)

L149-150: Please specify exactly what is meant by "the AI reconstruction follows a similar approach." Similar to what?

Response: We have removed the ambiguous description of "similar approach" and have provided a detailed description of the AI-based reconstruction workflow used in this study in Section 2.3, "AI Training and Reconstruction Methods."

Changes to the manuscript: ~~*In this study, the AI reconstruction follows a similar approach.*~~ *This study builds upon the work of Kadow et al. (2020), whose primary contribution was the introduction of a PConv-based framework for climate data reconstruction. Kadow et al. mainly demonstrated the feasibility and accuracy of PConv for climate field reconstruction at the global scale with a resolution of 5° × 2.5°, and showed that the method can effectively capture the spatial patterns of ENSO-related*

sea surface temperature anomalies. The PConv approach performs well for missing grid points over the low and mid-latitude areas, but provides limited assessment for polar regions or areas with severe observational deficiencies.

In this study, we introduce targeted improvements for the Antarctic region after 1961. On the one hand, additional Antarctic station data are incorporated to construct the input samples, ensuring sufficient valid data support after 1961. On the other hand, considering the coarse spatial resolution (5° × 2.5°) and the pronounced land–sea contrasts in Antarctic grid cells, the Antarctic land mask is adjusted. Based on these modifications, more detailed validation and analysis are conducted for both global and Antarctic domains to assess the performance and limitations of the model under different training datasets and different SST products.

L168: Provide more details on the regridding process. What exact resampling method was used?

Response: We have added a description of the interpolation method used in the regridding process.

Changes to the manuscript: ~~*To ensure comparability across different data sources, all external datasets, including observations, reanalysis products, and reconstructed data, were first regridded to a common 72×72 regular grid consistent with the AI reconstruction output, with a uniform spatial resolution of 5°×2.5°.*~~ *To ensure comparability across different data sources, all external benchmark ST datasets are first remapped to a spatial resolution of 5° × 2.5° consistent with the AI reconstruction inputs and outputs, using bilinear interpolation, resulting in a 72 × 72 regular grid.*

L205-211: The authors claim that the AI reconstruction performance under the Merge-E mask is better than that of Merge-H, attributing the smoother reconstructed fields in Antarctica to the AI being "influenced to some extent by the colour, texture, and style features at the edges of missing regions." This conclusion is debatable. Because ERSST is a spatially complete, interpolated dataset, it lacks the vast oceanic gaps present in datasets like HadSST. Therefore, the "smoother" performance observed in Merge-E is likely not due to PConv's enhanced inpainting capabilities, but rather to the statistical interpolation already performed within the ERSST dataset itself. Please revisit and clarify this discussion.

Response: Thank you for the comment. The original wording may have been misleading. What we intended to convey is that PConv is influenced by the smoother characteristics of ERSST, which leads to a smoother representation in the Antarctic region for Merge-E compared to Merge-H, rather than reflecting an inherent advantage of PConv itself. In the revised manuscript, we have removed this potentially misleading statement to avoid confusion for the reader.

Changes to the manuscript:

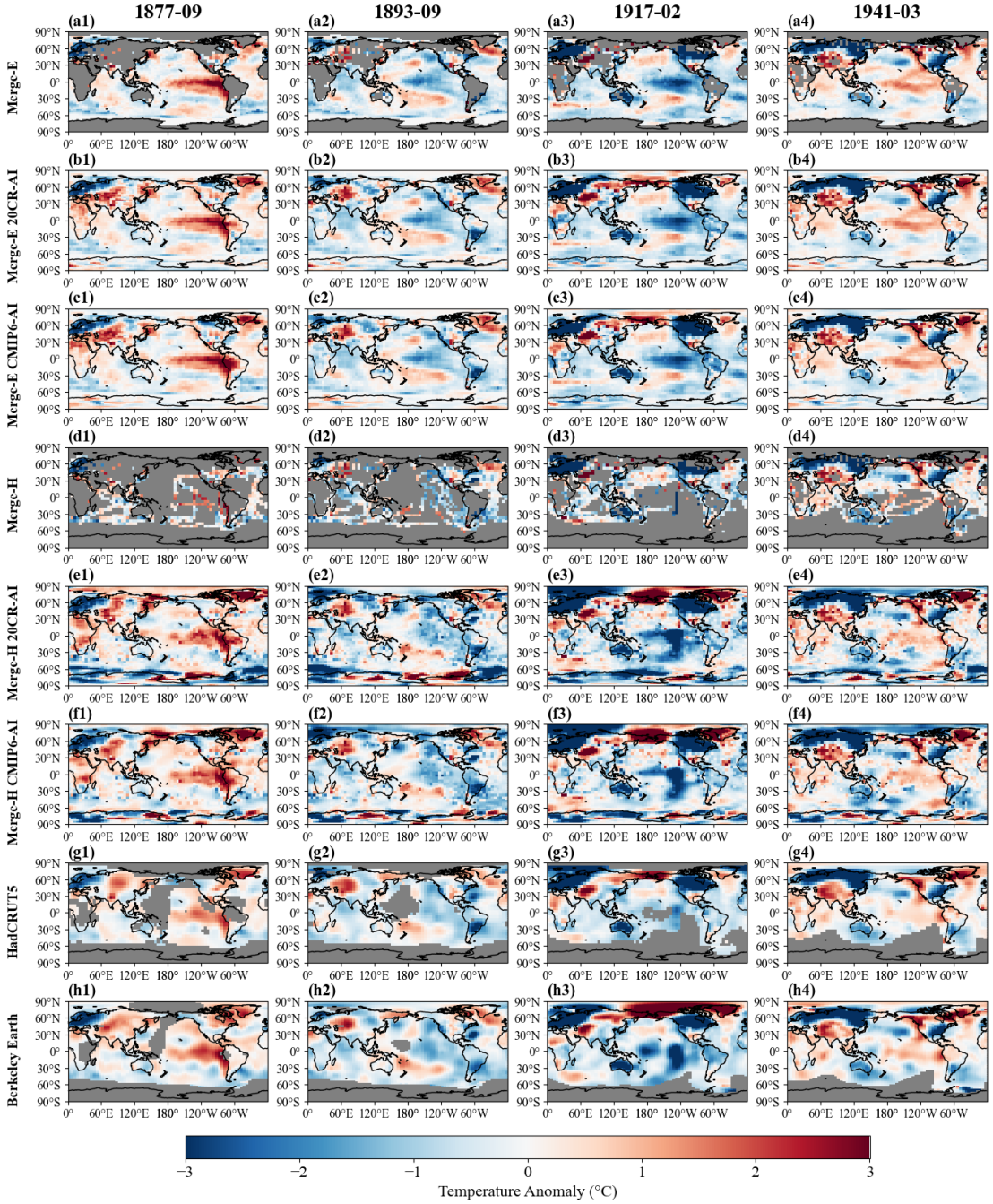

Figure 2: Global temperature anomaly fields before and after reconstruction for four typical ~~months~~ ENSO events. (a1–a4) Merge-E original; (b1–b4) Merge-E 20CR-AI; (c1–c4) Merge-E CMIP6-AI; (d1–d4) Merge-H original; (e1–e4) Merge-H 20CR-AI; (f1–f4) Merge-H CMIP6-AI; (g1–g4) HadCRUT5; (h1–h4) Berkeley Earth.

~~*The global spatial patterns of the four reconstructed fields are largely consistent across most regions; however, in high-latitude areas with extremely sparse early observations, the spatial distributions of Merge-E and Merge-H reconstructions show some differences. The reconstruction performance of the model improves as the coverage of original valid data in the model validations (Fig. S4, S5 and S6). Compared to Merge-H, Merge-E exhibits higher spatial coverage and fewer missing gaps in polar regions, particularly over Antarctica and its surrounding seas. This leads to better AI reconstruction performance under the Merge-E mask than under Merge-H (Fig. S4 and S6). Visually, the reconstructed*~~

~~*fields in Antarctica and adjacent regions are smoother in Merge-E than in Merge-H (Fig. 2), indicating that during image inpainting, the AI reconstruction is influenced to some extent by the colour, texture, and style features at the edges of missing regions in the two different datasets (Liu et al., 2018; Nazeri et al., 2019), thereby leading to distinct reconstructed features in the early periods when large areas of missing data occur in Merge-E and Merge-H.*~~ *In Fig. 2 (b1–b4, c1–c4), Merge-E uses the complete ocean component from ERSSTv6, which is not reconstructed by the AI model; therefore, the ENSO spatial patterns in these panels are directly derived from ERSSTv6 and are shown here for reference. During the strong El Niño event in September 1877, the AI model is able to reasonably reconstruct a coherent warming anomaly pattern over the equatorial western Pacific under data-sparse conditions (Fig. 2 d1, e1, f1). This pattern is characterized by pronounced warm anomalies in the tropical central and eastern Pacific, with alternating warm and cold anomalies distributed zonally along the equator, reflecting a typical ENSO signal.*

For the ENSO reconstructions in September 1893, February 1917, and March 1941 (Fig. 2 d2–d4, e2–e4, f2–f4), the spatial patterns produced by the AI model are generally consistent with those from HadCRUT5 (Fig. 2 g2–g4) and Berkeley Earth (Fig. 2 h2–h4), both of which also use HadSST-based ocean components. However, the spatial distribution of warm and cold anomalies along the equatorial Pacific, namely the classic "warm–cold tongue" structure, is less pronounced than in ERSSTv6 (Fig. 2 a2–a4). This discrepancy primarily arises from inherent differences between HadSST and ERSSTv6.

~~*Moreover, the AI reconstruction effectively reproduces characteristic climate events in key years. In particular, the results shown in Figures 2 (e1, f1) clearly capture the strong El Niño event of 1877, characterized by significantly positive SST anomalies in the central and eastern tropical Pacific and a distinct east–west dipole pattern of warm and cold anomalies along the equator, reflecting the typical signal of the El Niño–Southern Oscillation (ENSO). This demonstrates that the model is capable of reconstructing large-scale spatial patterns and temporal evolution of the temperature field even under extremely sparse observational coverage, indicating robust performance and spatial consistency.*~~

L215-217: Highlighting the single-year result of the 1877 El Niño is a good visual check, but it is not sufficient to claim "robust performance and spatial consistency." Moreover, to my knowledge, the infilled HadCRUT5 dataset can also represent this 1877 event. Additional spatial validation is needed to support this claim.

Response: We agree that the analysis of the spatial pattern of El Niño sea surface temperature anomalies in 1877 represents only a single case and is insufficient to demonstrate the overall performance of the dataset in reconstructing ENSO events. Therefore, we have added spatial validation analyses using HadCRUT5 and Berkeley Earth in Figure 2. In addition, we have analyzed the Oceanic Niño Index (ONI) of the reconstructed data in comparison with other benchmark datasets, in order to provide a more comprehensive evaluation of the dataset's ability to reconstruct ENSO events.

Changes to the manuscript:

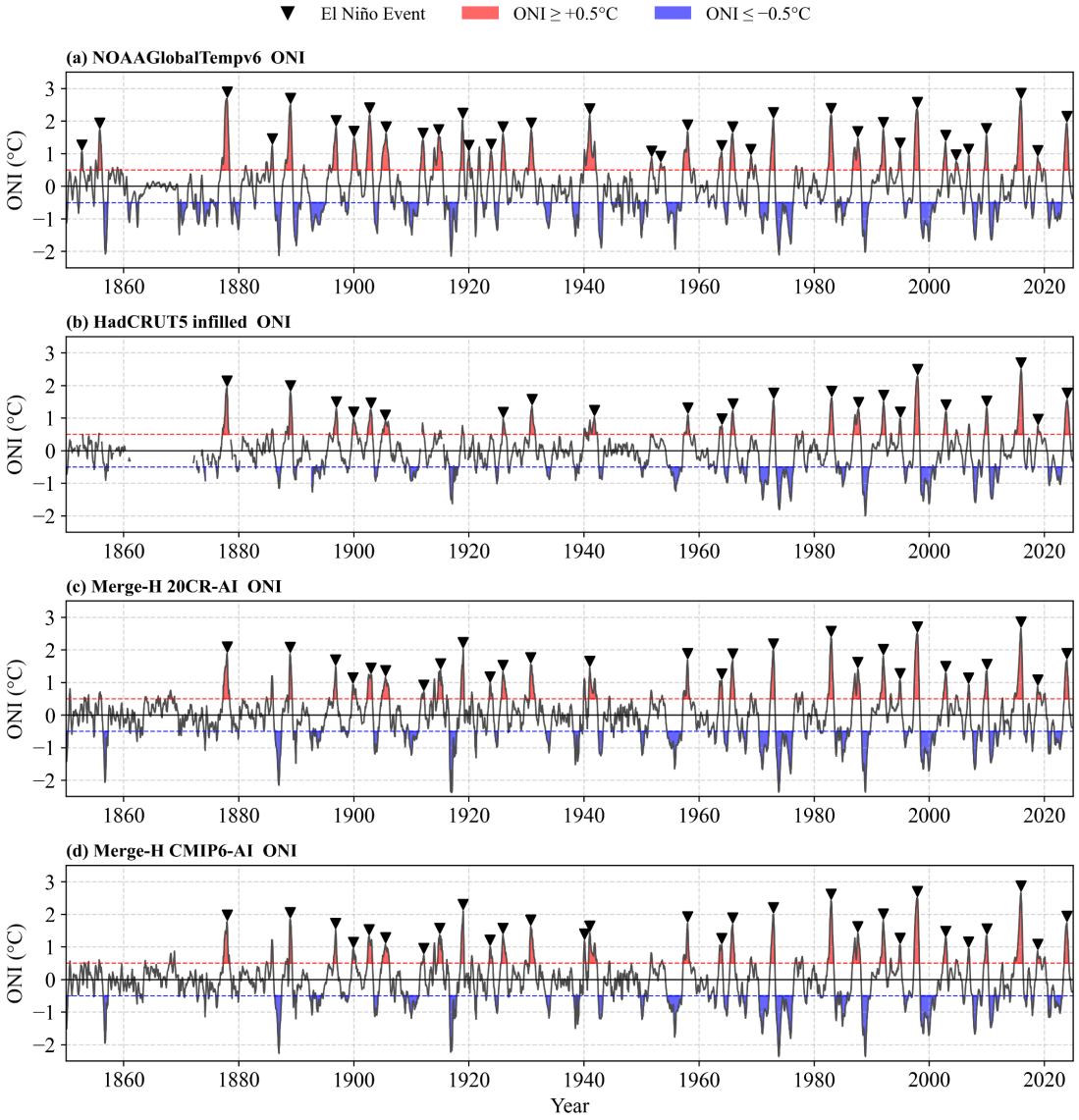

Figure 3: Ocean Niño Index (ONI) time series indicating ENSO events. (a) Time series of ONI from NOAAGlobalTempv6 over 1850–2024, with black triangles indicating El Niño events defined as ONI exceeding 0.5 °C for at least five consecutive months; (b–d) same as (a) but for HadCRUT5 infilled, Merge-H 20CR-AI, and Merge-H CMIP6-AI, respectively.

The two AI models are able to effectively capture the spatial patterns of ENSO, while also reproducing historical ENSO events. As shown in Fig. 3(a–d), the positive and negative phases of the ONI, which serves as an indicator of ENSO variability, are generally consistent with those from NOAAGlobalTempv6 and HadCRUT5. The AI-based reconstruction also fills the gaps in the ONI during 1910–1920 in HadCRUT5 infilled, where incomplete spatial coverage over the Niño 3.4 region led to biases or missing values. In addition, the reconstruction identifies the documented El Niño events of 1911–1912, 1914–1915, and 1918–1919 (Yu et al., 2013), which are also evident in NOAAGlobalTempv6 (Fig. 3a).

A few weak El Niño events show slight discrepancies compared with HadCRUT5 due to small differences in the amplitude of the Niño 3.4 index, indicating minor estimation biases in the AI-

reconstructed index. It is also noteworthy that the ENSO amplitude in NOAAGlobalTempv6 is generally stronger than that in HadCRUT5 prior to 1950, which arises from differences in the underlying sea surface temperature datasets used in these products. Consequently, the Merge-H results based on HadSST are overall more consistent with HadCRUT5. In summary, the AI models perform well in representing the spatial structure of ENSO, they are also capable of reproducing historical ENSO events through the ONI.

L224-253: The detailed analysis comparing Merge-H and Merge-E feels somewhat redundant. Since ERSST is already reconstructed and spatially complete, applying AI reconstruction to it does not add much novel insight. The authors might significantly condense this section to simply justify the choice of Merge-H rather than present it as a core comparative result.

Response: We agree that this section in the original manuscript was overly verbose. We have streamlined it accordingly and provided a clearer justification for the selection of Merge-H.

Changes to the manuscript:

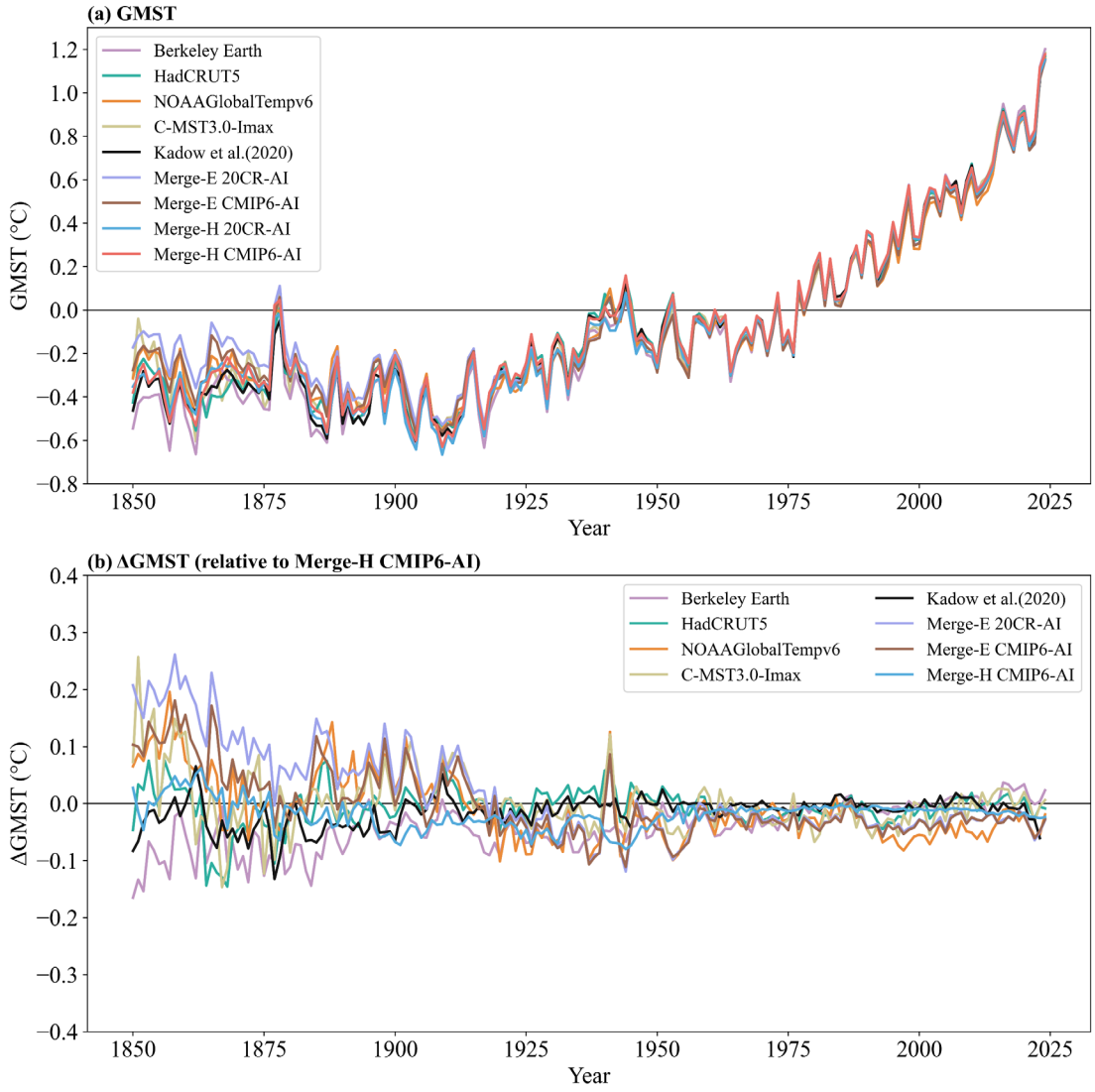

Figure 4: Global mean surface temperature (GMST) time series from 1850 to 2024 (relative to the 1961–1990 climatology). (a) GMST; (b) GMST differences from Merge-H CMIP6-AI.

Table 1: Trends of GMST over different periods and 95% confidence intervals (°C per decade), the global warming level (GWL) denotes the increase of GMST (°C) in 2024 relative to the 1850–1900 reference period, the dataset of Kadow et al. (2020) covers the period up to 2023, and its GWL is GMST in 2023 relative to the 1850–1900 baseline period.

Dataset/Period	1850–2024	1900–2024	1950–2024	1979–2024	GWL
Berkeley Earth	0.070±0.006	0.107±0.008	0.163±0.104	0.206±0.024	1.62
HadCRUT5	0.065±0.006	0.100±0.008	0.156±0.015	0.203±0.023	1.53
NOAAGlobalTempv6	0.058±0.006	0.098±0.008	0.155±0.013	0.196±0.024	1.45
C-MST3.0-Imax	0.061±0.006	0.098±0.008	0.159±0.014	0.210±0.022	1.50
Kadow et al. (2020)	*0.065±0.006*	*0.100±0.008*	*0.151±0.013*	*0.188±0.022*	*1.43*
Merge-E 20CR-AI	0.052±0.007	0.097±0.008	0.157±0.013	0.199±0.023	1.37
Merge-E CMIP6-AI	0.057±0.007	0.099±0.008	0.157±0.013	0.199±0.023	1.43
Merge-H 20CR-AI	0.064±0.006	0.105±0.008	0.155±0.014	0.196±0.023	*1.50*
Merge-H CMIP6-AI	0.064±0.006	0.101±0.008	0.156±0.014	0.199±0.023	1.52

~~*The AI reconstruction results from Merge-E and Merge-H exhibit overall consistent trends; however, certain differences exist during periods with sparse early observations. Prior to the 1890s, the global ST anomalies in the Merge-E reconstruction are generally higher than those in the Merge-H reconstruction. Given the significant contribution of SST to global annual mean temperature variability, this difference is mainly attributable to the differing SST datasets used in the two reconstructions. PConv rely heavily on the spatial structures present in regions with valid observations when reconstructing climate fields containing extensive missing data (Liu et al., 2018; Reichstein et al., 2019; Toms et al., 2020). In the Merge-E scenario, when early land observations are extremely sparse but oceanic grid points exhibit relatively complete spatial coverage (Fig. S3), the dominant spatial gradients, covariance structures, and anomaly patterns learned by the model during training are inherently governed by the ocean. According to the general properties of deep learning, convolutional neural networks preferentially learn the most frequent and statistically stable features in the input data (LeCun et al., 2015). Consequently, during 1850–1890, when land observations are limited, the PConv model is inevitably dominated by oceanic signals during feature extraction. This causes the model to "fill" missing regions with spatial structures resembling those of the ocean, which are typically smoother and warmer, thereby producing systematic biases in the land driven by ocean-dominated features. Around the 1890s, however, the substantial increase in land-based observation stations (Wei et al., 2025) enriches the spatial information over land. As the spatial patterns and variability of land anomalies begin to occupy sufficient weight in the training data, the PConv model becomes capable of simultaneously learning both land and ocean features. This reduces the early-period dominance of oceanic coverage and allows the reconstruction to gradually converge toward a more realistic combined land–ocean spatial structure. In the experiments described above, the SST product ERSSTv6 in Merge-E already incorporates ANN-based spatial infilling, resulting in smoother and more homogeneous oceanic features (Huang et al., 2025a, 2025b). This further amplifies the dominance of oceanic characteristics during feature learning. In contrast, in the Merge-H scenario, both land and ocean fields contain extensive missing data in the early period (Wei et al., 2025; Kennedy et al., 2019), meaning that no single domain provides a strong dominant signal. As a result, the PConv model is more likely to learn spatial*~~

~~*structures jointly constrained by both the land and the ocean domains, thereby reducing the risk of domain-specific biases, particularly biases associated with the ocean. Consequently, the Merge-E reconstruction exhibits pronounced overfitting to oceanic structures before the late nineteenth century, producing warmer biases and ultimately yielding a lower long-term trend and global warming level (GWL) than that of Merge-H (Table 1).*~~ *The AI reconstruction results from Merge-E and Merge-H exhibit certain differences during periods with sparse early observations (Fig. 4b). Prior to 1920, the GMST reconstructed by the Merge-E configuration is generally 0.1–0.2 °C higher than that of Merge-H, which is primarily attributed to differences in the underlying SST datasets used in the two approaches.*

When dealing with climate fields containing large amounts of missing data, partial convolutional neural networks (PConv) rely heavily on the spatial structures of the regions covered by valid observations during the learning process (Liu et al., 2018). When early land-based observations are extremely sparse, while ocean grid cells have relatively complete spatial coverage, the model predominantly learns temperature gradients, covariance structures, and anomaly patterns from oceanic data during training (LeCun et al., 2015; Reichstein et al., 2019; Toms et al., 2020). Consequently, under conditions of scarce land observations during 1850–1920, the Merge-E reconstruction is consistently warmer than other datasets by 0.1–0.3 °C over the low and mid-latitudes (Fig. S10), leading to lower long-term trends and GWL estimates compared with Merge-H (Table 1). This bias reflects an over-reliance on ocean-dominated features under early sparse observational conditions, resulting in systematic inconsistencies relative to other reconstruction schemes.

After 1920, however, as the number of land-based observation stations increases substantially (Wei et al., 2025), the PConv model is able to jointly learn both land and ocean characteristics, thereby reducing the bias introduced by early ocean-dominated training and allowing the reconstruction to gradually converge toward a more realistic coupled land–ocean structure. At the same time, since ERSSTv6 already incorporates ANN-based spatial infilling (Huang et al., 2025a, 2025b), the additional information introduced by the AI reconstruction in Merge-E is limited. Based on the above results, we adopt the more physically consistent merging scheme, the Merge-H product generated by combining C-LSAT2.1 with HadSST4, as the primary focus of our analysis.

L253-254: It is difficult to visually distinguish among the various products. Consider analyzing and plotting the residuals (differences) instead, which would make the variations much clearer to the reader.

Response: We agree that the overlap of multiple time series in the original figures made it difficult for readers to clearly discern the differences among them. Therefore, in the revised manuscript, we plot the differences of other datasets relative to Merge-H CMIP6-AI in "Figure 4 (GMST)," "Figure 5 (Zonal Mean Temperature Comparison)," "Figure 7 (Regional Land Temperature Series)," and "Figure 8 (Antarctic Temperature Series)," thereby providing a clearer representation of the discrepancies among the datasets.

L255-256: Only one out of the four AI reconstructions shows a GWL > 1.5. Given this, the conclusion drawn here feels too confident. It would be appropriate to discuss the spread/uncertainty among the models.

Response: The GWL of Merge-H 20CR-AI reported in Table 1 of the original manuscript was a typographical error and has been corrected to 1.50 °C in the revised version. For the two cases in Merge-E where the GWL does not reach 1.5 °C, we have provided a detailed explanation in the revised manuscript, clarifying that this discrepancy arises from differences in the sea surface temperature datasets rather than from model-induced uncertainty.

We agree that systematically quantifying model uncertainty or spread is important for assessing the reliability of the reconstruction results. However, the PConv-based reconstruction framework employed in this study is inherently a deterministic single-inference model and does not directly provide a mechanism for characterizing model uncertainty. Rigorous quantification of uncertainty (e.g., through repeated stochastic sampling or probabilistic modeling) would require the introduction of additional methodological frameworks, such as ensemble approaches or probabilistic models, which fall beyond the scope of the present study.

We are currently investigating and exploring more appropriate approaches for uncertainty quantification, with the aim of conducting a systematic analysis and assessment of this issue in future work.

L259-270: Inferring historical climate dynamics (i.e., ENSO) purely from the Global Mean Temperature (GMT) time series in Fig. 3 seems to be an over-interpretation. The claim that this "further demonstrating the robustness and reliability of the AI framework" feels overstated.

Response: We agree with this comment and have reduced this section accordingly.

Changes to the manuscript: ~~At the decadal scale, all reconstruction sequences clearly reproduce the pronounced global warming associated with the extreme 1876–1877 El Niño event, which led to the warmest years prior to the 1940s~~

Tables 2 & 3: Please add the other reference datasets to these tables for comparison. Without them, it is difficult to objectively assess the performance and reliability of the AI reconstructions.

Response: We have added Berkeley Earth, HadCRUT5, and Kadow et al. (2020) datasets to Table 2 for comparison. In addition, considering that the revised manuscript already includes validation analyses of the datasets in terms of “zonal temperature comparisons,” “global warming magnitude,” and “regional land temperature series,” the “seasonal temperature series” in the original manuscript provided relatively limited additional information. Therefore, taking into account the manuscript length, we have removed Figure 5, Table 3, and the corresponding analyses related to the “seasonal temperature series.”

Changes to the manuscript:

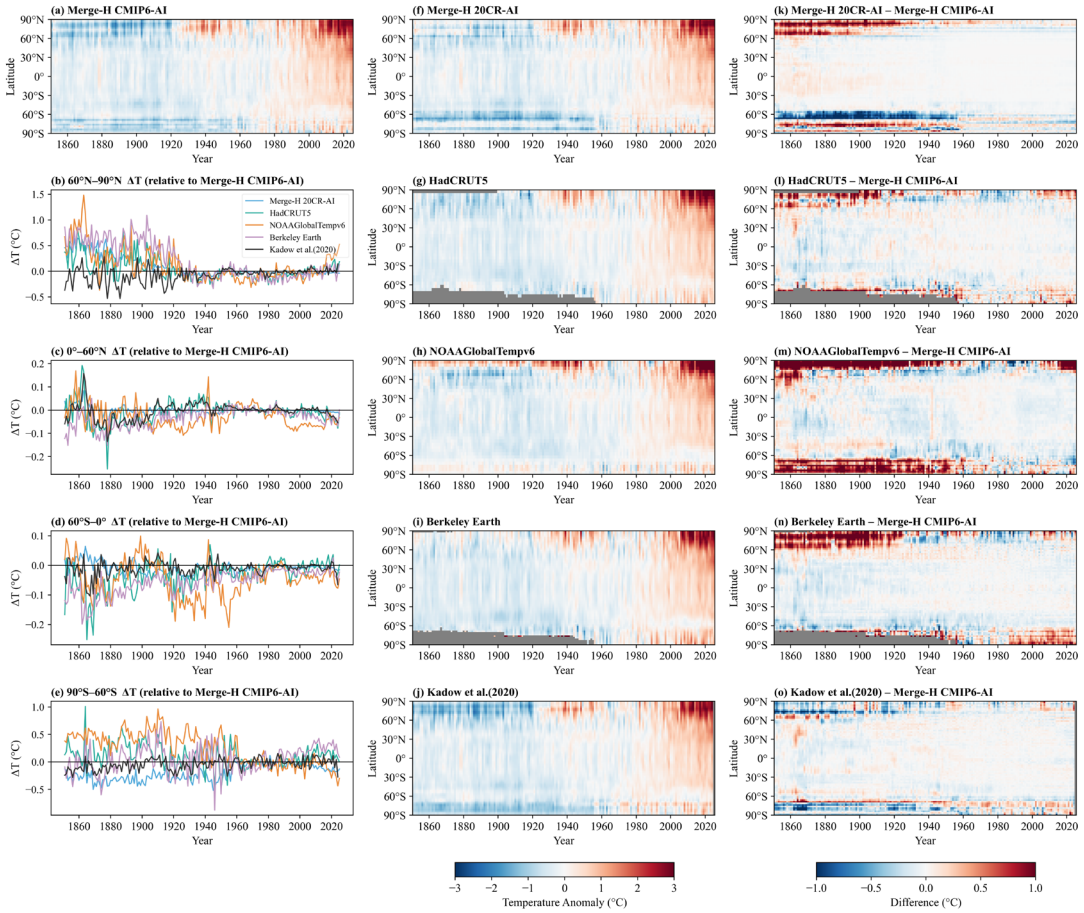

Figure 5 Zonal temperature comparison. (a) Zonal-mean temperature anomalies from Merge-H CMIP6-AI; (b–e) time series of zonal-mean temperature anomaly differences between other datasets and Merge-H CMIP6-AI across different latitude bands; (f–j) same as (a) for Merge-H 20CR-AI, HadCRUT5, NOAAGlobalTempv6, Berkeley Earth, and Kadow et al. (2020), respectively; (k–o) zonal-mean temperature anomaly differences between other datasets and Merge-H CMIP6-AI.

In the zonal temperature comparison across different datasets (Fig. 5), it can be seen that within the 60°S–60°N band, where observational sampling is relatively sufficient, the differences between each dataset and Merge-H CMIP6-AI generally remain within ±0.1°C to ±0.15 °C prior to the early 20th century (Fig. 5c, 5d). The Northern Hemisphere, characterized by a larger land fraction and denser observational networks, is substantially better constrained by observations than the predominantly ocean-covered Southern Hemisphere. Consequently, after 1900, the inter-dataset differences within the equator–60°N region rapidly decrease to within 0.1 °C. In contrast, within the equator–60°S region, this convergence is more delayed, with differences only gradually reducing to within 0.1 °C after 1950. Notably, only NOAAGlobalTempv6 exhibits a persistent cold bias of approximately 0.2 °C relative to the other benchmark datasets during 1920–1960 in this region, likely reflecting weaker constraints over the Southern Ocean.

In the Arctic region (Fig. 5b), prior to 1920, both Merge-H CMIP6-AI and Kadow et al. (2020) exhibit a cold bias of approximately 0.6 °C relative to NOAAGlobalTempv6 and Berkeley Earth, while this bias is smaller in Merge-H 20CR-AI. It is also observed that the AI reconstructions (Fig. 5l–5n) show varying degrees of cold bias in the Arctic compared with other datasets before 1920.

After 1930, as Arctic observations increase, the differences among datasets rapidly decrease to within 0.2 °C.

In the Antarctic region (Fig. 5e), during 1850–1960, Merge-H 20CR-AI exhibits a cold bias of approximately 0.7 °C relative to other datasets, with a pronounced cold anomaly near 60°S (Fig. 5f, 5k). In the early period of extremely sparse observations over the Southern Ocean, the Merge-H 20CR-AI reconstruction shows a relatively strong systematic cold bias. The Kadow et al. (2020) product and Merge-H CMIP6-AI also remain approximately 0.5 °C colder than NOAAGlobalTempv6, which provides full Antarctic coverage, while the larger variability in HadCRUT5 and Berkeley Earth is likely due to the lack of full spatial extrapolation south of 60°S prior to 1960. As Antarctic observations increase around 1961, differences between the two AI reconstructions rapidly converge to within 0.2 °C. Combined with the validation results in Fig. S6f and S6h, it can be further seen that Merge-H 20CR-AI exhibits a larger number of high-RMSE regions in Antarctica than Merge-H CMIP6-AI, indicating greater instability in its polar reconstruction capability prior to 1961.

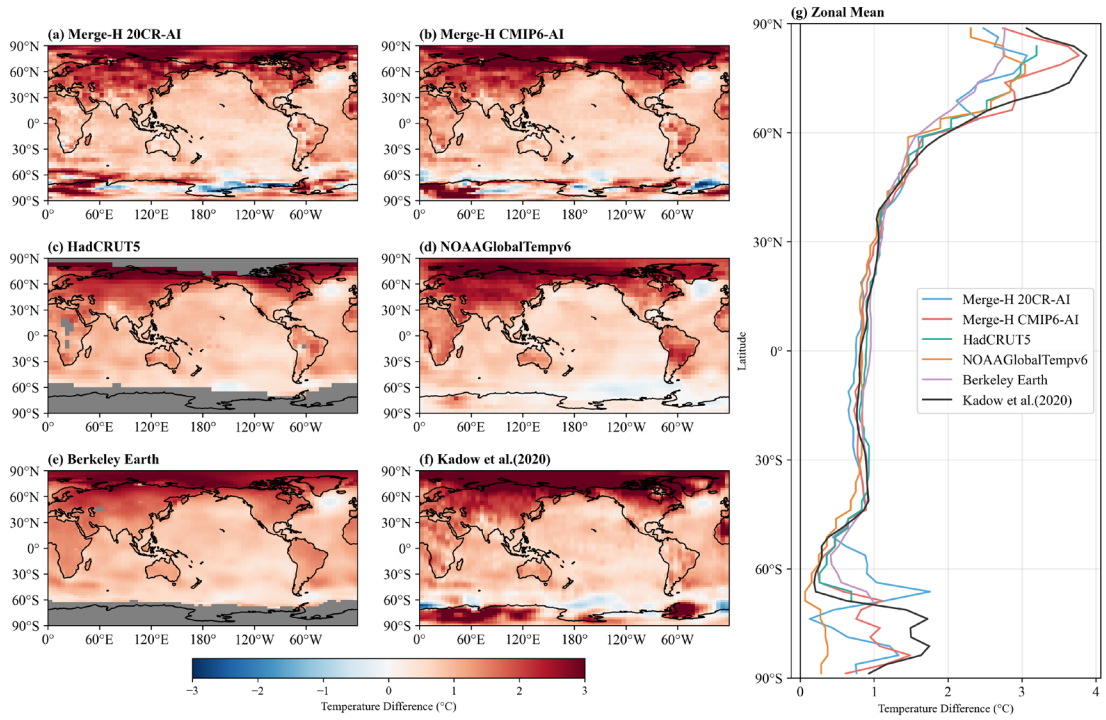

Figure 6 Global warming pattern. (a–f) Spatial distribution of mean global temperature warming over 2005–2020 relative to the 1850–1900 baseline period, (g) zonal-mean profile of warming magnitude.

Under the context of global warming, the spatial patterns of warming reconstructed by the two AI models are generally consistent with those of other datasets (Fig. 6). The main discrepancies are concentrated south of 60°S, where the AI-based reconstructions (Fig. 6a, 6b, 6f) exhibit spatial discontinuities in warming relative to NOAAGlobalTempv6. This may be attributed to the inherently

less smooth spatial temperature fields produced by the AI reconstruction during 1850–1900 (Fig. 1 e1, e2, f1, f2).

As shown in Fig. 6g, between 60°S and 80°S, Merge-H 20CR-AI displays zonal variations of ±1 °C relative to Merge-H CMIP6-AI, while the Kadow et al. (2020) reconstruction also shows a pronounced warming increase of up to 1.5 °C near 80°S, which is higher than all other datasets. This indicates that Merge-H 20CR-AI is less stable in this region, consistent with the findings in the previous zonal temperature comparison section. The representation of Arctic amplification by the AI models also differs among datasets (Fig. 6g), with inter-dataset differences of approximately ±1 °C near 80°S.

The two AI reconstructions based on HadSST SST data generally show higher consistency in warming patterns with HadCRUT5 and Berkeley Earth, which also use HadSST, compared to NOAAGlobalTempv6, which is based on ERSSTv6. In particular, near 110°W in the Southern Ocean adjacent to Antarctica, NOAAGlobalTempv6 exhibits a larger cooling magnitude than the other datasets. Although all datasets capture the North Atlantic "cold blob," the stronger cooling north of this region in NOAAGlobalTempv6 may represent an artifact of its underlying analysis procedure (Chan et al., 2025), a feature not observed in the other datasets.

Overall, the AI-based reconstructions are able to reasonably reproduce the magnitude and spatial distribution of global warming. However, caution is still warranted in regions covered by sea ice and subject to extremely sparse observational constraints, particularly in the polar regions.

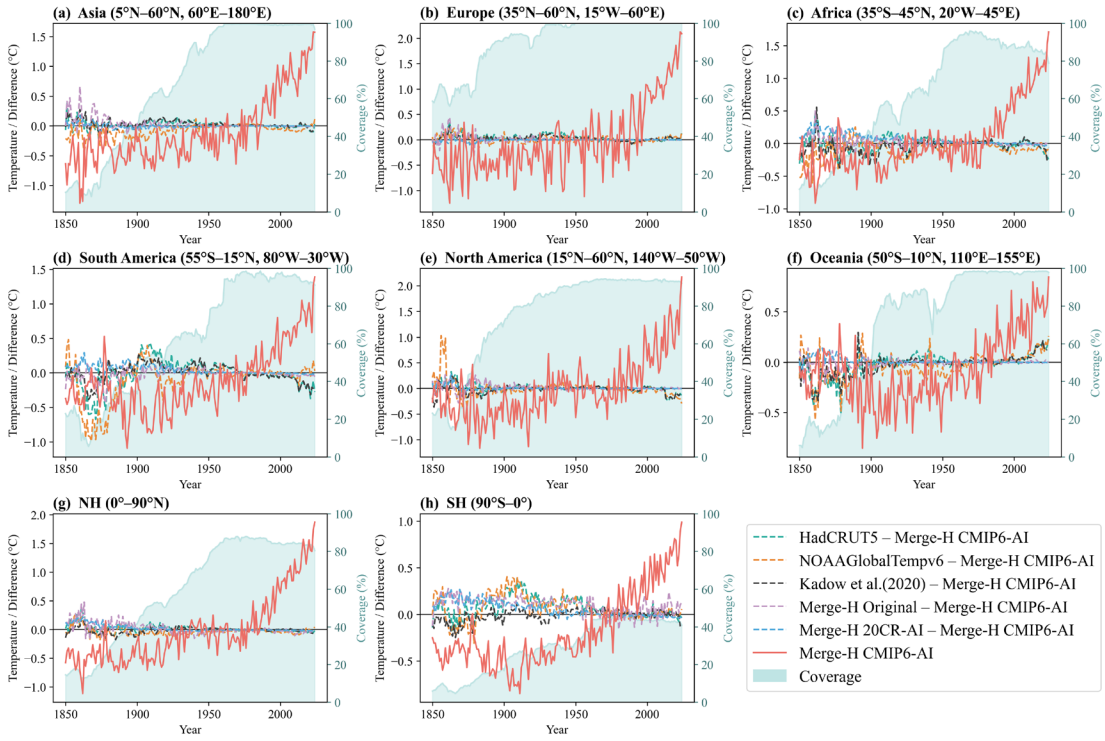

Figure 7: Annual mean surface temperature anomaly time series from 1850 to 2024 in different regions under the Merge-H Scenario. (a–h) Asia, Europe, Africa, South America, North America, Oceania, the Northern Hemisphere, and the Southern Hemisphere. Coverage indicates the original data availability in each region.

~~*Global reconstruction of land regions is one of the main focuses of this study. We first analyse the long-term variations of ST in different land regions before and after reconstruction (Fig. 4). In this analysis,*~~

~~*the original Merge-E dataset is compared with its two AI-based reconstructions. Considering the extremely sparse observational coverage over Antarctica, this subsection does not discuss the Antarctic region, which will be addressed in detail in Section 4 (Antarctic Reconstruction Results). As a reference for reconstruction performance, the effective data coverage of the original datasets in each region is also presented. Table 2 provides the linear trends of annual mean ST and their 95% CI for different regions from 1850 to 2024.*~~ *One of the key focuses of this study is the reconstruction of global land surface data (Fig. 7, Table 2). To evaluate the reconstruction performance, we also present the effective data coverage of the original datasets across different regions (Coverage in Fig. 7). In addition, Table 2 summarizes the linear trends of annual mean surface temperature over 1850–2024 for different regions, along with their 95% confidence intervals. The performance of the Antarctic land reconstruction is further evaluated in detail in Section 4, "Antarctic Reconstruction Results."*

Table 2: Temperature trends and 95% confidence intervals (CI) in different regions from 1850 to 2024 (°C per decade).

Region/Dataset	HadCRUT5	NOAA GlobalTempv6	Kadow et al. (2020)	Merge-H Original	Merge-H 20CR-AI	Merge-H CMIP6-AI
Asia	0.088±0.009	0.101±0.010	0.084±0.010	0.083±0.010	0.090±0.010	0.091±0.010
Europe	0.092±0.014	0.097±0.010	0.092±0.014	0.094±0.014	0.099±0.014	0.097±0.014
Africa	0.073±0.008	0.082±0.008	0.079±0.008	0.069±0.009	0.060±0.009	0.074±0.009
South America	0.067±0.007	0.090±0.009	0.058±0.008	0.068±0.009	0.055±0.010	0.064±0.010
North America	0.077±0.012	0.073±0.013	0.079±0.012	0.080±0.012	0.080±0.013	0.083±0.012
Oceania	0.054±0.008	0.050±0.009	0.050±0.008	0.041±0.008	0.035±0.009	0.038±0.008
NH	0.091±0.009	0.096±0.010	0.097±0.009	0.087±0.010	0.089±0.010	0.097±0.010
SH	0.059±0.006	0.059±0.006	0.068±0.006	0.054±0.007	0.050±0.008	0.063±0.007

~~*Figure 4 shows that as observational coverage increases, the differences between different reconstruction results decrease markedly. This feature is particularly evident in regions with increasingly abundant observational data, where the AI-reconstructed annual variations become more stable. In contrast, in periods or regions with sparse early observations and low data coverage, discrepancies among reconstructions are more pronounced. The differences between the two AI reconstructions are mainly concentrated in regions such as Africa and South America, where early observations were scarce. Due to the large extent of missing data in these areas, which corresponds to large holes in the input images that need to be filled, the accuracy of the AI reconstructions is significantly affected (Liu et al., 2018). Considering Table 2, the differences between the two AI reconstructions are unavoidable in some regions during periods of low early data coverage. For the Merge-H reconstruction, the two AI models exhibit slight differences in long-term trends, with 20CR-AI showing a lower temperature trend in all regions except Europe compared to CMIP6-AI. However, the trends remain within reasonable ranges relative to the pre-reconstruction data. Except for regions with substantial early data gaps, such as Africa, South America, and Oceania, most AI-reconstructed temperature trends exhibit stronger warming compared with pre-reconstruction trends.*~~

~~*On long timescales, all regions show significant warming trends, particularly since the 1970s, when the warming rate accelerated. Nevertheless, warming is uneven across continents due to differences in*~~

~~*response scales and persistence (Li et al., 2022). According to Table 2, in the Merge-H reconstruction scenario, Europe experienced the most pronounced warming between 1850 and 2024, with trends of 0.099 ± 0.014 and 0.097 ± 0.014 °C per decade, whereas Oceania experienced the smallest warming, with corresponding trends of 0.035 ± 0.009 and 0.038 ± 0.008 °C per decade. The warming rate in the SH is noticeably lower than in the NH. This north–south asymmetry under global warming results from significant differences in land–ocean distribution, with the Southern Ocean absorbing the majority of heat, leading to greater climate inertia in the SH (Hansen et al., 2010).*~~

~~*A comprehensive analysis of regional ST anomaly time series indicates that NH land areas contribute most significantly to global land warming. Within the NH, Europe, Asia, and North America are the primary contributors. In conditions of extremely low data coverage, the both reconstructions inevitably exhibit biases. However, as coverage increases, the AI reconstruction experiments show good consistency and stability in these regions.*~~

Figure 7 shows that as data coverage increases, the differences between the two reconstruction products and other datasets gradually decrease to within 0.1 °C after 1900. This feature indicates that in regions where observational data become progressively denser, the interannual differences among AI-based reconstructions tend to stabilize. In contrast, during earlier periods or in regions with sparse observations and low coverage, discrepancies among reconstruction products are mainly concentrated in Africa and South America, where early observational records are limited. When effective coverage is below 40% before 1900, Merge-H 20CR-AI is overall about 0.2 °C higher than Merge-H CMIP6-AI (Fig. 7c, 7d, 7h), which also leads to lower long-term trends in these regions for the former in Table 2.

Furthermore, during 1860–1870, when effective coverage in Africa, South America, and Oceania (Fig. 7c, 7d, 7f) drops below a critical threshold (~10%), and considering the strong spatial sampling heterogeneity and lack of inland temperature information in these continents (Wei et al., 2025), the AI reconstructions show varying degrees of deviation from NOAAGlobalTempv6 and HadCRUT5. The most pronounced discrepancy occurs in South America (Fig. 7d), where values are approximately 0.8 °C higher than NOAAGlobalTempv6 and 0.5 °C higher than HadCRUT5. However, around 1920, a systematic negative bias of about 0.3 °C appears. Over the full 1850–2024 period (Table 2), the long-term trend of the AI reconstruction in South America is broadly consistent with HadCRUT5. In Oceania, both Merge-H Original and the two AI-based reconstructions show consistently higher values than other datasets, resulting in lower long-term trends (Table 2), which can be attributed to inherent temperature biases in early land surface products prior to 1900.

On long timescales, regional surface temperature anomalies exhibit significant warming trends, particularly since the 1970s, when warming rates accelerate markedly. However, land regions across continents show heterogeneous warming patterns due to differing response timescales and persistence characteristics (Li et al., 2022). As shown in Table 2, the two reconstruction schemes indicate the strongest warming over Europe during 1850–2024, with trends of 0.097 ± 0.014 °C/decade and 0.099 ± 0.014 °C/decade, respectively, while Oceania shows the weakest warming, at 0.035 ± 0.009 °C/decade and 0.038 ± 0.008 °C/decade. The warming rate in the Southern Hemisphere is significantly lower than that in the Northern Hemisphere, reflecting hemispheric asymmetry driven by the large land–ocean contrast and the strong heat uptake by the Southern Ocean, which enhances climate inertia in the Southern Hemisphere (Hansen et al., 2010).

From the integrated analysis of regional surface temperature anomaly series, it is evident that Northern Hemisphere land areas contribute most significantly to global land warming. Within the Northern Hemisphere, Europe, Asia, and North America are the dominant contributors. Under

conditions of extremely low data coverage, both reconstructions inevitably exhibit biases; however, as coverage increases after 1900, the AI-based reconstruction results show improved consistency and stability across these regions.

L285-286: It is difficult to visually confirm in Figure 4 that "the differences between different reconstruction results decrease markedly." Consider specifying timeframes and supporting them with a quantifiable metric.

Response: We have added a clearer temporal constraint and a quantitative analysis for the period during which the differences among the results are significantly reduced.

Changes to the manuscript:

Figure 7 shows that as data coverage increases, the differences between the two reconstruction products and other datasets gradually decrease to within 0.1 °C after 1900.

Figure 5: Please improve the visual clarity of this figure. The lines overlap significantly, making it difficult to distinguish the differences between the datasets.

Response: We acknowledge that in some figures, including Figure 5, overlapping time series make it difficult for readers to distinguish them clearly. Therefore, in the revised manuscript, we have added difference plots showing the temperature anomalies relative to Merge-H CMIP6-AI for the series in Figures 4, 5, 7, and 8.

L361-362: There appears to be a contradiction here: the text states that CMIP6-AI performs better than 20CR-AI, but Figure S6 seems to show the reverse. Please verify and clarify.

Response: The advantages, limitations, and stability of 20CR-AI and CMIP6-AI across different regions have been analyzed and discussed in Figures 5 and 6 of the revised manuscript, and corresponding conclusions have been provided.

L368-371: Please rephrase this statement for clarity. It is not entirely clear how the reference datasets are being quantitatively or qualitatively compared with the station data.

Response: We have expanded and clarified the description of both quantitative and qualitative comparison methods with the reference datasets and station observations.

Changes to the manuscript:

The comparisons include a quantitative assessment of Antarctic annual mean temperature time series (Fig. 8) and Antarctic subregional temperature series (Fig. S13), as well as a qualitative analysis of the spatial distribution of temperature trends (Figs. 9 and S14).

L389-390: Please specify the exact statistical method used to measure the significance here.

Response: We have explicitly clarified the significance testing method in the Section 2.4 “Data validation methods”.

Changes to the manuscript:

The GMST series of the benchmark datasets used in this study follow the standard products released

by their respective official sources. All regional annual mean temperature series are first computed using area-weighted spatial averaging, followed by temporal averaging on an annual basis, in accordance with the World Meteorological Organization (WMO) methodology. Linear trends are estimated using ordinary least squares, and their statistical significance is assessed using a two-sided t-test. The Ocean Niño Index (ONI) is defined following the NOAA Climate Prediction Center standard, as the 3-month running mean of temperature anomalies over the Niño 3.4 region (5°S–5°N, 170°W–120°W), calculated relative to a centered 30-year base periods that is updated every 5 years.

L423-424: The conclusion that the products show high spatiotemporal consistency is not fully supported because it has not been benchmarked against other references or methods. Basing this conclusion solely on GMT and regional trends is clearly insufficient.

Response: We acknowledge that the original statement of "high spatiotemporal consistency" was overly confident and exceeded the support provided by the available evidence. Therefore, we have conducted more detailed comparative analyses for Antarctica, including subregional temperature time series and the spatial distribution of seasonal trends. Accordingly, we have revised the conclusion to adopt a more cautious interpretation of "high spatiotemporal consistency."

Changes to the manuscript:

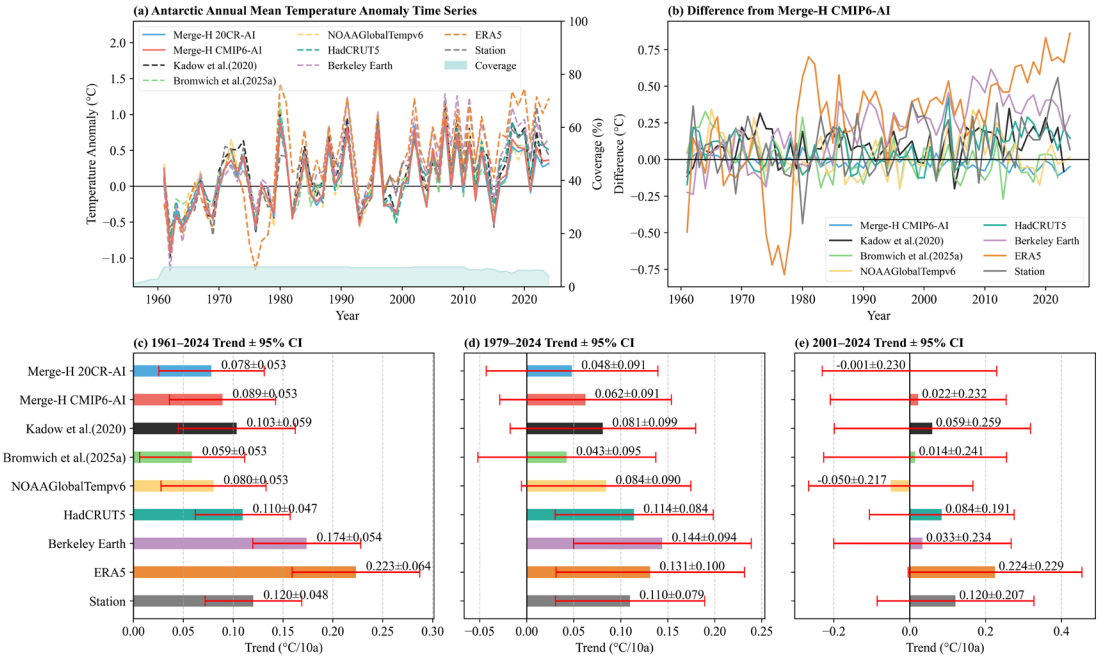

***Figure 8*: (a) Annual mean surface temperature anomaly time series over Antarctica from 1961 to 2024; *(b) Temperature difference from Merge-H CMIP6-AI;* (c–e) Linear trend of annual mean temperature and 95% confidence interval (°C per decade) during 1961–2024, 1979–2024 and 2001–2024.**

Figure 8 presents the annual mean ST anomalies over the Antarctic continent for 1961–2024, along with the linear temperature trends and their 95% CI for three representative periods. To comprehensively evaluate the reliability and performance of the AI reconstructions, the results were systematically compared with the arithmetic mean of 81 Antarctic observational stations used in this

study (Station in Figure 6), the reconstructions from Kadow et al. (2020) and Bromwich et al. (2025a), HadCRUT5, NOAAGlobalTempv6, Berkeley Earth, and the ERA5 reanalysis product. The comparisons include a quantitative assessment of Antarctic annual mean temperature time series (Fig. 8) and Antarctic subregional temperature series (Fig. S13), as well as a qualitative analysis of the spatial distribution of temperature trends (Figs. 9 and S14).

~~*Overall, the AI reconstructions exhibit high consistency with multiple observational and reconstruction products in terms of interannual variability (Fig. 8a). In particular, the reconstructed anomalies closely track the phases of strong cold and warm years, indicating that the AI method reliably captures interannual climate signals in the Antarctic region. During 1961–2024, the linear warming trends derived from the two AI reconstruction schemes are 0.078 ± 0.053 and 0.089 ± 0.053°C per decade, respectively (Fig. 8c). The trends are similar in magnitude, show minor differences, and are both statistically significant at the 0.05 level, demonstrating a pronounced warming trend over the Antarctic continent since the 1960s.*~~

~~*Notably, since 1979, the warming trends in the AI reconstructions remain non-significant, consistent with Bromwich and NOAAGlobalTempv6, whereas HadCRUT5, Berkeley Earth, and ERA5 still show significant warming trends, in agreement with the arithmetic mean of the station data (Station) (Fig. 6c). The higher trend estimates in these products may stem from biases introduced by statistical and reanalysis methods under sparse observational conditions in polar regions. Previous studies indicate that the ERA5 reanalysis system exhibits substantial seasonal biases in near-surface Antarctic temperatures, likely due to weak constraints on surface turbulent processes and increased assimilation errors under strong inversions and low friction conditions (Garza-Girón et al., 2024; Yang et al., 2025). Additionally, Berkeley Earth's spatial statistical extrapolation may produce systematic overestimation in high-latitude Antarctica (Rohde et al., 2020), although the exact contribution of statistical and assimilation methods to these biases remains to be quantified. These results highlight that systematic differences among data sources must be carefully considered when assessing long-term climate trends in polar regions.*~~

Overall, the AI reconstructions show high consistency with multi-source observations and reconstruction products (Bromwich et al., 2025a) in capturing interannual variability, particularly in the phase evolution of temperature anomalies during strong warm and cold events, which is largely synchronized with observations and representative reanalysis datasets (Fig. 8a). The differences between the two AI reconstruction approaches and Bromwich et al., NOAAGlobalTempv6, and HadCRUT5 are generally within 0.25 °C (Fig. 8b).

During 1961–2024, the linear warming trends derived from the two AI reconstructions are 0.078 ± 0.053 °C/decade and 0.089 ± 0.053 °C/decade, respectively. These magnitudes (Figs. 8c–8e) are close to those reported by Bromwich et al. (2020) and NOAAGlobalTempv6, and are lower than those from HadCRUT5, Berkeley Earth, ERA5, and station-based datasets. However, it should be noted that the station data used to construct the Station-based series exhibit a highly uneven spatial distribution, with observations primarily concentrated in regions experiencing stronger warming (Fig. S3, Fig. 9). This sampling structure causes the Station-based estimates in Fig. 8 to better represent localized signals rather than the pan-Antarctic mean state, thereby potentially leading to an overestimation of the warming trend at the continental scale of Antarctica.

Since 1979, the trends from the AI reconstructions remain statistically insignificant, consistent with Bromwich et al. and NOAAGlobalTempv6, whereas HadCRUT5, Berkeley Earth, and ERA5 still exhibit statistically significant warming trends (Fig. 8d). This apparent overestimation in these products may

arise from limited observational sampling over Antarctica and spatial smoothing or extrapolation procedures, resulting in reduced regional representativeness (Morice et al., 2021; Rohde et al., 2020). As shown in Fig. S13e and S13h, this bias mainly originates from warming contributions over East Antarctica. In addition, Bromwich et al. (2024) show that the pronounced Antarctic warming trend in ERA5 prior to 1979 may be overestimated, primarily due to a cold bias in the ECMWF model. Given the extremely sparse observations over the Southern Ocean and limited assimilation capability of early satellite records, this bias cannot be effectively constrained. After 1979, ERA5 exhibits further enhanced warming trends near the 0° longitude coastal sector, the Ross Ice Shelf, and Marie Byrd Land, which further increases the overall temperature trend. These results highlight that systematic differences among datasets must be treated with caution when assessing long-term climate change trends in polar regions.

~~*Overall, the AI reconstruction results demonstrate reasonable temporal consistency, trend significance, and magnitude of variability, validating the feasibility and potential utility of AI-based climate reconstruction for the Antarctic region.*~~

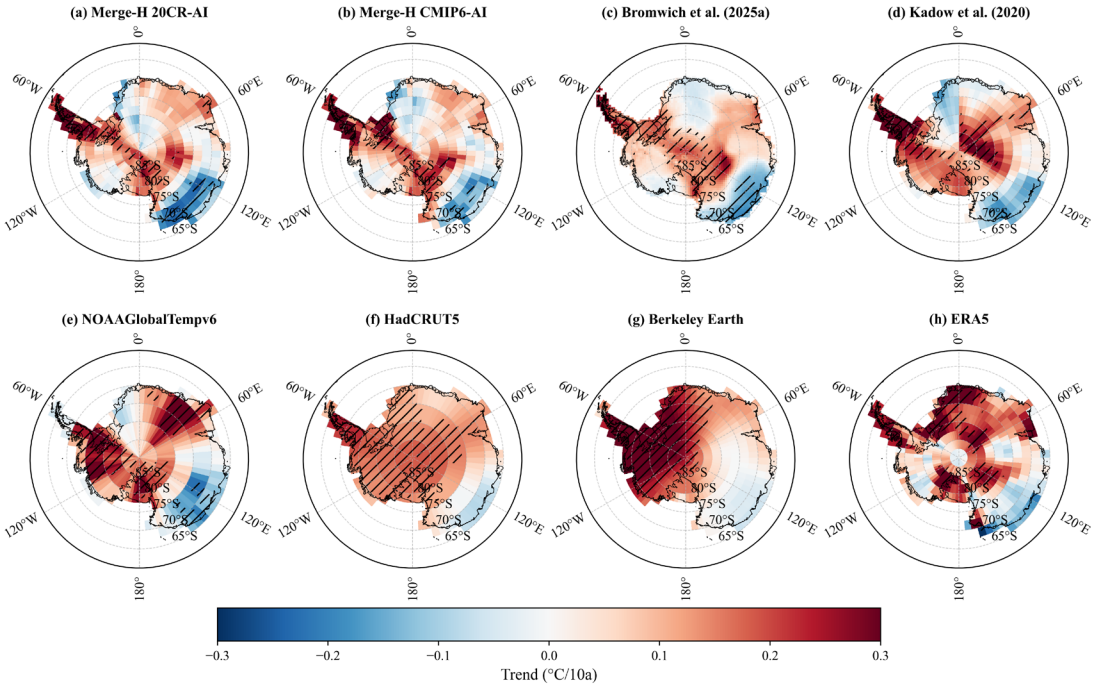

***Figure 9*: Spatial distribution of Antarctic annual mean surface temperature trends for 1979–2024 ("///" indicates regions where the temperature trend is statistically significant at the 0.05 level).**

The spatial pattern of Antarctic temperature change is of critical scientific importance for understanding regional climate variability and its potential driving mechanisms. Due to substantial differences in topography, sea ice coverage, circulation features, and ocean–atmosphere interactions across the Antarctic continent, temperature changes are not uniformly distributed but exhibit complex regional responses (Turner et al., 2005, 2019; Marshall, 2003, 2006). Figure 9 presents the spatial distribution of ST trends and their statistical significance for 1979–2024 derived from the two AI reconstruction schemes, together with the results from other representative datasets. ~~*Overall, both AI reconstruction schemes reveal significant warming over the Antarctic Peninsula, near the Ronne Ice Shelf, and the northeastern Ross Ice Shelf, while the Wilkes Land coast shows significant cooling. These spatial patterns are broadly consistent with the reconstructions of Bromwich, the NOAAGlobalTempv6*~~

~~*dataset, and reported observations or model simulations in these regions (Vaughan et al., 2003; Wang et al., 2025; Darelius et al., 2016; Clem et al., 2020; Sheehan et al., 2024), demonstrating the reliability of AI reconstructions in capturing key climate signals in Antarctica.*~~

~~*In contrast, the ERA5 product exhibits some spatial discrepancies: it shows significant warming near Queen Maud Land and west of the Ronne Ice Shelf, whereas cooling along the Wilkes Land coast is not statistically significant. The HadCRUT5, Berkeley Earth, and GISTEMPv4 global temperature reconstructions rely on limited ground-based observations in Antarctica and extend spatially using different statistical methods (optimal interpolation, EOF, or spatially weighted averaging) (Morice et al., 2021; Rohde et al., 2020; Lenssen et al., 2019). As a result, their temperature trend fields are smoother and have weaker spatial gradients. All three datasets indicate significant warming over the Antarctic Peninsula, with HadCRUT5 and Berkeley Earth also showing extensive warming near the South Pole, while GISTEMPv4 indicates significant warming east of the Ross Ice Shelf and in eastern Queen Maud Land. Notably, NOAAGlobalTempv6 also shows warming in eastern Queen Maud Land, whereas the two AI reconstruction schemes, reconstruction from Bromwich, and ERA5 reveal positive but statistically insignificant trends in this region, which lies in the transitional zone of Antarctic continental temperature variability.*~~

~~*In summary, the AI reconstructions successfully capture the main regional temperature trends and their statistical significance across Antarctica, reflecting climate signals consistent with observations and reanalysis data while exhibiting plausible spatial variability in data-sparse areas. This demonstrates that AI-based climate reconstruction not only provides high temporal consistency but also possesses strong potential for spatial applications, offering a novel tool for understanding Antarctic climate change processes and their underlying drivers.*~~

Overall, all datasets consistently reveal significant warming over the Antarctic Peninsula, the vicinity of the Ross Ice Shelf, and the northeastern sector of the Ross Ice Shelf (see Fig. S12 for Antarctic geographical references), while significant cooling is observed along the coast of Wilkes Land. In particular, the Antarctic Peninsula exhibits pronounced warming, with the AI reconstructions yielding trends of 0.301 ± 0.115 °C/decade and 0.311 ± 0.113 °C/decade, respectively (Fig. S13). For the annual mean time series, all datasets show biases of approximately ±1 °C. In the relatively large regions of West Antarctica and East Antarctica, the differences between the AI reconstructions and other datasets are within 0.5 °C and 0.25 °C, respectively, and the regional temperature trends are generally consistent with the Station-based estimates.

The spatial pattern of Antarctic temperature trends is broadly consistent with the reconstruction results of Bromwich et al. and NOAAGlobalTempv6, and is also in agreement with observational evidence or model-based studies in these regions (Vaughan et al., 2003; Wang et al., 2025; Darelius et al., 2016; Clem et al., 2020; Sheehan et al., 2024). Notably, the reconstruction of Kadow et al. (2020) (Fig. 9d) exhibits a clear artifact near 0° longitude, where the temperature trend shifts abruptly from significant warming east of 0° to cooling west of it, indicating a pronounced spatial discontinuity. This issue is largely mitigated in the present reconstruction (Figs. 9a, 9b), which, consistent with Bromwich et al. and NOAAGlobalTempv6, does not show widespread significant warming near 0° longitude. This highlights that AI-based reconstructions of Antarctic climate signals require adequate observational sampling support.

In contrast, the ERA5 product shows certain spatial differences. It exhibits significant warming over Queen Maud Land and west of the Ross Ice Shelf, while cooling along the Wilkes Land coast does not pass statistical significance testing. HadCRUT5 and Berkeley Earth display relatively smooth spatial

patterns of temperature trends over Antarctica, with weaker spatial gradients, and fail to capture the cooling signal over Queen Maud Land and the significant cooling along Wilkes Land coastal regions identified in Bromwich et al. (2025a) and NOAAGlobalTempv6. This is also one of the reasons why the temperature trends in Fig. 8d are higher in the former two datasets than in the latter two. It is worth noting that NOAAGlobalTempv6 also shows significant warming in eastern Queen Maud Land, while the AI reconstructions, Bromwich et al., and ERA5 all indicate a certain warming signal in this region. However, since this area lies within a transitional zone of cold and warm variability over the Antarctic interior, the warming signal does not reach statistical significance.

Overall, the AI reconstructions successfully capture the spatial distribution of temperature trends and their statistical significance across major Antarctic regions. They not only reproduce climate signals broadly consistent with observational and reanalysis datasets, but also exhibit reasonable spatial variability in data-sparse regions. This suggests that AI-based climate reconstruction methods provide robust and reliable temporal evolution of Antarctic climate after 1961, when observational constraints become available, and also demonstrate good skill in representing spatial structures. These results provide a new approach for better understanding Antarctic climate variability and its underlying driving mechanisms.

Section 5 (Limitations and future perspectives): As the ESSD journal is data-focused, this section should focus specifically on data limitations, uncertainties, and the irreplaceability of the generated datasets. Broad discussions of the future of AI methodologies are somewhat outside the journal's primary scope and should be minimized.

Response: We appreciate this comment. As a data description paper, we should place greater emphasis on the inherent limitations of the dataset. Accordingly, in the "Limitations and future perspectives" section, we have reduced the discussion on AI development.

Changes to the manuscript:

~~*Esteves et al. (2023) reported that scaling spherical convolutional neural networks (Scaling Spherical CNNs) can directly model data on the sphere, achieving competitive results in weather forecasting tasks, demonstrating the potential of spherical convolutional frameworks in atmospheric sciences. However, high-resolution global climate reconstruction with such methods remains significantly limited by current GPU memory capacity and the high cost of high-performance computing. Overall, balancing spherical geometric accuracy with computational feasibility while maintaining global spatial continuity is a key direction for future AI climate reconstruction. Integrating spherical deep learning structures with physical constraints may enable more realistic and accurate reconstructions of polar climate variability.*~~ *To address this issue, Esteves et al. (2023) reported that scaling spherical convolutional neural networks (Scaling Spherical CNNs) can be used to model spherical data; however, their application remains constrained by current GPU memory limitations and the high computational cost of high-performance computing resources.*

~~*Furthermore, combining AI's ability to capture nonlinear features and local spatial patterns with the interpretability, computational transparency, and uncertainty quantification advantages of statistical methods, forming "AI plus physics-driven" or "AI plus statistical fusion" frameworks, may represent a promising direction for climate reconstruction.*~~

~~*AI method demonstrates strong potential and scalability in climate reconstruction, yet its application in modeling and reconstructing climate systems remains an evolving area. The physical consistency and long-term stability of current AI reconstruction outputs still require further validation under more stringent constraints. Future research should incorporate climate dynamics and energy balance constraints while preserving the nonlinear fitting advantages of deep learning to ensure interpretability and generalizability in complex climate contexts. With continued advancement in high-performance computing and climate big data, AI reconstruction methods may achieve higher resolution and stronger constraint global climate field reconstructions. By integrating spherical neural networks, physical information, and multi-source data fusion frameworks, future AI reconstruction systems could more accurately reproduce climate evolution in data-sparse regions, supporting studies on polar amplification, local climate change, and data monitoring, detection, and evaluation. Overall, AI is expected to play an increasingly important role in climate science, providing a robust technological foundation for understanding past and projecting future global and regional climate change.*~~

Overall, while the core concept of this study is compelling and timely, the manuscript's current presentation, the robustness of its validation methods, and the understanding of the underlying datasets do not yet meet the necessary threshold for publication.

Response: We sincerely thank the reviewer again for the highly constructive and comprehensive comments. These valuable suggestions not only helped us identify shortcomings in the original manuscript, but also prompted us to systematically refine and improve the structure, validation framework, and understanding of the dataset.

In response to the key issues raised above, we have made substantial revisions and additions in the revised manuscript, including: optimizing the overall structure and presentation to improve clarity and logical flow; strengthening the validation and evaluation methods by conducting more rigorous and systematic assessments of the reconstruction results from multiple scales and perspectives; and further deepening the analysis and discussion of the characteristics and limitations of the underlying datasets.

Through these revisions, we aim to achieve higher standards in terms of result reliability, methodological rigor, and scientific interpretation, thereby providing stronger support for the main conclusions of this study. We believe that the revised manuscript has been significantly improved in both completeness and persuasiveness. We greatly appreciate the reviewer’s careful evaluation and professional guidance, which have been instrumental in enhancing the quality of this work.

Added References

Bromwich, D., Ensign, A., Wang, S. and Zou X.: Major Artifacts in ERA5 2-m Air Temperature Trends Over Antarctica Prior to and During the Modern Satellite Era. Geophys. Res. Lett., 51(21), https://doi.org/10.1029/2024GL111907, 2024.

Chan, D., Chan, S. C., Siddons, J. T., Cable, A., Faulkner, A., Kent, E. C., Gebbie, G., and Huybers, P.: DCENT- I: A Globally Infilled Extension of the Dynamically Consistent ENsemble of Temperature Dataset. Geosci. Data J., 13:e70054. https://doi.org/10.1002/gdj3.70054, 2026.

Chan, D., and Huybers, P.: Correcting Observational Biases in Sea Surface Temperature Observations Removes Anomalous Warmth During World War II. J. Clim., 34(11): 4585–4602. https://doi.org/10.1175/JCLI-D-20-0907.1, 2021.

Kent, E. C., and Kennedy, J. J.: Historical Estimates of Surface Marine Temperatures. Annu. Rev. Mar. Sci., 13: 283–311. https://doi.org/10.1146/annurev-marine-042120-111807, 2021.

Li, Z., Li, Q., Jiao, B., Xu, Q., Wei S., Ru, X., Si, P., Chao, L., Zhang, H., Lin, J., Liao, L., Zhang, H., Huang, B., and Jones, P.: An integrated uncertainty framework for the China-MST 3.0 global surface temperature dataset. J. Geophys. Res. Atmos., accepted.

Xie, S., Wei, S., Li, Z., and Li, Q.: Recent Changes in Antarctic Surface Air Temperature Based on the Fusion of Satellite and In-situ Measurements. Theor. Appl. Climatol., https://doi.org/10.1007/s00704-026-06130-0, in press.

Yu, J. Y. and Kim, S. T.: Identifying the types of major El Niño events since 1870. Int. J. Climatol. 33, 2105–2112. https://doi.org/10.1002/joc.3575, 2013.

Changes to Supplementary Information

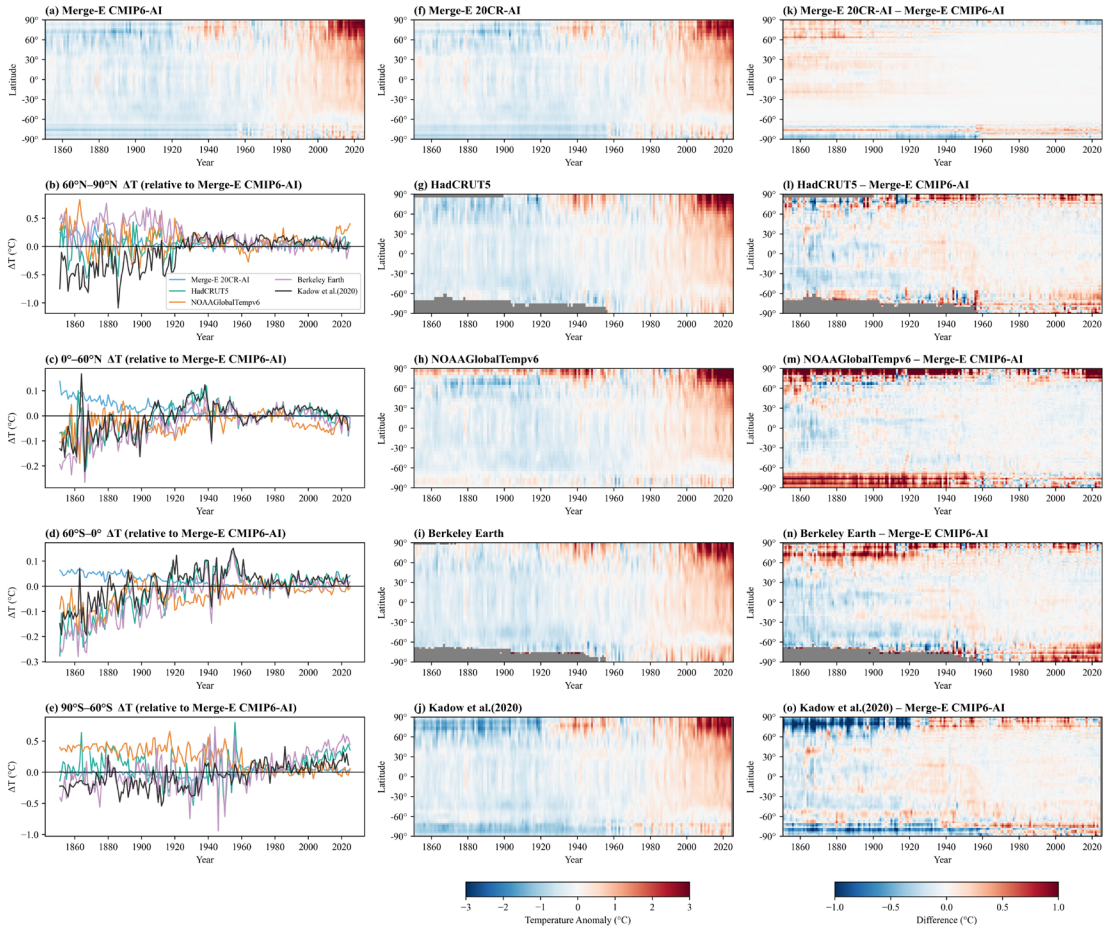

Figure S10 Zonal temperature comparison. (a) Zonal-mean temperature anomalies from Merge-E CMIP6-AI; (b–e) time series of zonal-mean temperature anomaly differences between other datasets and Merge-E CMIP6-AI across different latitude bands; (f–j) same as (a) for Merge-E 20CR-AI, HadCRUT5, NOAAGlobalTempv6, Berkeley Earth, and Kadow et al. (2020), respectively; (k–o) zonal-mean temperature anomaly differences between other datasets and Merge-E CMIP6-AI.

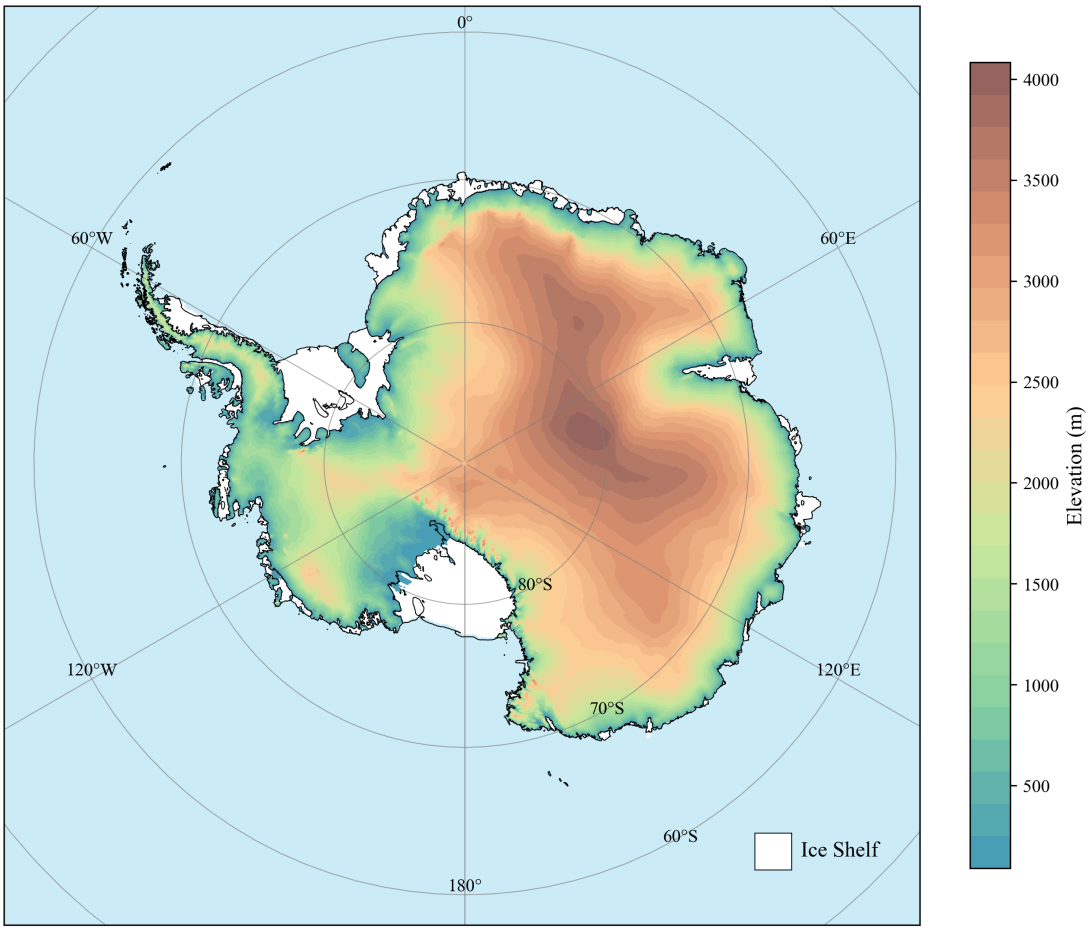

Figure S11 Topographic elevation map of the Antarctic continent.

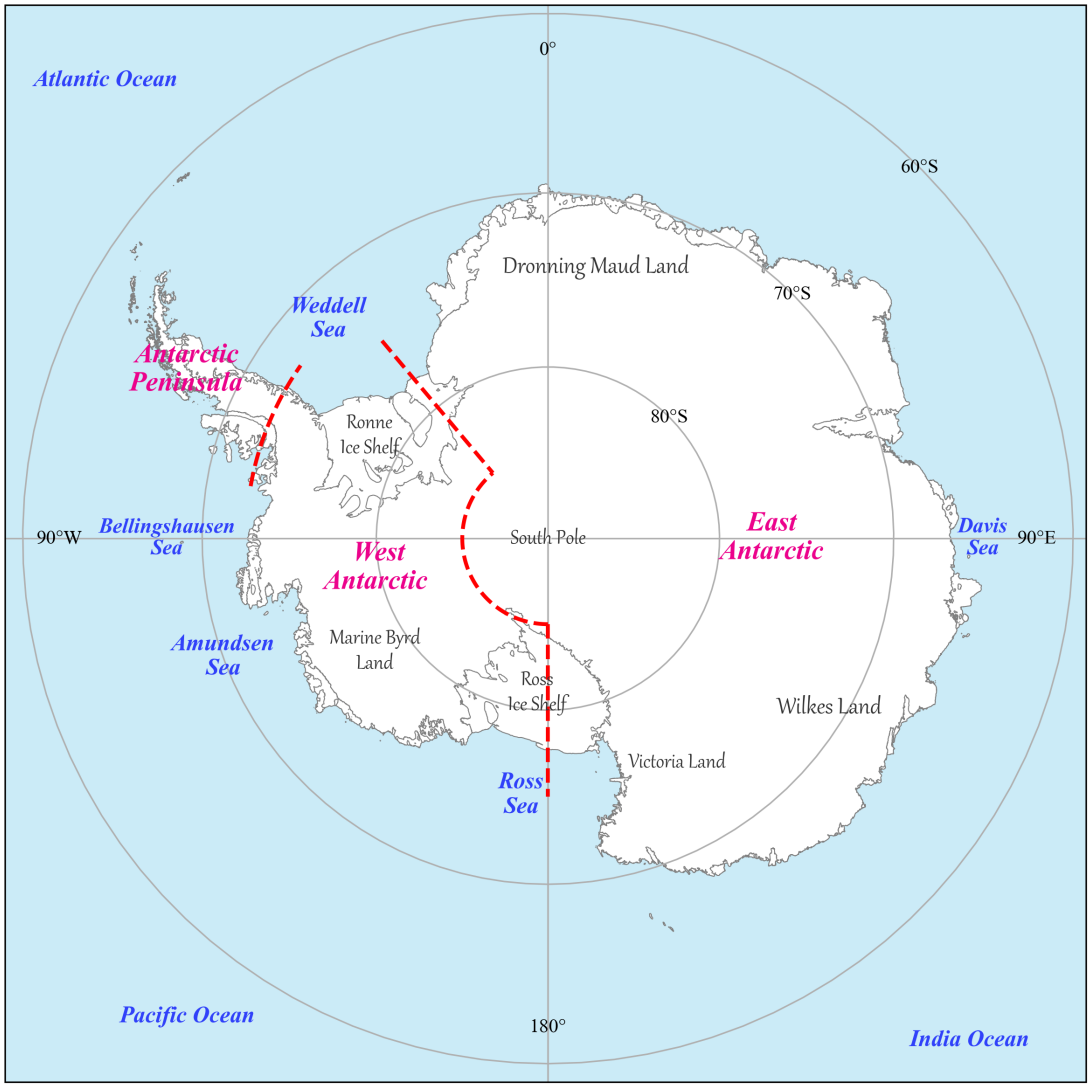

Figure S12. Geographic location map of the Antarctic continent (the red dashed lines indicate the boundaries dividing the Antarctic Peninsula, West Antarctica, and East Antarctica).

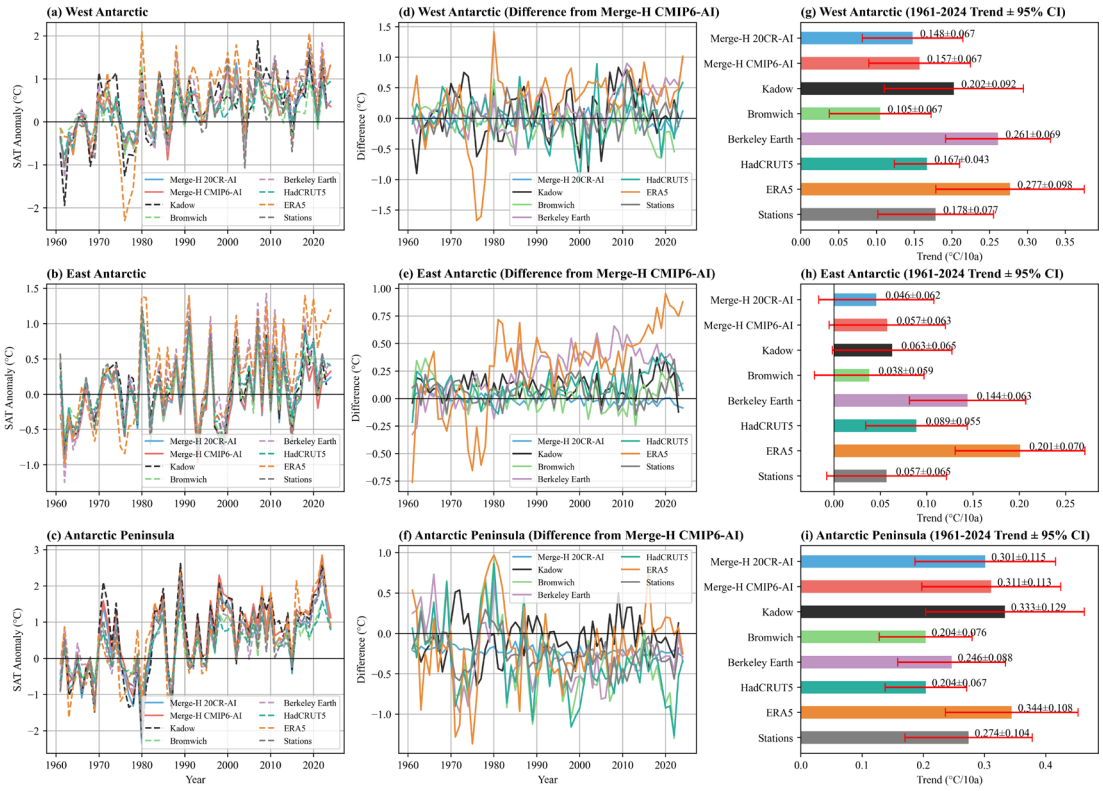

Figure S13. Area mean temperature time series and trends for Antarctic subregions.

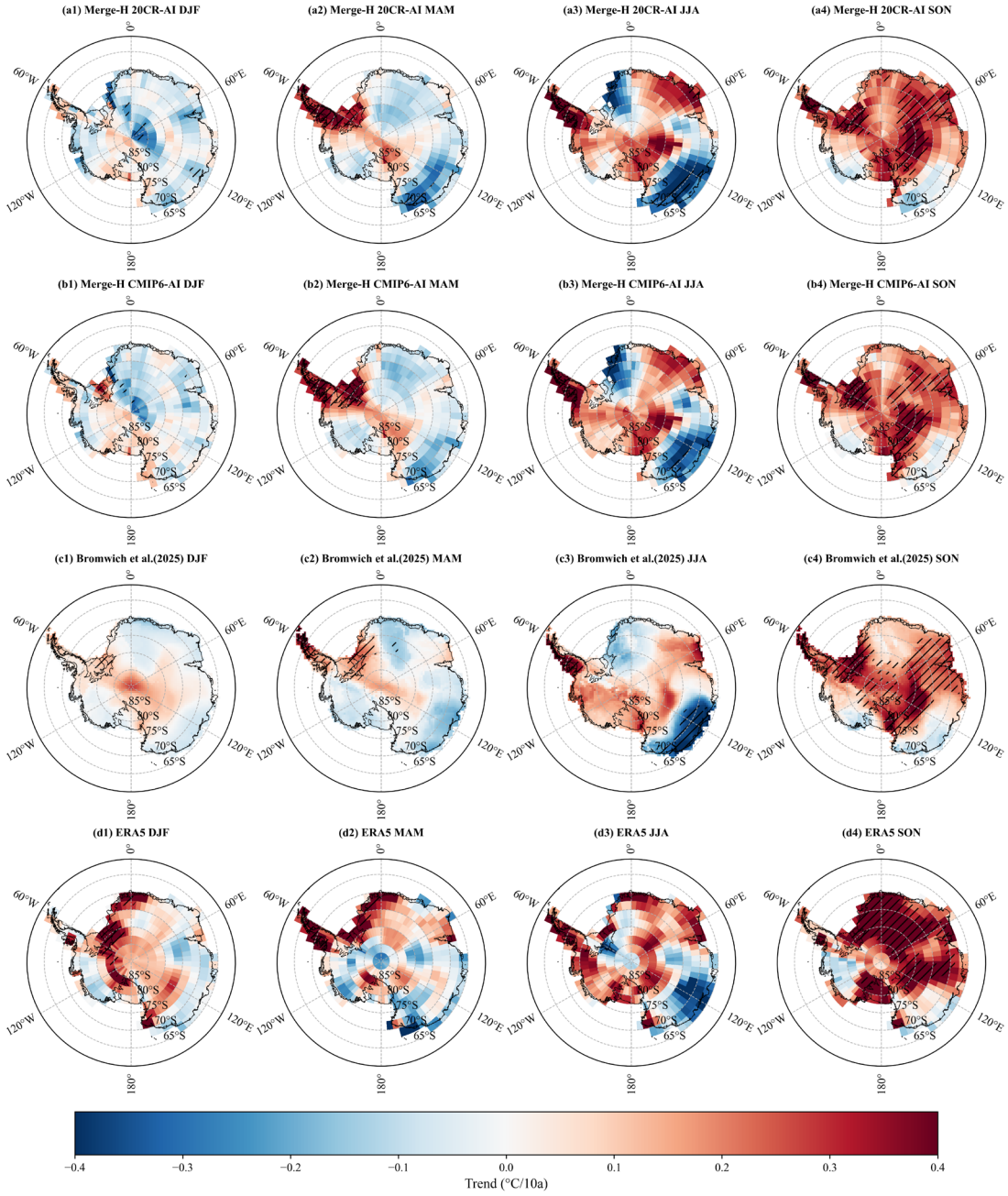

Figure S14. Spatial distribution of seasonal mean temperature trends over Antarctica during 1979–2024.