Response to Reviewer#3

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:

The manuscript presents an AI-based reconstruction of global surface temperature fields since 1850, with a particular focus on improving data coverage over Antarctica; while the application of advanced deep learning techniques to infill missing data is timely and relevant, several aspects require substantial clarification and improvement before the manuscript can be considered for publication.

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.

Major Issues

1. Unclear novelty relative to existing AI-based products

The novelty of the work needs to be more explicitly articulated. AI-based infilling approaches have already been applied in similar contexts; for example, ERSSTv6 (used in the Merge-E framework) already incorporates ANN-based spatial infilling. The authors should clearly explain how their approach differs from, and improves upon, existing methodologies.

Response: This study applies the PConv method to reconstruct climate fields based on previous work. In the revised manuscript, we briefly summarize the contributions of earlier studies as well as the incremental advances made in this work.

In addition, ERSSTv6 employs an artificial neural network (ANN) for spatial extrapolation over the ocean, and thus represents an ocean-only component. After being merged with the land component C-LSAT, it is no longer compatible with the PConv-based framework used in this study. A detailed explanation has been provided in Section 3.1 (Characteristics of the Global Reconstruction Results).

Changes to the Manuscript:

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.

2. Explanation of key concepts

Several important concepts are introduced without adequate explanation, assuming a level of prior knowledge that may not be shared by all readers. Examples include terms such as reanalysis (line 48) and polar amplification (line 57). These should be briefly but clearly defined when first introduced to ensure the manuscript is self-contained and accessible.

Response: We acknowledge that some technical terms were not sufficiently explained. To improve accessibility for a broader readership, we have added brief definitions when the terms "reanalysis" and "polar amplification" are first introduced in the manuscript.

Changes to the Manuscript:

reanalysis products (It refers to the use of data assimilation techniques to integrate multi-source observations with numerical models, producing meteorological fields that are temporally and spatially continuous and consistent)

Arctic amplification (In the context of global warming, temperature changes in the Arctic are significantly greater than the global average)

3. Insufficient validation and limited spatial comparison of results

At the end of Section 3.1 (lines 269–270), the authors state that the results demonstrate "the robustness and reliability of the AI framework for long-term climate reconstruction." However, this conclusion appears not fully supported by the presented analyses. The evaluation in 3.1 is primarily based on globally averaged quantities, which are insufficient to robustly assess spatial performance and regional consistency. A more comprehensive validation is required. In particular, the manuscript would benefit from the inclusion of spatially diagnostics, such as global maps of temperature trends (similar to the ones computed for Antartica) and their statistical significance, comparing the newly reconstructed products with existing datasets over different time periods. This would allow a clearer assessment of whether the proposed framework consistently reproduces known spatial patterns and regional features, rather than only matching average global statistics.

Response: We acknowledge that the previous validation framework was not sufficiently comprehensive. In the revised manuscript, we have conducted a more thorough evaluation, including regional and spatial analyses such as the ENSO spatial pattern and the Oceanic Niño Index (ONI), zonal temperature comparisons, and global warming pattern assessments.

Changes to the Manuscript:

3.1 ~~Characteristics of the Global Reconstruction Results~~ ENSO spatial patterns and the Ocean Niño Index (ONI)

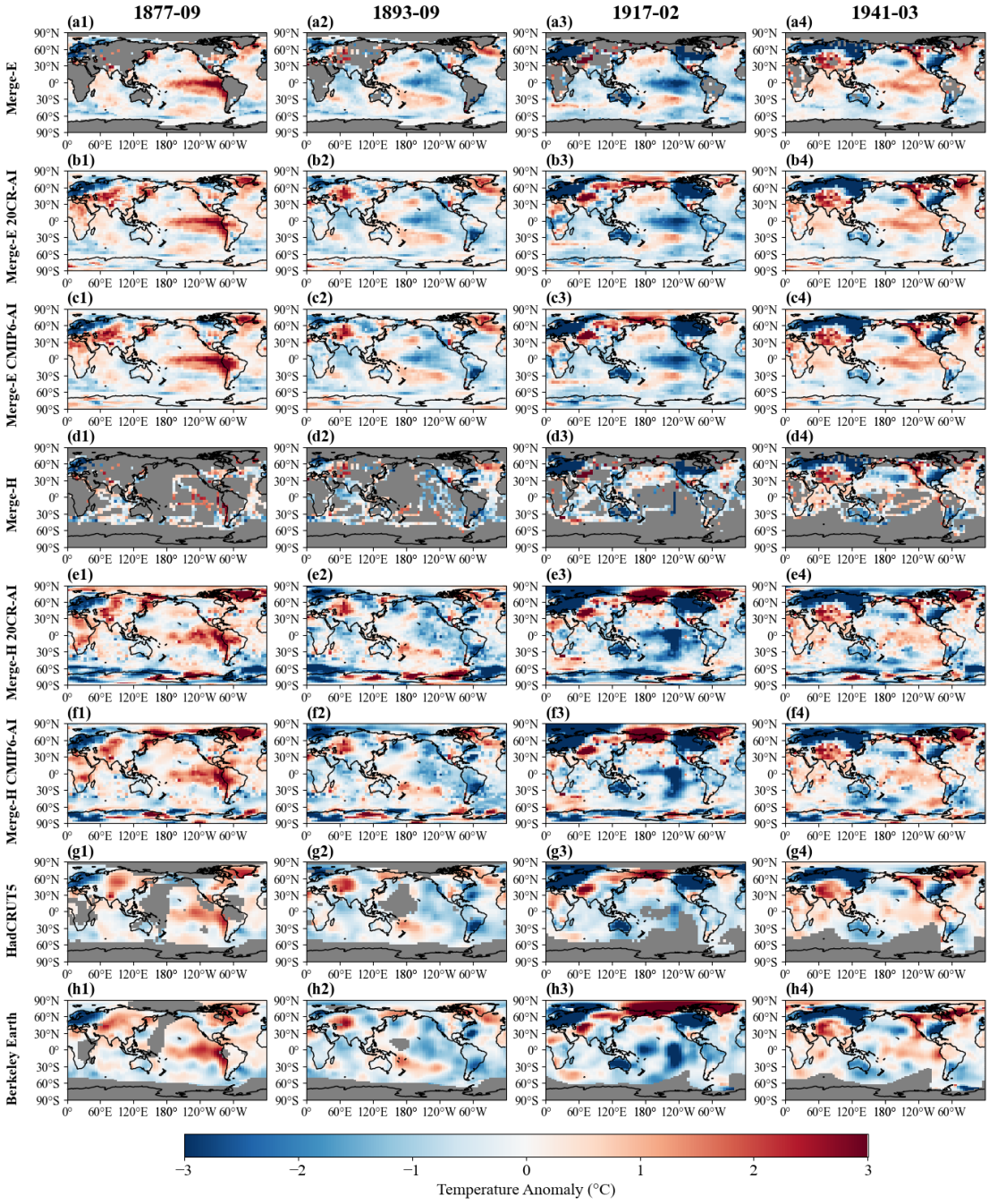

Figure 2: **Global temperature anomaly fields *(relative to the 1960–1990 climatology)* before and after reconstruction for four typical ~~months~~ *ENSO events*. *(a1–a4) Merge-E original; (b1–b4) Merge-E 20CR-AI; (c1–c4) Merge-E CMIP6-AI; (d1–d4) Merge-H original; (e1–e4) Merge-H 20CR-AI; (f1–f4) Merge-H 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.*~~

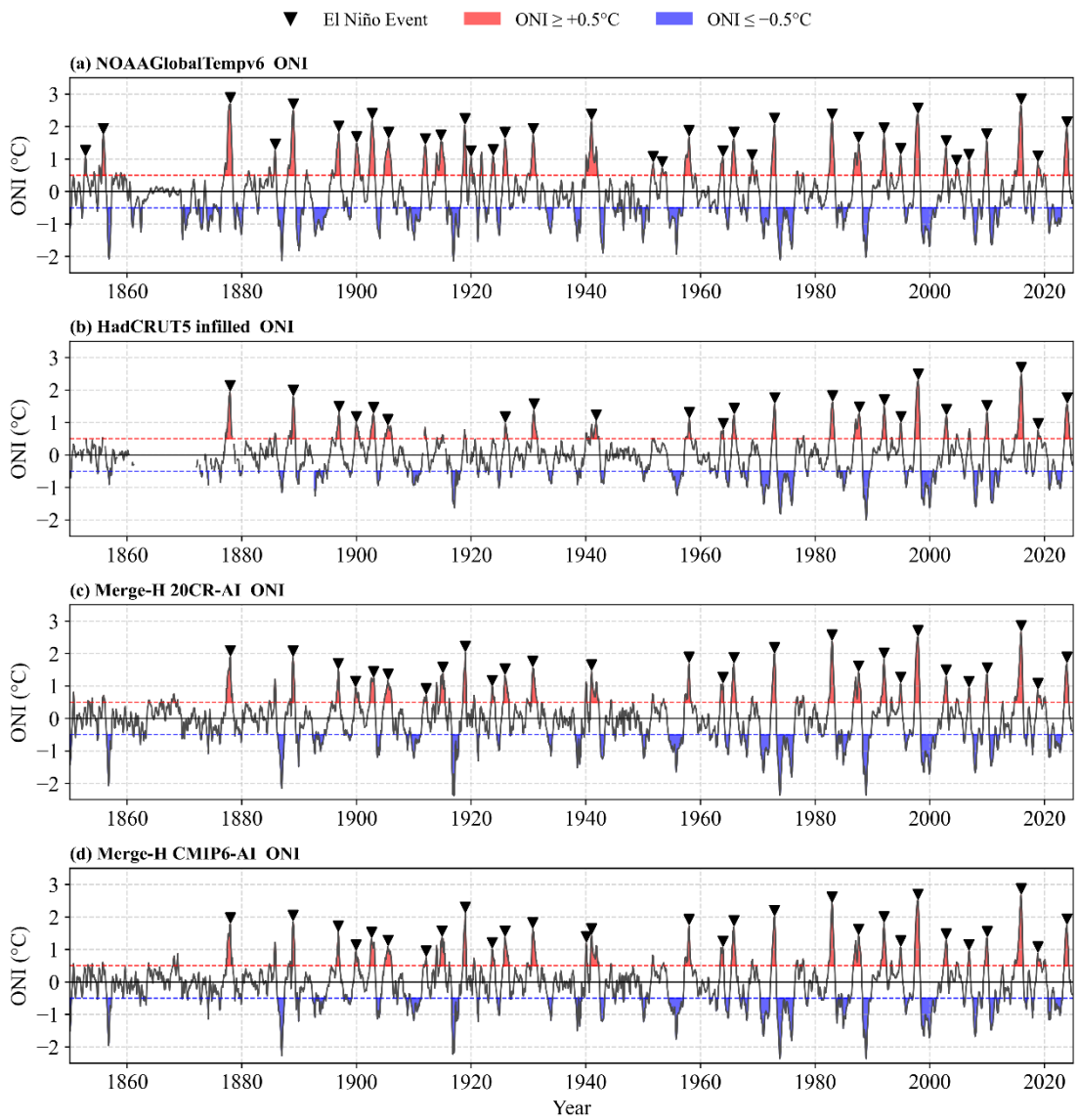

Figure 3: Ocean Niño Index (ONI) time series indicating ENSO events. (a) Time series of ONI from NOAAGlobalTempv6 over 1850–2024, with black triangles indicating El Niño events defined as ONI exceeding 0.5 °C for at least five consecutive months; (b–d) same as (a) but for HadCRUT5 infilled, Merge-H 20CR-AI,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

3.3 Comparison of zonal temperature

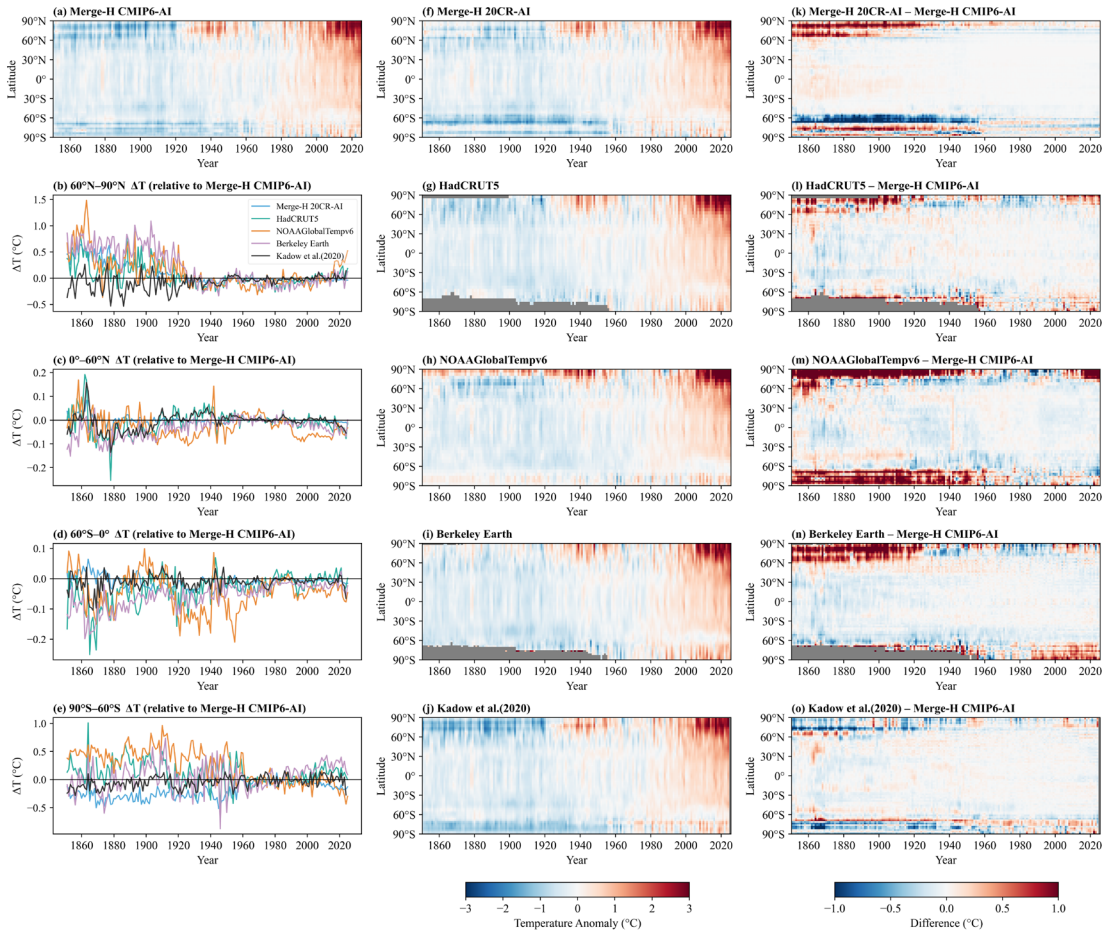

Figure 5 Zonal temperature comparison. (a) Zonal-mean temperature anomalies from Merge-H CMIP6-AI; (b–e) time series of zonal-mean temperature anomaly differences between other datasets and Merge-H CMIP6-AI across different latitude bands; (f–j) same as (a) for Merge-H 20CR-AI, HadCRUT5, NOAAGlobalTempv6,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

3.4 Global warming pattern

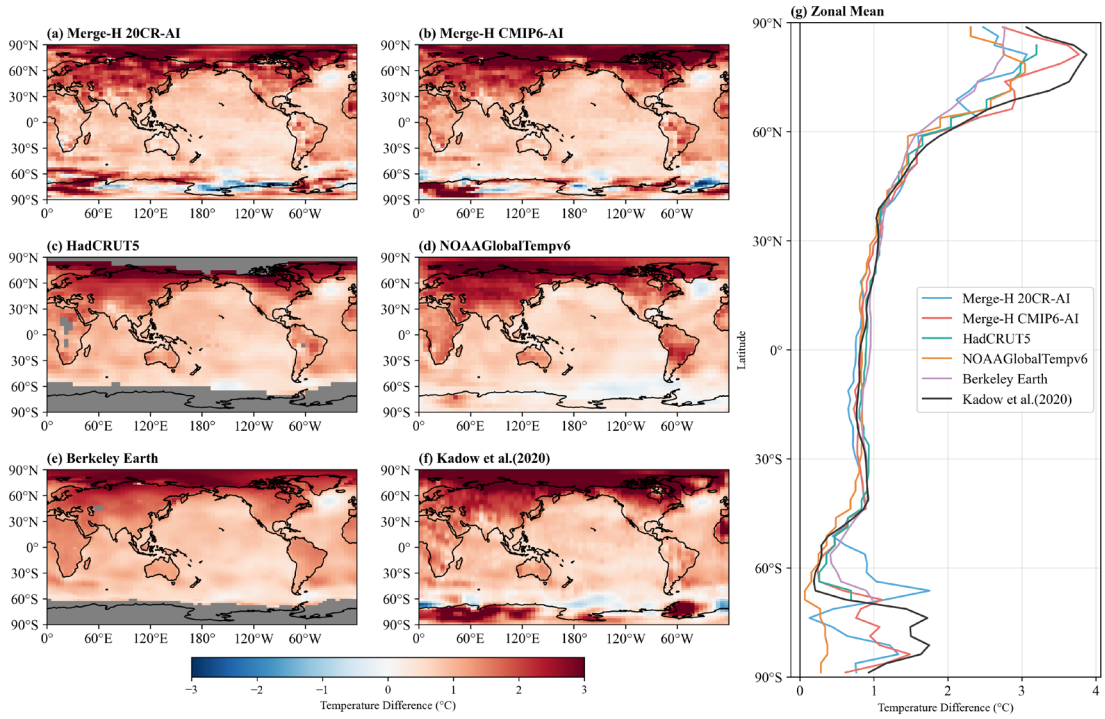

Figure 6 Global warming pattern. (a–f) Spatial distribution of mean global temperature warming over 2005–2020 relative to the 1850–1900 baseline period, (g) zonal-mean profile of warming magnitude.

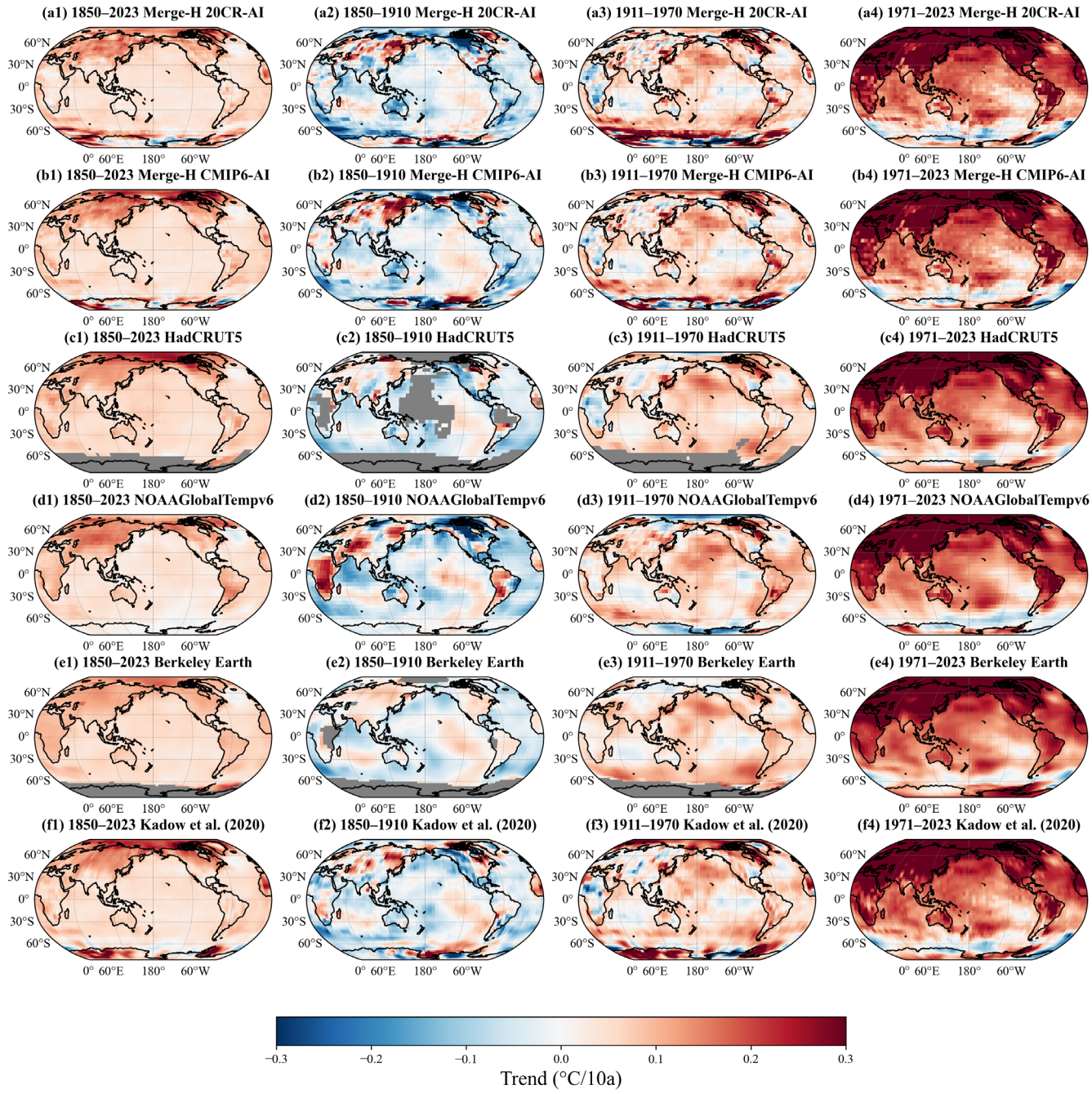

Figure 7 Spatial distribution of global surface temperature trends for 1850–2023, 1850–1910, 1911–1970, and 1971–2023. (a1–a4) Merge-H 20CR-AI, (b1–b4) Merge-H CMIP6-AI, (c1–c4) HadCRUT5, (d1–d4) NOAAGlobalTempv6, (e1–e4) Berkeley Earth, (f1–f4) reconstruction from Kadow et al. (2020).

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 (In the context of global warming, temperature changes in the Arctic are significantly greater than the global average) 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 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.

Figure 7 presents the spatial distribution of long-term warming trends from 1850 to 2023 across different datasets, as well as the warming trend patterns for three sub-periods: 1850–1910, 1911–1970, and 1971–2023. The results from Merge-H 20CR-AI and Merge-H CMIP6-AI are generally consistent with the reconstruction by Kadow et al. (2020). The spatial gradient of LSAT trends reconstructed by these three datasets is less uniform compared to other datasets, particularly across the continents during 1911–1970. Such differences are attributed to the nonlinear reconstruction capability of PConv. A notable difference between the AI-based reconstructions in this study and NOAAGlobalTempv6 lies in the 1850–1910 period. Specifically, NOAAGlobalTempv6 shows a widespread warming trend over the African continent, whereas the warming magnitude and spatial extent in other datasets. All datasets consistently indicate that, under the background of global warming, the magnitude of the warming trend during 1971–2023 is clearly higher than those in the earlier periods of 1850–1910 and 1911–1970.

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.

4. Lack of methodological clarity (trend estimation, significance, and data remapping)

Several key methodological steps are not described with sufficient detail, which limits reproducibility and makes it difficult to fully assess the robustness of the results. A first important point concerns the data remapping procedure. For instance, at line 126, line 151 and then again at lines 167-169 it is stated that observations have been regridded, but the exact method used for this remapping is not described. It should be clearly specified whether nearest-neighbour assignment, bilinear interpolation, area-weighted remapping, or another method is applied. Given the potential impact of this choice on the resulting fields, this step requires a precise methodological description. In addition, the procedures used for trend estimation and statistical significance testing are insufficiently documented. The manuscript should clearly specify:

◦ the method used to compute trends (e.g. ordinary least squares, robust regression, Theil-Sen non parametric approach, or other)

◦ how statistical significance is assessed

◦ and whether field significance or any correction for spatial autocorrelation and multiple testing has been considered (see Wilks et al, 2016)

Wilks, D. S., 2006: On "Field Significance" and the False Discovery Rate. J. Appl. Meteor. Climatol., 45, 1181–1189, https://doi.org/10.1175/JAM2404.1

Moreover, at line 353 the calculation of monthly observed anomalies at stations is not clearly described. It is unclear whether these anomalies are computed using the same reference period as the gridded datasets, or whether they are based on station-specific climatologies. This choice is crucial for consistency and comparability and should be explicitly stated in the methods section.

Finally, in Figure 6 caption it is stated that, for some products (marked with an asterisk), trends are computed over a shorter period ending in 2022. This introduces an inconsistency in the comparison, since trend estimates should be computed over identical time periods to ensure comparability. If this dataset is to be included in the analysis, then trends for all other products should be consistently truncated to 2022 (excluding 2023–2024), or otherwise a clear justification and sensitivity analysis should be provided.

Response: In the Methods section, we have added unified descriptions of the procedures used, including regridding, trend estimation, and significance testing. In addition, we acknowledge that multiple testing and spatial autocorrelation may lead to an overestimation of significance at individual grid points. However, as this study focuses on inter-dataset comparisons, we consistently applied the relatively simple ordinary least squares (OLS) regression. More sophisticated approaches, such as the AR(1) model commonly used in IPCC Assessment Reports to better quantify the significance and uncertainty of warming trends, were not adopted.

The calculation method referred to in line 353 has been added.

We acknowledge that, for the datasets marked with an asterisk in Figure 6, the data are only available up to 2022. Extending the trend calculation to 2024 would introduce inconsistencies; therefore, the trend period has been uniformly truncated to end in 2022.

Changes to the Manuscript:

(1) *Resolution standardization:* ~~*All datasets used for training and reconstruction were regridded to a common spatial resolution of 5°×2.5°, resulting in a 72×72 grid.*~~ *The input datasets used for training are uniformly remapped from higher resolution to a spatial resolution of 5° × 2.5° using conservative interpolation, resulting in a regular grid composed of 72 × 72 grid cells.*

The method involves extracting the gridded data from the reconstructed product over the same time period as the station observations, and then performing a consistency check after removing the mean from each respective time series.

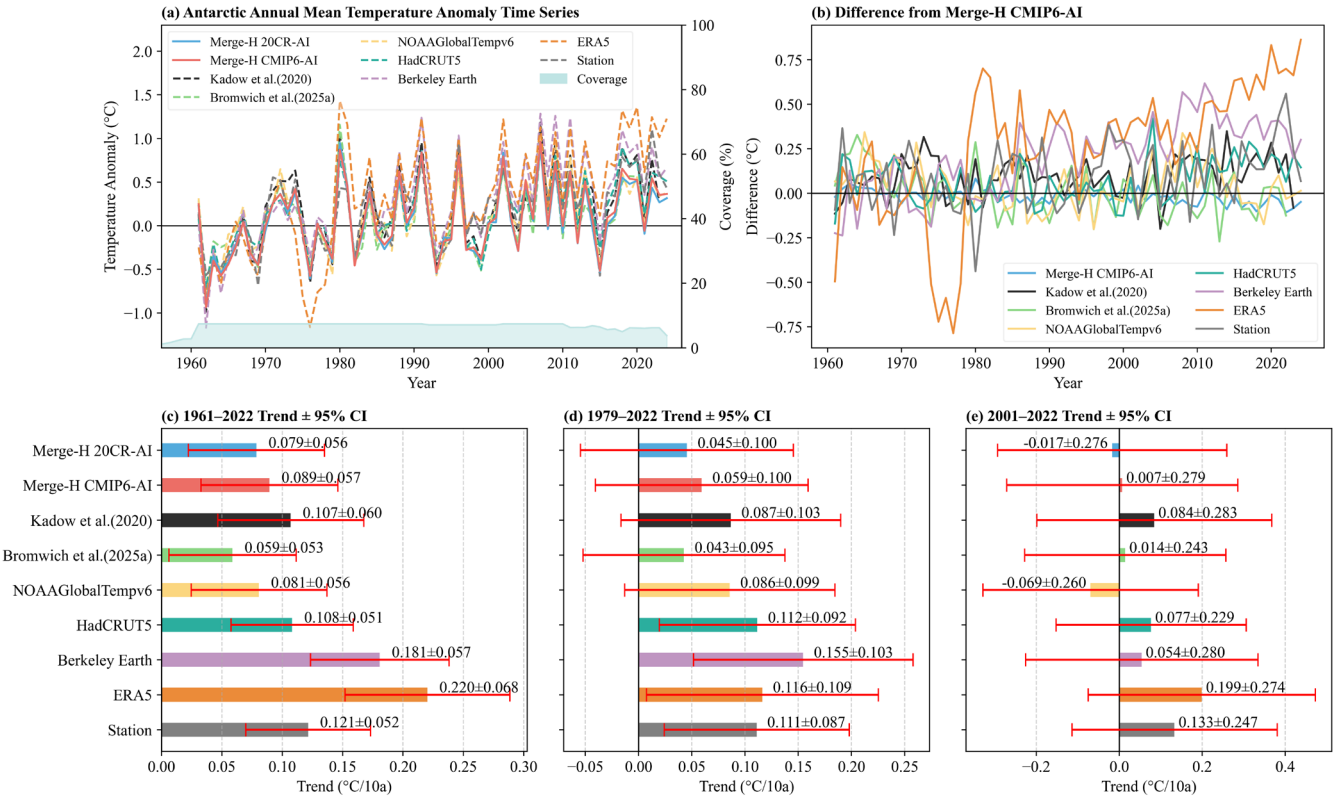

***Figure 9*: (a) Annual mean surface temperature anomaly time series over Antarctica from 1961 to 2024, *and coverage refers to the proportion of Antarctic grid cells that are occupied by observational grid points before reconstruction*; *(b) Temperature difference from Merge-H CMIP6-AI;* (c–e) Linear trend of annual mean temperature and 95% confidence interval (°C per decade) during 1961–2022, 1979–2022 and 2001–2022.**

2.4 ~~Model Post-Processing~~ Data validation methods

~~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.

…

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.

5. Over-reliance on supplementary material

Important methodological details, figures and dataset descriptions are frequently shifted to the supplementary material, despite being essential for understanding the study. This reduces the accessibility and self-consistency of the manuscript. For example, a comprehensive table listing all precipitation products used in the study, including their type, temporal coverage, and spatial extent, would be essential in the manuscript itself rather than being confined to the supplementary material. More generally, the main text should better synthesize and integrate such information instead of too often referring the reader to supplementary material.

Response: We have added a table following the Data Resources section summarizing the temperature datasets used in this study, including their type, temporal coverage, resolution, and corresponding references.

Changes to the Manuscript:

Table 1 List of information on the various data used in this paper.

Data	Type	Resolution	Time	Reference
20CR	Grid	Monthly/0.7°×0.7°	1850–2015	Slivinski et al. (2019)
CMIP6	Grid	Monthly/-	1850–2014	Eyring et al. (2016)
C-LSAT2.1	Grid	Monthly/5°×5°	1850–2024	Wei et al. (2025)
ERSSTv6	Grid	Monthly/5°×5°	1850–2024	Huang et al.(2025a)
HadSST4	Grid	Monthly/5°×5°	1850–2024	Kennedy et al. (2019)
SCAR READER	Station	Monthly	-	Turner et al. (2004)
GHCNv4 QCF and QFE	Station	Monthly	-	Menne et al. (2018)
OSU Polar Meteorology Group	Station	Monthly	1958–2022	Bromwich et al. (2025b)
Météo-France	Station	Monthly	2014–2016	Météo-France (2025)
University of Wisconsin-Madison	Station	Monthly	1958–2022	South Pole Meteorology Office (2025)
NIWA	Station	Monthly	2016–2022	NIWA (2025)

6. Use of informal language and incorrect syntaxisThe manuscript contains several informal expressions that should be revised to meet the standards of a journal such as Earth System Science Data. A thorough language editing is required to improve clarity, precision, and overall readability.

Examples: ◦ Line 30 and 38: commas are used where a full stop or a proper conjunction would be more appropriate; the sentence structure should be revised for clarity.

◦ Line 37: a space is missing after the end of the sentence.

◦ Line 53: avoid the use of the Saxon genitive; a more formal scientific construction (e.g., “of the …”) is preferable.

◦ Line 63: Expressions such as “fields often hard to capture the overall” are informal and elliptic; these should be reformulated in a clearer and more explicit way (e.g., “fields often struggle to capture the overall …”).

◦ Line 83: The expression “with as complete a spatial coverage as possible”, while grammatically correct, is stylistically awkward; a clearer formulation such as “with the most complete spatial coverage possible” or “with spatial coverage as complete as possible” is recommended.

◦ Lines 330, 332, and 335: inconsistent terminology is used for trends (e.g. “fastest”, “slowest”, “lower”). These should be replaced with more scientific and consistent expressions referring to trend magnitude (e.g. “higher/lower trend magnitude” instead of “fastest warming”).

Response: We thank the reviewer for the careful reading and for pointing out the informal expressions in the manuscript. We have revised and refined the relevant sections accordingly. In addition, after incorporating comparative analyses of the ENSO spatial pattern and the Oceanic Niño Index (ONI), zonal temperature comparisons, and global warming patterns, we found that the Section 3.3 (Characteristics of seasonal surface temperature changes) did not provide additional insights. Therefore, this section has been removed, and the corresponding content in the original manuscript (Lines 330, 332, and 335) has been deleted.

Changes to the Manuscript:

Line 30: Global surface temperature (ST) is one of the most fundamental variables in the climate system, directly reflecting the state of the ~~Earth's~~ global energy balance, and it plays a central role in the monitoring and assessment of climate change.

Line 38: The evolution of the atmosphere and ocean follows the fundamental physical laws of mass, momentum, and energy conservation; therefore, these constraints imply that the climate field exhibits a certain degree of spatial and temporal continuity and predictability.

Line 37: ...become particularly necessary (Vose et al., 2021; Morice et al., 2021). The evolution of the atmosphere and ocean follows the fundamental...

Line 53: ~~The~~ Antarctica holds an irreplaceable position in the global climate system, and its enormous glacier masses store a substantial portion of ~~the world's~~ global freshwater and play a major role in determining future sea level change.

Line 63: As a result, observation-based regional temperature fields often ~~hard~~ struggle to capture the overall spatial structure of ST variability.

Line 83: ...with ~~as complete a spatial coverage as possible~~ spatial coverage as complete as possible.

Minor issues:

• In general, the Introduction would benefit from a short concluding paragraph outlining the structure of the paper. Providing a brief roadmap of the subsequent sections would improve readability and help the reader navigate the manuscript more effectively.

Response: We have added an overview of the manuscript structure at the end of the Introduction to improve readability.

Changes to the Manuscript:

The remainder of this paper is organized as follows. Section 2 describes the data and methods, including data resources, the land–sea merging method, the AI training and reconstruction process, and the data validation methods. Section 3 presents the global reconstruction results, including the spatial pattern of ENSO and the Oceanic Niño Index (ONI), global mean surface temperature (GMST), zonal temperature comparisons, global warming patterns, and regional land surface air temperature comparisons. Section 4 presents the Antarctic temperature reconstruction results. Section 5 discusses the limitations and future perspectives. Sections 6 and 7 provide information on data and code availability, respectively. Section 8 concludes the paper.

• Line 68: the phrase "with different datasets as weights" is unclear and should be reformulated. It is not evident how datasets are used as weights.

Response: We have revised the summary of this section in the literature review.

Changes to the Manuscript: ~~*Nicolas et al. (2014) and Bromwich et al. (2025a) applied ordinary kriging with different datasets as weights to spatially extrapolate Antarctic ST, effectively mitigating the problem of sparse observational coverage.*~~ *Nicolas et al. (2014) and Bromwich et al. (2025a) reconstructed Antarctic surface temperature by using reanalysis data to characterize the spatial covariance structure, and then applying an ordinary kriging framework to derive optimal interpolation weights, thereby enabling spatial extrapolation under sparse observational coverage.*

• Line 96: the concept of temperature anomalies is introduced here for the first time, but the reference period used to compute these anomalies is not specified until later in the manuscript (line 154). It would improve clarity to define anomalies explicitly at first mention, including a clear statement of what they represent and the reference period over which they are calculated. Also specify the climatological reference period when using the term anomalies in figures (e.g. Figure 2).

Response: We have explicitly specified the climatological reference period when temperature anomalies first appear, and we have also indicated the reference period in the caption of Figure 2.

Changes to the Manuscript:

Building on these previous efforts, the present study applies PConv reconstruction framework for global ST anomaly (relative to the 1960–1990 climatology) fields.

Figure 2: Global temperature anomaly (relative to the 1960–1990 climatology) 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.

• Lines 113 and 115: the term Historical is used inconsistently, once with capitalisation and quotation marks and once in lowercase without quotes. It is unclear whether these refer to the same concept or to distinct definitions. The authors should clarify this point explicitly. If the same concept is intended, a consistent notation should be used throughout; if different concepts are meant, the distinction should be clearly explained.

Response: To avoid potential misunderstanding of the term "historical," we removed its first occurrence and changed the second "Historical" to historical. In this context, "historical simulations" refers to a specific experiment within the CMIP6 model framework.

Changes to the Manuscript:

In addition, two ~~historical~~ monthly ST anomaly (relative to the 1961–1990 climatology) datasets are employed to construct training sets for the AI reconstructions. The Twentieth Century Reanalysis version 3 (20CR; Slivinski et al., 2019), which provides monthly ST fields for 1850–2015, is used to train the "20CR-AI" model, whereas the ~~"Historical"~~ historical simulations from the Coupled Model Intercomparison Project Phase 6 (CMIP6), offering monthly ST fields for 1850–2014, are used to train the "CMIP6-AI" model. The 20CR dataset contains 80 ensemble members, whereas the CMIP6 dataset contains 105 ensemble members (Table S1)

• Figure 1: there are typographical errors within the figure, where the label “Trian” appears instead of “Train” in multiple instances. These should be corrected to ensure consistency and avoid confusion.

Response: “Trian” is likely a typographical error for “Train”, and we have corrected this in Figure 1.

Changes to the Manuscript:

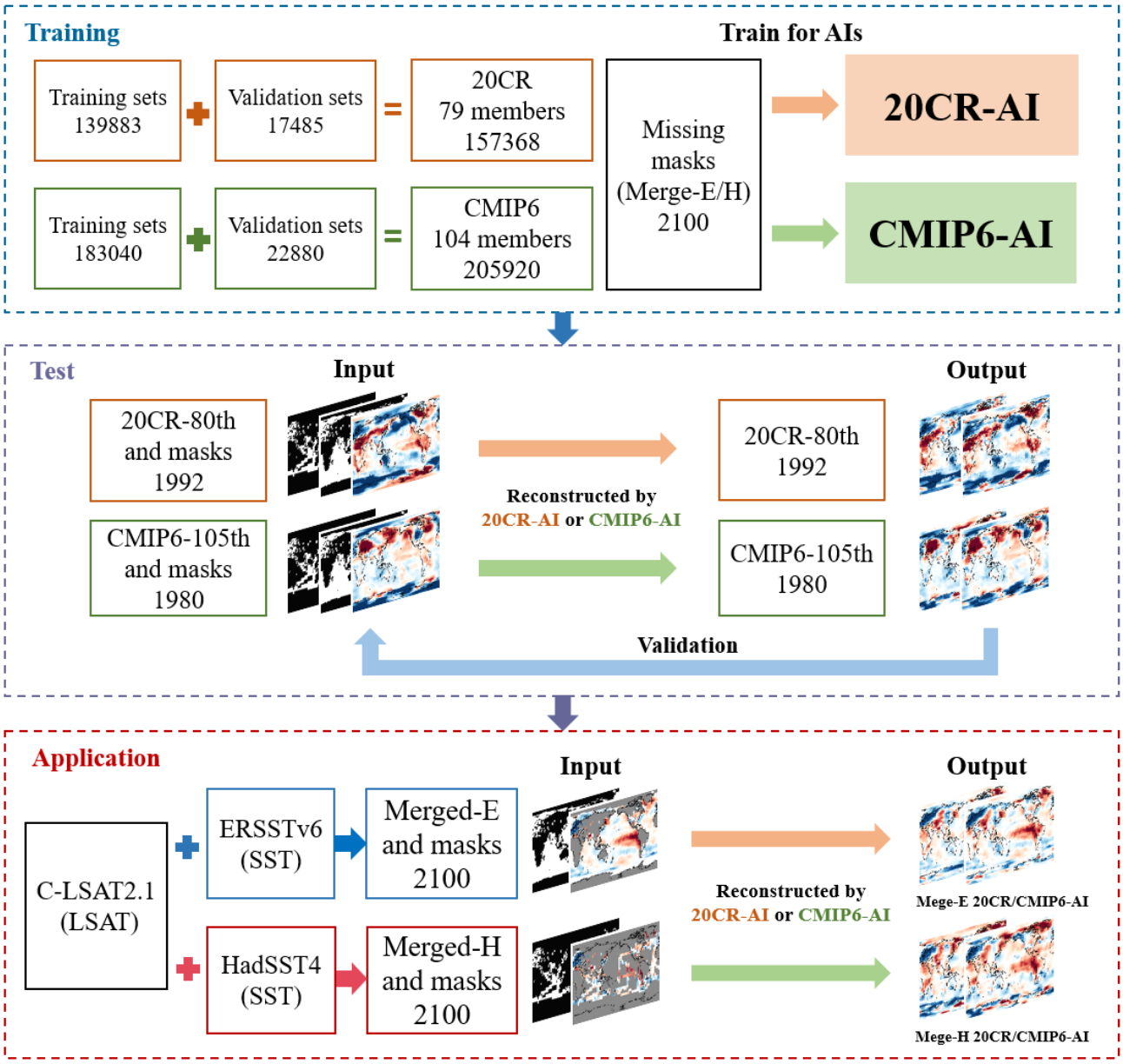

***Figure 1*: AI training and reconstruction *(The numbers shown in the data boxes represent the sample size of the monthly anomaly fields, corresponding to the total number of months)*.**

• Line 154: the reference to (2014) appears in parentheses without sufficient context, making its meaning unclear. It is not evident whether this refers to a citation, a dataset version, or a specific time period. The authors should clarify its meaning and ensure consistent and explicit referencing throughout the manuscript.

Response: We acknowledge that the “(2014)” in this section was unclear, and we have revised it accordingly.

Changes to the Manuscript:

Sample partitioning: Monthly temperature anomaly fields from 20CR (80 ensemble members) for 1850–2015 and CMIP6 (105 ensemble members) for 1850–2014 ~~(2014)~~ relative to the 1961–1990 climatology were divided into training and testing sets.

• Section 2.4 (opening sentences): the beginning of this section appears to repeat content already described previous Section of the methodology (lines 151–152). It is unclear whether this refers to a distinct procedure or the same one restated. If it is the same procedure, the text should be revised to avoid redundancy and improve coherence between sections; if instead two different procedures are intended, the distinction should be made explicit and clearly explained.

Response: We have revised the methodological descriptions in these two sections.

Changes to the Manuscript:

Resolution standardization: ~~All datasets used for training and reconstruction were regridded to a common spatial resolution of 5°×2.5°, resulting in a 72×72 grid.~~ The input datasets used for training are uniformly remapped from higher resolution to a spatial resolution of 5° × 2.5° using conservative interpolation, resulting in a regular grid composed of 72 × 72 grid cells.

~~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.

• Line 253: the statement referring to "two AI models" is ambiguous. It is not clear whether this refers to two distinct neural network architectures or to the same model architecture trained on different datasets. This point should be clarified explicitly, as it is essential for understanding the experimental design and the interpretation of the results.

Response: We have replaced the term "two AI models" with "Merge-H 20CR-AI and Merge-H CMIP6-AI."

Changes to the Manuscript:

~~It is noteworthy that the two AI models show small differences in their long-term reconstructions and trends under the Merge-H scenario~~ The Merge-H 20CR-AI and Merge-H CMIP6-AI show only minor differences in long-term reconstructed series and trends, and are generally consistent with the HadCRUT5 record

• Figure 4 and the subsequent paragraph: it is not sufficiently clear how the comparison between the original Merge-H dataset and the AI-reconstructed version is performed. In particular, it is unclear whether the averaged values are computed over the same spatial domain, or whether global means from the original dataset (which contains missing data) are being compared directly with fully reconstructed fields. The latter would not represent a fair comparison due to the differing spatial coverage. If, as would be appropriate, a common mask is applied and only grid points present in the original dataset are used for both products, this procedure should be explicitly stated and described in detail. In general, the methodology for ensuring a consistent

and fair comparison between incomplete and fully reconstructed fields needs to be clarified more thoroughly to avoid potential biases in the interpretation of the results.

Response: Regarding Figure 4 and the related analyses in the original manuscript, we agree that comparing incomplete observational data with fully reconstructed fields may introduce biases due to differences in spatial sampling. However, in meteorological and climate research, when observational datasets have incomplete spatial coverage, it is a widely adopted practice to compute regional mean time series based on the available grid cells. The main purpose of this approach is to make full use of existing observational information and characterize the mean variability of the represented region.

In this study, all datasets were processed using the same spatial domain and area-weighted averaging, and a consistent land mask was applied to ensure comparability. The primary objective is to evaluate the overall regional mean climate signal after full-field reconstruction. Therefore, using the complete spatial domain for averaging is more consistent with the goals of this study.

• Lines 300–301: two trend values are reported, but it is not clearly indicated which one corresponds to the Merge-H 20CR-AI product and which one refers to the Merge-H CMIP6 AI product. The attribution should be explicitly clarified in the text to avoid ambiguity and ensure correct interpretation of the results.

Response: The wording in this section was unclear, and we have revised it for clarity.

Changes to the Manuscript:

~~*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.*~~

As shown in Table 3, the Merge-H 20CR-AI and Merge-H CMIP6-AI 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.

• Lines 314–318: the current sentence appears to describe features and variability that are common across all seasons. It may therefore be more appropriate to move these general considerations to an earlier section where annual-scale results are discussed. Conversely, the seasonal-specific differences would be better addressed in the context of Figure 5, where the seasonal decomposition is explicitly presented.

Response: As noted in the previous response, the Section 3.3 (Characteristics of seasonal surface temperature changes) has been completely removed; therefore, this issue no longer exists.

• Line 320: the statement referring to “larger interannual fluctuations” would benefit from quantitative support. It would be helpful to either explicitly quantify these fluctuations or to clearly refer to an existing figure or table where their magnitude is reported, in order to substantiate the claim that they are the largest among the considered cases.

Response: As noted in the previous response, the Section 3.3 (Characteristics of seasonal surface temperature changes) has been completely removed; therefore, this issue no longer exists.

• Line 320: the sentence “this contrast is mainly attributed to the dominance of the NH” appears incomplete or at least unclear in its current form. It is not specified what the dominance of the Northern Hemisphere refers to in this context (e.g. global mean signal, trend structure, variability). The sentence should be completed or reformulated to explicitly state the variable or process being influenced.

Response: As noted in the previous response, the Section 3.3 (Characteristics of seasonal surface temperature changes) has been completely removed; therefore, this issue no longer exists.

• Line 321: the text refers to “land temperatures”, but it is not clear whether a formal separation between land and sea surface temperatures has been performed throughout the analysis. If such a distinction has been made, this is not clearly described in the methods section nor consistently reflected in the presentation of the results. Moreover, it is not evident whether dedicated maps or diagnostics are available to assess this separation and its implications. As already noted, the inclusion of spatially explicit analyses (e.g. trend maps or interannual variability fields) would be essential to properly support and quantify these statements, which in their current form appear insufficiently substantiated.

Response: As noted in the previous response, the Section 3.3 (Characteristics of seasonal surface temperature changes) has been completely removed; therefore, this issue no longer exists.

• Line 346: the statement that the coverage is “10%” is not clearly defined. It is unclear what the 100% reference corresponds to in this context (e.g. one station per grid cell, full spatial sampling, or complete dataset coverage). The method used to quantify this percentage should be explicitly clarified, as well as whether it refers to station coverage or dataset coverage. Without this clarification, the interpretation of this metric remains ambiguous.

Response: We acknowledge that the meaning of the 10% coverage was not clearly defined. Here, “coverage” refers to the proportion of observed grid cells over the Antarctic continent prior to reconstruction. We have added an explanatory note in the figure caption.

Changes to the Manuscript:

Figure 8: (a) Annual mean surface temperature anomaly time series over Antarctica from 1961 to 2024, and coverage refers to the proportion of Antarctic grid cells that are occupied by observational grid points before reconstruction; (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.

• Line 353: the term “Effective Grid Cell” is introduced without a clear definition. It is not explained whether this refers to a different concept from the previously used “grid cells,” nor what criteria make a grid cell “effective.” This should be explicitly defined in the methods section, clarifying its meaning and how it differs (if at all) from the standard grid cell definition used throughout the manuscript.

Response: We acknowledge that the term “Effective Grid Cell” may be misleading. Here, we intended to refer to grid cells located outside the observed grid points prior to reconstruction. We have revised the explanation accordingly.

Changes to the Manuscript:

we selected 14 independent Antarctic stations that were not used in the reconstruction and were located outside the ~~effective grid cells~~ spatial coverage of the original observational grid before reconstruction.

• Conclusions (line 286): the reference to "spatial patterns" is too broad, as the spatial evaluation appears to have been conducted only for Antarctica. The statement should therefore be qualified to avoid overgeneralisation, or the analysis should be extended to other regions if a global assessment of spatial patterns is intended.

Response: In the response above, we noted that we have conducted additional spatial analyses of the results. In the response to the next question, we provide the revised version of this section.

• Final paragraph of Conclusions: the content appears somewhat repetitive with respect to earlier parts of the Conclusions section. A restructuring would improve clarity and impact, by removing redundancy and making the final message more concise and effective. In particular, key findings and implications should be stated once, in a more streamlined form, avoiding restatement of results already discussed

Response: We have condensed this section. Finally, we would like to once again thank the reviewer for the detailed comments and suggestions, which have greatly improved the linguistic coherence, lexical precision, writing professionalism, and overall readability of the manuscript.

Changes to the Manuscript:

This study employed an AI model (PConv) to reconstruct global ST fields from two fused observational datasets. The results demonstrate that the AI reconstructions ~~shows high consistency~~ are broadly consistent with multiple representative climate datasets. ~~in terms of interannual variability, long-term temperature trends, and spatial patterns, validating the effectiveness and reliability of deep learning-based image inpainting approaches for climate reconstruction tasks.~~ Both Merge-H 20CR-AI and Merge-H CMIP6-AI datasets exhibit high temporal and spatial continuity in low- and mid-latitude regions and over Antarctica after 1961, providing a solid foundation for extending long-term climate records, assessing polar climate change, and supporting climate monitoring, detection, and attribution. ~~A comprehensive comparison of the global mean ST time series derived from different reconstruction schemes in this study indicates that the datasets exhibit largely consistent long-term trends. In particular, for Antarctica after 1961, the AI reconstruction aligns well with both observational and reanalysis data, indicating that the AI-based approach provides reliable support for reconstructing continuous global ST fields since the mid-19th century.~~

…

Based on the Merge-H AI reconstruction schemes, this study developed spatially complete global monthly ST anomaly datasets for 1850–2024 with a spatial resolution of 5°×2.5°, termed the China global Artificial Intelligence Reconstructed Surface Temperature$_{20CR/CMIP6}$ (C-AIRST$_{R/M}$) datasets, which are reconstructed independently using the 20CR-AI and CMIP6-AI schemes based on the merged C-LSAT2.1 and HadSST4. ~~Both datasets exhibit high temporal and spatial continuity, providing a solid foundation for extending long-term climate records, assessing polar climate change, and supporting climate monitoring, detection, and attribution.~~ Overall, this study demonstrates the potential and application value of AI in climate data reconstruction. With further advancements in deep learning, physics-informed learning, and high-performance computing, future AI-based climate reconstruction frameworks are expected to achieve breakthroughs in global continuity and high resolution, offering more robust scientific support for understanding the evolution of the Earth's climate system.

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.