

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:

I thank the authors for their detailed response and substantial revisions. The manuscript has improved considerably, especially through the expanded validation and the better justification of the produced product. However, after reading the revised manuscript carefully, I still have several important concerns. Some arise from the new analyses added in revision, while others concern the interpretation and practical value of the proposed dataset. In my view, these issues should be addressed before the manuscript can be considered suitable for publication in ESSD.

Response: We sincerely thank the reviewer for the careful review of our manuscript and for providing detailed and constructive comments. These valuable suggestions have been instrumental in further improving the quality of our manuscript. In accordance with the reviewer's suggestions, we have conducted a comprehensive and careful revision of the manuscript and have addressed each comment with corresponding responses and modifications. The detailed responses and the specific revisions made to the manuscript are presented point-by-point below. In the revised manuscript, the text highlighted in blue indicates the revised content, the red strikethrough text indicates the deleted content, and the black text represents the original content.

Comments from reviewer:

1. I appreciate that the authors have now added a clearer discussion of Kadow et al. (2020), which was missing or insufficiently acknowledged in the previous version despite the close methodological similarity. However, the revised manuscript still raises a broader concern in how closely related prior work is acknowledged and used to frame novelty.

In particular, the newly added ENSO validation, including the spatial 'warm-cold tongue' comparison in Fig. 2 and the ONI/Nino-index time-series comparison in Fig. 3, bears a high similarity in purpose and structure to the validation framework used in Fig.S12 in Qian et al. (2026, arXiv:2602.16515), which should be acknowledged and discussed.

Moreover, the authors discuss several AI-based reconstruction methods in the Introduction and in the final discussion on future perspectives, but the manuscript largely omits the most recent studies. Recent works based on generative models such as Qian et al. (2026) and Li et al. (2026, arXiv:2603.16272) are now promising to AI-based climate reconstruction and should be included in the discussion to provide a more accurate and up-to-date account of the field.

Response: We acknowledge that the omission of these two recent studies in the original manuscript was a shortcoming of our work. In the revised manuscript, we have incorporated discussions of these studies and added the corresponding citations in Section 1 (Introduction) and Section 5 (Limitations and Future Perspectives).

Changes to the manuscript:

Line 108–110 in track change:

Qian et al. (2026) improved the accuracy of global historical climate reconstruction by integrating a Video Diffusion Model with a U-Net framework. Li et al. (2026a) applied a generative diffusion model to perform probabilistic reconstruction of SST fields, providing a new approach for quantifying uncertainty in historical climate data.

Line 724–731 in track change:

Furthermore, with the rapid development of AI, generative diffusion models have demonstrated the potential to further improve the accuracy of historical global climate field reconstruction by employing a Video U-Net architecture to jointly model the temporal evolution of multiple spatial fields (Qian et al., 2026). By learning the probability distribution of SST fields under observational constraints and generating multiple physically plausible reconstructions, these models provide an alternative framework for probabilistic temperature field reconstruction and uncertainty quantification (Li et al., 2026a). Therefore, future developments in historical climate field reconstruction should not only account for the spherical geometry of the Earth but also preserve the spatiotemporal physical consistency of climate fields, thereby enabling more accurate and reliable historical temperature reconstructions together with probabilistic uncertainty

quantification.

Comments from reviewer:

2. The observational product reconstructed here combines land surface air temperature with sea surface temperature over the ocean. However, the AI models appear to be trained on near-surface air temperature fields from 20CR and CMIP6. This raises a basic but important question: does training on tas and reconstructing a blended SAT/SST observational product introduce systematic inconsistencies, especially near coastlines and sea-ice margins? The authors should explicitly discuss whether the SAT/SST mismatch affects the learned spatial structures and the reconstructed fields.

Relatedly, the training data are model-based or reanalysis-based products. CMIP6 and 20CR contain known regional biases, particularly in polar regions. The manuscript should test and discuss more clearly how such biases may propagate into the final reconstruction. This is important for a data product intended for broad use.

Response: We thank the reviewer for raising the important issue regarding the potential propagation of biases from the 20CR and CMIP6 training datasets into the reconstruction model. To address this concern, we have added Fig. S5 and Fig. S9 in the Supplementary Information to present relevant results, and have provided a detailed discussion of this issue in Section 5 (Limitations and Future Perspectives).

Changes to the manuscript:

Line 732–764 in track change:

The stability of AI reconstruction results largely depends on the spatial patterns of the training samples and the availability of original valid data. ~~In this study, the AI-based reconstruction approach inevitably exhibits systematic biases in polar regions during the early period without observational records. This indicates that while deep learning methods possess strong nonlinear fitting capabilities, they are constrained by the spatial structure and statistical characteristics of the input data. Moreover, we found that when missing regions are excessively large and contain no valid data, the AI model's prior correlation and RMSE deteriorate substantially, as observed in the pre-1961 polar regions. After introducing station observations post-1961 to increase effective coverage, the prior correlation improved and RMSE decreased, particularly for Merge-H, which had larger missing regions than Merge-E (Fig. S6). This finding emphasizes that while AI reconstruction can generate plausible climate fields under sparse observational conditions, sufficient image information or spatial constraints are required. Because the AI model does not explicitly include energy conservation or atmospheric dynamics constraints, extrapolations to data-sparse regions, such as early periods and polar regions, heavily rely on surrounding temperature spatial patterns, that is, the edge features of the missing regions (Liu et al., 2018). Thus, land-ocean boundaries, complex terrain, and extensive missing data areas are major sources of AI reconstruction errors. Future studies may improve reconstructions by incorporating terrain corrections, adjusting land-ocean weighting, introducing local dynamical parameterizations, or using physics-informed loss functions.~~ In this study, sensitivity experiments based on the PConv model indicate that the reconstructions may still exhibit a certain degree of systematic bias in the polar regions during the early period when observational records are sparse (Text S1; Figs. S5, S9). In particular, during periods with weak observational constraints (before the 1920s in the Arctic and before the 1960s in Antarctica), the model trained using 20CR and CMIP6 may introduce varying degrees of cold bias, which can reach approximately 0.5°C under the Merge-H configuration (Fig. S9). This phenomenon is likely attributable to two main factors. First, the extremely limited availability of early polar observations provides insufficient spatial constraints for the model, leading to substantially increased uncertainty when extrapolating over large data-sparse regions (Fig. 6). Second, systematic biases inherent in the training datasets may also propagate into the reconstruction through the deep learning process. For example, the limited observational constraints in the early polar regions within the 20CR reanalysis, together with biases in CMIP6 simulations of the polar climate state and its variability, may affect the spatial relationships learned by the model. Consequently, similar early cold biases are not unique to C-AIRST_M; they are also evident in the reconstruction of Kadow et al. (2020), which exhibits an approximately 0.5°C cold bias relative to other datasets (Fig. 6).

Therefore, the reconstruction results for these regions and periods with weak observational constraints should be regarded as exploratory estimates with considerable uncertainty and interpreted with caution when analyzing climate trends and temperature anomalies. Future improvements could be achieved by incorporating additional early proxy

records (e.g., ice cores, tree rings, and marine sediment records) and by adopting higher-resolution and more physically consistent reanalysis datasets as training constraints, thereby further improving the reliability of early climate field reconstructions in polar regions.

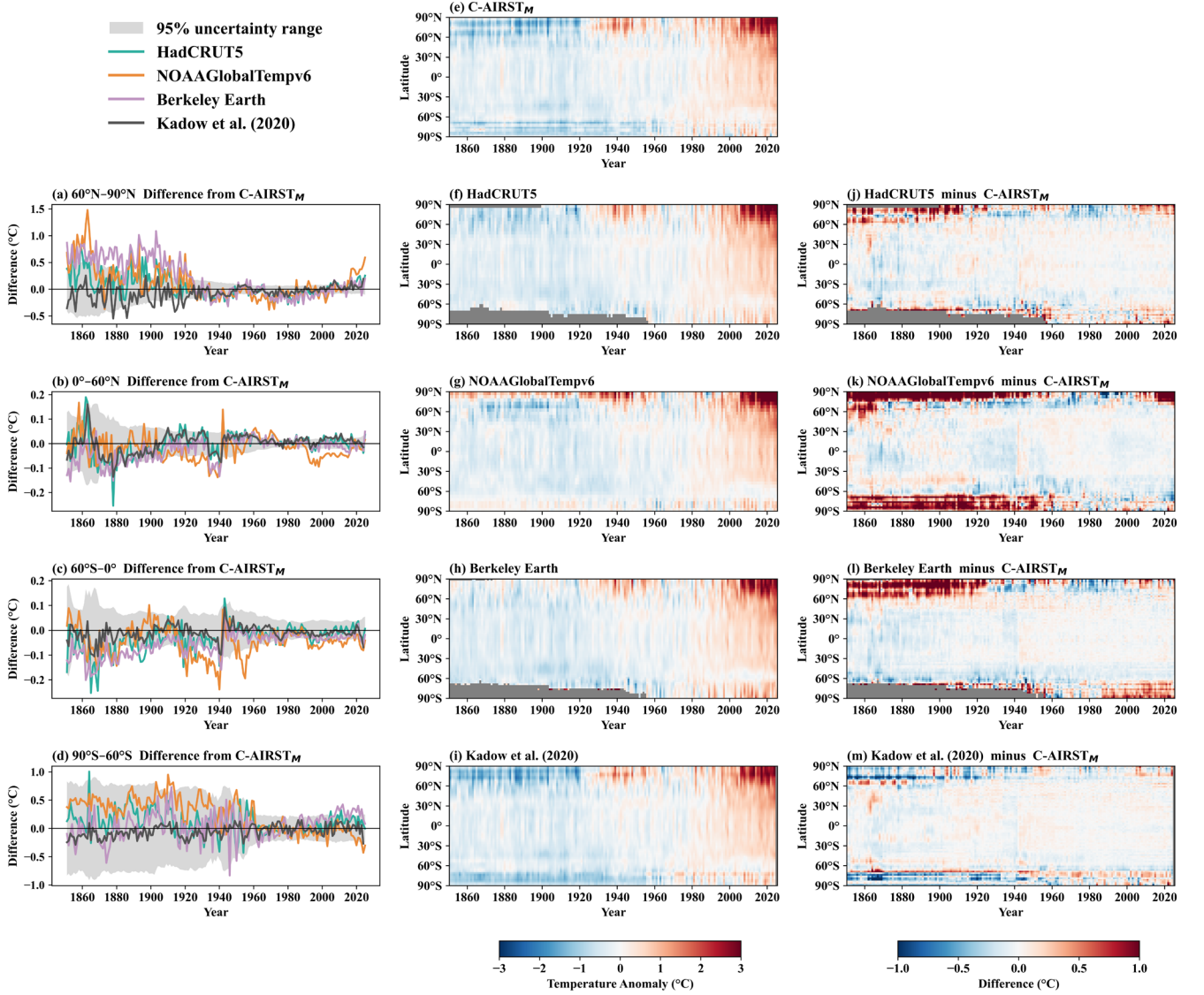


Figure 6 Zonal temperature comparison. The gray shading indicates the 95% uncertainty range estimated for C-AIRST_M. (a–d) time series of zonal-mean temperature anomaly differences between other datasets and C-AIRST_M across different latitude bands; (e–i) Zonal-mean temperature anomalies from C-AIRST_M, HadCRUT5, NOAAGlobalTempv6, Berkeley Earth, and Kadow et al. (2020), respectively; (j–m) zonal-mean temperature anomaly differences between other datasets and C-AIRST_M.

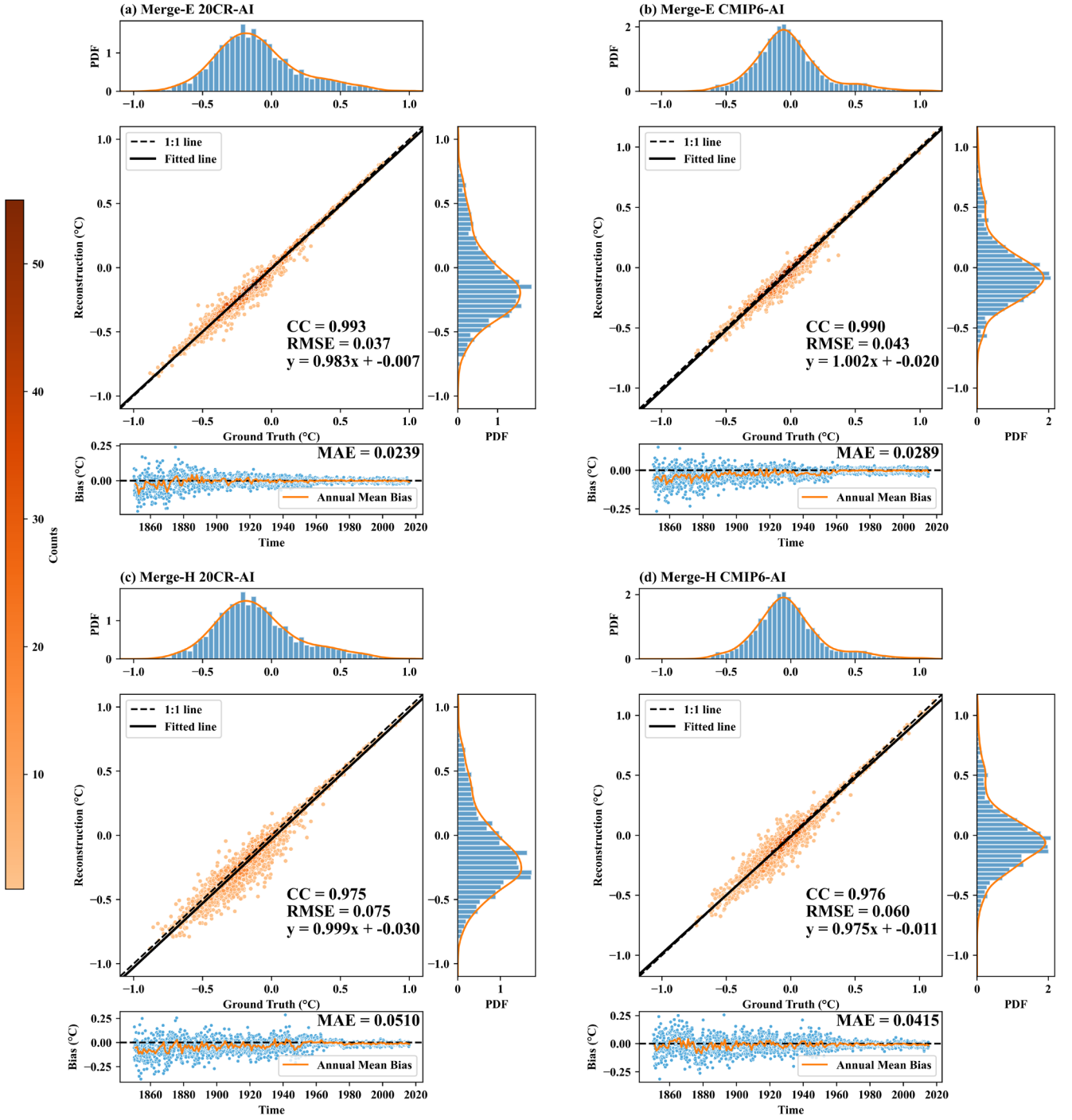


Figure S5: Model evaluation results for different sensitivity experiments. Each dot represents the monthly global area-weighted mean temperature from 1850 to 2014. The reconstruction performance was evaluated using the correlation coefficient (CC), root mean square error (RMSE), and mean absolute error (MAE). The histogram represents the probability density function (PDF) of the distribution. (a) Comparison between the Merge-E 20CR-AI reconstruction and the its original test dataset. (b–d) Same as (a), but for Merge-E CMIP6-AI, Merge-H 20CR-AI, and Merge-H CMIP6-AI, respectively.

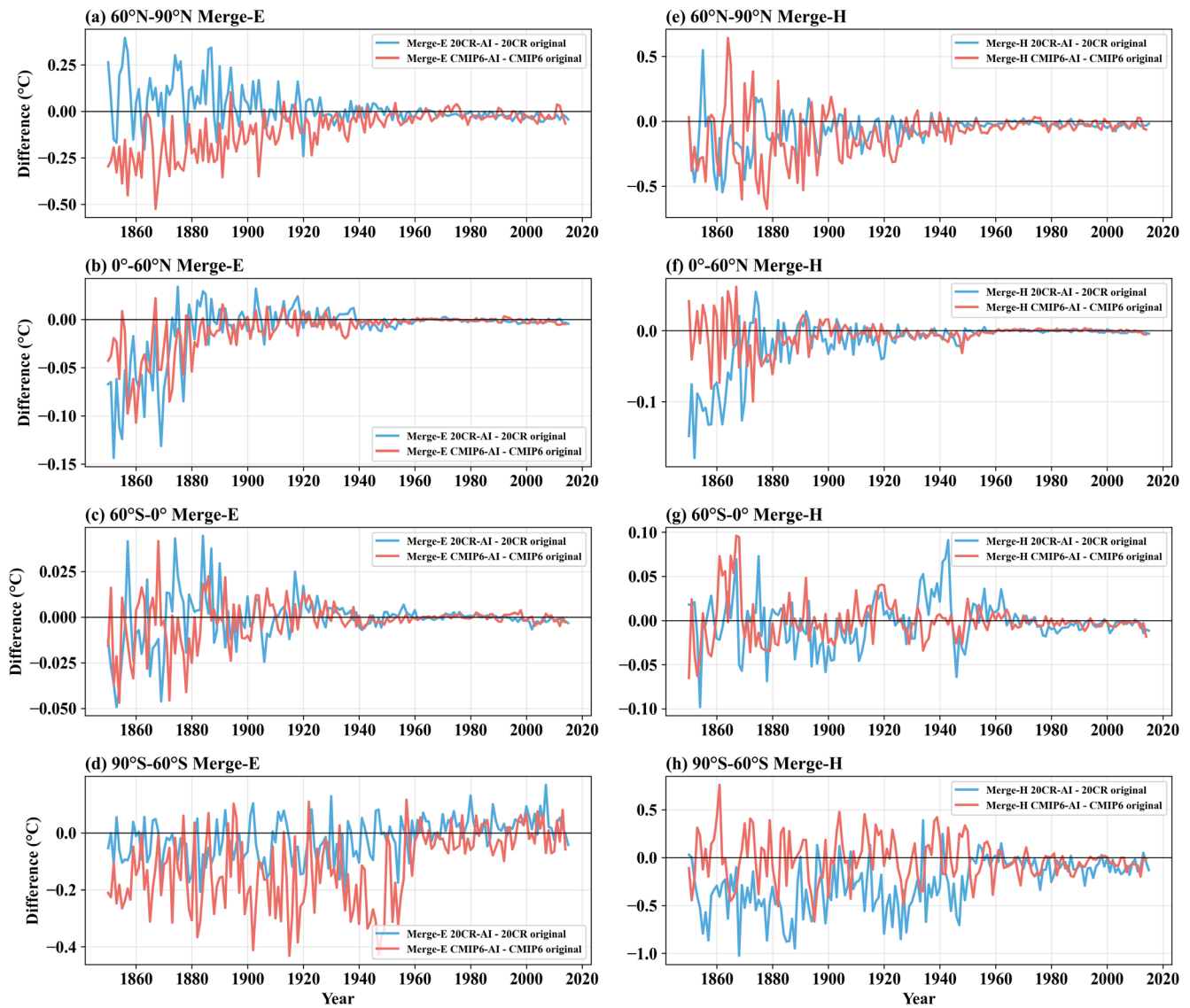


Figure S9: Annual mean difference time series across different latitude bands from 1850 to 2014. (a–d) Differences between the reconstructed test sets and original test sets under the Merge-E mask for the two models; (e–h) same as (a–d), but under the Merge-H mask.

Comments from reviewer:

3. The treatment of uncertainty remains strongly insufficient and remains the main unresolved issue about the revised dataset. The confidence intervals reported in the tables reflect only the statistical uncertainty of fitted linear trends, rather than uncertainty in the reconstructed data product itself. This distinction should be made explicit throughout the manuscript.

This is particularly important for the preferred Merge-H product, which uses HadSST4 as the ocean component. HadSST4 provides extensive uncertainty information, including measurement and sampling uncertainty, bias-adjustment uncertainty, covariance-based uncertainty estimates, and a 200-member ensemble. However, the authors do not clearly explain which HadSST4 field was used, whether the ensemble information was used, or why these uncertainty estimates were not propagated into the final reconstruction. Given that the ocean component strongly affects GMST, marine variability, and the early historical record, the uncertainty information already provided by HadSST4 should be explicitly incorporated into the analysis, or at minimum used to define uncertainty ranges or quality flags for the reconstructed product. It is difficult to justify not using the available HadSST4 uncertainty information in an ESSD data product.

Regarding the modeling uncertainty, the authors acknowledge that the PConv model itself is deterministic. However, this does not preclude the development of practical uncertainty estimates. Methodological uncertainty could still be addressed, for example, through a stochastic framework similar to that used in DCENT-I, or through sensitivity-based estimates based on alternative training datasets. Without such information, users cannot assess where the product is reliable and where it should be treated as exploratory.

It is noted that the expectations for uncertainty characterization differ between methodological demonstrations and operational data products. Earlier studies, such as Kadow et al. (2020) and Bochow et al. (2025), mainly demonstrated the feasibility of AI-based infilling. By contrast, the present manuscript is framed as an ESSD data-product paper and recommends a specific reconstruction for community use. In this context, uncertainty information is not merely an optional extension of the method, but an essential part of making the dataset interpretable and usable by the community.

Response: Regarding the uncertainty quantification, we developed an uncertainty quantification framework based on the approaches proposed by Kendall and Gal (2017) and Lakshminarayanan et al. (2017). Briefly, C-LSAT2.1 was combined with 200 HadSST4.2.0.0 ensemble members separately to generate 200 input ensembles, which were used to characterize the propagation of SST observational uncertainty. In addition, we adopted the Deep Ensemble method by independently training five PConv models to represent the uncertainty arising from the model itself. The 200 input ensembles were then reconstructed using the five independently trained PConv models, resulting in 1000 reconstructed ensemble members for uncertainty estimation. Since the input dataset (HadSST4.2.0.0) has been updated and uncertainty quantification has been newly incorporated into this study, the final data product and the corresponding descriptions in the manuscript have also been updated accordingly. Regarding the land observational uncertainty that was not explicitly incorporated into the uncertainty quantification, we have added a sensitivity discussion in Section 2.4 (Uncertainty Quantification).

Changes to the manuscript:

Line 149–163 in track change:

~~In this study, LSAT from C-LSAT2.1 are separately merged with the sea surface temperature (SST) products ERSSTv6 and the Met Office Hadley Centre's sea surface temperature (HadSST4.1.1.0) dataset (Kennedy et al., 2019), and the resulting merged datasets, referred to as “Merge E” and “Merge H”, are used as the original inputs for the AI-based reconstruction. Both datasets provide monthly mean ST from 1850 to 2024.~~

~~In addition, two 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 simulations from the Coupled Model Intercomparison Project Phase 6 (CMIP6; Eyring et al., 2016), 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). In this study, C-LSAT2.1 was used as the land surface air temperature (LSAT) component, while the 200-member HadSST4.2.0.0 sea surface temperature (SST) ensemble provided by the UK Met Office Hadley Centre (Kennedy et al., 2019) was used as the SST component. These datasets have a monthly temporal resolution and span the period from 1850 to 2024. A total of 105 ensemble members from Phase 6 of the Coupled Model Intercomparison Project (CMIP6; Eyring et al., 2016) historical simulations (Table S1) were used as training samples for the PConv model.~~

Table 1 List of information on the various data used in reconstruction 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)
HadSST4.2.0.0	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)

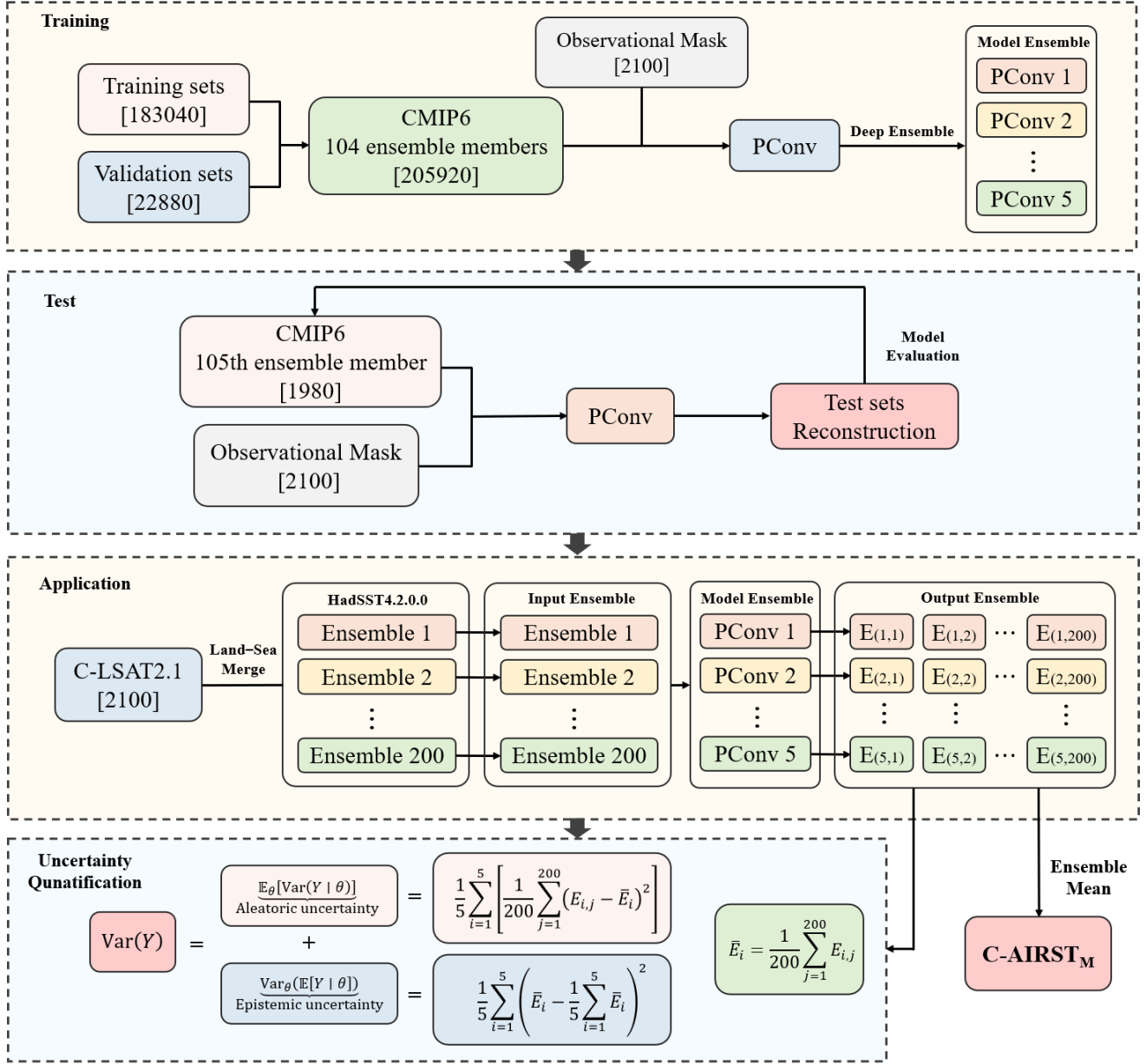


Figure 1: AI training and reconstruction process. The number in square brackets (e.g., [183040]) denotes the total number of the monthly temperature anomaly fields. In the output ensemble, $E(i, j)$ denotes the (j) -th ensemble member generated by the (i) -th PConv model.

Line 220–315 in track change:

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 ($5^{\circ} \times 2.5^{\circ}$) 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. The main contribution of this study is the integration of input observational uncertainty, represented by the HadSST4 ensemble members, and model uncertainty, represented by the Deep Ensemble approach, into a unified reconstruction framework. The law of total variance is then used to decompose and quantify these two sources of uncertainty, thereby providing predictive uncertainty estimates for reconstructed result. In addition to producing reconstructed global historical surface temperature fields, the proposed framework also provides associated uncertainty estimates, offering valuable confidence information for subsequent climate change analyses and the interpretation of reconstruction results.

The reconstruction workflow is illustrated in Figure 1 (~~adapted from Extended Data Fig. 1 of Kadow et al., 2020~~), and consists of the following steps:

- (1) Resolution standardization: The input datasets used for training are uniformly remapped from higher resolution to a spatial resolution of $5^{\circ} \times 2.5^{\circ}$ using conservative interpolation, resulting in a regular grid composed of 72×72 grid cells.
- (2) Sample partitioning: ~~Monthly temperature anomaly fields from 20CR (80 ensemble members) for 1850–2015 and CMIP6 (105 ensemble members) for 1850–2014 relative to the 1961–1990 climatology were divided into training and testing sets. One ensemble member was randomly selected as the test set, while the remaining 79 members from 20CR were used to train the 20CR-AI model, and 104 CMIP6 members were used to train the CMIP6-AI model.~~ The monthly temperature anomaly fields (relative to the 1961–1990 climatology) from 105 CMIP6 historical simulation ensemble members over the period 1850–2014 were used to construct the sample dataset. One ensemble member was randomly selected and reserved as the test set, while the remaining 104 CMIP6 ensemble members were used as the training set for the PConv model. The training samples were randomly partitioned, with 8/9 used for model training and 1/9 for validation. This resulted in 183,040 training samples and 22,880 validation samples for the AI model.
- ~~(3) Training sample construction: The training samples were randomly partitioned, with 8/9 used for model training and 1/9 for validation. This resulted in 139,883 training samples and 17,485 validation samples for the 20CR-AI model, and 183,040 training samples and 22,880 validation samples for the CMIP6-AI model.~~
- (4) Mask configuration: Observational data ~~from Merge-E and Merge-H~~ obtained by merging C-LSAT2.1 and HadSST4.2.0.0, ~~covering monthly values from 1850 to 2024~~ with monthly data spanning 1850–2024, were used as masks for missing values for the ~~respective~~ reconstruction ~~schemes~~. Observed grid points were marked as “1” and missing grid points as “0”, serving as inputs for model training and validation.
- (5) Model training and fine-tuning: ~~Both AI models were trained for 500,000 iterations with a learning rate of 0.0002, followed by an additional 500,000 iterations with a reduced learning rate of 0.00006. The batch size was set to 16, and computations were performed on an NVIDIA GeForce RTX 4060 GPU for approximately 8 hours.~~ During the PConv model training stage, five independent PConv models were trained using different random seeds (Seed = 1–5) for network initialization. Each model was trained for 500,000 iterations with a learning rate of 0.0002, followed by an additional 500,000 iterations with a reduced learning rate of 0.00006. The batch size was set to 16, and each training run required approximately 8 hours of computation on an NVIDIA GeForce RTX 4060 GPU.
- (6) Model evaluation: The CMIP6 test ensemble member was input into the single trained PConv model, and the reconstruction was compared with the original test ensemble member to evaluate model performance (see Text S1).
- (7) Reconstruction: ~~The datasets Merge-E and Merge-H, which contain missing values, are both input into the trained 20CR-AI and CMIP6-AI models for inference, with the output being globally complete monthly surface temperature datasets at a spatial resolution of $5^{\circ} \times 2.5^{\circ}$ for the period 1850–2024.~~ C-LSAT2.1 was merged with each of the 200 ensemble members from HadSST4.2.0.0 to generate 200 input ensemble members. Each input ensemble member was then input into the five trained PConv models, producing a total of 1,000 globally complete monthly temperature reconstructions at a spatial resolution of $5^{\circ} \times 2.5^{\circ}$ for the period 1850–2024. The ensemble mean of all reconstructions was taken as the final reconstructed dataset, referred to as China global Artificial Intelligence Reconstructed Surface Temperature_{CMIP6} (C-AIRST_M).

2.4 Uncertainty Quantification

In deep learning, predictive uncertainty can generally be decomposed into aleatoric uncertainty and epistemic uncertainty (Kendall and Gal, 2017). Aleatoric uncertainty arises from the inherent randomness of the observation process and measurement noise, whereas epistemic uncertainty reflects uncertainty in model parameters due to limited knowledge of the underlying data-generating process (Abdar et al., 2021).

In this study, the 200 ensemble members provided by HadSST4.2.0.0 were used to approximate the observational uncertainty of sea surface temperature (SST). The HadSST4 ensemble members are generated using an observational error model that accounts for multiple sources of uncertainty, including measurement errors, sampling errors, and bias-

adjustment uncertainties. Each ensemble member represents a physically plausible realization of the SST field constrained by the available observations (Kennedy et al., 2019). The C-LSAT2.1 dataset was merged separately with each HadSST4 ensemble member to construct 200 model input ensembles. Because these ensemble members collectively represent the irreducible uncertainty associated with the observation process, they were regarded as the aleatoric uncertainty of the input data, which was propagated through the reconstruction model via ensemble propagation (Kendall and Gal, 2017).

Epistemic uncertainty, also referred to as model uncertainty, was estimated using the Deep Ensemble approach, which is one of the most effective methods for quantifying epistemic uncertainty in deep learning (Lakshminarayanan et al., 2017). Specifically, five PConv models with identical network architectures, training datasets, and hyperparameters were independently trained using different random seeds for parameter initialization. The model with the best validation performance from each training run was retained to form the ensemble. Owing to different random initializations and stochastic optimization trajectories, the independently trained models converge to different parameter configurations, thereby reflecting uncertainty in model parameter estimation (Lakshminarayanan et al., 2017).

The 200 input ensembles were subsequently fed into each of the five trained PConv models, resulting in a total of 1,000 reconstructed temperature fields. This reconstruction ensemble simultaneously incorporates the propagated observational uncertainty from the input data and the prediction variability arising from different model parameterizations, thereby jointly representing both aleatoric and epistemic uncertainty. The ensemble mean of the 1,000 reconstructions was adopted as the final reconstructed temperature product. The total predictive uncertainty was decomposed according to the Law of Total Variance (Valdenegro-Toro and Mori, 2022):

$$\text{Var}(Y) = \underbrace{\mathbb{E}_{\theta}[\text{Var}(Y | \theta)]}_{\text{Aleatoric uncertainty}} + \underbrace{\text{Var}_{\theta}(\mathbb{E}[Y | \theta])}_{\text{Epistemic uncertainty}}$$

Here, θ denotes the parameters of each independently trained PConv model in the Deep Ensemble, and Y represents the reconstructed temperature field. The first term, $\mathbb{E}_{\theta}[\text{Var}(Y | \theta)]$, represents the propagated observational (aleatoric) uncertainty. For each fixed model θ , the variance of the predicted outputs was first calculated from the 200 ensemble predictions, yielding $\text{Var}(Y | \theta)$. The average of these variances across the five models was then taken to obtain $\mathbb{E}_{\theta}[\text{Var}(Y | \theta)]$. The second term, $\text{Var}_{\theta}(\mathbb{E}[Y | \theta])$, represents the model (epistemic) uncertainty. Specifically, for each model θ , the ensemble mean prediction $\mathbb{E}[Y | \theta]$ was first computed from the 200 reconstructed outputs. The variance among the mean predictions of the five independently trained models was then calculated to quantify the uncertainty arising from differences in model parameters.

The sum of these two components yields the total predictive variance. Its square root corresponds to the predictive standard deviation σ , and the 95% predictive uncertainty interval is reported as 1.96σ .

It should be noted that the total predictive uncertainty estimated in this study represents an approximation of the true predictive uncertainty. Although the observational uncertainty of the SST input data is propagated through the reconstruction model, the observational errors from C-LSAT2.1 are not explicitly incorporated into the uncertainty quantification framework. According to the uncertainty assessment of the C-LSAT2.1 dataset by Li et al. (2026b), the maximum 1σ station, sampling, and bias uncertainties are approximately 0.065°C , 0.025°C , and 0.005°C , respectively, in 1850, and all decrease to below approximately 0.02°C after the early twentieth century. Their omission is expected to have only a minor influence on the total predictive uncertainty estimated in this study. Consequently, the estimated uncertainty mainly reflects the contributions from SST input uncertainty and model uncertainty. The resulting total predictive uncertainty can therefore be regarded as a lower-bound estimate of the true predictive uncertainty, and the actual uncertainty over land regions is likely to be larger than that reported in this study.

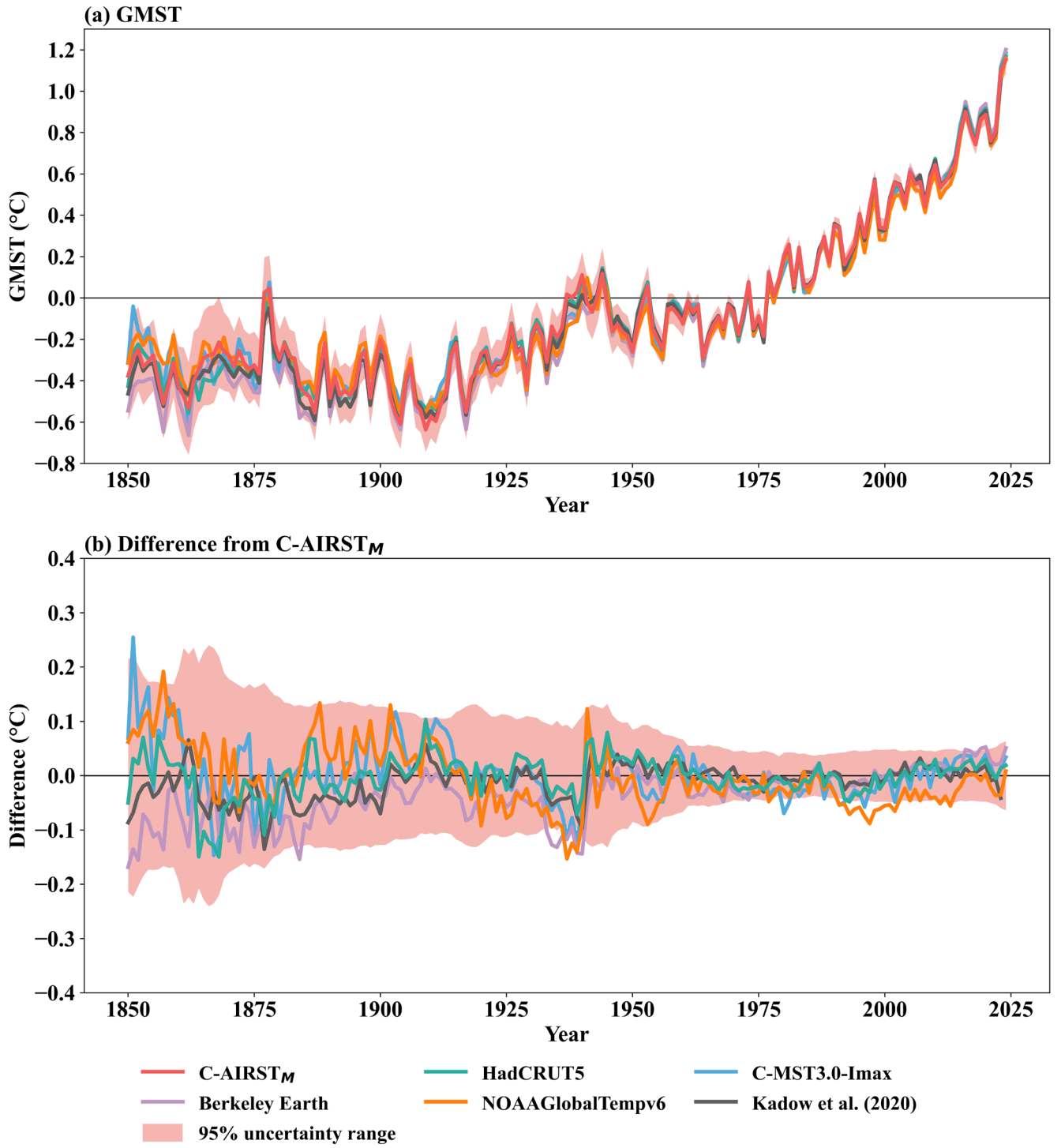


Figure 4: Global mean surface temperature (GMST) time series from 1850 to 2024 (relative to the 1961–1990 climatology). The red shading indicates the 95% uncertainty range estimated for C-AIRST_M. (a) GMST; (b) GMST differences from C-AIRST_M.

Table 2: 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
C-AIRST _M	0.064±0.006	0.101±0.008	0.155±0.013	0.194±0.023	1.49
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

Line 402–454 in track change:

The Global mean surface temperature (GMST) time series reconstructed by AI is shown in Figure 4, reflecting the long-term variations of global mean temperature from 1850 to 2024. ~~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. S8), leading to lower long-term trends and GWL estimates compared with Merge-H (Table 2). 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.~~

From 1850 to the early 20th century, global mean temperature remained generally stable (Fig. 4a), with occasional short-term cooling signals associated with enhanced volcanic activity and slight decreases in solar radiation (Fang et al., 2022). From the 1910s to the 1940s, the global mean temperature experienced the first rapid and sustained warming phase of the 20th century, with a pronounced increase in magnitude. From the mid-20th century to the mid-1970s, the warming trend weakened, followed by a renewed and more pronounced warming phase from the late 1970s onward, which continues to the present, reflecting the dominant influence of anthropogenic greenhouse gas emissions (IPCC, 2021). ~~Notably, none of the AI-based reconstruction series exhibit a significant “warming slow down” during 1998–~~

2012, consistent with the findings of Li et al. (2021).

The ~~Merge-H-20CR-AI and Merge-H-CMIP6-AI~~ C-AIRST_M show only minor differences in long-term reconstructed series and trends, and are generally consistent with the HadCRUT5 record (Fig. 4b, Table 2), with overall biases within 0.15 °C. After 1960, differences between the AI reconstructions and other datasets further decrease to within 0.05 °C. ~~Around 1940, several datasets exhibit a warm bias of approximately 0.12 °C relative to the Merge-H-CMIP6-AI reconstruction (Fig. 4b), which is attributed to non-standard temperature measurement practices during World War II, including changes in sea surface temperature observation methods (Chan and Huybers, 2021; Kent and Kennedy, 2021).~~

~~Both Merge-H-AI reconstructions yield a long-term trend of 0.064 ± 0.006 °C/decade over 1850–2024 (Table 2). The global warming level (GWL) for 2024 derived from the Merge-H-CMIP6-AI reconstruction reaches 1.52 °C, slightly higher than the 1.50 °C estimated from the Merge-H-20CR-AI reconstruction. This indicates that both AI-based reconstructions suggest that the year 2024 is approaching or has already reached the 1.5 °C global warming threshold relative to the pre-industrial period (1850–1900). During 1930–1940, C-AIRST_M was approximately 0.1 °C warmer than the other datasets, whereas it was slightly cooler during the 1940–1945 World War II period (Fig. 4b). These differences mainly arise from the revised historical SST bias corrections implemented in HadSST4.2.0.0 (Sandford and Rayner, 2026). In particular, the systematic warm bias introduced by the widespread use of Engine Room Intake (ERI) measurements during World War II was corrected (Chan and Huybers, 2021; Kent and Kennedy, 2021). In addition, a truncation error in the Kobe SST records resulted in an approximately 0.4 °C cold bias during 1930–1940, which was corrected in HadSST4.2.0.0 (Sandford and Rayner, 2026). This SST correction also contributed to a higher estimate of the GMST in C-AIRST_M during the same period. The long-term warming trends reconstructed by the AI model for different periods during 1850–2024 (Table 1) are generally consistent with the estimates from other global temperature datasets, including Berkeley Earth, HadCRUT5, and NOAA GlobalTempv6. The AI reconstruction also yields a Global Warming Level (GWL) of 1.49 °C, with only small differences relative to the other datasets, indicating that global warming in 2024 had approached the 1.5 °C warming threshold relative to the pre-industrial period (1850–1900). Overall, the AI reconstruction series shown in Fig. 4 reasonably reproduce the phase-dependent characteristics of global climate evolution over the past 175 years at interannual scales.~~

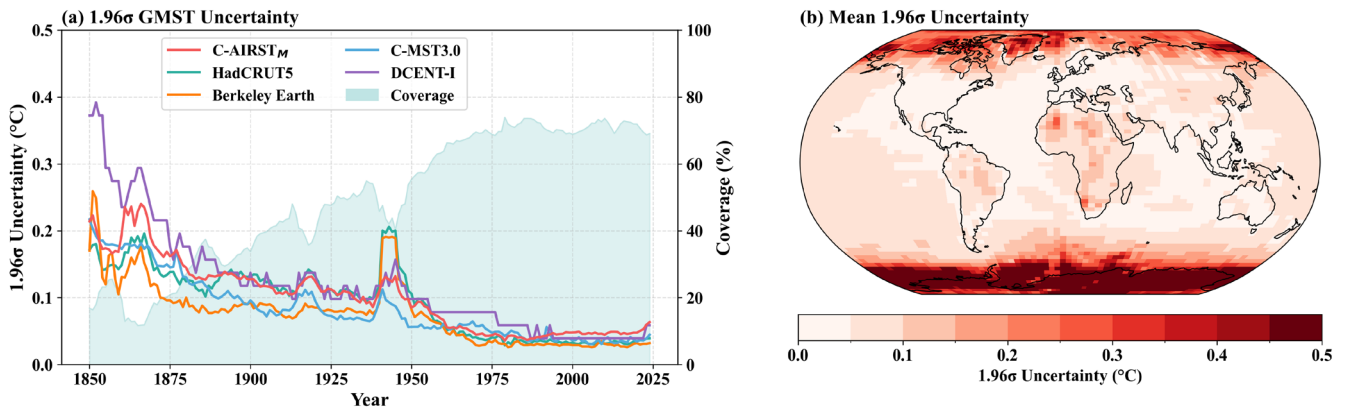


Figure 5: Uncertainty estimation. (a) Time series of the 1.96σ uncertainty in GMST from different datasets; (b) spatial distribution of the mean 1.96σ uncertainty estimated for C-AIRST_M.

Line 456–471 in track change:

Regarding the uncertainty estimation of GMST (Fig. 5), the uncertainty quantified in this study is not completely comparable to that of mainstream observational datasets, which generally consider multiple sources of uncertainty, including measurement errors, sampling errors, bias adjustments, and incomplete spatial coverage. The predictive uncertainty estimated here incorporates the propagation of input uncertainty represented by the 200 ensemble members of HadSST4.2.0.0, as well as the model uncertainty arising from the deep learning framework itself (see Section 2.4 for details on uncertainty quantification). Despite the differences in uncertainty sources, the estimated uncertainty in GMST is generally comparable to that of existing mainstream datasets. The 1.96σ uncertainty gradually decreases from approximately 0.2 °C in the 1850s to approximately 0.05 °C after the 1960s (Fig. 5a). In addition, uncertainty exhibits temporary increases during the two World War periods (the 1910s and 1940s), likely associated with reduced global

observational coverage.

The spatial distribution of uncertainty (Fig. 5b) shows that the polar regions have substantially higher uncertainty than the low- and mid-latitude regions due to the long-term scarcity of observational records. The uncertainty in the polar regions generally exceeds 0.3°C , whereas that in most low- and mid-latitude regions remains below 0.2°C . Notably, Antarctica exhibits substantially larger uncertainty than the Arctic, both in terms of the magnitude of uncertainty and the spatial extent of high-uncertainty regions. This difference is primarily related to the much sparser observational network and insufficient historical observations in Antarctica compared with the Arctic.

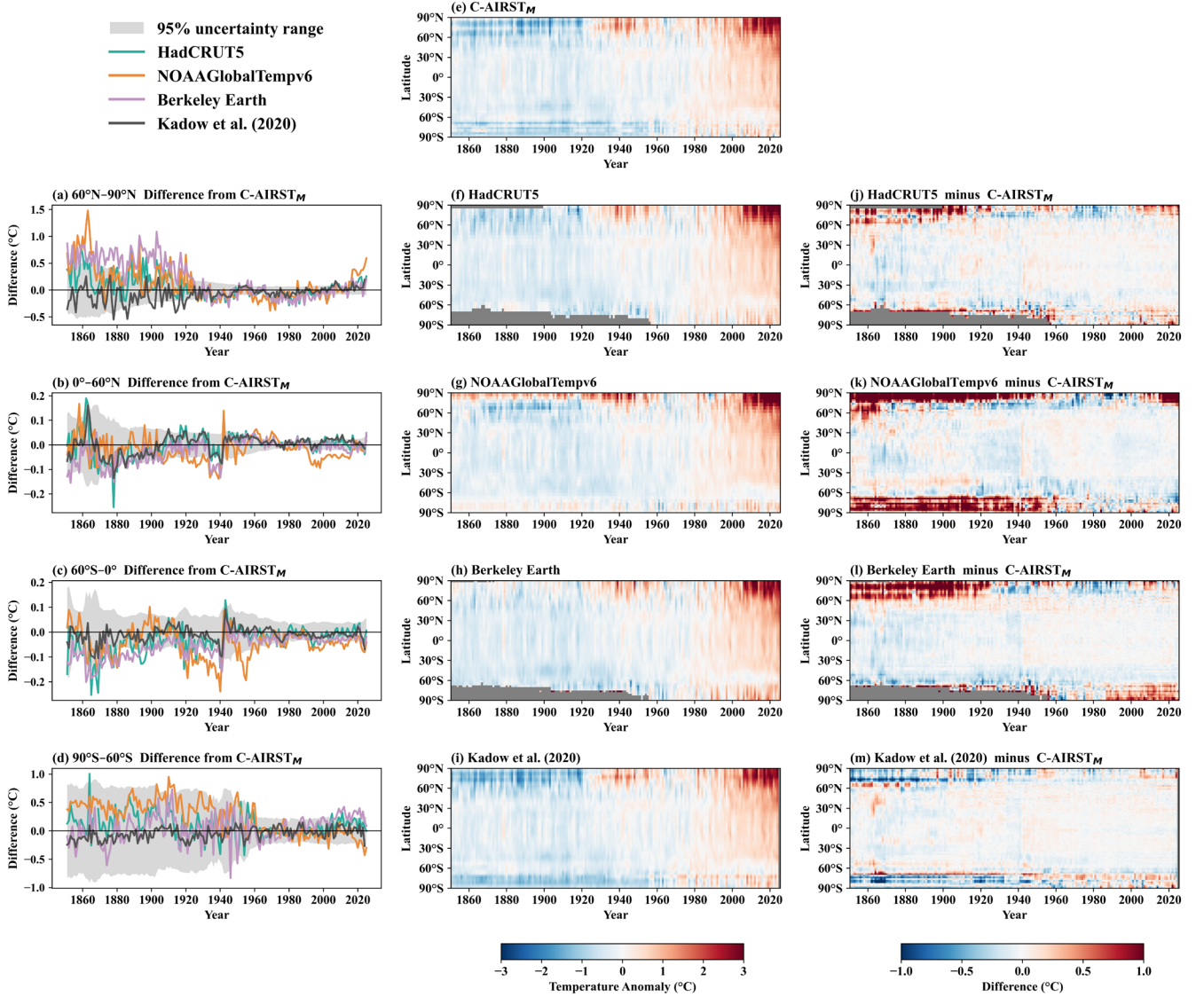


Figure 6 Zonal temperature comparison. The gray shading indicates the 95% uncertainty range estimated for C-AIRST_M. (a–d) time series of zonal-mean temperature anomaly differences between other datasets and C-AIRST_M across different latitude bands; (e–i) Zonal-mean temperature anomalies from C-AIRST_M, HadCRUT5, NOAAGlobalTempv6, Berkeley Earth, and Kadow et al. (2020), respectively; (j–m) zonal-mean temperature anomaly differences between other datasets and C-AIRST_M.

Line 478–506 in track change:

In the zonal temperature comparison across different datasets (Fig. 6), 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 C-AIRST_M generally remain within $\pm 0.1^{\circ}\text{C}$ to $\pm 0.2^{\circ}\text{C}$ prior to the early 20th century (Figs. 6b, 6c). The Northern Hemisphere (NH), characterized by a larger land fraction and denser observational networks, is substantially better constrained by observations than the predominantly ocean-covered Southern Hemisphere (SH). 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 1960s. The 95% uncertainty intervals estimated by C-AIRST_M in both regions remain within $\pm 0.1^{\circ}\text{C}$, indicating relatively low uncertainty. 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. 6a), prior to 1920, both C-AIRST_M and Kadow et al. (2020) exhibit a cold bias of approximately 0.6 °C relative to HadCRUT5, NOAAGlobalTempv6 and Berkeley Earth, while this bias is smaller in Merge-H 20CR-AI. It is also observed that the AI reconstructions (Figs. 6j–6l) 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. 6d), during 1850–1960, C-AIRST_M exhibits a cold bias of approximately 0.5 °C relative to other datasets. However, the estimates from different datasets are not entirely consistent, 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 C-AIRST_M 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. S4f and S4h, 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.

The above analyses suggest that the reconstructions exhibit some systematic cold bias in the early Arctic (before 1920s) and Antarctica (before 1960s), where observational constraints are relatively weak. Correspondingly, the 95% uncertainty ranges in these periods are also relatively large, reaching approximately $\pm 0.5^{\circ}\text{C}$ in the Arctic (Fig. 6a) and $\pm 0.8^{\circ}\text{C}$ in Antarctica (Fig. 6d). Consequently, the reconstructed temperatures for these periods and regions should be interpreted as exploratory and treated with appropriate caution.

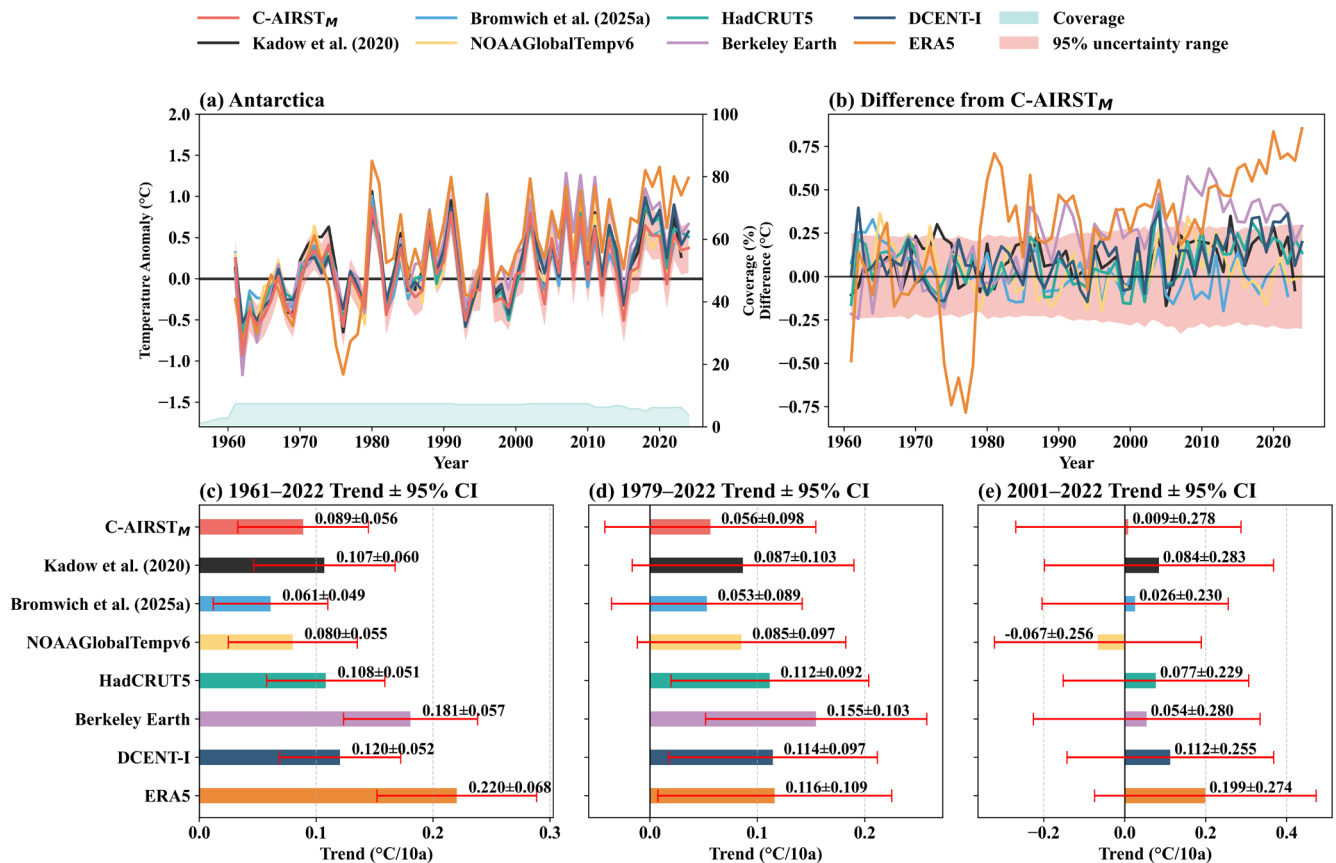


Figure 10: (a) Annual mean surface temperature anomaly time series over Antarctica from 1961 to 2024, and coverage refers to the proportion of Antarctic grid cells that are occupied by observational grid points before reconstruction; (b) Temperature difference from C-AIRST_M; (c–e) Linear trend of annual mean temperature and 95% confidence interval (°C per decade) during 1961–2022, 1979–2022 and 2001–2022.

Line 629–665 in track change:

Fig. 10 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 periods. To comprehensively evaluate the reliability and performance of the AI reconstructions, the results were systematically compared with ~~the spatially weighted mean of 81 Antarctic observational stations used in this study (after mapping the station data onto grid cells; Station in Figure 9)~~, the reconstructions from Kadow et al. (2020) and Bromwich et al. (2025a), HadCRUT5, NOAAGlobalTempv6, Berkeley Earth, DCENT-I and the ERA5 reanalysis product. The comparisons include a quantitative assessment of Antarctic annual mean temperature time series (Fig. 10) and Antarctic subregional temperature series (Fig. S13), as well as a qualitative analysis of the spatial distribution of temperature trends (Figs. 11 and S14).

Overall, ~~the AI reconstructions~~ C-AIRST_M shows 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. 10a). C-AIRST_M shows good agreement with the reconstruction from Bromwich et al. (2025a), NOAAGlobalTempv6, and HadCRUT5, with differences generally within 0.25°C (Fig. 10b), and its 95% uncertainty interval remained around 0.25°C.

During 1961–2022, the linear warming trends derived from ~~Merge-H 20CR-AI and Merge-H CMIP6-AI~~ C-AIRST_M is ~~0.079 ± 0.056 °C/decade and 0.089 ± 0.057 °C/decade, respectively~~. The magnitude (Fig. 10c) is close to those reported by Bromwich et al. (2020) and NOAAGlobalTempv6, and are lower than those from HadCRUT5, Berkeley Earth, DCENT-I, and 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 (Figs. S3 and 10). This sampling structure causes the Station-based estimates in Fig. 9 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.~~ During 1979–2022, the trend from C-AIRST_M remains statistically insignificant, consistent with Bromwich et al. (2025a) and NOAAGlobalTempv6, whereas HadCRUT5, Berkeley Earth, and ERA5 still exhibit statistically significant warming trends (Fig. 10d). 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. S13, this bias mainly originates from warming contributions over East Antarctica and West 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 (see Fig. S12 for Antarctic geographical references), 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. ~~Both the AI reconstructions~~ C-AIRST_M and all datasets ~~except ERA5~~ indicate that Antarctic warming is not statistically significant ~~from 2001 to 2022~~ (Fig. 10d). During this period of accelerated global mean temperature increase (Fig. 4a), no significant warming is detected over Antarctica, suggesting that internal climate variability continues to exert a stronger influence on Antarctic climate than external forcing, warranting further attention.

Comments from reviewer:

4. The manuscript's structure remains difficult to follow because Merge-E, Merge-H, 20CR-AI, and CMIP6-AI are all discussed as if they are parallel final products, while the authors ultimately recommend Merge-H CMIP6-AI. I suggest focusing the main text on the recommended product and moving Merge-E and other non-preferred configurations largely to the Supplement as sensitivity experiments. A short decision table explaining why Merge-H CMIP6-AI is preferred would help readers.

Response: We acknowledge that including the results of the Merge-E, Merge-H, 20CR-AI, and CMIP6-AI experiments in the main text may reduce the readability of the manuscript. These experiments are more appropriately presented as sensitivity experiments and model evaluations in the Supplementary Information. Therefore, we have removed these results from the main text and transferred them to the Supplementary

Information. In the revised manuscript, only C-AIRST_M is presented and described as the final reconstruction product. Accordingly, related contents throughout the manuscript have been revised.

In addition, we have added descriptions of the sensitivity experiments and the corresponding decision-making process in the Supplementary Information. All these modifications are summarized in the final section in this response, “Changes to Supplementary Information.”

Changes to the manuscript:

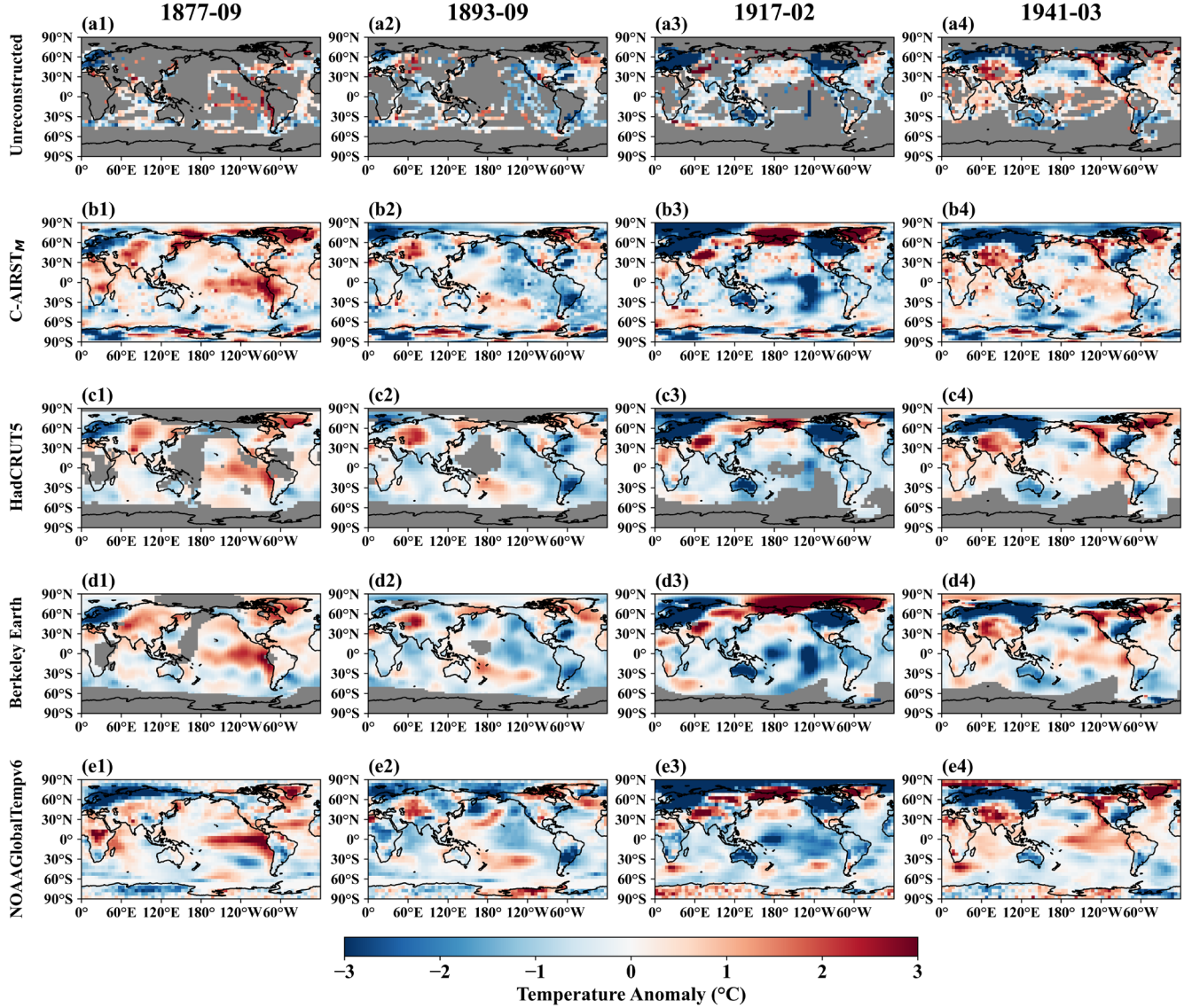


Figure 2: Global temperature anomaly (relative to the 1961–1990 climatology) fields before and after reconstruction for four typical ENSO events. (a1–a4) Unreconstructed C-AIRST_M; (b1–b4) C-AIRST_M; (c1–c4) HadCRUT5; (d1–d4) Berkeley Earth; (e1–e4) NOAA GlobalTempv6. From left to right, the columns show the temperature anomaly fields for September 1877, September 1893, February 1917, and March 1941, respectively.

Line 351–375 in track change:

Prior to the 19th century, the C-LSAT2.1 dataset contained substantial missing data in Asia, South America, Africa, and Antarctica due to the sparse distribution of land-based stations, and this situation was improved considerably after the 19th century (Wei et al., 2025). In addition, Antarctic observational stations began recording data progressively after the 1957/1958 International Geophysical Year; however, observational coverage in Antarctica remains extremely sparse (Wang et al., 2023). ~~As shown in Figure 2 and S3, the original ERSSTv6 dataset exhibits higher ocean data coverage compared to HadSST4.~~ As shown in Fig. 2, the ocean coverage of NOAA GlobalTempv6 is higher than that of the unreconstructed C-AIRST_M, which is attributed to the ocean component included in the former (ERSSTv6) employs an ANN reconstruction method, which allows for more accurate and stable reconstruction of SST in sparsely observed regions, reduces excessive smoothing, and better preserves spatial variability (Huang et al., 2025a, 2025b). In contrast,

HadSST4 primarily relies on statistical interpolation and ensemble-based methods, which aim to correct biases and estimate uncertainties within grid cells, without performing large-scale spatial reconstruction (Kennedy et al., 2019).

~~Considering the differences between the two SST datasets, this study performed AI-based reconstruction using two distinct merged datasets to investigate potential discrepancies under different conditions. Figures 2 (b1–b4, e1–e4, e1–e4, f1–f4) show the complete temperature anomaly fields reconstructed from Merge E and Merge H using the 20CR AI and CMIP6 AI models. In Fig. 2 (b1–b4, e1–e4), 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 b1). 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 b2–b4), the spatial patterns produced by the AI model are generally consistent with those from HadCRUT5 (Fig. 2 c2–c4) and Berkeley Earth (Fig. 2 d2–d4), 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 e2–e4). This discrepancy primarily arises from inherent differences between HadSST and ERSSTv6.

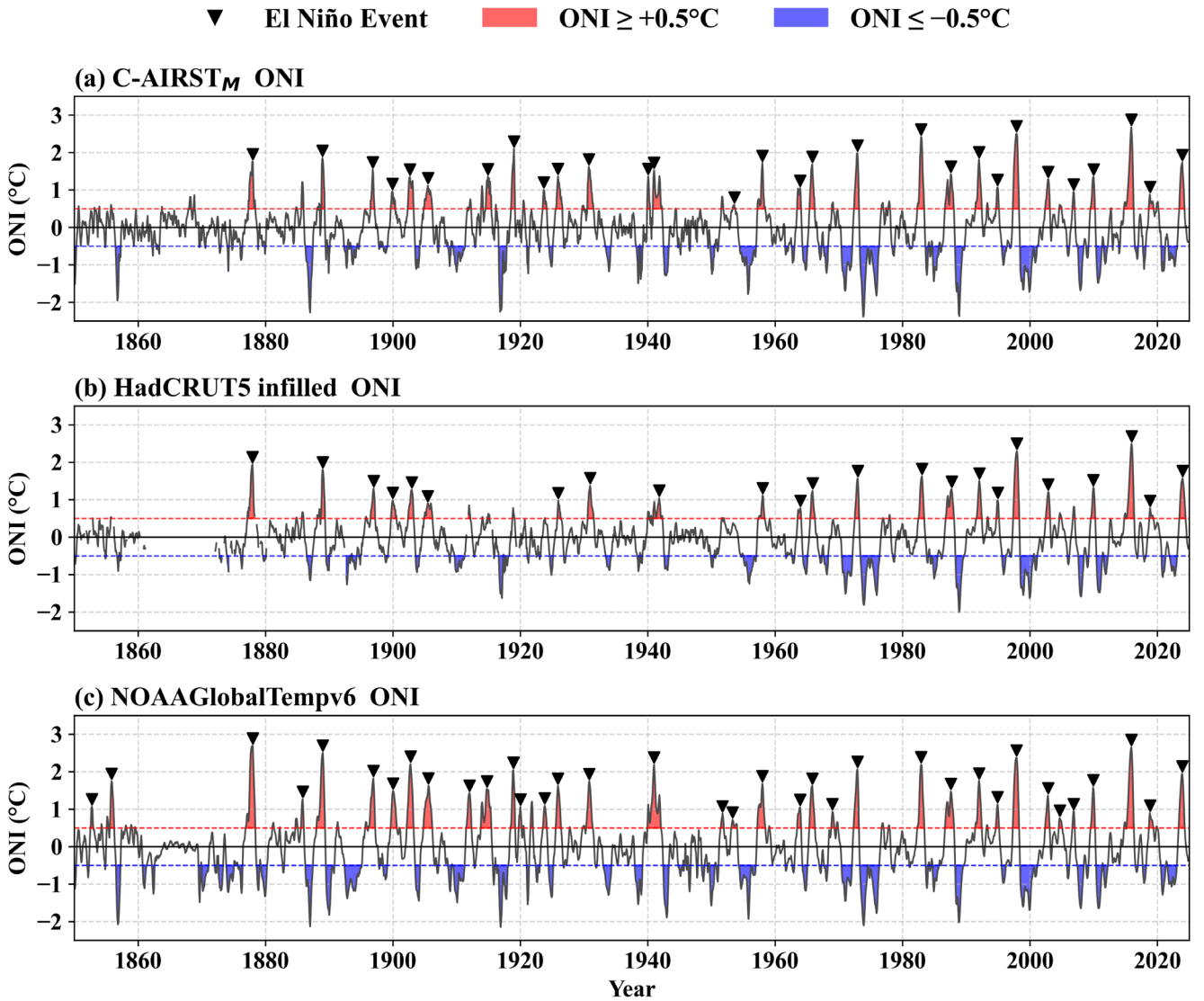


Figure 3: Ocean Niño Index (ONI) time series indicating ENSO events. (a) Time series of ONI from C-AIRST_M over 1850–2024, with black triangles indicating El Niño events defined as ONI exceeding 0.5 °C for at least five consecutive months; (b–c) same as (a) but for HadCRUT5 infilled, and NOAAGlobalTempv6,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

Line 380–393 in track change:

~~The two AI models are~~ AI model is 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. 3c).

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, and this discrepancy is primarily attributable to differences in the SST bias correction methods and spatial infilling approaches employed by the two datasets (Morice et al., 2021; Huang et al., 2025a; 2025b). Consequently, the Merge-H results C-AIRST_M 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.

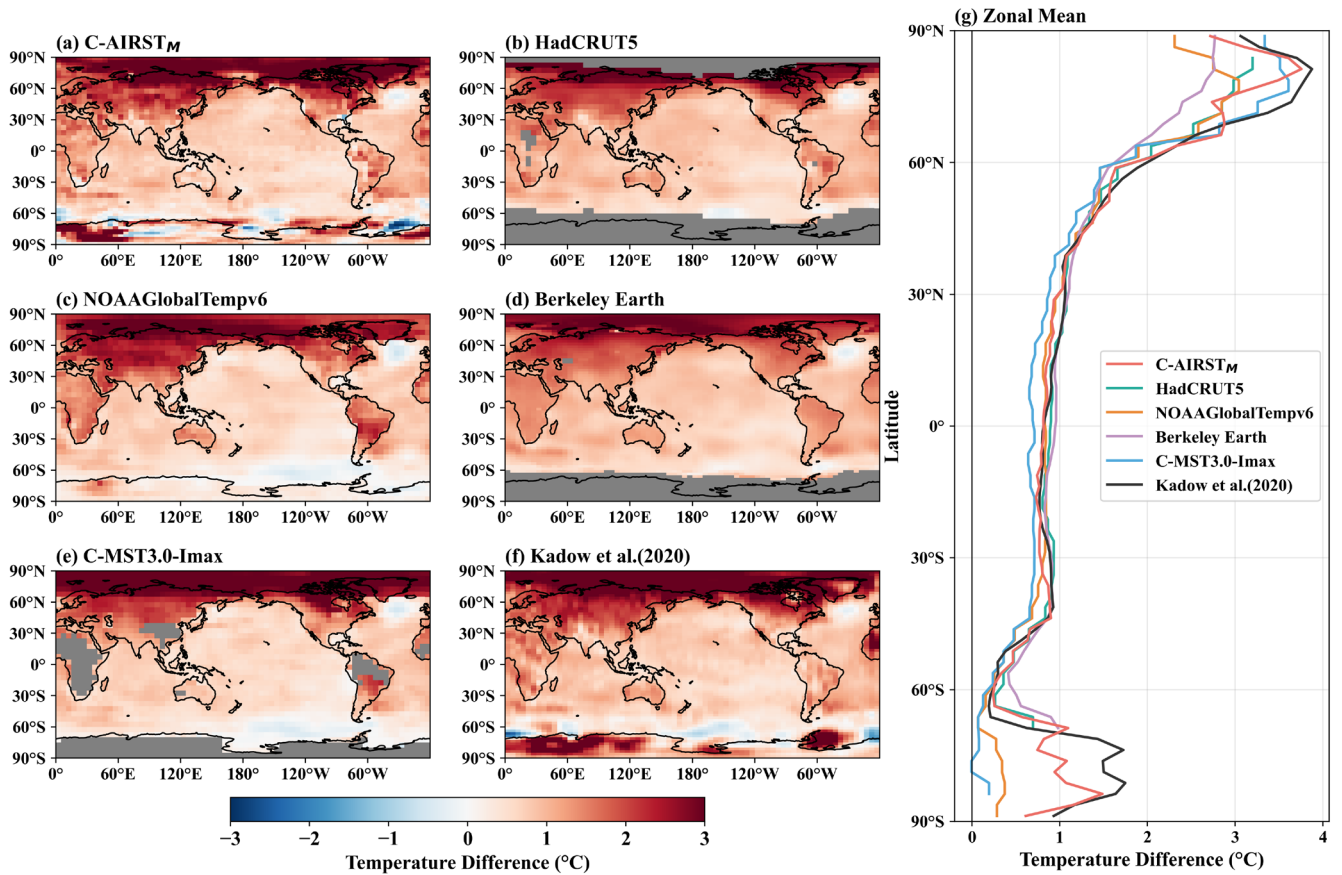


Figure 7 Global warming pattern. (a–f) Spatial distribution of mean global temperature warming over 2005–2020 relative to the 1850–1900 baseline period, (g) zonal-mean profile of warming magnitude.

Line 511–549 in track change:

Under the context of global warming, the spatial pattern of warming level from ~~reconstructed by the two AI model~~ C-AIRST_M is generally consistent with those of other datasets (Fig. 7). The main discrepancies are concentrated south of 60°S, where the PConv-based reconstruction (Figs. 7a, 7f) 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 (Figs. 2 b1, b2).

~~As shown in Fig. 7g, between 60°S and 80°S, Merge-H 20CR AI displays zonal variations of $\pm 1^\circ\text{C}$ 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 $\pm 1^{\circ}\text{C}$ near 80°S .~~

The ~~two AI reconstructions~~ C-AIRST_M and reconstruction from Kadow et al. (2020) 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.

The latitudinal warming amplitudes further reveal the consistency and differences among the reconstructed products in representing polar warming (Fig. 7g). C-AIRST_M and the reconstruction from Kadow et al. (2020) exhibit slightly stronger warming amplitudes near 80°N than the other datasets. In Antarctica, the warming amplitude estimated by C-AIRST_M is higher than that of NOAAGlobalTempv6 by approximately 0.7°C in the polar regions, respectively. These differences mainly arise from variations in the estimated baseline climatology for 1850–1900 among different datasets.

Furthermore, the zonal distribution of warming (Fig. 7g) shows a pronounced Arctic amplification effect, with warming in the Arctic substantially exceeding the global mean level. During 2005–2020, relative to the pre-industrial baseline period of 1850–1900, the cumulative warming in the Arctic region reached approximately $2.6\text{--}4.0^{\circ}\text{C}$, markedly higher than the corresponding global mean warming over the same period.

Fig. 8 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-H20CR-AI and Merge-H-CMIP6-AI are~~ C-AIRST_M is generally consistent with the reconstruction by Kadow et al. (2020). The spatial gradient of LSAT trends reconstructed by these two 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.

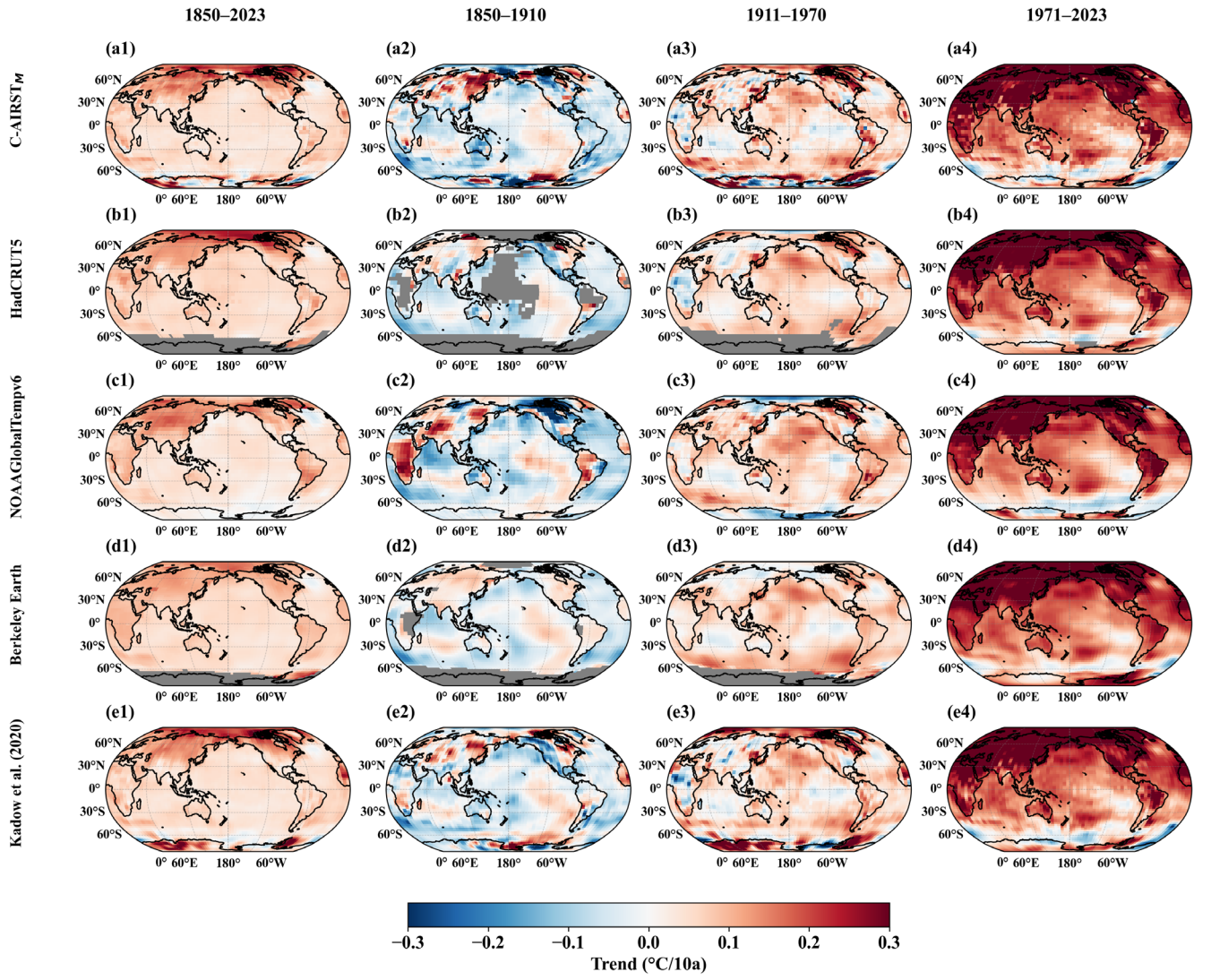


Figure 8 Spatial distribution of global surface temperature trends for 1850–2023, 1850–1910, 1911–1970, and 1971–2023. (a1–a4) C-AIRST_M, (b1–b4) HadCRUT5, (c1–c4) NOAAGlobalTempv6, (d1–d4) Berkeley Earth, (e1–e4) reconstruction from Kadow et al. (2020).

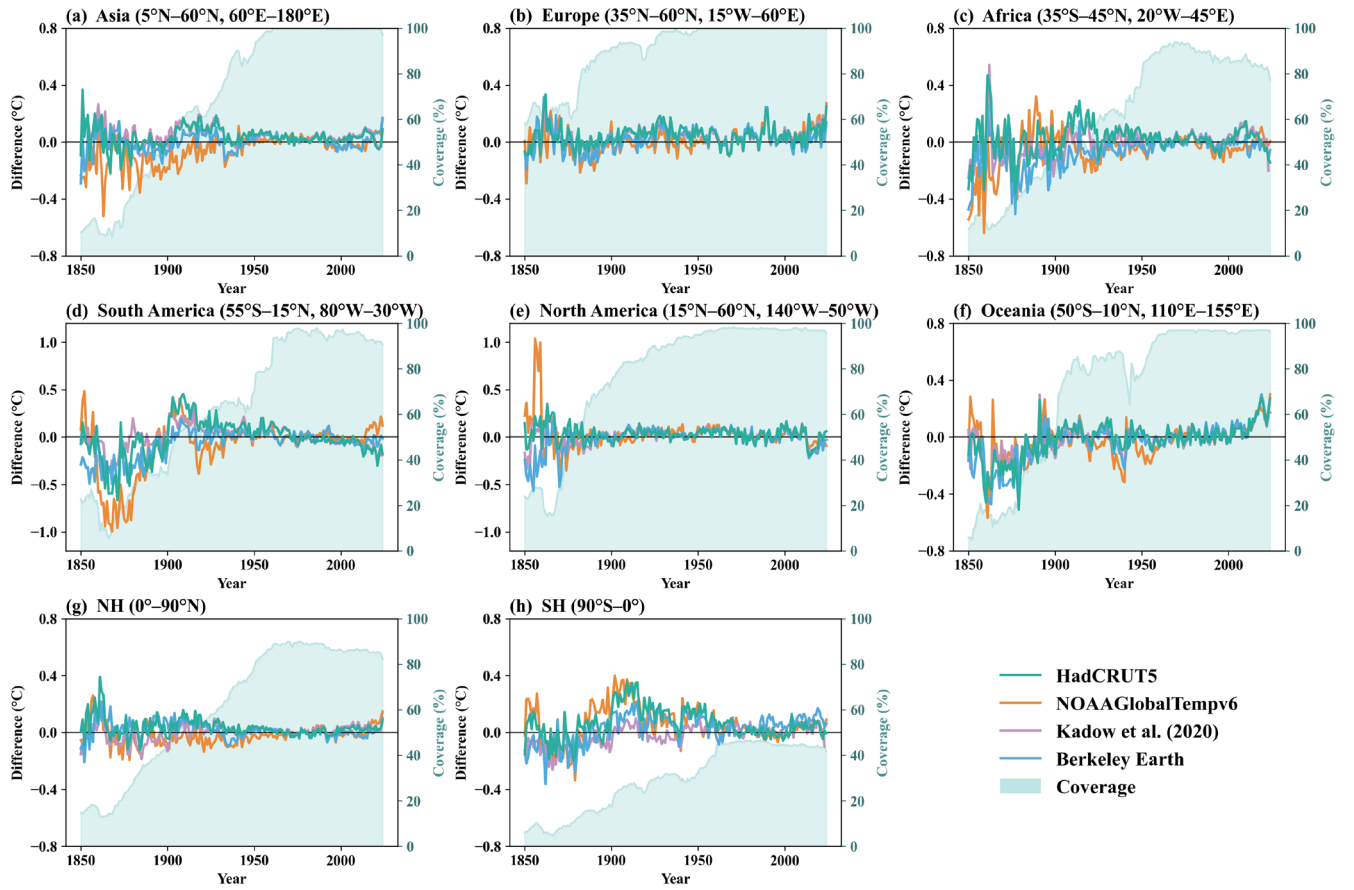


Figure 9: Annual mean surface temperature anomaly differences relative to C-AIRST_M from 1850 to 2024 in different land regions. (a–h) Asia, Europe, Africa, South America, North America, Oceania, the Northern Hemisphere (NH), and the Southern Hemisphere (SH). Coverage indicates the unreconstructed data availability in each region.

Line 559–572 in track change:

One of the key focuses of this study is the reconstruction of global land surface data (Fig. 9, Table 3). To evaluate the reconstruction performance, we also present the effective data coverage of the original datasets across different regions (Coverage in Fig. 9). In addition, Table 3 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.”

Fig. 9 shows that across different land regions as data coverage increases, the differences between the two reconstruction products C-AIRST_M and other datasets gradually decrease to within 0.1 °C as observational coverage exceeds approximately 50%. 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-H20CR-AI is overall about 0.2 °C higher than Merge-HCMIP6-AI C-AIRST_M exhibits differences of varying magnitudes relative to other datasets, ranging from 0.4 to 1.0 °C (Figs. 9c, 9d, 9h), which also leads to lower long-term trends in these regions for the former in Table 3.

Table 3: Temperature trends and 95% confidence intervals (CI) in different regions from 1850 to 2024 (°C per decade).

Region/Dataset	C-AIRST _M	HadCRUT5	NOAA GlobalTempv6	Berkeley Earth	Kadow et al. (2020)
Asia	0.088±0.010	0.088±0.009	0.101±0.010	0.090±0.010	0.084±0.010
Europe	0.090±0.014	0.092±0.014	0.097±0.010	0.094±0.013	0.092±0.014
Africa	0.072±0.009	0.073±0.008	0.082±0.008	0.082±0.008	0.079±0.008
South America	0.061±0.010	0.067±0.007	0.090±0.009	0.079±0.007	0.058±0.008
North America	0.079±0.012	0.077±0.012	0.073±0.013	0.087±0.011	0.079±0.012
Oceania	0.039±0.008	0.054±0.008	0.050±0.009	0.058±0.008	0.050±0.008
NH	0.095±0.009	0.091±0.009	0.096±0.010	0.091±0.009	0.097±0.009
SH	0.060±0.007	0.059±0.006	0.059±0.006	0.070±0.007	0.068±0.006

Line 575–597 in track change:

Furthermore, during 1860–1870, when effective coverage in Africa, South America, and Oceania (Figs. 9c, 9d, 9f) drops below a critical threshold (~20%), and considering the strong spatial sampling heterogeneity and lack of inland temperature information in these continents (Wei et al., 2025), the C-AIRST_M shows varying degrees of deviation from NOAA GlobalTempv6 and HadCRUT5. The most pronounced discrepancy occurs in South America (Fig. 9d), where values are approximately 0.8 °C higher than NOAA GlobalTempv6 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 3), the long-term trends of C-AIRST_M across different regions are broadly consistent with those of HadCRUT5. However, C-AIRST_M exhibits a weaker warming trend than other datasets over Oceania. 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 3), 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 3, the Merge-H 20CR-AI and Merge-H CMIP6-AI C-AIRST_M indicates the strongest warming over Europe during 1850–2024, with trend of 0.090 ± 0.014 °C/decade and 0.099 ± 0.014 °C/decade, respectively, while Oceania shows the weakest warming, at 0.039 ± 0.008 °C/decade and 0.038 ± 0.008 °C/decade. The warming rate in the SH is significantly lower than that in the NH, 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 early 20th century, the AI-based reconstruction results show improved consistency and stability across these regions.

Table 4: Validation between Antarctic station observations and monthly temperature anomalies from AI-reconstructed grid cells under two reconstruction schemes (r: correlation coefficient; RMSE: root mean square error, °C)

Station	Latitude	Longitude	Elevation(m)	Coverage	Merge-H 20CR-AI		C-AIRST _M	
					r	RMSE	r	RMSE
Dome A	−80.4	77.4	4084	2006–2019	0.52	1.28	0.63	1.18
Drescher	−79.2	−19.0	27	1993–2003	0.60	0.72	0.56	0.77
Ferrel	−77.9	170.8	45	2007–2023	0.84	0.65	0.88	0.56
G3	−70.9	69.9	84	2002–2020	0.49	0.79	0.40	0.84
GC41	−71.6	111.3	2763	1984–2005	0.80	0.57	0.96	0.44
General Belgrano	−78.0	−38.8	32	1956–1978	0.70	0.76	0.69	0.77
GF08	−68.5	102.1	2125	1987–2007	0.87	0.46	0.87	0.46
LG10	−71.3	59.2	2619	1993–2005	0.73	0.43	0.65	0.44
LG35	−76.0	65.0	2345	1994–2007	0.70	0.51	0.62	0.55
LG59	−73.5	76.9	2565	1994–2003	0.63	0.45	0.48	0.51
Mount Siple	−73.2	−127.1	230	1993–2006	0.84	0.34	0.84	0.34
Nordenskiöld	−73.1	−13.4	497	1995–2019	0.82	0.31	0.82	0.31
Russkaya	−74.7	−136.9	100	1981–1989	0.73	0.66	0.68	0.70
Thurston Island	−72.5	−97.6	225	2011–2021	0.74	0.56	0.83	0.45
Average (14 stations)					0.72	0.61	0.71	0.59

Line 603–623 in track change:

Among the 81 Antarctic stations used in this study (Table S2 and Fig. S1), the vast majority began recording Antarctic climate data after the 1957/1958 International Geophysical Year. Even so, the effective data coverage in Antarctica remained only around 8% after 1961 (Fig. 10a). During model validation for Antarctic data (Fig. S8), the reconstructed results from both AI models showed high performance, with correlation coefficients with the test dataset reaching approximately 0.9 and RMSE below 1.3 °C after 1961. Due to the scarcity of observational data prior to 1961, validation and assessment of Antarctic data for earlier years are not feasible. Therefore, this study focuses on the systematic validation and analysis of Antarctic reconstructed data from 1961 onward.

To better evaluate the performance of the AI model in reconstructing ST over the Antarctic continent since 1961 and to ensure the validity of the verification, we selected 14 independent Antarctic stations that were not used in the reconstruction and were located outside the spatial coverage of the original observational grid before reconstruction. The monthly observed anomalies at these stations were compared with the corresponding reconstructed anomalies to assess the consistency and statistically test whether the reconstructions reliably reproduced temperature variability (Monaghan et al., 2008; Nicolas et al., 2014; Bromwich et al., 2025a). 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. The results indicate that most stations exhibit high correlation with the reconstructed grid cell temperature series, with RMSE below 0.84 °C except for Dome A (Table 4). Dome A, located at the highest point of the East Antarctic Plateau (Table S2, Fig. S11), is influenced by surface winds controlled by synoptic-scale circulation, and its temperature records can reach extreme lows, making it more challenging for the reconstruction method to capture its variability (Scambos et al., 2018). Overall, the average correlation between the 14 stations and ~~the two reconstruction schemes~~ C-AIRST_M is 0.71, with corresponding mean RMSE value of 0.59 °C. ~~Although there are slight differences among the reconstruction schemes, the CMIP6-AI reconstruction performs slightly better than the 20CR-AI reconstruction, and both schemes show good agreement with the observational data.~~

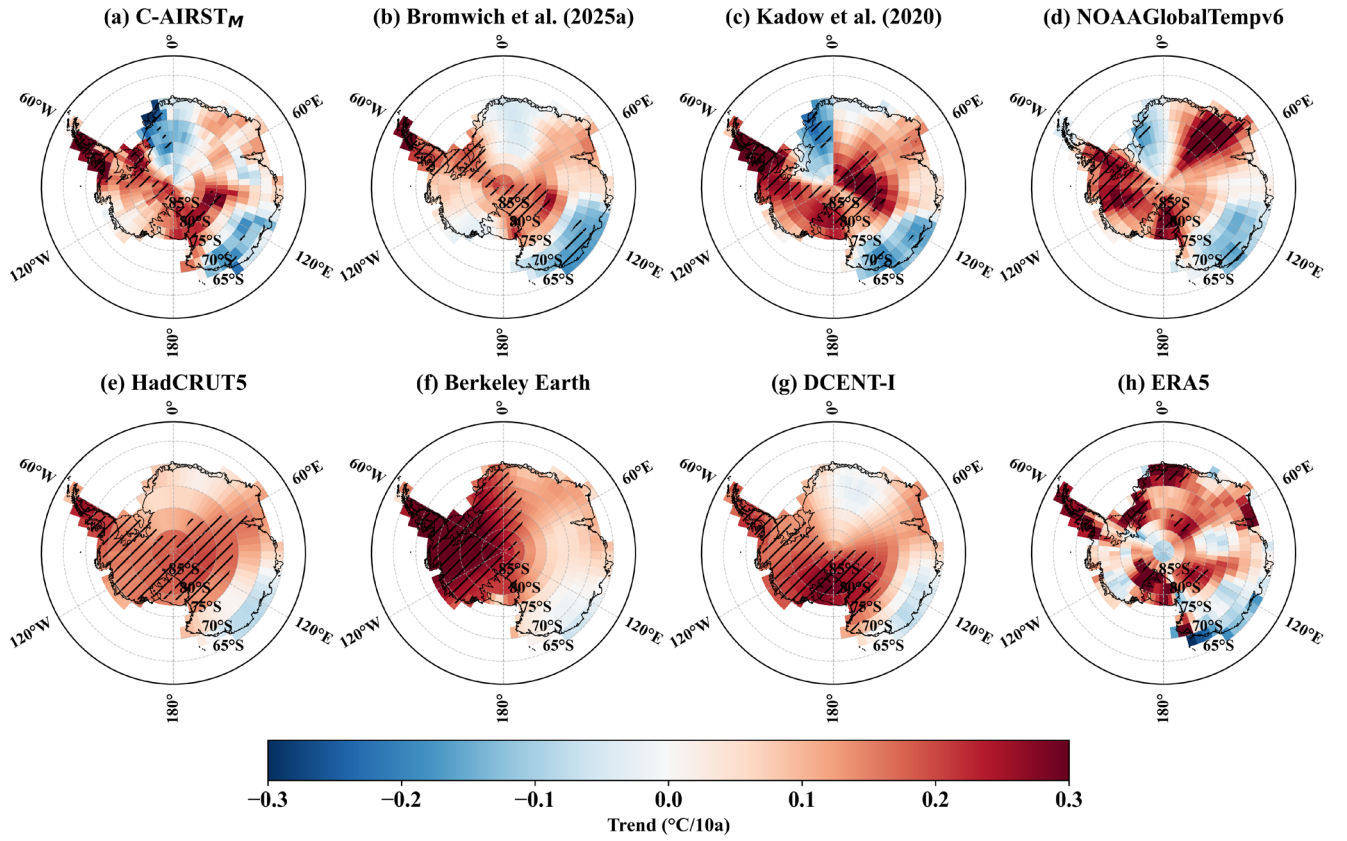


Figure 11: Spatial distribution of Antarctic annual mean surface temperature trends for 1979–2022 (“///” indicates regions where the temperature trend is statistically significant at the 0.05 level).

Line 669–705 in track change:

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). Fig. 11 presents the spatial distribution of ST trends and their statistical significance for 1979–2022 derived from C-AIRST_M, together with the results from other representative datasets.

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, while significant cooling is observed along the coast of Wilkes Land. In particular, the Antarctic Peninsula exhibits pronounced warming, with ~~the AI-reconstructions~~ C-AIRST_M yielding trend of ~~0.301 ± 0.115 °C/decade and 0.276 ± 0.109 °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 C-AIRST_M 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 NOAA GlobalTempv6, 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. 11c) 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~~ C-AIRST_M (Fig. 11a), which, consistent with Bromwich et al. and NOAA GlobalTempv6, 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, Berkeley Earth and DCENT-I 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 NOAA GlobalTempv6. This is also one of the reasons why the temperature trends in Fig. 10d are higher in the former three datasets than in the latter two. It is worth noting that NOAA GlobalTempv6 also shows significant warming in eastern Queen Maud Land, while C-AIRST_M, Bromwich et al. (2025a), 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~~ have the potential to provide a reliable representation of the 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.

Line 771–869 in track change:

8 Conclusions

This study employed an AI model (PConv) to reconstruct global ST fields from ~~two fused~~ observational datasets and performed a corresponding uncertainty estimation. Based on this framework, we developed a monthly global surface temperature dataset spanning 1850–2024, named the China global Artificial Intelligence Reconstructed Surface Temperature_{CMIP6} (C-AIRST_M). The results demonstrate that ~~the AI reconstructions~~ C-AIRST_M is broadly consistent with multiple representative climate datasets. ~~Both Merge-H 20CR-AI and Merge-H CMIP6-AI datasets~~ It exhibits high good temporal and spatial continuity in low- and mid-latitude regions and over Antarctica after 1961, ~~providing a solid foundation suggesting that it can serve as a useful dataset~~ for extending long-term climate records, assessing polar climate change, and supporting climate monitoring, detection, and attribution. The main conclusions of this study are as follows:

~~(1) The image reconstruction method based on Partial Convolution Neural Networks (PConv) can effectively fill in missing global surface air temperature information in the mid-to-low latitudes and capture large-scale spatial temperature anomaly structures such as ENSO. During periods of sparse early observations, both 20CR-AI and CMIP6-AI, which are large-sample-based AI reconstruction models, are able to reconstruct ENSO spatial patterns and reproduce historical ENSO events reasonably well. However, the estimation of ENSO variability shows some biases prior to 1960 and is slightly overestimated after 1979.~~

~~(2) The coverage of observational data is a key factor influencing the reliability of AI reconstruction results. During periods of sparse observations over land, different reconstruction schemes AI reconstructions exhibit certain systematic discrepancies. As observational coverage significantly improves after 1900 the early 20th century, the different reconstructions gradually converge and show better consistency. Overall, AI-based reconstructions can well characterize the long-term variability of global land surface temperature; however, results prior to 1900s in regions with low observational coverage should be interpreted with caution.~~

~~(3) The AI reconstruction results for Antarctica since 1961 are highly reliable. Since 1961, C-AIRST_M has reasonably captured the main characteristics of observed variability over Antarctica. Despite an effective observational coverage of only ~8% in Antarctica after 1961, the models achieve correlation coefficients of around 0.9 in this region, with RMSE generally below 1.3 °C the mean correlation coefficient between C-AIRST_M and independent observation stations is 0.71, with a mean RMSE of 0.59 °C. The AI-reconstructed temperature series are consistent with previous studies in both temporal evolution and spatial distribution, indicating that deep learning models can effectively reconstruct surface air temperature variability in high-latitude regions when sufficient climatic information is available.~~

~~(4) The Merge-H 20CR-AI and Merge-H CMIP6-AI products show relatively unstable performance in the Arctic prior to 1930 and in Antarctica prior to 1960. Although 20CR-AI performs slightly better than CMIP6-AI on the test set overall, the two models exhibit different strengths and weaknesses in data-sparse polar regions. Moreover, 20CR-AI shows larger~~

variability than CMIP6-AI. Therefore, considering overall robustness, it is recommended to preferentially use the more balanced dataset, Merge-H CMIP6-AI.

(5) The reconstructed temperature variations in the early Arctic (before the 1920s) and Antarctica (before the 1960s) should be interpreted with caution, as the estimated uncertainties exceed approximately 0.5°C and 0.8°C , respectively, and potential cold biases may exist during these periods. Therefore, the C-AIRST_M reconstruction for these early periods should be considered as providing exploratory insights rather than definitive estimates.

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 $5^{\circ}\times 2.5^{\circ}$, 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. 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.

(1) The PConv image inpainting framework effectively fills missing grid values and reproduces large-scale ENSO spatial structures under sparse early observational coverage. For pre-1900 periods with limited terrestrial station data, C-AIRST_M reasonably recovers historical El Niño and La Niña signals consistent with HadCRUT5 and Berkeley Earth. C-AIRST_M fills the 1910–1920 data gap in the Niño 3.4 index shown by HadCRUT5 and accurately captures well-documented historical El Niños events. Zonal comparisons indicate inter-dataset discrepancies are controlled within $\pm 0.1^{\circ}\text{C}$ after 1900 across the 60°S – 60°N band; large divergences (0.4 – 1.0°C) persist over Africa and South America before 1900 when observational coverage is below 40%, which requires cautious interpretation of early continental results.

(2) Over 1850–2024, C-AIRST_M yields a GMST trend of $0.064 \pm 0.006^{\circ}\text{C}$ per decade, nearly identical to HadCRUT5 ($0.065 \pm 0.006^{\circ}\text{C}$) and DCENT-I ($0.065 \pm 0.006^{\circ}\text{C}$), slightly lower than Berkeley Earth ($0.070 \pm 0.006^{\circ}\text{C}$) and marginally higher than NOAA GlobalTempv6 ($0.058 \pm 0.006^{\circ}\text{C}$). The 2024 pre-industrial (1850–1900) warming level of C-AIRST_M reaches 1.49°C , falling within the 1.43 – 1.62°C range of all benchmark datasets. C-AIRST_M accurately reproduces the four-stage GMST evolution: stable temperatures pre-1910, rapid warming 1910–1940, mild cooling mid-20th century, and accelerated warming post-1970. A small $\sim 0.1^{\circ}\text{C}$ warm bias of C-AIRST_M during 1930–1940 originates from revised ERI measurement bias corrections in HadSST4.2.0.0. The 95% GMST predictive uncertainty gradually declines from $\sim 0.2^{\circ}\text{C}$ in the 1850s to $\sim 0.05^{\circ}\text{C}$ after 1960, with temporary uncertainty spikes during two World Wars driven by reduced global observational sampling. Spatially, polar regions feature uncertainty exceeding 0.3°C , and Antarctic uncertainty is universally higher than the Arctic due to far sparser historical stations.

(3) Before 1920, Arctic reconstructions from C-AIRST_M show a systematic cold bias of $\sim 0.6^{\circ}\text{C}$ relative to HadCRUT5, NOAA GlobalTempv6 and Berkeley Earth; for 1850–1960, Antarctic fields are $\sim 0.5^{\circ}\text{C}$ colder than NOAA GlobalTempv6, which delivers full Antarctic spatial extrapolation. The 95% uncertainty bands hit $\pm 0.5^{\circ}\text{C}$ for pre-1920 Arctic and $\pm 0.8^{\circ}\text{C}$ for pre-1960 Antarctica, meaning these early polar outputs are exploratory rather than definitive. For continental land areas with sufficient post-1900 observations, inter-dataset differences shrink below 0.1°C . Among all continents, Europe exhibits the strongest 1850–2024 warming trend ($0.090 \pm 0.014^{\circ}\text{C/decade}$), while Oceania has the weakest trend ($0.039 \pm 0.008^{\circ}\text{C}$). Consistent with all reference datasets, Southern Hemisphere warming rates are markedly lower than Northern Hemisphere counterparts, driven by vast ocean heat uptake.

(4) Although the effective observational coverage of Antarctica remains only $\sim 8\%$ after 1961, C-AIRST_M achieves an average correlation coefficient of 0.71 and mean RMSE of 0.59°C against 14 unused independent Antarctic stations. Unlike Kadow et al. (2020), which shows artificial abrupt longitudinal trend shifts near 0° longitude, C-AIRST_M produces smooth, physically reasonable Antarctic spatial patterns consistent with Bromwich et al. (2025a) and NOAA GlobalTempv6. Over 1961–2022, the continental mean Antarctic warming trend of C-AIRST_M is $0.089 \pm 0.057^{\circ}\text{C}$ per decade, statistically significant, matching Bromwich et al. (2025a) and NOAA GlobalTempv6. C-AIRST_M successfully captures the robust warming signal over the Antarctic Peninsula ($0.276 \pm 0.109^{\circ}\text{C/decade}$) and significant cooling along Wilkes Land coast. For 2001–2022, no statistically significant Antarctic warming is detected by C-

AIRST_M, highlighting the dominant role of internal climate variability over external anthropogenic forcing in recent Antarctic change.

A minor limitation is that C-LSAT2.1 land temperature is treated as a deterministic input without ensemble error propagation. According to Li et al. (2026b), the maximum 1 σ land observational uncertainty reaches only 0.065 °C in 1850 and falls below 0.02 °C after the early 20th century, exerting only a minor influence on total predictive uncertainty. As such, the uncertainty values provided by C-AIRST_M represent a lower bound of the full true uncertainty across all land grids. The PConv network adopts planar two-dimensional convolution, which may introduce minor spatial artifacts near polar regions and global map boundaries due to the spherical-to-planar projection. All AI-based reconstructions including C-AIRST_M display non-negligible cold biases over sparsely observed pre-1920 Arctic and pre-1960 Antarctica, originating from insufficient spatial constraints and inherent biases within 20CR/CMIP6 training data. Future work will incorporate spherical convolution architectures, physical-informed loss functions, and paleoclimate proxy records (ice cores, tree rings) to mitigate high-latitude distortion and early-period biases. Additionally, an ensemble version of C-LSAT will be developed to propagate land observational uncertainty and achieve complete full-field uncertainty quantification for global temperature fields.

Overall, this study verifies the great potential of ensemble deep learning for long-term climate field reconstruction. The C-AIRST_M dataset delivers spatially complete monthly global surface temperature anomalies at 5°×2.5° resolution spanning 1850–2024, with uncertainty estimates in gridded temperature products. C-AIRST_M demonstrates reasonable temporal continuity and generally consistent spatial structures across mid/low latitudes and post-1961 Antarctica, providing valuable support for extending long climate records, polar climate diagnosis, and global climate monitoring, detection and attribution research.

Comments from reviewer:

5. Some statements still remain too strong, including 'highly reliable,' 'robust and reliable temporal evolution,' and 'excellent performance.' Given the limited uncertainty quantification and the sparse independent validation in polar regions, these claims should be softened, especially for Antarctica.

Response: We acknowledge that these statements were overly strong, and we have therefore weakened the corresponding expressions throughout the manuscript.

Changes to the manuscript:

Line 209–211 in track change:

The PConv method has ~~demonstrated excellent performance in reconstructing~~ been successfully applied to the reconstruction of globally missing climate data, including ST anomaly fields, wind fields, and surface solar radiation (Kadow et al., 2020; Zhou et al., 2022; Jiao et al., 2023).

Line 701–705 in track change:

This suggests that AI-based climate reconstruction methods ~~provide robust and reliable temporal evolution of Antarctic climate after 1961~~ have the potential to provide a reliable representation of the 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.

Line 792–793 in track change:

~~The AI reconstruction results for Antarctica since 1961 are highly reliable~~ Since 1961, C-AIRST_M has reasonably captured the main characteristics of observed variability over Antarctica.

Comments from reviewer:

Overall, I appreciate the substantial effort the authors have made in revising the manuscript. However, uncertainty quantification is central to the value of this dataset and cannot be addressed only through conceptual discussion. The

authors need to provide a concrete treatment of input-data uncertainty and reconstruction-method uncertainty, or at least a defensible practical uncertainty framework for users. Otherwise, the added value of the product beyond applying an existing AI framework to existing high-quality datasets remains insufficiently demonstrated.

Response: We sincerely thank the Editor and all reviewers for their careful attention to our study and for providing numerous constructive and insightful comments and suggestions. These valuable comments not only helped us identify the shortcomings of the original manuscript but also provided important guidance for further improving the research content, enhancing the quality of the data product, and strengthening the scientific rigor of the study.

We have always attached great importance to this work. Both during the preparation of the manuscript and after receiving the review comments, we have devoted substantial time and effort to systematically investigating the relevant literature, conducting in-depth discussions, performing additional experiments, carrying out visualization-based validations, and further analyzing the results. We have also comprehensively checked and revised both the data product and the manuscript accordingly. We highly value every comment and suggestion provided by the reviewers. In response to all comments raised by the Editor and reviewers during the two rounds of review, we have carefully addressed each point individually and made corresponding revisions in the revised manuscript.

We believe that these revisions have further improved the reliability of the dataset, the completeness of the manuscript, and the credibility of the results, leading to a substantial enhancement of the overall quality of the study. Once again, we sincerely appreciate the rigorous, professional, and constructive comments provided by the Editor and all reviewers. We genuinely hope that the revised manuscript will meet the requirements for publication.

Added References

- Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., and Nahavandi, S.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges, *Inform. Fusion*, 76, 243–297, <https://doi.org/10.1016/j.inffus.2021.05.008>, 2021.
- Kendall, A. and Gal, Y.: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? in: *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, Long Beach, CA, USA, 4–9 December 2017, 30, 5574–5584, 2017.
- Lakshminarayanan, B., Pritzel, A., and Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles, in: *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, Long Beach, CA, USA, 4–9 December 2017, 30, 6402–6413, 2017.
- Li, H., Wang, Y., Yang, K., Huang, G., Xia, X., Chen, Z., Tao, W., Lyu, C., Chen, L., Zhang, M., Hu, K., Gong, H., Fu, D., and Lin W.: Probabilistic reconstruction of global sea surface temperature using generative diffusion models, *arXiv preprint arXiv:2603.16272*, <https://doi.org/10.48550/arXiv.2603.16272>, 2026a.
- Qian, Z., Liu, T., Bathiany, S., Yang, S., Hess, P., Bochow, N., Burmester, C., Gelbrecht, M., Groenke, B., and Boers, N.: Generative deep learning improves reconstruction of global historical climate records, *arXiv preprint arXiv:2602.16515*, <https://doi.org/10.48550/arXiv.2602.16515>, 2026.
- Sandford, C., and Rayner, N.: Addressing the World War 2 Warm Anomaly in HadSST.4.2.0.0, *Int. J. Climatol.*, e70388, <https://doi.org/10.1002/joc.70388>, 2026.
- Valdenegro-Toro, M. and Mori, D. S.: A deeper look into aleatoric and epistemic uncertainty disentanglement, in: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, New Orleans, LA, USA, 19–20 June 2022, 1508–1516, <https://doi.org/10.1109/CVPRW56347.2022.00173>, 2022.

Changes to Supplement Information:

Text S1: Sensitivity experiments and model evaluation.

As shown in Fig. S3, we designed a series of sensitivity experiments using different Sea Surface Temperature (SST) datasets and training datasets to evaluate the performance of the reconstruction framework. Specifically, C-LSAT2.1 (Wei et al., 2025) was merged separately with ERSSTv6 (Huang et al., 2025a; 2025b) and the median field of HadSST4.2.0.0 (Kennedy et al., 2019) following the land–ocean merging method described in Section 2.2 (LSAT and SST merging method). These two merged datasets were referred to as Merge-E and Merge-H, respectively. Furthermore, two independent sets of training samples were used to train the PConv models: 80 ensemble members from The Twentieth Century Reanalysis version 3 (20CRv3) dataset (Slivinski et al., 2019) and 105 ensemble members from Coupled Model Intercomparison Project Phase 6 (CMIP6) historical simulations (Table S1; Eyring et al., 2016). The corresponding trained models were denoted as 20CR-AI and CMIP6-AI, respectively. Under the masks provided by Merge-E and Merge-H, the 20CR-AI and CMIP6-AI models were independently used to reconstruct the corresponding test sets, followed by model evaluation. This resulted in four sensitivity experiment configurations: Merge-E 20CR-AI, Merge-E CMIP6-AI, Merge-H 20CR-AI, and Merge-H CMIP6-AI (Fig. S3).

During model training and evaluation, the leave-one-out strategy described in Section 2.3 (AI training and reconstruction process) was adopted, where one ensemble member was withheld as the test dataset while the remaining members were used for model training. Since these sensitivity experiments were designed to investigate the impacts of different observational configurations and training datasets on reconstruction performance, only one training and reconstruction experiment was conducted for each configuration. The evaluation metrics included Bias, Correlation Coefficient (CC), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE).

The corresponding results are presented in Figs. S4–S9. Overall, the test performance under the Merge-E configuration was generally superior to that under the Merge-H configuration (Fig. S4). This difference is mainly attributed to the fact that the ocean component of Merge-E, ERSSTv6, applies an Artificial Neural Network (ANN)-based spatial extrapolation method (Huang et al., 2025a; 2025b) and incorporates more extensive SST information (Fig. S4). Therefore, Merge-E and Merge-H are not completely comparable due to differences in their SST data coverage and reconstruction methods.

For the Merge-E configuration, the performances of Merge-E 20CR-AI and Merge-E CMIP6-AI were generally comparable, with Merge-E 20CR-AI showing slightly better performance. Specifically, Merge-E 20CR-AI achieved higher reconstruction accuracy for the global area-weighted monthly temperature anomalies than Merge-E CMIP6-AI, with improved CC, RMSE, and MAE values (Fig. S5). It also showed better spatial correlation and spatial RMSE performance (Figs. S7 and S8), whereas the two configurations exhibited comparable results in terms of the averaged temporal correlation and RMSE (Fig. S6). These results indicate that when observational coverage is relatively high, the 20CR-AI model performs slightly better than CMIP6-AI. However, analysis of the final reconstruction products revealed that Merge-E may have learned excessive oceanic characteristics during the early period (before the 20th century). Specifically, the spatial patterns of ocean temperature anomalies may have been propagated into poorly observed terrestrial regions, resulting in warm biases of approximately 0.1–0.2°C over low- and mid-latitude land areas compared with NOAA GlobalTempv6, despite both datasets using the same SST component (ERSSTv6). Therefore, these biases are primarily attributed to the land component, suggesting that the Merge-E configuration may introduce unrealistic land temperature signals during early periods (Fig. S10).

For the Merge-H configuration, Merge-H CMIP6-AI generally performed better than Merge-H 20CR-AI. The former showed improved reconstruction accuracy for global area-weighted monthly temperature anomalies (Fig. S5), as well as better averaged temporal correlation and RMSE metrics (Fig. S6). In contrast, Merge-H 20CR-AI exhibited slightly better performance in terms of averaged spatial correlation and spatial RMSE, although the differences between the two configurations were relatively small. Although both monthly spatial correlation coefficients and grid-point temporal correlation coefficients can be used to evaluate reconstruction performance, this study places greater emphasis on grid-point temporal correlation because it directly assesses the capability of the reconstruction to capture the temporal evolution of climate variability at each grid point.

Considering the overall performance of different sensitivity experiments, Merge-H CMIP6-AI was selected as the final configuration. Specifically, C-LSAT2.1 merged with HadSST4.2.0.0 was used as the observational input, and the PConv model trained with CMIP6 historical simulations was adopted for reconstruction. Reconstruction and uncertainty estimation presented in the main text was performed using this configuration, resulting in the China global Artificial Intelligence Reconstructed Surface Temperature_{CMIP6} (C-AIRST_M).

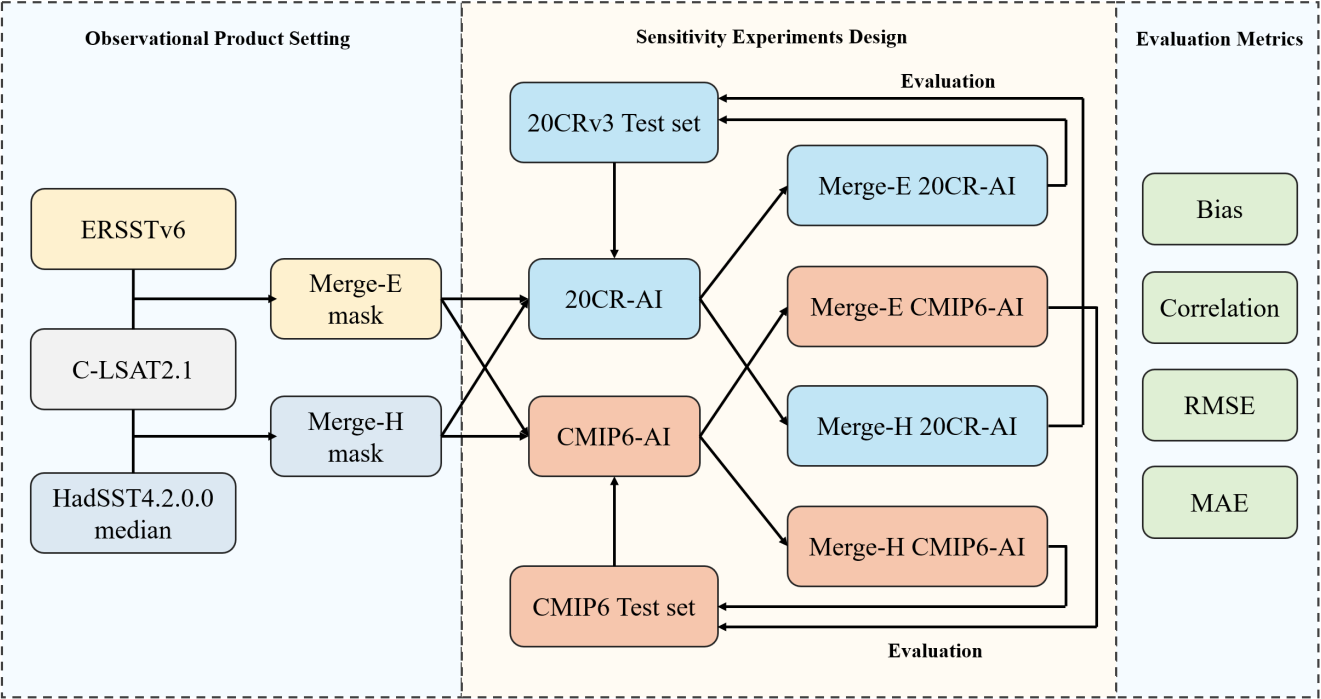


Figure S3: Sensitivity experiments and model evaluation.

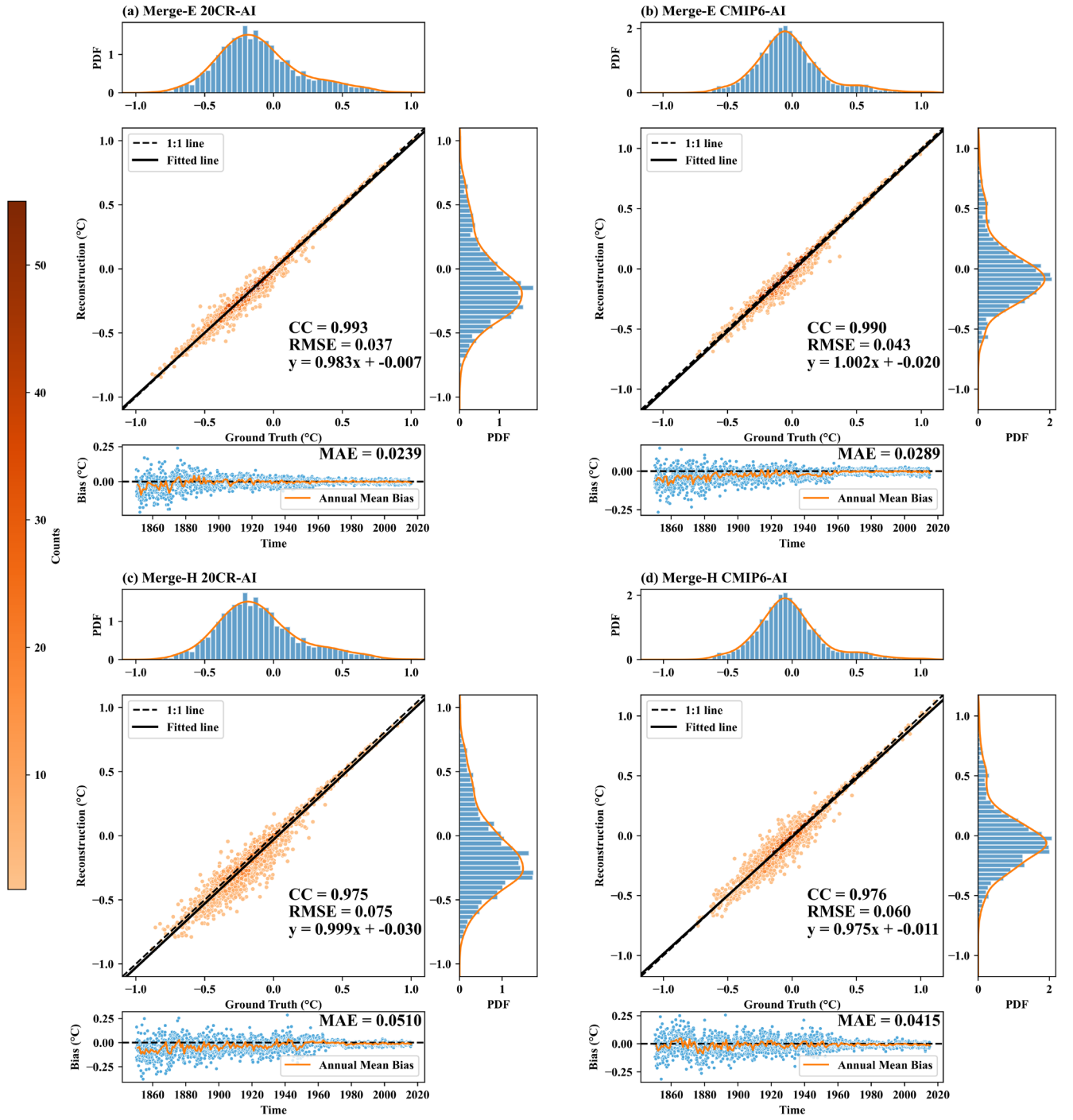


Figure S5: Model evaluation results for different sensitivity experiments. Each dot represents the monthly global area-weighted mean temperature from 1850 to 2014. The reconstruction performance was evaluated using the correlation coefficient (CC), root mean square error (RMSE), and mean absolute error (MAE). The histogram represents the probability density function (PDF) of the distribution. (a) Comparison between the Merge-E 20CR-AI reconstruction and the its original test dataset. (b–d) Same as (a), but for Merge-E CMIP6-AI, Merge-H 20CR-AI, and Merge-H CMIP6-AI, respectively.

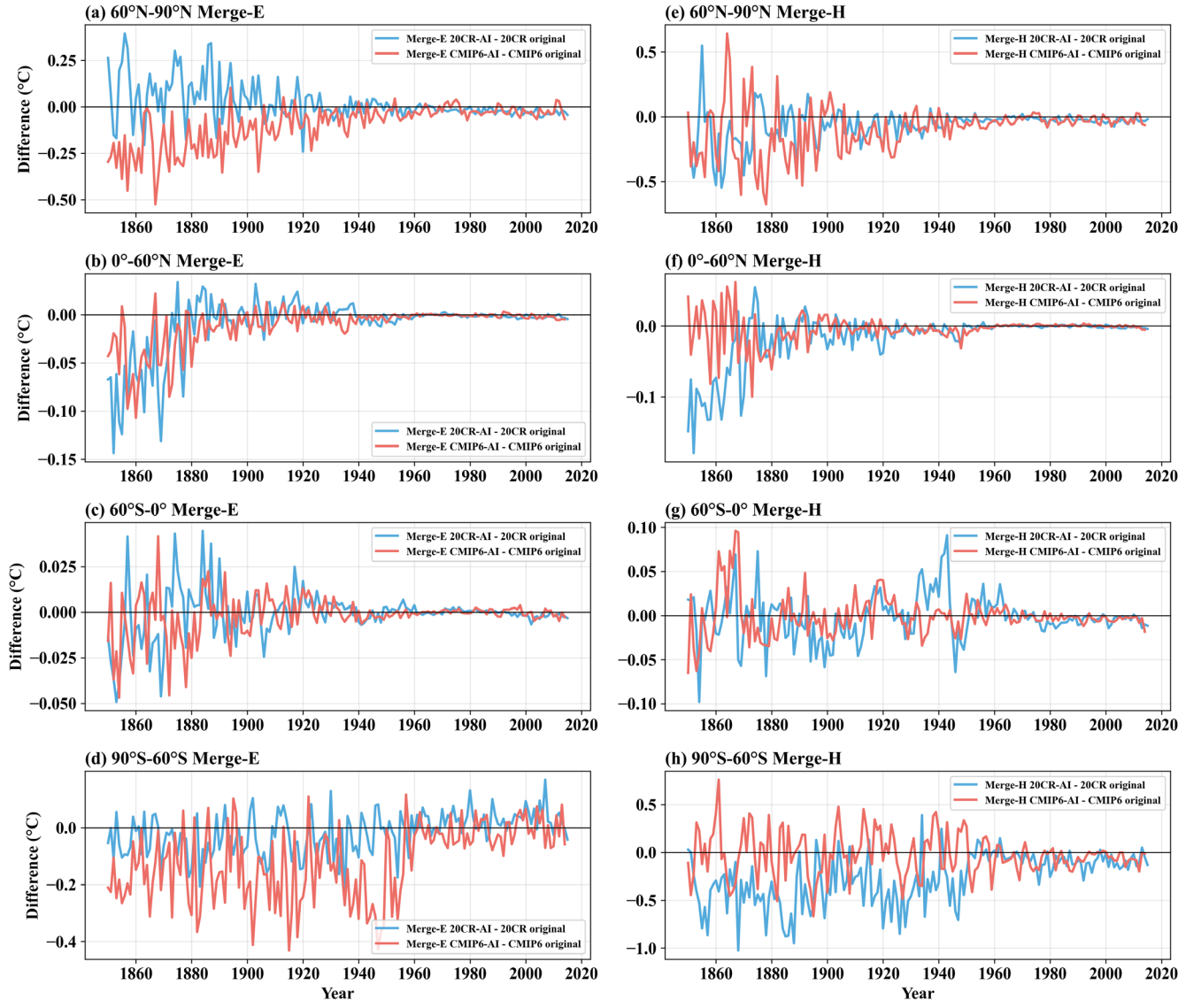


Figure S9: Annual mean difference time series across different latitude bands from 1850 to 2014. (a–d) Differences between the reconstructed test sets and original test sets under the Merge-E mask for the two models; (e–h) same as (a–d), but under the Merge-H mask.

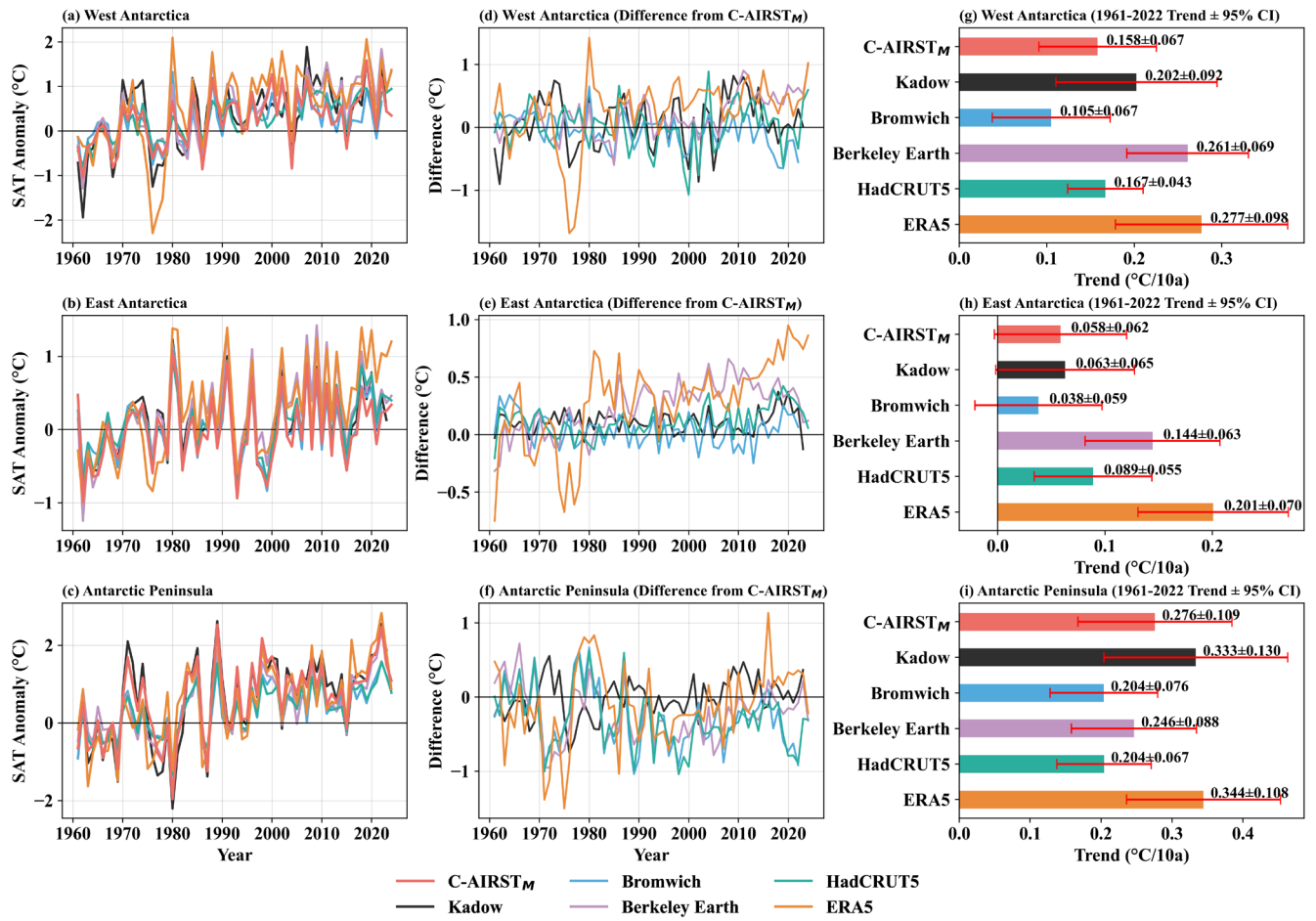


Figure S13: Area mean temperature time series and trends for Antarctic subregions.

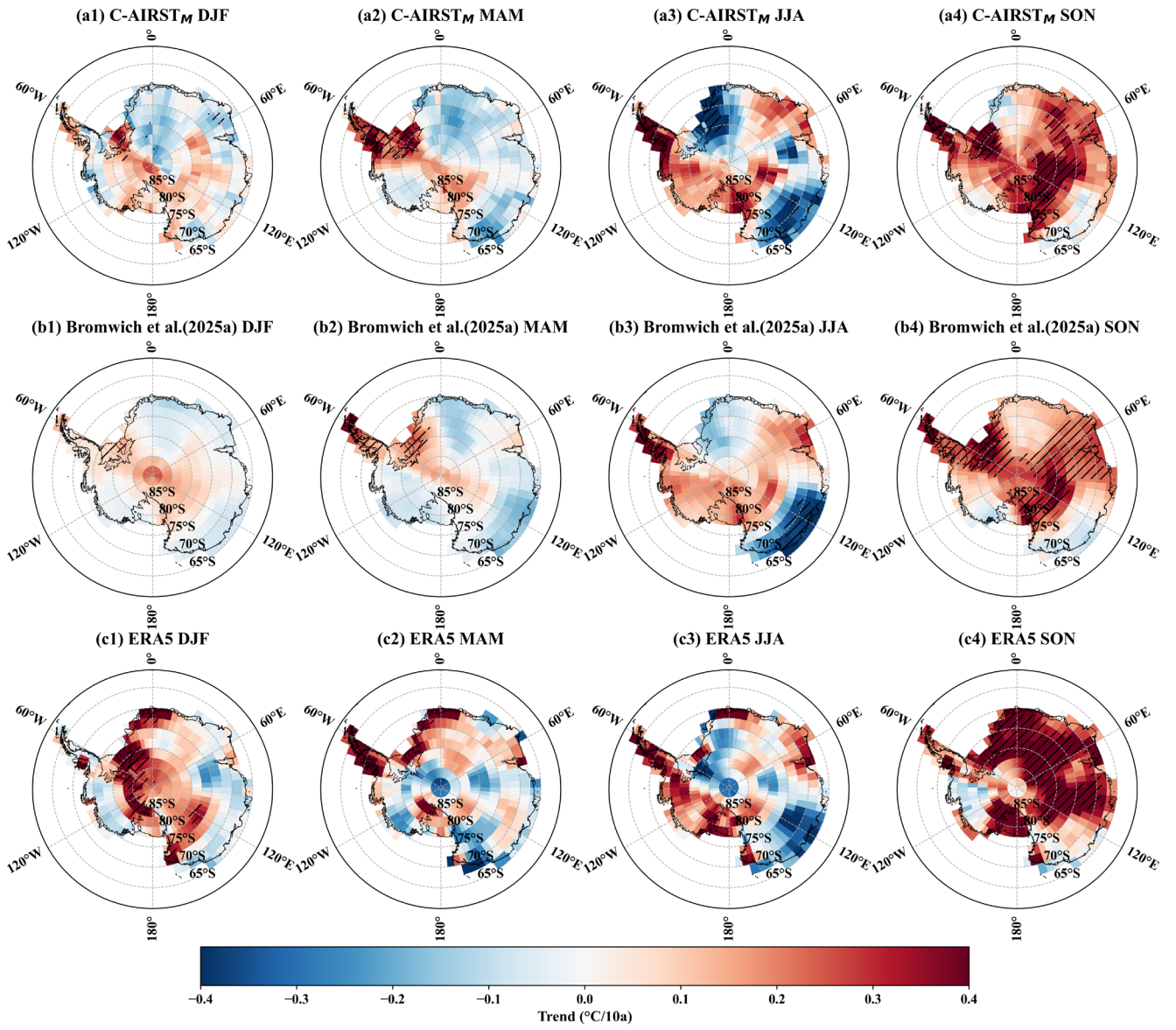


Figure S14: Spatial distribution of seasonal mean temperature trends over Antarctica during 1979–2022.