ESSD-2025-717 review

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

Major Issues

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

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

3. **Insufficient validation and limited spatial comparison of results**
   At the end of Section 3.1 (lines 269–270), the authors state that the results demonstrate "the robustness and reliability of the AI framework for long-term climate reconstruction." However, this conclusion appears not fully supported by the presented analyses. The evaluation in 3.1 is primarily based on globally averaged quantities, which are insufficient to robustly assess spatial performance and regional consistency.

   A more comprehensive validation is required. In particular, the manuscript would benefit from the inclusion of spatially diagnostics, such as global maps of temperature trends (similar to the ones computed for Antartica) and their statistical significance, comparing the newly reconstructed products with existing datasets over different time periods. This would allow a clearer assessment of whether the proposed framework consistently reproduces known spatial patterns and regional features, rather than only matching average global statistics.

4. **Lack of methodological clarity (trend estimation, significance, and data remapping)**
   Several key methodological steps are not described with sufficient detail, which limits reproducibility and makes it difficult to fully assess the robustness of the results.

   A first important point concerns the data remapping procedure. For instance, at line 126, line 151 and then again at lines 167-169 it is stated that observations have been regridded, but the exact method used for this remapping is not described. It should be clearly specified whether nearest-neighbour assignment, bilinear interpolation, area-weighted remapping, or another method is applied. Given the potential impact of this choice on the resulting fields, this step requires a precise methodological description.

In addition, the procedures used for trend estimation and statistical significance testing are insufficiently documented. The manuscript should clearly specify:

- the method used to compute trends (e.g. ordinary least squares, robust regression, Theil-Sen non parametric approach, or other),
- how statistical significance is assessed,
- and whether field significance or any correction for spatial autocorrelation and multiple testing has been considered (see Wilks et al, 2016)

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

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

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

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

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

   Examples:

   - Line 30 and 38: commas are used where a full stop or a proper conjunction would be more appropriate; the sentence structure should be revised for clarity.
   - Line 37: a space is missing after the end of the sentence.
   - Line 53: avoid the use of the Saxon genitive; a more formal scientific construction (e.g., “of the …”) is preferable.

- Line 63: Expressions such as *"fields often hard to capture the overall"* are informal and elliptic; these should be reformulated in a clearer and more explicit way (e.g., *"fields often struggle to capture the overall …"*).
- Line 83: The expression "with as complete a spatial coverage as possible", while grammatically correct, is stylistically awkward; a clearer formulation such as "with the most complete spatial coverage possible" or "with spatial coverage as complete as possible" is recommended.
- Lines 330, 332, and 335: inconsistent terminology is used for trends (e.g. "fastest", "slowest", "lower"). These should be replaced with more scientific and consistent expressions referring to trend magnitude (e.g. "higher/lower trend magnitude" instead of "fastest warming").

Minor issues:

- In general, the Introduction would benefit from a short concluding paragraph outlining the structure of the paper. Providing a brief roadmap of the subsequent sections would improve readability and help the reader navigate the manuscript more effectively.
- *Line 68:* the phrase *"with different datasets as weights"* is unclear and should be reformulated. It is not evident how datasets are used as weights.
- *Line 96:* the concept of temperature anomalies is introduced here for the first time, but the reference period used to compute these anomalies is not specified until later in the manuscript (line 154). It would improve clarity to define anomalies explicitly at first mention, including a clear statement of what they represent and the reference period over which they are calculated. Also specify the climatological reference period when using the term anomalies in figures (e.g. Figure 2).
- *Lines 113 and 115*: the term Historical is used inconsistently, once with capitalisation and quotation marks and once in lowercase without quotes. It is unclear whether these refer to the same concept or to distinct definitions. The authors should clarify this point explicitly. If the same concept is intended, a consistent notation should be used throughout; if different concepts are meant, the distinction should be clearly explained.
- *Figure 1:* there are typographical errors within the figure, where the label *"Trian"* appears instead of *"Train"* in multiple instances. These should be corrected to ensure consistency and avoid confusion.
- *Line 154:* the reference to *(2014)* appears in parentheses without sufficient context, making its meaning unclear. It is not evident whether this refers to a citation, a dataset version, or a specific time period. The authors should clarify its meaning and ensure consistent and explicit referencing throughout the manuscript.
- *Section 2.4 (opening sentences):* the beginning of this section appears to repeat content already described previous Section of the methodology (lines 151–152). It is unclear whether this refers to a distinct procedure or the same one restated. If it is the same procedure, the text should be revised to avoid redundancy and improve coherence between

sections; if instead two different procedures are intended, the distinction should be made explicit and clearly explained.

- *Line 253:* the statement referring to *“two AI models”* is ambiguous. It is not clear whether this refers to two distinct neural network architectures or to the same model architecture trained on different datasets. This point should be clarified explicitly, as it is essential for understanding the experimental design and the interpretation of the results.
- *Figure 4 and the subsequent paragraph:* it is not sufficiently clear how the comparison between the original Merge-H dataset and the AI-reconstructed version is performed. In particular, it is unclear whether the averaged values are computed over the same spatial domain, or whether global means from the original dataset (which contains missing data) are being compared directly with fully reconstructed fields. The latter would not represent a fair comparison due to the differing spatial coverage.

  If, as would be appropriate, a common mask is applied and only grid points present in the original dataset are used for both products, this procedure should be explicitly stated and described in detail. In general, the methodology for ensuring a consistent and fair comparison between incomplete and fully reconstructed fields needs to be clarified more thoroughly to avoid potential biases in the interpretation of the results.
- *Lines 300–301:* two trend values are reported, but it is not clearly indicated which one corresponds to the Merge-H 20CR-AI product and which one refers to the Merge-H CMIP6-AI product. The attribution should be explicitly clarified in the text to avoid ambiguity and ensure correct interpretation of the results.
- Lines 314–318: the current sentence appears to describe features and variability that are common across all seasons. It may therefore be more appropriate to move these general considerations to an earlier section where annual-scale results are discussed. Conversely, the seasonal-specific differences would be better addressed in the context of Figure 5, where the seasonal decomposition is explicitly presented.
- *Line 320:* the statement referring to “larger interannual fluctuations” would benefit from quantitative support. It would be helpful to either explicitly quantify these fluctuations or to clearly refer to an existing figure or table where their magnitude is reported, in order to substantiate the claim that they are the largest among the considered cases.
- *Line 320:* the sentence *“this contrast is mainly attributed to the dominance of the NH”* appears incomplete or at least unclear in its current form. It is not specified what the dominance of the Northern Hemisphere refers to in this context (e.g. global mean signal, trend structure, variability). The sentence should be completed or reformulated to explicitly state the variable or process being influenced.
- *Line 321:* the text refers to “land temperatures”, but it is not clear whether a formal separation between land and sea surface temperatures has been performed throughout the analysis. If such a distinction has been made, this is not clearly described in the methods section nor consistently reflected in the presentation of the results. Moreover, it is not evident whether dedicated maps or diagnostics are available to assess this separation and its implications. As already noted, the inclusion of spatially explicit analyses (e.g. trend maps

or interannual variability fields) would be essential to properly support and quantify these statements, which in their current form appear insufficiently substantiated.

- *Line 346:* the statement that the coverage is “10%” is not clearly defined. It is unclear what the 100% reference corresponds to in this context (e.g. one station per grid cell, full spatial sampling, or complete dataset coverage). The method used to quantify this percentage should be explicitly clarified, as well as whether it refers to station coverage or dataset coverage. Without this clarification, the interpretation of this metric remains ambiguous.
- *Line 353:* the term “Effective Grid Cell” is introduced without a clear definition. It is not explained whether this refers to a different concept from the previously used “grid cells,” nor what criteria make a grid cell “effective.” This should be explicitly defined in the methods section, clarifying its meaning and how it differs (if at all) from the standard grid cell definition used throughout the manuscript.
- *Conclusions (line 286):* the reference to “spatial patterns” is too broad, as the spatial evaluation appears to have been conducted only for Antarctica. The statement should therefore be qualified to avoid overgeneralisation, or the analysis should be extended to other regions if a global assessment of spatial patterns is intended.
- *Final paragraph of Conclusions:* the content appears somewhat repetitive with respect to earlier parts of the Conclusions section. A restructuring would improve clarity and impact, by removing redundancy and making the final message more concise and effective. In particular, key findings and implications should be stated once, in a more streamlined form, avoiding restatement of results already discussed.