Comments on "BHRR v1.0: a two-stage Transformer framework for simultaneous spatial restoration and quantile-function bias correction of climate model temperature fields"

by Young Hoon Song, Hyung Ju Kim, and Eun-Sung Chung

The manuscript presents BHRR v1.0, a two-stage Transformer-based framework for post-processing climate-model temperature fields by sequentially combining Restormer-based spatial restoration with Vision Transformer (ViT)-based quantile-function bias correction. The framework is designed to improve both the spatial fidelity and statistical distribution of daily near-surface temperature fields while preserving climate-change signals in future projections through equidistant cumulative distribution function (CDF) matching. The method is evaluated over Oceania using ACCESS-CM2 historical and future simulations and is benchmarked against the widely used NEX-GDDP-CMIP6 bias-corrected dataset. The manuscript is generally well organized, clearly written, and addresses a topic that falls within the scope of *Geoscientific Model Development*. However, substantial clarification and additional analysis are needed regarding the methodological motivation and novelty, the design and validation of the machine-learning framework, the preservation of future climate-change signals, and the generalizability and scientific interpretation of the results. In its current form, these issues limit confidence in the robustness and broader applicability of the proposed approach. I therefore recommend **major revision**, with specific comments provided below.

Major comments

- **Introduction (Lines 69–75):** This paragraph would benefit from clarification. It presents downscaling and bias correction as the two standard components of operational post-processing, which may overstate how routinely the two approaches are coupled. Although many applications require higher-resolution and more accurate climate products, downscaling and bias correction are distinct post-processing methods with different objectives. Downscaling enhances spatial resolution and regional detail, whereas bias correction reduces systematic model biases relative to a reference dataset. Either approach may be applied independently depending on the application. Please revise the wording to avoid implying that the two methods are necessarily or routinely applied together in operational systems.
- **Section 2.3: methodological clarification:** Several aspects of the proposed machine-learning framework require further clarification: (a) The workflow first interpolates coarse-resolution GCM fields to the target grid before applying Restormer. Because interpolation introduces smoothing and can affect fine-scale structure, please specify the exact interpolation method and discuss how sensitive the restoration performance is to this choice. Would alternatives such as bicubic interpolation, nearest-neighbor interpolation, or conservative remapping materially affect the results or introduce different artifacts? (b) The description of the quantile-correction stage is difficult to follow. Please clarify how the Restormer-derived and reference-based quantile maps are constructed, what quantities the ViT predicts, and how the predicted quantile map is subsequently used in equidistant CDF matching. It would also help to explain, at the workflow level, why a ViT is introduced to learn the quantile-function translation rather than directly constructing empirical model-to-reference quantile mappings from the historical simulations and observations. (c) The restoration stage appears to assume that the reference dataset defines the target high-resolution spatial structure. Please clarify whether the reference dataset must always

have finer resolution than the GCM input and whether the target resolution is determined by the reference dataset. This assumption is important for assessing the applicability of the framework to other variables, regions, and observational products.

- **Novelty and benchmarking:** The proposed workflow combines spatial restoration and bias correction in two sequential deep-learning stages. Because sequential downscaling and bias correction are already well established in climate post-processing, including BCSD and related approaches, the manuscript should more clearly identify what is fundamentally novel beyond the use of specific neural-network architectures. In particular, what additional capability is introduced by the Restormer → ViT-based quantile translation → equidistant CDF-matching workflow compared with existing sequential or integrated methods? The comparison with NEX-GDDP-CMIP6 provides a useful benchmark against a widely used statistical product; however, comparison with one or more representative deep-learning-based downscaling or post-processing methods, if feasible, would better demonstrate the added scientific value of BHRR.
- **Motivation for quantile mapping:** The motivation for quantile mapping could be developed more fully. The manuscript states that quantile mapping is widely used because it corrects the full distribution rather than only the mean error. This is important, but distribution-based methods are also valuable because climate-model biases may vary across the distribution and may evolve under external forcing. Methods such as equidistant quantile mapping are designed to correct biases across the distribution, including the tails, while better preserving projected climate-change signals than a simple mean-bias correction. Please expand the discussion to explain why a distribution-based approach is preferable in this application.
- **Added value of the ViT quantile-function translation:** Conventional empirical quantile mapping can be constructed directly by pairing historical model and reference quantile functions and then applying the resulting transfer relationship to future projections. In BHRR, the historical quantile functions are first computed from the Restormer output and the reference dataset, after which a ViT is trained to translate the Restormer-derived quantile map into a predicted reference-based quantile map. The predicted map is then used for rank-preserving equidistant CDF matching. Although the manuscript explains why distribution-aware correction is preferable to direct pixel-wise temperature regression, it remains unclear what additional capability is provided by learning the quantile-function translation with a ViT rather than directly estimating the empirical model-to-reference mapping. For example, does the ViT improve spatial coherence across neighboring grid cells, regularize noisy empirical quantiles, improve tail estimation, or generalize more effectively to unseen samples? Because the ViT adds substantial complexity, a direct comparison with a non-learned empirical quantile-mapping baseline, or a clearer justification of the learned translation, would help establish its added value. This is one of the central methodological questions raised by the framework. A concise discussion in the conclusion comparing the method with traditional and existing machine-learning approaches, and summarizing its limitations, would also be useful.
- **Generalizability and application to future climates:** A major concern is the generalizability of the proposed framework. The supervised-learning formulation assumes the availability of paired low-resolution GCM simulations and high-resolution observation-based reference data over a sufficiently long historical period. While this assumption is reasonable for near-surface temperature, it may not hold for other variables or regions with limited observations. In addition, the framework is trained using only one GCM (ACCESS-CM2) and one reference dataset (PGFv3), producing a model- and dataset-specific mapping rather than demonstrating a broadly

transferable post-processing framework. Please discuss the extent to which the method could be transferred to other GCMs, variables, regions, and reference products. Furthermore, both Restormer and ViT are trained exclusively on historical simulations and historical reference data and then applied to future projections. This assumes that the historical model–reference relationship and the learned low-to-high-resolution mapping remain valid under changing climate conditions. Future climates, particularly under strong forcing, may contain spatial patterns and distributions outside the historical training range. The ETCCDI and Sen's-slope analyses do not directly demonstrate reliable out-of-distribution generalization. These assumptions and limitations should be discussed explicitly.

- **Evaluation of climate-change signal preservation:** The authors assess whether the framework preserves scenario-dependent climate-change signals by applying it to SSP2-4.5 and SSP5-8.5. This is an appropriate and potentially valuable test; however, the current diagnostics mainly show that SSP5-8.5 produces stronger warming and larger changes in extremes than SSP2-4.5 and that BHRR is broadly comparable with NEX-GDDP-CMIP6. These results demonstrate scenario sensitivity, but they do not directly quantify preservation of the original ACCESS-CM2 climate-change signal. First-order diagnostics should include direct comparisons of future-minus-historical changes in raw ACCESS-CM2 and BHRR, differences between SSP5-8.5 and SSP2-4.5 before and after correction, and explicit measures of changes in signal magnitude and spatial pattern. For a quantile-based method, it would also be valuable to evaluate whether quantile-specific changes are preserved. I also question the statement associated with Fig. 7 that differences are negligible before approximately 2050. This may be reasonable for NEX-GDDP-CMIP6, but the BHRR curves appear to show a systematic separation between SSP2-4.5 and SSP5-8.5 during approximately 2020–2050, particularly for Tas. Please clarify whether the downscaling and bias-correction procedure amplifies the response to external forcing relative to raw ACCESS-CM2 and NEX-GDDP-CMIP6, and provide additional quantitative analysis and discussion.
- **Interpretation of the Results section:** Much of the Results section is limited to describing features shown in the figures, while the broader implications for the proposed method are not sufficiently developed. The authors should discuss more clearly: (1) how well the machine-learning-based downscaling and bias-correction framework reproduces and preserves future scenario-dependent climate signals, and (2) why BHRR differs from NEX-GDDP-CMIP6 and the raw driving GCM, including whether the differences arise from the restoration stage, the quantile-based correction, or other methodological choices. More interpretation is needed to explain what the results imply about the strengths, limitations, and possible sources of error in the framework. Additional examples are provided in the minor comments below.

Minor comments

1. The title uses the phrase "simultaneous spatial restoration and quantile-function bias correction." However, the method consists of two clearly separated stages: Restormer is trained first, followed by ViT, and the two networks are trained independently and applied sequentially. "Sequential" therefore appears more accurate than "simultaneous."
2. The tables are numbered out of sequence. Table 2 (ViT hyperparameters) is introduced before Table 1 (evaluation metrics). Please renumber or reorder the tables according to their first appearance in the text.

3. PSNR and SSIM are not defined before their first use. Please define both abbreviations when they are first introduced. Similarly, Section 3.1.2 introduces Q–Q plots without explanation; a brief description of this diagnostic in Section 2.4 would improve readability.
4. The manuscript describes the optimization settings and notes that model performance was monitored during training, but no training diagnostics are presented. It would be helpful to include representative training and validation results for both the Restormer and ViT stages to demonstrate reliable convergence and the absence of substantial overfitting, for example by showing training and validation loss curves together with the evolution of relevant evaluation metrics.
5. Line 550: There is a stray bold period at the end of Section 3.2.
6. The middle-row B panel in Fig. 6, corresponding to the TNn bias for BHRR, appears unusual and warrants further examination. It shows pronounced checkerboard or striping patterns, abrupt land–ocean discontinuities, broad negative biases over the surrounding ocean, and large positive biases over Australia. Could the authors explain the origin of these features and verify that they are not artifacts of tiling, masking, regridding, or other preprocessing steps? In addition, the discussion of Fig. 6 in Lines 531–549 mainly describes where BHRR and NEX-GDDP-CMIP6 perform relatively well or poorly for individual ETCCDI indices, but the overall scientific message is not sufficiently clear. Please summarize the principal conclusion—for example, whether BHRR provides a consistent improvement across indices and regions or whether its benefits are limited to particular variables, extremes, or locations.
7. Similarly, the discussion of Figs. 7 and 8 is largely descriptive, while the main scientific message is not sufficiently clear. What do these results imply about the performance and added value of the proposed machine-learning framework? In particular, the authors should explain whether the figures demonstrate that BHRR preserves the driving GCM's climate-change signal, improves the spatial and temporal representation of future warming relative to NEX-GDDP-CMIP6, or introduces amplification or other modifications to the scenario-dependent response. As noted in the major comments, the authors should discuss the physical or methodological reasons for the differences between BHRR and NEX-GDDP-CMIP6 rather than only describing the plotted features. Similar clarification and interpretation are needed for Sections 3.3.2 and 3.3.3.
8. Figure 12 appears to repeat much of the information already shown in Fig. 7, with the main addition being the $\pm 1\sigma$ spatial-standard-deviation envelope. Please clarify the distinct scientific purpose of Fig. 12 and what additional insight it provides. The text should also state clearly that the envelope represents spatial heterogeneity among grid cells rather than uncertainty in the projected domain mean.