For data description manuscript [essd-2026-180]

1st round author's response to the reviewer #1

General comment. Given that near-surface humidity and temperature are critical drivers of glacier ablation (surface energy balance), permafrost degradation, and alpine ecosystem evolution, this manuscript addresses the severe scarcity of centennial-scale, high-resolution datasets over the Tibetan Plateau by proposing an innovative hybrid-structure deep learning downscaling approach based on FourCastNet. Nevertheless, the current manuscript still exhibits several critical limitations that must be addressed through a major revision before it can be considered for publication.

Authors' response: We sincerely thank the reviewer for the positive assessment of the scientific significance and potential value of TPHH. We agree that the original submission required greater transparency regarding reproducibility, the interpretation of reconstruction performance, station-data dependence, and the observational limitations of the early historical period.

In response, we have revised the response and manuscript materials to distinguish clearly among reconstruction-performance assessment, breakpoint and trend diagnostics, DEM-sensitivity analysis, and proxy-window evaluation. We have clarified that the 135-station CMA network comprises 125 stations with observations extending to the pre-1979 period and 10 additional stations with observations only from 1979 onward, and that all comparisons with these stations are described as station-based consistency evaluations. We revised the description of FourCastNet as a data-driven statistical downscaling model that learns terrain-related spatial relationships from the TPMFD target fields. The newly added analyses provide reconstruction-performance, breakpoint, trend, DEM-sensitivity, and proxy-window diagnostics. These results support temporal consistency and large-scale plausibility, but do not establish absolute accuracy or a comprehensive uncertainty estimate for the observationally sparse pre-1950 period. Locations: revised manuscript, Sections 2–6; Supplementary Table S1 and Figures S24–S47.

Major comment 1. The paper utilizes FourCastNet for data reconstruction, but the manuscript currently lacks details regarding the hyperparameter tuning process and specific parameter configurations. For an ESSD manuscript, reproducibility is a critical criterion. The authors are strongly encouraged to provide a comprehensive hyperparameter table and specify the training environment details in the main text or appendix.

Authors' response: We thank the reviewer for emphasizing reproducibility. We agree that the original description did not provide sufficient information on the FourCastNet configuration, the production settings, or the computational environment.

We have added Supplementary Table S1, which reports the final production configuration, including the input and output variables, AFNO/FourCastNet backbone, patch size, hidden dimension and layers, FNO blocks, optimizer and beta parameters, base learning rate and scheduler, global batch size, loss function, total epochs, nine-fold cross-validation strategy, spatial resolution, and grid dimensions. The final configuration was selected using validation performance during the 1979–2023 CRU_TS–TPMFD overlap period.

We have also added the training environment, including the operating system, Python, PyTorch, CUDA, GPU, CPU, and system-memory information. The released repository contains the environment specification and configuration files used for the production runs.

The preprocessing, model-training, inference, cross-validation, DEM-sensitivity, breakpoint, trend, and figure-generation scripts have been released at https://github.com/Sawyer000/TPHH. Revision locations: Methods; Code and Data Availability; Supplementary Table S1.

Major comment 2. The authors utilized standard deterministic metrics (R^2, RMSE, MAE, and IOA) to quantify reconstruction errors. However, this approach treats meteorological station observations and the TPMFD target dataset as the absolute "ground truth," neglecting the inherent instrumental errors and the systematic uncertainties embedded within the forcing products themselves. To enhance the physical reliability of TPHH for cryospheric and hydrological applications, the authors should account for these error sources and provide spatiotemporal uncertainty estimates for the reconstructed dataset.

Authors' response: We sincerely thank the reviewer for this important distinction. We agree that neither TPMFD nor station observations are error-free. TPMFD is a merged forcing product that inherits uncertainty from WRF simulations, ERA5, and station-based corrections, while station observations include instrumental, sampling, and representativeness errors.

We have therefore replaced error-free reference terminology with "target dataset" or "reference dataset" for TPMFD and with "station observations" for the CMA records.

We have also revised the scope of the added evaluation. We do not present the new results as a multi-source uncertainty product. Instead, we report reconstruction-performance diagnostics during the TPMFD overlap period, breakpoint and trend diagnostics for historical temporal consistency, and DEM-sensitivity diagnostics.

The corrected Table 3 reports the arithmetic means of R^2, RMSE, MAE, and IOA across the same nine folds contained in the **Table S2** of supplementary material. The corresponding gridded metric NetCDF files have been deposited with the TPHH data release at https://doi.org/10.57760/sciencedb.36169 [Version 3]. **NOTE: We have submitted the new version of the data repository DOI. It has been under review since July 20th and is expected to be formally released within a week.**

Spatial performance was evaluated for 1979–2008 and 1994–2023 using TPMFD as the reference. These maps characterize where TPHH agrees more or less closely with the target fields during the overlap period; they are not interpreted as uncertainty estimates for 1901–1978.

For the historical period, the Pettitt and monthly trend analyses are used to diagnose temporal stability, breakpoint behavior, and large-scale plausibility. They do not provide direct observational error estimates for the pre-1950 period.

Accordingly, the revised text states: "The newly added analyses provide reconstruction-performance, breakpoint, trend, DEM-sensitivity, and proxy-window diagnostics. These results support temporal consistency and large-scale plausibility, but do not establish absolute accuracy or a comprehensive uncertainty estimate for the observationally sparse pre-1950 period."

Revision locations: Section 4.1.1 and corrected Table 3; Section 4.3.2; Discussion Section 6.3; Supplementary Table S2 and Figures S24–S47; data files at https://doi.org/10.57760/sciencedb.36169 [Version 3].

Major comment 3. The description of the different station numbers (87, 125, 135) is very confusing to follow, and their respective roles remain unclear. In particular, it is difficult to assess whether the validation stations are fully independent from the TPMFD target dataset, which itself incorporates CMA station observations. The authors should explicitly clarify:

1) which stations were used in TPMFD construction;

2) which stations were excluded from training;

3) which stations were used for truly independent evaluation, and

4) Whether there is any duplication of data information between the target dataset and the validation procedure.

In addition, Figure 7 appears to use the 125/135 CMA stations rather than the 87 independent stations mentioned earlier in the manuscript. The rationale for this choice is unclear. If independent stations are available, it would seem more appropriate to use them for the primary validation analyses in Figure 7 to avoid potential dependence between TPMFD and the validation dataset.

Authors' response: We sincerely thank the reviewer for identifying the ambiguity in the station descriptions. We agree that the CMA stations cannot be assumed to be independent of TPMFD because TPMFD itself incorporates CMA and NCEI/ISD information.

We have removed the earlier 87-station category from the revised manuscript and response. The revised manuscript describes a total of 135 CMA stations: 125 stations have observations extending to the pre-1979 period, whereas the remaining 10 stations have observations only from 1979 onward.

The whole 135-station CMA network was not ingested directly as point-level training samples by TPHH, instead the model was trained with gridded CRU_TS predictors and gridded TPMFD target fields. Nevertheless, indirect information overlap may occur because CMA observations contributed to TPMFD production.

We have therefore relabeled all evaluations based on these CMA networks as "station-based consistency evaluations" rather than as independent validation. Because a definitive station-level cross-match with the complete TPMFD production-station inventory was not available, we do not classify any of these CMA stations as independent and explicitly acknowledge possible information overlap.

Figure 7 now presents station-based consistency results for the pre-1979 subset of up to 125 stations and for the post-1979 network of up to 135 stations. The results quantify agreement with available station records but are not used to claim independence from the TPMFD information chain.

Figure 1 has been revised to show the 125 stations with observations extending to the pre-1979 period as circles and the 10 additional stations observed only from 1979 onward as triangles; together, these stations constitute the 135-station network available after 1979.

The revised Discussion explicitly states that potential information overlap may lead station comparisons to overestimate independence and that the central and western high-elevation Plateau remain weakly constrained by direct station observations.

This revised terminology separates model cross-validation against TPMFD, station-based consistency evaluation against CMA records, and proxy-based temporal plausibility evaluation against tree-ring chronologies.

Revision locations: Method Section 2.3; Results Section 4.2 and Figure 7; Discussion Section 6.3.

Major comment 4. The reconstructed 1901–1978 historical climate fields rely entirely on a deep learning model trained on the 1979–2023 modern period. This implicitly assumes statistical stationarity in the relationship between the coarse CRU predictors and the fine-scale TPMFD targets across vastly different climatic regimes. However, this critical assumption is not rigorously evaluated. Prior to 1979, station coverage over the Tibetan Plateau was extremely sparse, and early CRU grids heavily rely on large-scale interpolations that lack real mesoscale gradients. Applying a model overly optimized for modern terrain features to these early interpolated fields risks generating artificial spatial features or "homogenized" artifacts. The brief mention in Section 6.3 is insufficient. The authors must:

(1) refrain from treating the reliability of the early-century reconstruction on par with the post-1979 period, and

(2) substantially expand the discussion on how sparse early observations may affect the robustness, spatial variability, and uncertainty of the reconstructed 1901–1978 climate fields.

Authors' response: We sincerely thank the reviewer for highlighting the stationarity assumption and the weak observational constraint on the 1901–1978 reconstruction, particularly before 1950. We agree that the early high-resolution fields should not be interpreted with the same confidence as the post-1979 target-constrained period.

We have revised the manuscript to distinguish the post-1979 overlap period, the 1950–1978 historical period with increasing station availability, and the pre-1950 observationally sparse period. The revised text explains that FourCastNet transfers statistical relationships learned from modern CRU_TS–TPMFD pairs and cannot recover early mesoscale variability that is absent from the coarse CRU_TS predictors.

Because direct pre-1950 error estimation is not possible with the available observations, we did not perform a direct before-versus-after-1950 error comparison. Instead, we added three indirect diagnostics: target-referenced performance maps for two modern 30-year windows, grid-cell Pettitt tests for five 30-year windows, and monthly trend analyses over the same windows. These analyses assess stability and plausibility but do not quantify pre-1950 reconstruction error. These tests evaluate temporal stability and large-scale plausibility rather than absolute early-period accuracy.

For the 1961–1990 window containing the 1978/1979 transition, the fractions of valid grid cells with Pettitt $p < 0.05$ are 4.0% for temperature, 2.1% for specific humidity, and 21.9% for pressure. Of all valid grid cells, 1.43% for temperature, 0.03% for specific humidity, and 8.56% for pressure had a Pettitt-selected breakpoint in 1978 or 1979 with $p < 0.05$. In contrast, the larger fractions for pressure (21.9% significant grid cells and 8.56% with a significant 1978/1979 breakpoint) preclude the same conclusion for pressure, we therefore treat pressure as showing greater transition sensitivity and do not interpret the Pettitt results as evidence of uniform temporal continuity.

The revised Discussion states that these diagnostics support regional-scale temporal consistency but cannot validate fine-scale pre-1950 variability. Early-century local patterns, especially over the western and central high-elevation Plateau, should therefore be interpreted cautiously. Revision locations: Section 4.3.2; Discussion Section 6.3; Supplementary Figures S30–S47.

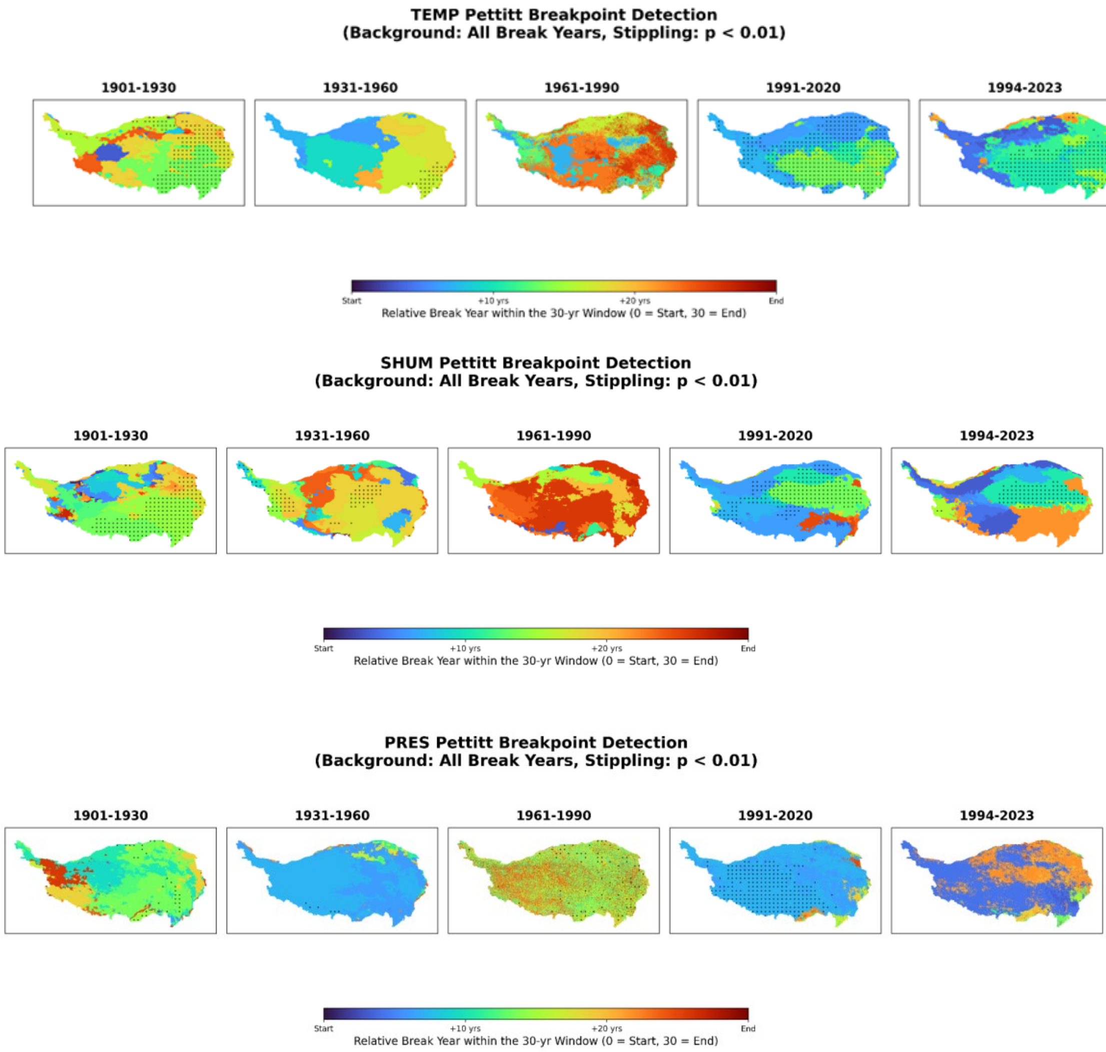


Figure S30−S32. Grid-cell Pettitt breakpoint diagnostics for temperature, specific humidity, and surface pressure across five 30-year windows. The panels report detected breakpoint years and associated p values, the 1961–1990 window is used to assess the spatial prevalence of breakpoints near the 1978/1979 transition.

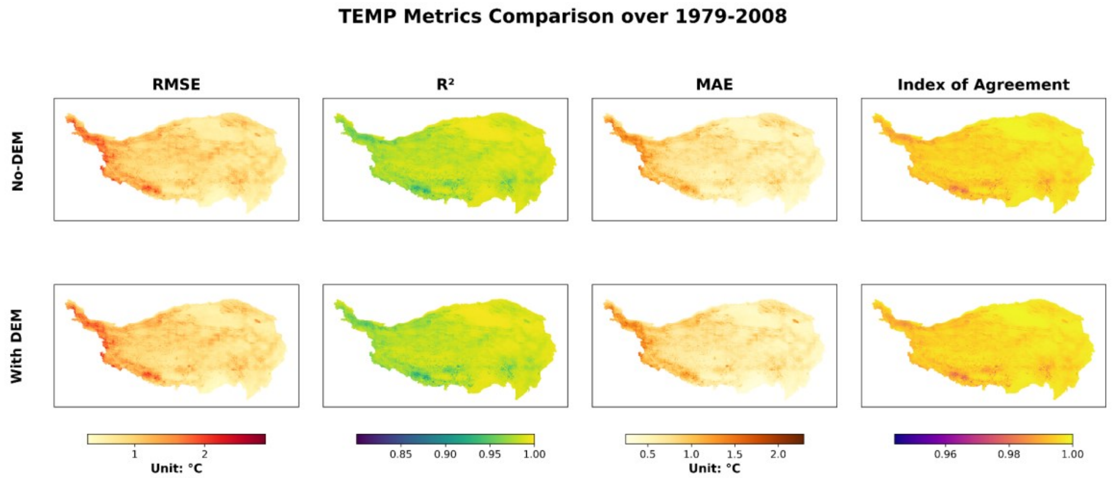
Major comment 4.1 The spatial representativeness of the existing 135 validation stations exhibits significant limitations (Figure 1). Most stations are clustered in the eastern and southern Tibetan Plateau, particularly within relatively low-elevation valleys, while the central high-elevation interior and western Tibetan Plateau remain sparsely monitored or largely unrepresented. As a result, the current validation framework may not adequately assess model performance over the full Tibetan Plateau domain, especially in data-scarce high-altitude regions where climatic and topographic conditions differ substantially from the station-dense areas.

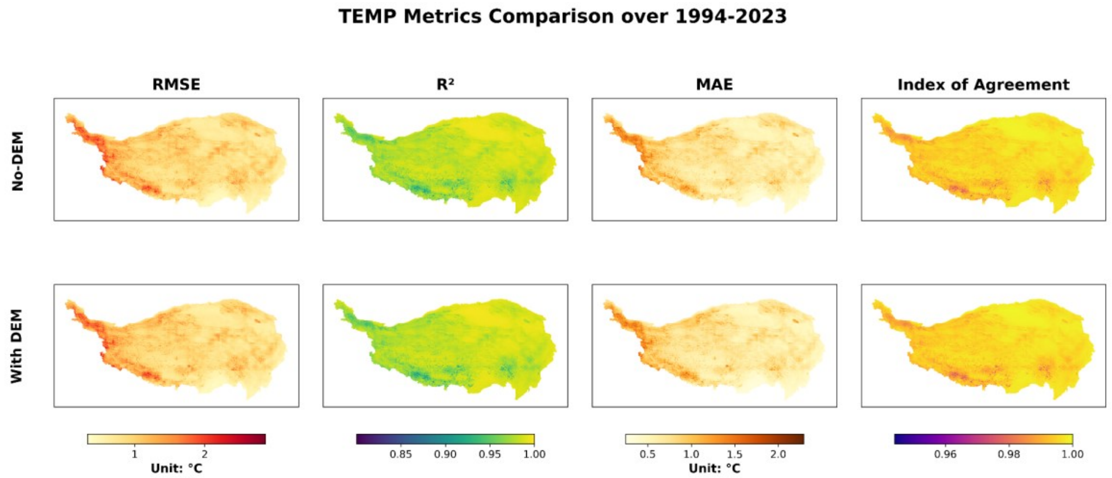
Authors' response: We thank the reviewer for emphasizing the uneven spatial representativeness of the station network. We agree that most long-term stations are concentrated in the eastern and southern Plateau and in accessible valleys, while the western Plateau and the Qiangtang interior remain sparsely observed.

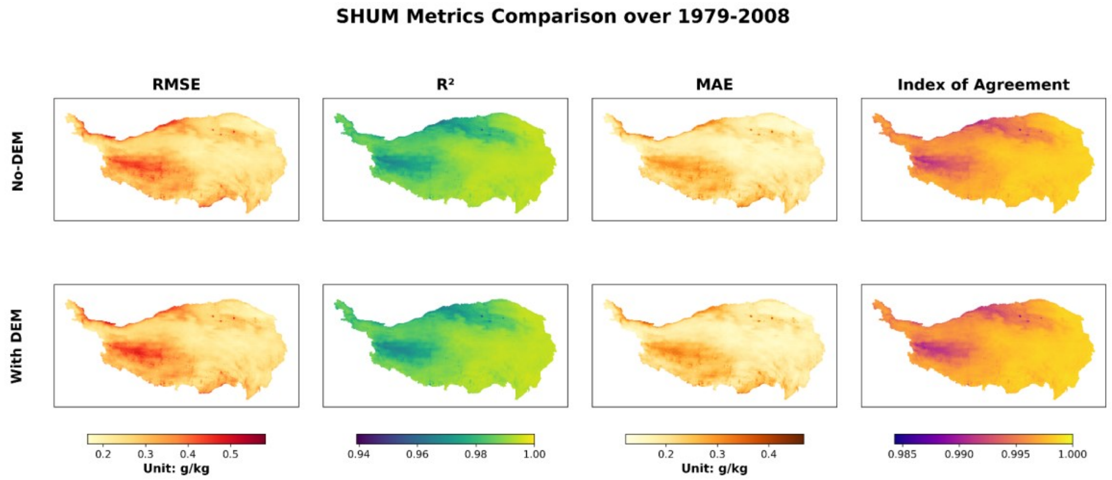
We have revised the validation terminology so that the results for the pre-1979 subset of up to 125 stations and the post-1979 network of up to 135 stations are presented as station-based consistency evaluations. The manuscript now states explicitly that these station comparisons cannot be generalized to the entire Plateau.

To address this concern, we added grid-cell R^2, RMSE, MAE, and IOA maps against TPMFD for 1979–2008 and 1994–2023, together with DEM ablation experiment maps. These are target-referenced reconstruction-performance and sensitivity maps, not direct measurements of reliability in unmonitored regions.

The revised Discussion identifies the western and central high-elevation Plateau as less observationally constrained and advises caution for local applications there. Locations: revised manuscript, Sections 4.1–4.2 and Figures 6–7; Discussion Section 6.3; Supplementary Figures S24–S29.







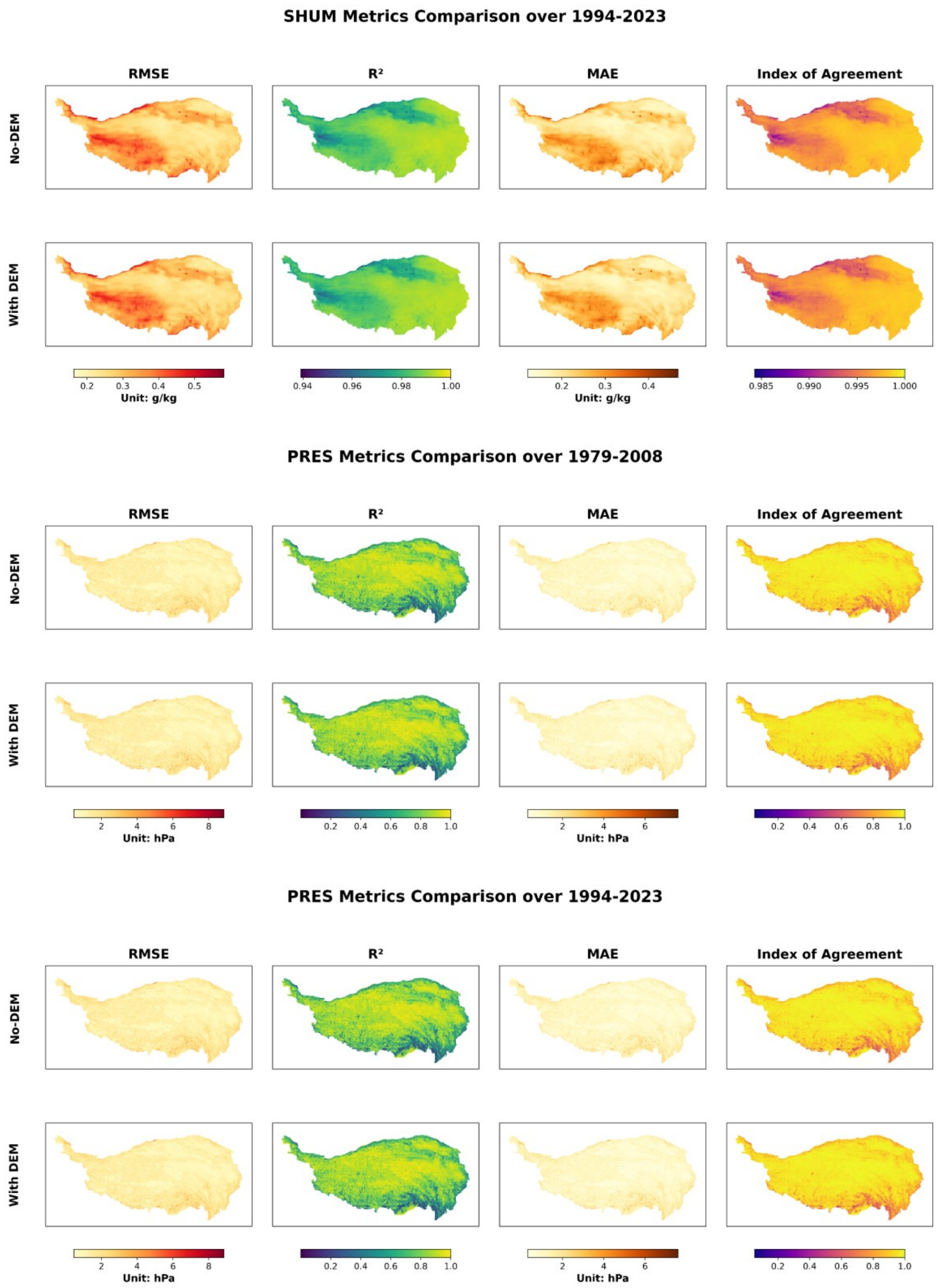


Figure S24-29. Spatial reconstruction-performance and DEM-sensitivity diagnostics for temperature, specific humidity, and surface pressure during 1979–2008 and 1994–2023, evaluated against TPMFD using R^2, RMSE, MAE, and IOA.

Major comment 4.2 For a century-long climate dataset, the representation of long-term climate variability and trends carries the highest scientific value. The authors are recommended to provide spatial distribution maps of the long-term trends for near-surface humidity. Further, these trends should be horizontally compared with existing centennial historical datasets (e.g., 20CRv3, the ERA5-Land back-extension, or older statistical downscaling products) to demonstrate the robustness of TPHH in capturing large-scale climate wetting/drying signals over the past 120 years.

Authors' response: We sincerely thank the reviewer for emphasizing the scientific value of long-term humidity trends. We agree that spatially resolved trend diagnostics are important for assessing the large-scale temporal behavior of TPHH.

We have added monthly specific-humidity trend and p-value maps for five 30-year windows: 1901–1930, 1931–1960, 1961–1990, 1991–2020, and 1994–2023. These maps show the spatial heterogeneity and temporal dependence of wetting and drying tendencies without treating any single window as representative of the full 1901–2023 period.

We evaluated the feasibility of comparisons with the products suggested by the reviewer, but none provided a directly comparable near-surface specific-humidity variable spanning the full 1901–2023 period with harmonized variable definitions, units, and spatial coverage. We therefore did not perform a nominal comparison based on incompatible variables or periods.

The revised response and manuscript consequently present the new humidity maps as TPHH trend diagnostics rather than as cross-product validation. Their interpretation is limited to large-scale plausibility, and the observationally sparse early period is discussed cautiously. Revised locations: Section 4.3.2; Discussion Section 6.3; Supplementary Figures S38–S42.

Major comment 4.3 The rationale for selecting the single-point validation location (91.1°E, 31.5°N) in Figure 11 is unclear. Since climatic variability over the Tibetan Plateau is highly heterogeneous, a single-point example is insufficient to demonstrate temporal continuity or reconstruction robustness over the entire domain. The selected point may also represent a favorable region with relatively smooth terrain and strong model performance. It is necessary to conduct further analysis using multiple representative sites from different climatic and topographical regions on the Tibetan Plateau, so as to more reliably assess the temporal continuity and reliability of the reconstruction results.

Authors' response: We thank the reviewer for pointing out that the original single-point example at 91.1° E, 31.5° N could not represent the climatic and topographic diversity of the Tibetan Plateau.

We have removed the single-point example and replaced it with a multi-site mean consistency diagnostic. For each month, model values were interpolated to the available station locations and averaged using the same valid-station support as the observations. The revised Figure 11 presents temperature and specific-humidity series for the 135-station network where observations are available.

Figure 11 is used only to examine network-mean agreement and temporal behavior during the station-observed period. It is not presented as an independent validation of pre-1950 conditions, because the station network provides little or no direct constraint for much of that period.

We acknowledge that a multi-site mean does not replace a multi-site analysis across every climatic and topographic region suggested by the reviewer. Rather than overstate the evidence, we now combine this network-mean diagnostic with the Plateau-mean series, gridded performance maps, Pettitt breakpoint analysis, and trend diagnostics, while retaining an explicit limitation regarding local early-century variability. Revised locations: Section 4.3.2 and Figure 11; Discussion Section 6.3.

Major comment 4.4 As a manuscript centered around datasets and reconstruction methods, scientific reproducibility is paramount. However, the current "Data availability" section of the manuscript only provides download link for the dataset, without mentioning any public addresses for the core code of the deep learning model. It is strongly recommended that the authors follow the open science policy of ESSD and provide publicly accessible DOI links.

Authors' response: We thank the reviewer for emphasizing that both data and code access are required for reproducibility.

We have released the preprocessing, normalization, FourCastNet training, historical inference, nine-fold cross-validation (**Table S2**), DEM-sensitivity, Pettitt, trend, proxy-window, and figure-generation scripts.

The revised Code and Data Availability statement reads:

1) The TPHH dataset, gridded metric, DEM-sensitivity, Pettitt, and trend NetCDF files are available at https://doi.org/10.57760/sciencedb.36169 [Version 3].
2) The source code is available at https://github.com/Sawyer000/TPHH.
3) The code repository provides the final production configuration, environment specification, and workflow documentation.

Revised locations: Code and Data Availability; Supplementary Table S1 and S2.

Minor comment 1. The major acronyms should be fully expanded upon their first appearance in both the Abstract and the main text. The acronyms like "CRU" and "CMA" are not spelled out at their first mention. Please check the entire text for similar errors.

Authors' response: We thank the reviewer for this suggestion. We have expanded each major acronym at first appearance in both the Abstract and the main text and then used the abbreviation consistently.

Specifically, CRU_TS is introduced as "Climatic Research Unit Time-Series", CMA as "China Meteorological Administration", and TPMFD, TPHH, DEM, AFNO, ViT, GCM, CMIP6, RMSE, MAE, IOA, and LOESS are defined at their first occurrences. Revised locations: Abstract and Introduction; Methods.

Minor comment 2. Line 70, CRU or CRU_TS?

Authors' response: The intended dataset is CRU_TS. We have standardized the name as CRU_TS throughout.

Minor comment 3. To help readers visually evaluate the spatial independence of the validation network, the authors are strongly encouraged to upgrade Figure 1 by:

1) Utilizing distinct colors or geometric shapes to differentiate the historical stations prior to 1979 from the modern ones;
2) Clearly marking the truly independent validation stations;
3) Highlighting the location of the single-point validation grid used in Figure 11 (91.1°E, 31.5°N);

Authors' response: We thank the reviewer for this helpful suggestion. We have revised Figure 1 to distinguish the pre-1979 station network (125 stations with valid records) from the post-1979 network (135 stations with valid records) using different symbols and colors.

Because the CMA records may overlap indirectly with information used in TPMFD production, the figure and text no longer label either network as an independent validation set. Both are identified as station networks used for station-based consistency evaluation.

The original single-point marker has been removed because Figure 11 is now a multi-site mean consistency diagnostic rather than a single-location example.

Revised Figure 1. Geographical setting of the Tibetan Plateau. The map illustrates the complex topography and elevation gradients (left), and the study area's location within the broader Asian context (right). And spatial distribution of the 135 ground-based meteorological stations.

Minor comment 4. Despite repeated discussions throughout the manuscript regarding the profound impacts of terrain effects on downscaling meteorological fields over the rugged Tibetan Plateau, static terrain channels such as Digital Elevation Model (DEM), slope, or aspect are not explicitly included in the model's actual input variables (predictors). Why is this?

Authors' response: We thank the reviewer for asking why static topographic predictors were not included in the final model. We conducted a controlled DEM ablation experiment before selecting the production configuration.

The experiment compared two otherwise identical configurations:

1) CRU_TS meteorological predictors without a static DEM channel;

2) The same predictors with high-resolution DEM as an additional static channel. Both configurations used the same training periods, preprocessing, architecture, loss, and evaluation procedure. Across temperature, specific humidity, and pressure, adding DEM changed the spatially averaged R^2 and IOA by less than approximately 0.001 and produced only very small changes in RMSE and MAE for both 1979–2008 and 1994–2023 (**Figure S24–S29**).

Under the tested configuration, adding an explicit DEM channel did not materially improve predictive performance, indicating that it supplied little additional predictive information beyond the learned CRU_TS–TPMFD mapping. We therefore retained the meteorological-predictor-only configuration. Revised locations: Methods and Discussion; Supplementary Figures S24–S29.

Minor comment 5. The manuscript frequently describes the TPMFD dataset as the "ground truth". In geoscientific and remote sensing literature, "ground truth" is strictly reserved for direct, in-situ physical measurements (e.g., weather station or flux tower observations). Since TPMFD is itself a blended/assimilated forcing product rather than primary direct observations, calling it "ground truth" is scientifically inaccurate. To maintain academic rigor, the authors should replace this term with "reference dataset" or "target dataset" throughout the text.

Authors' response: We thank the reviewer for this important terminology correction. TPMFD is a multi-source gridded forcing product derived from WRF simulations, ERA5, and station-based corrections and should not be treated as error-free.

We have replaced "ground truth" with "target dataset" or "reference dataset" when referring to TPMFD and use "station observations" for the CMA records. Evaluations against TPMFD are described as target-referenced reconstruction-performance assessments. The revised text also acknowledges TPMFD's inherited uncertainties and the potential information overlap between TPMFD and CMA station records. Revised locations: Abstract, Methods, Results, figure captions, and Discussion.

Minor comment 6. In Section 2 ("Comparable datasets"), the authors could explicitly provide the specific temporal coverage/time range for each climate product mentioned in the manuscript.

Authors' response: We have added the temporal coverage of each comparison product in Section 2.

Minor comment 7. Line 353: ERA5 (0.1°)-> ERA5 (0.1°×0.1°)

Authors' response: We have revised the ERA5 resolution to "0.1° × 0.1°".

Minor comment 8. The font of "temp, (b) shum, and (c) pres" in the caption of Figure 11 is inconsistent with the rest of the caption.

Authors' response: We have standardized the typography in the revised Figure 11 caption.

For data description manuscript [essd-2026-180]

1^{st} round author's response to the reviewer #2

General comment. This study addresses a critical challenge in high-altitude climatology by constructing TPHH, a long-term (1901–2023), high-resolution (1/30°) near-surface humidity, temperature, and surface pressure dataset for the Tibetan Plateau. The work is highly significant as it bridges the pre-satellite era data gap, providing much-needed spatial detail for a region characterized by extreme topographic complexity and elevation-dependent warming. The methodological innovation lies in the application of the FourCastNet architecture, which synergistically integrates Vision Transformers and Adaptive Fourier Neural Operators to map coarse-resolution signals from the CRU dataset onto a high-resolution grid governed by terrain constraints. The validation effort is commendable, employing not only a network of 135 independent meteorological stations but also an innovative biological cross-verification using 41 tree-ring proxy groups to ensure temporal fidelity in the early 20th century. This dataset could provide a foundational material for studying long-term hydrological cycles, glacier dynamics, and ecosystem responses to climate change on the Tibetan Plateau. To meet the stringent standards of ESSD, it is essential that the authors thoroughly address the concerns raised below, specifically regarding the physical grounding of the FourCastNet architecture and the reliability of reconstructions in the data-sparse western Tibetan Plateau.

Authors' response: We sincerely thank the reviewer for the positive assessment of the scientific value of TPHH and its potential applications. We also appreciate the reviewer's emphasis on the interpretation of the FourCastNet framework and on the reliability of the observationally sparse western and early-period reconstruction.

We have revised the response and manuscript to describe FourCastNet as a data-driven statistical downscaling model. The model does not impose thermodynamic or dynamical equations through its architecture or loss function; instead, it learns statistical relationships between coarse CRU_TS predictors and the terrain-related spatial patterns represented in the TPMFD target fields.

We have removed language that could imply hard physical enforcement and now distinguish learned terrain-related spatial relationships from explicit physical constraints.

We have also limited the reliability assessment to the evidence actually provided: target-referenced reconstruction-performance diagnostics, Pettitt breakpoint analysis, monthly trend analysis, DEM sensitivity, station-based consistency evaluation, and tree-ring proxy-window diagnostics.

The 135-station CMA network comprises 125 stations with observations extending to the pre-1979 period and 10 additional stations with observations only from 1979 onward; all station comparisons are described as station-based consistency evaluations because indirect information overlap with TPMFD cannot be excluded. Tree-ring analyses are presented as proxy-based support for temporal plausibility rather than validation of local-scale early-century accuracy.

The newly added analyses provide reconstruction-performance, breakpoint, trend, DEM-sensitivity, and proxy-window diagnostics. These results support temporal consistency and large-scale plausibility, but do not establish absolute accuracy or a comprehensive uncertainty estimate for the observationally sparse pre-

1950 period. Revised locations: Sections 2–6; Supplementary Table S1 and Figures S8–S47.

Major comment 1. Quantification of uncertainty in the pre-1950s observational void. The authors utilize CRU TS 4.08 as the primary predictor. It is well-documented that before the 1950s, ground-based observations within the Tibetan Plateau were virtually non-existent, and the CRU signal in this period relies heavily on spatial interpolation from distant stations outside the plateau. While the deep learning model successfully learns fine-scale "terrain textures" from the TPMFD training period (1979–2023), there is a risk that the reconstructed high-resolution features in the early 20th century are merely reflections of static topographic constraints rather than actual climatic fluctuations. The authors must provide a deeper discussion or a sensitivity analysis comparing the model's error characteristics before and after the densification of the observation network (e.g., around 1950).

Authors' response: We sincerely thank the reviewer for emphasizing the pre-1950 observational void. We agree that this is the least observationally constrained part of TPHH and that fine-scale early-century patterns cannot be interpreted on the same basis as post-1979 fields.

We have added a dedicated discussion explaining that FourCastNet can refine the spatial expression of signals present in CRU_TS but cannot recover mesoscale early-period variability that is absent from the coarse predictors. Some fine-scale features may therefore reflect stable statistical spatial modes learned from the modern TPMFD period.

A direct comparison of reconstruction errors before and after 1950 was not possible because suitable pre-1950 observations are unavailable. We therefore added indirect stability and plausibility diagnostics—Pettitt tests and monthly trends for five 30-year windows, plus target-referenced performance maps for two modern windows—and expanded the Discussion to explain this limitation. These analyses do not quantify pre-1950 error. These are stability and plausibility diagnostics rather than direct pre-1950 error estimates.

For the 1961–1990 window containing the 1978/1979 transition, 4.0% of valid temperature grid cells, 2.1% of specific-humidity grid cells, and 21.9% of pressure grid cells have Pettitt $p < 0.05$. Of all valid grid cells, significant Pettitt-selected breakpoints in 1978 or 1979 occurred in 1.43% for temperature, 0.03% for specific humidity, and 8.56% for pressure.

The revised manuscript states that these results support regional-scale temporal consistency, especially for temperature and humidity around the transition, but cannot establish absolute accuracy for the pre-1950 period or validate local fine-scale variability in the western and central high-elevation Plateau. Revised locations: Section 4.3.2; Discussion Section 6.3; Supplementary Figures S30–S47.

Major comment 2. Theoretical justification of physical consistency. The manuscript repeatedly emphasizes that the hybrid-structure deep learning approach is physically-consistent. However, FourCastNet is essentially a data-driven super-resolution model. While it effectively captures "terrain governance," it is unclear whether the model architecture or loss functions explicitly enforce thermodynamic laws (e.g., the Clausius-Clapeyron equation or the relationship between air pressure and hypsometry). The authors should clarify if this consistency is merely a spatial correlation with topography or if the model satisfies fundamental meteorological equations during the downscaling process.

Authors' response: We thank the reviewer for this important conceptual clarification. We agree that FourCastNet is a data-driven model and that its architecture and loss function do not explicitly impose the Clausius–Clapeyron relationship, the hypsometric equation, surface-energy balance, water-vapor conservation, or other meteorological governing equations.

We have revised the terminology throughout the Abstract, Introduction, Methods, Results, Discussion, and Conclusions to describe the method as data-driven statistical downscaling that reproduces terrain-related spatial relationships learned from TPMFD.

The model learns a mapping between coarse-resolution CRU_TS predictors and high-resolution TPMFD target fields during 1979–2023. Consequently, the fine-scale structures in TPHH are interpreted as learned statistical patterns associated with elevation gradients, mountain–valley contrasts, and regional hydroclimatic organization, rather than as fields produced by explicit equation-based constraints.

The DEM ablation experiment provides an additional diagnostic: explicitly adding DEM changes R^2 and IOA by less than 0.001 and yields only very small changes in RMSE and MAE. This result indicates that the explicit DEM channel supplied little additional predictive information beyond the learned CRU_TS–TPMFD mapping, and it does not demonstrate enforcement of physical laws. Revised locations: Abstract and Introduction; Methods Section 3.1; Discussion; Supplementary Figures S24–S29.

Major comment 3. Spatial Bias in Validation and Representative Uncertainty。 As shown in Figure 1 and Figure 7, the 135 validation stations are predominantly clustered in the eastern and southern edges or river valleys of the plateau. The western interior and the high-altitude "No Man's Land" (Qiangtang) are almost entirely unrepresented in the ground-truth network. Given that the model's RMSE is higher in complex terrain, the authors should provide a spatial uncertainty map or an analysis of residuals across elevation gradients to quantify the reliability of the data in these unmonitored high-altitude regions.

Authors' response: We sincerely thank the reviewer for emphasizing the spatial bias of the station network. We agree that the eastern and southern concentration of stations limits direct evaluation over the western interior and Qiangtang.

We have added spatial maps of R^2, RMSE, MAE, and IOA for temperature, specific humidity, and pressure during 1979–2008 and 1994–2023, together with DEM/no-DEM comparisons. These maps provide spatially complete target-referenced reconstruction-performance and sensitivity diagnostics during the TPMFD overlap period.

We have also revised Figure 1 and the station terminology. Figure 1 now shows the 125 stations with observations extending to the pre-1979 period as circles and the 10 additional stations observed only from 1979 onward as triangles; together they form the 135-station network available after 1979.

The revised Discussion identifies the western and central high-elevation Plateau as weakly constrained by direct observations. The metric maps are not presented as direct uncertainty estimates for unmonitored areas, and users are advised to interpret local early-period variability cautiously. Revised locations: Sections 4.1–4.2 and Figures 1, 6, and 7; Discussion Section 6.3; Supplementary Figures S24–S29.

Major comment 4. Phenological logic of tree-ring variable selection (equation 5). In the ternary regression model for tree-ring validation, the authors selected April-May-June (AMJ) humidity as a key factor. The choice of this specific seasonal window requires stronger phenological justification. In many parts of the Tibetan Plateau, tree growth may be more sensitive to JJA (Summer) humidity or have lagged responses to winter moisture. The authors should explain why the AMJ window was selected as the primary moisture constraint across all 41 proxy groups.

Authors' response: We thank the reviewer for requesting a clearer phenological basis for the AMJ humidity predictor. We agree that a single seasonal window should not be assumed to be optimal at every chronology.

We have expanded the Methods to explain that AMJ represents early-growing-season moisture conditions relevant to cambial reactivation, leaf expansion, root water uptake, and the initiation of xylem formation in many cold and seasonally dry tree-line environments.

We compared AMJ, MJJ, and JJA windows across the proxy groups shown in Supplementary Figures S9–S23. Group 3 is presented separately in Figure 12 for the selected window. The full group-by-group results are provided in Supplementary Figures S9–S13 (AMJ), S14–S18 (MJJ), and S19–S23 (JJA).

The revised text describes AMJ as a representative early-growing-season moisture indicator rather than a universally optimal window. The tree-ring analysis is presented as proxy-based evidence of temporal plausibility, not as proof of local climate accuracy. For the representative Group 3 example in Figure 12, the revised panel reports Corr. = 0.844 and $p < 0.001$. Revised locations: Methods Section 3.5 and Results Section 4.4; Figure 12; Supplementary Figures S8–S23.

Minor comment 1. Line 22 & Supplement (Fig S1, S2 titles): There is an inconsistency in the dataset's acronym. The supplement uses "THPP" while the main text uses "TPHH". Please standardize to TPHH throughout.

Authors' response: We have corrected "THPP" to "TPHH" in Supplementary Figures S1–S3 and checked the acronym throughout the manuscript and supplementary captions. Location: Supplementary Figures S1–S3 and captions.

Minor comment 2. Line 34: The text mentions "87 independent ground-based meteorological stations", but Line 56 and Figure 1 refer to "135 stations". Please verify the exact number used for historical validation (1901–1978).

Authors' response: We have removed the earlier 87-station category. The revised manuscript reports a total of 135 CMA stations, of which 125 have observations extending to the pre-1979 period and the remaining 10 have observations only from 1979 onward. Because indirect overlap with TPMFD cannot be excluded, neither network is labeled as an independent validation set. Revised locations: Section 2.3 and Figures 1 and 7.

Minor comment 3. Line 225 (Equation 1): The formatting for the R2 formula is slightly garbled in the PDF, with overlapping characters in the denominator. Please re-typeset using standard LaTeX.

Authors' response: We have re-typeset Equation 1 using standard mathematical notation and checked the denominator, summation limits, and superscripts in the revised PDF.

Minor comment 4. Line 290: The text states, "As demonstrated in Table 1...", but the performance metrics are actually presented in Table 3. This is a citation error.

Authors' response: We have corrected the citation from "Table 1" to "Table 3".

Minor comment 5. Lines 505, 520, 530 of References: Several citations are listed for 2025 or 2026 (e.g., Jiang et al., 2025; O'Neill et al., 2025). Please indicate if these are preprints or currently in press.

Authors' response: We have checked the cited 2025/2026 references and identified each as a published article, preprint, or in-press item as appropriate, with DOI and bibliographic details added where available.

Minor comment 6. Figure 6 (RMSE Heatmap): The color scale for specific humidity is quite narrow (max 0.40 g/kg). It is difficult to discern spatial variance in the plateau interior. Consider using a more contrast-enhancing color ramp for the humidity panel.

Authors' response: We have revised the specific-humidity RMSE color scale in Figure 6 to improve contrast over the Plateau interior while preserving the full data range.

Minor comment 7. Figure 12: The Pearson correlation coefficient (0.82) is unexpected high. Please include the significance level (p-value) to strengthen the statistical argument.

Authors' response: We have revised Figure 12 and its caption to report Corr. = 0.844 and $p < 0.001$ for the representative Group 3 experiment.

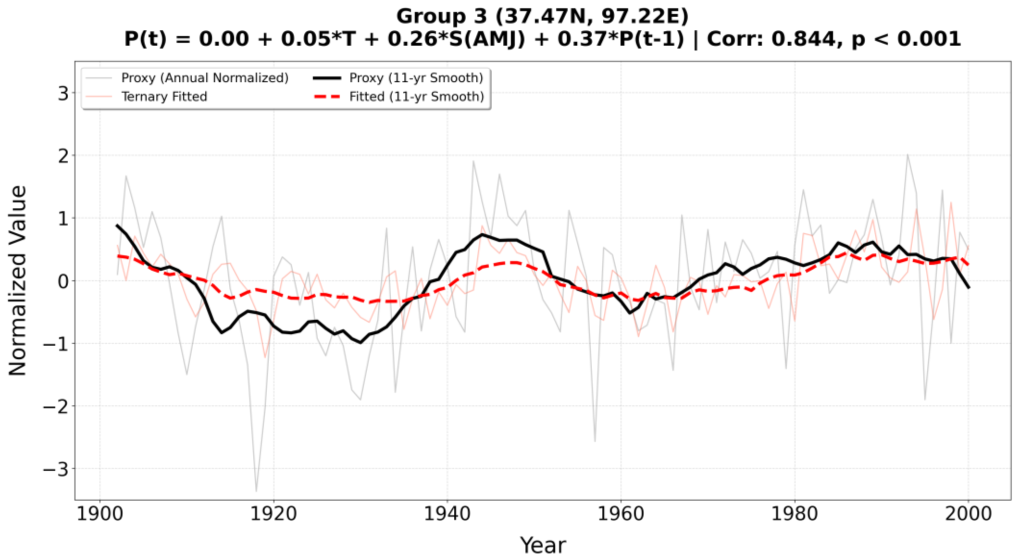


Figure 12. Temporal consistency between the actual tree-ring proxy and the TPHH-based ternary fitted series (1901–2000). The fitted series is generated via a standardized ternary regression model integrating TPHH JJA temperature (T), AMJ specific humidity (S), and lag -1 biological memory of the tree-ring proxy (P). Bold lines denote the 11-year moving average, highlighting the low-frequency correlation. See Supplementary **Figures S9–S23** for the complete validation across proxy groups.