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.

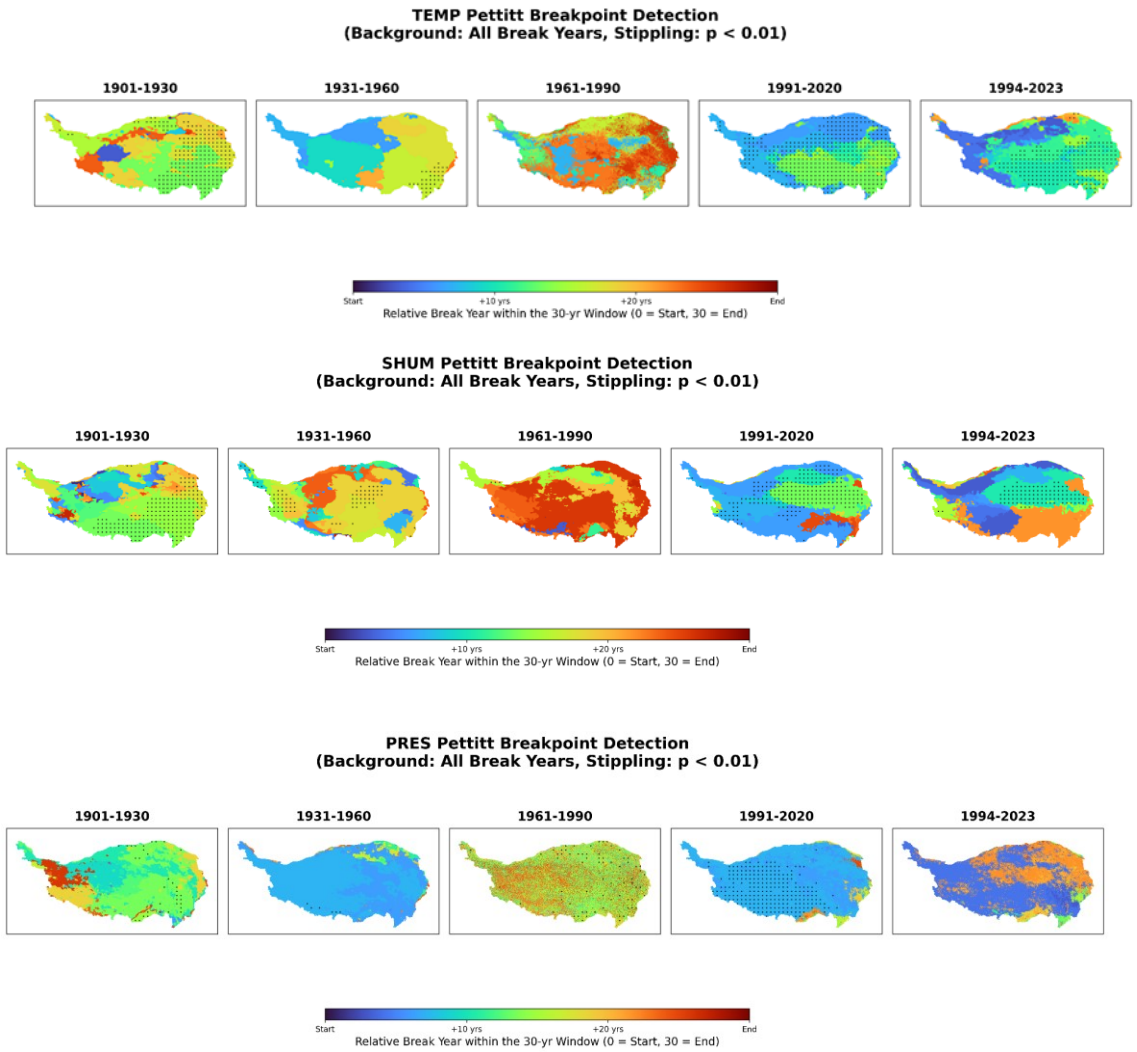


Figure S30–S32. Grid-cell Pettitt breakpoint diagnostics for temperature, specific humidity, and surface pressure across five 30-year windows. The panels report detected breakpoint years and associated p values, the 1961–1990 window is used to assess the spatial prevalence of breakpoints near the 1978/1979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

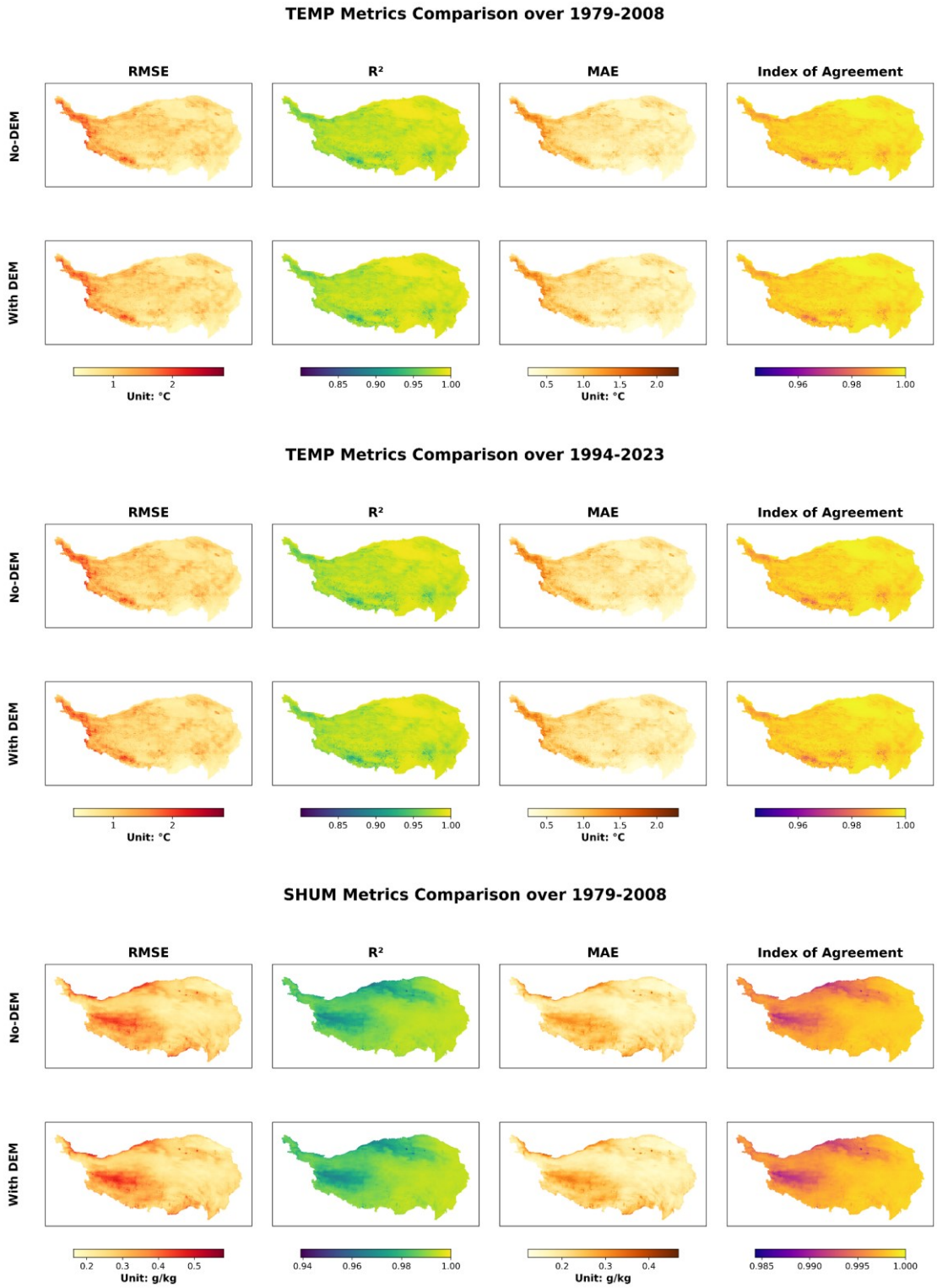
Major comment 4.1 The spatial representativeness of the existing 135 validation stations exhibits significant limitations (Figure 1). Most stations are clustered in the eastern and southern Tibetan Plateau, particularly within relatively low-elevation valleys, while the central high-elevation interior 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.



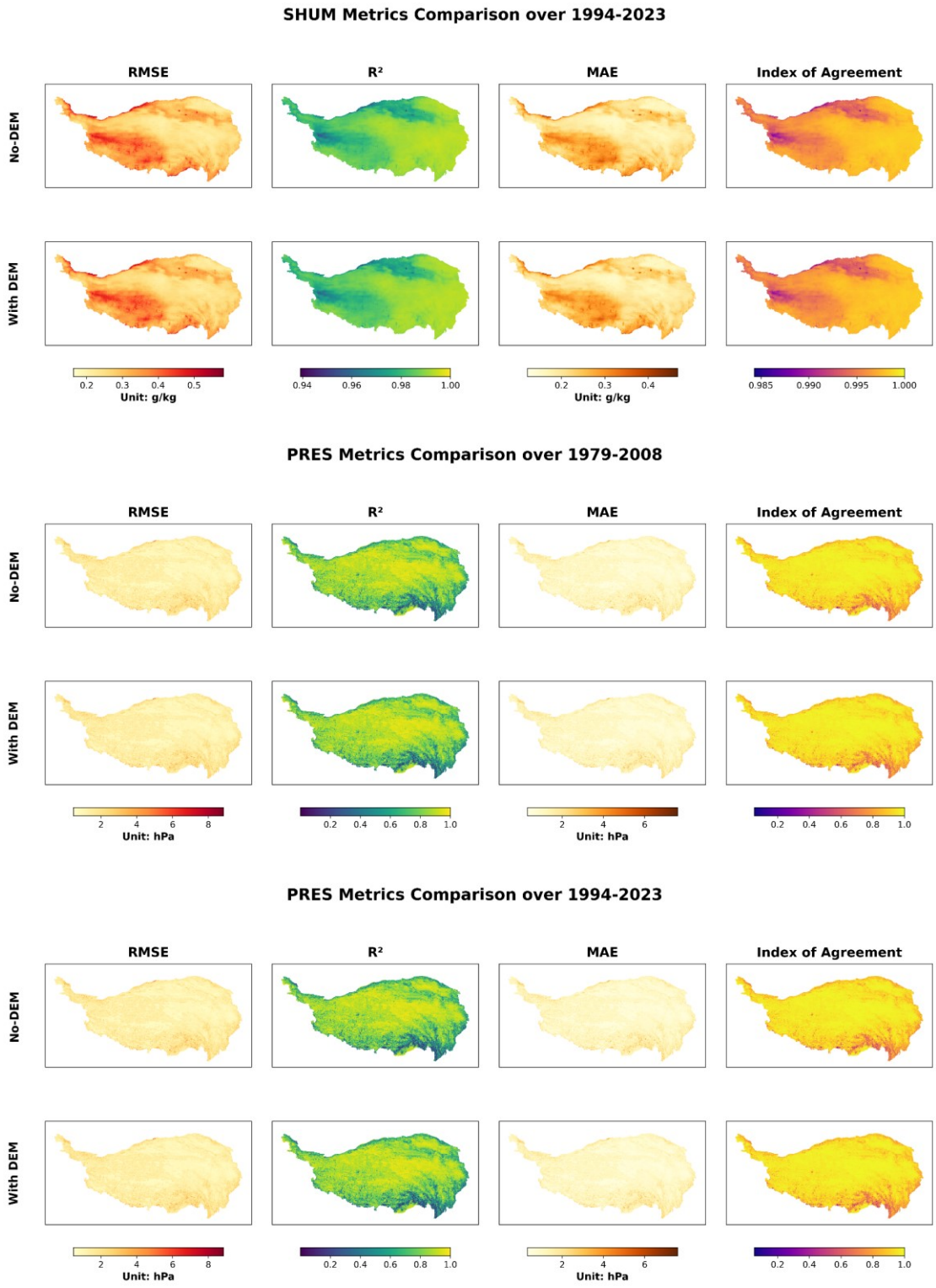


Figure S24-29. Spatial reconstruction-performance and DEM-sensitivity diagnostics for temperature, specific humidity, and surface pressure during 1979–2008 and 1994–2023, evaluated against TPMFD using R^2, RMSE, MAE, and IOA.

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.