

PAPER • OPEN ACCESS

Model architecture and calibration period sensitivity in controlled super-resolution reconstruction of near-surface air temperature

To cite this article: Jangho Lee and Sara Shamekh 2026 *Mach. Learn.: Earth* **2** 025017

View the [article online](#) for updates and enhancements.

You may also like

- [Robust deep learning-based downscaling of mean and extreme precipitation over the Indian subcontinent](#)
Midhun Murukesh and Pankaj Kumar
- [Interpolation of environmental data using deep learning and model inference](#)
Chibuike Chiedozie Ibebuchi and Itohan-Osa Abu
- [Enhancing subseasonal 2 m temperature forecast correction and downscaling over north China based on terrain-informed deep learning](#)
Xinyu Liu, Junjie Ma, Qiyu Zhang et al.

MACHINE LEARNING

Earth



PAPER

OPEN ACCESS

RECEIVED
30 July 2026

REVISED
23 September 2026

ACCEPTED FOR PUBLICATION
24 September 2026

PUBLISHED
7 October 2026

Original content from
this work may be used
under the terms of the
[Creative Commons
Attribution 4.0 licence](#).

Any further distribution
of this work must
maintain attribution to
the author(s) and the title
of the work, journal
citation and DOI.



Model architecture and calibration period sensitivity in controlled super-resolution reconstruction of near-surface air temperature

Jangho Lee* and Sara Shamekh

Courant Institute School of Mathematics, Computing, and Data Science, New York University, New York, New York 10012, United States of America

* Author to whom any correspondence should be addressed.

E-mail: jangho.lee@nyu.edu

Keywords: downscaling, climate change, RCAN, SRCNN, XGBoost

Abstract

Super-resolution reconstruction recovers fine-scale fields from coarse inputs, but its reliability across periods depends on whether relationships learned historically remain transferable. Using a controlled ERA5-Land 2 m temperature super-resolution reconstruction experiment over Africa, we compare U-Net, RCAN, super-resolution convolutional neural network, XGBoost, and quantile delta mapping (QDM) trained on the 1950s, 1980s, and 2010s and evaluated on the same daily 2021–2025 data. In the multiyear evaluation, spatial convolutional models outperform the pixel-wise baseline, with U-Net achieving the lowest mean root mean squared error across calibration periods in every year. In percentile-based hot-context subsets, XGBoost exhibits substantially larger errors and stronger dependence on training-period recency. U-Net maintains the lowest error, while RCAN has smaller calibration-period spread than U-Net. QDM has the smallest spread but higher error than the unadjusted coarse input. Within this same-source reconstruction setting, calibration-period sensitivity is both model- and distribution-dependent: average performance can conceal failures in the hot tail, and model selection should evaluate calibration-period sensitivity and extreme-temperature behavior separately.

1. Introduction

Near-surface air temperature is one of the most widely used variables in climate-impact research, as it directly influences heat exposure, agricultural productivity, ecosystem functioning, energy demand, and many other processes that unfold at regional to local scales (Parry 2019, Spector *et al* 2019, Mahato 2021, Lee and Dessler 2022). However, the spatial resolution of global climate model output is often too coarse to capture the temperature variability most relevant to these applications. Local temperature patterns are shaped not only by large-scale atmospheric conditions, but also by topography, proximity to water bodies, vegetation, soil properties, land cover, and other surface characteristics (Hathway and Sharples 2012, Wong *et al* 2012, Konarska *et al* 2016, Sun and Zhang 2016). These factors can generate substantial temperature differences over distances much shorter than a typical global climate model grid cell, which is generally on the order of 100 km. As a result, assessments based directly on coarse-resolution model output may provide climate information at a scale that is poorly aligned with practical decision-making. Downscaling, which translates coarse-resolution climate information to finer spatial and/or temporal scales, is therefore essential for producing climate projections that better reflect local environmental conditions (Casanueva *et al* 2019, Jiang *et al* 2023, Georgescu *et al* 2024, Lee *et al* 2024).

Downscaling methods are generally divided into dynamical and statistical approaches. Dynamical downscaling uses a regional climate model to simulate atmospheric processes at a finer resolution within the boundary conditions provided by a global climate model (Hong and Kanamitsu 2014, Xue *et al* 2014, Xu *et al* 2019, Hall *et al* 2024). Because it explicitly represents physical processes and interactions, dynamical downscaling can provide internally consistent fields for multiple climate variables. Producing

long simulations at high spatial resolution is computationally costly. These computational requirements often limit the size of the ensemble that can be generated and make it difficult to explore the full range of uncertainty in future climate projections. Statistical downscaling, by contrast, can provide a more computationally efficient alternative (Schoof 2013, Wilby and Dawson 2013, Sun *et al* 2024). Rather than explicitly simulating regional atmospheric dynamics, it estimates a relationship between coarse-scale climate information and a finer-scale target. Once that relationship has been learned, the model can be applied to large datasets at relatively low computational cost. This makes statistical downscaling attractive for applications requiring high-resolution projections from many climate models, scenarios, or ensemble members.

Within this statistical framework, a wide range of methods has been developed. Earlier approaches, including regression-based models, analog methods, and bias-correction techniques, generally imposed relatively simple or explicitly specified relationships between large-scale climate information and local variables (Wilby *et al* 2004). In contrast, recent machine learning methods, such as random forests and gradient-boosted decision trees (He *et al* 2024, Yuan *et al* 2025), can capture nonlinearities, thresholds, and interactions among predictors, while deep learning models can learn spatial structures and high-dimensional predictor–predictand relationships directly from data (Wang *et al* 2021, Daust *et al* 2026, Kheir *et al* 2023, Tomasi *et al* 2025). More recently, deep learning models and generative models, including convolution-based models, generative adversarial networks and diffusion-based approaches, have further expanded the ability of downscaling models to produce fine-scale fields with realistic spatial patterns and distributional characteristics (Hosseini Baghanam *et al* 2024, Jha *et al* 2025, Lee and Park 2025, Lee and Shamekh 2026, Ling *et al* 2024, Rosu *et al* 2025, Saoulis *et al* 2025, Zhang *et al* 2025, 2026).

Despite these methodological advances, statistical downscaling models are typically trained under historical climate conditions but are ultimately intended to be applied to future climates (Gutiérrez *et al* 2013, Dixon *et al* 2016, Lanzante *et al* 2018). This creates a setting that differs from conventional statistical prediction problems, where the training and prediction data are generally assumed to arise from the same underlying distribution. In future-climate applications, the target temperature distribution may shift under global warming, and changes may occur not only in the mean state but also in variability, extremes, and the upper tail of the distribution. At the same time, the relationships between large-scale atmospheric predictors and local surface responses may evolve. Statistical downscaling therefore depends on the assumption that the relationships learned from the historical climate remain sufficiently stable when the model is applied to another climate state. This assumption is often described as stationarity. If it does not hold, a model that performs well in historical validation may not remain reliable as the climate moves away from the conditions used for training.

These methodological advances also raise the question of how model flexibility affects temporal transferability. Greater flexibility may help a model learn robust spatial or physical relationships that remain useful under future climate states, but it may also allow the model to fit regime-specific patterns that are unlikely to hold outside the historical training distribution. Yet many studies demonstrate the accuracy of a particular downscaling method or compare several methods within a common historical period. As a result, it remains unclear whether recent improvements in statistical downscaling have also improved robustness to climate shifts, or whether they mainly provide better interpolation within the historical distribution. It is also unclear whether the model that performs best overall is necessarily the model whose performance changes least when the calibration and application climates become more different. Resolving this uncertainty is important because model evaluation based only on average historical accuracy may favor a method that is not well suited to future-climate applications.

To address this gap, we investigate whether calibration-period sensitivity in controlled super-resolution reconstruction of near-surface air temperature depends on model choice. Rather than treating nonstationarity as a uniform limitation shared by all statistical approaches, we examine whether different model formulations vary in their ability to retain predictive skill as the climate conditions used for application move away from those used for calibration. Specifically, we assess whether increasing temporal separation between training and testing periods leads to systematic performance degradation, whether the magnitude of this degradation differs among models, and whether the model with the best overall historical accuracy is also the model that is least sensitive to calibration-period selection. By comparing both predictive accuracy and calibration-period sensitivity, this study provides a controlled comparison of super-resolution reconstruction models, motivated by future-climate downscaling applications, where the key concern is not only how well a model reproduces observed conditions, but also how reliably the learned relationship persists outside the period in which it was established.

2. Data and method of analysis

2.1. ERA5-land reanalysis

This study uses ERA5-Land reanalysis data (Muñoz-Sabater *et al* 2021) to construct a controlled super-resolution reconstruction experiment for near-surface air temperature. ERA5-Land provides land-surface climate variables on a regular latitude–longitude grid with a spatial resolution of $0.1^\circ \times 0.1^\circ$. The study domain covers Africa and the surrounding land regions, extending from 40.0° N to 35.0° S and from 20.0° W to 55.0° E in the original grid.

The target variable is 2 m air temperature (T2M), sampled at 3 hourly temporal resolution. The original ERA5-Land T2M field is treated as the high-resolution reference, and a coarse-resolution input is generated by spatially aggregating the same field to $1.5^\circ \times 1.5^\circ$. The super-resolution reconstruction task is therefore defined as reconstructing the 0.1° T2M field from degraded coarse T2M and a limited set of auxiliary predictors.

Four ERA5-Land variables are used as physical predictors: coarse T2M, high-vegetation leaf area index (LAI-H), low-vegetation leaf area index (LAI-L), and geopotential (GPH). Coarse T2M provides the large-scale thermal condition, while LAI-H and LAI-L describe vegetation-related surface effects that can influence near-surface temperature through evapotranspiration, shading, and surface energy partitioning. GPH is used as a static terrain-related predictor associated with elevation and topographic structure. Latitude and sine/cosine encodings of day of year and hour complete the nine input channels used by the learned models.

ERA5-Land provides spatially complete fields for examining how super-resolution reconstruction models behave when calibrated in one climate period and applied to another. Future climate down-scaling is ultimately applied to gridded model output, which requires spatially complete and physically consistent predictor and target fields. Using variables from the same reanalysis product provides a controlled setting in which the coarse input, high-resolution target, and auxiliary predictors are internally consistent.

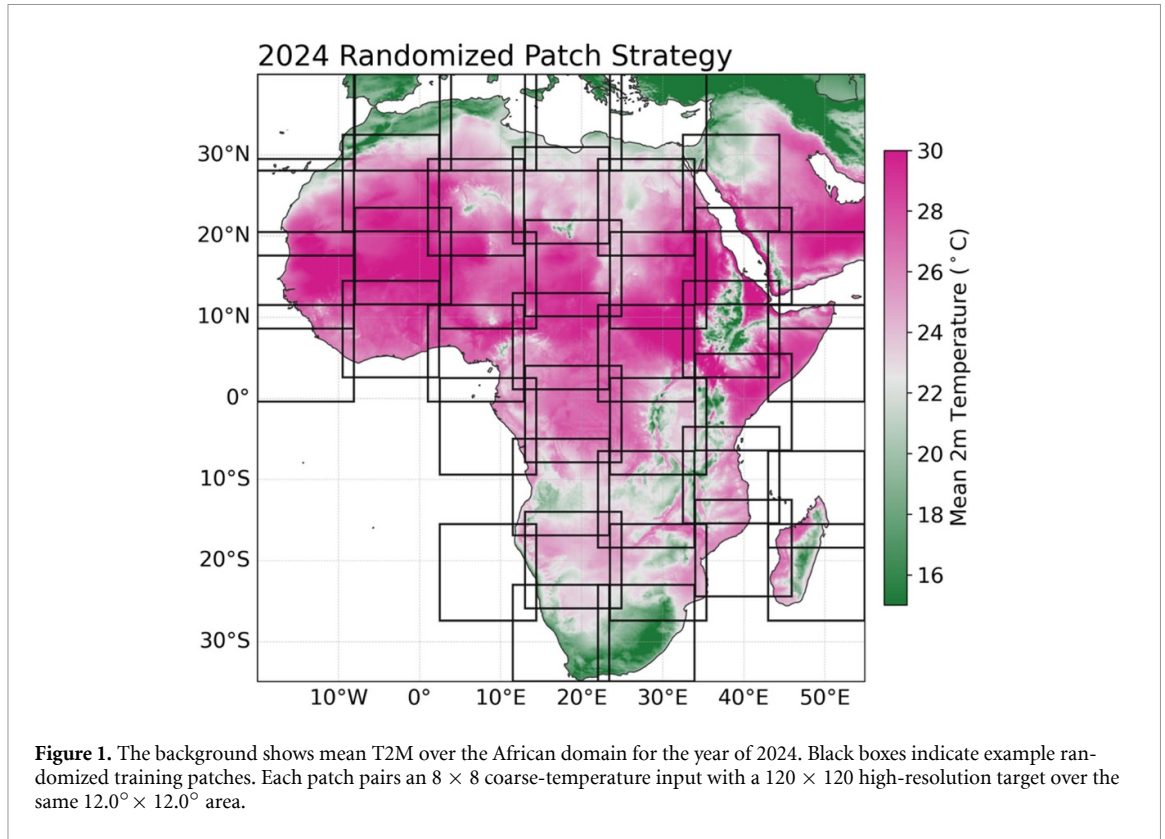
The predictor set is intentionally limited. Variables such as humidity, wind, radiation, soil moisture, and upper-air atmospheric fields are excluded, although they may contain useful information for local temperature prediction. Including many dynamically related predictors could make the reconstruction problem easier but would also make it harder to compare the model implementations and calibration periods. The purpose is to test whether different super-resolution reconstruction models can recover fine-scale T2M structure from a common coarse-temperature input and a small set of surface and geographic predictors.

Inputs and targets for the learned models are normalized before training using statistics computed from the corresponding calibration period. The models are trained in a direct prediction setting: they predict the high-resolution T2M field itself rather than the residual between the high-resolution target and the coarse input. Model outputs are transformed back to physical temperature units before evaluation.

2.2. Study design and data preprocessing

The study design is built to evaluate whether super-resolution reconstruction models trained under different historical climate conditions retain their skill when applied to a later climate state. Three historical calibration periods are considered: 1950s (1950–1959), 1980s (1980–1989), and 2010s (2010–2019). These periods represent temporally separated climate states within the ERA5-Land record. The primary evaluation uses daily 12:00 UTC fields from 2021–2025: 1826 dates and 30 non-overlapping tiles per date, giving 54 780 common date–tile cases. Each method is calibrated separately for the three calibration periods and then evaluated against the same 2021–2025 reference data. Performance differences are reported as calibration-period sensitivity, reflecting the combined changes in climate conditions, predictor support, reanalysis constraints, and training samples across decades. The full 3 hourly 2025 evaluation is retained as a complementary case study. It is worth noting that the 1950s provide an early-calibration case with a cooler climate and sparser observational constraints than the later periods, such as satellites. Comparing all three decades captures model behavior across this range of calibration conditions.

Patch-based samples are used instead of training on the full spatial field at once. Each sample pairs an 8×8 coarse-temperature patch with a 120×120 high-resolution target patch over the same $12.0^\circ \times 12.0^\circ$ area. The patch-based approach is necessary for both computational efficiency and sample diversity. Training directly on the full African domain for decade-long 3 hourly data would be inefficient and would provide fewer distinct spatial training samples. By extracting local patches, the models can



learn from many different spatial and temporal contexts while keeping each training sample computationally manageable.

Training patches are selected using a reproducible randomized patch strategy. A spatial jitter of two coarse cells, equivalent to approximately 3.0° , is applied to shift patch locations, and patch overlap is controlled with a maximum overlap of 30 high-resolution cells, also approximately 3.0° . This prevents the training samples from being fixed to a rigid spatial grid and reduces repeated exposure to nearly identical neighboring patches. The randomized shifting also reduces the chance that the models simply memorize fixed patch identities or location-specific patterns, encouraging them to learn a more general coarse-to-fine relationship. Figure 1 illustrates this randomized patch sampling strategy over the African domain, with boxes indicating example training patches superimposed on the mean T2M field.

2.3. U-Net

U-Net is used as an additional spatial convolutional model with an encoder–decoder structure and skip connections (Ronneberger *et al* 2015). It receives the same nine-channel predictor patch as the other learned models and produces a one-channel high-resolution T2M field. The coarse-temperature input is already repeated onto the fine grid during preprocessing, so the network reconstructs temperature on the 120×120 input grid.

The encoder extracts features at progressively coarser spatial resolutions, while the decoder returns them to the original grid. The implementation uses feature widths of 4, 8, and 16 channels, corresponding to spatial dimensions of 120×120 , 60×60 , and 30×30 . Two max-pooling stages connect these resolutions.

Let X denote the normalized nine-channel input patch and e_l the encoder feature map at resolution level l . The encoder is written as:

$$e_0 = E_0(X), \quad e_l = E_l(P(e_{l-1})), \quad l = 1, 2 \quad (1)$$

where E_l is a convolutional block with its own learned weights, and P denotes 2×2 max pooling with stride 2. Each block contains two dense 3×3 convolutions, each followed by a rectified linear unit (ReLU). A one-cell padding preserves spatial dimensions within each block. Thus, pooling changes the spatial resolution, while the deeper encoder blocks increase the number of feature channels. The deepest feature map, e_2 , forms the bottleneck between the encoder and decoder.

Table 1. Model capacity and fitting settings for each calibration decade.

Model	Capacity/structure	Training samples per decade	Validation samples	Fitting
U-Net	7753 parameters; widths 4/8/16; two pooling levels	1544 299–1544 776 patches	4096 patches	AdamW; 126–128 selected epochs
RCAN	6437 parameters; width 12; one residual attention block	1544 299–1544 776 patches	4096 patches	AdamW; 122–128 selected epochs
SRCNN	4509 parameters; widths 6/4; kernels 9/1/5	1544 299–1544 776 patches	4096 patches	AdamW; 118–126 selected epochs
XGBoost	512 depth-four trees	24 708 784–24 716 416 pixel rows	131 072 pixel rows	Boosted squared-error regression
QDM	Month $\times$ 30 tile empirical distributions	Daily-noon valid pixels pooled by calibration decade	No predictive tuning	Additive distribution mapping

Starting from $d_2 = e_2$, the decoder upsamples the deeper features and combines them with encoder features at the corresponding resolution. The decoder and final temperature prediction are written as:

$$d_l = D_l([e_l, U_l(d_{l+1})]), l = 1, 0 \quad (2)$$

$$\hat{Y} = C(d_0) \quad (3)$$

where d_l denotes the decoded feature map, U_l is bilinear interpolation to the spatial dimensions of e_l , and the brackets denote concatenation along the channel dimension. D_l is a decoder block containing the same two-convolution and ReLU sequence as an encoder block, but with separate learned weights. The decoder reduces the concatenated features to 8 channels at level 1 and 4 channels at level 0. C is a learned 1×1 convolution, and $\hat{Y}$ denotes the predicted normalized T2M field.

The skip connections make the encoder features at each resolution directly available during reconstruction. They combine higher-resolution features with features processed through the coarser bottleneck. These connections concatenate feature maps rather than adding the input temperature to the output. This distinguishes the encoder–decoder pathway from the residual feature additions used in RCAN, while providing another spatial formulation for the same reconstruction task.

The model has 7753 trainable parameters, including convolutional biases. It uses no batch normalization, dropout, channel-attention module, or final output activation. Temperature is predicted directly in normalized units and transformed back to physical units for evaluation. The loss, optimization settings, and calibration-period fitting procedure are reported in table 1 and its accompanying text.

2.4. Residual channel attention network (RCAN)

The RCAN is used as the attention-based super-resolution reconstruction model (Zhang *et al* 2018). It receives a multi-channel predictor patch, including coarse T2M, LAI-H, LAI-L, and GPH, and predicts a one-channel high-resolution T2M field. Its fully convolutional structure preserves the spatial organization of each patch and uses neighboring-grid information to reconstruct fine-scale temperature patterns.

The architecture consists of a head convolution, a residual group, a body-tail convolution, and a final output convolution. The head convolution projects the predictors into 12 learned feature channels. The residual group contains one residual channel attention block, in which convolutional layers learn fine-scale corrections while residual connections retain the large-scale thermal information contained in the incoming features.

The key component of RCAN is the channel attention mechanism. After convolutional feature extraction, each feature channel is summarized by global average pooling:

$$z_c = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W F_c(i, j) \quad (4)$$

where F_c is the c th feature channel, and H and W are the spatial dimensions of the feature map. This operation compresses each feature channel into a single value representing its average activation over the patch. We also test an ablated version in which this global pooling step is removed, but the main performance patterns did not change materially. We therefore retain global average pooling because it follows the standard RCAN formulation and reduces the influence of local pixel-scale noise on the channel-attention weights. The resulting channel summary vector is then passed through a small gating network:

$$s = \sigma(W_2 \delta(W_1 z)) \quad (5)$$

where δ denotes a nonlinear activation function and σ denotes the sigmoid function. The reduction factor in the attention module is set to 4, so the gating network first reduces and then restores the channel dimension before producing the final channel weights.

The attention weights are used to rescale the feature map channel by channel:

$$F_c^{\text{att}}(i, j) = s_c F_c(i, j) . \quad (6)$$

Thus, feature channels that are more relevant for a given sample receive greater emphasis, while less informative channels are suppressed. The attention weights are learned in feature space rather than being assigned directly to the raw predictors. Therefore, the mechanism is interpreted as adaptive weighting of learned feature representations derived from coarse temperature, vegetation, terrain, geographic, and temporal information.

After the residual group, a body-tail convolution further processes the attention-refined features. A long residual connection adds the output of the residual group back to the earlier feature representation, allowing the model to preserve background thermal structure while focusing the deeper layers on fine-scale reconstruction. Finally, the output convolution maps the 12 hidden feature channels to a single high-resolution T2M field.

2.5. Super-resolution convolutional neural network (SRCNN)

The SRCNN is used as a non-attention convolutional baseline (Dong *et al* 2015). SRCNN has a simpler architecture than RCAN. It consists of three convolutional stages: feature extraction, nonlinear mapping, and reconstruction. The first convolution uses a relatively large kernel to extract local spatial features from the multi-channel input. This step allows the model to detect broad spatial patterns such as coarse-temperature gradients, vegetation contrasts, coastal transitions, and terrain-related temperature differences. The second convolution acts as a channel-mixing or bottleneck layer, combining the extracted features into a more compact representation. The final convolution reconstructs the high-resolution temperature field from this intermediate feature representation.

In this study, SRCNN follows the sequence of a 9×9 convolution, a nonlinear activation, a 1×1 convolution, another nonlinear activation, and a final 5×5 convolution. The 9×9 convolution provides a wider local receptive field, while the 1×1 convolution mixes information across feature channels without expanding the spatial window. The final 5×5 convolution converts the learned feature representation into the predicted temperature field. This structure allows SRCNN to represent spatially coherent temperature patterns while remaining much simpler than RCAN.

SRCNN applies its learned convolutional filters in the same way to all samples, whereas RCAN adaptively reweights feature channels through its attention mechanism. SRCNN therefore provides a non-attention CNN baseline for comparison with RCAN.

2.6. XGBoost

XGBoost is used as a pixel-wise tabular baseline. Each valid land pixel is treated as an independent training sample, and the model predicts high-resolution T2M from the local predictor vector. It uses the same predictor variables as the convolutional models, with spatial information represented by the predictors at the individual pixel.

This model is included because tree-based gradient boosting is a strong nonlinear regression method for tabular data (Chen and Guestrin 2016). It captures nonlinear relationships between local predictors and temperature, providing a useful tabular comparison with the spatial convolutional models.

For fair comparison, XGBoost predictions are evaluated using the same evaluation patch locations, masks, and time steps as the convolutional models. Although the model itself predicts each pixel independently, its outputs are reorganized into the same patch structure used by U-Net, RCAN, and SRCNN. This ensures that performance is compared on common evaluation samples.

2.7. Quantile delta mapping (QDM)

Additive QDM (Cannon *et al* 2015) is included as a distributional benchmark. For each calendar month and spatial tile, empirical distributions of coarse and high-resolution calibration temperatures are estimated from valid pixels in daily 12:00 UTC fields. For an evaluation temperature, its percentile in the pooled 2021–2025 coarse-input distribution is mapped to the corresponding high-resolution calibration quantile; the difference between the evaluation temperature and the coarse calibration quantile is then added. Distributions use 0.1 K histogram bins from 200 to 340 K. This implementation uses the evaluation coarse-input distribution and is therefore transductive; high-resolution evaluation temperatures are reserved for scoring. QDM uses temperature, month, and tile rather than the full nine-channel predictor set. The unadjusted coarse field, repeated on the fine grid, provides an additional reference.

Table 1 summarizes model capacity and fitting. U-Net, RCAN, and SRCNN use masked mean-squared error, AdamW with learning rate 10^{-4} and weight decay 10^{-4} , batch size 256, and a maximum of 128 epochs. XGBoost uses squared error, at most 512 depth-four trees, learning rate 0.03, and row and column subsampling of 0.8; 16 training and 32 validation pixels are sampled per patch. All use one fitted seed and early stopping patience 16, with configurations fixed in advance. Validation uses a random calibration-patch holdout, capped at 4096 patches. Normalization precedes that internal split.

2.8. Evaluation metrics and calibration-period sensitivity

Model evaluation is conducted on finite valid-land pixels using the same 2021–2025 daily reference samples for all methods and calibration periods. In addition to standard error statistics, such as root mean squared error (RMSE), the complementary 3 hourly 2025 analysis uses the structural similarity index (SSIM) to evaluate whether a model preserves the spatial organization of the high-resolution T2M field. This is important because super-resolution reconstruction must recover coherent spatial patterns removed during coarse graining, rather than simply reduce pointwise error.

The primary metrics pool 536 915 214 valid-land pixel scores, with equal weight per pixel. Daily 12:00 UTC samples evaluate instantaneous temperature at a common UTC time, corresponding to different local solar times across the domain. RMSE differences have identical numerical values in K and °C.

For a predicted T2M field x and reference field y , SSIM (Wang *et al* 2004) is defined as

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (7)$$

where μ_x and μ_y are patch-wide means, σ_x^2 and σ_y^2 are patch-wide variances, σ_{xy} is the patch-wide covariance, and C_1 and C_2 are small constants used for numerical stability. Higher SSIM indicates that the prediction better preserves the spatial structure of the reference T2M field. We use SSIM as a complement to error-based metrics because two models can have similar average errors but differ in how well they reproduce gradients, texture, and coherent sub-grid temperature patterns.

To quantify the calibration-period sensitivity of each model, we compare performance across the three calibration periods under the same evaluation samples. Let Q_p denote a performance score for a model trained on calibration period p , where p denotes the 1950s, 1980s, or 2010s calibration period. The calibration-period spread is defined as

$$S_{\text{abs}}(Q) = \max_p(Q_p) - \min_p(Q_p) \quad (8)$$

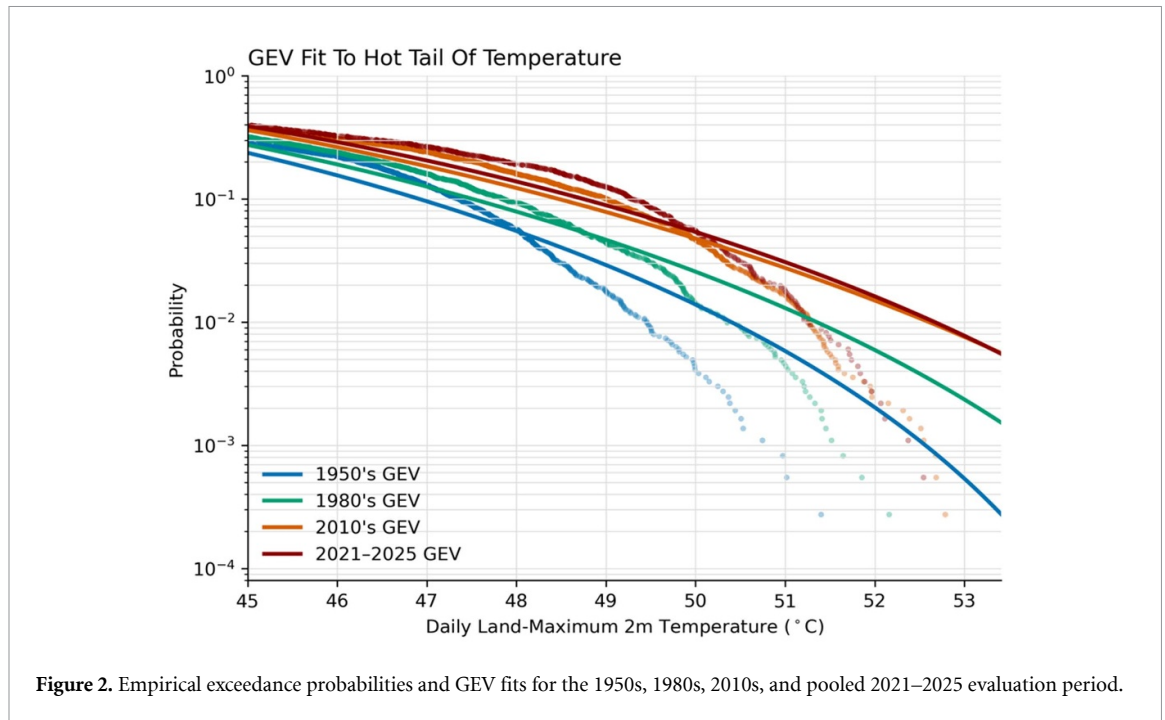
This spread measures how strongly the model performance changes when the calibration period is changed. A small spread indicates that the model is relatively insensitive to calibration-period selection, whereas a large spread indicates stronger calibration-period dependence.

For the full and extreme-temperature evaluations, we also use a relative spread to avoid interpreting larger absolute changes as greater calibration-period sensitivity simply because one model has a larger error scale. For a non-negative error metric E_p , the relative spread is defined as

$$S_{\text{rel}}(E) = \frac{\max_p(E_p) - \min_p(E_p)}{\bar{E}} \times 100 \quad (9)$$

where $\bar{E}$ is the arithmetic mean error across the three calibration periods for the same model and evaluation subset. Equation (9) summarizes the largest separation among the three errors relative to their mean, allowing comparison across models with different error scales. A coefficient of variation summarizes deviations around the mean; with three calibration periods, both provide descriptive summaries. We complement the relative range with paired period contrasts to assess the direction and uncertainty of the differences. The relative range is applied to non-negative errors and is sensitive to mean values near zero.

Hot-context subsets are defined by the P90, P95, and P99 percentiles of reference patch-maximum temperature pooled across all 2021–2025 daily evaluation cases. The thresholds are 43.28 °C, 44.98 °C, and 47.81 °C, selecting 5478, 2739, and 548 date–tile cases, respectively. All valid pixels in each selected patch are scored, characterizing reconstruction across hot spatial contexts. Selection is fixed from the reference and shared by all methods. P99 spans 352 d and eight tiles, with 59.49% of its cases in one tile.



Evaluation-sampling uncertainty is estimated with 2000 paired bootstrap replicates. Within each year, overlapping seven-day blocks are sampled with replacement and truncated to the observed day count; 30 tiles are sampled with replacement once per replicate. Crossed day and tile weights are shared by all methods and subsets; tile weights are also shared across years. This combines moving-block resampling with crossed resampling of sampling dimensions (Owen 2007). Two-sided pointwise 95% percentile intervals are obtained for RMSE, relative spread, within-model period contrasts, and between-model contrasts of the arithmetic mean of the three period-specific RMSEs. These pointwise intervals quantify evaluation-sampling uncertainty conditional on the fitted models, QDM distributions, and observed thresholds, with tiles treated as separate resampling units. Block lengths of 1, 14, and 30 d assess sensitivity to the temporal choice.

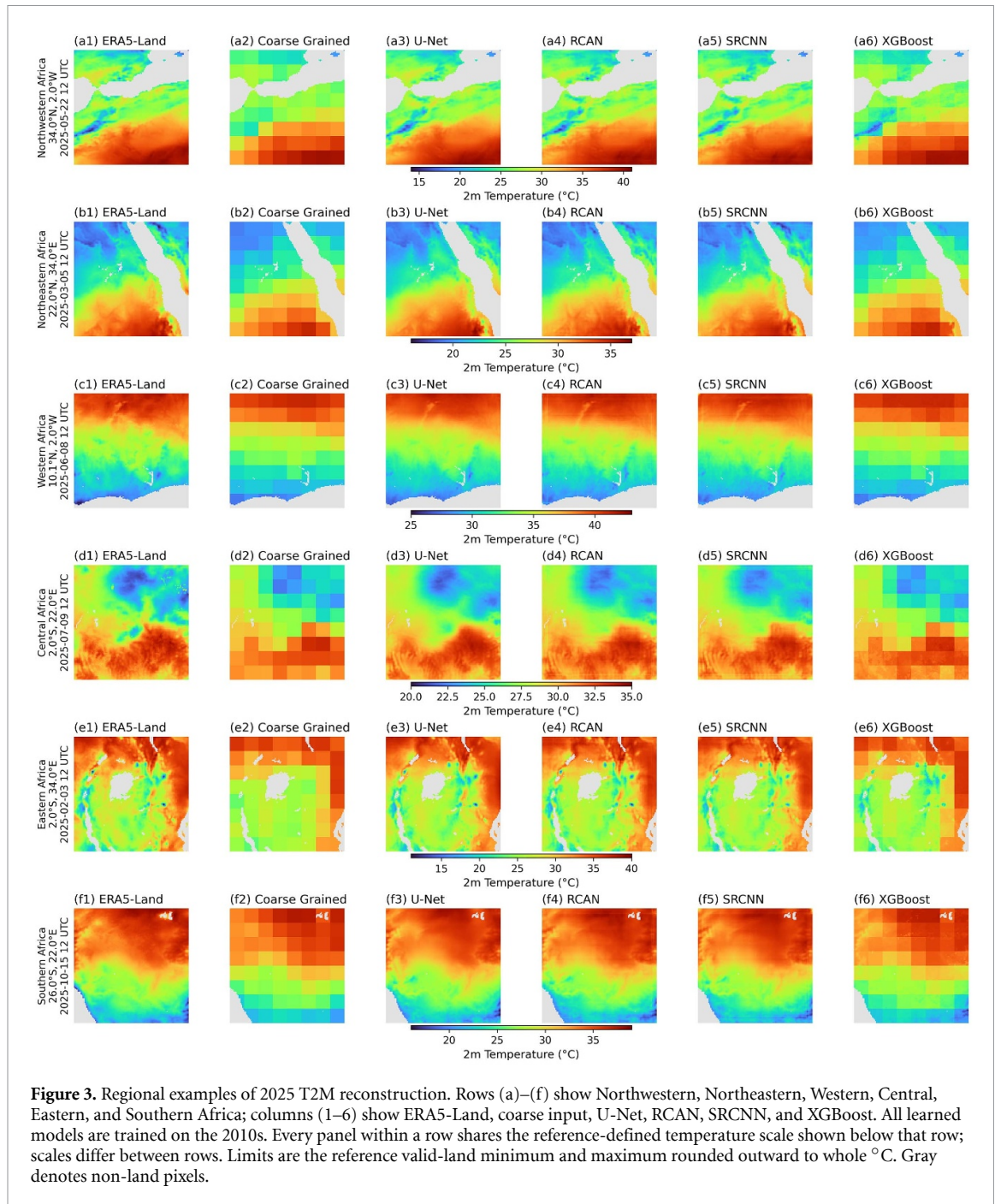
3. Results

3.1. Temporal shift in T2M distribution

Figure 2 compares the hot-tail distributions of daily land-maximum T2M for the three calibration periods, 1950s, 1980s, and 2010s, and for the pooled 2021–2025 evaluation period. Daily land maxima are calculated across all valid land pixels and available 3-hourly fields within each UTC day. Dots indicate empirical exceedance probabilities, while solid lines indicate generalized extreme value (GEV) fits, which provide a compact description of upper-tail behavior. Each historical period spans 10 years, while the evaluation distribution pools all five years used in the primary evaluation.

The fitted curves show a clear shift toward warmer upper-tail conditions over time. The 1950s distribution decays most rapidly, indicating that very high daily land-maximum T2M values are relatively rare in the earliest calibration period. The 1980s distribution shifts toward warmer conditions, and the 2010s distribution shows a further increase in exceedance probability at high temperatures. The 2021–2025 curve lies slightly above the 2010s curve over most of the hot tail, while the empirical distributions are closely aligned at the highest temperatures. This indicates that hot-tail T2M conditions during 2021–2025 are less represented in the older calibration periods, whereas the 2010s provide a closer match to the evaluation climate.

This distributional shift provides context for the calibration-period comparison. A model trained on the 1950s or 1980s learns the relationship between coarse T2M, LAI-H, LAI-L, GPH, and high-resolution T2M under a cooler calibration climate than the one encountered during 2021–2025. The multiyear evaluation below assesses predictive accuracy and calibration-period sensitivity across these differing temperature distributions.



3.2. Qualitative examples of super-resolution reconstruction

Figure 3 illustrates reconstruction in six 2025 cases spanning Northwestern, Northeastern, Western, Central, Eastern, and Southern Africa. U-Net, RCAN, and SRCNN recover spatial gradients and fine-scale structures removed by coarse graining. XGBoost retains more block-like patterns in its pixel-wise reconstruction. The 2010s-trained models are compared on the same cases. U-Net has the lowest displayed-case RMSE in five cases and RCAN in the northeastern case. These high-contrast examples illustrate spatial behavior. Quantitative comparisons across common evaluation samples follow in section 3.3.

The six $12^\circ \times 12^\circ$ tiles are fixed geographically, with centers at 34.05° N, 2.05° W; 22.05° N, 33.95° E; 10.05° N, 2.05° W; 1.95° S, 21.95° E; 1.95° S, 33.95° E; and 25.95° S, 21.95° E, respectively. Within each tile, spatial temperature contrast is defined as the difference between the valid-land P95 and P5 in each of the 365 daily 12:00 UTC reference fields in 2025. We select the date nearest the within-tile P90 of these contrasts, resolving ties by the earliest date. Selection is based on spatial contrast in the reference field.

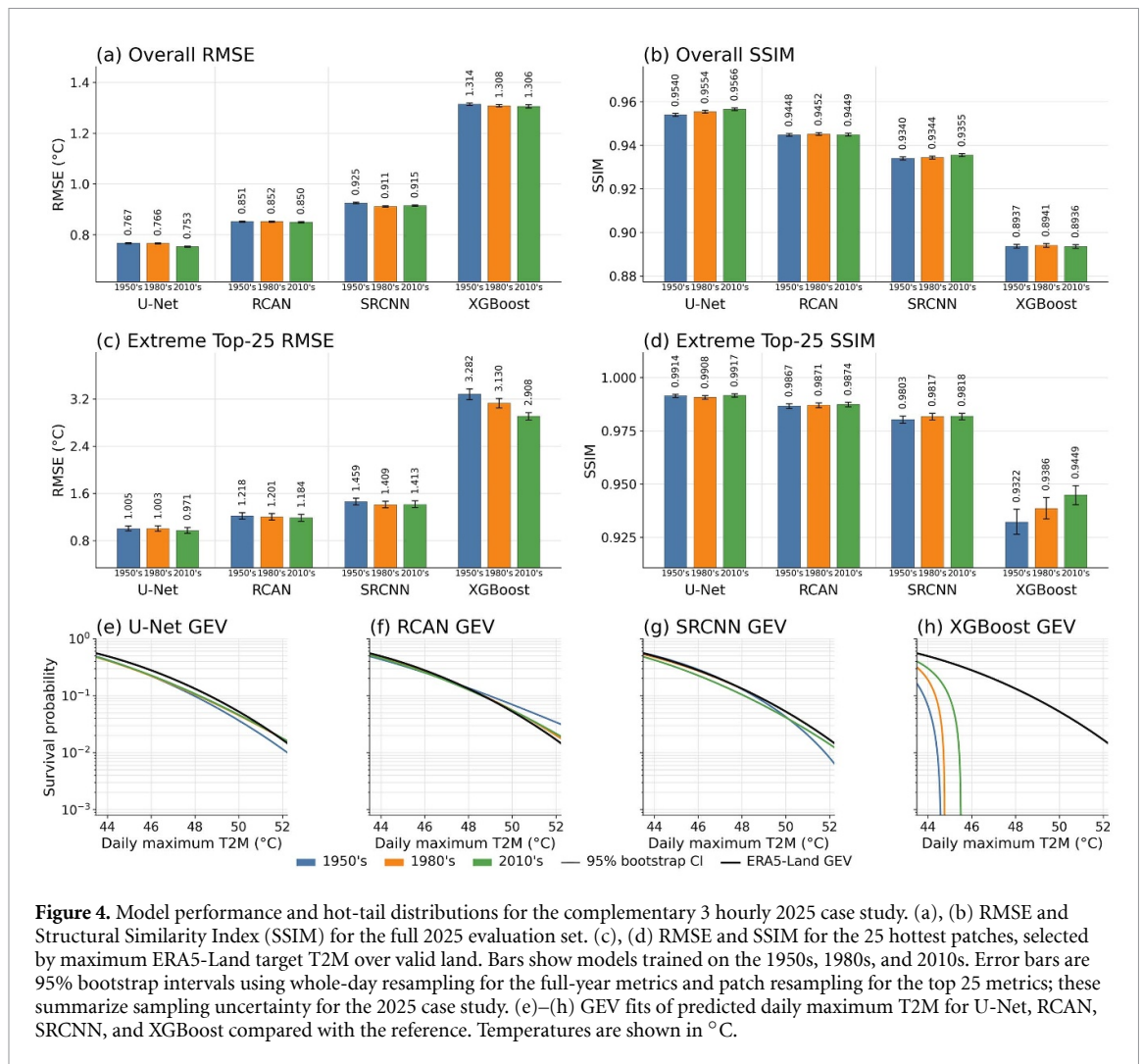


Figure 4. Model performance and hot-tail distributions for the complementary 3 hourly 2025 case study. (a), (b) RMSE and Structural Similarity Index (SSIM) for the full 2025 evaluation set. (c), (d) RMSE and SSIM for the 25 hottest patches, selected by maximum ERA5-Land target T2M over valid land. Bars show models trained on the 1950s, 1980s, and 2010s. Error bars are 95% bootstrap intervals using whole-day resampling for the full-year metrics and patch resampling for the top 25 metrics; these summarize sampling uncertainty for the 2025 case study. (e)–(h) GEV fits of predicted daily maximum T2M for U-Net, RCAN, SRCNN, and XGBoost compared with the reference. Temperatures are shown in °C.

3.3. Model-dependent super-resolution reconstruction skill and calibration-period sensitivity

Model performance is evaluated primarily using the common daily 2021–2025 dataset and its P90, P95, and P99 hot-context subsets. The full 3 hourly 2025 evaluation and the 25 hottest 2025 patches are retained as a complementary case study (figure 4). For the latter, patches are ranked by the maximum high-resolution ERA5-Land T2M among valid land pixels, and the 25 hottest patches are retained. Reference-based selection ensures that all methods within each comparison are evaluated on identical cases. Primary inference on hot-context performance uses the larger percentile subsets and paired block bootstrap.

Figure 4 compares model performance in the full-year and extreme evaluations. For the full 2025 dataset, differences among model implementations are much larger than differences among calibration periods. U-Net has the lowest RMSE, followed by RCAN, SRCNN, and XGBoost. The model ranking is therefore consistent across calibration periods.

The overall SSIM results show the same ordering. The CNN-based models therefore reproduce both pointwise T2M and spatial structure more accurately than the pixel-wise model. Calibration-period sensitivity is small in the full-year evaluation, with RCAN showing the smallest RMSE spread. Domain-averaged metrics thus show larger differences among these model implementations than among calibration periods.

The complementary top-25 evaluation accentuates the differences among models. U-Net retains the lowest error, while XGBoost deteriorates most strongly. The performance gap between the CNN models and XGBoost is therefore substantially larger in the hot tail.

Calibration-period sensitivity is also more evident under extreme heat. XGBoost RMSE decreases monotonically as the calibration period becomes more recent, whereas RCAN remains comparatively stable. This contrast persists after normalizing the spread by each model's mean extreme RMSE: XGBoost remains substantially more sensitive to the calibration period than the spatial CNNs.

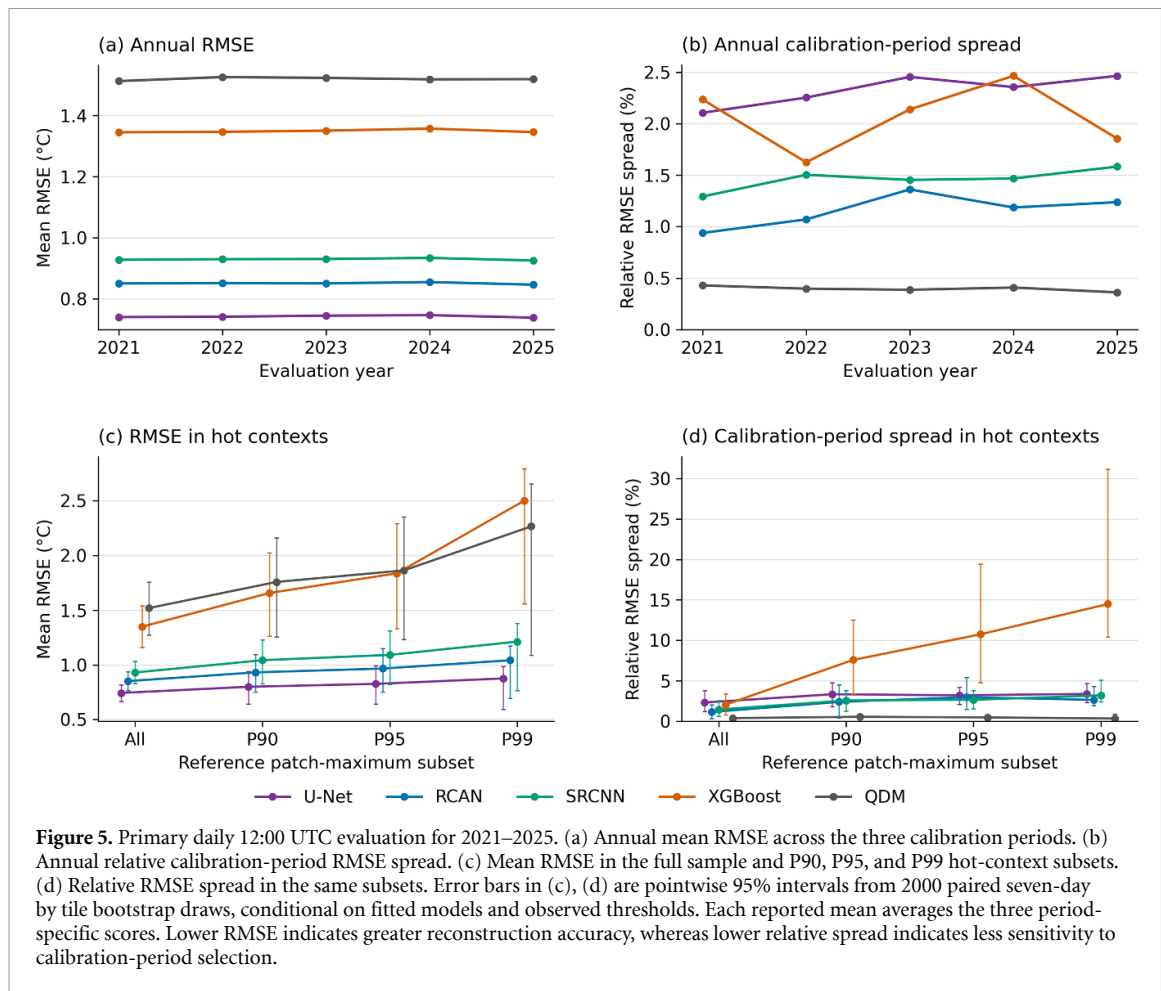


Figure 5. Primary daily 12:00 UTC evaluation for 2021–2025. (a) Annual mean RMSE across the three calibration periods. (b) Annual relative calibration-period RMSE spread. (c) Mean RMSE in the full sample and P90, P95, and P99 hot-context subsets. (d) Relative RMSE spread in the same subsets. Error bars in (c), (d) are pointwise 95% intervals from 2000 paired seven-day by tile bootstrap draws, conditional on fitted models and observed thresholds. Each reported mean averages the three period-specific scores. Lower RMSE indicates greater reconstruction accuracy, whereas lower relative spread indicates less sensitivity to calibration-period selection.

Extreme SSIM provides a complementary measure. U-Net maintains the strongest spatial agreement, while RCAN and SRCNN also show similar spatial skill across calibration periods. XGBoost's spatial agreement improves with more recent calibration. Its spatial similarity alongside high RMSE suggests that XGBoost captures some broad spatial organization while underestimating the magnitude of extreme T2M. Joint interpretation of RMSE and SSIM therefore distinguishes spatial-pattern similarity from accurate reconstruction of hot-tail temperature magnitude.

The GEV fits in figures 4(e)–(h) show the corresponding changes in predicted daily maximum T2M distributions. U-Net and RCAN follow the ERA5-Land reference tail across calibration periods, while SRCNN shows somewhat greater separation among periods. XGBoost strongly truncates the upper tail: its survival curves decay rapidly and do not extend to the highest reference temperatures. The 2010s-trained XGBoost curve lies closer to the reference than the curves trained on earlier periods, but all three underestimate the upper-tail magnitude.

This truncation is consistent with the extreme-temperature bias. XGBoost has a cold bias across all three calibration periods, with greater underestimation for the earlier-trained models. This behavior is consistent with the limited extrapolation capacity of tree-based models, which predict from partitions learned within the training distribution rather than extrapolating smoothly beyond it. The larger bias of the 1950s-trained model is consistent with a greater distributional separation between the 1950s calibration climate and the 2025 hot tail.

A residual XGBoost sensitivity test reduces the apparent relative RMSE variation across calibration periods but increases absolute RMSE by approximately 1.5 times. Interpreting the relative spread alongside absolute RMSE reveals the associated loss of predictive skill.

In the primary multiyear evaluation, U-Net has the lowest mean RMSE across calibration periods in every evaluation year and hot-context subset (figure 5). The ordering U-Net, RCAN, SRCNN, XGBoost, and QDM is unchanged across years. The paired U-Net-minus-RCAN difference is -0.108 °C overall (pointwise 95% interval -0.122 °C to -0.096 °C) and -0.166 °C at P99 (-0.193 °C to -0.095 °C). Both U-Net and RCAN show larger errors in the hottest contexts, with a smaller proportional increase for U-Net. RCAN has a smaller descriptive relative RMSE spread than U-Net in every subset, whereas

QDM has the smallest spread overall. QDM has higher RMSE than the unadjusted coarse input in every subset. QDM therefore combines low calibration-period sensitivity with higher absolute error.

At P99, U-Net's 1950s-minus-2010s RMSE difference is $0.0297\text{ }^{\circ}\text{C}$ (pointwise 95% interval 0.0110 to $0.0358\text{ }^{\circ}\text{C}$). For RCAN, the corresponding difference is $0.0276\text{ }^{\circ}\text{C}$ ($0.0077\text{ }^{\circ}\text{C}$ – $0.0361\text{ }^{\circ}\text{C}$), while the 1980s-minus-2010s difference is $0.0098\text{ }^{\circ}\text{C}$ ($-0.0166\text{ }^{\circ}\text{C}$ – $0.0189\text{ }^{\circ}\text{C}$). The intervals are positive for the 1950s–2010s differences in both models, whereas RCAN's 1980s–2010s interval includes zero. These conditional, unadjusted comparisons quantify differences in predictive performance across calibration periods.

Overall, the results show that calibration-period sensitivity depends on both model implementation and the part of the T2M distribution being evaluated. Under average conditions, spatial CNNs outperform the pixel-wise baseline, while calibration-period effects remain small. Under extreme heat, XGBoost becomes substantially less accurate and more dependent on the calibration period. U-Net has the lowest error, while RCAN has smaller relative calibration-period spread than U-Net. Model selection should therefore consider calibration-period sensitivity and hot-tail error in addition to average accuracy.

4. Summary and conclusion

This study examines whether calibration-period sensitivity in controlled super-resolution reconstruction of near-surface air temperature differs among model implementations. Using ERA5-Land T2M over Africa, we compare models trained on temporally separated historical periods and evaluated under the same 2021–2025 daily sample, with the full 3 hourly 2025 data providing a complementary case study.

The largest differences in average reconstruction skill occur among model implementations, with U-Net giving the lowest RMSE. The spatial CNNs reconstruct fine-scale T2M patterns more effectively than the pixel-wise baseline, highlighting the spatial component of the super-resolution reconstruction task. Calibration-period differences remain small in the full-year evaluation but become clearer under extreme heat. In the upper tail of T2M, the pixel-wise model shows larger errors, stronger underestimation, and greater dependence on training-period recency, indicating poorer performance in hot contexts and greater sensitivity to calibration-period selection.

RCAN has smaller relative calibration-period spread than U-Net, while U-Net has lower error, including under hot conditions. QDM has the smallest spread and higher RMSE than the unadjusted coarse input. These results support treating calibration-period sensitivity and predictive accuracy as separate evaluation criteria, with hot-tail performance providing additional information beyond average accuracy.

This result motivates broader evaluation for future-climate downscaling. Models are often selected using historical or domain-averaged validation accuracy, but such evaluation may conceal failures under distribution shift. Model selection should also consider temporal transferability, extreme-condition behavior, and stability across calibration periods. These criteria are particularly important for climate-impact applications, where underestimating local hot extremes may have greater consequences than comparable errors under typical conditions.

The lower CNN errors are consistent with a role for spatial representation in this controlled super-resolution reconstruction. Pixel-wise nonlinear regression captures local predictor–T2M relationships, while CNNs additionally represent spatial gradients and coherent sub-grid thermal patterns through neighboring-grid context. RCAN further reweights learned feature channels according to local conditions.

Several limitations remain. The coarse T2M input is generated by aggregating the high-resolution ERA5-Land reference, giving a controlled, same-source assessment of spatial reconstruction and calibration-period sensitivity. Independent reanalysis or GCM inputs would introduce cross-product biases that may affect model ordering. Differences among calibration decades also combine changes in climate, predictor support, reanalysis constraints, and calibration samples. These factors jointly contribute to the observed calibration-period sensitivity.

The 2021–2025 evaluation covers a recent climate window over Africa, with the hottest subset concentrated in a limited part of the domain. The reported bootstrap intervals quantify evaluation-sampling uncertainty conditional on the fitted models and observed thresholds. The comparisons describe the tested model implementations as a whole. Further targeted comparisons can clarify the separate effects of attention, residual connections, model capacity, and predictor selection. Broader testing across years, regions, climate regimes, and independent reanalysis or GCM inputs is needed to assess transferability beyond the present reconstruction setting.

The main conclusion is that evaluation of controlled super-resolution reconstruction should extend beyond performance under historical or average conditions. Reconstructed temperatures depend on the

training data and on how the model represents the coarse-to-fine relationship. Treating calibration-period sensitivity and hot-context accuracy as separate model-selection criteria provides a practical framework for comparing performance under temporally shifted and extreme-temperature conditions.

Acknowledgment

This work was carried out thanks to the support by Wellcome (Grant No. 309864/Z/23/Z). The views expressed herein do not necessarily represent those of UNESCO, ICTP or Wellcome.

Data availability statement

The ERA5-Land data used in this study can be accessed via Climate Data Store (<https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land?tab=overview>), under ERA5-Land dataset archive (DOI: <https://10.24381/cds.e2161bac>).

The data that support the findings of this study are openly available at the following URL/DOI: <https://10.24381/cds.e2161bac>.

ORCID iD

Jangho Lee  0000-0002-8942-1092

References

- Cannon A J, Sobie S R and Murdock T Q 2015 Bias correction of GCM precipitation by quantile mapping: how well do methods preserve changes in quantiles and extremes? *J. Clim.* **28** 6938–59
- Casanueva A, Kotlarski S, Herrera S, Fischer A M, Kjellstrom T and Schwierz C 2019 Climate projections of a multivariate heat stress index: the role of downscaling and bias correction *Geosci. Model Dev.* **12** 3419–38
- Chen T and Guestrin C 2016 Xgboost: a scalable tree boosting system Paper presented at the proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining
- Daust K, Monahan A and Mahony C 2026 Generative adversarial networks for deep learning downscaling of temperature, humidity, and precipitation *Artif. Intell. Earth Syst.* **5** e250095
- Dixon K W, Lanzante J R, Nath M J, Hayhoe K, Stoner A, Radhakrishnan A, Balaji V and Gaitán C F 2016 Evaluating the stationarity assumption in statistically downscaled climate projections: is past performance an indicator of future results? *Clim. Change* **135** 395–408
- Dong C, Loy C C, He K and Tang X 2015 Image super-resolution using deep convolutional networks *IEEE Trans. Pattern Anal. Mach. Intell.* **38** 295–307
- Georgescu M, Broadbent A M and Krayenhoff E S 2024 Quantifying the decrease in heat exposure through adaptation and mitigation in twenty-first-century US cities *Nat. Cities* **1** 42–50
- Gutiérrez J M, San-martín D, Brands S, Manzanar R and Herrera S 2013 Reassessing statistical downscaling techniques for their robust application under climate change conditions *J. Clim.* **26** 171–88
- Hall A, Rahimi S, Norris J, Ban N, Siler N, Leung I R, Ullrich P, Reed K A, Prein A F and Qian Y 2024 An evaluation of dynamical downscaling methods used to project regional climate change *J. Geophys. Res. Atmos.* **129** e2023JD040591
- Hathway E and Sharples S 2012 The interaction of rivers and urban form in mitigating the Urban Heat Island effect: a UK case study *Build. Environ.* **58** 14–22
- He H, Wang J, Ding J and Wang L 2024 Spatial downscaling of precipitation data in arid regions based on the XGBoost-MGWR model: a case study of the Turpan–Hami region *Land* **13** 448
- Hong S-Y and Kanamitsu M 2014 Dynamical downscaling: fundamental issues from an NWP point of view and recommendations *Asia-Pac. J. Atmos. Sci.* **50** 83–104
- Hosseini Baghanam A, Nourani V, Bejani M and Ke C-Q 2024 Improving the statistical downscaling performance of climatic parameters with convolutional neural networks *J. Water Clim. Change* **15** 1772–96
- Jha S K, Gupta V, Sharma P J, Mishra A and Joshi S 2025 Deep learning super-resolution for temperature data downscaling: a comprehensive study using residual networks *Front. Clim.* **7** 1572428
- Jiang T, Krayenhoff E S, Voogt J A, Warland J, Demuzere M and Moede C 2023 Dynamically downscaled projection of urban outdoor thermal stress and indoor space cooling during future extreme heat *Urban Clim.* **51** 101648
- Kheir A M, Elnashar A, Mosad A and Govind A 2023 An improved deep learning procedure for statistical downscaling of climate data *Heliyon* **9** e18200
- Konarska J, Holmer B, Lindberg F and Thorsson S 2016 Influence of vegetation and building geometry on the spatial variations of air temperature and cooling rates in a high-latitude city *Int. J. Climatol.* **36** 2379–95
- Lanzante J R, Dixon K W, Nath M J, Whitlock C E and Adams-smith D 2018 Some pitfalls in statistical downscaling of future climate *Bull. Am. Meteorol. Soc.* **99** 791–803
- Lee J, Berkelhammer M, Wilson M D, Love N and Cintron R 2024 Urban land surface temperature downscaling in Chicago: addressing ethnic inequality and gentrification *Remote Sens.* **16** 1639
- Lee J and Dessler A E 2022 The impact of neglecting climate change and variability on ERCOT's forecasts of electricity demand in Texas *Weather Clim. Soc.* **14** 499–505

- Lee J and Park S Y 2025 WGAN-GP-based conditional GAN (cGAN) with extreme critic for precipitation downscaling in a key agricultural region of the Northeastern US *IEEE Access* **13** 46030–41
- Lee J and Shamekh S 2026 How far can we downscale? Resolution limits and physical interpretability of diffusion models for African precipitation *Mach. Learn. Earth* **2** 015018
- Ling F, Lu Z, Luo J-J, Bai L, Behera S K, Jin D, Pan B, Jiang H and Yamagata T 2024 Diffusion model-based probabilistic downscaling for 180-year East Asian climate reconstruction *npj Clim. Atmos. Sci.* **7** 131
- Mahato A 2021 Climate change and its impact on agriculture *Acta Hort. Regiotect.* **24** 50–57
- Muñoz-Sabater J *et al* 2021 ERA5-Land: a state-of-the-art global reanalysis dataset for land applications *Earth Syst. Sci. Data* **13** 4349–83
- Owen A B 2007 The pigeonhole bootstrap (<https://doi.org/10.1214/07-AOAS122>)
- Parry M L 2019 *Climate Change and World Agriculture* (Routledge)
- Ronneberger O, Fischer P and Brox T 2015 U-net: convolutional networks for biomedical image segmentation. paper presented at the international conference on medical image computing and computer-assisted intervention
- Rosu P, Bahng M, Jiang E, Zhu R and Tarokh V 2025 A PDE-informed latent diffusion model for 2-m temperature downscaling (arXiv:2510.23866)
- Saoulis A A, Lucas C, Lord N S, Addor N and Moraga J S 2025 Diffusion models for climate data surpass alternative statistical downscaling techniques *authorea preprints*
- Schoof J T 2013 Statistical downscaling in climatology *Geogr. Compass* **7** 249–65
- Spector J T, Masuda Y J, Wolff N H, Calkins M and Seixas N 2019 Heat exposure and occupational injuries: review of the literature and implications *Curr. Environ. Health Rep.* **6** 286–96
- Sun R and Zhang B 2016 Topographic effects on spatial pattern of surface air temperature in complex mountain environment *Environ. Earth Sci.* **75** 621
- Sun Y, Deng K, Ren K, Liu J, Deng C and Jin Y 2024 Deep learning in statistical downscaling for deriving high spatial resolution gridded meteorological data: a systematic review *ISPRS J. Photogramm. Remote Sens.* **208** 14–38
- Tomasi E, Franch G and Cristoforetti M 2025 Can AI be enabled to perform dynamical downscaling? A latent diffusion model to mimic kilometer-scale COSMO5. 0_CLM9 simulations *Geosci. Model Dev.* **18** 2051–78
- Wang F, Tian D, Lowe L, Kalin L and Lehrter J 2021 Deep learning for daily precipitation and temperature downscaling *Water Resour. Res.* **57** e2020WR029308
- Wang Z, Bovik A C, Sheikh H R and Simoncelli E P 2004 Image quality assessment: from error visibility to structural similarity *IEEE Trans. Image Process.* **13** 600–12
- Wilby R L, Charles S P, Zorita E, Timbal B, Whetton P and Mearns L O 2004 Guidelines for use of climate scenarios developed from statistical downscaling methods. Supporting material of the intergovernmental panel on climate change, available from the DDC of IPCC TG CIA, 27 p 27
- Wilby R L and Dawson C W 2013 The statistical downscaling model: insights from one decade of application *Int. J. Climatol.* **33** 1707–19
- Wong N H, Tan C L, Nindiyani A D S, Jusuf S K and Tan E 2012 Influence of water bodies on outdoor air temperature in hot and humid climate ICSDC 2011: integrating sustainability practices in the construction industry pp 81–89
- Xu Z, Han Y and Yang Z 2019 Dynamical downscaling of regional climate: a review of methods and limitations *Sci. China Earth Sci.* **62** 365–75
- Xue Y, Janjic Z, Dudhia J, Vasic R and De Sales F 2014 A review on regional dynamical downscaling in intraseasonal to seasonal simulation/prediction and major factors that affect downscaling ability *Atmos. Res.* **147** 68–85
- Yuan W, Hu S, Zhan C, Wang G and Luo Y 2025 Machine learning land surface temperature downscaling method based on Landsat 9 and Sentinel-2 satellite feature interaction *Geo-Spatial Inf. Sci.* 1–22
- Zhang H, Tang B-H, Hu T, Jiang Y and Li Z-L 2026 PGDM: physically guided diffusion model for land surface temperature downscaling *Remote Sens. Environ.* **340** 115432
- Zhang L, Lin X, Pan Z, Zhao X, Bhattarai A and Wu X 2025 Diffusion models and bias correction in statistical downscaling: a review and critical analysis. Paper presented at the 2025 18th international conference on advanced computer theory and engineering (ICACTE)
- Zhang Y, Li K, Li K, Wang L, Zhong B and Fu Y 2018 Image super-resolution using very deep residual channel attention networks *Paper presented at the European Conf. on Computer Vision*