

Evaluation of conditional diffusion for cross-sensor land surface temperature downscaling across African cities

Jangho Lee & Sara Shamekh

To cite this article: Jangho Lee & Sara Shamekh (14 Oct 2026): Evaluation of conditional diffusion for cross-sensor land surface temperature downscaling across African cities, International Journal of Remote Sensing, DOI: [10.1080/01431161.2026.2747322](https://doi.org/10.1080/01431161.2026.2747322)

To link to this article: <https://doi.org/10.1080/01431161.2026.2747322>



© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 14 Oct 2026.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)



Evaluation of conditional diffusion for cross-sensor land surface temperature downscaling across African cities

Jangho Lee and Sara Shamekh

Courant Institute School of Mathematics, Computing, and Data Science, New York University, New York, NY, USA

ABSTRACT

Fine-resolution land surface temperature (LST) is needed to characterize urban thermal patterns, but satellite observations typically trade spatial detail for temporal coverage. This study evaluates whether conditional diffusion can reconstruct ECOSTRESS-scale LST from coarse-resolution METEOSAT observations and fine-resolution environmental descriptors. Paired ECOSTRESS and METEOSAT scenes from six African cities were used for model development and retrospective evaluation. The 212 development-city evaluation scenes from 2022 were excluded from training but contributed to validation and checkpoint selection, while Johannesburg was designated as the geographic holdout and did not contribute to training or normalization. Each model predicts the residual between ECOSTRESS and aligned METEOSAT LST and adds this correction to the observed coarse thermal field. Conditional diffusion using 50-step DDIM sampling was compared with SRCNN, a deterministic U-Net, and the uncorrected METEOSAT input using radiometric, structural, multiscale, and spatial-consistency diagnostics. Predictions were reconstructed without ECOSTRESS-derived clipping. Across the 212 scenes, DDIM-50 with a fixed 10-member ensemble reduced mean RMSE from 5.143 K for METEOSAT to 4.936 K and median RMSE from 3.838 to 3.482 K. The paired mean contrast is -0.207 K (95% interval -0.400 to -0.011); after excluding two scenes flagged by a reference-only quality-control rule, the paired contrast retains the same direction. Ensemble averaging improved pixel-wise stability but reduced part of the fine-scale variability. These results show that the no-clipping comparison supports a small average radiometric improvement from the fixed-member diffusion product while making the trade-off between stable temperature reconstruction and preservation of local spatial detail explicit.

ARTICLE HISTORY

Received 27 July 2026
Accepted 4 October 2026

KEYWORDS

Urban meteorology;
diffusion model;
downscaling

1. Introduction

Urban thermal environments are spatially heterogeneous. Vegetation, impervious surfaces, water bodies, surface albedo, topography, building density, and urban morphology

CONTACT Jangho Lee ✉ jangho.lee@nyu.edu Courant Institute School of Mathematics, Computing, and Data Science, New York University, 251 Mercer Street, New York, NY 10012, USA

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

interact to produce localized hot and cool areas, terrain-related gradients, and urban-rural contrasts. Many of these patterns occur at spatial scales much finer than those resolved by kilometre-scale thermal products, yet they are central to understanding how land cover and urban form influence surface heating within and around cities (Cilek and Cilek 2021; Liao, Hong, and Heo 2021; Rahimi, Barghjelveh, and Dong 2021; Schwaab et al. 2021; Yang et al. 2020). Fine-resolution LST therefore provides an important spatial layer for urban climate analysis, land-atmosphere studies, ecological monitoring, and the assessment of thermal contrasts associated with vegetation, water, terrain, bare ground, and the built environment (Sobrino et al. 2012; Zhou et al. 2018).

The need for spatially detailed thermal information is particularly evident in rapidly changing and observationally sparse urban regions. African cities encompass humid tropical, high-elevation, semi-arid, and dryland environments, and their thermal patterns are shaped by very different combinations of terrain, vegetation, water availability, exposed soil, and urban development. At the same time, spatially dense in situ meteorological and environmental monitoring remains limited in many locations. Satellite observations consequently provide one of the few spatially consistent sources for comparing surface thermal structure across multiple cities.

A persistent limitation of satellite thermal remote sensing is the trade-off between spatial resolution and temporal availability. High-resolution instruments can resolve neighbourhood-scale variability but are constrained by orbital sampling and retrieval availability. ECOSTRESS, for example, provides LST at approximately 70 m resolution and can reveal detailed thermal contrasts, but its International Space Station orbit produces irregular coverage for any individual city, with no fixed revisit interval (Chang et al. 2025; Fisher et al. 2020). Geostationary observations occupy the opposite end of this trade-off. METEOSAT provides frequent and spatially extensive thermal monitoring over Africa, but its kilometre-scale spatial support cannot resolve many local urban and environmental contrasts (Atitar and Sobrino 2008; Trigo et al. 2008).

LST downscaling seeks to bridge this scale mismatch by combining a coarse thermal observation with finer-resolution information about the land surface. Statistical and semi-empirical approaches have related coarse LST to vegetation, built-up intensity, water, albedo, elevation, and other environmental predictors. Furthermore, machine-learning and deep-learning methods have been used to represent nonlinear relationships between thermal conditions and surface properties (Bah et al. 2022; Bechtel, Zakšek, and Hoshyaripour 2012; Bonafoni 2016; Hutengs and Vohland 2016; Lee et al. 2024; W. Li et al. 2019; Mukherjee, Joshi, and Garg 2014, 2015; Yoo et al. 2020). These studies demonstrate that auxiliary variables contain substantial information about sub-pixel LST variability. They also show that performance depends on environmental context and that a model can achieve a favourable mean error while still smoothing the local amplitude or spatial arrangement that motivates downscaling. A closely related body of research addresses the same spatial-temporal trade-off through multi-sensor and spatiotemporal fusion (Gao et al. 2006; Weng, Fu, and Gao 2014; X. Zhu et al. 2010, 2016).

Image super-resolution provides a related methodological foundation. Synthetic degradation experiments derive a low-resolution input from the target image, so the degradation process and input-target correspondence are known. Cross-sensor reconstruction must also accommodate differences in spatial support, acquisition time, viewing geometry, retrieval characteristics, and missing-data patterns between METEOSAT and

ECOSTRESS. It estimates a fine-resolution field when the corresponding high-resolution observation is unavailable, whereas synthetic experiments recover a target from an input derived from that same target. This distinction motivates evaluation with paired observations and a path to reference-independent inference.

Most conventional statistical and neural downscaling models produce a single deterministic estimate for a given set of inputs. Such point estimates can minimize average loss while attenuating high-frequency variability when several fine-resolution configurations are compatible with the same coarse observation. Conditional diffusion offers a different formulation by learning an iterative denoising process for the residual field conditioned on the observed coarse LST and auxiliary predictors (Ho, Jain, and Abbeel 2020; Saharia et al. 2022). Diffusion-based methods are increasingly used in satellite-image generation and remote sensing super-resolution (Wang and Sun 2025). However, it remains necessary to determine whether the generated detail improves agreement with independently observed fine-scale thermal structure, whether those improvements extend beyond pixel-wise error, and how ensemble averaging changes the balance between local amplitude and smoothness.

This study asks three questions: whether conditional diffusion improves observation-based METEOSAT-to-ECOSTRESS downscaling; whether any gain extends beyond pixel-wise error to spatial structure, local amplitude, and distributional extremes; and how ensemble averaging changes those properties. These questions focus the analysis on the features that motivate fine-resolution LST reconstruction, including terrain gradients, localized hot and cool areas, and contrasts associated with vegetation, water, bare soil, and the built environment.

We formulate METEOSAT-to-ECOSTRESS downscaling as conditional residual reconstruction. Each model predicts the fine-scale correction from the aligned METEOSAT field and high-resolution environmental descriptors, then adds that correction to the coarse thermal input. Conditional diffusion is compared with SRCNN and DetUNet using radiometric error, structural similarity, distributional tails, multiscale coefficient behaviour, perturbation tests, and physical consistency checks. This design distinguishes broad thermal consistency, pixel-wise accuracy, local structure, spatial amplitude, and extreme temperatures. A central contribution of this study is the complementary evaluation framework, which distinguishes radiometric agreement, local structure, distributional extremes, multiscale variability, and physical consistency rather than ranking models by a single score.

2. Data and methods

2.1. Study area

The development domains comprise Addis Ababa (Ethiopia), Kano (Nigeria), Kinshasa (Democratic Republic of the Congo), Nairobi (Kenya), Niamey (Niger), and Touba (Senegal). Together they represent pronounced differences in elevation, land cover, urban extent, and surrounding landscape. Addis Ababa and Nairobi provide elevated and topographically complex settings; Kinshasa provides a humid environment with prominent water and vegetation contrasts; and Kano, Niamey, and Touba provide dryland or semi-arid settings with extensive bare ground and

different urban-rural transitions. This environmental diversity is important because a downscaling relationship that is useful in one surface context may not transfer directly to another.

For each city, the analysis domain includes the urban core together with peri-urban and surrounding rural areas. The domains therefore contain both intra-urban variability and broader urban-rural contrasts rather than isolating only built-up pixels. [Figure 1](#) shows the city locations and summarizes the corresponding elevation and land-cover structure, including differences in terrain, built area, vegetation, water, cropland, rangeland, and bare ground.

2.2. Thermal remote sensing data

The study combines two thermal remote sensing products with complementary spatial and temporal characteristics: ECOSTRESS (Fisher et al. [2020](#); Hulley et al. [2021](#)) and METEOSAT-derived LST (Freitas et al. [2009](#); Schmetz et al. [2002](#)). ECOSTRESS supplies the fine-resolution satellite reference used to define and evaluate the target residual, whereas METEOSAT supplies the observed coarse-resolution thermal field that is available to every model at inference.

ECOSTRESS is a thermal infrared radiometer mounted on the International Space Station. Its approximately 70 m LST product resolves fine-scale surface patterns that are not visible in kilometre-scale observations, including terrain-related gradients, water-related cooling, urban-rural contrasts, and localized hot and cool features. Coverage of a specific city is nevertheless intermittent because of orbital overpass timing, swath geometry, clouds, and retrieval availability (Chang et al. [2025](#); Fisher et al. [2020](#)).

METEOSAT is a geostationary meteorological satellite system that provides frequent and spatially continuous thermal observations over Africa. The METEOSAT-derived LST field used here has a nominal spatial resolution of approximately 3 km at the sub-satellite point for the relevant SEVIRI channels and an effectively coarser footprint away from nadir. It can represent broad thermal evolution over time, but its native spatial support cannot resolve many neighbourhood-scale temperature contrasts (Atitar and Sobrino [2008](#); Trigo et al. [2008](#)).

Each ECOSTRESS scene is paired with a temporally matched METEOSAT field, which is mapped to the ECOSTRESS analysis grid and retained as both a model input and a no-learning baseline. These paired observations support learning the ECOSTRESS-referenced correction from coarse thermal conditions and fine-resolution environmental descriptors.

ECOSTRESS serves as the fine-resolution satellite reference. The target residual therefore combines sub-grid thermal variability with cross-sensor differences in retrieval, emissivity, viewing geometry, acquisition time, and spatial support; the reported errors quantify agreement with ECOSTRESS. Independent in situ measurements would be needed to separate absolute surface-temperature error from sensor differences.

[Figure 2](#) illustrates the complete input-target setting. ECOSTRESS supplies the fine-resolution reference LST, METEOSAT supplies the observed coarse thermal field, and the auxiliary layers provide spatial information on vegetation, built-up surfaces, water, albedo, elevation, and building morphology. The framework uses these layers jointly rather than attempting to infer fine-scale structure from the coarse thermal image alone.

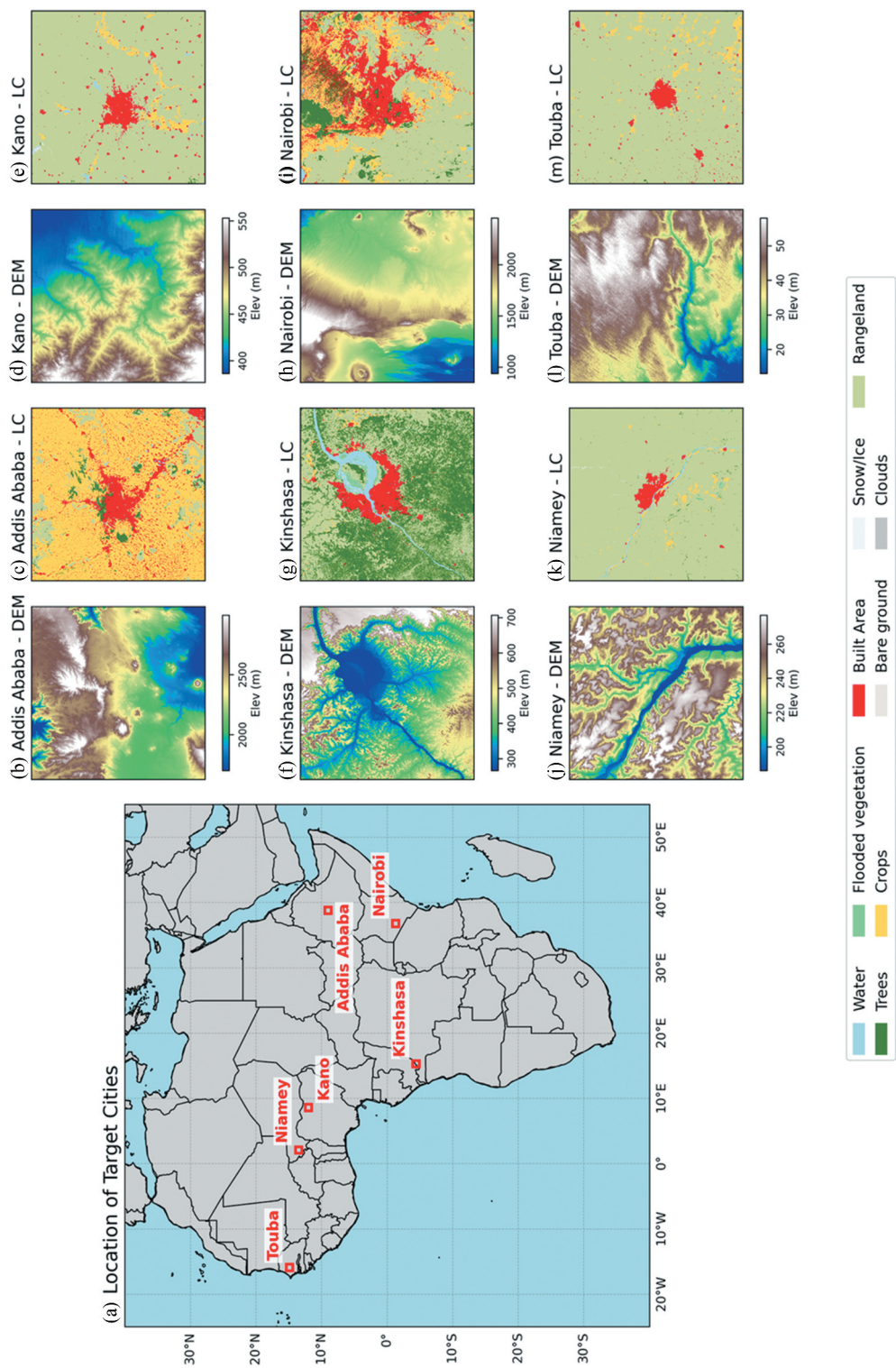


Figure 1. Study area and environmental heterogeneity of the target African cities. The map identifies the six development cities. Digital elevation model and land-cover panels show differences in terrain, built area, vegetation, water, cropland, rangeland, bare ground, and surrounding landscape context.

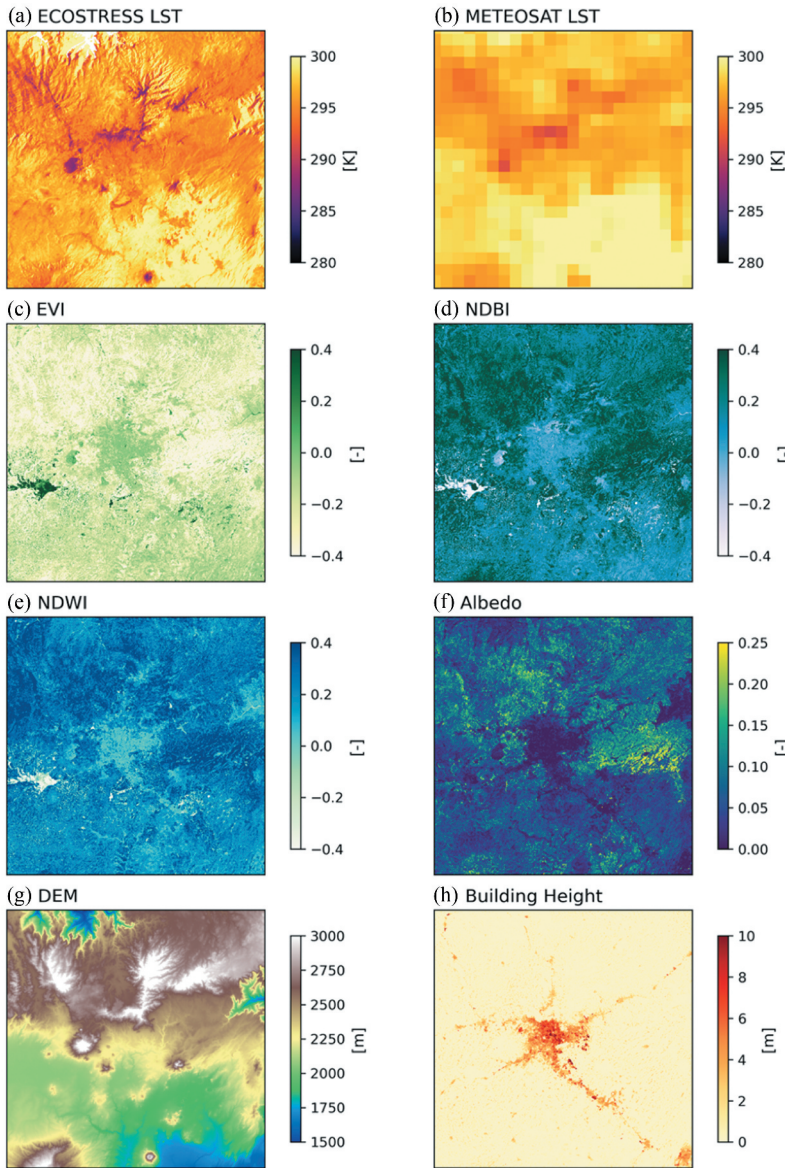


Figure 2. Example thermal and auxiliary input layers. ECOSTRESS supplies the fine-resolution satellite reference, METEOSAT supplies the observed coarse-resolution LST field, and the auxiliary predictors describe vegetation, built-up surfaces, water, albedo, terrain, and building-height structure. Optical descriptors are interpolated monthly composites rather than simultaneous acquisitions.

2.3. Auxiliary variables

The 13-channel predictor tensor contains the aligned METEOSAT LST field; Landsat-derived enhanced vegetation index (EVI), normalized difference built-up index (NDBI), normalized difference water index (NDWI), and albedo; elevation from the Copernicus GLO-30 digital elevation model; building fraction, mean building height, and building-

height standard deviation from the Global Building Atlas; and sine-cosine encodings of day of year and UTC hour of day. EVI represents vegetation condition, NDBI provides information on built-up and impervious surfaces, NDWI indicates water or moisture-related contrast, albedo represents surface reflectance, the DEM represents topographic control, and the building variables describe three-dimensional urban structure (Du et al. 2014; P. Li, Xiao, and Feng 2018; Simard et al. 2024; Trevisani, Skrypitsyna, and Florinsky 2023; Venter et al. 2020; X. X. Zhu et al. 2025).

The retained optical composite rasters have 100 m spacing before alignment to the 0.001° analysis grid. Elevation is spatially averaged to that grid; building polygons supply cell fractions and building-height moments before nearest-neighbour alignment. The nominal 10 m land-cover raster provides cartographic context and is not a model predictor. Re-gridding does not create independent information below the source spatial support.

The high-resolution predictors supply surface context that METEOSAT does not resolve, while METEOSAT anchors the scene's large-scale thermal state. Together they guide LST reconstruction at METEOSAT observation times using information on vegetation, water, the built environment, terrain, and seasonal and diurnal timing.

Temporal support differs among descriptors. Elevation is static, and building morphology is treated as time invariant using predominantly 2019 imagery with 2018 fallback. The optical descriptors are cloud-screened monthly median composites interpolated between nominal midpoints. Across the 215 daytime candidates, the nearer midpoint was a median of 8 days (maximum 15 days) from the thermal observation, and the farther midpoint was a median of 23 days (maximum 31 days). These offsets describe composite support rather than individual acquisition age, and missing months may draw on older observations. The mismatch matters most when vegetation, moisture, or land cover changes rapidly.

2.4. Data preprocessing and dataset split

All thermal and auxiliary layers are prepared in a common scene-level framework before patch extraction. For each ECOSTRESS acquisition, METEOSAT observations within two hours before and two hours after the ECOSTRESS overpass are considered. When valid METEOSAT observations are available on both sides of the overpass, the field is linearly interpolated to the ECOSTRESS acquisition time. When only one valid observation is available within the window, the nearest field is used. If no valid METEOSAT observation is available, the corresponding area is excluded through the valid-observation mask.

The archive retains the matched METEOSAT field used for each paired scene. Neighbouring hourly fields are unavailable for a separate interpolation-sensitivity test, so rapid morning transitions remain a source of uncertainty.

The temporally matched METEOSAT field and every auxiliary predictor are spatially aligned to a common EPSG:4326 analysis grid with 0.001° spacing, approximately 111 m north-south. Thermal and optical fields use nearest-neighbour mapping, while elevation is spatially averaged. This grid spacing differs from the native approximately 70 m ECOSTRESS product. Pixels affected by ECOSTRESS cloud masks, incomplete swath coverage, retrieval gaps, missing METEOSAT values, or unavailable predictor values are excluded from the common finite evaluation support. Training patches require complete

Table 1. Collocated scene inventory by city and UTC year, and 2022 daytime eligibility.

City	2018	2019	2020	2021	2022	2023	2024	2025	2022 day/used
Addis Ababa	3	35	48	50	53	59	64	68	26/26
Kano	9	55	78	67	54	67	54	88	44/44
Kinshasa	0	28	26	33	26	19	24	31	21/21
Nairobi	5	30	70	75	63	63	67	90	34/32
Niamey	9	53	93	66	72	84	57	103	56/56
Touba	7	6	13	35	47	64	49	92	34/33
Johannesburg	22	94	100	114	90	99	63	100	43/43

data, whereas prediction windows require at least 90% finite values in ECOSTRESS and each required spatial predictor; invalid cells are excluded from scoring.

The retained collocated archive spans July 2018 through December 2025 and contains 3,034 scenes: 2,352 from the six development cities and 682 from Johannesburg (South Africa), which is designated as the geographic holdout. The intended scene-level roles assign 2020 to validation, 2022 to testing, and the other years to training. [Table 1](#) summarizes the city-year inventory and reports the numbers of daytime 2022 candidates and eligible scenes.

The retained arrays differ from the intended scene-level partition. No development-city scene from 2022 entered the stored diffusion training arrays; 405,687 development-city 2022 patches entered the stored validation arrays and therefore contributed to check-point selection. The Johannesburg evaluation scenes did not enter training or contribute to normalization and are treated as the geographic holdout. Diffusion used the complete stored arrays, whereas SRCNN and DetUNet used fixed seed-42 quarter-samples drawn from those arrays. Accordingly, the 2022 development-city comparison is retrospective rather than an untouched test, while Johannesburg provides the geographic-holdout analysis.

A window-based strategy converts valid portions of each scene into local samples. Full ECOSTRESS scenes often contain large invalid regions because of clouds, incomplete swaths, or unavailable predictors. Treating a full scene as one sample would therefore discard many usable local observations and would be computationally inefficient. Overlapping patches increase the number and diversity of local examples and allow the models to learn how the coarse thermal state and environmental context relate to the ECOSTRESS-referenced residual at neighbourhood scales.

Each valid patch contains 64×64 pixels, corresponding to approximately 7.1 km north–south on the analysis grid. Training patches are extracted with a stride of 16 pixels in the development cities. Inference uses the same 64×64 window size with a stride of 32 pixels, after which the patch predictions are combined to reconstruct the scene.

[Figure 3](#) illustrates prediction-style masking and patch selection. Training patches require complete data, whereas prediction windows allow limited missing support. The effective number of windows therefore varies with cloud and swath coverage.

2.5. Problem formulation

The downscaling task is formulated as residual correction rather than direct prediction of absolute LST. Let T_{ECO} denote the ECOSTRESS reference LST and $T_{\text{MET}}^{\uparrow}$ denote the METEOSAT-derived LST after spatial alignment to the ECOSTRESS grid. The superscript $\uparrow$

Patching strategy

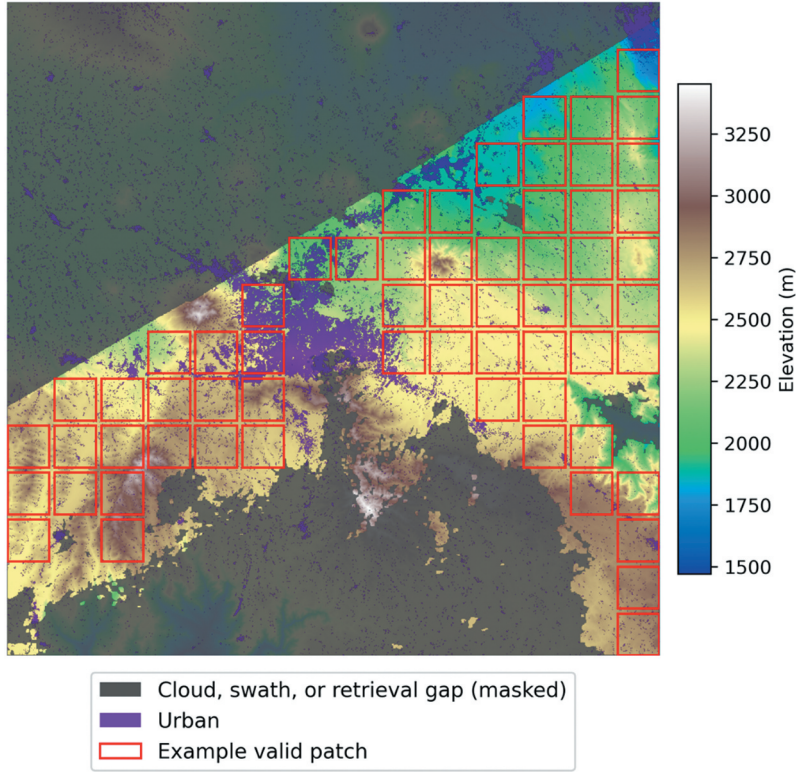


Figure 3. Window-based valid-patch extraction. The illustrative windows allow at most 10% missing values and use a 72-pixel stride; training patches instead require complete data. Masked areas denote cloud, swath, or retrieval gaps. Urban overlays provide visual context and do not filter the windows.

indicates that the coarse METEOSAT field has been mapped to the ECOSTRESS analysis grid before comparison. The raw residual target is defined as

$$\Delta T = T_{\text{ECO}} - T_{\text{MET}}^{\uparrow} \quad (1)$$

For each scene, the model predicts a residual from the aligned METEOSAT field and auxiliary predictors. ECOSTRESS defines the supervised target and evaluation reference. At inference, the predicted residual is converted to kelvin and added to METEOSAT without clipping to the ECOSTRESS temperature range or to any other output bound. ECOSTRESS temperature values are used for scoring but do not truncate or transform the reconstructed field.

Normalization means and standard deviations were calculated from all finite pixels in the stored training arrays, separately for the residual target and each predictor channel. These statistics include the retained training rows but exclude all development-city 2022 rows and the Johannesburg evaluation rows. The stored values were applied unchanged during validation and inference. To place the residual target on a common dimensionless scale for model training, the scene-specific residual ΔT is standardized using its stored-training mean and standard deviation:

$$x_0 = \frac{\Delta T - \mu_{\Delta T}}{\sigma_{\Delta T}} \quad (2)$$

where x_0 is the normalized clean residual, and $\mu_{\Delta T}$ and $\sigma_{\Delta T}$ are the global stored training mean and standard deviation of ΔT . For each predictor channel c ,

$$\tilde{X}_c = \frac{X_c - \mu_c}{\sigma_c} \quad (3)$$

where X_c is the original predictor field for channel c , and μ_c and σ_c are the corresponding global stored training mean and standard deviation for that predictor channel. The normalized multi-channel predictor stack is denoted by $\tilde{X}$. After a model predicts a normalized residual, the prediction is converted back to kelvin as

$$\widehat{\Delta T} = \sigma_{\Delta T} \hat{X}_0 + \mu_{\Delta T} \quad (4)$$

and the final high-resolution LST estimate is reconstructed as

$$\hat{T}_{HR} = T_{MET}^{\dagger} + \widehat{\Delta T} \quad (5)$$

where $\hat{X}_0$ is the predicted normalized residual, $\widehat{\Delta T}$ is the predicted residual in K, and $\hat{T}_{HR}$ is the final downscaled LST. This reconstruction step is shared by all learned models and keeps target definition and unit conversion consistent. Residual addition retains a coarse thermal baseline but does not enforce conservation after aggregation back to coarse support.

2.6. Downscaling models

2.6.1. Conditional diffusion model

The proposed model is a conditional diffusion model trained on normalized residual patches. Let m index a 64×64 patch in a mini-batch and let $\tilde{X}_m$ denote the normalized predictor stack for that patch. The diffusion model receives two types of input: a noisy residual field and the normalized predictor stack. The noisy residual carries the current diffusion state, while the predictor stack provides the conditioning information from METEOSAT and the auxiliary variables. The model outputs a prediction of the clean normalized residual.

Let T denote the total number of diffusion steps. In this study, T is set to 1000. For patch m , a diffusion step $t_m \in \{1, \dots, T\}$ is sampled, and Gaussian noise is added to the clean residual target $x_{0,m}$ to create a noisy residual $x_{t_m,m}$:

$$x_{t_m,m} = \sqrt{\bar{\alpha}_{t_m}} x_{0,m} + \sqrt{1 - \bar{\alpha}_{t_m}} \epsilon_m, \epsilon_m \sim \mathcal{N}(0, I) \quad (6)$$

Here, $x_{t_m,m}$ is the normalized residual patch after noise has been added at diffusion step t_m , ϵ_m is a Gaussian noise patch with the same spatial shape as $x_{0,m}$, and I is the identity covariance. The scalar $\bar{\alpha}_{t_m}$ controls how much of the clean residual remains at step t_m . It is defined from the diffusion noise schedule as

$$\alpha_t = 1 - \beta_t, \bar{\alpha}_t = \prod_{s=1}^t \alpha_s \quad (7)$$

where β_t is the noise variance at diffusion step t , α_t is the signal-retention factor, and $\bar{\alpha}_t$ is the cumulative signal-retention factor from step 1 to t . A cosine schedule is used for β_t , with offset 0.008 and noise variances clipped to $[10^{-4}, 0.9999]$, gradually increasing the noise level across diffusion steps.

The conditional diffusion network is denoted by f_θ , where θ represents its learnable parameters. The architecture is a U-Net-style encoder-decoder with residual convolutional blocks, skip connections, sinusoidal timestep embeddings, and a self-attention module at the bottleneck. The noisy residual and predictor stack are concatenated at the input, and timestep information is injected through the residual blocks. The network predicts the clean normalized residual directly:

$$\hat{x}_{0,m} = f_\theta(x_{t_m,m}, t_m, \tilde{X}_m) \quad (8)$$

Here, the inputs to f_θ are the noisy residual $x_{t_m,m}$, the diffusion step t_m , and the normalized predictor stack $\tilde{X}_m$. The output $\hat{x}_{0,m}$ is the predicted clean normalized residual. This is a direct x_0 -prediction formulation, not a noise-prediction formulation.

The diffusion stem uses 3×3 kernels and 64 channels. Encoder widths are 64/128/256, with one residual block containing two 3×3 convolutions per stage, eight-group normalization, and SiLU. Four-head self-attention lies between two bottleneck residual blocks; the sinusoidal timestep embedding maps 64 to 256 dimensions. The decoder uses 2×2 transposed convolutions and a 1×1 output head. Dropout is absent.

The diffusion training objective includes the expectation over randomly sampled diffusion steps and Gaussian noise:

$$\mathcal{L}_{diff}(\theta) = \frac{1}{4096} \mathbb{E}_{m,t_m,\epsilon_m} \left[\|x_{0,m} - f_\theta(x_{t_m,m}, t_m, \tilde{X}_m)\|_2^2 \right] \quad (9)$$

where the expectation is taken over sampled training patches, diffusion steps, and Gaussian noise. In practice, this expectation is approximated by the mini-batch mean over N patches, and $\|\cdot\|_2^2$ denotes the summed squared residual error over all pixels in the patch. The factor $1/4096$ converts this sum to a per-pixel mean. At each training iteration, t_m is drawn uniformly from $\{1, \dots, T\}$ and independent Gaussian noise $\epsilon_m \sim \mathcal{N}(0, I)$ is sampled for each patch to form $x_{t_m,m}$. An exponential moving average (EMA) of the model weights, with decay 0.995, is used during validation and prediction to stabilize the generated outputs.

During inference, the trained reverse process generates one or more normalized residual samples for each scene. Let $\hat{x}_{0,m}^{(k)}$ denote the k -th generated normalized residual sample for patch m . Each generated sample is converted to physical units and added to METEOSAT:

$$\hat{T}_{HR,m}^{(k)} = T_{MET,m}^\uparrow + \left(\sigma_{\Delta T} \hat{x}_{0,m}^{(k)} + \mu_{\Delta T} \right) \quad (10)$$

where $T_{MET,m}^\uparrow$ is the aligned METEOSAT LST patch and $\hat{T}_{HR,m}^{(k)}$ is the high-resolution LST prediction corresponding to diffusion sample k . Because each diffusion realization begins from a sampled Gaussian field, multiple realizations can be averaged to form an ensemble prediction. To avoid ambiguity between the sample index and the ensemble size, let K_{ens} denote the number of diffusion samples included in the ensemble:

$$\hat{T}_{HR,m}^{(K_{ens})} = \frac{1}{K_{ens}} \sum_{k=1}^{K_{ens}} \hat{T}_{HR,m}^{(k)}, K_{ens} \in \{1, 2, 5, 10, 15, 25\} \quad (11)$$

Member seeds are declared before evaluation and member identity is preserved through patch stitching. $K = 1$ uses the first fixed member, $K = 5$ averages the first five, and $K = 10$ averages all ten. No member is ranked or selected using ECOSTRESS. The fixed-member DDIM-50 protocol is used for the all-scene comparison and all displayed diffusion fields.

DDIM-50 (Song, Meng, and Ermon 2020) accelerates inference from the trained diffusion model by sampling 50 selected points from the 1,000-step cosine schedule. Ancestral DDPM adds Gaussian noise during every nonfinal reverse step, whereas DDIM with $\eta=0$ is deterministic for a fixed initial Gaussian field; different seeds still produce distinct ensemble members. Reducing each draw from 1,000 to 50 network evaluations makes fixed-member analysis across the full scene set practical without retraining. The reversed schedule spans indices 0–999 and returns the predicted clean residual at the final step.

Each scene uses ten fixed Gaussian seeds declared before inference. The $K = 10$ mean is the primary diffusion product, while $K = 1$ and $K = 5$ characterize how averaging changes the selected-scene reconstructions.

2.6.2. SRCNN baseline

SRCNN is used as a compact deterministic convolutional baseline. It receives the same normalized predictor stack $\tilde{X}_m$ as input and predicts one normalized residual field. The SRCNN mapping is denoted by g_{ϕ}^{SRCNN} , where ϕ represents the SRCNN model parameters:

$$\hat{x}_{0,m}^{SRCNN} = g_{\phi}^{SRCNN}(\tilde{X}_m) \quad (12)$$

Here, $\hat{x}_{0,m}^{SRCNN}$ is the SRCNN-predicted normalized residual for patch m . The model is deterministic, meaning that a given input patch always produces one fixed output. The SRCNN architecture uses an initial convolution, a sequence of residual convolutional blocks, and a final convolutional head. If $h_{m,0}$ denotes the initial feature map and B_{ℓ} denotes residual block ℓ , the residual-block update can be summarized as

$$h_{m,\ell} = h_{m,\ell-1} + \lambda B_{\ell}(h_{m,\ell-1}), \ell = 1, \dots, L \quad (13)$$

where $h_{m,\ell}$ is the feature map after residual block ℓ , L is the number of residual blocks, and λ is a residual scaling factor. In this implementation, $L = 4$ and $\lambda = 0.1$. The predicted residual is then produced by the final convolutional layer.

SRCNN uses width 32 and 3×3 convolutions in its stem, four residual blocks, and output head. Each residual block uses LeakyReLU(0.2), Dropout2d(0.1), and residual scaling 0.1. After the four blocks, the stem feature is added to the block output, followed by a 32-channel 3×3 refinement convolution, LeakyReLU(0.2), and a 3×3 one-channel output convolution.

SRCNN is trained using a mean absolute error objective on the normalized residual target:

$$\mathcal{L}_{SRCNN}(\phi) = \frac{1}{4096N} \sum_{m=1}^N \|x_{0,m} - g_{\phi}^{SRCNN}(\tilde{X}_m)\|_1 \quad (14)$$

where $\|\cdot\|_1$ denotes the sum of absolute residual errors over all pixels in the patch. Division by 4096 gives the per-pixel mean, as also used for DetUNet. After prediction, the normalized residual is converted to physical units and added to METEOSAT using the same reconstruction procedure as the diffusion model:

$$\hat{T}_{HR,m}^{SRCNN} = T_{MET,m}^{\dagger} + \left(\sigma_{\Delta T} \hat{x}_{0,m}^{SRCNN} + \mu_{\Delta T} \right) \quad (15)$$

This baseline tests whether a lightweight deterministic convolutional model can learn fine-scale residual corrections from the multi-source predictor stack.

2.6.3. Deterministic U-Net baseline

Deterministic U-Net (DetUNet) is used as a stronger deterministic image-to-image regression baseline. It belongs to the same U-Net-style encoder-decoder family as the diffusion backbone and also uses residual blocks and self-attention to represent multi-scale spatial structure. However, unlike the diffusion model, DetUNet does not use a noisy residual input or timestep embedding; it maps the predictor stack directly to one deterministic residual estimate. The DetUNet mapping is denoted by g_{ψ}^{UNet} , where ψ represents the model parameters:

$$\hat{x}_{0,m}^{UNet} = g_{\psi}^{UNet}(\tilde{X}_m) \quad (16)$$

The U-Net structure combines encoder features, bottleneck features, and decoder features through skip connections. Let E_1 , E_2 , and E_3 denote encoder blocks; let A denote the attention operation at the deepest encoder stage; and let D_2 , D_1 , and D_0 denote decoder blocks. A simplified representation of the forward pass is:

$$e_{m,1} = E_1(\tilde{X}_m), e_{m,2} = E_2(e_{m,1}), e_{m,3} = E_3(e_{m,2}) \quad (17)$$

$$b_m = A(e_{m,3}) \quad (18)$$

$$d_{m,2} = D_2([b_m, e_{m,3}]), d_{m,1} = D_1([d_{m,2}, e_{m,2}]),$$

$$d_{m,0} = D_0([d_{m,1}, e_{m,1}]) \quad (19)$$

$$\hat{x}_{0,m}^{UNet} = W_o * d_{m,0} + b_o \quad (20)$$

In these equations, $e_{m,1}$, $e_{m,2}$, and $e_{m,3}$ are encoder feature maps; b_m is the bottleneck feature map after attention; $d_{m,2}$, $d_{m,1}$, and $d_{m,0}$ are decoder feature maps; $[\cdot, \cdot]$ denotes channel-wise concatenation through skip connections; W_o and b_o are the final convolution weights and bias; and the asterisk symbol denotes convolution. This structure allows the model to combine local detail with broader spatial context.

DetUNet uses a 7×7 stem, encoder widths 32/64/128, eight-group normalization, LeakyReLU(0.2), and Dropout2d(0.3). Average pooling and bilinear upsampling connect the stages, and a 1×1 head predicts the residual. The bottleneck expands to 256 channels; four-head attention acts at the deepest encoder and first 64-channel decoder stage.

Table 2. Configuration and data exposure of the retained neural protocols.

Protocol item	Conditional diffusion	SRCNN	Deterministic U-Net
Parameters	4,995,713	4,387,297	5,087,457
Core network	Residual U-Net; widths 64/128/256; bottleneck attention	Width-32 CNN; four residual blocks	U-Net; widths 32/64/128; attention
Training objective	Normalized-residual MSE; direct x_0 prediction	Normalized-residual MAE	Normalized-residual MAE
Optimizer and learning rate	AdamW; 2×10^{-4} constant	AdamW; 3×10^{-4} constant	AdamW; OneCycle maximum 5×10^{-4}
Weight decay; batch size	10^{-4} ; 64	10^{-3} ; 128	10^{-2} ; 128
Maximum epochs / patience / best epoch	500 / 20 / 45	500 / 30 / 42	100 / 15 / 19
Loader rows training / validation	2,158,561 / 825,872	539,641 / 206,468	539,641 / 206,468
Reconstruction	EMA 0.995; DDIM-50 in main comparison	Deterministic forward pass; eight-way averaging	Deterministic forward pass; eight-way averaging

DetUNet is trained using a mean absolute error objective on the normalized residual target:

$$\mathcal{L}_{\text{UNet}}(\psi) = \frac{1}{4096N} \sum_{m=1}^N \|x_{0,m} - g_{\psi}^{\text{UNet}}(\tilde{x}_m)\|_1 \quad (21)$$

The final physical LST prediction is reconstructed as

$$\hat{T}_{HR,m}^{\text{UNet}} = T_{MET,m}^{\uparrow} + \left(\sigma_{\Delta T} \hat{x}_{0,m}^{\text{UNet}} + \mu_{\Delta T} \right) \quad (22)$$

DetUNet is included as the primary deterministic deep-learning baseline because it has greater capacity to represent multi-scale spatial structure than SRCNN while still producing a single deterministic residual estimate for each input patch.

Table 2 summarizes model size, optimization, data exposure, and reconstruction. Because the objectives and training schedules differ, the results compare the retained end-to-end protocols.

Best epochs come from the retained training histories. Deterministic row counts follow the original seed-42 shuffle-then-quarter rule, and retained validation includes development-city observations from 2022. All models use the same 13 predictors and output one residual field; the spatial and quantitative diagnostics use fixed-member DDIM-50 predictions without output clipping.

Checkpoint selection uses EMA normalized-residual MSE for diffusion and normalized-residual MAE for the deterministic models. Fixed integer seeds reproduce the DDIM and bootstrap analyses, with effective scene seeds retained in the manifests. Table 2 records the retained epochs and stopping settings. Hardware identifiers and complete random-number-generator states for the historical training runs were not retained, so exact bitwise reproduction of those runs cannot be claimed.

2.7. Prediction and evaluation metrics

2.7.1. Evaluation design, reconstruction, and sampling

The primary retrospective evaluation uses 212 eligible development-city daytime scenes from 2022. A separate geographic-holdout analysis uses 43 eligible Johannesburg scenes that did not enter training or normalization. Spatial diagnostics use 12 development-city

and two Johannesburg cases selected for high within-group coverage. Table 1 gives the full candidate and eligibility counts.

All methods are scored on common finite support within each scene. METEOSAT is the no-learning baseline; RMSE, MAE, bias, and total WaveSim are computed for METEOSAT, SRCNN, DetUNet, and DDIM-50 K = 10. Scene metrics receive equal weight so greater valid coverage does not dominate. For the 12 spatial-diagnostic scenes, approximate coarse regions are inferred from the block structure of the nearest-neighbour-mapped METEOSAT field. ECOSTRESS, METEOSAT, and DDIM-50 K = 10 are averaged over identical common finite pixels within each region; regions with fewer than 25 common pixels are omitted. Bias, RMSE, and Pearson correlation are calculated across regions within each scene and then averaged with equal scene weight. These inferred regions approximate rather than reproduce native sensor footprints.

Scenes are divided into 64×64 windows. Patch predictions are combined with the normalized outer product of two 64-point Tukey windows (taper 0.5), reducing overlap-edge artefacts; deterministic models also use eight geometric test-time transforms. The stitched prediction is

$$\hat{T}(u) = \frac{\sum_{m \in \mathcal{M}(u)} w_m(u) \hat{T}_m(u)}{\sum_{m \in \mathcal{M}(u)} w_m(u)} \quad (23)$$

where the weighted sum includes all patches covering each valid pixel.

Inference uses fixed, outcome-independent member seeds and applies no output clipping. SRCNN and DetUNet each yield one stitched field; DDIM-50 K = 10 averages ten stitched members whose identities remain fixed across patches. METEOSAT and every learned method are scored on the same finite within-scene support.

A matched sampler check uses the same 12 development-city scenes, retained patch coordinates, initial Gaussian fields, finite scoring support, equal scene weighting, and absence of clipping for ancestral DDPM-1000 and DDIM-50 at K = 1, 5, and 10. Both samplers use one model object and forward path, and each matched member pair receives the same initial Gaussian field. K = 1 uses member 01; K = 5 and K = 10 are pixelwise means of fixed stitched members 01–05 and 01–10. The retained ECOSTRESS-and-predictor availability rule defines the evaluation windows, but ECOSTRESS temperature values do not enter either sampler. This design compares the reverse samplers while holding member identity, initialization, support, and aggregation fixed.

2.7.2. Pixel-wise, structural, and distributional metrics

Model performance is computed against ECOSTRESS over the common valid mask. Equation (24) defines signed error as predicted minus reference LST at each valid pixel:

$$e_i = \hat{T}_i - T_i \quad (24)$$

Bias is the mean signed error:

$$\text{Bias} = \frac{1}{n} \sum_{i=1}^n e_i \quad (25)$$

Root-mean-square error (RMSE) is:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n e_i^2} \quad (26)$$

Bias summarizes the mean warm or cool offset, whereas RMSE gives greater weight to large pixel errors. Positive bias denotes a warmer prediction and negative bias a cooler prediction. Bias closer to zero and lower RMSE indicate better radiometric agreement, but neither establishes whether fine-scale structures occur at the correct locations. These metrics are computed scene by scene before equal-scene aggregation.

Mean absolute error (MAE) is the mean absolute prediction error over the same mask and is reported in kelvin. Compared with RMSE, MAE is less sensitive to a small number of large errors.

The structural similarity index (SSIM) compares local luminance, contrast, and structure:

$$\text{SSIM}(T, \hat{T}) = \frac{(2\mu_T\mu_{\hat{T}} + C_1)(2\sigma_{T\hat{T}} + C_2)}{(\mu_T^2 + \mu_{\hat{T}}^2 + C_1)(\sigma_T^2 + \sigma_{\hat{T}}^2 + C_2)} \quad (27)$$

Local means, variances, and covariance enter Equation (27), while C_1 and C_2 stabilize the ratio. Higher SSIM, approaching one, indicates closer local structural agreement with ECOSTRESS.

SSIM uses the finite ECOSTRESS range, a 7×7 uniform window, sample covariance, data range = 1, $K_1 = 0.01$, and $K_2 = 0.03$, giving $C_1 = 0.0001$ and $C_2 = 0.0009$. Cells outside common support are filled with 0.5; a fully finite 64×64 sensitivity checks this support treatment.

Cold- and hot-tail biases compare the predicted and ECOSTRESS first and 99th percentiles, respectively:

$$\text{EB1} = Q_{0.01}(\hat{T}) - Q_{0.01}(T) \quad (28)$$

$$\text{EB99} = Q_{0.99}(\hat{T}) - Q_{0.99}(T) \quad (29)$$

EB1 and EB99 values closer to zero indicate better tail agreement. Positive values mean the predicted percentile is too warm and negative values mean it is too cool. These distributional summaries do not identify whether the corresponding hot or cool pixels occur at the correct locations.

2.7.3. WaveSim diagnostics

WaveSim provides a complementary multi-scale coefficient diagnostic (Accarino et al. 2025). A two-dimensional wavelet decomposition of ECOSTRESS and predicted fields gives scale-specific representations. The overall score combines magnitude, displacement, and coefficient-distribution components:

$$\text{WaveSim}(X, Y) = \sum_{s=1}^S w_s \left[M_s(\tilde{X}_s, \tilde{Y}_s)^a D_s(\tilde{X}_s, \tilde{Y}_s)^\beta S_s(\tilde{X}_s, \tilde{Y}_s)^\gamma \right] \quad (30)$$

Equation (30) combines four coefficient groups with weights that sum to one; its magnitude, displacement, and structure terms have separate exponents. We use a three-level two-dimensional db4 decomposition with symmetric boundaries. The three detail

levels and level-3 approximation each receive weight 0.25, all exponents equal one, detail-orientation scores are averaged before multiplication, and the stabilizer is 10^{-10} .

Magnitude similarity compares mean squared wavelet-coefficient energy:

$$M_s = 1 - \frac{|E(\tilde{X}_s) - E(\tilde{Y}_s)|}{E(\tilde{X}_s) + E(\tilde{Y}_s) + \varepsilon}, E(Z) = \frac{1}{|Z|} \sum_j Z_j^2 \quad (31)$$

For a group containing N coefficients, energy is their mean squared value and a small constant prevents division by zero. Magnitude similarity approaches one when predicted and reference energies agree, and decreases as their energy magnitudes diverge.

Displacement similarity compares row- and column-wise marginal energy distributions through Jensen–Shannon divergence with natural logarithms:

$$D_s = (1 - \text{JSD}(p_s^r, q_s^r))(1 - \text{JSD}(p_s^c, q_s^c)) \quad (32)$$

Reference and prediction row and column marginals are evaluated separately at each scale. Higher displacement similarity indicates closer marginal energy distributions but does not locate individual thermal features.

Structure similarity compares centred, normalized vectors of sorted coefficients:

$$S_s = \frac{1}{2}(1 + \cos(\bar{x}_s, \bar{y}_s))P_s \quad (33)$$

Cosine similarity is combined with a penalty that lowers the score when coefficient distributions differ strongly. Because sorting removes position, higher structure similarity indicates closer coefficient distributions rather than preservation of spatial arrangement.

The structure penalty is

$$P_s = \text{clip}\left\{1 - \frac{\|a^* - b^*\|_2/n}{\text{mean}(|a^*|) + \text{mean}(|b^*|) + \varepsilon}, 0, 1\right\} \quad (34)$$

In Equation (34), a^* and b^* are the sorted, flattened coefficient vectors and n is their common length; cosine similarity is bounded between -1 and 1 . Before decomposition, cells outside common finite support are filled with each field's valid-cell mean.

The metrics emphasize complementary properties: bias, RMSE, and MAE describe radiometric error; SSIM describes local structure; EB1 and EB99 describe distributional tails; and WaveSim describes multiscale coefficient behaviour. Because WaveSim's displacement and sorted-coefficient components can saturate without preserving exact feature location, it is interpreted with maps, controlled perturbations, radiometric and structural metrics, and a fully finite-block sensitivity.

2.7.4. Statistical comparisons

Statistical comparisons use paired within-scene differences. The primary DDIM-50 K = 10 minus METEOSAT RMSE contrast uses 50,000 fixed-seed scene bootstraps for 95% intervals. A sensitivity analysis excludes whole scenes using a reference-only rule based solely on the fraction of supported ECOSTRESS pixels outside 200–400 K; no learned prediction or model error enters the rule. Pixels, ensemble members, and overlapping patches are not treated as inferential replicates.

3. Results and discussion

3.1. Spatial detail and ensemble smoothing

Figure 4 establishes the spatial behaviour behind the aggregate results. In the example Addis Ababa case, the comparison separates the broad thermal state supplied by METEOSAT from the neighbourhood-scale pattern reconstructed by the learned methods using the same no-clipping DDIM-50 product evaluated quantitatively.

METEOSAT captures the broad warm-cool organization but retains a coarse-grid imprint and suppresses ridge-valley contrasts. SRCNN introduces limited sharpening, while DetUNet recovers more continuous local texture; both remain smoother than ECOSTRESS. The fixed-member DDIM-50 outputs express more of the ridge network, localized cool areas, and surrounding texture. Major patterns persist across $K = 1$, $K = 5$, and $K = 10$, which is consistent with the conditioning fields controlling much of the spatial organization. Averaging mainly changes local amplitude: $K = 1$ retains the strongest departures, and the larger composites progressively damp small-scale variability. The mapped fields use the same fixed-member DDIM protocol as the quantitative analysis and contain no output clipping.

The residual formulation helps explain this behaviour. Each learned method starts from the same aligned METEOSAT field and estimates a correction, so the coarse observation provides a common thermal background while the environmental descriptors organize

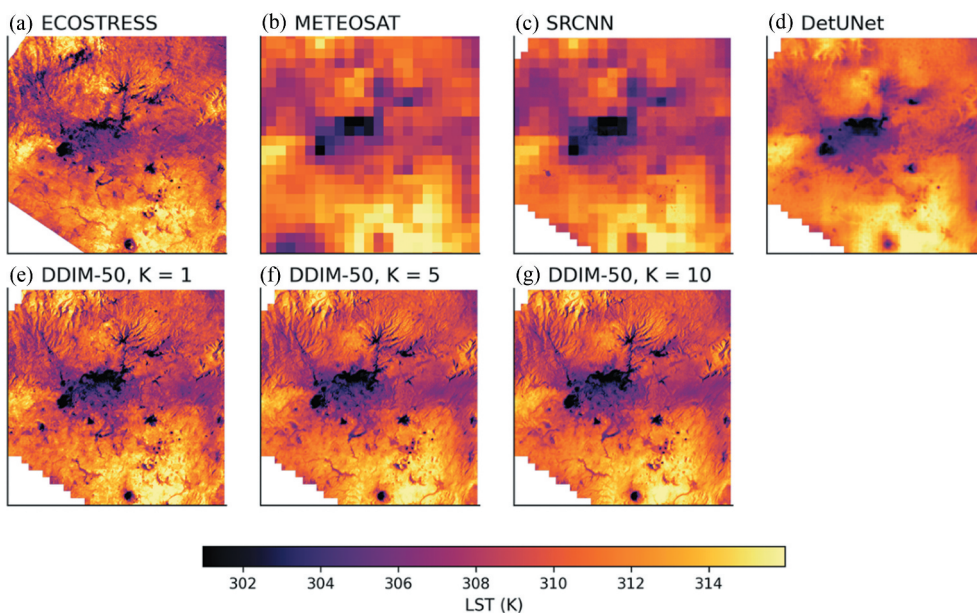


Figure 4. Qualitative comparison of ECOSTRESS, METEOSAT, deterministic baselines, and fixed-member DDIM-50 outputs for Addis Ababa on 6 April 2022 at 06:59:19 UTC. $K = 1$ uses the first predeclared member, $K = 5$ averages the first five, and $K = 10$ averages all ten after stitching. No learned panel uses ECOSTRESS-derived clipping; numerical comparisons use common finite support. The common 1st–99th percentile color limits are used only for display and do not alter the evaluated arrays or scores.

the sub-grid response. The sharper diffusion field therefore reflects a richer correction rather than replacement of the regional temperature pattern. This anchor promotes large-scale agreement but does not guarantee correct local amplitude, placement, or distribution, so the maps must be interpreted with radiometric, structural, and multiscale diagnostics.

The reconstruction checks support this reading at both patch and coarse scales. Tukey-weighted blending limits patch-boundary discontinuities, so the visible texture is not interpreted as a stitching artefact. The coarse-support comparison then distinguishes conservation of the observed background from movement towards the fine-resolution reference. After aggregation to the same inferred coarse regions, DDIM-50 $K = 10$ differs from METEOSAT by a bias of $+0.220$ K, RMSE of 1.903 K, and correlation of 0.889 . Against ECOSTRESS aggregated to those regions, its bias is $+0.677$ K, RMSE is 2.253 K, and correlation is 0.819 ; the corresponding METEOSAT-to-ECOSTRESS values are $+0.457$ K, 2.543 K, and 0.806 . The reconstruction therefore moves towards the aggregated reference in both RMSE and correlation on these cases, although exact coarse-scale conservation is not imposed and the regions approximate nearest-neighbour mapped sensor support.

Figure 5 isolates the effect of fixed-member DDIM-50 averaging across the 12 selected scenes without output clipping. Mean RMSE decreases from 3.113 K at $K = 1$ to 2.872 K at $K = 5$ and 2.842 K at $K = 10$, while total WaveSim falls from 0.691 to 0.462 . Because the same predeclared members are reused across K , these changes reflect averaging within the retained model. Averaging cancels member-specific departures and produces a more stable central estimate while suppressing part of the wavelet-scale amplitude that distinguishes individual realizations.

The spread-error analysis shows that this variance reduction should not be read as uncertainty calibration. The equal-scene central-90% interval covers 48.7% of ECOSTRESS

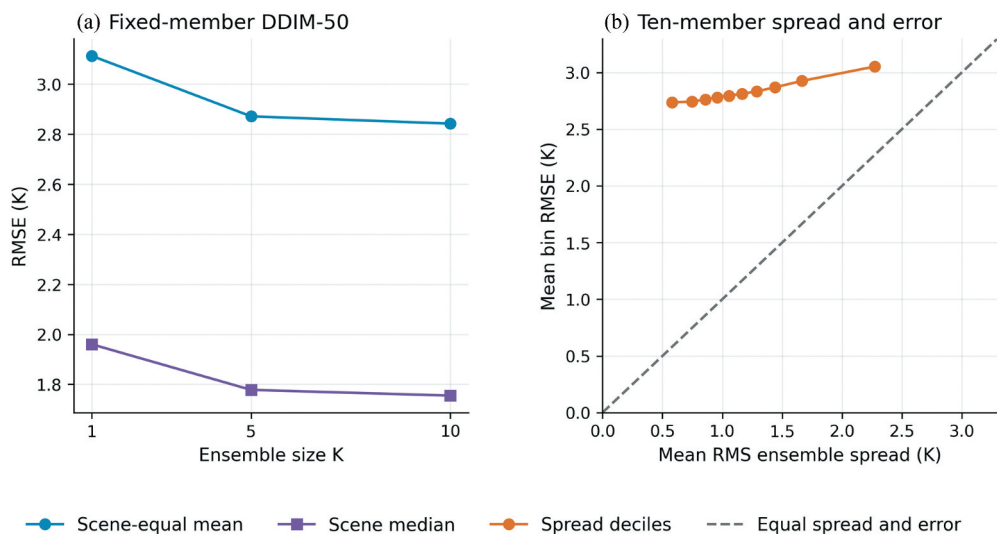


Figure 5. No-clipping DDIM-50 ensemble diagnostics on the 12 selected development-city scenes. (a) Equal-scene mean and median RMSE for fixed-member $K = 1, 5$, and 10 fields. (b) Equal-scene spread-decile comparison from the same ten stitched members; the dashed line denotes equal spread and error and is not a calibration claim.

values, so the ensemble characterizes sensitivity to the diffusion realization rather than calibrated prediction bounds. In the matched sampler comparison, 12-scene mean RMSE values are 3.241, 2.856, and 2.800 K for ancestral DDPM-1000 at $K = 1, 5$, and 10 , respectively; the corresponding DDIM-50 values are 3.113, 2.872, and 2.842 K. METEOSAT is 3.205 K on the same support. At $K = 10$, the paired DDPM-minus-DDIM difference is -0.042 K with a 95% interval of -0.211 to $+0.078$. All values use predeclared member identities on common finite support, and the displayed fields use the same fixed-member DDIM protocol.

3.2. All-scene performance and metric trade-offs

The all-scene analysis asks whether the spatial behaviour seen in the illustrative cases persists across the much larger variation among observations. The 12 examples were selected for high within-group coverage and provide useful spatial illustrations but do not represent the full difficulty range. Across all 212 eligible development-city scenes, the method distributions are broad and substantially overlapping. Scene-to-scene variation is larger than the average separation among learned models, while the aggregate summaries show a consistent but modest benefit from learned residual correction.

All methods are evaluated without output clipping on identical finite support within each scene, and scenes receive equal weight. This matters because observations with more valid pixels would otherwise dominate and because the same scene often remains relatively easy or difficult for every method. Paired contrasts remove much of this shared scene difficulty by comparing methods for the same thermal state, spatial support, and ECOSTRESS observation. SRCNN and DetUNet contribute one stitched field per scene, whereas DDIM-50 $K = 10$ uses the same ten predeclared member identities in every scene. A paired scene-range-clipped shadow calculation from the identical raw predictions gives a no-clipping-minus-shadow-clipped mean RMSE of $+0.028$ K for DDIM-50 $K = 10$. No clipped field enters the primary results.

Table 3 reveals two levels of improvement. Every learned method reduces mean RMSE relative to METEOSAT, and DDIM-50 $K = 10$ gives the lowest value, 4.936 versus 5.143 K, together with the smallest absolute mean bias. SRCNN nevertheless has a slightly lower mean MAE than DDIM-50, 3.614 versus 3.641 K. The corresponding RMSE medians for METEOSAT, SRCNN, DetUNet, and DDIM-50 $K = 10$ are 3.838, 3.664, 3.579, and 3.482 K. Most of the mean RMSE separation is therefore between learned correction and the unmodified coarse input, with a smaller incremental difference among the neural workflows.

Figure 6 shows a mean paired DDIM-50 $K = 10$ difference from METEOSAT of -0.207 K, with a 95% scene-bootstrap interval of -0.400 to -0.011 . DDIM-50 has lower RMSE in 63.2% of scenes. This supports a negative mean difference under scene-level resampling, although it does not imply that every scene or city improves.

A quality-control rule defined independently of model performance flags a scene when more than 1% of supported ECOSTRESS pixels fall outside 200–400 K. It flags two scenes: Addis Ababa on 14 June 2022 (1.84%) and Kinshasa on 15 May 2022 (4.45%). Excluding them gives a paired contrast of -0.207 K (-0.403 to -0.009), compared with -0.207 K for all 212 scenes. The paired contrast retains the same sign after reference-QC exclusion, so the two flagged observations do not determine its direction.

Table 3. Equal-scene mean and median performance statistics on common prediction support across all 212 eligible development-city scenes. Learned predictions and METEOSAT are evaluated without output clipping.

Model	N	RMSE mean (K)	RMSE median (K)	MAE mean (K)	MAE median (K)	Bias mean (K)	Bias median (K)	WaveSim mean	WaveSim median
METEOSAT	212	5.143	3.838	3.764	3.038	−0.278	+0.367	0.364	0.332
SRCNN	212	4.984	3.664	3.614	2.767	+0.461	+0.441	0.369	0.350
Deterministic U-Net	212	4.962	3.579	3.630	2.731	+0.359	+0.479	0.368	0.348
DDIM-50 K = 10 (fixed members 01–10)	212	4.936	3.482	3.641	2.700	+0.164	+0.231	0.427	0.414

WaveSim provides the multiscale counterpart to the radiometric result. Mean total WaveSim increases from 0.364 for METEOSAT to 0.427 for DDIM-50 K = 10. Considered with the lower RMSE, this indicates that the diffusion output does more than create a smoother version of METEOSAT: its wavelet-coefficient behaviour also becomes more similar to ECOSTRESS. A displaced or smoothed field can nevertheless retain similar coefficient behaviour, so WaveSim cannot independently confirm exact feature placement. Confidence is strongest when maps, radiometric errors, structural metrics, and WaveSim agree.

The ensemble-size sensitivity reproduces the trade-off seen in the selected scenes. Under the no-clipping, fixed-member protocol, averaging acts as variance reduction for pixel-wise error and as a smoother of the stochastic residual. K = 10 is retained for the primary comparison as a defined reconstruction product rather than a universally optimal ensemble size. An application centred on stable mean temperature can favour a larger ensemble, while one centred on local variability may retain a smaller ensemble and evaluate its greater realization sensitivity.

Distributional tails expose behaviour that scene means cannot describe. A model can improve RMSE and mean bias while still missing a small set of very cool or very warm locations. Percentile bias describes the predicted temperature distribution, but it does not require the predicted extremes to occur at the same pixels as the ECOSTRESS extremes; a warm shift of the whole distribution can also make a hot-tail value appear favourable. Reference-defined, location-conditioned errors are therefore needed to test whether specific cool surfaces and hot spots are reconstructed rather than only a similar histogram. The deterministic baselines and DDIM use different objectives, optimization histories, data exposure, and tuning effort, so the results compare retained end-to-end workflows rather than isolating architecture. Future work should compare these protocols with the DisTrad/TsHARP lineage (Agam et al. 2007; Kustas et al. 2003), evaluated on identical support and under a common tuning budget.

3.3. City-level variation and calendar-group examples

Across the pooled 212-scene evaluation, DDIM-50 with K = 10 lowers mean RMSE relative to the uncorrected METEOSAT input. This improvement is not uniform across cities: city-mean RMSE is lower in Addis Ababa, Nairobi, Niamey and Touba and higher in Kano and Kinshasa. The April–September and October–March cases in Figure 7 provide spatial context for this variation rather than representing common climatic seasons. Together,

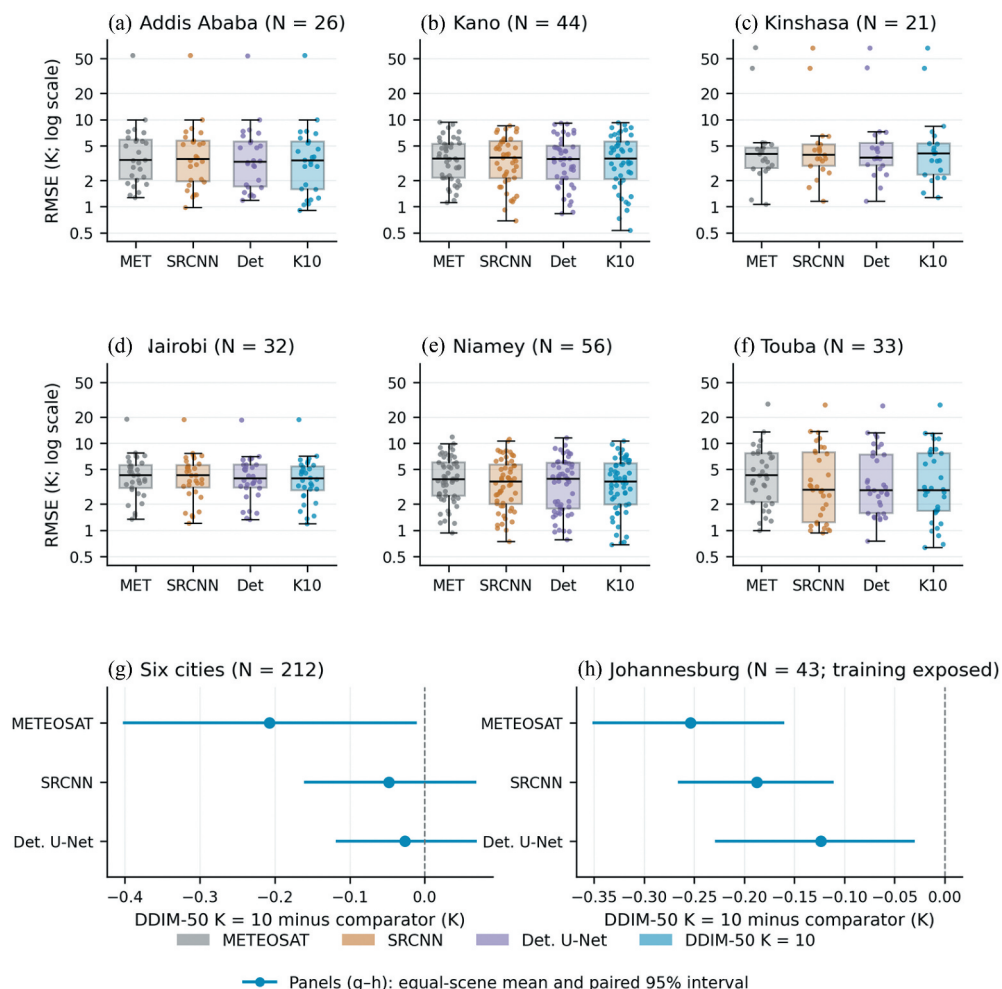


Figure 6. No-clipping per-scene RMSE distributions and paired model contrasts. (a–f) all 112 eligible development-city scenes separated by city; dots show every scene, boxes show the median and interquartile range, and the shared logarithmic scale retains all high-error scenes. (g) DDIM-50 K = 10 minus each comparator across the 212 development-city scenes. (h) The same contrasts across 43 Johannesburg scenes, which form the geographic-holdout evaluation. Points and lines are equal-scene mean differences and paired scene-bootstrap 95% intervals; negative values favor DDIM-50.

the city summaries and maps show how the same residual model responds as the dominant surface contrasts change.

The improving cities span both elevated terrain in Addis Ababa and Nairobi and dryland environments in Niamey and Touba, while the two cities without mean RMSE improvement include dryland Kano and humid Kinshasa. The response therefore does not follow a simple division by elevation, moisture regime, or land-cover class. It reflects the scene-specific relationship among the METEOSAT thermal background, the fine-resolution descriptors, and the ECOSTRESS-referenced residual. This explains how the pooled mean can be favourable even though the magnitude and direction of change differ among cities.

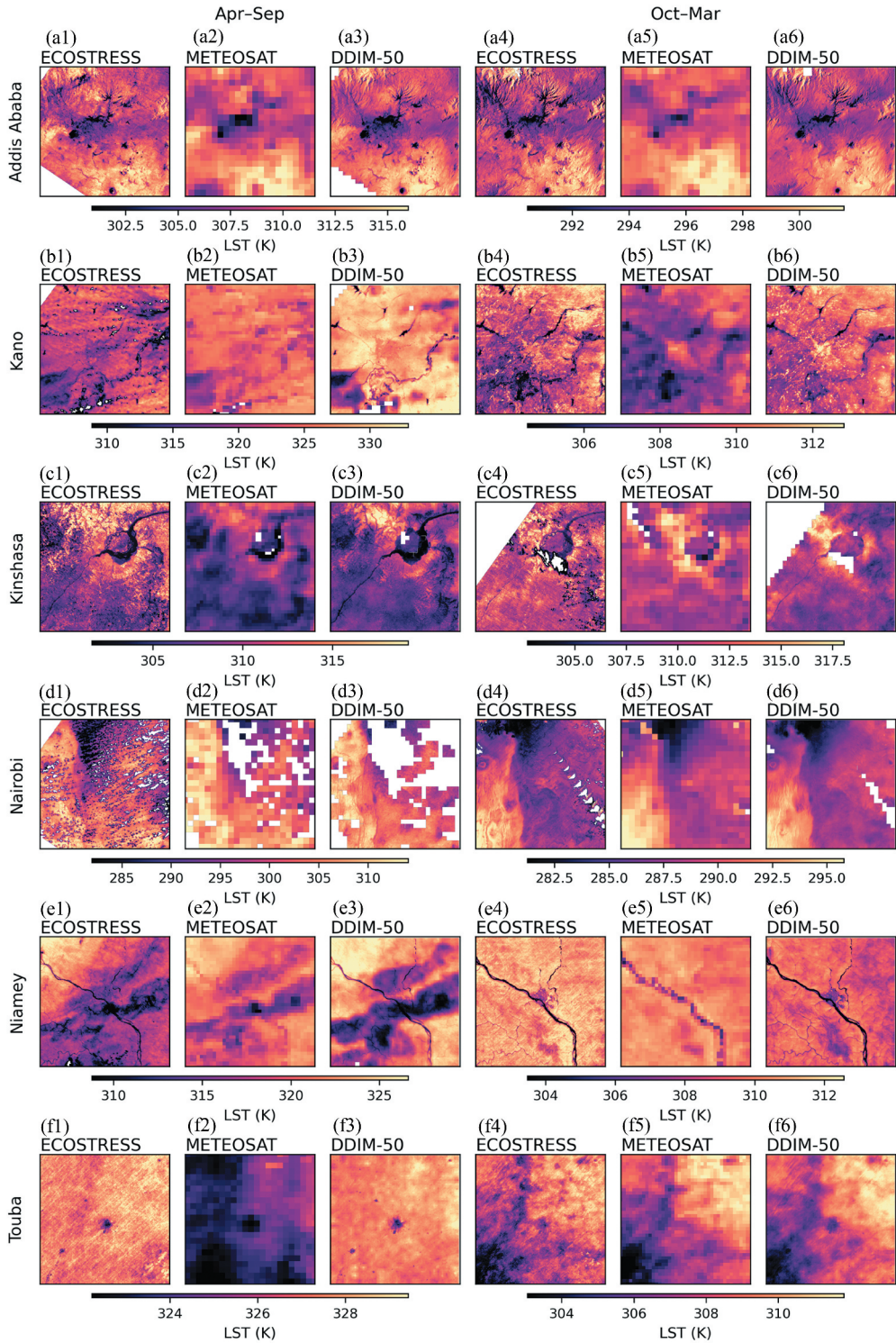


Figure 7. Cross-city comparison of ECOSTRESS, METEOSAT, and fixed-member DDIM-50 $K = 10$ LST for the April–September and October–March illustrative cases. Rows correspond to Addis Ababa, Kano, Kinshasa, Nairobi, Niamey, and Touba. The displayed DDIM fields use the same no-clipping protocol as Table 3; numerical evaluation uses common finite support. Within each calendar-group triplet, common 1st–99th percentile color limits are used only for display and do not alter the evaluated arrays or scores.

Figure 7 uses the same fixed-member, no-clipping DDIM-50 $K=10$ product as Table 3. In Addis Ababa and Nairobi, the reconstructed fields follow terrain-organized gradients and ridge–valley patterns that are largely absent from METEOSAT. In Kinshasa, water and vegetation boundaries organize much of the fine-scale contrast, whereas Kano, Niamey, and Toubia show dryland texture, bare-ground variability, and urban–rural transitions. These qualitatively different geometries are consistent with the predictor stack guiding the correction rather than applying a uniform sharpening filter. Persistent landscape features recur across the two calendar groups, while the absolute thermal range and broad warm–cool background change with the METEOSAT observation. Recognizable structure and lower city-average RMSE remain distinct outcomes, so the maps are interpreted with the city summaries rather than as independent performance evidence.

The reconstructed fields are most informative for analyses of relative fine-scale LST contrast at METEOSAT observation times, including urban–rural gradients and terrain-, water-, and vegetation-related patterns. Compact hot or cool surfaces warrant greater attention because ensemble averaging can attenuate their amplitude even when the surrounding field improves. These products remain radiometric land-surface temperature estimates; translation to near-surface air temperature, human heat exposure, or thermal comfort requires corresponding meteorological, exposure, and validation data (Good 2016; Marzban, Sodoudi, and Preusker 2018).

Taken together, the city results show that the learned correction is conditional on local setting rather than a uniform sharpening operation. The reconstructed fields are most informative when radiometric and structural evidence agree for the setting of interest, while cloudy or incomplete inputs remain outside the valid prediction support. Repeated landscape descriptors may contribute both transferable relationships and persistent city-specific patterns; the present experiment does not isolate those contributions. The comparison is retrospective because development-city observations from 2022 contributed to validation and checkpoint selection. A future partition that excludes evaluation scenes from training, validation, normalization, and model selection is required for prospective temporal and geographic transfer tests.

3.4. Johannesburg performance and cold-tail behaviour

Johannesburg tests whether the average reconstruction behaviour extends beyond the six development cities. Its 43 eligible daytime scenes form the geographic-holdout evaluation and did not contribute to training or normalization. This provides a stronger basis than the illustrative maps alone for asking whether the learned relationship among the coarse thermal state, environmental context, and fine-scale residual remains useful in a different urban domain. The analysis also asks whether improved scene-average performance represents the full temperature distribution, because an aggregate gain can remain uneven for localized surface conditions.

Without output clipping, DDIM-50 $K=10$ lowers mean RMSE from 4.432 K for METEOSAT to 4.178 K, while the medians are 2.494 and 2.176 K. The paired mean difference is -0.254 K with a 95% interval of -0.349 to -0.163 . Mean total WaveSim increases from 0.308 for METEOSAT to 0.363 for DDIM-50 $K=10$. Both radiometric and multiscale comparisons therefore support improved average agreement in the geographic holdout, although they do not imply uniform improvement at every location.

Figure 8 shows how the all-scene result appears in two examples from the same no-clipping DDIM-50 product. The displayed fields contain finer texture and warm-cool contrasts than METEOSAT while retaining the broad thermal organization of the coarse observation. Aggregate scene summaries can nevertheless coexist with errors at rare hot or cool locations because these pixels occupy only a small part of the scene. Location-conditioned diagnostics therefore remain necessary when an application depends on extreme surfaces rather than on the dominant scene area.

The Johannesburg comparison further demonstrates why no single metric is sufficient. A coarse field can have a moderate scene-mean bias while failing to represent where hot and cool features occur. Conversely, a generated field can improve spatial organization without being best at every percentile statistic. The Johannesburg results provide a geographic-holdout diagnostic in a distinct city domain and show how the retained

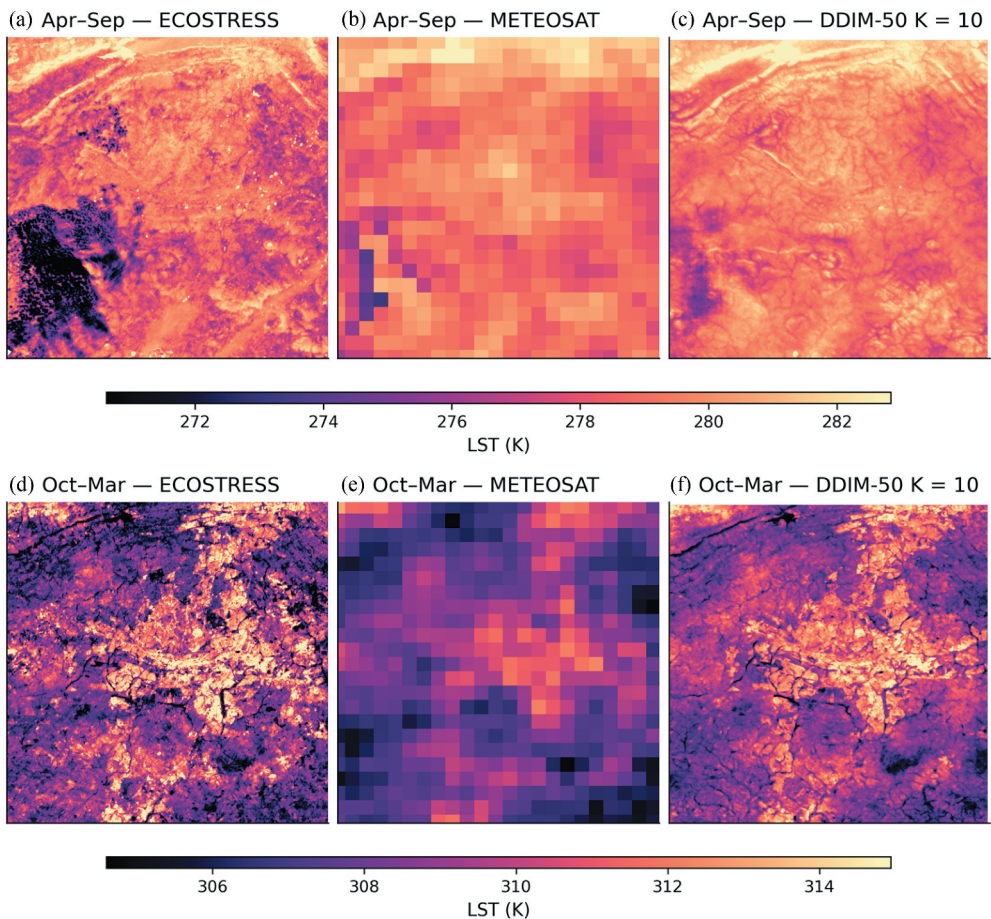


Figure 8. Geographic-holdout examples over Johannesburg for the April–September and October–March cases. The displayed field is the fixed-member, no-clipping DDIM-50 $K = 10$ product evaluated in this analysis. Johannesburg is designated as the geographic holdout. Within each calendar-group triplet, common 1st–99th percentile color limits are used only for display and do not alter the evaluated arrays or scores.

relationship behaves outside the six development cities. Broader geographic generalization will require evaluation in additional fully separated cities.

4. Summary and conclusion

This study evaluates conditional diffusion for observation-based downscaling of land surface temperature (LST) from METEOSAT to the ECOSTRESS spatial scale across environmentally diverse African cities. The approach reconstructs the residual between the aligned METEOSAT observation and the ECOSTRESS reference rather than predicting absolute LST independently. METEOSAT supplies the observed large-scale thermal state, while fine-resolution descriptors of vegetation, water, built surfaces, elevation, building morphology, and time guide the spatial organization of the sub-grid correction. Applying the same residual formulation and evaluation support to diffusion, SRCNN, and deterministic U-Net allows the added value of learned reconstruction to be separated from the information already contained in the coarse thermal input.

Across the 212 eligible daytime scenes from the six development cities, all three learned methods reduce mean RMSE relative to the uncorrected METEOSAT input. The no-clipping DDIM-50 $K = 10$ product has mean and median RMSE of 4.936 and 3.482 K, compared with 5.143 and 3.838 K for METEOSAT. The paired mean difference is -0.207 K (95% interval -0.400 to -0.011). Excluding two scenes flagged by a reference-only physical-range rule gives -0.207 K (-0.403 to -0.009); the comparison retains the same direction after this reference-only exclusion. The deterministic networks achieve similar radiometric performance, and SRCNN has a marginally lower mean MAE than DDIM-50. The clearest overall result is therefore that learned residual correction improves the coarse input, while diffusion provides an additional but smaller advantage when radiometric balance and multiscale similarity are considered together.

The city-level results show that this pooled improvement is not spatially uniform. Relative to the uncorrected METEOSAT input, DDIM-50 with $K = 10$ lowers city-mean RMSE in Addis Ababa, Nairobi, Niamey and Touba but produces small increases in Kano and Kinshasa. These differences do not follow a simple division by elevation, moisture regime, or land-cover type. Instead, reconstruction performance depends on how the observed coarse thermal state, environmental descriptors, and ECOSTRESS-referenced residual relate within each setting. The spatial comparisons support this interpretation: terrain-organized gradients are prominent in Addis Ababa and Nairobi, water and vegetation boundaries shape the Kinshasa case, and the dryland cities contain different combinations of bare-soil texture and urban–rural transitions. The method therefore acts as a context-dependent thermal correction rather than a uniform image-sharpening operation.

The reconstruction diagnostics clarify how the fine-resolution fields are produced. Tukey-weighted blending limits patch-boundary discontinuities, so the resulting texture is not interpreted primarily as an artefact of patch seams. On the 12 illustrated scenes, aggregation to inferred coarse regions gives a DDIM-to-ECOSTRESS RMSE of 2.253 K and correlation of 0.819, compared with 2.543 K and 0.806 for METEOSAT against the same aggregated reference. The reconstruction therefore moves towards the aggregated reference in both RMSE and correlation on these cases. Exact conservation to the original coarse field is not imposed.

The diffusion experiments further demonstrate that ensemble size changes the scientific character of the output. Increasing K lowers pixel-wise RMSE because averaging suppresses member-specific departures and stabilizes their shared component. At the same time, total WaveSim decreases as local wavelet-scale amplitude is attenuated. Larger ensembles consequently provide a smoother and more stable central estimate, whereas smaller ensembles or individual members retain stronger localized variability. A matched no-clipping comparison at $K = 1, 5$, and 10 yields ancestral DDPM-1000 mean RMSE values of 3.241, 2.856, and 2.800 K and DDIM-50 values of 3.113, 2.872, and 2.842 K . This check isolates the reverse sampler while holding initialization, member identity, support, clipping, and aggregation fixed. The maps and headline values use the same predeclared fixed-member DDIM fields. Ensemble size should therefore be selected according to the intended use of the LST field rather than treated as a universally optimizable setting.

Johannesburg provides a separate geographic assessment using all 43 eligible daytime scenes in the geographic holdout. Under no-clipping inference, DDIM-50 with $K = 10$ lowers mean RMSE from 4.432 to 4.178 K and median RMSE from 2.494 to 2.176 K relative to METEOSAT. Mean total WaveSim increases from 0.308 to 0.363. The evaluated Johannesburg scenes did not enter training or normalization. These results provide evidence outside the six development cities, while one geographic holdout cannot establish broad geographic generalization.

A central contribution of this study is the use of complementary diagnostics to distinguish different aspects of reconstruction quality. RMSE, MAE, and bias quantify radiometric agreement; SSIM measures local structural correspondence; WaveSim evaluates multiscale coefficient behaviour; and distributional and location-conditioned tail metrics assess temperature extremes. These measures do not necessarily move in the same direction. A smoother field can achieve lower pixel error while losing local amplitude, and a field can reproduce a similar coefficient distribution without placing individual features in the correct locations. Evaluating the methods through these complementary perspectives provides a more informative account of downscaling performance than assigning overall quality from a single score.

Several limitations define the scope of these findings. Development-city observations from 2022 did not enter training or normalization but did contribute to validation and checkpoint selection, so those results describe retrospective cross-sensor agreement rather than an untouched temporal test. The Johannesburg scenes form the geographic-holdout evaluation. ECOSTRESS is a fine-resolution satellite reference rather than ground truth, so retrieval uncertainty, emissivity assumptions, viewing geometry, acquisition-time mismatch, and differences in sensor support may all contribute to the modelled residual. No-clipping inference removes direct use of the scene ECOSTRESS range, but geographic evaluation remains limited to one city and primarily daytime conditions, and the raw diffusion ensemble spread is not calibrated predictive uncertainty. The reconstructed fields represent surface temperature and should not be interpreted directly as near-surface air temperature or human heat exposure. Future work should therefore use fully separated data partitions, additional geographic holdouts, night-time and extreme-temperature observations, independent in situ measurements, and calibrated probabilistic diagnostics.

Furthermore, the all-scene evaluation uses DDIM-50 to make repeated fixed-member inference computationally feasible. The reported results therefore characterize an accelerated 50-step sampler rather than the maximum performance potentially attainable from the trained 1,000-step diffusion schedule. Increasing the number of DDIM steps could alter radiometric accuracy or the recovery of fine-scale and extreme-temperature features, although improvement is not guaranteed and was not tested systematically. Evaluating sensitivity to longer DDIM trajectories remains an important direction for future work, particularly because a 1,000-step trajectory would require approximately twenty times as many network evaluations per ensemble member as DDIM-50. Such tests should retain the same no-clipping, outcome-independent protocol.

Despite these limitations, the study demonstrates that the fixed-member diffusion product generated without target-derived clipping or outcome-based member selection at inference provides a modest average reduction in retrospective RMSE relative to coarse METEOSAT while reconstructing spatial variation that the original input cannot resolve. Its value lies both in the observation-anchored residual framework and in showing how radiometric accuracy, spatial structure, multiscale variability, coarse-support behaviour, ensemble averaging, and quality control should be evaluated together. By using inference without target-derived clipping or outcome-based member selection, this work makes the size and limits of the gain visible and provides a useful methodological foundation for more reliable fine-resolution thermal reconstruction in heterogeneous and data-sparse urban regions.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was carried out with support from Wellcome [Grant 309864/Z/23/Z]. The views expressed herein do not necessarily represent those of UNESCO, ICTP, or Wellcome.

Data availability statement

The source Earth-observation and geospatial products used in this study are publicly available from their original providers: ECOSTRESS ECO_L2T_LSTE Version 2 from the NASA Land Processes Distributed Active Archive Centre through AppEEARS (https://doi.org/10.5067/ECOSTRESS/ECO_L2T_LSTE.002; <https://appeears.earthdatacloud.nasa.gov/>); the Copernicus Land Monitoring Service Land Surface Temperature 2018–present, global 3 km hourly Version 3 product (<https://doi.org/10.2909/1012ec6a-a073-4cb5-95f4-95dad6f4f423>); Landsat Collection 2 Level-2 surface reflectance from the U.S. Geological Survey (<https://www.usgs.gov/landsat-missions/landsat-collection-2-surface-reflectance>); Copernicus DEM GLO-30 (<https://doi.org/10.5270/ESA-c5d3d65>); and GlobalBuildingAtlas (<https://doi.org/10.14459/2025mp1782307>).

The code supporting this study is openly available on Zenodo at <https://doi.org/10.5281/zenodo.23125291>.

References

- Accarino, G., V. Acquaviva, S. Shamekh, D. Watson-Parris, and D. Lawrence. 2025. "WaveSim: A Wavelet-Based Multi-Scale Similarity Metric for Weather and Climate Fields." *arXiv preprint arXiv:2512.14656*.
- Agam, N., W. P. Kustas, M. C. Anderson, F. Li, and C. M. Neale. 2007. "A Vegetation Index Based Technique for Spatial Sharpening of Thermal Imagery." *Remote Sensing of Environment* 107 (4): 545–558. <https://doi.org/10.1016/j.rse.2006.10.006>.
- Atitar, M., and J. A. Sobrino. 2008. "A Split-Window Algorithm for Estimating LST from Meteosat 9 Data: Test and Comparison with in situ Data and MODIS LSTs." *IEEE Geoscience & Remote Sensing Letters* 6 (1): 122–126.
- Bah, A. R., H. Norouzi, S. Prakash, R. Blake, R. Khanbilvardi, and C. Rosenzweig. 2022. "Spatial Downscaling of GOES-R Land Surface Temperature Over Urban Regions: A Case Study for New York City." *Atmosphere* 13 (2): 332. <https://doi.org/10.3390/atmos13020332>.
- Bechtel, B., K. Zakšek, and G. Hoshyaripour. 2012. "Downscaling Land Surface Temperature in an Urban Area: A Case Study for Hamburg, Germany." *Remote Sensing* 4 (10): 3184–3200. <https://doi.org/10.3390/rs4103184>.
- Bonafoni, S. 2016. "Downscaling of Landsat and MODIS Land Surface Temperature Over the Heterogeneous Urban Area of Milan." *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 9 (5): 2019–2027. <https://doi.org/10.1109/JSTARS.2016.2514367>.
- Chang, Y., Q. Weng, J. A. Vooigt, and J. Xiao. 2025. "Urban Thermal Anisotropies by Local Climate Zones: An Assessment Using Multi-Angle Land Surface Temperatures from ECOSTRESS." *Remote Sensing of Environment* 322:114705. <https://doi.org/10.1016/j.rse.2025.114705>.
- Cilek, M. U., and A. Cilek. 2021. "Analyses of Land Surface Temperature (LST) Variability Among Local Climate Zones (LCZs) Comparing Landsat-8 and ENVI-Met Model Data." *Sustainable Cities and Society* 69:102877. <https://doi.org/10.1016/j.scs.2021.102877>.
- Du, Z., W. Li, D. Zhou, L. Tian, F. Ling, H. Wang, Y. Gui, and B. Sun. 2014. "Analysis of Landsat-8 OLI Imagery for Land Surface Water Mapping." *Remote Sensing Letters* 5 (7): 672–681. <https://doi.org/10.1080/2150704X.2014.960606>.
- Fisher, J. B., B. Lee, A. J. Purdy, G. H. Halverson, M. B. Dohlen, K. Cawse-Nicholson, A. Wang, R. G. Anderson, B. Aragon, and M. A. Arain. 2020. "ECOSTRESS: NASA's Next Generation Mission to Measure Evapotranspiration from the International Space Station." *Water Resources Research* 56 (4): e2019WR026058.
- Freitas, S. C., I. F. Trigo, J. M. Bioucas-Dias, and F.-M. Gottsche. 2009. "Quantifying the Uncertainty of Land Surface Temperature Retrievals from SEVIRI/Meteosat." *IEEE Transactions on Geoscience & Remote Sensing* 48 (1): 523–534. <https://doi.org/10.1109/TGRS.2009.2027697>.
- Gao, F., J. Masek, M. Schwaller, and F. Hall. 2006. "On the Blending of the Landsat and MODIS Surface Reflectance: Predicting Daily Landsat Surface Reflectance." *IEEE Transactions on Geoscience & Remote Sensing* 44 (8): 2207–2218. <https://doi.org/10.1109/TGRS.2006.872081>.
- Good, E. J. 2016. "An in situ-Based Analysis of the Relationship Between Land Surface 'Skin' And Screen-Level Air Temperatures." *Journal of Geophysical Research Atmospheres* 121 (15): 8801–8819. <https://doi.org/10.1002/2016JD025318>.
- Ho, J., A. Jain, and P. Abbeel. 2020. "Denoising Diffusion Probabilistic Models." *Advances in Neural Information Processing Systems* 33:6840–6851. https://proceedings.neurips.cc/paper_files/paper/2020/hash/4c5bfc8584af0d967f1ab10179ca4b-Abstract.html.
- Hulley, G. C., F. M. Göttsche, G. Rivera, S. J. Hook, R. J. Freepartner, M. A. Martin, K. Cawse-Nicholson, and W. R. Johnson. 2021. "Validation and Quality Assessment of the ECOSTRESS Level-2 Land Surface Temperature and Emissivity Product." *IEEE Transactions on Geoscience & Remote Sensing* 60:1–23. <https://doi.org/10.1109/TGRS.2021.3079879>.
- Hutengs, C., and M. Vohland. 2016. "Downscaling Land Surface Temperatures at Regional Scales with Random Forest Regression." *Remote Sensing of Environment* 178:127–141. <https://doi.org/10.1016/j.rse.2016.03.006>.
- Kustas, W. P., J. M. Norman, M. C. Anderson, and A. N. French. 2003. "Estimating Subpixel Surface Temperatures and Energy Fluxes from the Vegetation Index–Radiometric Temperature

- Relationship." *Remote Sensing of Environment* 85 (4): 429–440. [https://doi.org/10.1016/S0034-4257\(03\)00036-1](https://doi.org/10.1016/S0034-4257(03)00036-1).
- Lee, J., M. Berkelhammer, M. D. Wilson, N. Love, and R. Cintron. 2024. "Urban Land Surface Temperature Downscaling in Chicago: Addressing Ethnic Inequality and Gentrification." *Remote Sensing* 16 (9): 1639. <https://doi.org/10.3390/rs16091639>.
- Li, P., C. Xiao, and Z. Feng. 2018. "Mapping Rice Planted Area Using a New Normalized EVI and SAVI (NVI) Derived from Landsat-8 OLI." *IEEE Geoscience & Remote Sensing Letters* 15 (12): 1822–1826. <https://doi.org/10.1109/LGRS.2018.2865516>.
- Li, W., L. Ni, Z.-I. Li, S.-B. Duan, and H. Wu. 2019. "Evaluation of Machine Learning Algorithms in Spatial Downscaling of MODIS Land Surface Temperature." *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 12 (7): 2299–2307. <https://doi.org/10.1109/JSTARS.2019.2896923>.
- Liao, W., T. Hong, and Y. Heo. 2021. "The Effect of Spatial Heterogeneity in Urban Morphology on Surface Urban Heat Islands." *Energy & Buildings* 244:111027. <https://doi.org/10.1016/j.enbuild.2021.111027>.
- Marzban, F., S. Sodoudi, and R. Preusker. 2018. "The Influence of Land-Cover Type on the Relationship Between NDVI–LST and LST–T Air." *International Journal of Remote Sensing* 39 (5): 1377–1398. <https://doi.org/10.1080/01431161.2017.1402386>.
- Mukherjee, S., P. Joshi, and R. D. Garg. 2014. "A Comparison of Different Regression Models for Downscaling Landsat and MODIS Land Surface Temperature Images Over Heterogeneous Landscape." *Advances in Space Research* 54 (4): 655–669. <https://doi.org/10.1016/j.asr.2014.04.013>.
- Mukherjee, S., P. Joshi, and R. D. Garg. 2015. "Evaluation of LST Downscaling Algorithms on Seasonal Thermal Data in Humid Subtropical Regions of India." *International Journal of Remote Sensing* 36 (10): 2503–2523. <https://doi.org/10.1080/01431161.2015.1041175>.
- Rahimi, E., S. Barghjvelveh, and P. Dong. 2021. "Quantifying How Urban Landscape Heterogeneity Affects Land Surface Temperature at Multiple Scales." *Journal of Ecology and Environment* 45 (1): 22. <https://doi.org/10.1186/s41610-021-00203-z>.
- Saharia, C., J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi. 2022. "Image super-resolution via iterative refinement." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 45 (4): 4713–4726. <https://doi.org/10.1109/TPAMI.2022.3204461>.
- Schmetz, J., P. Pili, S. Tjemkes, D. Just, J. Kerkmann, S. Rota, and A. Ratier. 2002. "Supplement to an Introduction to Meteosat Second Generation (MSG)." *Bulletin of the American Meteorological Society* 83 (7): 992. <https://doi.org/10.1175/BAMS-83-7-Schmetz-2>.
- Schwaab, J., R. Meier, G. Mussetti, S. Seneviratne, C. Bürgi, and E. L. Davin. 2021. "The Role of Urban Trees in Reducing Land Surface Temperatures in European Cities." *Nature Communications* 12 (1): 6763. <https://doi.org/10.1038/s41467-021-26768-w>.
- Simard, M., M. Denbina, C. Marshak, and M. Neumann. 2024. "A Global Evaluation of Radar-Derived Digital Elevation Models: SRTM, NASADEM, and GLO-30." *Journal of Geophysical Research Biogeosciences* 129 (11): e2023JG007672. <https://doi.org/10.1029/2023JG007672>.
- Sobrinho, J. A., R. Oltra-Carrió, G. Sòria, R. Bianchi, and M. Paganini. 2012. "Impact of Spatial Resolution and Satellite Overpass Time on Evaluation of the Surface Urban Heat Island Effects." *Remote Sensing of Environment* 117:50–56. <https://doi.org/10.1016/j.rse.2011.04.042>.
- Song, J., C. Meng, and S. Ermon. 2020. "Denoising Diffusion Implicit Models." *arXiv preprint arXiv:2010.02502*.
- Trevisani, S., T. N. Skrypitsyna, and I. V. Florinsky. 2023. "Global Digital Elevation Models for Terrain Morphology Analysis in Mountain Environments: Insights on Copernicus GLO-30 and ALOS AW3D30 for a Large Alpine Area." *Environmental Earth Sciences* 82 (9): 198. <https://doi.org/10.1007/s12665-023-10882-7>.
- Trigo, I. F., L. F. Peres, C. C. DaCamara, and S. C. Freitas. 2008. "Thermal Land Surface Emissivity Retrieved from SEVIRI/Meteosat." *IEEE Transactions on Geoscience & Remote Sensing* 46 (2): 307–315. <https://doi.org/10.1109/TGRS.2007.905197>.
- Venter, Z. S., S. L. Scott, P. G. Desmet, and M. T. Hoffman. 2020. "Application of Landsat-Derived Vegetation Trends Over South Africa: Potential for Monitoring Land Degradation and Restoration." *Ecological Indicators* 113:106206. <https://doi.org/10.1016/j.ecolind.2020.106206>.

- Wang, C., and W. Sun. 2025. "Semantic Guided Large Scale Factor Remote Sensing Image Super-Resolution with Generative Diffusion Prior." *ISPRS Journal of Photogrammetry & Remote Sensing* 220:125–138. <https://doi.org/10.1016/j.isprsjprs.2024.12.001>.
- Weng, Q., P. Fu, and F. Gao. 2014. "Generating Daily Land Surface Temperature at Landsat Resolution by Fusing Landsat and MODIS Data." *Remote Sensing of Environment* 145:55–67. <https://doi.org/10.1016/j.rse.2014.02.003>.
- Yang, J., Y. Zhan, X. Xiao, J. C. Xia, W. Sun, and X. Li. 2020. "Investigating the Diversity of Land Surface Temperature Characteristics in Different Scale Cities Based on Local Climate Zones." *Urban Climate* 34:100700. <https://doi.org/10.1016/j.uclim.2020.100700>.
- Yoo, C., J. Im, S. Park, and D. Cho. 2020. "Spatial Downscaling of MODIS Land Surface Temperature: Recent Research Trends, Challenges, and Future Directions." *Korean Journal of Remote Sensing* 36 (4): 609–626.
- Zhou, D., J. Xiao, S. Bonafoni, C. Berger, K. Deilami, Y. Zhou, S. Frolking, R. Yao, Z. Qiao, and J. A. Sobrino. 2018. "Satellite Remote Sensing of Surface Urban Heat Islands: Progress, Challenges, and Perspectives." *Remote Sensing* 11 (1): 48. <https://doi.org/10.3390/rs11010048>.
- Zhu, X., J. Chen, F. Gao, X. Chen, and J. G. Masek. 2010. "An Enhanced Spatial and Temporal Adaptive Reflectance Fusion Model for Complex Heterogeneous Regions." *Remote Sensing of Environment* 114 (11): 2610–2623. <https://doi.org/10.1016/j.rse.2010.05.032>.
- Zhu, X., E. H. Helmer, F. Gao, D. Liu, J. Chen, and M. A. Lefsky. 2016. "A Flexible Spatiotemporal Method for Fusing Satellite Images with Different Resolutions." *Remote Sensing of Environment* 172:165–177. <https://doi.org/10.1016/j.rse.2015.11.016>.
- Zhu, X. X., S. Chen, F. Zhang, Y. Shi, and Y. Wang. 2025. "GlobalBuildingAtlas: An Open Global and Complete Dataset of Building Polygons, Heights and LoD1 3D Models." *Earth System Science Data* 17 (12): 6647–6668. <https://doi.org/10.5194/essd-17-6647-2025>.