Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Knowledge-Guided Probabilistic Forecasting of Urban Land-Surface Temperature in Three Moroccan Cities

**Merieme Debbagh**

debbagh.merieme@gmail.com

Université Sultan Moulay Slimane

**youness madani**

Université Sultan Moulay Slimane

**Belaid Bouikhalene**

Université Sultan Moulay Slimane

**Research Article**

**Keywords:** Urban heat forecasting, Land surface temperature, Probabilistic forecasting, Knowledge-guided modeling, Directional residuals, Calibration, Morocco

**Posted Date:** October 30th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11217781/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Knowledge-Guided Probabilistic Forecasting of Urban Land-Surface Temperature in Three Moroccan Cities

Merieme Debbagh[1*], Youness Madani[1] and Belaid Bouikhalene[1]

[1]Computer Science Department, Sultan Moulay Slimane University, Beni Mellal, Morocco.

*Corresponding author(s). E-mail(s): debbagh.merieme@gmail.com;
Contributing authors: youness.madani@usms.ma;
b.bouikhalene@usms.ma;

**Abstract**

Urban heat forecasting is important for climate adaptation, yet reliable multi-city prediction remains difficult because short-term meteorology interacts with persistent urban form, this study evaluates a knowledge-guided probabilistic framework for next-day land surface temperature (LST) forecasting in Casablanca, Marrakech, and Tangier, a seven-day dynamic sequence and six static descriptors feed a heteroscedastic recurrent baseline, while a matched-seed warm-start residual introduces prescribed directional contributions for built-up, impervious, and vegetation fractions. A sentinel-clean audit retained 365,500 valid daily targets from 440 verified cells and reproduced 24,902 canonical forecasting sequences, the point benchmark evaluates six models over five seeds on identical train, validation, and test target sets (14,917/4,008/5,977), using minimum validation RMSE for checkpoint selection, the probabilistic dynamic+static baseline achieved the lowest mean test RMSE and MAE, **2.998 ± 0.043** and **2.313 ± 0.017** °C; the proposed model ranked second in RMSE and third in MAE at **3.034±0.119** and **2.349±0.090** °C, in the separately preserved validation-Gaussian-NLL-selected paired analysis, the 95% moving-block-bootstrap intervals for RMSE and Gaussian NLL included zero, whereas MAE favored the baseline by 0.0336 °C.
The structured mean residual nevertheless satisfies the prescribed directions exactly and raises complete-predictor consistency by 7.6, 10.4, and 24.7 percentage points for built-up, impervious, and vegetation responses at the primary perturbation, calibration and hot-tail analyses do not show an overall uncertainty advantage, and transfer to an unseen city is mixed.

The contribution is therefore a reproducible accuracy–physical-consistency trade-off within a probabilistic forecasting pipeline rather than a claim of predictive or uncertainty superiority.

**Keywords:** Urban heat forecasting; Land surface temperature; Probabilistic forecasting; Knowledge-guided modeling; Directional residuals; Calibration; Morocco

# 1 Introduction

Urban heat is an increasingly consequential manifestation of climate change in cities, elevated surface and near-surface temperatures intensify heat stress, increase cooling demand, and can amplify inequalities in thermal exposure, at the same time, the growing availability of meteorological, remotely sensed, and geospatial urban data has made data-driven modeling of urban thermal dynamics substantially more feasible, recent reviews nevertheless describe a field in which methodological progress has been uneven, [1] report rapid growth in machine-learning applications for urban canopy-layer air-temperature prediction but identify continuing weaknesses in uncertainty reporting, benchmark consistency, and operational validation, the study [2] reach a similar conclusion for urban heat-island prediction, where evaluation remains dominated by point-accuracy comparisons, broader reviews [3, 4] further argue that predictive modeling should be assessed in relation to the full chain from sensing and forecasting to mitigation, adaptation, and decision support.

Recent empirical work confirms that urban heat modeling has moved well beyond descriptive mapping, [5] show that explicit urban-configuration information can improve land surface temperature prediction, thereby linking thermal prediction to the physical structure of the city, [6] use deep learning to estimate thermal consequences of projected urban sprawl, extending the task toward scenario-based prediction, morphology-aware work under local climate zones likewise demonstrates that urban form carries systematic thermal information [7], multi-source and multi-city studies have further expanded the forecasting setting by combining meteorological histories, urban morphology, and distributed observations [8–10].

Two methodological issues are particularly important, the first is physical plausibility; vegetation is generally associated with cooling through shading and evapotranspiration, whereas greater built-up and impervious fractions are often associated with enhanced heat storage and reduced evaporative cooling under otherwise comparable conditions, these relationships are heterogeneous rather than universal causal laws, and post-hoc interpretability studies show that their magnitudes can vary across local climate zones and urban contexts [11–13], nevertheless, most forecasting models still treat such variables as unrestricted covariates. When a directional prior is scientifically intended, an unconstrained predictor may therefore respond in a direction that conflicts with that prior even if its average error is low.

A broader machine-learning literature provides mechanisms for incorporating such prior structure, theory-guided and scientific-knowledge-guided learning explicitly integrate domain knowledge into data-driven models through inputs, architecture,

objectives, or hybrid components [14, 15]; shape-constrained methods, including deep lattice networks and constrained monotonic neural networks, show that user-specified monotone relationships can be enforced by construction [16, 17].
The second issue is probabilistic reliability; environmental predictions may be used under spatial heterogeneity, measurement imperfections, and distribution shift, so a probabilistic output should be evaluated rather than treated as automatically informative, in an Earth-observation setting, the study [18] demonstrate the importance of calibration and uncertainty assessment for deep models under distribution shift. The methodological implication is relevant here: reporting a predictive variance is not, by itself, evidence of superior uncertainty quality, proper scores, interval behavior, hot-tail performance, and transfer must be examined separately from point accuracy.
Against this background, this study develops and audits a knowledge-guided probabilistic forecasting framework for next-day LST in Casablanca, Marrakech, and Tangier, the objective is not to maximize a single benchmark score or to claim universal physical correctness, rather, the study asks whether a transparent, directionally structured residual can improve physically intended response behavior while keeping predictive performance close to a strong unconstrained probabilistic baseline, the evaluation is designed around exact target alignment, matched seeds, separated checkpoint-selection tracks, dependence-aware paired inference, mechanism-focused interventions, calibration diagnostics, and cross-city stress tests.
The main contributions are:

- A probabilistic next-day LST forecaster combines seven-day meteorological and thermal-history sequences with static urban descriptors and augments a strong recurrent baseline with explicitly structured mean and variance residuals.
- The architectural claim is stated at its correct scope: the structured residual obeys the prescribed directions for built-up, impervious, and vegetation fractions, whereas the unconstrained base branch prevents a claim of global end-to-end monotonicity.
- A sentinel-clean data-lineage audit establishes a reproducible canonical benchmark, and six five-seed models are evaluated on exactly the same 14,917/4,008/5,977 train/validation/test targets, including two graph/spatial and two modern temporal forecasting families.
- The accuracy–physical-consistency trade-off is examined using paired moving-block bootstrap inference, complete-predictor directionality tests, frozen component interventions, independently retrained structural ablations, calibration and hot-tail diagnostics, city-wise and leave-one-city-out analyses, and a selective 2022 secondary holdout.

## 2 Related Work and Research Gap

Research on data-driven urban heat has progressed from retrospective thermal mapping toward predictive, morphology-aware, and increasingly operational modeling. Recent reviews nevertheless indicate that methodological development has been uneven, [1] identify rapid growth in machine-learning applications for urban canopy-layer temperature prediction but also note persistent weaknesses in uncertainty

reporting, benchmark consistency, and operational validation, [2] reach a similar conclusion for urban heat-island prediction, where model assessment remains dominated by point-error comparisons, broader reviews [3, 4] argue that urban heat modeling should be evaluated across the chain from sensing and prediction to mitigation, adaptation, and decision support, while work on local climate zones emphasizes that thermal processes vary substantially across morphological settings [19], spatial heterogeneity is therefore a core modeling issue rather than a secondary source of noise.

## 2.1 Data-driven prediction and forecasting of urban thermal conditions

Machine-learning studies have established that urban thermal variables can be modeled directly from remotely sensed, meteorological, and geospatial information, [5] show that explicit representation of urban configuration improves land surface temperature prediction, demonstrating that built form is relevant to both accuracy and interpretation, [6] use deep learning to estimate the thermal consequences of urban sprawl, extending urban heat modeling from retrospective description toward scenario-based prediction, the study [20] examine thermal-field variation under urban expansion, while the work [21] address high-resolution LST downscaling in a rapidly urbanizing semi-arid region.

A related body of work moves closer to operational or multi-city forecasting, the studies [8] and [9] demonstrate the value of citizen weather-station observations for urban temperature prediction, broadening the observational basis beyond conventional monitoring networks, while [10] compare GRU, DNN, and ANN models using urban morphological variables across European cities, providing an explicitly multi-city perspective, [22] combine thermal, morphological, and activity-related information for high-resolution air-temperature modeling, [23] develops a machine-learning framework for spatiotemporal urban temperature mapping under typical meteorological conditions; and [24] analyze urban heat from a spatiotemporal modeling perspective.

## 2.2 Urban morphology, local climate zones, and interpretable learning

Urban morphology is widely recognized as a major determinant of intra-urban thermal variation, [7] analyze air temperature through local climate zones and morphology-aware machine learning, [25, 26] examine how urban structure and development intensity relate to thermal conditions, and other work connects morphology to high-resolution temperature mapping [22].

Interpretability-oriented studies further clarify these relationships, [11] use interpretable machine learning to assess urban scene characteristics in New York City, [12] employ a CatBoost–SHAP framework to analyze heterogeneous morphology effects across local climate zones, and [13] use XGBoost–SHAP to investigate built-environment influences on urban heat islands.

The morphology literature is also closely connected to planning and mitigation, [27] combines machine learning and geospatial analysis to support surface urban heat-island mitigation in Riyadh, the work in [28] develop a machine-learning framework

for heat mitigation in informal settlements, [29] examine policy implementation for urban heat-island mitigation, [30] identify multiscale nonlinear effects of urban green space, and [31] study seasonal inequity in heat exposure. Additional studies investigate nighttime heat-island patch structure [32] and the thermal role of public open space using multisource data [33].

## 2.3 Knowledge-guided and shape-constrained learning

The idea of integrating scientific knowledge into machine learning is broader than urban heat forecasting, theory-guided data science frames domain knowledge as a source of structure that can improve scientific consistency, interpretability, or generalization [14], a later survey organizes scientific-knowledge integration across architecture, loss functions, hybrid modeling, and other mechanisms for engineering and environmental systems [15], within this broad family, shape-constrained learning addresses a more specific problem: enforcing known qualitative relationships such as monotonicity, deep lattice networks can impose partial monotonicity with respect to user-specified inputs [16], while constrained monotonic neural networks provide another architecture-level route to monotone function approximation [17].

## 2.4 Spatiotemporal forecasting architectures and probabilistic evaluation

General time-series and graph-learning research provides several architectures relevant to urban thermal forecasting, STGCN combines graph and temporal convolutions to model observations linked by a fixed spatial structure [34], MTGNN learns directed inter-variable relations and integrates graph propagation with dilated temporal convolutions [35], PatchTST represents time series as temporal patches processed by a channel-independent Transformer [36], whereas TimesNet maps one-dimensional temporal variation into a two-dimensional multi-periodic representation [37], these model families are useful references because urban temperature evolves through both temporal persistence and spatial coupling, their effectiveness, however, depends on the forecast horizon, history length, graph definition, availability of static descriptors, and tuning protocol.

Probabilistic evaluation remains less developed in urban heat forecasting than point prediction, in the broader environmental domain, [18] show that calibration and uncertainty assessment are important when deep models are evaluated under distribution shift, the implication for urban heat is straightforward: a forecasting system should not be judged only by RMSE or MAE when it also produces a predictive distribution, proper scores, empirical coverage, interval width, and high-temperature behavior provide complementary evidence.

## 2.5 Research gap and study positioning

The literature leaves four connected methodological requirements insufficiently resolved in combination, first, benchmark comparisons require exact alignment of target observations, split years, metrics, random seeds, and checkpoint-selection rules;

otherwise, rankings can be confounded by the evaluated cohort, second, the established physical relevance of urban morphology is rarely translated into an explicitly audited directional mechanism within a probabilistic urban-heat forecaster, third, evaluations often emphasize average point accuracy without jointly examining calibration, high-temperature behavior, transfer across cities, or the dependence structure of spatiotemporal errors, fourth, post-hoc explanations are commonly reported without mechanism-focused interventions that test whether a proposed structural component is actually responsible for the claimed response behavior.

This study addresses those requirements through a deliberately narrow contribution, it combines a probabilistic dynamic–static recurrent predictor with a structured residual whose prescribed directional effects apply exactly to the residual branch, not to the complete predictor, the empirical design uses a verified canonical cohort and evaluates the baseline, the full proposed model, and four contemporary graph or temporal reference implementations on identical target observations over five matched seeds, point-performance comparisons are complemented by city-stratified moving-block bootstrap inference, physical-directionality audits, frozen-component interventions, independently retrained structural ablations, calibration and hot-tail diagnostics, a secondary temporal holdout, and leave-one-city-out evaluation, our contribution is to quantify an accuracy–physical-consistency trade-off under a reproducible multi-city forecasting protocol.

Table 1 summarizes the methodological progression of the field and the unresolved issue motivating each component of the present design.

**Table 1** Synthesis of representative related work and the methodological requirements motivating the present study.

| Research emphasis | Representative studies | Established contribution | Unresolved methodological issue |
|---|---|---|---|
| Urban thermal prediction and downscaling | Tanoori et al. [5]; Fu et al. [6]; Zhang et al. [20]; Tahooni et al. [21] | Demonstrates that remotely sensed, meteorological, and urban-configuration variables support direct prediction of urban thermal conditions at multiple resolutions. | Results are difficult to compare across heterogeneous targets, resolutions, sampling schemes, horizons, and validation protocols. |
| Operational and multi-city forecasting | Romero Rodríguez et al. [8]; Castro Medina et al. [9]; Tehrani et al. [10]; Chen and Han [22]; Shen [23]; Sampaio et al. [24] | Extends urban heat modeling toward temporally resolved, multisource, and multi-city prediction. | Physically guided structure, probabilistic output, and exact target-aligned benchmarking are rarely integrated in one framework. |
| Morphology, local climate zones, and interpretability | Fan et al. [7]; Lau et al. [25, 26]; Li et al. [11]; Liu et al. [12]; Zhang et al. [13] | Establishes that urban-form effects are heterogeneous, physically meaningful, and important for both prediction and explanation. | Directional knowledge is usually assessed post hoc rather than encoded and tested as an architectural mechanism. |
| Knowledge-guided and shape-constrained learning | Karpatne et al. [14]; Willard et al. [15]; You et al. [16]; Runje and Shankaranarayana [17] | Provides general mechanisms for integrating scientific knowledge and for imposing user-specified monotonic relationships by construction. | These general methods do not by themselves establish how a residual-level directional constraint behaves inside probabilistic multi-city urban LST forecasting. |
| Planning and mitigation applications | Alghamdi [27]; Hussainzad and Gou [28]; Wang and Liu [31]; Zhong et al. [30]; Wang et al. [29, 32]; Shi et al. [33] | Connects thermal analysis to green-space design, exposure equity, mitigation policy, and urban planning. | The emphasis is generally diagnostic or prescriptive rather than probabilistic next-day forecasting and reliability assessment. |
| Graph and modern temporal forecasting | STGCN [34]; MTGNN [35]; PatchTST [36]; TimesNet [37] | Provides mechanisms for spatial dependence, adaptive graph learning, patch-based attention, and multi-periodic temporal representation. | Reported performance is sensitive to history length, graph construction, input availability, tuning, and the evaluated target cohort. |
| Probabilistic environmental prediction | Zhang et al. [18]; methodological concerns identified in reviews [1, 2] | Shows that calibration, distribution shift, and predictive reliability are essential beyond point accuracy. | Urban heat studies rarely combine probabilistic forecasting with calibration, hot-tail, transfer, and dependence-aware inference. |

# 3 Methodology

The methodological framework was designed so that data construction, model development, benchmark comparison, and reliability assessment follow a transparent and internally consistent protocol, Figure 1 summarizes the complete workflow, from multi-source data integration and sentinel-clean validation to canonical-cohort evaluation, physical-response diagnostics, and robustness analysis.

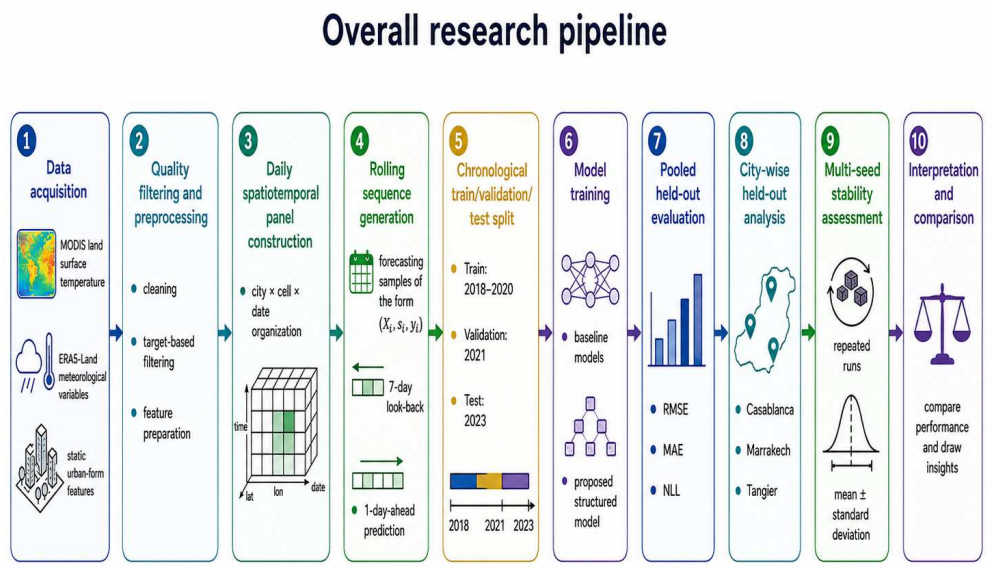



**Fig. 1** Overall workflow from multi-source data construction and sentinel-clean validation to canonical-cohort training, aligned comparison, physical-response diagnostics, and reliability analysis.

## 3.1 Study area, data sources, and verified lineage

The analysis covers Casablanca, Marrakech, and Tangier, daily land-surface temperature (LST) is obtained from the MODIS/Terra MOD11A1 Collection 6.1 product (Google Earth Engine collection `MODIS/061/MOD11A1`), a daily product with 1-km native spatial resolution; the LST values are converted to degrees Celsius during preprocessing, daily meteorological drivers are taken from ERA5-Land Daily Aggregates (Google Earth Engine collection `ECMWF/ERA5_LAND/DAILY_AGGR`) and harmonized with the cell-level daily panel.

The ten dynamic predictors used by the forecasting models are `t2m_c`, `td2m_c`, `rh2m_pct`, `wind10_ms`, `ssrd_wm2`, `tp_mm`, `sp_hpa`, `lst_c`, `sin_doy`, and `cos_doy`; the prediction target is next-day `target_lst_c`.

The six static descriptors are longitude, latitude, elevation (`elev_m`), vegetation fraction, built-up fraction, and impervious fraction.

Daily city–year extraction was performed in Google Earth Engine, with ERA5-Land variables sampled into the spatial extraction framework and missing LST represented upstream by sentinel values that are removed by the audited cleaning procedure.

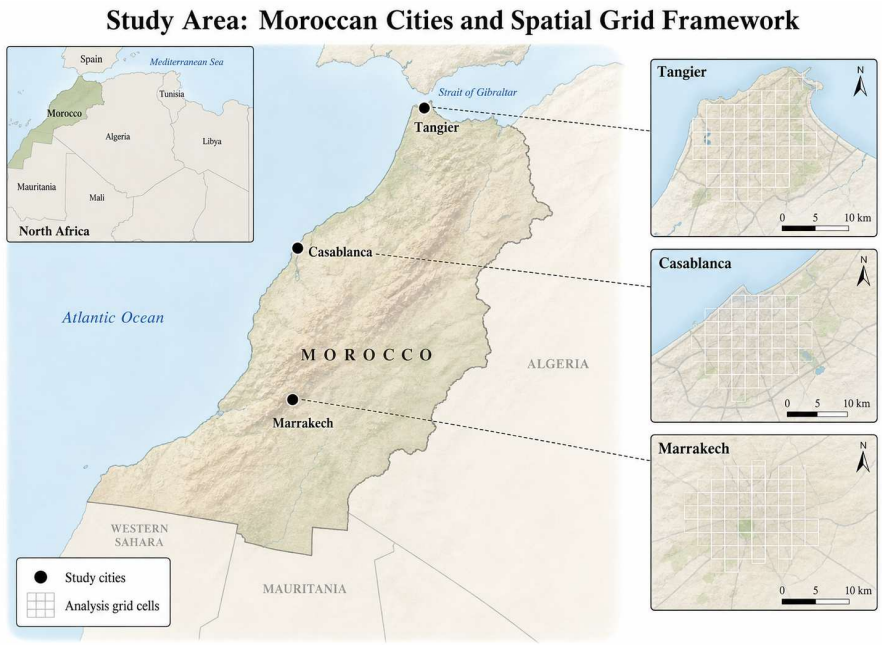



**Fig. 2** Study area and spatial grid framework for Casablanca, Marrakech, and Tangier. The verified modeling panel contains 94, 97, and 249 retained cells, respectively.

The data-lineage audit distinguishes verified downstream quantities from historical upstream assertions, the merged city–cell–date panel contains 964,040 rows, and a legacy target-clean representation contained 370,780 rows, a sentinel-clean reconstruction removed 5,280 invalid target sentinel values and contains 365,500 rows, the canonical train, validation, and test sequences are exactly unchanged between the legacy and sentinel-clean versions, including keys, ordering, features, and targets. The final retained set contains 440 cells: 94 in Casablanca, 97 in Marrakech, and 249 in Tangier.

To clarify the provenance of the final modeling dataset, Figure 3 summarizes the verified row-level transitions from the merged panel to the sentinel-clean representation, the diagram distinguishes reproducible cleaning operations from upstream filtering stages that could not be reconstructed retrospectively.

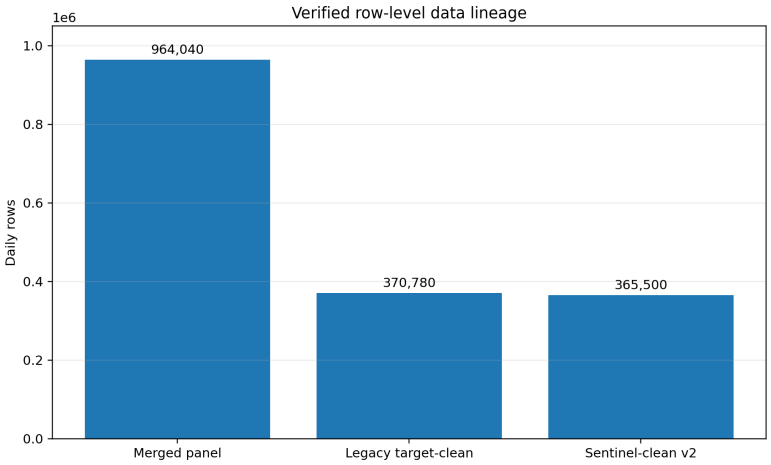



**Fig. 3** Verified row-level data lineage. Sentinel cleaning removes 5,280 invalid target rows without changing the canonical forecasting sequences.

Figure 4 complements the lineage audit by illustrating how the heterogeneous thermal, meteorological, temporal, and static spatial variables were harmonized into the daily spatiotemporal panel and subsequently transformed into rolling next-day forecasting samples.

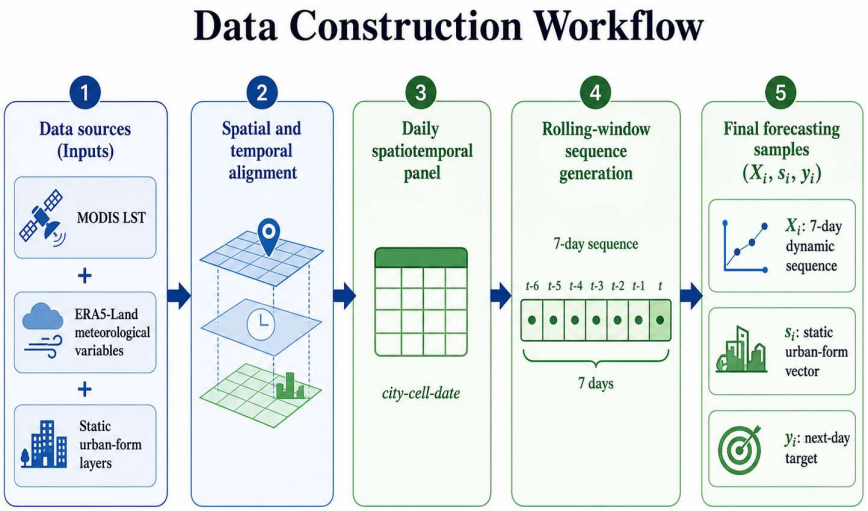



**Fig. 4** Construction of the daily spatiotemporal panel and rolling forecasting samples from heterogeneous data sources.

## 3.2 Canonical forecasting samples

The look-back length is $L = 7$ days and the forecast horizon is $H = 1$ day, a sample $(X_i, s_i, y_i)$ is retained only when the seven input dates are consecutive, all required

dynamic and static inputs are valid, and the next-day target is observed, the normalizer is fitted exclusively on the training split: means and standard deviations for the dynamic variables, static variables, and target are estimated from training data and then applied unchanged to validation and test data. Predictions are transformed back to physical units for evaluation, and predictive variances are rescaled consistently with the squared target scale.
The primary chronological split uses 2018–2020 for training, 2021 for validation, and 2023 for testing, yielding 14,917, 4,008, and 5,977 canonical targets, respectively; the corresponding city counts are reported in Table 2.

**Table 2** Canonical sequence counts by split and city.

| Split | Casablanca | Marrakech | Tangier | Total |
|---|---|---|---|---|
| Training | 2,724 | 5,179 | 7,014 | 14,917 |
| Validation | 925 | 1,806 | 1,277 | 4,008 |
| Test | 1,350 | 2,219 | 2,408 | 5,977 |

Figure 5 illustrates the structure of each canonical forecasting sample, a seven-day sequence of dynamic meteorological and thermal variables is combined with a static urban-form vector to predict the land-surface temperature of the following day.

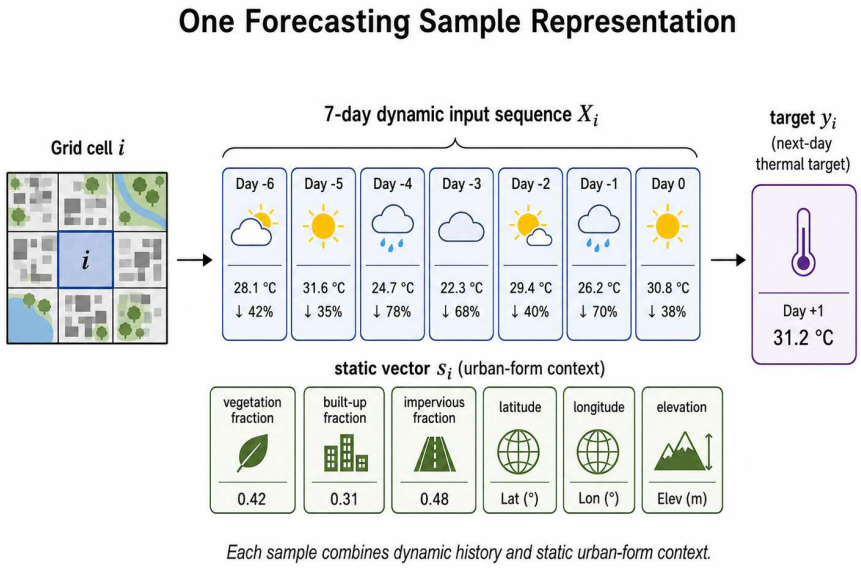



**Fig. 5** Representation of a seven-day dynamic sequence, a static urban-form vector, and the next-day LST target.

## 3.3 Probabilistic base model

A GRU maps the dynamic sequence to $h = E_\theta(X)$. A feed-forward branch maps the complete normalized static vector to $u = U_\eta(s)$, and a fusion module gives $z =$

$F_\rho([h, u])$, linear heads output a base mean $\mu_0(X, s)$ and a pre-softplus variance score $q_0(X, s)$, with

$$v_0(X, s) = \text{softplus}(q_0(X, s)) + \varepsilon, \qquad \varepsilon = 10^{-4}. \tag{1}$$

The score $q_0$ is a pre-softplus variance score and is not interpreted as a mathematical log variance, the recurrent models are trained by Gaussian negative log-likelihood (NLL), checkpoint reporting is track-specific: the six-model point benchmark uses the minimum-validation-RMSE checkpoints, whereas the original probabilistic and mechanism-focused analyses retain the minimum-validation-Gaussian-NLL checkpoints.

## 3.4 Knowledge-guided residual model

The static variables are partitioned as $s = (s^+, s^-, s^0)$, where $s^+$ contains built-up and impervious fractions, $s^-$ contains vegetation fraction, and $s^0$ contains longitude, latitude, and elevation.
Figure 6 summarizes this physically motivated grouping into positively constrained, negatively constrained, and unconstrained contextual variables.

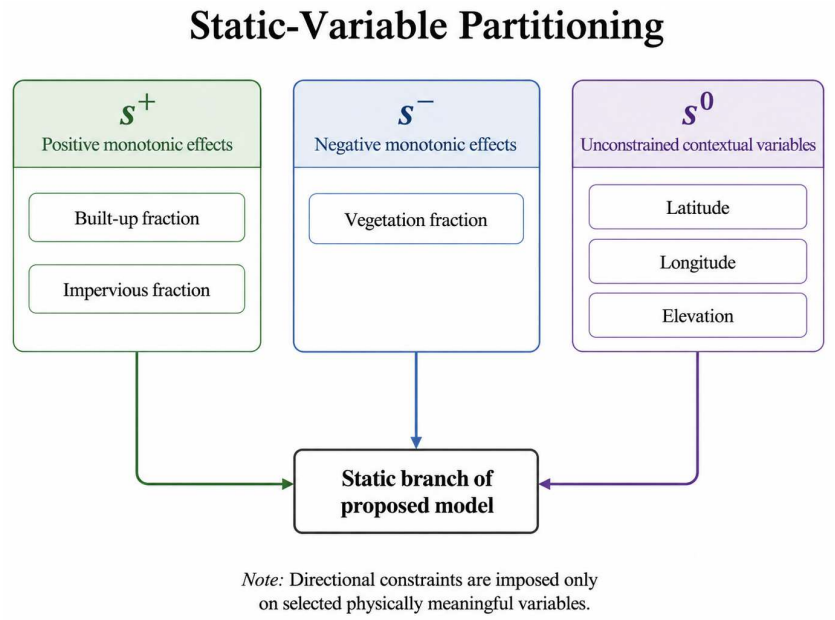



**Fig. 6** Partition of static variables used by the knowledge-guided residual branch.

A residual context is formed from the dynamic embedding and the unconstrained context variables,

$$c = C_\psi([h, U_0(s^0)]). \tag{2}$$

For each constrained fraction, the nonnegative quadratic basis is $m(x) = (x, x^2)$. Softplus-parameterized gates satisfy $\alpha, \beta, \widetilde{\alpha}, \widetilde{\beta} \geq 0$. The structured residuals are

$$r_\mu = \langle \alpha(c), m^+(s^+) \rangle - \langle \beta(c), m^-(s^-) \rangle, \tag{3}$$

$$r_q = \langle \widetilde{\alpha}(c), m^+(s^+) \rangle + \langle \widetilde{\beta}(c), m^-(s^-) \rangle. \tag{4}$$

Positive learnable scales $\lambda_\mu$ and $\lambda_q$ produce

$$\mu(X, s) = \mu_0(X, s) + \lambda_\mu r_\mu, \tag{5}$$

$$q(X, s) = q_0(X, s) + \lambda_q r_q, \qquad v(X, s) = \text{softplus}(q(X, s)) + \varepsilon. \tag{6}$$

The predictive law is $Y \mid X, s \sim \mathcal{N}(\mu(X, s), v(X, s))$.

Figure 7 summarizes the implemented knowledge-guided probabilistic architecture and the separation between the unconstrained base pathway and the structured residual pathway.

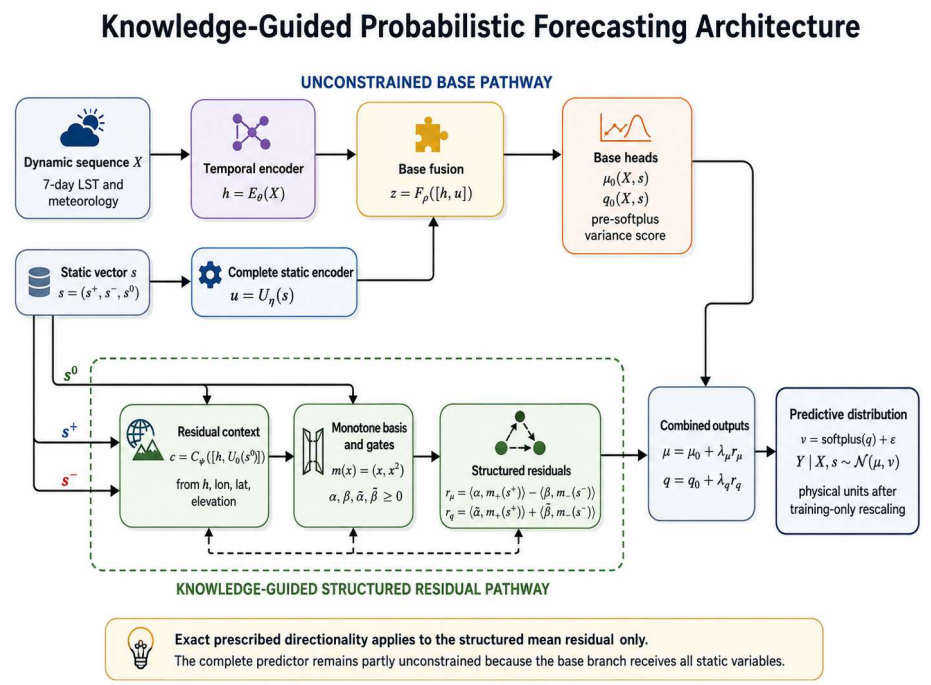



**Fig. 7** Matched-seed knowledge-guided probabilistic architecture.

For fixed context, $r_\mu$ is nondecreasing in built-up and impervious fractions and nonincreasing in vegetation fraction, exact monotonicity is not claimed for the complete mean because $\mu_0$ also receives the complete static vector.

## 3.5 Matched-seed warm-start optimization

Each proposed run is initialized from the baseline checkpoint trained with the same random seed (7, 13, 23, 37, or 73). Optimization uses Adam without weight decay, the baseline configuration uses hidden size 64, dropout 0.1, batch size 256, learning rate $10^{-3}$, gradient clipping at 1.0, a maximum of 30 epochs, and early-stopping patience 10, for the proposed model, the GRU hidden size and residual-context hidden size are

both 64, dropout is 0.1, and the mean and variance residual scales are initialized to 0.1 and learned through their positive parameterization.
The proposed warm-start run has a maximum of 50 total epochs, during the first six epochs, all parameters are frozen except the enabled nonnegative gates and learnable residual scales; this stage uses Adam with learning rate $10^{-3}$, the remaining enabled modules are then unfrozen for joint fine-tuning at learning rate $3 \times 10^{-4}$, a ReduceLROnPlateau scheduler halves the stage-2 learning rate after three validation plateaus down to a minimum of $10^{-5}$, batch size remains 256, gradient clipping is 1.0, weight decay is zero, and early-stopping patience is 10, this matched-seed design removes the earlier shared-checkpoint confound.
The optimization objective is :

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left[ \frac{(y_i - \mu_i)^2}{2v_i} + \frac{1}{2} \log v_i \right] + 10^{-3}(\lambda_\mu + \lambda_q). \tag{7}$$

All normalization statistics are inherited consistently from the matched baseline checkpoint, the official recurrent result artifacts record CPU execution, but the exact historical processor model, operating-system version, and package versions were not preserved; current audit-environment versions are therefore not attributed retrospectively to the original training runs.

# 4 Theoretical Analysis

The analysis is restricted to properties supported by the implemented architecture, the directional guarantee applies exactly to the structured residual correction, the complete mean also contains an unconstrained base forecast that receives the full static vector; therefore, exact end-to-end monotonicity is not claimed, the results below establish residual directionality and sufficient regularity conditions for bounded sensitivity and predictive-distribution continuity.

## 4.1 Assumptions

**Assumption 1** (Compact normalized domain) *The admissible input set $\mathcal{X} \times \mathcal{S}$ is compact.*

**Assumption 2** (Regular base predictor) *The base functions $\mu_0(X, s)$ and $q_0(X, s)$ are Lipschitz on $\mathcal{X} \times \mathcal{S}$.*

**Assumption 3** (Regular residual context) *The dynamic encoder, context-static encoder, and fusion map defining c are Lipschitz on their admissible domains.*

**Assumption 4** (Monotone bounded basis and gates) *The basis maps $m^+$ and $m^-$ are componentwise nondecreasing, Lipschitz, and bounded on the compact fraction domain. The gates $\alpha$, $\beta$, $\widetilde{\alpha}$, and $\widetilde{\beta}$ are nonnegative, Lipschitz, and bounded. The scales $\lambda_\mu$ and $\lambda_q$ are finite and nonnegative.*

Nonnegativity of the gates and residual scales is enforced through softplus parameterization, the compact-domain and Lipschitz assumptions are sufficient analytical conditions for the finite trained modules; the study does not claim to have estimated global Lipschitz constants empirically.

## 4.2 Directional structure of the residual

**Proposition 1** (Directional structure of the residual) *For fixed context $c$, the residual*

$$r_\mu(c, s^+, s^-) = \langle \alpha(c), m^+(s^+) \rangle - \langle \beta(c), m^-(s^-) \rangle \tag{8}$$

*is componentwise nondecreasing in $s^+$ and componentwise nonincreasing in $s^-$.*

*Proof* Every coordinate of $m^+$ and $m^-$ is nondecreasing, while $\alpha(c)$ and $\beta(c)$ are componentwise nonnegative, hence the positive inner product is nondecreasing in $s^+$ and the negative inner product is nonincreasing in $s^-$. □

The proposition concerns only the residual correction, the total mean would inherit exact monotonicity only if the base mean were independent of the corresponding constrained variable or changed in the same direction; neither condition is imposed.

## 4.3 Sensitivity and variance regularity

**Theorem 2** (Sensitivity of the complete mean) *Under Assumptions 1–4, a finite $L_\mu > 0$ exists such that*

$$|\mu(X, s) - \mu(\widetilde{X}, \widetilde{s})| \le L_\mu \big( \|X - \widetilde{X}\| + \|s - \widetilde{s}\| \big). \tag{9}$$

*Proof* The base mean is Lipschitz by assumption, products of bounded Lipschitz basis and gate functions are Lipschitz on a compact domain; finite scaling and addition preserve Lipschitz continuity. □

**Theorem 3** (Variance regularity) *Under the same assumptions, the pre-softplus variance score $q(X, s)$ and predictive variance $v(X, s)$ are Lipschitz on the admissible domain and $v(X, s) \ge \varepsilon > 0$.*

*Proof* The variance-score argument follows the same bounded-product reasoning as the mean case, Softplus is globally 1-Lipschitz, and adding $\varepsilon$ guarantees strict positivity. □

## 4.4 Stability of the predictive distribution

**Theorem 4** (Predictive-distribution stability) *Let $P_{X,s} = \mathcal{N}(\mu(X, s), v(X, s))$. Under Assumptions 1–4, a finite $L_P > 0$ exists such that*

$$W_2(P_{X,s}, P_{\widetilde{X},\widetilde{s}}) \le L_P \big( \|X - \widetilde{X}\| + \|s - \widetilde{s}\| \big). \tag{10}$$

*Proof* For one-dimensional Gaussian distributions,

$$W_2^2(\mathcal{N}(\mu_1, v_1), \mathcal{N}(\mu_2, v_2)) = (\mu_1 - \mu_2)^2 + (\sqrt{v_1} - \sqrt{v_2})^2. \tag{11}$$

The mean and variance are Lipschitz, and $v \geq \varepsilon$ makes the square-root map Lipschitz over the attained range. □

Figure 8 links the formal statements established above to the corresponding architectural components, it distinguishes the exact directional property of the structured residual from the assumption-conditional sensitivity and predictive-distribution continuity results for the complete model.

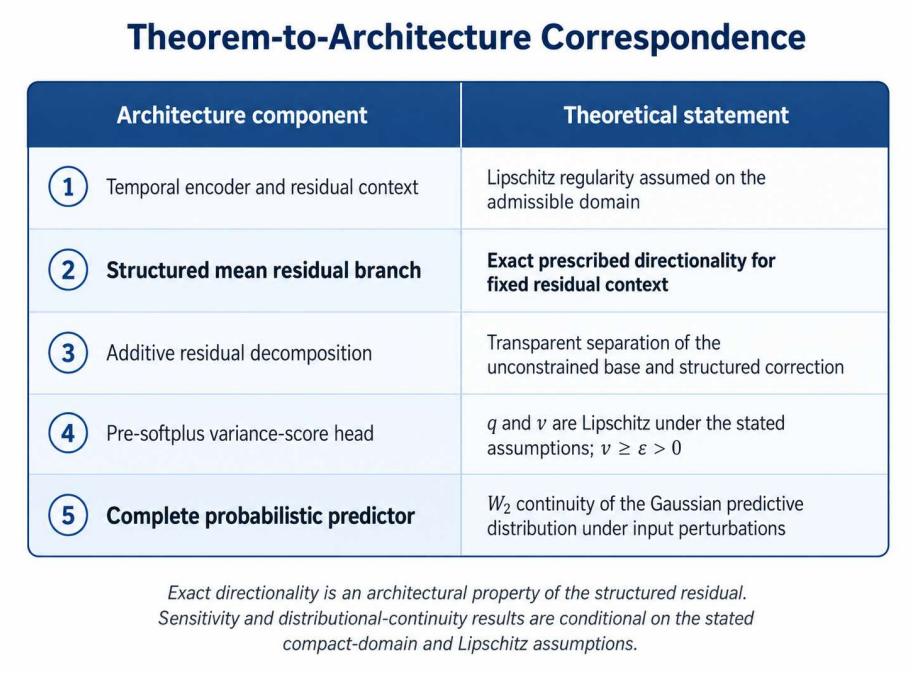



**Fig. 8** Correspondence between architectural components and theoretical statements.

# 5 Experimental Design

## 5.1 Protocol audit and evaluation cohorts

The project initially contained results from two different supervised cohorts, the canonical sequence cohort contains 14,917/4,008/5,977 train/validation/test targets, earlier graph and journal-origin experiments used a broader masked graph cohort containing 185,210/61,620/64,012 targets.

The canonical targets are a strict subset of the graph targets, with 170,293, 57,612, and 58,035 additional graph-only targets. Consequently, metrics from the two cohorts are not ranked as identical-target head-to-head results.

The broader-cohort results are retained only as contextual historical evidence in

Appendix A.
The primary benchmark re-runs four contemporary references on the canonical cohort and combines them with the existing baseline and proposed runs. Exact target identity is verified by city, cell identifier, target date, and target value. All four aligned reference models and all five seeds use identical canonical cohort hashes. Graph-only leakage is zero, checkpoints reload strictly, and test metrics are computed only after validation-based checkpoint selection.
Figure 9 summarizes the two deliberately separated checkpoint-selection tracks used in the final analysis: validation RMSE for the aligned six-model point benchmark and validation Gaussian NLL for the recurrent probabilistic and mechanism-focused analyses.

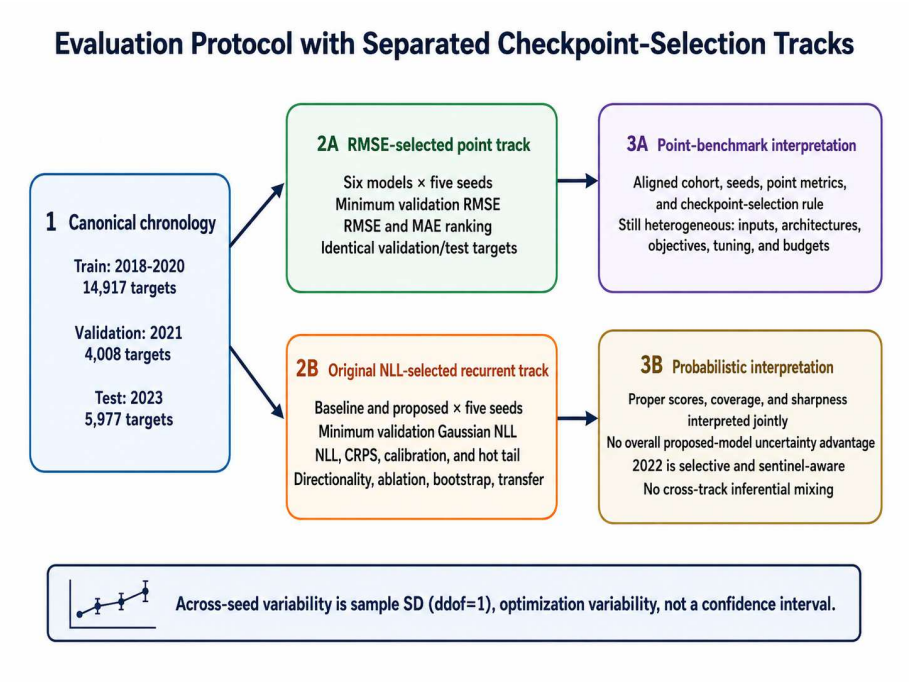



**Fig. 9** Chronological evaluation structure with separated checkpoint-selection tracks.

## 5.2 Aligned six-model point benchmark

The direct comparison contains exactly six models:

1. the probabilistic dynamic+static GRU baseline;
2. the full matched-seed proposed model;
3. an STGCN-style graph/spatial model [34];
4. an MTGNN-style adaptive graph/spatial model [35];
5. a PatchTST-style modern temporal model [36]; and
6. a TimesNet-style modern temporal model [37].

For the six-model point benchmark, all models are evaluated using seeds 7, 13, 23, 37, and 73 on exactly the same canonical target observations, with identical physical-unit RMSE and MAE definitions and minimum validation RMSE checkpoint selection, the probabilistic dynamic+static baseline and the full proposed model use their dedicated RMSE-selected checkpoint tracks for this comparison.
The original validation-Gaussian-NLL-selected baseline and proposed tracks are retained separately for probabilistic and mechanism-focused analyses, including Gaussian NLL, CRPS, calibration, hot-tail evaluation, directionality tests, structural ablations, paired bootstrap inference, transfer, and the secondary 2022 analysis, results from the two checkpoint-selection tracks are not mixed within a single inferential comparison.
The recurrent baseline and proposed model receive the ten dynamic features and all six static features, STGCN-style and MTGNN-style retain the 440-node dynamic graph context but are supervised only on canonical targets and omit the six static descriptors, PatchTST-style and TimesNet-style receive the canonical seven-day ten-feature dynamic sequences and likewise omit the static descriptors, thus, the target observations, seeds, point metrics, and checkpoint-selection criterion are aligned, whereas architecture-specific inputs remain different and are reported explicitly.

## 5.3 Inference, physical-response tests, and ablations

For the original validation-Gaussian-NLL-selected recurrent track, the baseline and proposed models have paired predictions for every test target, a city-stratified moving-block bootstrap over target dates uses 10,000 replicates and block lengths of 7, 14, and 28 dates, the primary block length is 14, and differences are defined as proposed minus baseline, to avoid interpreting a non-significant result as evidence of practical equivalence, confidence intervals are reported alongside explicit analysis margins, following the general distinction between null-hypothesis testing and equivalence reasoning [38]. The declared margins are $\pm 0.10$ °C for RMSE and MAE and $\pm 0.05$ for Gaussian NLL. They are study-specific interpretation thresholds rather than externally validated operational tolerances: 0.10 °C is approximately 3.4% of the baseline test RMSE on this track, while the NLL margin is treated only as a small analysis-scale change relative to the observed NLL level. No formal two-one-sided equivalence test is performed; an interval lying inside these margins is therefore described as consistent with practical closeness, not as proof of statistical equivalence.
Physical-response consistency is evaluated by increasing built-up and impervious fractions and decreasing vegetation fraction by perturbations $\delta \in \{0.01, 0.05, 0.10\}$, holding the remaining inputs fixed, the structured residual and complete predictor are evaluated separately, frozen-checkpoint interventions remove or scale structured components without retraining; independently retrained mean-only and variance-only models provide a complementary mechanism test.

## 5.4 Uncertainty, transfer, and secondary holdout

Probabilistic performance is assessed through Gaussian NLL, CRPS, calibration error, interval width, and interval score on the complete 2023 test set and on the top 10% and

top 5% hottest targets, leave-one-city-out experiments test transfer to a completely unseen city, a sentinel-clean reconstruction of 2022 yields 1,779 valid sequences and is used only as a secondary post-hoc holdout because it is selective and city-imbalanced.

# 6 Results

## 6.1 Data retention and reproducibility

Table 3 summarizes the verified lineage, sentinel cleaning changes the daily-row count but leaves every canonical forecasting sequence exactly unchanged, the retained 440-cell panel is verified, whereas the historical upstream 2,074- and 1,429-cell stages cannot be reconstructed from the available artifacts, within the retained cells, 275 contribute to the selective 2022 holdout and 165 do not, the largest standardized difference between these two within-retained groups concerns 2022 LST availability (SMD 1.566); the largest static differences are substantially smaller (absolute SMDs approximately 0.35–0.41).

**Table 3** Verified data lineage and canonical benchmark counts.

| Statistic | Value |
|---|---|
| Merged daily rows | 964,040 |
| Legacy target-clean rows | 370,780 |
| Sentinel-clean v2 rows | 365,500 |
| Removed invalid target sentinel rows | 5,280 |
| Verified retained cells | 440 |
| Canonical sequences | 24,902 |
| Training / validation / test | 14,917 / 4,008 / 5,977 |
| Canonical legacy–v2 equivalence | Exact |

## 6.2 RMSE-selected six-model predictive benchmark

All six models in Table 4 use the same 4,008 validation targets and 5,977 test targets, the same five random seeds, the same physical-unit RMSE and MAE definitions, and minimum validation RMSE checkpoint selection. The probabilistic dynamic+static baseline achieves the lowest mean test RMSE and MAE.

The full proposed model achieves the second-lowest mean test RMSE and the third-lowest mean test MAE, while MTGNN-style attains the second-lowest mean MAE. STGCN-style is less accurate than the recurrent models, while PatchTST-style and TimesNet-style are weaker in the short seven-day dynamic-only setting used here.

**Table 4** Definitive RMSE-selected five-seed point benchmark on the canonical cohort. Values are mean ± sample SD; lower is better. Seed SD describes optimization variability, not a confidence interval.

| Model | Parameters | Val RMSE | Val MAE | Test RMSE | Test MAE | RMSE rank | MAE rank |
|---|---|---|---|---|---|---|---|
| **Prob. dynamic+static baseline** | 23,426 | 2.668 ± 0.100 | 2.063 ± 0.061 | **2.998 ± 0.043** | **2.313 ± 0.017** | 1 | 1 |
| Full proposed | 32,720 | **2.664 ± 0.121** | **2.058 ± 0.073** | 3.034 ± 0.119 | 2.349 ± 0.090 | 2 | 3 |
| Aligned MTGNN-style | 39,553 | 2.809 ± 0.037 | 2.123 ± 0.045 | 3.055 ± 0.109 | 2.327 ± 0.092 | 3 | 2 |
| Aligned STGCN-style | 25,473 | 2.801 ± 0.095 | 2.168 ± 0.052 | 3.087 ± 0.024 | 2.406 ± 0.023 | 4 | 4 |
| Aligned PatchTST-style | 106,113 | 3.147 ± 0.119 | 2.431 ± 0.046 | 3.365 ± 0.109 | 2.686 ± 0.109 | 5 | 6 |
| Aligned TimesNet-style | 99,841 | 3.041 ± 0.102 | 2.361 ± 0.122 | 3.523 ± 0.150 | 2.668 ± 0.150 | 6 | 5 |

Figure 10 provides the corresponding visual comparison of test RMSE and MAE, with across-seed variability shown descriptively by sample-standard-deviation error bars.

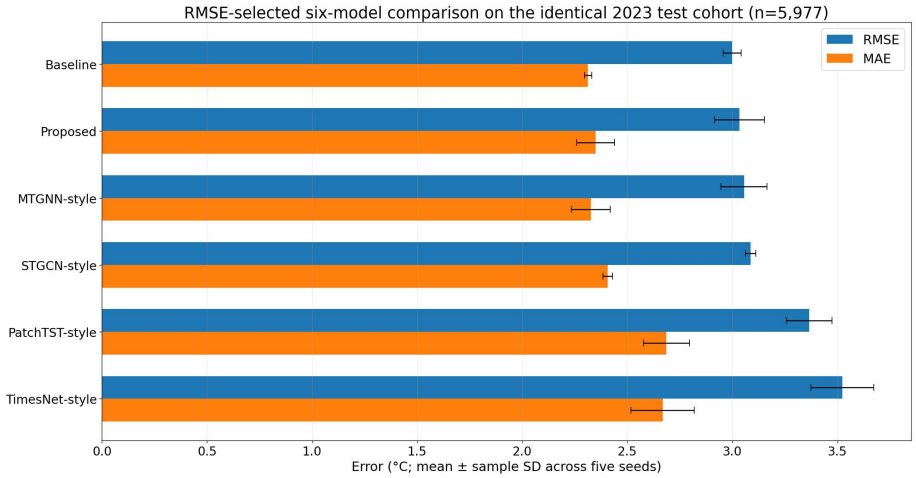



**Fig. 10** RMSE-selected test RMSE and MAE on the identical canonical 2023 target cohort (n = 5,977). Error bars denote sample standard deviation across five training seeds.

The four aligned references retain exact per-seed runtime measurements. Across five seeds, STGCN-style required approximately 275 s of training and 25 s of evaluation, MTGNN-style 268 s and 34 s, PatchTST-style 3,506 s and 204 s, and TimesNet-style 2,278 s and 158 s, respectively, on the available Apple MPS environment. Baseline and proposed runtimes were not recorded in their official artifacts, so no complete six-model speed ranking is made.

## 6.3 Paired baseline–proposed inference on the original NLL-selected track

Table 5 reports the original validation-Gaussian-NLL-selected paired moving-block bootstrap, the 95% intervals for RMSE and Gaussian NLL include zero, whereas the

MAE interval is entirely positive and favors the unconstrained baseline by 0.0336 °C. The complete intervals for all three metrics remain within the declared study-specific practical-closeness margins, with the same qualitative pattern for block lengths of 7, 14, and 28 dates.

**Table 5** Paired city-stratified moving-block bootstrap on the original validation-Gaussian-NLL-selected 2023 predictions (proposed minus baseline; 10,000 replicates; primary block length 14 dates).

| Metric | Difference | 95% CI low | 95% CI high | $p$ | Interpretation |
|---|---|---|---|---|---|
| RMSE | 0.0120 | -0.0207 | 0.0325 | 0.702 | CI includes 0 |
| MAE | 0.0336 | 0.0106 | 0.0550 | 0.003 | Baseline lower |
| Gaussian NLL | 0.0063 | -0.0121 | 0.0161 | 0.779 | CI includes 0 |

The proposed model supports an accuracy–consistency interpretation, with the MAE difference.

## 6.4 Directional consistency and mechanism evidence

The structured mean residual satisfies the prescribed directions for 100% of evaluated test perturbations at $\delta = 0.01$, 0.05, and 0.10, the complete predictor is not globally monotone because the unconstrained base branch can oppose the residual, at the primary $\delta = 0.05$, complete-predictor consistency increases from 35.28% to 42.87% for built-up fraction, from 51.50% to 61.90% for impervious fraction, and from 42.65% to 67.33% for vegetation, built-up consistency remains below 50%, which is reported explicitly rather than hidden by the exact residual result.
Table 6 reports the independently retrained structural ablations on the same canonical test cohort.

**Table 6** Retrained structural ablations and complete-predictor directional consistency at $\delta = 0.05$.

| Model | RMSE | MAE | NLL | CRPS | Built-up | Impervious | Vegetation |
|---|---|---|---|---|---|---|---|
| Baseline | 2.9755 | 2.2936 | 2.5425 | 1.6493 | 35.28% | 51.50% | 42.65% |
| Full proposed | 2.9876 | 2.3272 | 2.5488 | 1.6631 | **42.87%** | **61.90%** | **67.33%** |
| Mean-structure only | 2.9722 | 2.3049 | 2.5496 | 1.6504 | 37.43% | 52.78% | 59.22% |
| Variance-structure only | 2.9751 | 2.2985 | 2.5440 | 1.6516 | 35.73% | 52.10% | 46.96% |

Figure 11 visualizes the corresponding complete-predictor directional-consistency rates for the baseline, full proposed model, and independently retrained structural ablations.

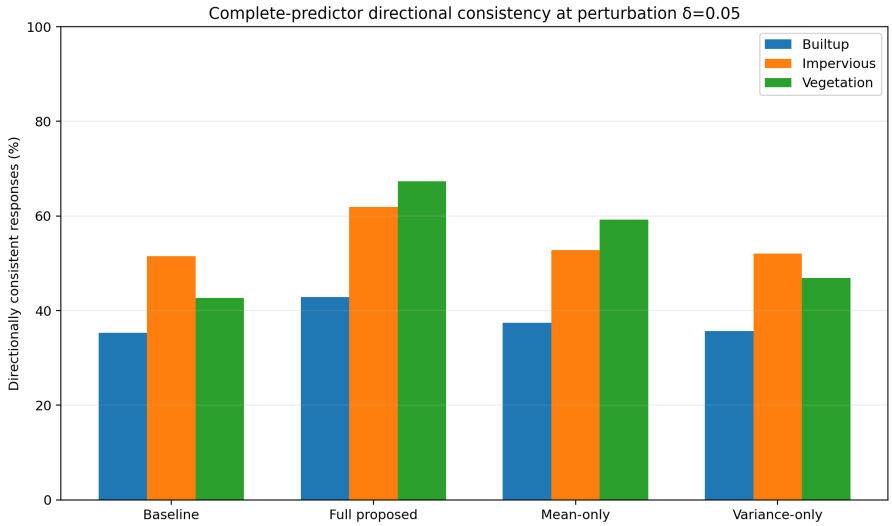



**Fig. 11** Complete-predictor directional consistency for the independently retrained models.

The retrained mean-only model remains numerically close to the baseline on RMSE, MAE, NLL, and CRPS while improving directional consistency by 2.15, 1.28, and 16.57 percentage points for built-up, impervious, and vegetation responses.
The full model shows larger directional-consistency gains than mean-only, but it also has slightly higher RMSE, MAE, and CRPS, because no separate inferential test was performed for these ablation contrasts, the comparison is reported descriptively.
Frozen-checkpoint interventions provide a complementary decomposition, relative to the embedded base branch, the full structured mean raises directional consistency by approximately 10.51, 11.05, and 24.03 percentage points while incurring a 0.0346 °C MAE penalty, removing or scaling individual mean terms changes only the corresponding physical response, increasing the mean-residual scale from zero to one produces a smooth increase in all three consistency measures and a gradual MAE increase (Figure 12).

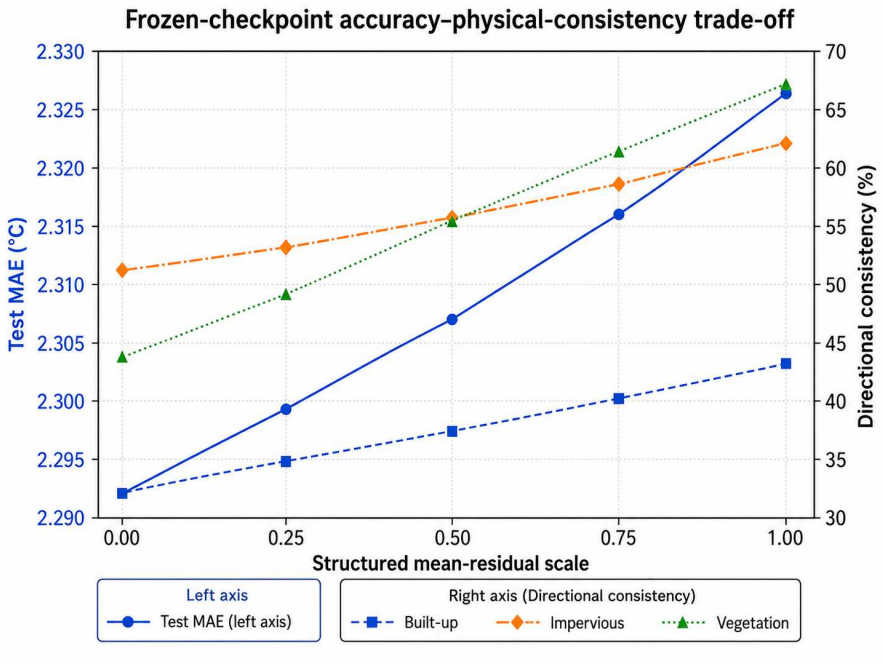



**Fig. 12** Frozen-checkpoint scaling diagnostic.

## 6.5 Uncertainty calibration and extreme heat

On the full test set, the baseline and proposed models have similar calibration summaries (Table 7), the proposed model has slightly lower mean and maximum calibration error, but its intervals are wider and its interval score, CRPS, and NLL are slightly worse.

**Table 7** Full-test probabilistic summary. Lower is better except interval width, which measures sharpness rather than accuracy.

| Model | Mean CE | Max CE | Norm. width | Int. score | CRPS | NLL |
|---|---|---|---|---|---|---|
| Baseline | 0.0230 | 0.0419 | 0.3000 | 10.8730 | **1.6493** | **2.5425** |
| Proposed | **0.0220** | **0.0370** | 0.3059 | 10.8873 | 1.6631 | 2.5488 |

The proposed intervals are wider by approximately 0.145, 0.186, and 0.221 °C at nominal 80%, 90%, and 95% levels, in the hottest subsets, both models under-cover, and the baseline is consistently better on calibration error, interval score, CRPS, and NLL (Table 8).

**Table 8** Probabilistic performance on the hottest test targets.

| Subset | Model | Mean cal. error | Interval score | CRPS | NLL |
|---|---|---|---|---|---|
| Top 10% | Baseline | **0.166** | **14.59** | **2.391** | **2.966** |
| | Proposed | 0.186 | 15.29 | 2.501 | 3.030 |
| Top 5% | Baseline | **0.268** | **19.32** | **2.945** | **3.363** |
| | Proposed | 0.290 | 20.26 | 3.068 | 3.437 |

Figure 13 summarizes the CRPS pattern across the full test cohort and the two hottest target subsets.

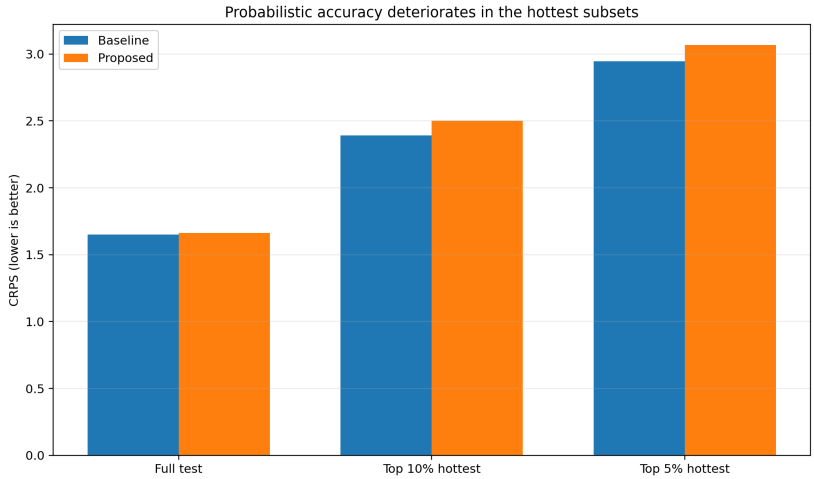



**Fig. 13** CRPS on the complete test set and the hottest subsets.

## 6.6 City heterogeneity and transfer

Under pooled training, the five-seed baseline/proposed mean RMSEs are 3.483/3.509 in Casablanca, 2.872/2.883 in Marrakech, and 2.750/2.754 in Tangier, the proposed model is therefore not clearly stronger in any pooled city after the matched-seed correction, although the differences in Marrakech and Tangier are small.

The leave-one-city-out stress test is mixed (Table 9), the proposed model is close to the baseline for Casablanca and Marrakech and has slightly lower NLL for Casablanca, but it is clearly weaker when Tangier is unseen, within-city pooled performance therefore cannot be interpreted as transfer capability.

**Table 9** Leave-one-city-out transfer to the unseen city.

| Held-out city | Model | $n$ | RMSE | MAE | NLL |
|---|---|---|---|---|---|
| Casablanca | Baseline | 1,350 | 3.560 | 2.890 | 2.982 |
| | Proposed | 1,350 | 3.596 | 2.930 | **2.969** |
| Marrakech | Baseline | 2,219 | 3.361 | 2.601 | 2.790 |
| | Proposed | 2,219 | 3.363 | 2.602 | 2.795 |
| Tangier | Baseline | 2,408 | **3.135** | **2.375** | **2.645** |
| | Proposed | 2,408 | 3.419 | 2.602 | 2.669 |

Figure 14 visualizes the leave-one-city-out RMSE comparison and highlights the larger transfer degradation when Tangier is held out.

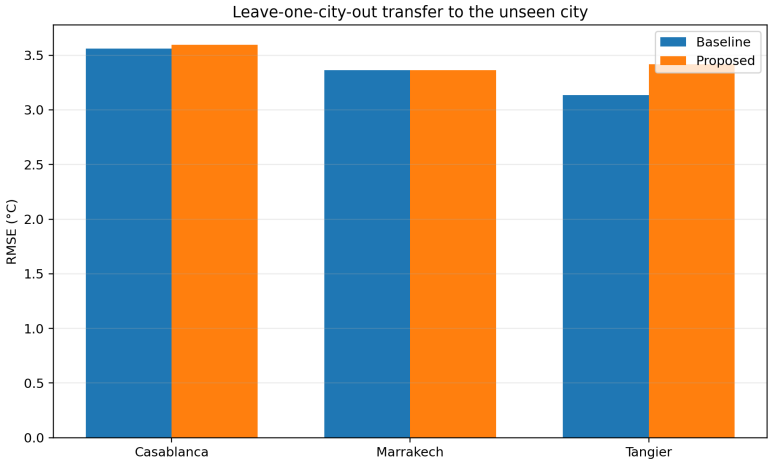



**Fig. 14** Leave-one-city-out RMSE.

## 6.7 Secondary 2022 holdout

The sentinel-clean 2022 reconstruction contains 1,779 valid sequences: 96 from Casablanca, 697 from Marrakech, and 986 from Tangier, only 275 of the 440 retained cells contribute, and the dates are selective rather than year-round.
The holdout is therefore secondary and post hoc, not a replacement for the 2023 primary test, the baseline and proposed results are close: RMSE $3.0876 \pm 0.0292$ versus $3.1043 \pm 0.0458$, MAE $2.4178 \pm 0.0151$ versus $2.4338 \pm 0.0319$, and NLL $2.6654 \pm 0.0671$ versus $2.6647 \pm 0.0490$, both models under-cover, and the analysis does not identify a meaningful winner.

# 7 Discussion

## 7.1 What the aligned benchmark establishes

The corrected benchmark changes the evidential status of the comparative claims. Earlier journal-origin and graph-family results were obtained on a broader masked graph cohort and cannot support direct ranking against the canonical recurrent models. After alignment, all six main models predict the same 5,977 test observations and use minimum validation RMSE checkpoint selection. Under this point-benchmark protocol, the probabilistic dynamic+static baseline is the strongest point-accuracy model, while the proposed model ranks second on mean RMSE and third on mean MAE; MTGNN-style ranks second on mean MAE. The original validation-Gaussian-NLL-selected recurrent track is kept separate and is used for paired dependence-aware inference and probabilistic analyses.

The aligned graph/spatial and temporal models provide useful reference points rather than claims about the original published architectures under exhaustive tuning. MTGNN-style is competitive on MAE, STGCN-style is less accurate than the recurrent models, and the PatchTST-style and TimesNet-style implementations are weaker in this seven-day dynamic-only setting. Seed standard deviations are reported descriptively as optimization variability and are not used for inferential ranking. Because these are compact style implementations with frozen configurations, the findings should not be generalized to longer histories, alternative tuning budgets, additional static inputs, or official reproductions.

## 7.2 Physical structure as an explicit trade-off

The principal positive result is mechanistic rather than rank based, the structured residual satisfies the prescribed directions exactly and improves the complete predictor's response consistency, especially for vegetation and imperviousness, the complete model nevertheless remains non-monotone because the base branch receives all static variables, built-up consistency remains below 50%, showing that the residual correction is not strong enough to override the base response in most built-up perturbations, this limitation is important: the model adds a directionally structured component, but it does not convert the entire forecaster into a shape-constrained network.

Frozen interventions and independent retraining support a coherent mechanism interpretation, the mean residual accounts for most of the physical-response gains, and scaling it produces a smooth accuracy–consistency frontier, the variance residual provides no demonstrated empirical advantage, the mean-only retrained model has slightly lower mean RMSE than the baseline on the preserved NLL-selected track, but those aggregate values are used only as a mechanism diagnostic; no new primary model is selected from the post-hoc ablations, the full proposed configuration remains the pre-specified primary structured model and provides the largest complete-predictor directionality gains.

This positioning also clarifies the relationship to general monotonic and knowledge-guided learning, fully shape-constrained architectures can guarantee monotonicity of

the complete mapping for selected variables [16, 17], by contrast, the present residual design intentionally leaves the base model unconstrained, its benefit is therefore not a stronger theoretical guarantee; it is the ability to introduce a transparent directional correction into an existing probabilistic recurrent forecaster and then quantify the resulting predictive cost and response-consistency benefit.

## 7.3 Probabilistic and hot-tail limitations

Full-test calibration summaries are slightly lower for the proposed model on mean and maximum calibration error, but its intervals are wider and its proper scores are not improved, these descriptive differences do not establish a general uncertainty advantage, in the hottest 10% and 5% of targets, the proposed model is consistently worse on the reported probabilistic diagnostics and both models under-cover, the probabilistic architecture should therefore be described as producing an evaluable predictive distribution, not as improving hot-tail reliability or calibration, tail-aware objectives, quantile models, conformal methods, and city-specific recalibration are natural extensions, but they require prospective evaluation rather than post-hoc tuning on the current test set.

## 7.4 Data and external-validity limitations

The study covers only three Moroccan cities and 440 quality-filtered cells with unequal city representation, the upstream transition from the historical 2,074 initial and 1,429 candidate cells to the verified 440-cell panel cannot be reconstructed because stage-specific identifiers and exclusion records are unavailable, the 2022 holdout is selective and city-imbalanced, and all sequence observations remain spatiotemporally correlated, cross-city transfer is particularly weak for Tangier, these limitations constrain claims about national representativeness and deployment under distribution shift.
The aligned contemporary models are compact style implementations with frozen configurations, not official reproductions or exhaustive tuning exercises, they use the same targets, seeds, checkpoint criterion, and point metrics, but the graph and temporal references omit the six static features used by the recurrent models.
The benchmark therefore establishes exact target comparability, not equal information content or equal optimization budgets, similarly, historical hardware and software details are incomplete for the recurrent runs, so reproducibility is strongest at the level of retained code, configurations, cohort hashes, checkpoints, and outputs rather than full computational-environment recreation.

## 7.5 Implications and future work

The evidence supports the use of physically structured probabilistic forecasting when directional response behavior is itself a scientific objective and a small point-accuracy cost is acceptable, future work should recover or rebuild an auditable upstream spatial-cell selection ledger, extend the panel to additional Moroccan climates and years, evaluate richer urban morphology and local climate zones, and develop city-adaptive or hierarchical transfer, a base branch that excludes the constrained variables would permit exact end-to-end directional guarantees, bringing the architecture closer to fully

shape-constrained learning, tail-aware probabilistic objectives and formal recalibration should be evaluated on pre-specified future periods rather than repeatedly inspected test subsets.

# 8 Conclusion

This study evaluates knowledge-guided probabilistic next-day LST forecasting in Casablanca, Marrakech, and Tangier, a sentinel-clean audit verifies 365,500 valid daily target observations from 440 retained cells and exactly reproduces 24,902 canonical chronological forecasting sequences, the six-model point benchmark is aligned on target cohort, seeds, physical-unit point metrics, and minimum validation RMSE checkpoint selection.

The probabilistic dynamic+static baseline is the strongest point-accuracy model, with the lowest mean test RMSE and MAE, the proposed model ranks second in RMSE and third in MAE, on the separately preserved validation-Gaussian-NLL-selected recurrent track, the 95% paired moving-block-bootstrap intervals include zero for RMSE and Gaussian NLL, while MAE favors the baseline by 0.0336 °C, all three intervals fall inside the declared study-specific practical-closeness margins, which is reported as descriptive evidence of closeness rather than a formal equivalence-test result, the structured mean residual nevertheless satisfies its directional design exactly and improves complete-predictor consistency, most strongly for vegetation and imperviousness, no general calibration, hot-tail, or cross-city advantage is established.

The contribution is therefore an evidence-bounded demonstration that directional prior structure can be added to a strong probabilistic urban-heat forecaster while explicitly measuring its predictive cost, response-consistency benefit, and uncertainty limitations, the work does not claim a new general monotonic-learning theory, complete-predictor monotonicity, national representativeness, or predictive and uncertainty superiority.

# CRediT authorship contribution statement

**Merieme Debbagh:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review & editing.

**Youness Madani:** Conceptualization, Methodology, Supervision, Writing – review & editing.

**Belaid Bouikhalene:** Methodology, Supervision, Project administration, Writing – review & editing.

# Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Data availability

The processed data, canonical forecasting sequences, source code, configuration files, and analysis artifacts supporting the findings of this study are available from the corresponding author upon reasonable request. The underlying MODIS/Terra MOD11A1 and ERA5-Land products are publicly available through Google Earth Engine.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Declaration of generative AI

During the preparation of this work, the authors used OpenAI ChatGPT to assist with language refinement, organization, and improvement of manuscript clarity, after using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

## Appendix A Historical different-cohort context

The reproducibility audit identified 13 completed historical forecasting references evaluated on the broader masked graph-target cohort (185,210/61,620/64,012 targets) rather than the canonical sequence cohort (14,917/4,008/5,977 targets), the canonical targets are a strict subset of the graph cohort, these historical metrics remain useful for implementation debugging and contextual sensitivity, but they are not included in the main numerical ranking and are not used to claim that the proposed model outperforms journal-origin or modern forecasting families; the aligned STGCN-style, MTGNN-style, PatchTST-style, and TimesNet-style experiments replace those cross-cohort comparisons in the main paper.

## Appendix B Frozen mean-residual scaling diagnostic

Table B1 reports the complete frozen-checkpoint mean-residual scaling values underlying the mechanism diagnostic discussed in the main text.

**Table B1** Frozen-checkpoint mean-residual scaling diagnostic. The scale was not selected using the test set.

| Scale | Test RMSE | Test MAE | Built-up | Impervious | Vegetation |
|---|---|---|---|---|---|
| 0.00 | 2.9730 | 2.2926 | 32.37% | 50.86% | 43.29% |
| 0.25 | 2.9744 | 2.2991 | 34.74% | 53.22% | 49.07% |
| 0.50 | 2.9774 | 2.3071 | 37.35% | 55.89% | 55.38% |
| 0.75 | 2.9817 | 2.3165 | 40.11% | 58.76% | 61.38% |
| 1.00 | 2.9876 | 2.3272 | 42.87% | 61.90% | 67.33% |

## References


[1] Wang, H., Yang, J., Chen, G., Ren, C., Zhang, J.: Machine learning applications on air temperature prediction in the urban canopy layer: A critical review of 2011–2022. Urban Climate **49**, 101499 (2023) https://doi.org/10.1016/j.uclim.2023.101499

[2] Ahmed, A.N., AlDahoul, N., Aziz, N.A., Huang, Y.F., Sherif, M., El-Shafie, A.: The urban heat island effect: A review on predictive approaches using artificial intelligence models. City and Environment Interactions **28**, 100234 (2025) https://doi.org/10.1016/j.cacint.2025.100234

[3] Snaiki, R., Merabtine, A.: Recent advances on machine learning techniques for urban heat island applications: A review and new horizons. Sustainable Cities and Society **134**, 106943 (2025) https://doi.org/10.1016/j.scs.2025.106943

[4] Gaur, A., Deb, C.: Machine learning methods and approaches for urban heat island (uhi) assessment: A comprehensive review. Renewable and Sustainable Energy Reviews **234**, 116903 (2026) https://doi.org/10.1016/j.rser.2026.116903

[5] Tanoori, G., Soltani, A., Modiri, A.: Machine learning for urban heat island (uhi) analysis: Predicting land surface temperature (lst) in urban environments. Urban Climate **55**, 101962 (2024) https://doi.org/10.1016/j.uclim.2024.101962

[6] Fu, S., Wang, L., Khalil, U., Cheema, A.H., Ullah, I., Aslam, B., Tariq, A., Aslam, M., Alarifi, S.S.: Prediction of surface urban heat island based on predicted consequences of urban sprawl using deep learning: A way forward for a sustainable environment. Physics and Chemistry of the Earth, Parts A/B/C **135**, 103682 (2024) https://doi.org/10.1016/j.pce.2024.103682

[7] Fan, C., Zou, B., Li, J., Wang, M., Liao, Y., Zhou, X.: Exploring the relationship between air temperature and urban morphology factors using machine learning under local climate zones. Case Studies in Thermal Engineering **55**, 104151 (2024) https://doi.org/10.1016/j.csite.2024.104151

[8] Romero Rodríguez, L., Guerrero Delgado, M.C., Castro Medina, D., Sánchez Ramos, J., Álvarez Domínguez, S.: Forecasting urban temperatures through crowdsourced data from citizen weather stations. Urban Climate **57**, 102021 (2024) https://doi.org/10.1016/j.uclim.2024.102021

[9] Castro Medina, D., Guerrero Delgado, M.C., Sánchez Ramos, J., Palomo Amores, T., Romero Rodríguez, L., Álvarez Domínguez, S.: Empowering urban climate resilience and adaptation: Crowdsourcing weather citizen stations-enhanced temperature prediction. Sustainable Cities and Society **101**, 105208 (2024) https://doi.org/10.1016/j.scs.2024.105208

[10] Tehrani, A.A., Veisi, O., Kia, K., Delavar, Y., Bahrami, S., Sobhaninia, S., Mehan,

A.: Predicting urban heat island in european cities: A comparative study of gru, dnn, and ann models using urban morphological variables. Urban Climate **58**, 102061 (2024) https://doi.org/10.1016/j.uclim.2024.102061

[11] Li, H., Cai, Y., Yang, Y., Cai, X.: Assessing the influence of urban scene characteristics on urban heat island: An interpretable machine learning approach in new york city. Urban Climate **62**, 102542 (2025) https://doi.org/10.1016/j.uclim.2025.102542

[12] Liu, Q., Hang, T., Yunfei, W., *et al.*: Unveiling differential impacts of multi-dimensional urban morphology on heat island effect across local climate zones: Interpretable catboost-shap machine learning model. Building and Environment **275**, 112574 (2025) https://doi.org/10.1016/j.buildenv.2025.112574

[13] Zhang, M., Zhou, D., Xu, D., Liu, K., Gu, Z., Yang, Y., Zhang, Q., Chen, L.: How do built environment factors influence urban heat islands across local climate zones? evidence from an interpretable xgboost-shap model. Sustainable Cities and Society **136**, 107078 (2026) https://doi.org/10.1016/j.scs.2025.107078

[14] Karpatne, A., Atluri, G., Faghmous, J.H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., Kumar, V.: Theory-guided data science: A new paradigm for scientific discovery from data. IEEE Transactions on Knowledge and Data Engineering **29**(10), 2318–2331 (2017) https://doi.org/10.1109/TKDE.2017.2720168

[15] Willard, J., Jia, X., Xu, S., Steinbach, M., Kumar, V.: Integrating scientific knowledge with machine learning for engineering and environmental systems. ACM Computing Surveys **55**(4), 1–37 (2022) https://doi.org/10.1145/3514228

[16] You, S., Ding, D., Canini, K., Pfeifer, J., Gupta, M.R.: Deep lattice networks and partial monotonic functions. In: Advances in Neural Information Processing Systems 30, pp. 2981–2989 (2017)

[17] Runje, D., Shankaranarayana, S.M.: Constrained monotonic neural networks. In: Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 29338–29353 (2023)

[18] Zhang, M., Fernández-Torres, M.-Á., Cohrs, K.-H., Camps-Valls, G.: Calibration and uncertainty quantification for deep learning-based drought detection. International Journal of Applied Earth Observation and Geoinformation **140**, 104563 (2025) https://doi.org/10.1016/j.jag.2025.104563

[19] Rahmani, N., Sharifi, A.: Urban heat dynamics in local climate zones (lczs): A systematic review. Building and Environment **267**, 112225 (2025) https://doi.org/10.1016/j.buildenv.2024.112225

[20] Zhang, M., Tan, S., Zhang, C., *et al.*: Machine learning in modelling the urban

thermal field variance index and assessing the impacts of urban land expansion on seasonal thermal environment. Sustainable Cities and Society **106**, 105345 (2024) https://doi.org/10.1016/j.scs.2024.105345

[21] Tahooni, A., Kakroodi, A.A., Kiavarz, M., Mansourian, H.: High-resolution urban lst downscaling via machine learning and shap: A case study in a rapidly urbanizing semi-arid region. Sustainable Cities and Society **134**, 106897 (2025) https://doi.org/10.1016/j.scs.2025.106897

[22] Chen, Y.-C., Han, J.-Y.: Modeling high-resolution air temperature variability with multi-source data: Predictive insights for urban heat mitigation strategies. Sustainable Cities and Society **131**, 106788 (2025) https://doi.org/10.1016/j.scs.2025.106788

[23] Shen, P.: Spatiotemporal mapping of urban air temperature and uhi under tmy condition: A reference station based machine learning approach. Energy and Buildings **343**, 115923 (2025) https://doi.org/10.1016/j.enbuild.2025.115923

[24] Sampaio, R.J., Rodriguez, D.A., Espíndola, R.P., Silva, F.P.: Spatiotemporal assessment and monitoring of urban heat islands in metropolitan areas using machine learning and downscaling. Sustainable Cities and Society **126**, 106365 (2025) https://doi.org/10.1016/j.scs.2025.106365

[25] Lau, T.-K., Chen, Y.-C., Lin, T.-P.: Application of local climate zones combined with machine learning to predict the impact of urban structure patterns on thermal environment. Urban Climate **52**, 101731 (2023) https://doi.org/10.1016/j.uclim.2023.101731

[26] Lau, T.-K., Lin, T.-P., *et al.*: Investigating the relationship between air temperature and the intensity of urban development using on-site measurement, satellite imagery and machine learning. Sustainable Cities and Society **98**, 104982 (2024) https://doi.org/10.1016/j.scs.2023.104982

[27] Alghamdi, A.S.: An integrated ml-powered geospatial analysis of surface urban heat island and its mitigation in riyadh city, saudi arabia. Urban Climate **64**, 102642 (2025) https://doi.org/10.1016/j.uclim.2025.102642

[28] Hussainzad, E.A., Gou, Z.: A machine learning framework for urban heat mitigation in informal settlements: Climate-resilient planning in kabul, afghanistan. Sustainable Cities and Society **133**, 106845 (2025) https://doi.org/10.1016/j.scs.2025.106845

[29] Wang, S.-Y., Ou, H.-Y., Chen, P.-C., Lin, T.-P.: Implementing policies to mitigate urban heat islands: Analyzing urban development factors with an innovative machine learning approach. Urban Climate **55**, 101868 (2024) https://doi.org/10.1016/j.uclim.2024.101868

[30] Zhong, Q., Li, Z., Zhu, J., Yuan, C.: Revealing multiscale and nonlinear effects of urban green spaces on heat islands in high-density cities: Insights from mspa and machine learning. Sustainable Cities and Society **120**, 106173 (2025) https://doi.org/10.1016/j.scs.2025.106173

[31] Wang, J., Liu, Z.: Mitigating urban heat exposure inequity with a seasonally adaptive machine-learning approach. Sustainable Cities and Society **136**, 107063 (2025) https://doi.org/10.1016/j.scs.2025.107063

[32] Wang, H., Yi, T., Lu, Y., Wang, Y., Wu, J.: Patterns of nighttime surface urban heat island patch in mega urban agglomerations: A case study in the pearl river delta, china. Sustainable Cities and Society **128**, 106465 (2025) https://doi.org/10.1016/j.scs.2025.106465

[33] Shi, Y., Mao, Y., Zhou, L.: Research on the relationship between public open space and the thermal environment in shanghai based on multisource data integration. Sustainable Cities and Society **126**, 106415 (2025) https://doi.org/10.1016/j.scs.2025.106415

[34] Yu, B., Yin, H., Zhu, Z.: Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, pp. 3634–3640 (2018). https://doi.org/10.24963/ijcai.2018/505

[35] Wu, Z., Pan, S., Long, G., Jiang, J., Chang, X., Zhang, C.: Connecting the dots: Multivariate time series forecasting with graph neural networks. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 753–763 (2020). https://doi.org/10.1145/3394486.3403118

[36] Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J.: A time series is worth 64 words: Long-term forecasting with transformers. In: International Conference on Learning Representations (2023)

[37] Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., Long, M.: Timesnet: Temporal 2d-variation modeling for general time series analysis. In: International Conference on Learning Representations (2023)

[38] Lakens, D.: Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. Social Psychological and Personality Science **8**(4), 355–362 (2017) https://doi.org/10.1177/1948550617697177