

Article

Benchmarking Model Complexity for Short-Term Surrogate Forecasting of High-Resolution Urban WRF Outputs During an Extreme Heatwave in Chongqing

Yanan Liu ¹, Maoyuan Chai ¹, Runjie Xie ¹, Hong Li ^{2,3,*} , Ruiqing Du ^{4,*} and Bao-Jie He ^{2,3,5,*} ¹ School of Architecture and Urban Planning, Chongqing Jiaotong University, Chongqing 400074, China² School of Architecture and Urban Planning, Chongqing University, Chongqing 400045, China³ Key Laboratory of New Technology for Construction of Cities in Mountain Area, Ministry of Education, Chongqing University, Chongqing 400045, China⁴ Atmosphere, Earth and Energy Division, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA⁵ School of Architecture, Design and Planning, The University of Queensland, Brisbane 4072, Australia

* Correspondence: lihom@cqu.edu.cn (H.L.); du5@llnl.gov (R.D.); baojie.he@cqu.edu.cn (B.-J.H.)

Abstract

High-resolution urban Weather Research and Forecasting (WRF) simulations resolve interactions among urban form, near-surface meteorology, atmospheric dynamics, and building-energy processes, but their computational cost limits repeated scenario analysis. This study compares eight pointwise temporal surrogates (Persistence, previous-day same-hour, Ridge, Random Forest, XGBoost, multilayer perceptron (MLP), gated recurrent unit (GRU), and long short-term memory (LSTM)) for recursive 24 h prediction of T_2 , Q_2 , W_{10} , and WRF-BEM diagnostic air-conditioning power density (AC) from a 120 h WRF-BEP/BEM-LCZ heatwave simulation over Chongqing. Model development used three expanding-window folds, fold-specific scaling, fixed monitoring cells, and five fixed seeds for stochastic families; the h97–h120 full-domain test remained untouched until all modeling choices had been finalized. Station comparisons provided a limited check on the physical plausibility of the WRF meteorological fields; AC was not validated against metered energy use. Model performance depended strongly on the target variable. Random Forest performed best for T_2 (1.081 ± 0.002 °C), LSTM for Q_2 (0.969 ± 0.063 g/kg), and MLP for AC (0.505 ± 0.187 W/m²), whereas deterministic Ridge was best for W_{10} (1.095 m/s). W_{10} had weaker temporal memory (median lag-1/lag-24 ACF 0.581/0.211). Ridge outperformed the Previous-day baseline for 75.0% of grid cells, although errors increased sharply in the top wind-speed decile and during rapid changes. For this single-event, single-city benchmark, recurrent models did not improve W_{10} accuracy over Ridge. Model choice depended on the target and the balance between accuracy, spatial fidelity, and computational cost; transfer to other events and meteorological regimes remains untested.



Academic Editor: Thomas Panagopoulos

Received: 15 August 2026

Revised: 10 September 2026

Accepted: 13 September 2026

Published: 16 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: urban WRF; surrogate modeling; model complexity; heatwave; temporal backtesting; Local Climate Zones; recursive forecasting

1. Introduction

Urbanization replaces vegetated and permeable land with buildings, roads, and other engineered surfaces. These changes alter radiative exchange, heat storage, aerodynamic roughness, turbulent transport, and the partitioning of energy into sensible and latent heat. They contribute to the urban heat island and can intensify thermal stress during heatwaves,

particularly where dense built form, limited evapotranspiration, and anthropogenic heat coincide. The Local Climate Zone (LCZ) framework classifies urban and natural landscapes using physically interpretable surface, structural, and thermal properties [1]. The World Urban Database and Access Portal Tools (WUDAPT) initiative has also promoted standardized urban morphological data for weather, climate, and environmental modeling [2]. A subsequent global 100 m LCZ map made consistent urban-form information more widely available for Earth-system and urban environmental applications [3]. The Weather Research and Forecasting (WRF) model can combine these data with mesoscale dynamics and urban canopy representations to simulate urban–atmosphere interactions [4].

High-resolution urban simulations are computationally expensive to repeat, so surrogate models may enable faster analysis of their outputs. Yet greater model complexity increases computational cost without necessarily improving predictions for every target variable. Here, we compare eight temporal surrogate families for short-term prediction during an extreme heatwave in Chongqing. We aim to assess the gains and limitations of complexity in a controlled case study, rather than to establish a general ranking across urban climates.

2. Literature Review

Urban WRF performance depends strongly on the representation of urban form and land-surface heterogeneity. Early developments in urban canopy modeling introduced physically based urban energy budgets [5], canyon-resolving single-layer formulations [6], and multilayer surface-exchange parameterizations [7]. These schemes underpin many current mesoscale urban models. Multilayer urban canopy schemes and the Building Effect Parameterization/Building Energy Model (BEP/BEM) explicitly describe exchanges involving roofs, walls, streets, indoor thermal conditions, and building energy demand [8,9]. LCZ-informed WRF configurations have improved the representation of intra-urban thermal contrasts during heatwaves [10]. Urban morphology and land-use specifications can materially affect simulated near-surface meteorology and heat-stress fields [11], and enhanced urban representation can alter the simulated structure of large-scale heatwaves [12]. Recent WRF-BEP/BEM applications have also used detailed urban physics to assess metropolitan heat-mitigation measures [13]. Even local modifications to buildings can affect community microclimate [14]. In a mountainous city, this built-form heterogeneity coexists with terrain-driven differences in local circulation. Representing these processes also increases computational cost, limiting the number of scenarios, sensitivity experiments, and rapid-response assessments that can be performed with the full physical model.

Building-energy feedback adds scientific value to urban climate simulations but also increases their computational burden. Air-conditioning systems remove heat from indoor spaces and reject it outdoors, creating a two-way climate–energy interaction that can reinforce urban warming during hot weather [15]. WRF-BEP/BEM has been evaluated against measurements of building energy use and can reproduce city-scale variability in energy demand when the relevant urban parameters are adequately represented [16]. Earth-system modeling has also emphasized spatial variation in air-conditioning adoption and its contribution to anthropogenic heat [17]. A surrogate for urban WRF-BEP/BEM outputs should therefore be assessed for both near-surface meteorology and model-diagnosed quantities such as instantaneous air-conditioning load or power density. The aim is to retain the temporal and spatial information of the physical model while reducing the cost of repeated evaluation.

Urban thermal research ranges from localized heating strategies for personal comfort [18] to spatial assessment of heat exposure. Recent machine-learning studies have predicted urban land-surface temperature from urban configuration and land cover [19],

efficiently estimated high-resolution urban microclimates for adaptation assessment [20], and quantified nonlinear associations between multidimensional urban morphology and heat risk across hundreds of cities [21]. These applications identify relationships among landscape structure, thermal conditions, and heat exposure, but they do not necessarily emulate the temporal evolution of a process-based atmospheric model. In numerical-model emulation, the training target is the output of an expensive simulation, and the task is to reproduce the model response under a specified forecast protocol. Accordingly, this study treats WRF outputs as surrogate reference targets, rather than observations or observational ground truths.

Applications of machine learning in weather, climate, and Earth-system modeling now include forecasting, emulation, parameter estimation, and extreme-event analysis [22]. Hybrid approaches have trained machine-learning components within a general circulation model and achieved stable weather and climate simulations [23]. Earlier work on random-forest parameterization also showed that relatively simple algorithms can remain stable in prognostic settings [24]. For urban applications, Meyer et al. [25] developed a neural-network emulator that reproduced the collective behavior of several urban land-surface models and operated within WRF. Other work has used generative methods for reliable high-resolution climate and weather downscaling [26], Fourier neural operators to super-resolve near-surface climate fields [27], and WRF-ML to connect WRF with machine-learning parameterizations [28]. WeatherBench 2 formalized common evaluation data, operationally informed metrics, and like-for-like comparison for data-driven weather prediction [29]. These studies cover parameterization replacement, field downscaling, model-response approximation, and prognostic emulation. The choice of architecture should therefore reflect the information available to the model and the intended evaluation.

Whether added model complexity reliably improves surrogate skill remains unresolved. Recent atmospheric and land-surface studies suggest that such improvements are not consistent. In a WRF/Noah-MP sensitivity-analysis framework, Di Santo et al. [30] found that the performance and robustness of several machine-learning families depended on the response being approximated, rather than increasing monotonically with complexity. Wesselkamp et al. [31] reported similarly task-dependent behavior when comparing LSTM, gradient-boosting, and feed-forward neural-network emulators for prognostic land-surface states. Beucler et al. [32] examined this issue using error-complexity Pareto fronts and model hierarchies. Lütjens et al. [33] further showed that internal variability can change the apparent ranking of linear and deep-learning climate emulators, especially for noisy variables. This concern is particularly relevant in recursive forecasting because a model optimized to minimize one-step error may lose accuracy when its predictions are fed back as inputs. A fair comparison therefore requires strong simple baselines, identical information content, leakage-free temporal evaluation, repeated runs for stochastic models, and explicit measurement of computational cost.

The development of large data-driven weather models has expanded the range of architectures available for emulation. Pangu-Weather used three-dimensional neural networks for global medium-range prediction [34], while GraphCast learned global atmospheric evolution on a graph mesh [35]. HRRRCast extended data-driven emulation to convection-allowing regional forecasts [36], and StormCast used generative diffusion methods for kilometer-scale convective evolution conditioned on detailed atmospheric states [37]. Graph methods also support other atmospheric tasks; GATES, for example, emulates transport from meteorological and surface information [38]. An urban canopy model has also been trained across diverse building morphologies and meteorological forcing to reproduce urban land-atmosphere exchanges [39]. Together with neural-operator methods for spatial climate fields [27], these studies address tasks that include spatial

connections, multivariable states, or training archives much larger than those available for the single-event pointwise forecast examined here.

Evaluation design is as important as the choice of architecture. Domain-mean RMSE alone cannot show whether a surrogate for short-term urban WRF forecasts derives its skill from temporal persistence, accumulates error with lead time, preserves spatial structure, or fails in the hot and high-load conditions of greatest interest during a heatwave. Domain averages can also conceal large differences among LCZs and land-cover classes. Weather-Bench 2 [29] and probabilistic AI-weather benchmarking [40] emphasize explicit baselines and evaluation protocols rather than reliance on a single summary score. When models have similar accuracy, computational cost may determine their practical value for scenario exploration or rapid screening. The land-surface comparison reported in [31] and the Pareto framework in [32] support an evaluation that combines temporal backtesting, lead-hour error, spatial-field fidelity, LCZ-stratified results, critical regimes, stochastic stability, and accuracy-complexity trade-offs.

The relative merits of these model families remain less clear for recursive forecasting of high-resolution urban WRF-BEP/BEM outputs during a single extreme-heat event. Comparisons in other areas of Earth-system modeling show why this question matters. Tree ensembles and neural networks can have different strengths in process emulation [41]; emulator-based inference can support land-surface model calibration [42]; physically structured Bayesian emulators can combine statistical flexibility with process constraints [43]; and machine-learning emulators may augment expensive regional climate simulations [44]. Urban machine-learning studies often relate land-surface or morphological predictors to outcomes [19–21], whereas advanced weather emulators increasingly rely on multivariate spatial fields and large training archives [34–39].

Short, event-focused simulations also play a role in WRF evaluation. Bhimireddy and Bhaganagar [45] assessed dynamic downscaling using observations from targeted convective periods in the Sagebrush tracer experiments and tested sensitivity to model configuration. That work provides methodological context for process-specific case studies, rather than evidence about urban, suburban, and rural land-use effects. Event-level results still require separate testing before they can support conclusions about seasonal climate or conditions at other locations.

This study asks whether more complex temporal models improve recursive prediction when each model receives the preceding 24 h of one WRF variable at the same grid cell. Eight families are compared for T_2 , Q_2 , W_{10} , and the WRF-BEM AC diagnostic in one 120 h Chongqing heatwave simulation. The analysis examines overall accuracy, changes with forecast lead time, spatial fidelity, and computational cost. The following sections describe the common development protocol and address these questions using a held-out 24 h period; the resulting rankings are specific to this case.

3. Materials and Methods

The workflow combines WRF simulation and evaluation, leakage-controlled surrogate development, and an independent full-domain test (Figure 1). Station observations provide a check on the physical plausibility of WRF, whereas surrogate models are trained and evaluated against held-out WRF output. Model structures and hyperparameters are finalized using h1–h96 before the untouched h97–h120 test period is accessed.

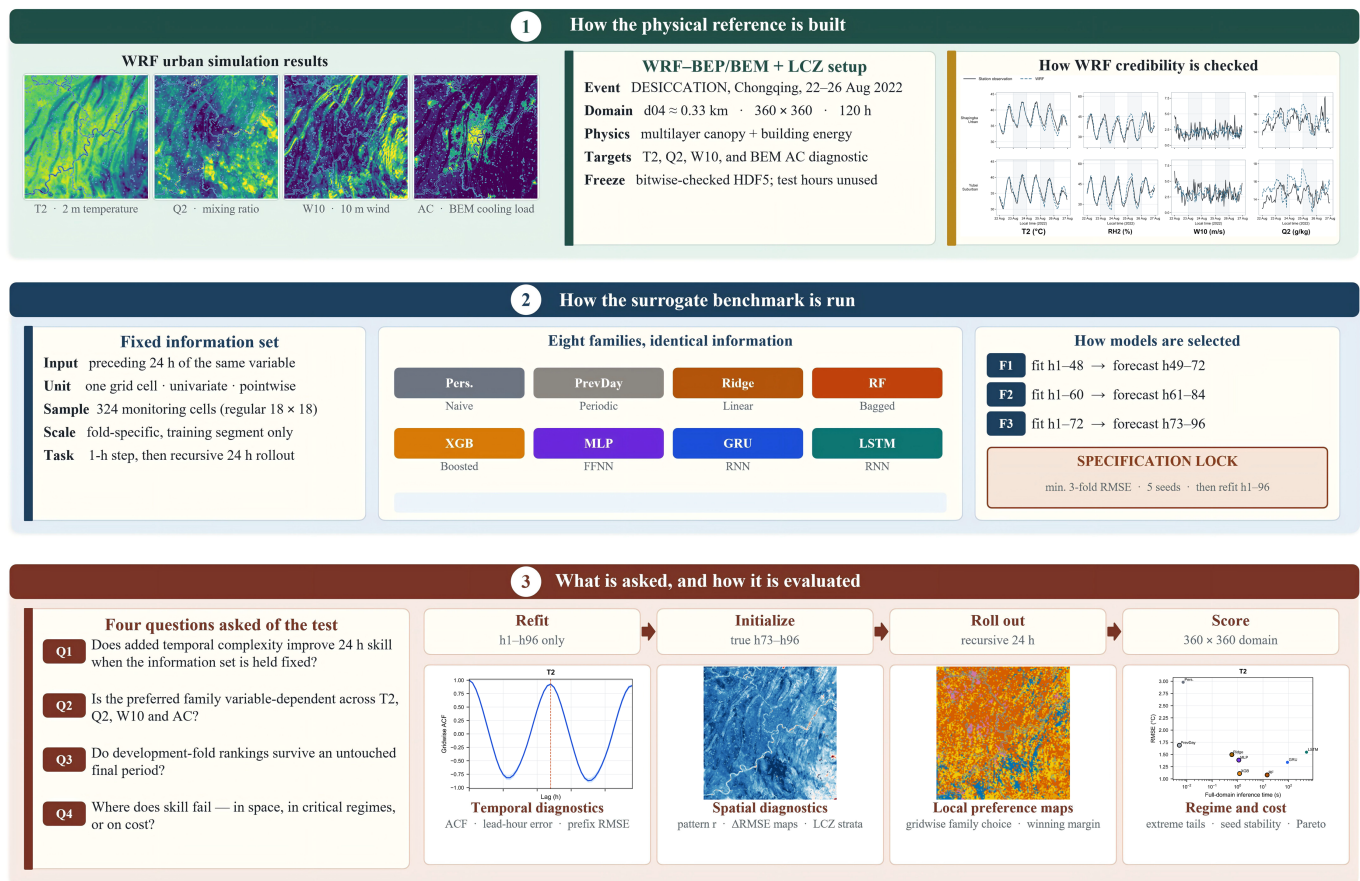


Figure 1. Study workflow.

3.1. August 2022 Heatwave Event and WRF Dataset

We used a WRF simulation of the August 2022 Chongqing heatwave covering the period from 00:00 local time on 22 August to 23:00 on 26 August 2022, a continuous 120 h period. The innermost urban domain had 360×360 grid cells with a spacing of approximately 0.33 km. Its urban configuration combines WRF with the Building Effect Parameterization/Building Energy Model and an LCZ-informed representation [4,8–10]. The surrogate targets are 2 m air temperature (T_2 , °C), 2 m water-vapor mixing ratio (Q_2 , g/kg), 10 m wind speed (W_{10} , m/s), and instantaneous WRF-BEM air-conditioning power density (AC, W/m^2). AC is a model diagnostic, not metered building electricity consumption.

The analysis used a dedicated HDF5 dataset containing T_2 , Q_2 , AC, scalar W_{10} , geographic coordinates, LCZ/land-cover class, urban category, and distance-to-river context. W_{10} is the 10 m horizontal wind-speed magnitude, defined as $W_{10} = \sqrt{U_{10}^2 + V_{10}^2}$ and expressed in m/s. It was derived from the same August 2022 WRF simulation and aligned with the same 120 hourly time steps as the other variables. Automated checks verified the $120 \times 360 \times 360$ array dimensions, finite non-negative values, time alignment, spatial metadata, and consistency of the $T_2/Q_2/AC$ fields throughout the analysis workflow. No final-test target was used for scaling or model selection.

The urban domain lies within Chongqing Municipality in southwestern China (Figure 2). Banan, Beibei, Bishan, Jiangjin, Shapingba, and Yubei are the districts containing the six reference stations; the simulation is evaluated across the full grid, not only within these six administrative units. Rural, suburban, and urban labels in Table 1 describe station environments, not uniform land-cover classes for entire districts. Built surfaces influence heat storage, evapotranspiration, aerodynamic drag, and building-energy exchange,

whereas vegetated and water surfaces have different energy and moisture budgets [1,5–9]. LCZ/land-cover classes represent these surface differences in the spatial diagnostics. Administrative outlines provide geographic orientation only; they are neither predictors nor units for aggregating metrics.

Table 1. Six-station WRF RMSE for 22–26 August 2022. Each value is calculated from the 120 paired hourly observations and WRF outputs at the corresponding station.

Station	Station Environment	T ₂ (°C)	RH ₂ (%)	W ₁₀ (m/s)	Q ₂ (g/kg)
Banan	Rural	0.984	4.108	1.323	1.628
Beibei	Suburban	1.459	4.320	1.307	1.115
Bishan	Urban	1.193	3.377	1.052	1.159
Jiangjin	Suburban	0.935	4.478	0.806	1.561
Shapingba	Urban	1.148	3.771	0.987	1.068
Yubei	Suburban	0.929	3.229	1.059	1.302

The selected period contains five complete diurnal cycles within the prolonged 2022 summer heatwave; it does not represent the entire event. The official Chongqing climate assessment reports an August mean temperature of 32.5 °C, 4.9 °C above normal and the highest since 1961, with the regional heatwave continuing from 24 July to 29 August [46]. The continuous 120 h record provides the sequence required for model development and a held-out full-day test under sustained extreme heat. The selected dates therefore place the study within a documented heatwave, but they do not by themselves establish seasonal representativeness or performance under other weather conditions.

The study-area locator map uses the Ministry of Natural Resources standard map of China, GS(2019)1652 [47], and DataV district outlines [48]. DataV coordinates were converted from GCJ-02 to approximate WGS84 to align the outlines with the WRF coordinates. These outlines provide geographic orientation; the WRF grid and LCZ fields define the analysis domain. Numbered symbols indicate nominal WRF grid references associated with the station records, not surveyed station coordinates.

3.2. Station-Based Evaluation of the WRF Simulation

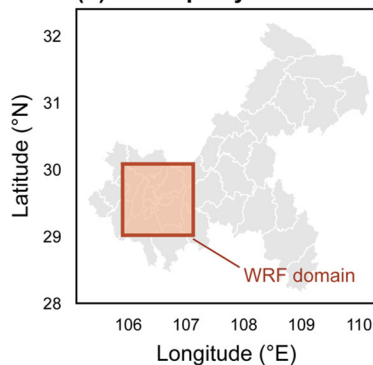
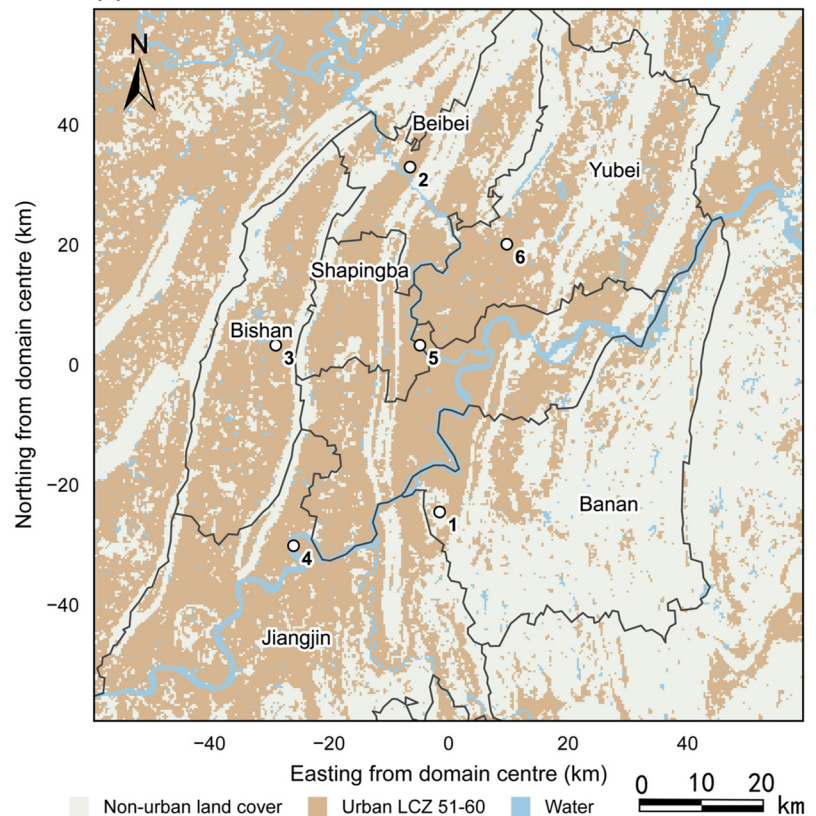
We checked the physical plausibility of the WRF reference configuration by comparing its outputs with observations from Banan, Beibei, Bishan, Jiangjin, Shapingba, and Yubei. The comparison covers the same 120 h period used for the surrogate benchmark, from 00:00 local time on 22 August to 23:00 on 26 August 2022, and includes 2 m air temperature (T₂), 2 m relative humidity (RH₂), 10 m wind-speed magnitude (W₁₀), and 2 m water-vapor mixing ratio (Q₂). Table 1 reports the station-specific RMSE over this period, and Figure 3 shows the corresponding hourly observations and WRF series.

These station comparisons provide a limited check on the physical plausibility of the WRF simulation, not a formal validation of the full simulated state or of the surrogate forecasts. The surrogate benchmark measures fidelity to held-out WRF outputs and should not be interpreted as direct observation-based forecast validation. The WRF-BEM AC diagnostic is not evaluated here because no corresponding metered building-energy observations were used.

(a) Chongqing in China

Approval No. GS(2019)1652

Source: Ministry of Natural Resources of China

(b) Municipality and WRF domain**(c) WRF domain and station districts**

Nominal grid references: 1 Banan 2 Beibei 3 Bishan 4 Jiangjin 5 Shapingba 6 Yubei

Figure 2. Study area and geographic context: (a) Chongqing in China; (b) Chongqing Municipality and the WRF domain; (c) WRF land cover and the six station districts. Station-environment categories in Table 1 do not describe entire districts. Numbered symbols indicate nominal station-associated WRF grid references. The map in panel (a) is based on the standard map from the Standard Map Service System of the Ministry of Natural Resources of China (approval No. GS(2019)1652) [47]. District outlines are from DataV [48]; LCZ and water classes are from WRF.

Station RMSE ranged from 0.929 to 1.459 °C for T_2 , 3.229 to 4.478% for RH_2 , 0.806 to 1.323 m/s for W_{10} , and 1.068 to 1.628 g/kg for Q_2 (Table 1). In the 22–26 August 2022 time series, WRF reproduced the dominant multiday temperature and humidity cycles at all six stations, although hourly discrepancies varied more for wind speed and Q_2 (Figure 3). The observed cycles and reported discrepancies provide limited support for treating the WRF meteorological fields as physically plausible reference data for this case. They do not establish accuracy at every grid cell or under other weather regimes, nor do they establish the accuracy of the surrogate forecasts against observations.

3.3. Forecast Task and Expanding-Window Backtest

Each trainable model uses the previous 24-hour values of one variable at one grid cell to predict its value in the next hour. Forecasts beyond one hour are recursive: each prediction is appended to the input history until a 24 h sequence is complete. Model selection uses

three expanding-window folds to reduce dependence on any single development interval. Fold 1 uses h1–h48 for fitting and forecasts h49–h72 from the WRF h25–h48 history; Fold 2 uses h1–h60 for fitting and forecasts h61–h84 from h37–h60; Fold 3 uses h1–h72 for fitting and forecasts h73–h96 from h49–h72. The folds preserve temporal order and sample successive phases of the event. This design follows recommendations to avoid random resampling for non-stationary time series and to use temporally ordered out-of-sample evaluation [49,50]. Table 2 summarizes the development folds and final test.

Table 2. Temporal development and final-test protocol.

Stage	Fit Hours	Initialization History	Forecast Target	Horizon
Fold 1	h1–48	h25–48	h49–72	24 h
Fold 2	h1–60	h37–60	h61–84	24 h
Fold 3	h1–72	h49–72	h73–96	24 h
Final	h1–96	h73–96	h97–120	24 h

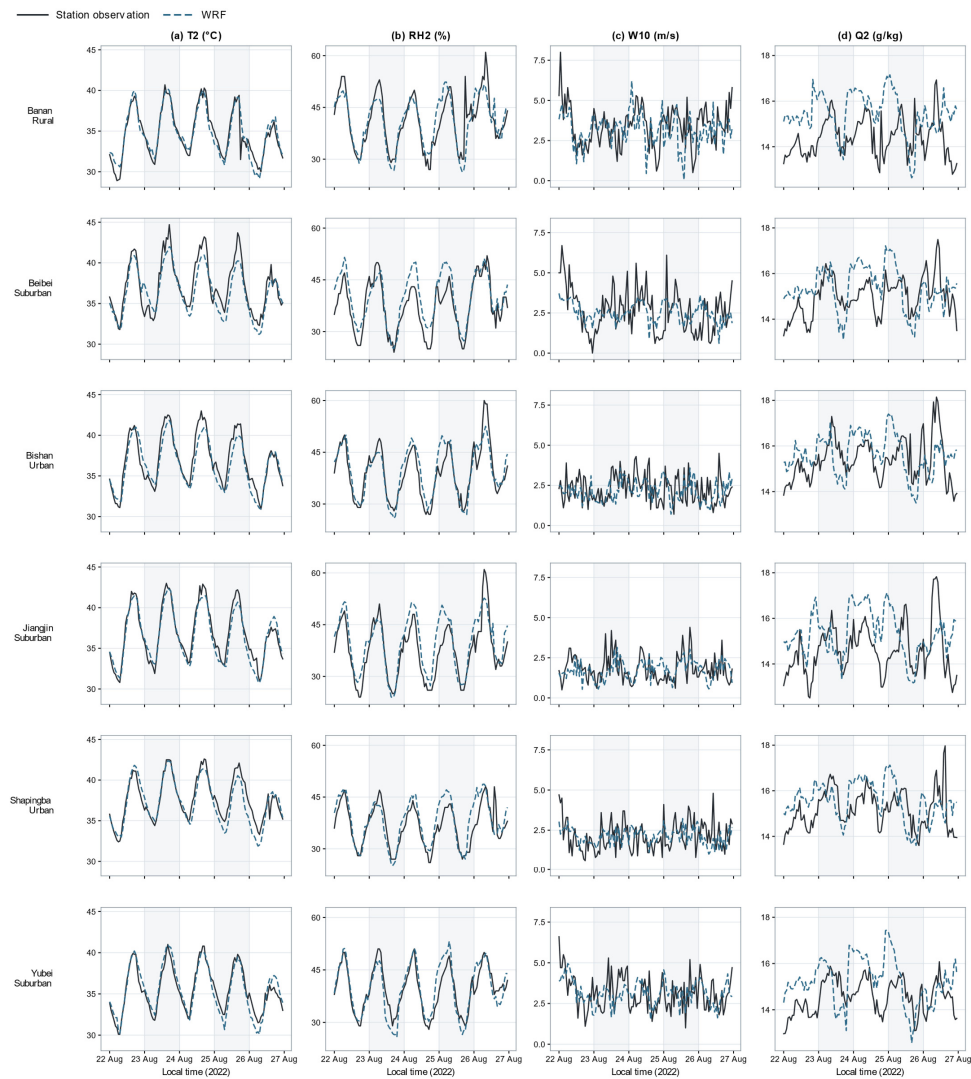


Figure 3. Station-based evaluation of the WRF simulation from 22 to 26 August 2022. Black solid lines show hourly station observations, and blue dashed lines show the corresponding WRF series for T₂, RH₂, W₁₀, and Q₂. Table 1 summarizes RMSE from the same 120 h series. It evaluates WRF credibility rather than the surrogate forecasts.

The 24 h history was chosen to include one complete diurnal cycle and to keep the available temporal input constant across model families and targets. This choice controls the input available in the benchmark; it does not imply that 24 h is optimal for every variable. In particular, W10 has weaker temporal persistence than T2 and AC, so a shorter input history for wind could be useful. Comparing history lengths would introduce an additional experimental factor and was not considered as part of this study.

All trainable families use the same regular 18×18 array of 324 monitoring cells during development. A separate scaler for each variable is fitted to the training segment of each fold. Hyperparameters are selected to minimize mean 24 h RMSE across the three folds, with ties among candidates within 0.5% of the best score resolved in favor of the simpler specification. Neural-network-stopping epochs are determined only during development and are then fixed for final refitting. Once all choices are finalized, the selected specification is refitted on h1–h96 and initialized with the WRF h73–h96 history to forecast h97–h120 over all 360×360 grid cells. The final test is excluded from hyperparameter selection, epoch selection, and model-family choice.

3.4. Model Families and Reproducibility

The benchmark includes naive, periodic, regularized-linear, bagged-tree, boosted-tree, feed-forward neural, and recurrent-neural families. This family-level comparison follows emulator studies in which performance depends on the target process and evaluation objective rather than improving monotonically with nominal complexity [30,31]. Table 3 lists the fixed structure and selection procedure for each family.

Table 3. Benchmark model families and selection procedures.

Family	Fixed Structure or Definition	Selection and Repetition
Persistence	Last available value	No fitting
Previous-day	Same hour from the previous 24 h	No fitting
Ridge	24-lag regularized autoregression	α selected from 0.01 to 100
Random Forest	Bagged regression trees	Two fixed candidate sets; five seeds
XGBoost	Gradient-boosted trees	Two fixed candidate sets; five seeds
MLP	24–64–32–1; ReLU; dropout 0.2	Five seeds; fixed final epochs
GRU	One recurrent layer; 64 hidden units	Five seeds; fixed final epochs
LSTM	Stacked 128/64 hidden units; dropout 0.2	Five seeds; fixed final epochs

Neural networks are trained with mean-squared-error loss, Adam optimization, and stopping criteria determined only during development. The fixed final epoch counts for T₂/Q₂/AC are 13/4/16 for MLP, 8/3/23 for GRU, and 9/2/12 for LSTM; the corresponding W₁₀ epoch counts are 8, 8, and 9. Random Forest, XGBoost, MLP, GRU, and LSTM use the same five random seeds: 42, 31,415, 73,017, 82,216, and 20,260,811. W₁₀ hyperparameters were selected independently across the same three development folds. The h97–h120 test was not used for tuning.

3.5. Metrics and Diagnostic Analyses

The primary error metrics are root mean square error (RMSE), mean absolute error (MAE), and mean error (ME). For N pairs of surrogate forecasts and WRF reference values, the metrics are calculated as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2} \quad (1)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i| \quad (2)$$

$$ME = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i) \quad (3)$$

Here, $\hat{y}_i$ is the surrogate prediction and y_i is the WRF reference value for sample i . Scores reported for 6, 12, and 24 h summarize cumulative prefixes from the same forecast origin, without selecting a favorable origin. Gridwise RMSE is calculated over the 24 final-test hours. Spatial fidelity is evaluated using hourly spatial Pearson correlation, centered RMSE, and the ratio of predicted to reference spatial standard deviation. These measures complement aggregate scores because weather and climate benchmarks call for common reference data, strong baselines, multiple metrics, and spatially explicit evaluation [29,51].

Spatial added value is measured against the relevant simple baseline using the following RMSE difference. Negative values denote an improvement.

$$\Delta RMSE = RMSE(trainable) - RMSE(baseline) \quad (4)$$

The remaining diagnostics include gridwise autocorrelation from lag 1 to lag 48, LCZ/land-cover stratification, distance-to-river classes, maps of the gridwise best-performing family and its winning margin, and $\Delta RMSE$ maps relative to a simple baseline. The W_{10} snapshot is the final-test hour with the largest spatial standard deviation in WRF W_{10} . W_{10} regimes are defined by quantiles of the final-test targets ($\leq P_{25}$, $P_{25}-P_{75}$, $P_{75}-P_{90}$, and $>P_{90}$) and absolute hourly tendencies ($\leq P_{50}$, $P_{50}-P_{90}$, and $>P_{90}$). These post hoc thresholds are used only for diagnostic analysis, not for tuning.

To identify where Q2 spatial amplitude was reduced, the spatial-fidelity calculation was repeated within each LCZ/land-cover class and river-distance bin. Predictions were first averaged over the five seeds. For each test hour, spatial standard deviation and Pearson correlation were then calculated across cells within each group. The resulting metrics were averaged over the 24 test hours. The predicted-to-WRF standard-deviation ratio and correlation were calculated for both LSTM and Random Forest, with cell counts recorded for each group. These within-group diagnostics differ from pooled RMSE and do not alter model selection. Very small groups are retained for completeness but provide an unstable basis for geographic comparison.

4. Results

4.1. Final Independent-Test Benchmark

No family performed best for all four variables. Random Forest achieved the lowest T_2 RMSE (1.081 ± 0.002 °C), LSTM the lowest Q_2 RMSE (0.969 ± 0.063 g/kg), and MLP the lowest AC RMSE (0.505 ± 0.187 W/m²). Deterministic Ridge performed best for W_{10} (1.095 m/s), narrowly ahead of Random Forest (1.098 ± 0.001 m/s) and XGBoost (1.103 ± 0.001 m/s). Mean RMSE was 1.105 ± 0.027 m/s for GRU and 1.147 ± 0.032 m/s for LSTM. Thus, the recurrent models did not improve W_{10} accuracy over Ridge in this event (Table 4; Figure 4).

Table 4. Full-domain RMSE for the independent h97–h120 test for T_2 , Q_2 , W_{10} , and AC. Values for stochastic families are reported as the mean $\pm$ sample standard deviation across five seeds.

Model	T_2 ($^{\circ}\text{C}$)	Q_2 (g/kg)	W_{10} (m/s)	AC (W/m^2)
Persistence	2.981	1.054	1.745	1.605
Previous-day	1.689	1.732	1.300	0.529
Ridge	1.497	1.125	1.095	0.526
Random Forest	1.081 ± 0.002	0.980 ± 0.014	1.098 ± 0.001	0.509 ± 0.001
XGBoost	1.109 ± 0.009	1.173 ± 0.017	1.103 ± 0.001	0.607 ± 0.007
MLP	1.385 ± 0.305	1.091 ± 0.115	1.119 ± 0.035	0.505 ± 0.187
GRU	1.339 ± 0.252	1.021 ± 0.136	1.105 ± 0.027	0.679 ± 0.194
LSTM	1.547 ± 0.350	0.969 ± 0.063	1.147 ± 0.032	1.177 ± 0.645

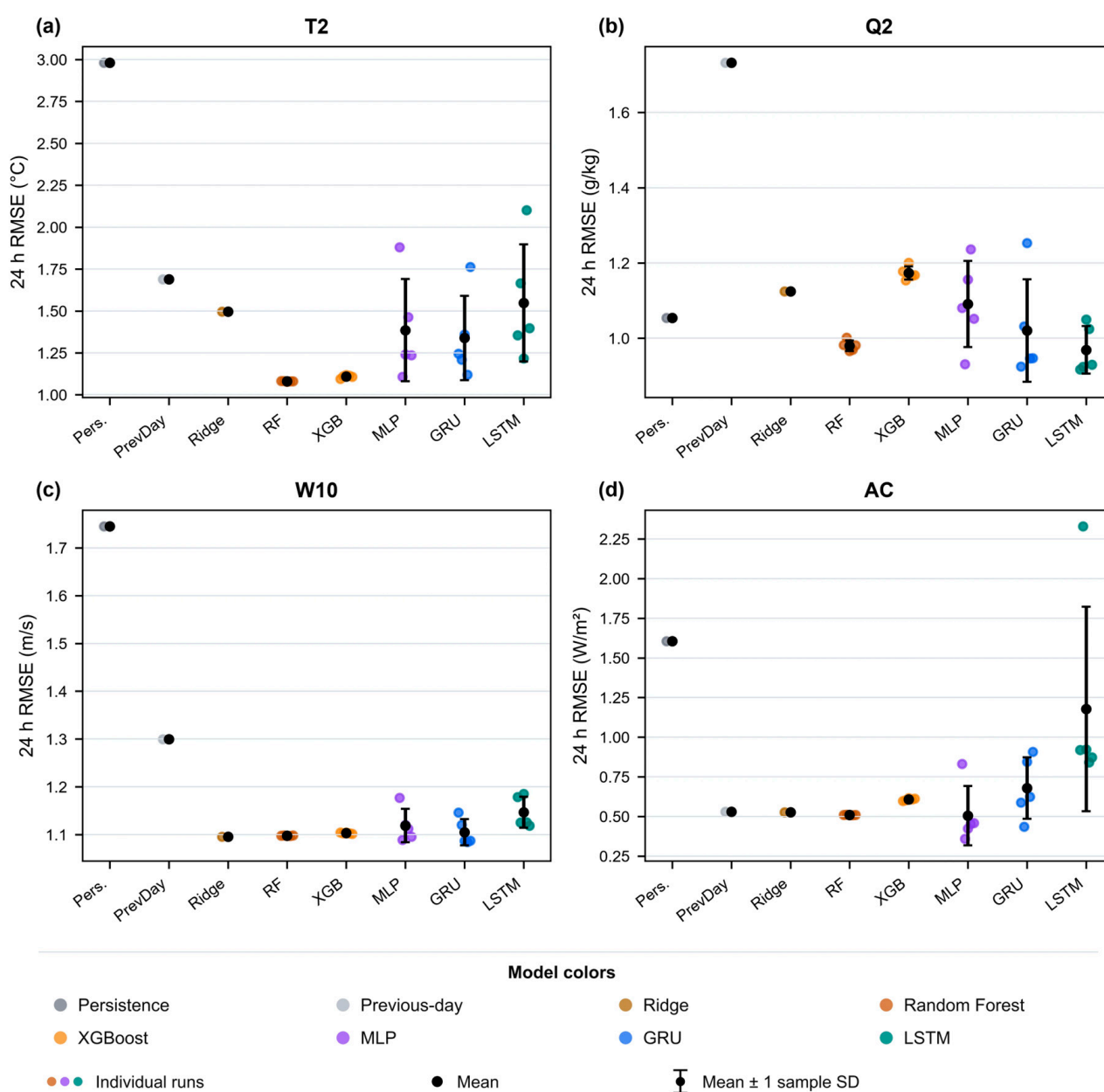


Figure 4. Independent-test 24 h RMSE for all eight model families and four WRF targets, including seed-level variability for stochastic models. (a) 2 m air temperature (T_2); (b) 2 m water-vapor mixing ratio (Q_2); (c) 10 m wind-speed magnitude (W_{10}); (d) WRF-BEM air-conditioning power density (AC).

4.2. Temporal Memory, Development Stability, and Lead Dependence

The variables had markedly different temporal structures. Median lag-1/lag-24 autocorrelation was 0.949/0.925 for T_2 , 0.857/0.093 for Q_2 , 0.581/0.211 for W_{10} , and 0.948/0.986 for AC. W_{10} retained moderate hour-to-hour memory but weak daily recurrence, unlike T_2 and AC. These autocorrelations indicate weaker temporal persistence for W_{10} ; they do not establish which alternative input window or predictors would improve its forecast (Figure 5).

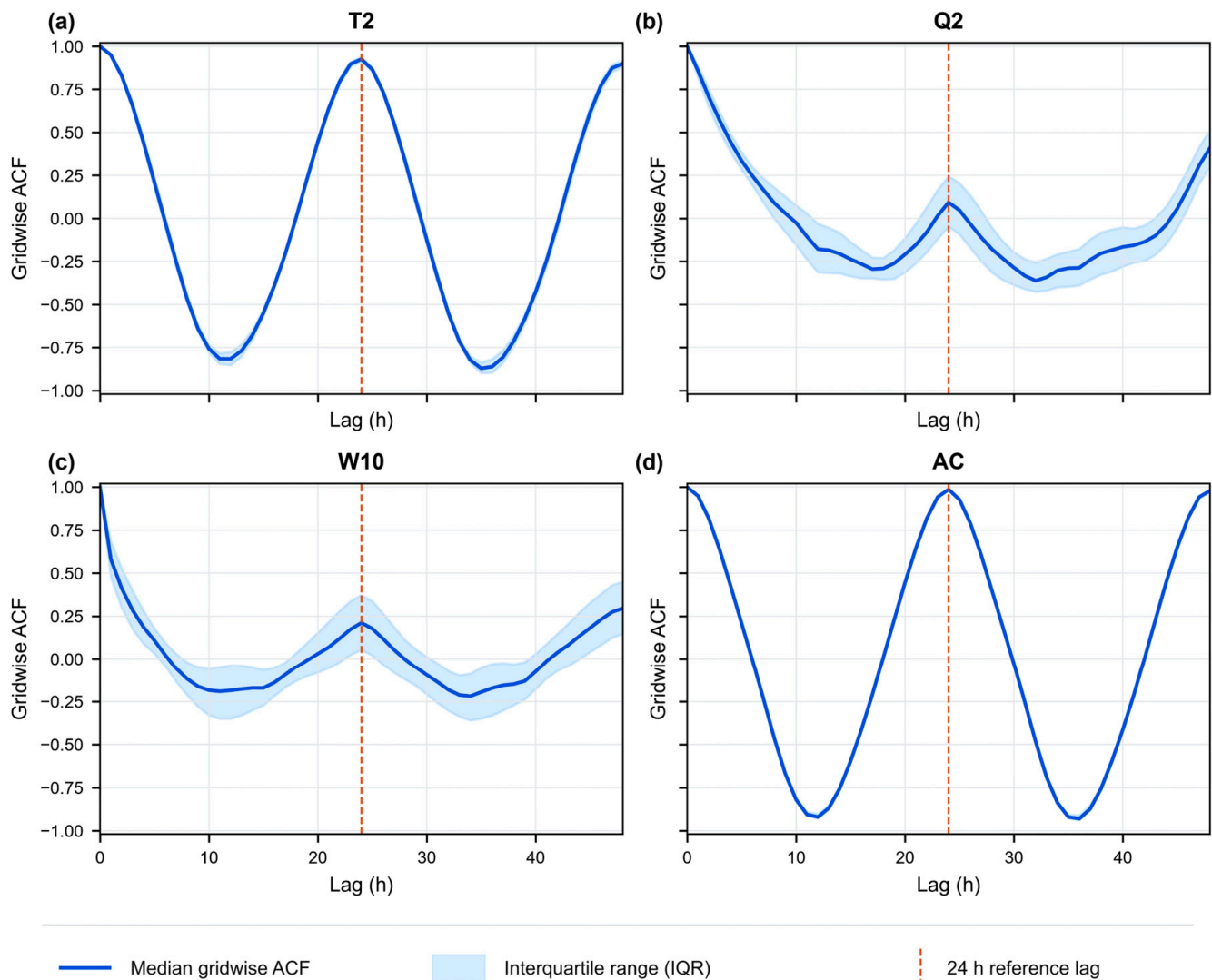


Figure 5. Gridwise autocorrelation functions (ACFs) calculated over development hours h1–h96, (a) 2 m air temperature (T_2); (b) 2 m water-vapor mixing ratio (Q_2); (c) 10 m wind-speed magnitude (W_{10}); (d) WRF-BEM air-conditioning power density (AC).

Development rankings did not necessarily carry over to the independent test. Across Folds 1–3, LSTM ranked first for T_2 , LSTM and GRU were effectively tied for Q_2 , and previous-day same-hour ranked first for AC. The untouched final period instead favored Random Forest, LSTM, and MLP for those variables. Q_2 also showed substantial fold-to-fold variability, so a single validation block would have produced a less reliable model choice. Figure 6 compares these development rankings with the final-test results.

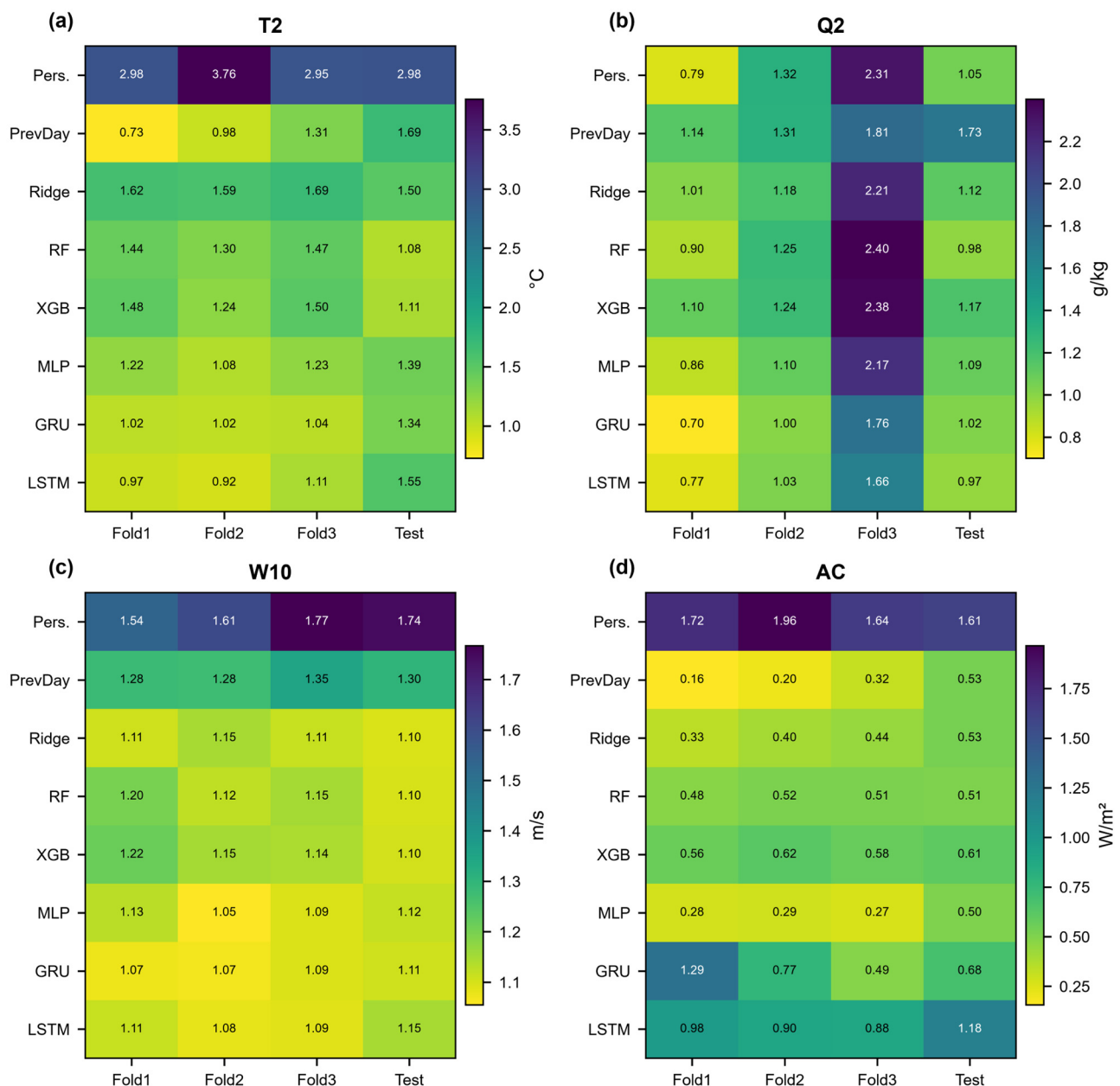


Figure 6. 24 h RMSE for the three development folds and the independent final test: (a) 2 m air temperature (T2); (b) 2 m water-vapor mixing ratio (Q2); (c) 10 m wind-speed magnitude (W10); (d) WRF-BEM air-conditioning power density (AC).

Lead-hour RMSE patterns differed across the four target variables (Figure 7). For T2, Random Forest RMSE increased from 0.442 °C at lead hour 1 to a maximum of 1.760 °C at lead hour 15, then declined to 0.721 °C at lead hour 24. Among the Q2 models shown, the differences in RMSE were greatest during the early forecast hours. Previous-day RMSE exceeded 2.0 g/kg from lead hours 1 to 4, while LSTM and Persistence produced substantially lower errors. At lead hour 24, LSTM had the lowest RMSE among the plotted Q2 models at 0.645 g/kg. AC followed a different pattern. Persistence RMSE increased at the middle and later forecast leads and reached 4.052 W/m² at lead hour 17, whereas MLP reached a much lower maximum of 1.102 W/m² at lead hour 9. For W10, Ridge and Random Forest produced nearly overlapping error profiles. Their RMSE values were close to 1.3 m/s around lead hours 4 and 5, dropped below 0.9 m/s at lead hour 20, and increased modestly thereafter.

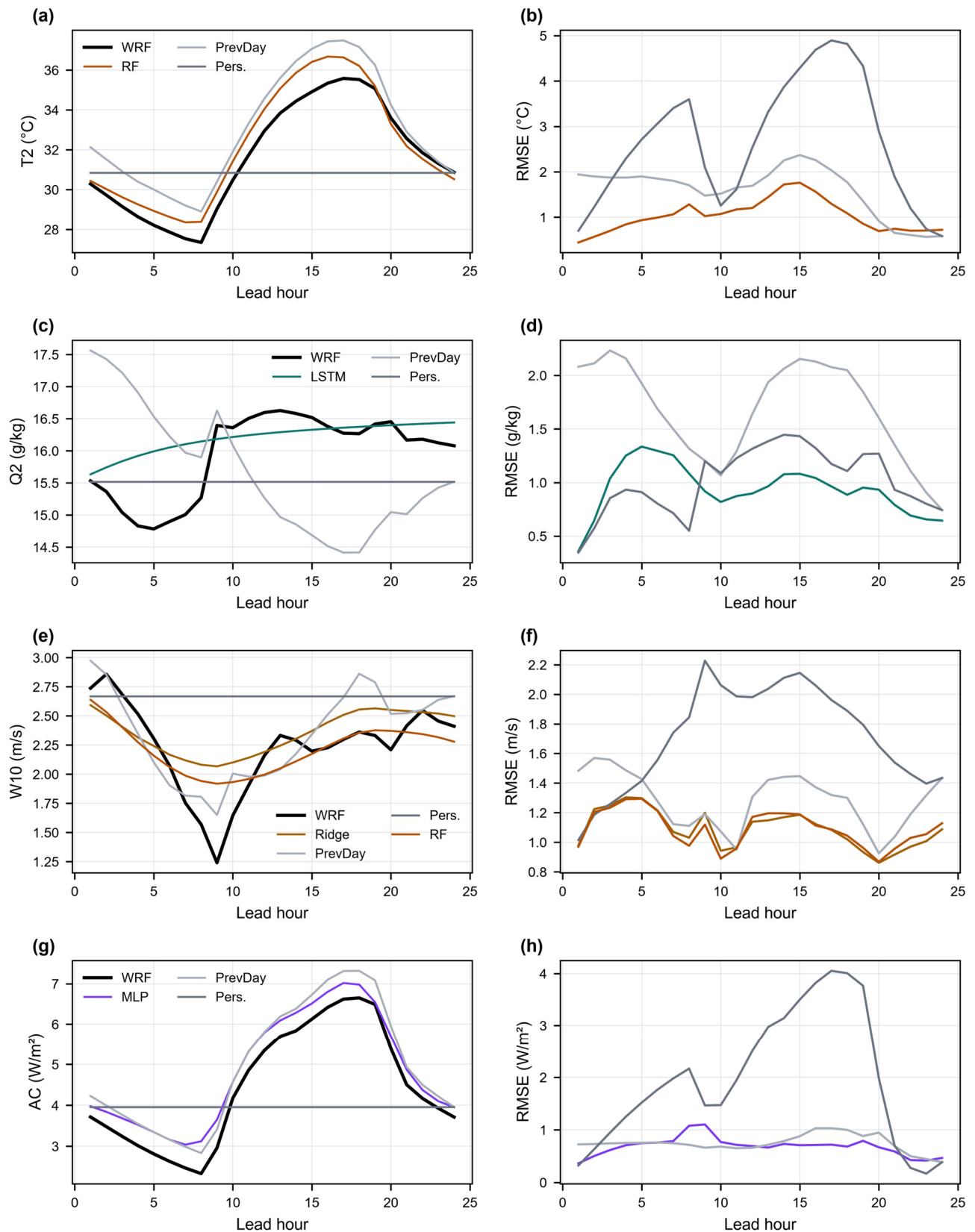


Figure 7. Domain-mean final-test trajectories and lead-hour RMSE. The W_{10} panels show WRF, Ridge, Previous-day, Persistence, and Random Forest; representative families are shown for the other variables. (a) mean 2 m air temperature (T2); (b) T2 RMSE; (c) mean 2 m water-vapor mixing ratio (Q2); (d) Q2 RMSE; (e) mean 10 m wind-speed magnitude (W10); (f) W10 RMSE; (g) mean WRF-BEM air-conditioning power density (AC); (h) AC RMSE.

Figure 7 reports RMSE separately at each lead hour, whereas Figure 8 reports cumulative RMSE over the stated 6, 12, or 24 h prefix, starting from lead hour 1. In this same-origin cumulative-prefix analysis, the preferred model changed as additional forecast hours were included (Figure 8). For T2, LSTM had the lowest cumulative RMSE over the 6 h prefix at 0.716 °C, XGBoost ranked first over 12 h at 0.937 °C, and Random Forest ranked first over the full 24 h prefix at 1.081 °C. Persistence achieved the lowest Q2 RMSE over both the 6 h and 12 h prefixes, with values of 0.767 and 0.922 g/kg, respectively, while LSTM had the lowest RMSE over the full 24 h prefix at 0.969 g/kg. For AC, Random Forest ranked first over the 6 h and 12 h prefixes, with RMSE values of 0.346 and 0.446 W/m², while MLP achieved the lowest 24 h cumulative RMSE at 0.505 W/m². For W10, XGBoost achieved marginally lower RMSE over the 6 h and 12 h prefixes, with RMSE values of 1.200 and 1.121 m/s, whereas Ridge achieved the lowest RMSE over the full 24 h prefix at 1.095 m/s. Because these scores are nested prefixes from a single forecast origin, they describe error accumulation within one recursive forecast rather than performance across independent forecast origins. The model with the lowest RMSE over a shorter prefix was therefore not always the best-performing model over the full 24 h period.

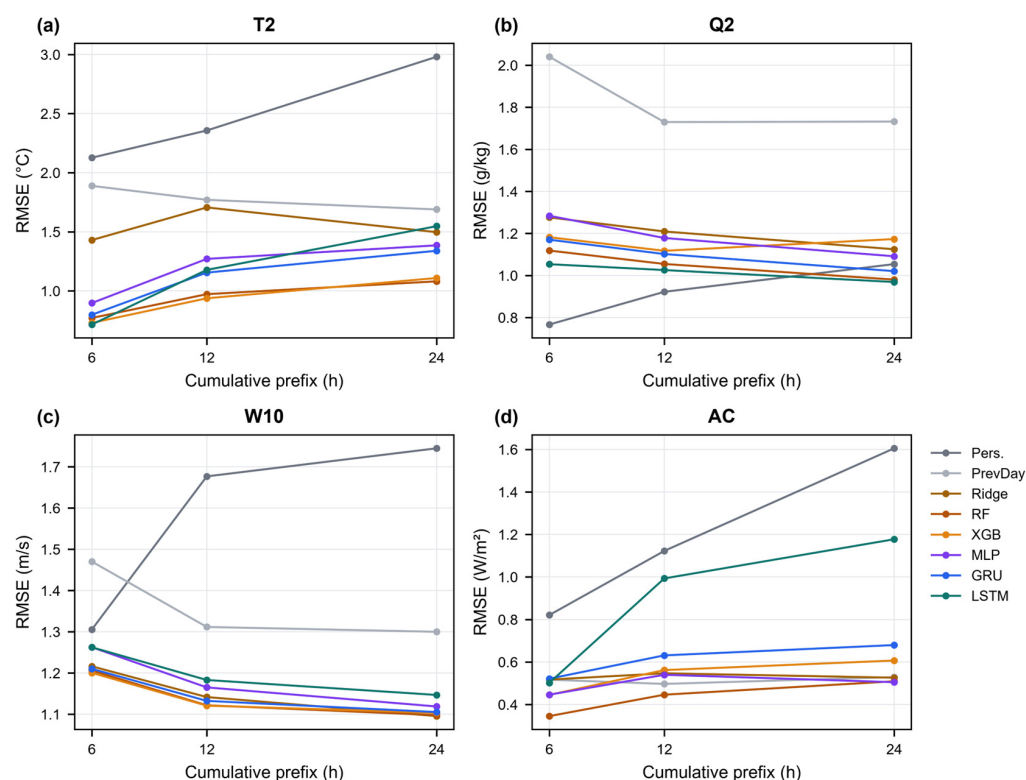
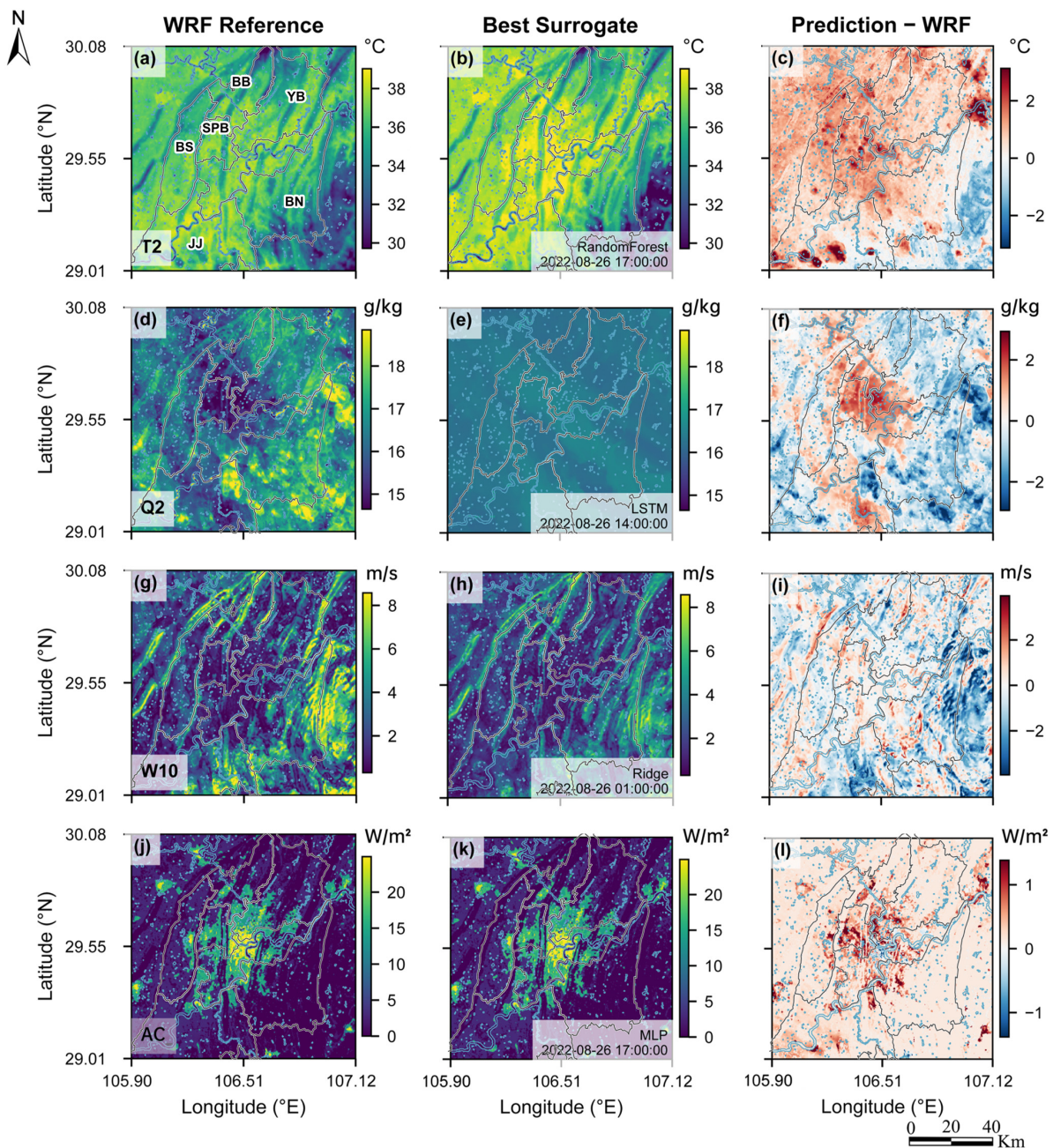


Figure 8. Same-origin cumulative-prefix RMSE at 6, 12, and 24 h for the four targets; these are prefixes from one forecast origin, not independent origins. (a) 2 m air temperature (T2); (b) 2 m water-vapor mixing ratio (Q2); (c) 10 m wind-speed magnitude (W10); and (d) WRF-BEM air-conditioning power density (AC).

4.3. Spatial Fidelity and Local Model Preference

Spatial-field fidelity varied substantially among the four targets (Figure 9). Across the final 24 h, Random Forest retained the T2 spatial pattern with a mean correlation of $r = 0.946$ and a mean spatial standard-deviation ratio of 1.001. MLP reproduced the AC field with similarly high fidelity: $r = 0.998$ and a standard-deviation ratio of 1.060. Q2 was less accurately resolved spatially. Although LSTM had the lowest domain-average RMSE, its mean spatial correlation was only $r = 0.315$ and its standard-deviation ratio was 0.298, consistent with the pronounced smoothing visible in the surrogate field. Ridge provided intermediate spatial

fidelity for W10, with a mean correlation of $r = 0.710$ and a standard deviation ratio of 0.687. The selected W10 map corresponds to event hour h98 (01:00 on 26 August), when the WRF spatial standard deviation reached 2.230 m/s. At this time, Ridge retained the broad structure of the wind field but damped some of the local spatial contrasts.



Districts: BN Banan BB Beibei BS Bishan JJ Jiangjin SPB Shapingba YB Yubei

Figure 9. Spatial-field fidelity at objectively selected final-test hours. Columns show WRF reference fields, the selected surrogate, and prediction minus WRF; rows show T2, Q2, W10, and AC. Predictions are averaged over five seeds for stochastic models. Thin outlines delineate the six districts containing the stations; abbreviations are defined beneath the panels. The outlines provide orientation only and do not define the spatial-fidelity metrics. (a) WRF 2 m air temperature (T2); (b) Random Forest T2 prediction; (c) T2 prediction minus WRF; (d) WRF 2 m water-vapor mixing ratio (Q2); (e) LSTM Q2 prediction; (f) Q2 prediction minus WRF; (g) WRF 10 m wind-speed magnitude (W10); (h) Ridge W10 prediction; (i) W10 prediction minus WRF; (j) WRF-BEM air-conditioning power density (AC); (k) MLP AC prediction; (l) AC prediction minus WRF.

Q2 spatial amplitude was reduced in both urban and non-urban cells: LSTM had mean spatial standard deviation ratios of 0.368 and 0.293, respectively (Table 5). Among well-represented classes, the ratio was 0.348 in urban class 59 (43,672 cells), 0.416 in urban class 56 (11,166 cells), and 0.241 over water (class 17; 5770 cells). The corresponding Random Forest ratios were 0.807, 0.838, and 0.800. Thus, the damping was not confined to a single surface class. Correlation provides a separate measure of spatial alignment: LSTM retained a high correlation across water cells (0.947) despite its low amplitude ratio, but its correlation within class 59 was only 0.073. Near the rivers, its ratio was 0.267 within 1 km and 0.352 at 1–3 km (Table 6). These descriptive contrasts do not establish a causal river effect, and the bin beyond 10 km contains only 88 cells.

Table 5. Q2 spatial fidelity by urban status and LCZ/land-cover class.

Group	Cells	LSTM SD Ratio	RF SD Ratio	LSTM r	RF r
Full domain	129,600	0.298	0.789	0.315	0.439
Urban LCZ 51–60	62,501	0.368	0.826	−0.019	0.088
Non-urban	67,099	0.293	0.823	0.478	0.566
LCZ 1	1633	0.389	0.990	0.176	0.124
LCZ 2	280	0.418	0.982	0.107	0.107
LCZ 4	364	0.442	1.002	0.146	0.156
LCZ 5	29,813	0.419	1.027	0.201	0.119
LCZ 6	324	0.372	0.785	0.032	0.051
LCZ 7	226	0.419	0.870	0.206	0.205
LCZ 8	234	0.436	0.944	0.168	0.154
LCZ 9	167	0.318	0.819	0.029	0.122
LCZ 10	2148	0.389	0.959	0.165	0.159
LCZ 11	25	0.788	1.756	0.346	0.347
LCZ 12	23,249	0.352	0.876	0.151	0.115
LCZ 14	2856	0.349	0.874	0.139	0.063
LCZ 16	10	0.732	1.190	0.140	0.151
LCZ 17	5770	0.241	0.800	0.947	0.952
LCZ 51	91	0.274	0.539	−0.235	0.051
LCZ 52	38	0.523	0.885	−0.320	−0.306
LCZ 53	14	0.613	0.962	−0.196	−0.131
LCZ 54	1083	0.546	0.980	−0.095	−0.164
LCZ 55	1672	0.496	0.915	−0.081	−0.114
LCZ 56	11,166	0.416	0.838	−0.009	−0.034
LCZ 58	4746	0.462	0.815	0.042	−0.043
LCZ 59	43,672	0.348	0.807	0.073	0.084
LCZ 60	19	0.428	0.939	0.291	0.085

RF denotes Random Forest. Ratios are predicted spatial standard deviation divided by WRF spatial standard deviation; r is Pearson correlation. Values are means over the 24 test hours after averaging predictions across five seeds. Urban cells are LCZ classes 51–60; all other valid cells form the non-urban group. Cell counts describe spatial support, not independent temporal samples. Sparse classes are retained for completeness and should be interpreted cautiously.

Gridwise model rankings varied among variables and across locations (Figure 10). The median relative winning margin was 5.0% for T2, 11.2% for Q2, 2.33% for W10, and 0.0% for AC. Model separation was weakest for AC, where 53.9% of grid cells had a margin below 1%. The locally best and second-best families therefore often had nearly identical RMSE. W10 also showed substantial local competition: 26.0% of cells had margins below 1%, while only 9.9% exceeded 10%. Random Forest, XGBoost, and Ridge were the gridwise winners over 21.1%, 20.5%, and 19.8% of the domain, respectively, even though Ridge had the lowest domain-average W10 RMSE. Domain-average rankings therefore did not apply uniformly across the domain. This distinction was particularly clear for AC, where

widespread near-ties make the designation of a single local winner less clear, and for W10, where several models performed similarly across large parts of the domain.

Table 6. Q2 spatial fidelity by distance to river.

Group	Cells	LSTM SD Ratio	RF SD Ratio	LSTM r	RF r
River 0–1 km	34,356	0.267	0.787	0.591	0.714
River 1–3 km	71,081	0.352	0.844	0.103	0.125
River 3–5 km	20,559	0.349	0.881	0.138	0.130
River 5–10 km	3516	0.325	0.898	−0.004	0.241
River > 10 km	88	0.154	1.627	0.162	0.267

RF denotes Random Forest. Metrics use the same calculation as in Table 5. River-distance bins are descriptive geographic groups and do not establish a causal river effect. The bin beyond 10 km has only 88 cells.

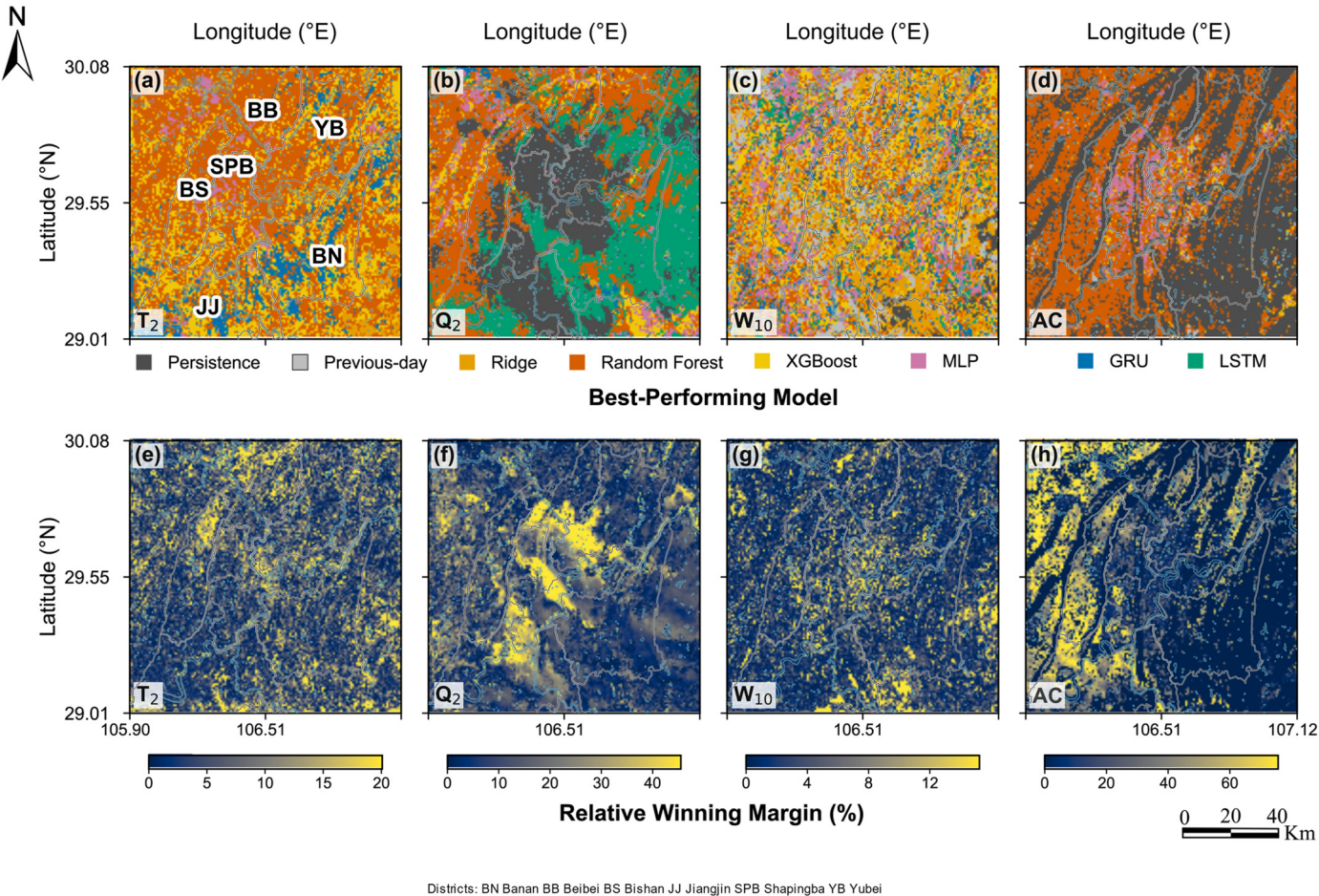


Figure 10. Spatial distribution of the best-performing model and its relative winning margin for T2, Q2, W10, and AC over the final 24 h test period. Upper panels show the model with the lowest gridwise RMSE; lower panels show its percentage advantage over the second-best model relative to the best-model RMSE. Small margins indicate near ties. Thin district outlines and the key to abbreviations provide geographic orientation, not district-level model rankings. (a) best-performing model for 2 m air temperature (T2); (b) best-performing model for 2 m water-vapor mixing ratio (Q2); (c) best-performing model for 10 m wind-speed magnitude (W10); (d) best-performing model for WRF-BEM air-conditioning power density (AC); (e) T2 relative winning margin; (f) Q2 relative winning margin; (g) W10 relative winning margin; (h) AC relative winning margin.

4.4. Spatial Added Value and Critical Operating Regimes

Compared with the Previous-day baseline, Random Forest reduced T_2 gridwise RMSE at 97.9% of cells, with a median ΔRMSE of -0.672 °C. LSTM reduced Q_2 RMSE relative to Persistence at 65.8% of cells, although the improvement was smaller in the static urban subset. MLP reduced AC RMSE relative to Previous-day at only 22.2% of the full domain; its mean ΔRMSE nevertheless became negative for urban/BEM-active cells and cells with high baseline error. Many non-urban or near-zero-load cells are already captured by simple rules, whereas high-load cells contribute disproportionately to squared error. For W_{10} , Ridge reduced gridwise RMSE relative to Previous-day at 75.0% of cells; mean and median ΔRMSE were -0.162 and -0.135 m/s, respectively (Figure 11).

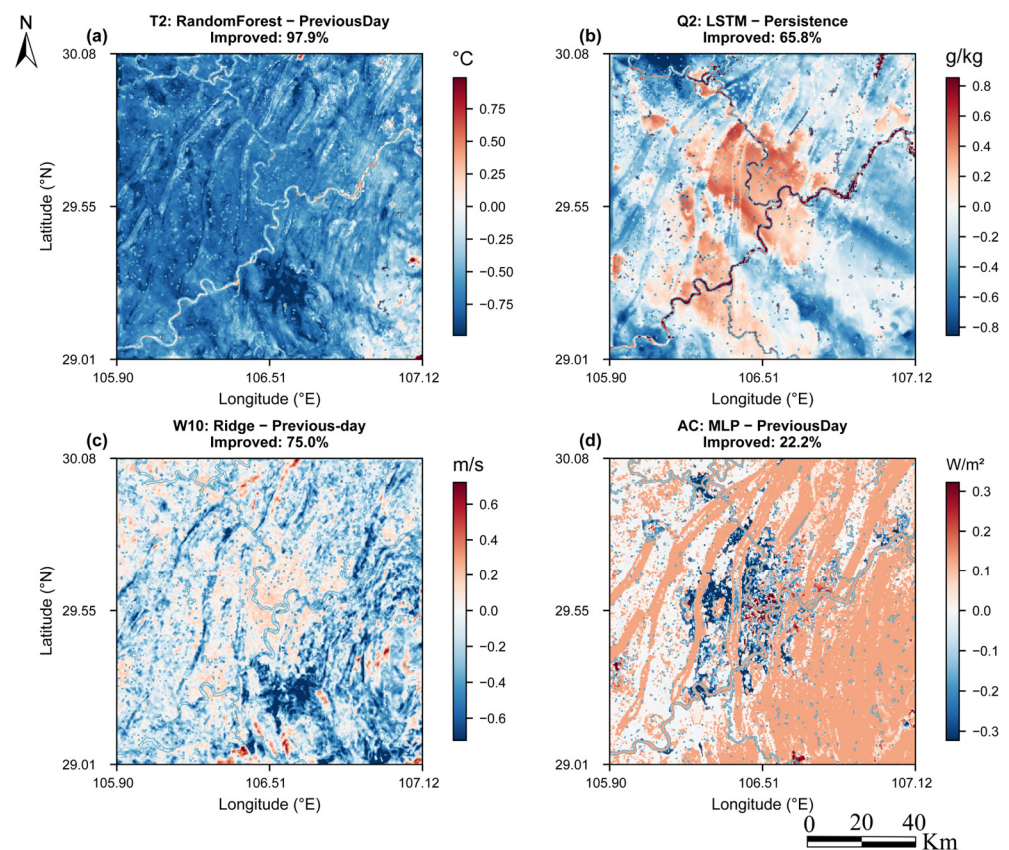


Figure 11. Spatial added value of the selected trainable family relative to the relevant simple baseline. Negative ΔRMSE values indicate improvement. (a) Random Forest minus Previous-day for 2 m air temperature (T_2); (b) LSTM minus Persistence for 2 m water-vapor mixing ratio (Q_2); (c) Ridge minus Previous-day for 10 m wind-speed magnitude (W_{10}); (d) MLP minus Previous-day for WRF-BEM air-conditioning power density (AC).

Regime-specific results help explain the aggregate scores (Figure 12). Random Forest remained best for T_2 in the full urban sample and the top 25%, 10%, and 5% WRF-temperature subsets. Its RMSE declined from 1.070 °C for all urban samples to 0.940 °C in the top 5%, while Persistence substantially underpredicted the hottest states. Random Forest performed best for Q_2 under small tendencies, whereas LSTM ranked first for medium and top-decile tendencies. For AC, Random Forest performed best at low and medium urban loads, but MLP performed best in the top-decile regime (1.432 W/m²), where Persistence reached 6.395 W/m². Figure 13 compares full-domain, urban/BEM-active, and urban-load AC results. W_{10} showed a different limitation. Ridge RMSE was 0.777 m/s for moderate winds, 0.988 for high winds, and 2.027 for the top wind-speed decile. It increased from 0.876 under small absolute tendencies to 1.742 m/s in the largest-tendency decile. Errors therefore

increased under strong winds and rapid changes; this regime comparison does not establish whether different predictors or model architectures would remove the deficit.

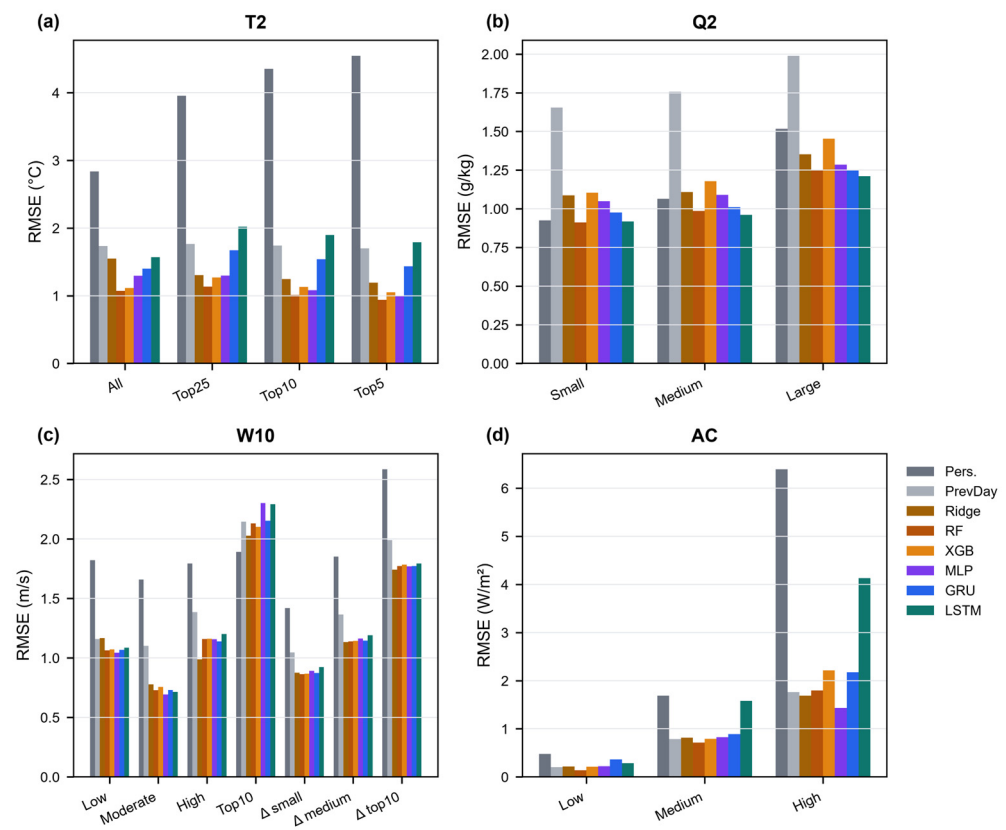


Figure 12. RMSE for objectively defined T₂ high-temperature, Q₂-tendency, W₁₀ wind-speed and absolute-tendency, and AC-load regimes. (a) 2 m air temperature (T₂) for all urban samples and the highest 25%, 10%, and 5% of WRF temperatures; (b) 2 m water-vapor mixing ratio (Q₂) under small, medium, and large absolute hourly tendencies; (c) 10 m wind-speed magnitude (W₁₀) by wind-speed class and absolute hourly tendency, with Δ identifying tendency-based groups; (d) WRF-BEM air-conditioning power density (AC) under low, medium, and high urban loads.

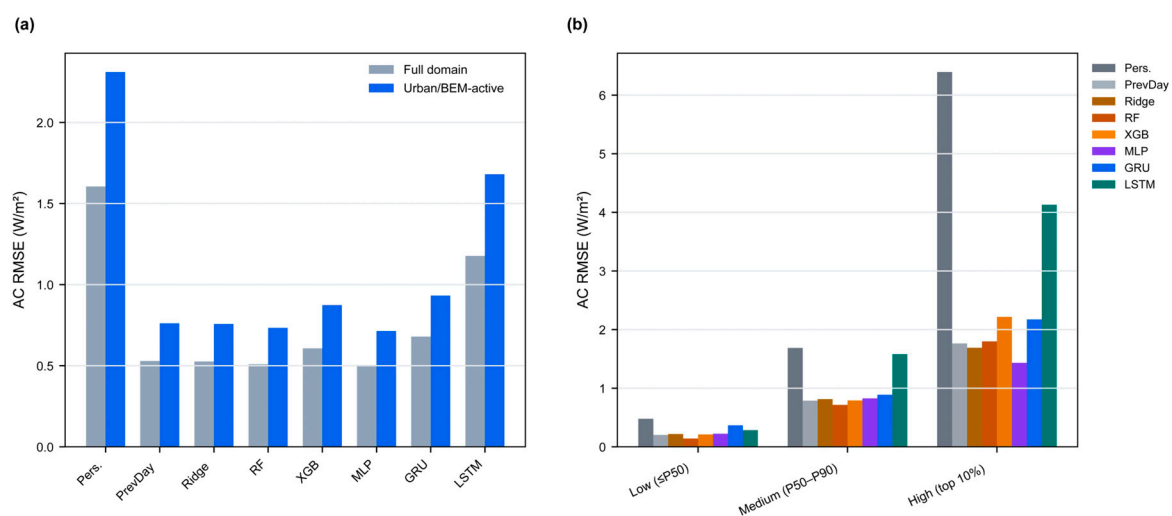


Figure 13. AC performance for the full domain compared with urban/BEM-active cells and across urban load regimes. (a) comparison between the full domain and urban/BEM-active grid cells for each model; (b) comparison among models under low, medium, and high urban loads.

4.5. LCZ, River Context, and Accuracy–Complexity Trade-Offs

Stratification by LCZ and land-cover class showed clear differences in surrogate performance between urban and non-urban areas (Figure 14). For AC, errors were close to zero in classes with little or no BEM activity because the reference load was negligible. Differences among the trainable models were more apparent in urban classes 51 to 60, where BEM simulated active building-energy demand. This pattern suggests that prediction difficulty depends partly on the spatial distribution of the target values, not only on the model family. For AC in particular, large non-urban areas may have low errors while contributing little to the high-load conditions of practical interest. Classwise RMSE values should also be interpreted alongside the class-support information in Figure 14, as estimates from sparsely represented classes are less stable.

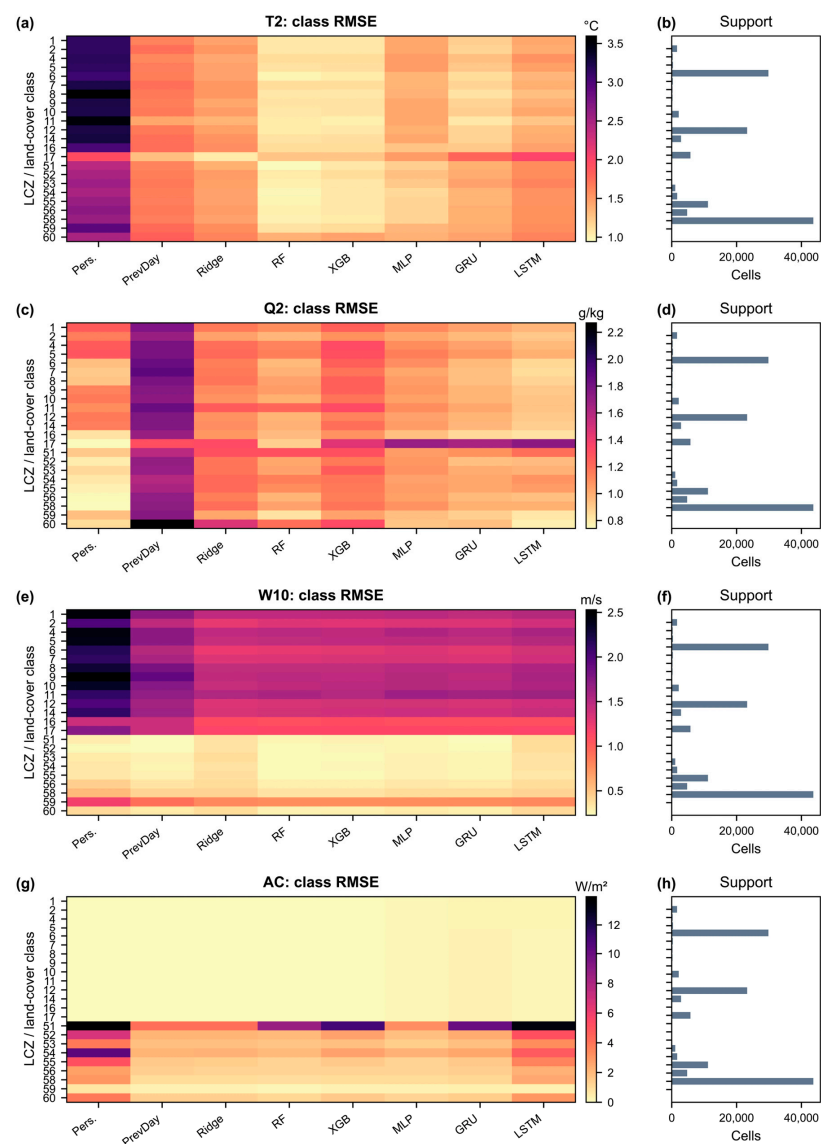


Figure 14. LCZ/land-cover RMSE and class support for T_2 , Q_2 , W_{10} , and AC. Cell counts are provided to help avoid overinterpreting results for sparsely represented classes. (a) classwise RMSE for 2 m air temperature (T_2); (b) grid-cell counts for the classes in panel (a); (c) classwise RMSE for 2 m water-vapor mixing ratio (Q_2); (d) grid-cell counts for the classes in panel (c); (e) classwise RMSE for 10 m wind-speed magnitude (W_{10}); (f) grid-cell counts for the classes in panel (e); (g) classwise RMSE for WRF-BEM air-conditioning power density (AC); (h) grid-cell counts for the classes in panel (g).

Stratification by distance to the river revealed a different form of spatial variation (Figure 15). These distance bins provide descriptive geographic context only and do not imply a causal river effect. For T2, Random Forest outperformed Persistence and Previous-day in every distance class, so its advantage was not limited to a specific river-proximity zone. Q2 performance varied more strongly among distance classes, particularly for LSTM. For AC, MLP and Previous-day produced similar errors in several bins. This result is consistent with the broader finding that simple temporal baselines remain competitive where building energy demand is low or close to zero. Figures 14 and 15 therefore indicate that surrogate performance varies with spatial context, although the magnitude and interpretation of this variation differ among target variables. No single geographic factor accounts for all of the observed differences.

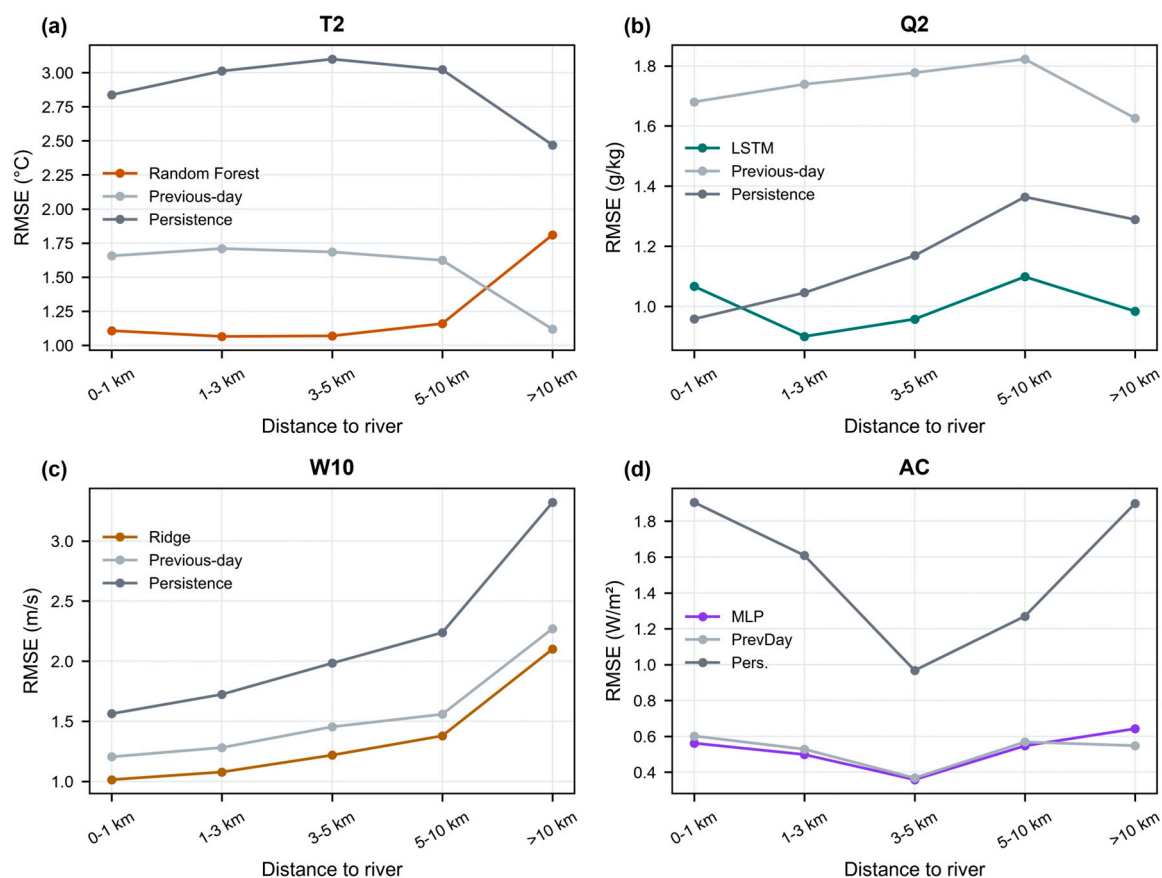


Figure 15. RMSE for the best-performing model and simple baselines across distance-to-river classes for four targets. The analysis provides descriptive geographic context and does not imply a causal river effect. (a) 2 m air temperature (T2), comparing Random Forest with Previous-day and Persistence; (b) 2 m water-vapor mixing ratio (Q2), comparing LSTM with the two baselines; (c) 10 m wind-speed magnitude (W10), comparing Ridge with the two baselines; (d) WRF-BEM air-conditioning power density (AC), comparing MLP with the two baselines.

The surrogate timings were measured on a CPU-based Windows workstation with two AMD EPYC 7532 processors, 32 physical cores per processor (64 in total), and 256 GB RAM. These timings apply to this implementation and workstation, not to the computational resources used for the underlying WRF simulation or to GPU inference.

Table 7 reports the measured computational requirements of the benchmark models. Ridge had negligible training costs and required an average of 0.544 s for full-domain inference. Random Forest was considerably more demanding, requiring 12.655 s for inference and producing the largest model file (232.107 MB). MLP was relatively lightweight,

with 3713 trainable parameters and an average inference time of 1.015 s, close to the 1.196 s required by XGBoost. The recurrent models were much slower. GRU required an average of 84.660 s for full-domain inference, while LSTM required 448.555 s and contained 116,801 trainable parameters. The additional complexity of these architectures therefore incurred a substantial computational cost, although it did not always produce a comparable gain in predictive accuracy.

Table 7. Mean measured complexity across T_2 , Q_2 , W_{10} , and AC. Times refer to final refitting and 24 h full-domain inference in the CPU-based computational environment used in this study.

Model	Train (s)	Inference (s)	Parameters	Size (MB)	Pareto Variables
Persistence	0.000	0.008	0	0.000	Q_2
Previous-day	0.000	0.004	0	0.000	T_2 , Q_2 , AC, W_{10}
Ridge	0.005	0.544	25	0.001	T_2 , AC, W_{10}
Random Forest	1.719	12.655	—	232.107	T_2 , Q_2
XGBoost	1.083	1.196	—	1.386	T_2
MLP	6.586	1.015	3713	0.018	T_2 , AC
GRU	57.763	84.660	12,929	0.055	—
LSTM	62.573	448.555	116,801	0.471	Q_2

The accuracy–efficiency trade-off differed by target (Figure 16). For T_2 , Random Forest achieved the lowest RMSE, while Previous-day, Ridge, XGBoost, and MLP offered less costly Pareto-optimal alternatives. For Q_2 , Persistence and Previous-day occupied the low-cost end of the front, followed by Random Forest and LSTM; LSTM achieved the lowest error but required substantially more inference time. For AC, Previous-day, Ridge, and MLP were Pareto optimal, with MLP achieving the lowest error for the WRF-BEM diagnostic at modest cost. For W_{10} , only Previous-day and Ridge were on the front: Ridge reduced RMSE from 1.300 to 1.095 m/s, while the nonlinear models offered no gain in accuracy to offset their additional cost. These trade-offs describe the present test period and implementation, not a hardware-independent ranking of model families.

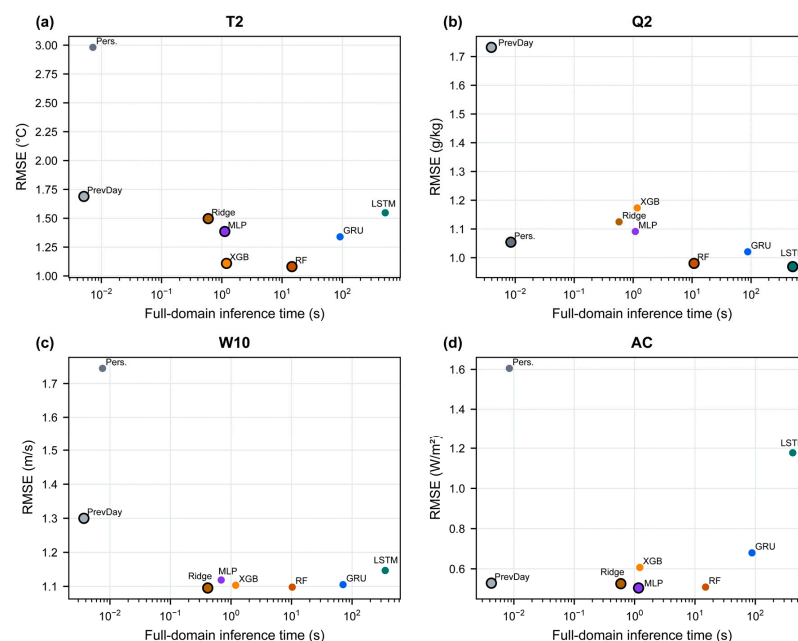


Figure 16. Final-test accuracy-complexity Pareto fronts for four targets. Black-edged markers are non-dominated for each variable. (a) 2 m air temperature (T_2); (b) 2 m water-vapor mixing ratio (Q_2); (c) 10 m wind-speed magnitude (W_{10}); (d) WRF-BEM air-conditioning power density (AC).

5. Discussion

The station comparison provides limited evidence that the WRF meteorological fields are physically plausible for this case. It does not formally validate the full simulation or establish the observation-based accuracy of its surrogates. The benchmark evaluates how closely models reproduce held-out WRF output, including any biases in that reference. AC is an instantaneous WRF-BEM diagnostic and has not been validated here against metered electricity use. These distinctions constrain the interpretation of the application-oriented findings below.

5.1. Complexity Is Variable- and Objective-Dependent

Within this single-event, single-city benchmark, the results do not support a monotonic relationship between model complexity and predictive value. This finding agrees with atmospheric and land-surface emulator comparisons in which the preferred family changes with the target process, internal variability affects rankings, and added complexity is not necessarily Pareto-efficient [30–33]. Tree ensembles performed particularly well for T_2 in the present experiment. The 24 h T_2 history contained strong daily memory, but forecasts for the final period still benefited from nonlinear correction. LSTM lowered domain-average Q_2 error, although its spatial smoothing suggests that much of the gain came from trajectory and bias adjustment rather than reconstruction of local moisture structure. A compact MLP reduced high-load AC errors; deeper recurrent models added cost and instability without improving domain-average accuracy. W_{10} provides a case-specific counterexample to the use of complexity as a default selection criterion: regularized linear autoregression outperformed every nonlinear and recurrent family despite the dynamical nature of wind.

For applications under conditions similar to this event, the results identify different candidates for further testing. Random Forest offers the lowest T_2 error and preserves spatial variability well. LSTM has the lowest aggregate Q_2 RMSE but smooths the moisture field, which limits its usefulness where local contrasts matter. MLP provides a comparatively inexpensive approximation of the WRF-BEM AC diagnostic, rather than a validated estimate of actual building electricity demand. Ridge has the lowest W_{10} error in this test. None of these choices have yet been evaluated across independent heatwaves, cities, or other weather regimes.

5.2. Why Temporal Backtesting Matters

A single validation block would have led to overconfident conclusions and, for some variables, different model choices. The three development folds reveal phase dependence within the heatwave, especially for Q_2 and the recurrent models. The difference between development and final-test rankings does not justify post hoc retuning; it indicates temporal non-stationarity within this short event. Time-series forecasting studies warn that random resampling can give misleading estimates of performance under temporal dependence or non-stationarity and recommend ordered holdouts or repeated out-of-sample evaluation with strong baselines [49,50]. Our expanding-window design follows that principle and reserves h97–h120 as an untouched final test. It improves robustness within the event but does not replace validation on independent meteorological events.

5.3. Spatial Fidelity, Critical Regimes, and Added Value

The spatial diagnostics explain why domain-average RMSE is insufficient. Q_2 is the clearest case: LSTM had the lowest mean RMSE but reproduced only a small fraction of the WRF spatial standard deviation. By contrast, Random Forest for T_2 and MLP for AC retained the main spatial patterns more faithfully. Maps of winning margin and Δ RMSE also show that the model with the lowest domain-average RMSE has little local advantage

over simpler alternatives across large areas. This result is consistent with climate-emulator benchmarks that evaluate spatially resolved fields [51] and with rapid emulators developed for spatially resolved temperature responses [52]. The issue is particularly relevant for AC because non-urban and near-zero-load cells occupy much of the domain, while high-load urban cells contribute more to squared error.

The regime diagnostics also place the aggregate rankings in context. Errors increased in the strongest winds and during rapid W10 changes, even though Ridge performed best on average. Spatial neighbors or additional meteorological variables might help represent these conditions, and a shorter history might better capture some wind variations. The present experiment does not assess these possibilities: all trainable models use the same 24 h univariate history, with no spatial-input, multivariable-input, or window-length ablation. The results therefore identify limitations of the tested forecasts without establishing their cause.

5.4. Relation to Modern Weather–AI Architectures

Modern weather models use atmospheric information at scales and resolutions that differ from those in this benchmark. Evaluations of global AI models show that their rankings vary with region and forecast objective [53]. GenCast [54] and Aurora [55] use much broader training archives and atmospheric information. Here, the input is limited to a local univariate history. Their reported skill cannot establish how they would perform on this event-level task. Spatial and multivariable surrogates are relevant candidates for future comparison, but this study does not demonstrate that they would resolve the W_{10} errors or the Q2 smoothing.

5.5. Limitations

This benchmark covers a single 120 h heatwave simulation in Chongqing, with one held-out 24 h test period. Its rankings may change across events, seasons, cities, or meteorological regimes. The three development folds sample different parts of the same event and are not independent climatological replications. All trainable families use a 24 h pointwise univariate history; alternative history lengths, spatial neighbors, and additional variables have not been tested. Their potential benefits remain hypotheses. The station comparison provides a limited check on WRF meteorological plausibility, while the AC diagnostic lacks metered-energy validation. Absolute computation times also depend on the hardware, software, and implementation used. Broader deployment would require evaluation using independent events and observations suited to the intended application.

6. Conclusions

- (1) In this single-event, single-city benchmark, the lowest 24 h RMSE was achieved by Random Forest for T2 (1.081 ± 0.002 °C), LSTM for Q2 (0.969 ± 0.063 g/kg), MLP for AC (0.505 ± 0.187 W/m²), and Ridge for W10 (1.095 m/s). The AC result concerns emulation of WRF-BEM diagnostic power density only; it does not establish accuracy against measured building electricity consumption.
- (2) W_{10} had moderate lag-1 memory but weak lag-24 memory (0.581/0.211). Ridge outperformed Previous-day at 75.0% of cells, retained 68.7% of WRF spatial variability on average, and showed much larger errors during strong winds and rapid changes.
- (3) Added recurrent complexity did not improve W10 RMSE over Ridge in this test, and the lowest aggregate Q2 error coincided with substantial spatial smoothing. These findings support evaluating spatial structure and computational cost alongside aggregate error in this case. Whether spatial or multivariable inputs would improve these forecasts remains untested.

- (4) The present results identify case-specific candidate surrogates, not a universal ordering of model families. Future comparisons should test alternative input windows and spatial or multivariable predictors across independent events and cities before drawing conclusions about transfer to other meteorological regimes.

Author Contributions: Conceptualization, Y.L., H.L., R.D. and B.-J.H.; methodology, Y.L., R.X., M.C. and R.D.; software, Y.L., R.X., R.D. and M.C.; validation, Y.L., M.C. and R.D.; formal analysis, Y.L., R.X. and M.C.; investigation, Y.L., H.L., and M.C.; data curation, Y.L., R.X. and R.D.; writing: original draft preparation, Y.L., R.X., H.L., B.-J.H., R.D. and M.C.; writing—review and editing, H.L., R.D., and B.-J.H.; visualization, Y.L., R.X., H.L., R.D., B.-J.H. and M.C.; supervision, Y.L., H.L., R.D. and B.-J.H.; project administration, Y.L.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Chongqing Municipal Social Science Planning Cultivation Project (No. 2025PY37), the Open Research Foundation of Chongqing Key Laboratory of Numerical Model-AI Integrated Forecast and Warning for Severe Weather (No. MAIN-2025031), the Chongqing Municipal Education Commission Science and Technology Project (No. KJZD-K202400706), the Chongqing Jiaotong University Municipal-Level Joint Graduate Training Base (No. JDLHPYJD2020038), and the Chongqing Municipality Construction Science and Technology Project (Chengke 2023 No. 2-17).

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors thank the National Supercomputing Center in Guangzhou and its Nansha subcenter for providing high-performance computing resources.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

WRF	Weather Research and Forecasting Model
BEP	Building Effect Parameterization
BEM	Building Energy Model
LCZ	Local Climate Zone
T ₂	2 m air temperature
Q ₂	2 m water-vapor mixing ratio
AC	WRF-BEM instantaneous air-conditioning power density
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
ME	Mean Error
MLP	Multilayer Perceptron
GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory
W ₁₀	10 m horizontal wind-speed magnitude

References

1. Stewart, I.D.; Oke, T.R. Local Climate Zones for Urban Temperature Studies. *Bull. Am. Meteorol. Soc.* **2012**, *93*, 1879–1900. [\[CrossRef\]](#)
2. Ching, J.; Mills, G.; Bechtel, B.; See, L.; Feddema, J.; Wang, X.; Ren, C.; Brousse, O.; Martilli, A.; Neophytou, M.; et al. WUDAPT: An Urban Weather, Climate, and Environmental Modeling Infrastructure for the Anthropocene. *Bull. Am. Meteorol. Soc.* **2018**, *99*, 1907–1924. [\[CrossRef\]](#)
3. Demuzere, M.; Kittner, J.; Martilli, A.; Mills, G.; Moede, C.; Stewart, I.D.; van Vliet, J.; Bechtel, B. A Global Map of Local Climate Zones to Support Earth System Modelling and Urban-Scale Environmental Science. *Earth Syst. Sci. Data* **2022**, *14*, 3835–3873. [\[CrossRef\]](#)

4. Chen, F.; Kusaka, H.; Bornstein, R.; Ching, J.; Grimmond, C.S.B.; Grossman-Clarke, S.; Loridan, T.; Manning, K.W.; Martilli, A.; Miao, S.; et al. The Integrated WRF/Urban Modelling System: Development, Evaluation, and Applications to Urban Environmental Problems. *Int. J. Climatol.* **2011**, *31*, 273–288. [\[CrossRef\]](#)
5. Masson, V. A Physically-Based Scheme for the Urban Energy Budget in Atmospheric Models. *Bound.-Layer Meteorol.* **2000**, *94*, 357–397. [\[CrossRef\]](#)
6. Kusaka, H.; Kondo, H.; Kikegawa, Y.; Kimura, F. A Simple Single-Layer Urban Canopy Model for Atmospheric Models: Comparison with Multi-Layer and Slab Models. *Bound.-Layer Meteorol.* **2001**, *101*, 329–358. [\[CrossRef\]](#)
7. Martilli, A.; Clappier, A.; Rotach, M.W. An Urban Surface Exchange Parameterisation for Mesoscale Models. *Bound.-Layer Meteorol.* **2002**, *104*, 261–304. [\[CrossRef\]](#)
8. Salamanca, F.; Krpo, A.; Martilli, A.; Clappier, A. A New Building Energy Model Coupled with an Urban Canopy Parameterization for Urban Climate Simulations-Part I: Formulation, Verification, and Sensitivity Analysis of the Model. *Theor. Appl. Climatol.* **2010**, *99*, 331–344. [\[CrossRef\]](#)
9. Salamanca, F.; Martilli, A. A New Building Energy Model Coupled with an Urban Canopy Parameterization for Urban Climate Simulations-Part II: Validation with One-Dimensional Off-Line Simulations. *Theor. Appl. Climatol.* **2010**, *99*, 345–356. [\[CrossRef\]](#)
10. Molnar, G.; Gyongyosi, A.Z.; Gal, T. Integration of an LCZ-Based Classification into WRF to Assess the Intra-Urban Temperature Pattern under a Heatwave Period in Szeged, Hungary. *Theor. Appl. Climatol.* **2019**, *138*, 1139–1158. [\[CrossRef\]](#)
11. Hu, H.; Zhang, X.; He, C.; Fertig, E.; Zarrin, R.; Neyestani, S.E.; Zhang, J. Quantifying Urban Morphology-Induced Uncertainty in Urban Meteorology and Heat Stress Simulations in Southern California. *J. Geophys. Res. Atmos.* **2026**, *131*, e2025JD045318. [\[CrossRef\]](#)
12. Patel, P.; Jamshidi, S.; Nadimpalli, R.; Aliaga, D.G.; Mills, G.; Chen, F.; Demuzere, M.; Niyogi, D. Modeling Large-Scale Heatwave by Incorporating Enhanced Urban Representation. *J. Geophys. Res. Atmos.* **2022**, *127*, e2021JD035316. [\[CrossRef\]](#)
13. Brousse, O.; Simpson, C.H.; Zonato, A.; Martilli, A.; Taylor, J.; Davies, M.; Heaviside, C. Cool Roofs Could Be Most Effective at Reducing Outdoor Urban Temperatures in London (United Kingdom) Compared with Other Roof Top and Vegetation Interventions: A Mesoscale Urban Climate Modeling Study. *Geophys. Res. Lett.* **2024**, *51*, e2024GL109634. [\[CrossRef\]](#)
14. Ding, D.; Xian, M.; Wang, H. Hidden Impacts of Elevator Addition on Community Microclimate: A Simulation-Based Case Study. *Urban Build. Sci.* **2025**, *1*, 2. [\[CrossRef\]](#)
15. Salamanca, F.; Georgescu, M.; Mahalov, A.; Moustou, M.; Wang, M. Anthropogenic Heating of the Urban Environment due to Air Conditioning. *J. Geophys. Res. Atmos.* **2014**, *119*, 5949–5965. [\[CrossRef\]](#)
16. Pappaccogli, G.; Giovannini, L.; Zardi, D.; Martilli, A. Assessing the Ability of WRF-BEP + BEM in Reproducing the Wintertime Building Energy Consumption of an Italian Alpine City. *J. Geophys. Res. Atmos.* **2021**, *126*, e2020JD033652. [\[CrossRef\]](#)
17. Li, X.C.; Zhao, L.; Oleson, K.; Zhou, Y.; Qin, Y.; Zhang, K.; Fang, B. Enhancing Urban Climate-Energy Modeling in the Community Earth System Model (CESM) through Explicit Representation of Urban Air-Conditioning Adoption. *J. Adv. Model. Earth Syst.* **2024**, *16*, e2023MS004107. [\[CrossRef\]](#)
18. Fang, H.; Jiang, J.; Qin, M.; Han, X.; Yuan, Y.; Sun, S. Toward Personalized Thermal Comfort: A Spatiotemporal Dynamic Heating Strategy for the Catalytic Combustion Local Heater. *Urban Build. Sci.* **2026**, *2*, 1. [\[CrossRef\]](#)
19. Tanoori, G.; Soltani, A.; Modiri, A. Machine Learning for Urban Heat Island (UHI) Analysis: Predicting Land Surface Temperature (LST) in Urban Environments. *Urban Clim.* **2024**, *55*, 101962. [\[CrossRef\]](#)
20. Buster, G.; Cox, J.; Benton, B.N.; King, R.N. Estimating the Impacts of Increasing Temperatures and the Efficacy of Climate Adaptation Strategies in Urban Microclimates with Deep Learning. *Urban Clim.* **2025**, *64*, 102603. [\[CrossRef\]](#)
21. Tang, J.; Hou, H.; Estoque, R.C.; Murayama, Y.; Zhang, X.; Wang, R.; Hu, T. Urban Heat Risk Assessment: A Machine Learning Analysis of Multidimensional Morphology in 293 Chinese Cities. *npj Urban Sustain.* **2026**, *in press*. [\[CrossRef\]](#)
22. Camps-Valls, G.; Fernandez-Torres, M.A.; Cohrs, K.H.; Höhl, A.; Castelletti, A.; Pacal, A.; Robin, C.; Martinuzzi, F.; Papoutsis, I.; Prapas, I.; et al. Artificial Intelligence for Modeling and Understanding Extreme Weather and Climate Events. *Nat. Commun.* **2025**, *16*, 1919. [\[CrossRef\]](#)
23. Kochkov, D.; Yuval, J.; Langmore, I.; Norgaard, P.; Smith, J.; Mooers, G.; Kloewer, M.; Lottes, J.; Rasp, S.; Dueben, P.; et al. Neural General Circulation Models for Weather and Climate. *Nature* **2024**, *632*, 1060–1066. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Yuval, J.; O’Gorman, P.A. Stable Machine-Learning Parameterization of Subgrid Processes for Climate Modeling at a Range of Resolutions. *Nat. Commun.* **2020**, *11*, 3295. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Meyer, D.; Grimmond, S.; Dueben, P.; Hogan, R.; van Reeuwijk, M. Machine Learning Emulation of Urban Land Surface Processes. *J. Adv. Model. Earth Syst.* **2022**, *14*, e2021MS002744. [\[CrossRef\]](#)
26. Rampal, N.; Gibson, P.B.; Sherwood, S.; Abramowitz, G.; Hobeichi, S. A Reliable Generative Adversarial Network Approach for Climate Downscaling and Weather Generation. *J. Adv. Model. Earth Syst.* **2025**, *17*, e2024MS004668. [\[CrossRef\]](#)
27. Jiang, P.; Yang, Z.; Wang, J.; Huang, C.; Xue, P.; Chakraborty, T.C.; Chen, X.; Qian, Y. Efficient Super-Resolution of Near-Surface Climate Modeling Using the Fourier Neural Operator. *J. Adv. Model. Earth Syst.* **2023**, *15*, e2023MS003800. [\[CrossRef\]](#)

28. Zhong, X.; Ma, Z.; Yao, Y.; Xu, L.; Wu, Y.; Wang, Z. WRF-ML v1.0: A Bridge between WRF v4.3 and Machine Learning Parameterizations and Its Application to Atmospheric Radiative Transfer. *Geosci. Model Dev.* **2023**, *16*, 199–209. [[CrossRef](#)]
29. Rasp, S.; Hoyer, S.; Merose, A.; Langmore, I.; Battaglia, P.; Russell, T.; Sanchez-Gonzalez, A.; Yang, V.; Carver, R.; Agrawal, S.; et al. WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models. *J. Adv. Model. Earth Syst.* **2024**, *16*, e2023MS004019. [[CrossRef](#)]
30. Di Santo, D.; He, C.; Chen, F.; Giovannini, L. ML-AMPSIT: Machine Learning-Based Automated Multi-Method Parameter Sensitivity and Importance Analysis Tool. *Geosci. Model Dev.* **2025**, *18*, 433–459. [[CrossRef](#)]
31. Wesselkamp, M.; Chantry, M.; Pinnington, E.; Choulga, M.; Boussetta, S.; Kalweit, M.; Bodecker, J.; Dormann, C.F.; Pappenberger, F.; Balsamo, G. Advances in Land Surface Forecasting: A Comparison of LSTM, Gradient Boosting, and Feed-Forward Neural Networks as Prognostic State Emulators in a Case Study with ecLand. *Geosci. Model Dev.* **2025**, *18*, 921–937. [[CrossRef](#)]
32. Beucler, T.; Grundner, A.; Shamekh, S.; Ukkonen, P.; Chantry, M.; Lagerquist, R. Distilling Machine Learning’s Added Value: Pareto Fronts in Atmospheric Applications. *Artif. Intell. Earth Syst.* **2025**, *4*, e240078. [[CrossRef](#)]
33. Lütjens, B.; Ferrari, R.; Watson-Parris, D.; Selin, N.E. The Impact of Internal Variability on Benchmarking Deep Learning Climate Emulators. *J. Adv. Model. Earth Syst.* **2025**, *17*, e2024MS004619. [[CrossRef](#)]
34. Bi, K.; Xie, L.; Zhang, H.; Chen, X.; Gu, X.; Tian, Q. Accurate Medium-Range Global Weather Forecasting with 3D Neural Networks. *Nature* **2023**, *619*, 533–538. [[CrossRef](#)] [[PubMed](#)]
35. Lam, R.; Sanchez-Gonzalez, A.; Willson, M.; Wirnsberger, P.; Fortunato, M.; Alet, F.; Ravuri, S.; Ewalds, T.; Eaton-Rosen, Z.; Hu, W.; et al. Learning Skillful Medium-Range Global Weather Forecasting. *Science* **2023**, *382*, 1416–1421. [[CrossRef](#)] [[PubMed](#)]
36. Abdi, D.; Jankov, I.; Madden, P.; Vargas, V.; Smith, T.A.; Frolov, S.; Flora, M.; Potvin, C. HRRRCast: A Data-Driven Emulator for Regional Weather Forecasting at Convection-Allowing Scales. *Artif. Intell. Earth Syst.* **2026**, *5*, 250061. [[CrossRef](#)]
37. Pathak, J.; Cohen, Y.; Garg, P.; Harrington, P.; Brenowitz, N.; Durran, D.; Mardani, M.; Vahdat, A.; Xu, S.; Kashinath, K.; et al. Kilometer-Scale Convection-Allowing Model Emulation Using Generative Diffusion Modeling. *Sci. Adv.* **2026**, *12*, eadv0423. [[CrossRef](#)]
38. Fillola, E.; Santos-Rodriguez, R.; Tunnicliffe, R.; Clark, J.N.; Keshtmand, N.; Ganesan, A.; Rigby, M. Enabling Fast Greenhouse Gas Emissions Inference from Satellites with GATES: A Graph-Neural-Network Atmospheric Transport Emulation System. *Geosci. Model Dev.* **2026**, *19*, 1893–1915. [[CrossRef](#)]
39. Meng, X.; Wang, Y.; Yang, Y.; Huang, Q.; Guo, J.; Lv, Y.; Dai, Z. Deep Learning-Based Urban Canopy Model for Complex Building Morphologies. *J. Adv. Model. Earth Syst.* **2026**, *18*, e2025MS005483. [[CrossRef](#)]
40. Brenowitz, N.D.; Cohen, Y.; Pathak, J.; Mahesh, A.; Bonev, B.; Kurth, T.; Durran, D.R.; Harrington, P.; Pritchard, M.S. A Practical Probabilistic Benchmark for AI Weather Models. *Geophys. Res. Lett.* **2025**, *52*, e2024GL113656. [[CrossRef](#)]
41. Natel, C.; Belda, D.M.; Anthoni, P.; Hass, N.; Rabin, S.; Arneth, A. Emulating Grid-Based Forest Carbon Dynamics Using Machine Learning: An LPJ-GUESS v4.1.1 Application. *Geosci. Model Dev.* **2025**, *18*, 4317–4333. [[CrossRef](#)]
42. Keetz, L.T.; Aalstad, K.; Fisher, R.A.; Poppe Teran, C.; Naz, B.; Pirk, N.; Yilmaz, Y.A.; Skarpaas, O. Inferring Parameters in a Complex Land Surface Model by Combining Data Assimilation and Machine Learning. *J. Adv. Model. Earth Syst.* **2025**, *17*, e2024MS004542. [[CrossRef](#)]
43. Bouabid, S.; Sejdinovic, D.; Watson-Parris, D. FaIRGP: A Bayesian Energy Balance Model for Surface Temperatures Emulation. *J. Adv. Model. Earth Syst.* **2024**, *16*, e2023MS003926. [[CrossRef](#)]
44. Kendon, E.J.; Addison, H.; Doury, A.; Somot, S.; Watson, P.A.G.; Booth, B.B.B.; Coppola, E.; Gutierrez, J.M.; Murphy, J.; Scullion, C. Potential for Machine Learning Emulators to Augment Regional Climate Simulations in Provision of Local Climate Change Information. *Bull. Am. Meteorol. Soc.* **2025**, *106*, E1175–E1203. [[CrossRef](#)]
45. Bhimreddy, S.R.; Bhaganagar, K. Performance Assessment of Dynamic Downscaling of WRF to Simulate Convective Conditions during Sagebrush Phase 1 Tracer Experiments. *Atmosphere* **2018**, *9*, 505. [[CrossRef](#)]
46. Chongqing Meteorological Service. Chongqing Climate Impact Assessment for August 2022. China Weather. 6 September 2022. Available online: <https://www.weather.com.cn/chongqing/qxfwcp/yqhpi/09/3560321.shtml> (accessed on 5 September 2026). (In Chinese).
47. Ministry of Natural Resources of the People’s Republic of China. Standard Map of China (English), 1:60,000,000; Approval No. GS(2019)1652. Available online: <http://bzdt.ch.mnr.gov.cn/> (accessed on 6 September 2026).
48. Alibaba Cloud DataV. Chongqing District Boundary Data. Available online: https://geo.datav.aliyun.com/areas_v3/bound/500000_full.json (accessed on 5 September 2026).
49. Cerqueira, V.; Torgo, L.; Mozetič, I. Evaluating Time Series Forecasting Models: An Empirical Study on Performance Estimation Methods. *Mach. Learn.* **2020**, *109*, 1997–2028. [[CrossRef](#)]
50. Hewamalage, H.; Ackermann, K.; Bergmeir, C. Forecast Evaluation for Data Scientists: Common Pitfalls and Best Practices. *Data Min. Knowl. Discov.* **2023**, *37*, 788–832. [[CrossRef](#)] [[PubMed](#)]

51. Watson-Parris, D.; Rao, Y.; Olivié, D.; Seland, Ø.; Nowack, P.; Camps-Valls, G.; Stier, P.; Bouabid, S.; Dewey, M.; Fons, E.; et al. ClimateBench v1.0: A Benchmark for Data-Driven Climate Projections. *J. Adv. Model. Earth Syst.* **2022**, *14*, e2021MS002954. [[CrossRef](#)]
52. Womack, C.B.; Giani, P.; Eastham, S.D.; Selin, N.E. Rapid Emulation of Spatially Resolved Temperature Response to Effective Radiative Forcing. *J. Adv. Model. Earth Syst.* **2025**, *17*, e2024MS004523. [[CrossRef](#)]
53. Liu, C.-C.; Hsu, K.; Peng, M.S.; Chen, D.-S.; Chang, P.-L.; Hsiao, L.-F.; Fong, C.-T.; Hong, J.-S.; Cheng, C.-P.; Lu, K.-C.; et al. Evaluation of Five Global AI Models for Predicting Weather in Eastern Asia and Western Pacific. *npj Clim. Atmos. Sci.* **2024**, *7*, 221. [[CrossRef](#)]
54. Price, I.; Sanchez-Gonzalez, A.; Alet, F.; Andersson, T.R.; El-Kadi, A.; Masters, D.; Ewalds, T.; Stott, J.; Mohamed, S.; Battaglia, P.; et al. Probabilistic Weather Forecasting with Machine Learning. *Nature* **2025**, *637*, 84–90. [[CrossRef](#)] [[PubMed](#)]
55. Bodnar, C.; Bruinsma, W.P.; Lucic, A.; Stanley, M.; Allen, A.; Brandstetter, J.; Garvan, P.; Riechert, M.; Weyn, J.A.; Dong, H.; et al. A Foundation Model for the Earth System. *Nature* **2025**, *641*, 1180–1187. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.