Weather-guided surrogate modeling of building energy use in Nepal and zero-shot generalization across South Asia

Aashish Katuwal[a,*]

[a] Department of Mechanical and Aerospace Engineering, Pulchowk Campus, Institute of Engineering, Tribhuvan University, Lalitpur, Nepal

* Corresponding author. E-mail address: 077bme001.aashish@pcampus.edu.np

Abstract

Building performance simulation is a standard tool for energy codes, retrofit screening, and design, but it is seldom used in Nepal or elsewhere in South Asia. This paper applies a recently published weekly weather-guided surrogate framework to Nepal, where ten typical-meteorological-year stations cover 72–2,300 m of elevation within roughly 200 km. A total of 1,100 EnergyPlus simulations of the U.S. Department of Energy (DOE) medium-office prototype were run at 22 locations. A transformer trained on the Nepali stations predicted weekly energy use at fully held-out Nepali stations with a symmetric mean absolute percentage error (SMAPE) of 3.1–8.7%. One change to the reference method was required: weather inputs were scaled to fixed physical bounds rather than to per-location statistics. This is a precondition for zero-shot use and also improved transfer between climates. With no local simulation, the Nepal-trained model predicted ten cities in the other seven South Asian countries at 3.1–5.9% weekly SMAPE, and sliding-window training with a three-seed ensemble reduced the regional mean to 4.0%. When the training network was cut to three stations, every unseen Nepali station stayed within 5.7% as long as the three spanned the hot, intermediate, and cold regimes, whereas removing the single cold-mountain station degraded accuracy at every cold city. Ulaanbaatar, designated in advance as outside this coverage, was consistently underpredicted; additional Nepali stations or training windows did not correct this, and augmentation made it worse. Coverage of climate regimes, rather than the volume of training data, determines where the model can be trusted.

Keywords: building energy modeling; surrogate model; transformer; EnergyPlus; cross-location generalization; Nepal; South Asia

1. Introduction

Buildings are responsible for roughly 37% of global energy- and process-related CO_2 emissions [1]. In South Asia the building stock is expanding rapidly as cities grow and mechanical cooling becomes more widely available [2,3]. Nepal follows the same pattern in the Kathmandu Valley, along the Terai corridor, and in the mid-hill towns, and it faces the same constraint as its neighbors: building performance simulation with tools such as EnergyPlus [4], which underpins energy codes, retrofit screening, and design-stage optimization elsewhere, is used far less than the growth of the stock would justify. The obstacle is mostly practical: simulating even one building archetype over a design space takes hours of computation and specialist effort for every climate considered, and Nepal adds a difficulty of its own. Its inhabited climates range from subtropical plains below 100 m to cold mountain settlements above 2,000 m [5], one of the steepest gradients anywhere, so there is no single representative climate to simulate. A tool that returned simulation-grade energy estimates in seconds, for any Nepali climate, from nothing more than a public

weather file would put many more analyses within reach of local practitioners, municipalities, and researchers.

Surrogate models are the usual answer to this problem. A machine learning model is trained once on a set of simulations and can then be queried in milliseconds [6–9]. Metamodels based on regression, Gaussian processes, and neural networks are now standard tools for design-space exploration [10–12], for retrofit analysis at the scale of building categories [13], and for AI-driven design optimization more generally [14]. Most of them, however, are tied to a single location, and the transfer-learning approaches studied so far usually assume that at least some data from the target site are available for fine-tuning [15,16]. Weather-informed surrogates that do generalize across locations have typically needed very large training geographies to do so: Rackes et al. [17] used 426 Brazilian locations and Westermann et al. [18] used 569 Canadian locations. Manmatharasan et al. [19] recently argued that temporal resolution, not geographic breadth, is what lets a single surrogate serve many climates from few training sites. By conditioning the model on weekly weather sequences instead of annual aggregates, the surrogate learns how weather drives load rather than memorizing site totals, and it can then predict at unseen locations after training on only one or two. Their demonstration covered ten Canadian cities in five climate zones. The question addressed here is what the framework can do for a country such as Nepal and, because Nepal's elevation range packs much of South Asia's climate diversity into one small national network, what it can do for the wider region.

To answer this, the weekly weather-guided surrogate framework was implemented end to end for Nepal and evaluated at three successively wider scopes. The first is prediction within Nepal. When the model is trained on subsets of ten SRC-TMYx stations between 72 m and 2,300 m, how accurately does it predict the weekly energy use of a standard office prototype at Nepali stations it has never seen? The second is sensitivity to the training corpus. As the network is thinned, does accuracy depend on how many stations remain or on which climate regimes they cover? This matters wherever weather data, simulation budgets, or modeling expertise are scarce. The third is regional generalization. Trained on Nepal alone, with no local simulations and no fine-tuning, can the surrogate predict energy use in the other seven South Asian countries, whose climates range from the equatorial Maldives to the cold, arid highlands of Afghanistan?

The paper makes four contributions. First, it presents a complete, laptop-scale implementation of weekly weather-guided surrogate modeling for Nepal, based on 1,100 verified EnergyPlus simulations at 22 locations and four model architectures, with open code and data; the surrogate reaches 3.1–8.7% weekly SMAPE at fully held-out Nepali stations. Second, a station-count sensitivity analysis on the same network shows that accuracy depends on the climate regimes the training stations span rather than on their number, which suggests how a national network should be prioritized when resources are limited. Third, it introduces an adaptation that mountainous terrain makes necessary, namely scaling weather inputs to fixed physical bounds instead of per-location statistics, explains why this also improves cross-climate transfer, and recommends it as a default for deployments that cover large elevation ranges or are intended for zero-shot use. Fourth, it reports a continental-scale zero-shot evaluation, with 3.1–5.9% weekly SMAPE across all eight South Asian countries from the ten-station Nepali corpus, and shows through pre-designated boundary probes and a corpus ablation that transfer quality follows climate-regime coverage, so that the domain of validity can be predicted before deployment.

2. Reference framework and methodological adaptations

The implementation follows Manmatharasan et al. [19], who write the surrogate as $\hat{E} = f(\varphi(W), B)$. A sequence encoder φ summarizes one week of hourly weather W, and the resulting embedding is combined with the building design parameters B to predict the energy use of that week. Their central result, that weekly resolution is what makes cross-location transfer possible, is the property that matters most for Nepal, so their experimental design was kept wherever possible: the same DOE medium-office prototype [20], the same arrangement of 50 Latin hypercube designs by 52 weeks per location, a closely matching set of fourteen envelope and operation parameters, the same three sequence architectures (a temporal convolutional network, a transformer encoder, and a decoupled autoencoder whose weather encoder is pretrained on all training locations and then frozen), and the same single-teacher transfer grids for mapping which climates can teach which. Where the Nepalese setting required a departure, it is stated explicitly; Table 1 lists the differences.

Table 1. Correspondence between the reference framework and this implementation.

Element	**Manmatharasan et al. [19]**	**This study**	**Reason for adaptation**
Prototype and design space	DOE medium-office prototype; 50 LHS design variants over 14 parameters	Same prototype and design-space size; closely matching parameter set	Comparability
Climate system	Ten Canadian cities in five climate zones, from mild coastal to subarctic	Ten Nepali SRC-TMYx stations with elevation-driven climates (72–2,300 m within about 200 km)	Target geography
Architectures	TCN, transformer with sinusoidal positional encoding, decoupled convolutional autoencoder	Same three encoders plus an MLP baseline; the transformer uses learned positional embeddings	Comparability
Weather normalization	Min-max scaler fitted on the training-location weather file, without clipping	Min-max against fixed physical bounds per channel (Section 4.3)	Per-location statistics are undefined for unseen cities and unstable across large elevation offsets
Hyperparameters	Tuned separately for each training instance	Fixed across all experiments; early stopping with patience 15	Hundreds of training runs on one CPU; reported accuracy is conservative
Same-location evaluation	A second TMY file of the same location for validation and same-location testing	Held-out designs within the training stations; one TMY file per station	A single TMY file per Nepali station is the available data (Section 7.5)
Evaluation scope	Transfer within Canada under single, paired, and three-zone training	Within-country transfer, corpus-size sensitivity, zero-shot evaluation in seven other nations, and two out-of-region probes	Research questions of Section 1
Metrics	SMAPE, RMSE, Pearson r	Same set plus annual SMAPE, CV-RMSE, NMBE, and R^2	Regional reporting practice; bias (NMBE) diagnoses coverage gaps

Two of these changes deserve comment, because without them the framework would be reproducible in Nepal but not actually usable there. The normalization change (Section 4.3) is forced by the geography and by the zero-shot objective itself. A surrogate that will be queried for cities that contributed no training data cannot normalize its inputs with statistics of those cities. The fixed-hyperparameter protocol gives up some per-instance accuracy in exchange for a procedure that any regional institution can rerun unchanged on ordinary hardware. The accuracy figures reported below are therefore conservative compared with a tuned deployment, which is appropriate for a study whose main purpose is to build capability.

3. Study area and data

3.1. Nepal: the training geography

Nepal's habitable terrain rises from the Terai plains at about 70 m, through the inner valleys and mid-hills, to cold mountain districts above 2,000 m, all within roughly 200 km from south to north. Within a single national weather network, this produces a climate span that elsewhere would take a continent to cover [5]. Ten SRC-TMYx stations [21] form the training pool (Table 2). They include four hot-plains sites, one transition valley, humid and temperate mid-hill stations, two cool-hill stations, and one cold mountain station, Jumla, which anchors the heating regime. Typical meteorological year files are the usual weather basis for annual building simulation [22], and the TMYx series provides them free of charge for every station used here.

Table 2. Nepali training stations (coordinates and elevations from the EPW file headers).

Station	Elevation (m)	Lat. (°N)	Climate role
Biratnagar	72	26.48	Hot Terai, east (hot pole)
Bhairahawa (Siddharthanagar)	109	27.51	Hot Terai, west
Simara	137	27.16	Hot Terai, central
Nepalgunj	165	28.10	Hottest extreme, far-west plains
Surkhet (Birendranagar)	720	28.60	Inner valley, transition
Pokhara	827	28.20	Humid mid-hill
Kathmandu (Tribhuvan Intl)	1,338	27.70	Temperate mid-hill
Okhaldhunga	1,720	27.30	Cool hill, east
Dadeldhura (Amargadhi)	1,848	29.30	Cool hill, west
Jumla	2,300	29.28	Cold mountain (cold pole)

3.2. South Asia: the evaluation geography

Twelve test cities (Table 3) were chosen before any simulation outside Nepal was run. They are stratified by climate regime and cover all eight SAARC countries: Nepal supplies the training data and the other seven supply the tests. Two probes outside the region were added to mark the limits of the method. Seoul is a cold continental climate at low elevation, which separates cold from altitude, and Ulaanbaatar is an extreme continental climate that was designated in advance as the expected failure case. The ground-truth simulations at the test cities exist only to score the surrogate; none of their data is used in training, validation, early stopping, or normalization. The simulated weekly energy ranges in the last column confirm the intended regimes. The equatorial cities have no low-load season (the lowest week in Malé, 9,704 kWh,

is far above the lowest weeks anywhere in the Nepali corpus), and the peak weeks in Ulaanbaatar need more than twice the largest energy value in the whole training corpus. In this quantitative sense, Ulaanbaatar lies outside what the training data cover.

Table 3. Evaluation cities (all SRC-TMYx [21]; energy ranges over 50 designs and 52 weeks).

City	Country	Station, elev.	Regime / role	Weekly energy (kWh)
Patna	India	Patna-Jayaprakash AP, 52 m	Indo-Gangetic plains	6,421–15,498
Lahore	Pakistan	Allama Iqbal Intl, 217 m	Indo-Gangetic, west	6,523–16,010
Dhaka	Bangladesh	Hazrat Shahjalal Intl, 9 m	Monsoon delta	6,550–14,351
Chennai	India	Chennai Intl, 16 m	Tropical	8,911–16,305
Colombo	Sri Lanka	Katunayake, 7 m	Equatorial	9,800–15,162
Malé	Maldives	Velana Intl, 1.8 m	Oceanic equatorial	9,704–14,714
Karachi	Pakistan	Jinnah Intl, 30 m	Hot semi-arid coast	6,630–15,804
Jodhpur	India	Jodhpur, 219 m	Hot desert	6,554–16,236
Kabul	Afghanistan	Kabul Intl, 1,791 m	Continental cold-arid	6,680–17,466
Thimphu	Bhutan	Paro Intl, 2,235 m	Himalayan cold	5,839–13,873
Seoul	South Korea	Seongnam AP, 25 m	Out-of-region probe	6,217–20,713
Ulaanbaatar	Mongolia	Chinggis Khaan Intl, 1,330 m	Out-of-region probe (pre-designated failure)	6,385–35,809

4. Methodology

4.1. Simulation campaign

A Latin hypercube design matrix [23] draws 50 variants of the DOE medium-office prototype [20] over fourteen parameters: wall, roof, and floor insulation, applied as no-mass R-value overrides; window U-value and solar heat gain coefficient by orientation; window-to-wall scaling; building orientation; exterior surface absorptances; internal equipment power density; and heating and cooling setpoints. The variants are written onto the EnergyPlus prototype with eppy [24] and simulated for a full year under each weather file. With 50 designs at 22 stations and cities this gives 1,100 runs, which were executed by a six-worker parallel batch runner that skips completed runs and can be resumed. EnergyPlus sometimes reports severe errors yet still completes the annual run, so every output was checked for exactly 8,760 hourly records before it entered the dataset; all 1,100 runs passed. Hourly facility energy (electricity and natural gas meters summed) is aggregated to calendar weeks. For each location this yields a weather tensor $W \in \mathbb{R}^{2600\times168\times8}$ (50 designs by 52 weeks of 168-hour sequences over eight EPW channels: dry-bulb and dew-point temperature, relative humidity, station pressure, global and diffuse horizontal irradiance, wind direction, and wind speed), a design matrix $B \in \mathbb{R}^{2600\times14}$, and weekly energy $E \in \mathbb{R}^{2600}$ in kWh.

4.2. Surrogate architectures

Four models share a single interface. A weather encoder φ(W) is concatenated with the min-max scaled design parameters B and passed to a feed-forward head that predicts weekly energy, which is standardized during training. The four encoders are a multilayer perceptron on flattened weather (36k parameters), which

serves as the baseline for the resolution ablation; a temporal convolutional network [25] (69k); a transformer encoder [26] with learned positional embeddings (97k); and the decoupled autoencoder of the reference framework, whose convolutional weather encoder is first pretrained to reconstruct the 520 distinct weeks of the ten Nepali stations and then frozen while the regression head is trained. Sequence encoders of this kind are the usual deep-learning tools for energy time series [27]. All models were implemented in PyTorch [28] and trained with Adam, with early stopping on validation mean squared error at a patience of 15. Hyperparameters were held fixed in every experiment.

4.3. Physically bounded weather normalization

Using the framework across Nepal's elevation range, and in particular for cities that contribute no training data, calls for a weather normalization that does not depend on location statistics. Each channel was therefore min-max scaled against fixed physical bounds wide enough to contain any plausible terrestrial value: dry-bulb temperature from −40 to +50 °C, dew point from −45 to +35 °C, relative humidity from 0 to 100%, station pressure from 45 to 105 kPa, irradiance from 0 to 1,400 W m^{-2}, wind direction from 0 to 360°, and wind speed from 0 to 40 m s^{-1}. This choice was made for two reasons. The first concerns correctness under elevation offsets. A min-max scaler fitted on one station maps the pressure of a station at a different elevation far outside the unit interval (station pressure at Jumla is roughly 77% of that at Biratnagar), and in pilot experiments this sent transferred predictions out of range. The second is less obvious and probably more broadly relevant: even a scaler fitted on the whole training corpus turns out to be a poor choice for transfer. Such a scaler stretches pressure, whose variance within a station is small but whose offsets between stations are large, across the full unit interval, and in doing so hands the network an almost perfect location-identity feature. Under physical bounds the same offsets occupy only about 0.4 of the interval, which returns relative weight to the channels that actually drive the thermodynamics. Section 5.3 quantifies the effect. Physically bounded normalization is suggested as a sensible default for weather-guided surrogates aimed at mountainous terrain or zero-shot use.

4.4. Training corpora

Four transformer corpora were used for the regional evaluation. They were chosen to test how corpus size and regime coverage affect transfer, with the full ten-station corpus (iv) as the reference: (i) two stations, Surkhet and Jumla (transition valley and cold pole); (ii) three stations, Biratnagar, Okhaldhunga, and Jumla (hot pole, mid-hill, and cold pole); (iii) eight stations, excluding Jumla and Dadeldhura, kept on purpose as a model with many stations but no cold regime; and (iv) all ten stations.

4.5. Evaluation protocol

Model selection used 20% of the designs, held out within the training stations only. The evaluation locations contribute no gradients, no early-stopping signal, and, because of the normalization in Section 4.3, no scaling statistics, so every cross-location figure in this paper is strictly zero-shot. Within Nepal, two experiments are reported: holdout transfer (train on eight stations, validate on a ninth, test on the tenth) and the ten-by-ten single-teacher grids for each architecture. At the regional scale, each corpus of Section 4.4 was scored on all 2,600 samples of each test city. The metrics are weekly and annual SMAPE, per-design Pearson correlation averaged over designs, RMSE, CV-RMSE, NMBE, and R^2, with CV-RMSE and NMBE following the conventions of ASHRAE Guideline 14 [29]. NMBE is given particular attention at

sites outside coverage, because coverage failures show up as systematic bias. Where the true weekly variance is small, as in the equatorial cities, correlation metrics are reported with an explicit caveat (Section 6.3). All computation ran on the CPU of one consumer laptop. The simulations took roughly 2 to 4 hours per city in the background, transformer training took 1.5 to 5.1 hours per corpus on calendar weeks (up to 73 hours for the largest sliding-window runs), and inference for a whole city takes less than one second.

4.6. Sliding-window augmentation and seed ensembling

Two refinements to training were evaluated as an ablation on the three-station corpus. Neither requires new simulations. In the first, sliding windows, training sequences are taken from the same EPW years and hourly simulation output as 168 h windows advanced in 24 h steps, which gives 359 windows per design and year (17,950 samples per station) instead of 52 calendar weeks, roughly a sevenfold increase in weather diversity. Windows that belong to held-out designs are excluded from training, and validation, early stopping, and all evaluation stay on the calendar-week datasets, so every reported metric remains directly comparable with the runs without augmentation. In the second, ensembling, three models are trained from different random initializations with the validation-design split held fixed, and their predictions are averaged; the standard deviation across members is kept as a simulation-free indicator of uncertainty, in the spirit of deep ensembles [30]. Hyperparameters are those of Section 4.2. The same protocol was then applied to the ten-station corpus, with three independently initialized runs and their three-member ensemble.

5. Results I: a national surrogate for Nepal

5.1. Holdout transfer

With eight stations for training, one for validation, and one fully held out, the transformer reaches a weekly SMAPE of 3.32% at Pokhara (annual 0.58%, R^2 0.88) and 3.10% at Nepalgunj (annual 0.31%, R^2 0.96). Both are interpolation cases inside the elevation ladder. At Jumla the weekly SMAPE is 8.71% (annual 5.93%, Pearson r 0.94); this is the hardest case, because the cold pole has to be extrapolated from stations no higher than 1,848 m. In practical terms, the surrogate gives simulation-grade weekly profiles at locations bracketed by the national network, and above the ceiling of the network it remains usable, with a degradation that can be quantified.

5.2. Sensitivity to the number of training stations

Thinning the training network shows how much of this accuracy depends on the number of stations. Two stations (Surkhet and Jumla) keep all eight unseen Nepali stations within 6.9% weekly SMAPE (Fig. 1). Three stations at the regime extremes (Biratnagar as the hot pole, Okhaldhunga as the mid-hill, Jumla as the cold pole) keep every unseen station within 5.7% weekly SMAPE, with annual errors mostly at or below 1% and Pearson r of at least 0.87. The full ten-station model, which sees every regime, remains the most accurate (0.9–2.1% weekly SMAPE on held-out designs within the country), but accuracy falls gradually rather than abruptly as the network is thinned, as long as the stations that remain still span the national regimes. The selection rule echoes the strategic multi-zone mixing examined in [19], where combining one mild, one intermediate, and one extreme Canadian zone gave robust transfer. Where a network has to be

prioritized, the first stations should cover the regimes the territory contains (its hot pole, its cold pole, and a midpoint); density can be added afterwards.

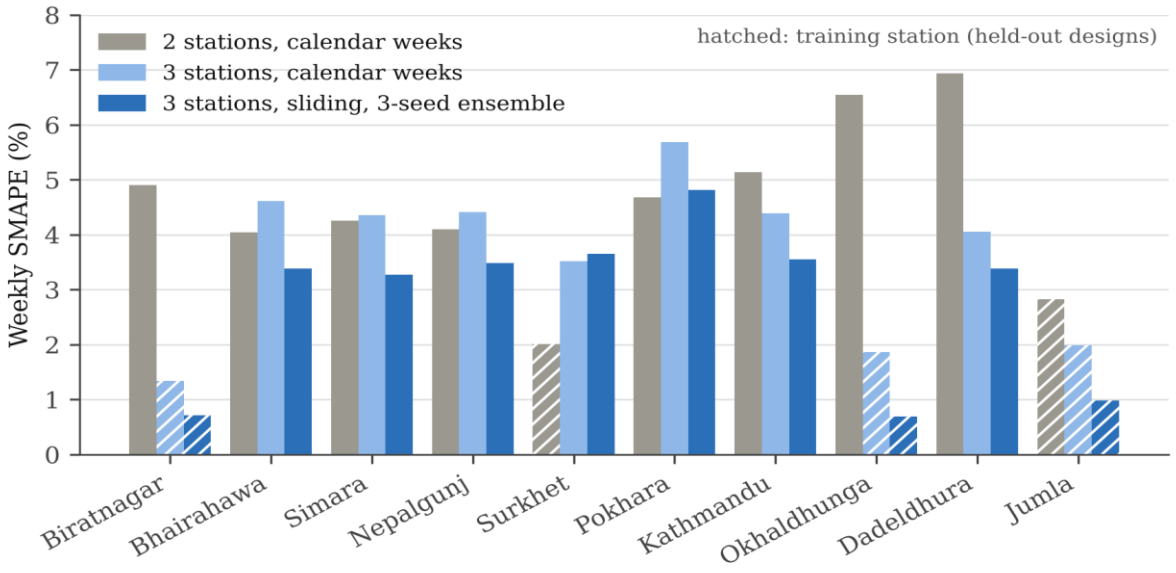


Fig. 1. Weekly SMAPE at the ten Nepali stations for the two-station corpus (Surkhet and Jumla), the three-station corpus (Biratnagar, Okhaldhunga, and Jumla), and the three-station sliding-window ensemble of Section 4.6. Hatched bars are training stations, evaluated on held-out designs; all other bars are fully unseen locations.

5.3. The normalization adaptation, quantified

With min-max normalization fitted on the training location, the single-teacher TCN grid is unusable off the diagonal. Errors saturate the SMAPE scale as the transferred pressure values leave the unit interval; the failed grid is archived with the released code. A scaler fitted on the full training corpus restores stability, but physically bounded normalization also improves transfer in cases where stability was never in doubt: the mean weekly SMAPE of the best TCN teacher over all ten targets drops from about 17.7% with the corpus-wide scaler to about 6.7% with physical bounds, which agrees with the location-identity mechanism described in Section 4.3. All results in this paper use physical bounds. The per-location convention worked well in the Canadian setting of [19], and nothing here contradicts that. What makes the normalization convention decisive here is the much larger station-pressure offset across Nepal's elevation span, together with the zero-shot objective.

5.4. Teacher grids and failure modes as guidance for corpus design

Across the ten-by-ten grids (Fig. 2), mid-elevation stations are the best single teachers under every architecture; for the transformer the best is Surkhet, with a mean weekly SMAPE of 5.4% and a worst cell of 13.1%. This reproduces the observation in [19] that middle climate zones generalize best, with elevation here playing the part of the zone ladder, and it agrees with their reading of extreme-climate training as a data-imbalance problem. Two failure modes recur when a single extreme-climate station teaches alone, and both can be diagnosed in advance. The first is extrapolation collapse from a hot-pole teacher. A TCN trained only on plains data predicts worse and worse toward the mountains (Nepalgunj to Jumla: 188.9% weekly SMAPE, NMBE of −111%), because cold weeks lie entirely outside its experience. The second is regime

inversion from a cold-pole teacher. A TCN or autoencoder trained only on Jumla is anticorrelated with the simulation at every warmer station (Pearson r between −0.45 and −0.89): it has learned the heating-regime rule that warmer weather means less energy and applied it where cooling dominates. Both modes persist in the TCN and autoencoder under every normalization tested. The transformer under physical bounds avoids both: its Jumla-teacher correlations are all positive (0.17 to 0.66) and its worst plains-teacher cell is 23.3%. Two deployment rules follow from this: do not train on a single climatic extreme, and prefer the transformer with physically bounded inputs. The latter is the configuration used in Section 6.

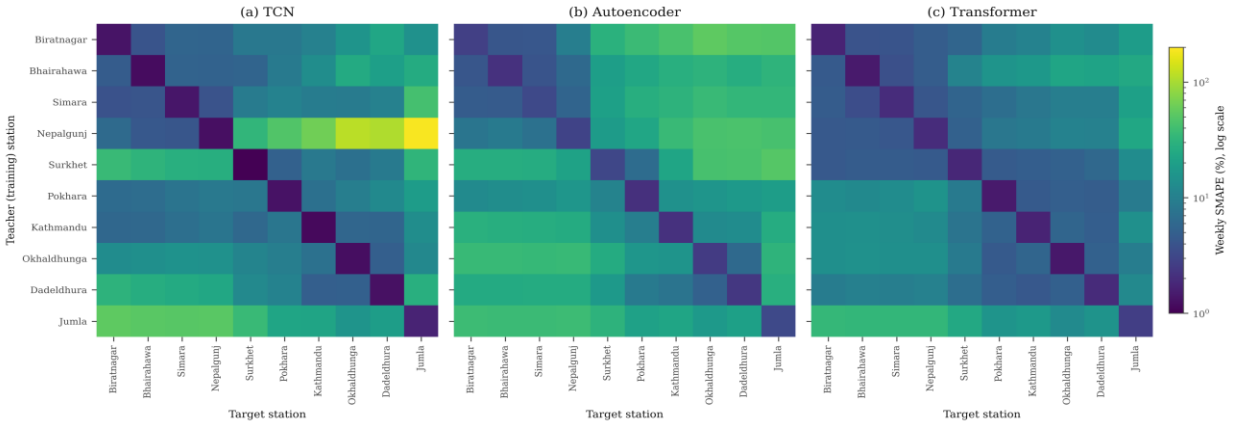


Fig. 2. Single-teacher transfer grids for the three sequence architectures under physically bounded normalization (weekly SMAPE, log color scale). Stations are ordered by elevation. The hot-pole (Nepalgunj) and cold-pole (Jumla) teacher rows fail in the TCN and autoencoder but remain bounded in the transformer.

6. Results II: zero-shot generalization across South Asia

6.1. Headline

Fig. 3 shows the regional evaluation. Every corpus that contains Jumla keeps all ten South Asian cities between 3.1% and 7.0% weekly SMAPE. Even the reduced three-station corpus of Section 5.2 performs well across the subcontinent: Patna 4.60%, Lahore 4.52%, Dhaka 5.21%, Chennai 4.35%, Colombo 4.50%, Malé 5.55%, Karachi 4.59%, Jodhpur 5.76%, Kabul 6.98%, and Thimphu 5.55%. The full ten-station model, which is the reference configuration, improves almost every city (3.13–5.92%) and also covers Seoul outside the region (5.16%, Pearson r 0.94, NMBE −0.3%). Annual errors are mostly below 3% (Table 4), a level of precision that is compatible with common screening and compliance uses.

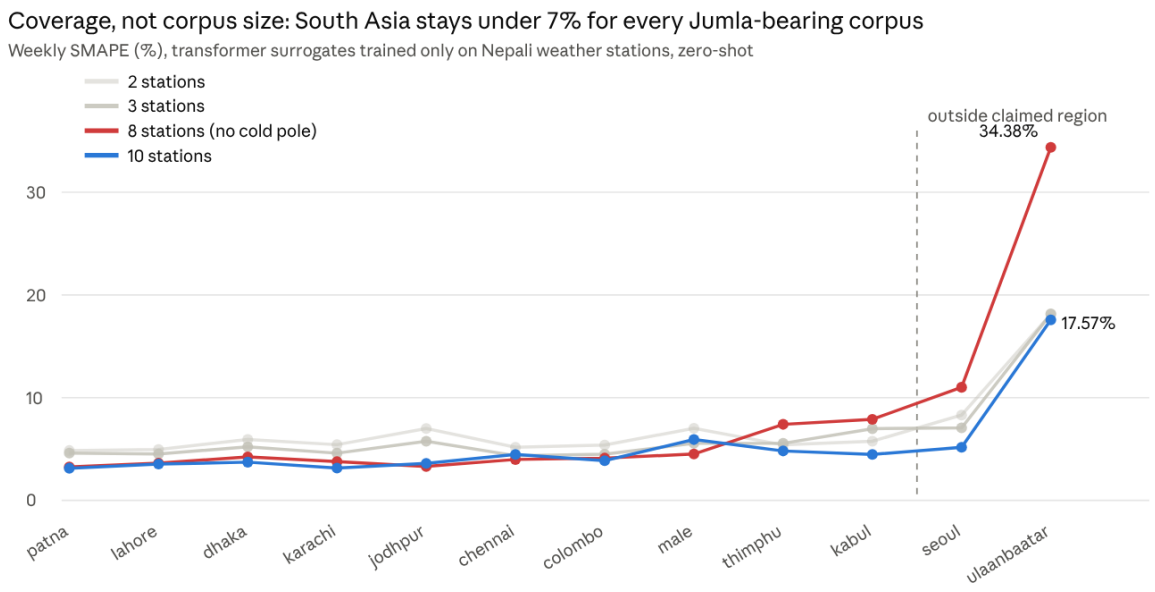


Fig. 3. Zero-shot weekly SMAPE at the twelve evaluation cities for the four Nepali training corpora (single transformer models, calendar-week training). The dashed line separates the ten South Asian cities from the two out-of-region probes.

6.2. Coverage governs transfer

The corpus ablation clarifies what drives these results. The eight-station model without the cold pole is the best model at every warm city, which confirms that more in-regime data helps. Yet it loses to the two-station model at every cold-leaning city (Thimphu 7.40% against 5.40%, Kabul 7.89% against 5.76%, Seoul 11.01% against 8.30%) and nearly doubles the error at Ulaanbaatar (34.38% against 18.12%). Among the corpora that do contain Jumla, on the other hand, the out-of-coverage probe hardly responds to corpus size: 18.12%, 18.15%, and 17.57% for two, three, and ten stations, with NMBE close to −19% in each case. This is the underprediction mechanism anticipated in Section 3.2, in which the peak heating weeks lie beyond every training target. The mean over the ten South Asian cities follows a shallow data-budget curve: 5.69% for two stations, 5.16% for three, and 4.07% for ten. One station of the right regime is worth more than six extra stations of the wrong regime, and the trustworthy domain of a deployment can therefore be read from the regime coverage of its training corpus before any local simulation is run.

6.3. The remaining gap: equatorial seasonality

Energy levels transfer to the deep tropics, where Malé stays at or below 7% under every corpus, but the weekly shape does not. The per-design correlation falls with latitude (Chennai 0.73, Colombo 0.57, Malé 0.09 for the ten-station model), and Malé is the only city where three stations beat ten. Part of this is the metric: when the true weekly variance is close to zero, correlation mostly measures noise, which is why SMAPE and NMBE carry the assessment there. Part of it is a real coverage gap: no Nepali station has cooling demand with almost no seasonality. The planned remedy is to add one tropical station to the corpus, which is listed as follow-up work in Section 8.

Table 4. Ten-station model at the twelve cities (weekly SMAPE is shown in Fig. 3).

City	Annual SMAPE (%)	NMBE (%)	Pearson r	R^2
Patna	1.18	1.19	0.98	0.95
Lahore	0.57	−0.43	0.98	0.96
Dhaka	1.15	1.15	0.94	0.89
Chennai	0.89	0.83	0.73	0.57
Colombo	1.66	1.43	0.57	0.48
Malé	3.21	−3.13	0.09	−0.21
Karachi	0.45	0.27	0.97	0.94
Jodhpur	1.68	−1.65	0.98	0.94
Kabul	1.08	0.42	0.91	0.84
Thimphu	1.09	0.42	0.91	0.85
Seoul	1.74	−0.27	0.94	0.87
Ulaanbaatar	20.41	−18.52	0.93	0.49

6.4. Training-set augmentation and ensembling

Table 5 and Fig. 4 report the ablation described in Section 4.6. On the three-station corpus, sliding windows alone bring the mean weekly SMAPE over the ten South Asian cities down from 5.16% to 4.28–4.62%, depending on the seed, and averaging the three seeds reaches 4.03%. That matches the ten-station calendar-week model (4.07%) with no additional simulation. Within Nepal, the ensemble keeps every unseen station at or below 4.81% weekly SMAPE, with annual errors no larger than 2.02%, and Pokhara, previously the hardest interpolation target, improves from 5.70% to 4.81%. Ten-station sliding runs (three seeds; 0.9–1.7% weekly on held-out designs at every training station) round out the comparison. Their South Asian means of 4.22–4.48% fall inside the single-seed band of the three-station corpus, and the three-seed ensemble reaches 3.96%, within 0.07 percentage points of the three-station ensemble. With augmentation, then, the data-budget curve is flat from three stations to ten. The full gradient still adds depth where its regimes apply: the ten-station calendar-week model and sliding ensemble predict Kabul at 4.47% and 4.94%, against 6.9–7.0% for the three-station counterparts. One ten-station seed reached 6.59%, however, so this gain is reliable only when seeds are averaged.

Outside coverage, and at its edge, the opposite happens. Augmentation degrades performance in every configuration and every seed. Ulaanbaatar worsens from about 18% to 22.5–30.3%, and the underprediction deepens to an NMBE of −24% to −30%. Seoul rises from 5.16% to 7.70–9.61% across the augmented ten-station seeds (8.32% for the ensemble). The ten-station corpora also lose the deep tropics: Malé reaches 5.34% with per-design correlation near zero, against 4.05% and $r = 0.44$ for the three-station ensemble. Augmentation sharpens the model inside its regime envelope and makes it more confidently wrong outside it, whatever the corpus size, which is consistent with the broader finding that data volume cannot substitute for coverage.

Ensemble disagreement gives some warning of this. On the three-station corpus, member spread rises steadily as coverage thins: 1.3–2.6% of the prediction across South Asia, 3.5% at Seoul, and 3.9% at Ulaanbaatar (Fig. 4c), an indicator that requires no simulation. On the ten-station corpus the spread rises only from 1.5–2.4% across South Asia to 2.5% at Seoul and 2.7% at Ulaanbaatar, an order of magnitude

below the error there, so the flag is informative only on the reduced corpus and must accompany, not replace, the envelope checks of Section 7.3. For deployment, the ten-station ensemble is the reference configuration (3.96%, and clearly better at Kabul). The three-station ensemble is a reduced-budget alternative that matches it on average at 30% of the simulation budget and does better in the tropics and at the two out-of-region probes.

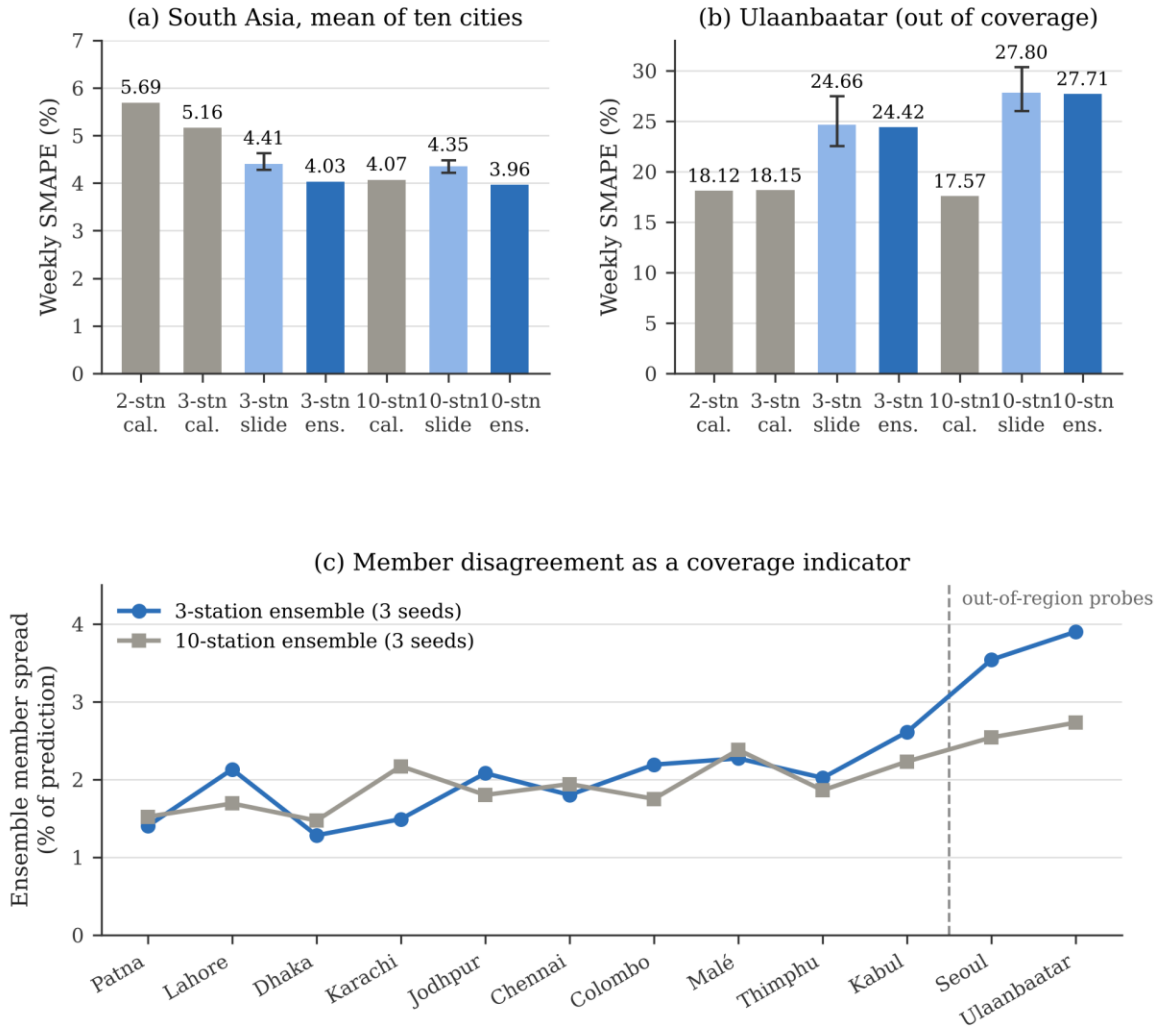


Fig. 4. Sliding-window augmentation and seed ensembling. (a) Mean weekly SMAPE over the ten South Asian cities and (b) at Ulaanbaatar, by training configuration (error bars span individual seeds). (c) Ensemble member spread by city: on the three-station corpus the spread rises monotonically as coverage thins, whereas on the ten-station corpus the spread rises only marginally, far below the error.

Table 5. Augmentation ablation (weekly SMAPE, %). SA mean is over the ten South Asian cities; Nepal max is the worst fully unseen Nepali station.

Configuration	Nepal max	SA mean	Malé	Kabul	Seoul	Ulaanbaatar
3 stations, calendar weeks, single model	5.70	5.16	5.55	6.98	7.06	18.15

Configuration	Nepal max	SA mean	Malé	Kabul	Seoul	Ulaanbaatar
3 stations, sliding, single model (3 seeds)	4.69–6.43	4.28–4.62	4.16–5.18	6.69–7.97	6.85–7.95	22.50–27.46
3 stations, sliding, 3-seed ensemble	4.81	4.03	4.05	6.89	6.71	24.42
10 stations, calendar weeks, single model	n/a	4.07	5.92	4.47	5.16	17.57
10 stations, sliding, single model (3 seeds)	n/a	4.22–4.48	5.28–6.30	4.91–6.59	7.70–9.61	25.97–30.34
10 stations, sliding, 3-seed ensemble	n/a	3.96	5.34	4.94	8.32	27.71

7. Discussion

7.1. What this enables in Nepal

The practical outcome of this study is a procedure that an institution in Kathmandu can carry out on one laptop in a few days. Download the ten public TMYx weather files of the national network, run about 500 background EnergyPlus simulations of the archetype of interest (about 150 if the budget allows only three regime-spanning stations), train one transformer overnight, and the result is a tool that returns simulation-grade weekly energy for that archetype at any Nepali location in under a second. The first applications are the ones that the simulation bottleneck currently rations: screening design spaces to support energy provisions in building regulations, ranking retrofits across municipal building portfolios in climates that were never simulated individually, and teaching in engineering programs where simulation time, not software, is the scarce resource. Because the inputs are a public weather file and fourteen design numbers, the tool also reaches the settings where simulation expertise is thinnest.

7.2. A regional capability from a national network

The zero-shot results show that Nepal's geography is an advantage here. The elevation range acts as a natural laboratory whose ten stations cover most of South Asia's climate regimes, so a surrogate trained in Nepal serves Patna, Lahore, Dhaka, Colombo, Karachi, Kabul, and Thimphu without a single simulation in those countries. For regional bodies and cross-border research groups this has a direct implication. One member's modest simulation campaign can seed a shared prediction capability, and the other members need only validate rather than build corpora of their own. The coverage principle also tells such a group where its next station should go: not toward more of what it already has, but toward the regime it lacks, which for South Asia is one equatorial station (Section 6.3).

7.3. Knowing where to trust the model

A prediction tool meant for practice has to state its own limits. The boundary probes show that those limits can be read in advance from two checks that require no simulation: whether the target's weather falls within the regime envelope of the training corpus, and whether the target's plausible load peaks fall within the training energy range. Ulaanbaatar fails both, and the surrogate fails there in the predicted direction (persistent underprediction) and by 18–30% depending on configuration; neither corpus size nor augmentation changes the character of the failure. Seoul passes the weather check under the full gradient

and is predicted at 5.2–9.6% depending on configuration. The ensemble of Section 4.6 adds a third, model-internal check that is valid on the reduced corpus. On the three-station corpus, disagreement among seed members rises from 1.3–2.6% of the prediction inside South Asia to 3.5% at Seoul and 3.9% at Ulaanbaatar, which flags thinning coverage without any ground truth (Fig. 4c). On the ten-station corpus the spread rises only marginally (to 2.7% at Ulaanbaatar, against an error of 28%), so the spread should always accompany, and never replace, the two envelope checks. Deployments should report all three alongside their predictions, so that the surrogate comes with a stated operating range rather than being used as a black box.

7.4. Methodological observations for similar implementations

Three observations are likely to apply beyond Nepal. First, physically bounded input normalization is a sensible default wherever a surrogate will cross large elevation ranges or be queried zero-shot. It is required for correctness in the second case and improved accuracy in both settings studied here. Second, corpus design should be stratified by regime. The transfer grids show that training on a single extreme produces characteristic, diagnosable failures (extrapolation collapse from hot-only corpora, inversion of seasonal logic from cold-only ones), and that a triad of hot pole, cold pole, and midpoint is the minimum that avoids them, in line with the strategic multi-zone selection examined in [19]. Third, among the reference architectures the transformer is the most robust under distribution shift. It remains well behaved even under single-extreme training, and it is the recommended choice when training data are few, consistent with its strong showing in the Canadian experiments of [19].

7.5. Limitations

Several limitations should be kept in mind. Ground truth is EnergyPlus output under typical meteorological year weather, not metered consumption, and the claims concern the simulated energy of a single archetype, the DOE medium-office prototype. Other typologies, notably Nepal's residential stock, require their own corpora, and the pipeline is released to make that straightforward. One TMY file per station leaves interannual variability and climate trends untested. Hyperparameters were fixed for protocol reproducibility, so the reported accuracy is conservative. Within-corpus diagonal figures reuse the designs that drove early stopping and are mildly optimistic, while every cross-location and zero-shot figure is free of this effect. Correlation metrics are uninformative in climates with little seasonality and are flagged where reported. The training network itself is small, ten stations within one country, and the reduced-corpus results rest on a single station combination each (one two-station and one three-station set), so they indicate the role of regime coverage rather than establish a general rule for station selection. Finally, the training latitude band is narrow (26.5–29.3° N), so day-length regimes far from it enter the evaluation only through the two out-of-region probes.

8. Conclusions

Weekly weather-guided surrogate modeling [19] was implemented for Nepal, and the results show that the country's compressed climate gradient turns a ten-station national weather network into a training corpus sufficient for simulation-grade weekly energy prediction nationwide. Fully held-out stations are predicted at 3.1–8.7% weekly SMAPE, and no worse than 5.7% (4.8% with ensembling) when the network is reduced to three regime-spanning stations. Without further simulation, the same model covers all eight South Asian countries: 3.1–5.9% zero-shot from ten stations, and 4.0% mean with sliding-window training and a three-

seed ensemble, a level the reduced three-station corpus also reaches. Making the framework usable on this geography required one adaptation, which is recommended for future implementations: scaling weather inputs to fixed physical bounds. This is a correctness requirement for zero-shot use and an accuracy improvement for cross-climate transfer. The domain of validity proved predictable in advance from climate-regime coverage. A probe designated in advance as out of coverage failed with the anticipated bias regardless of corpus size and worsened rather than improved under data augmentation; omitting the single cold-regime station degraded every cold-leaning prediction and nearly doubled the boundary error; and on the reduced corpus, ensemble disagreement rose steadily as coverage thinned, giving a warning signal that needs no simulation. The whole capability, including 1,100 simulations and all training and evaluation, fits on one laptop. The code, design matrices, and datasets will be released with a Zenodo DOI upon acceptance. Follow-up work will add one equatorial station to close the remaining gap in the tropics, replicate the annual-input baseline of [19] in this setting to complete the resolution comparison, extend the corpus to residential typologies, and formalize the coverage checks of Section 7.3 into a pre-deployment audit. For Nepal, and for data-scarce regions more generally, the encouraging conclusion is that the barrier to credible, climate-resolved building energy prediction is no longer compute or data volume but weather-station coverage of the climate regimes in which predictions are needed.

Data and code availability

All code (the simulation pipeline, configuration, and verification scripts), the Latin hypercube design matrix, the per-location weekly datasets, trained model checkpoints, and all metric files will be deposited on GitHub with a Zenodo DOI upon acceptance. Weather files are publicly available SRC-TMYx files from climate.onebuilding.org; the DOE medium-office prototype is distributed with EnergyPlus.

CRediT authorship contribution statement

Aashish Katuwal: Conceptualization, Methodology, Software, Investigation, Data curation, Formal analysis, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the author used Claude (Anthropic) to assist with code development, experiment planning, results analysis, and manuscript drafting. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Funding

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

References

[1] United Nations Environment Programme, 2022 Global Status Report for Buildings and Construction: Towards a Zero-emission, Efficient and Resilient Buildings and Construction Sector, UNEP, Nairobi, 2022.

[2] International Energy Agency, The Future of Cooling: Opportunities for Energy-Efficient Air Conditioning, IEA, Paris, 2018.

[3] R. Khosla, N.D. Miranda, P.A. Trotter, A. Mazzone, R. Renaldi, C. McElroy, F. Cohen, A. Jani, R. Perera-Salazar, M. McCulloch, Cooling for sustainable development, Nature Sustainability 4 (2021) 201–208.

[4] D.B. Crawley, L.K. Lawrie, F.C. Winkelmann, W.F. Buhl, Y.J. Huang, C.O. Pedersen, R.K. Strand, R.J. Liesen, D.E. Fisher, M.J. Witte, J. Glazer, EnergyPlus: creating a new-generation building energy simulation program, Energy and Buildings 33 (2001) 319–331.

[5] S. Bodach, W. Lang, J. Hamhaber, Climate responsive building design strategies of vernacular architecture in Nepal, Energy and Buildings 81 (2014) 227–242.

[6] H.-X. Zhao, F. Magoulès, A review on the prediction of building energy consumption, Renewable and Sustainable Energy Reviews 16 (2012) 3586–3592.

[7] K. Amasyali, N.M. El-Gohary, A review of data-driven building energy consumption prediction studies, Renewable and Sustainable Energy Reviews 81 (2018) 1192–1205.

[8] P. Westermann, R. Evins, Surrogate modelling for sustainable building design – a review, Energy and Buildings 198 (2019) 170–186.

[9] N.D. Roman, F. Bre, V.D. Fachinotti, R. Lamberts, Application and characterization of metamodels based on artificial neural networks for building performance simulation: a systematic review, Energy and Buildings 217 (2020) 109972.

[10] T. Østergård, R.L. Jensen, S.E. Maagaard, A comparison of six metamodeling techniques applied to building performance simulations, Applied Energy 211 (2018) 89–103.

[11] L. Van Gelder, P. Das, H. Janssen, S. Roels, Comparative study of metamodelling techniques in building energy simulation: guidelines for practitioners, Simulation Modelling Practice and Theory 49 (2014) 245–257.

[12] J.S. Hygh, J.F. DeCarolis, D.B. Hill, S.R. Ranjithan, Multivariate regression as an energy assessment tool in early building design, Building and Environment 57 (2012) 165–175.

[13] F. Ascione, N. Bianco, C. De Stasio, G.M. Mauro, G.P. Vanoli, Artificial neural networks to predict energy performance and retrofit scenarios for any member of a building category: a novel approach, Energy 118 (2017) 999–1017.

[14] P. Manmatharasan, G. Bitsuamlak, K. Grolinger, AI-driven design optimization for sustainable buildings: a systematic review, Energy and Buildings 332 (2025) 115440.

[15] G. Pinto, Z. Wang, A. Roy, T. Hong, A. Capozzoli, Transfer learning for smart buildings: a critical review of algorithms, applications, and future perspectives, Advances in Applied Energy 5 (2022) 100084.

[16] C. Fan, Y. Sun, K. Xiao, J. Zhao, Statistical investigations of transfer learning-based methodology for short-term building energy predictions, Applied Energy 262 (2020) 114499.

[17] A. Rackes, A.P. Melo, R. Lamberts, Naturally comfortable and sustainable: informed design guidance and performance labeling for passive commercial buildings in hot climates, Applied Energy 174 (2016) 256–274.

[18] P. Westermann, M. Welzel, R. Evins, Using a deep temporal convolutional network as a building energy surrogate model that spans multiple climate zones, Applied Energy 278 (2020) 115563.

[19] P. Manmatharasan, G. Bitsuamlak, K. Grolinger, High-resolution weather-guided surrogate modeling for data-efficient cross-location building energy prediction, Energy and Buildings 359 (2026) 117251.

[20] M. Deru, K. Field, D. Studer, K. Benne, B. Griffith, P. Torcellini, B. Liu, M. Halverson, D. Winiarski, M. Rosenberg, M. Yazdanian, J. Huang, D. Crawley, U.S. Department of Energy Commercial Reference Building Models of the National Building Stock, Technical Report NREL/TP-5500-46861, National Renewable Energy Laboratory, Golden, CO, 2011.

[21] L.K. Lawrie, D.B. Crawley, Development of Global Typical Meteorological Years (TMYx), climate.onebuilding.org, 2022 (accessed October 2026).

[22] D.B. Crawley, Which weather data should you use for energy simulations of commercial buildings?, ASHRAE Transactions 104 (2) (1998) 498–515.

[23] M.D. McKay, R.J. Beckman, W.J. Conover, A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics 21 (1979) 239–245.

[24] S. Philip, eppy: scripting language for EnergyPlus IDF files, https://github.com/santoshphilip/eppy (accessed October 2026).

[25] S. Bai, J.Z. Kolter, V. Koltun, An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, arXiv:1803.01271, 2018.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems 30, 2017, pp. 5998–6008.

[27] B. Lim, S. Zohren, Time-series forecasting with deep learning: a survey, Philosophical Transactions of the Royal Society A 379 (2021) 20200209.

[28] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., PyTorch: an imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32, 2019, pp. 8024–8035.

[29] ASHRAE, Guideline 14-2014: Measurement of Energy, Demand, and Water Savings, American Society of Heating, Refrigerating and Air-Conditioning Engineers, Atlanta, GA, 2014.

[30] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, in: Advances in Neural Information Processing Systems 30, 2017, pp. 6402–6413.