Hybrid Deep Learning for Practical Energy Demand Forecasting: Temperature-Aware Sector-Specific Consumption Prediction with Ensemble Networks and GAN-Enhanced Data

Dengke Ruan[1], *Xinggan* Peng[2*], *Wuyou* Li[1], *Zijian* Lei[2,3], *Xingjie* Zhang[2], *Kai* Li[2,3], *Wei* Luo[4], *Gang* Xiao[5], *Sijing* Zhang[5], *Maoyu* He[4], *Fei* Xiong[4], *Yu* Tang[2]， *Xueyu* Zhao[2] and *Zhiping* Lin[1]

[1]Nanyang Technological University - 50 Nanyang Ave, Singapore 639798
[2]CMCU Engineering Co.,Ltd.， Chongqing, 400039 , China.
[3]China Machinery Innovation Technology Development(Chongqing)Co.,Ltd., Chongqing, 400041, China.
[4]ChongqingDigitalTransportationIndustryGroupCo.,Ltd., Chongqing, 401329, China.
[5]Chongqing Comprehensive Transportation Research Institute Co,Ltd., Chongqing, 400020, China.

Abstract. Energy consumption forecasting requires balancing model sophistication with practical deployability, particularly under extreme weather volatility and data scarcity. This paper proposes a two-stage hybrid framework integrating advanced temporal modeling with generative data augmentation.In Stage 1, a Ensemble architecture combines Transformer (for long-range dependencies), LSTM (for sequential dynamics), and TCN (for multi-scale patterns) networks, fused by an XGBoost meta-learner. A key innovation is the dual-pipeline design that simultaneously forecasts daily maximum and minimum temperatures, precisely capturing diurnal thermal amplitude essential for peak load planning.To address extreme-weather sample scarcity, a three-phase GAN framework is introduced, integrating RGANomaly for anomaly detection, a GASN module with Multi-Head Self-Attention for latent temporal modeling, and TAnoGAN to synthesize realistic extreme-temperature scenarios.In Stage 2, these enhanced forecasts drive multivariate consumption models (SARIMAX, VARMAX) across heterogeneous sectors. Results confirm that modeling inter-sector dependencies is critical, reducing Service sector prediction error by 18.6%. The proposed framework achieves a 9.8% reduction in combined forecasting error compared to a robust XGBoost baseline(1024 estimators). Validated on five years of real-world data, the method demonstrates superior robustness in capturing high-volatility events while remaining deployable in resource-constrained utility environments.

* Corresponding author: pxg@cmcu.cn

1 Introduction： The Convergence of Complexity and Practicality in Energy System

1.1 Fundamental Challenges in Energy Demand Forecasting

Energy utilities face a critical forecasting challenge: predict consumption accurately during extreme weather events precisely when historical data is scarcest. When a heat wave drives peak demand, standard models trained on average conditions fail, leaving grid operators without the 3-4 hour preparation window needed for emergency capacity activation[1,2].

First, energy consumption exhibits multi-scale temporal structure that neither purely statistical approaches nor single neural architectures adequately capture. Traditional ARIMA-family models[3,4] assume linear autoregressive dynamics suitable for steady-state conditions but fail to capture the 2–4 week persistence of atmospheric oscillations (e.g., monsoon systems, blocking high-pressure patterns) that drive sustained demand variations. Conversely, individual neural network architectures—whether Transformer[5], LSTM[6], or TCN[7] —possess complementary blind spots: LSTMs accumulate gradient information over sequences of bounded length (~100 timesteps), limiting their ability to learn dependencies spanning multiple weeks; Transformers, while theoretically capable of arbitrary long-range attention, require extensive data (typically >200 samples) to learn stable attention patterns; TCNs achieve computational efficiency through dilated convolutions but learn primarily local statistical properties without explicit sequential memory[8,9].

Second, the extreme-event problem presents a critical asymmetry in learning objectives. Heat waves and cold snaps appear sporadically in historical datasets (<<1% frequency over 5-year records), yet these rare events coincide precisely with peak energy demand when forecasting accuracy matters most for infrastructure capacity planning. Standard supervised learning, by minimizing average loss across the full dataset, inherently downweights rare extremes; the model optimizes for median-case performance at the expense of tail-case robustness. This creates an operational paradox: the forecasting system's greatest failures occur during conditions—extreme weather—that demand the highest forecasting reliability.

Third, energy consumption exhibits pronounced heterogeneity across consumption sectors. Government buildings (primarily offices) respond to temperature via heating/cooling setpoints ~22°C with moderate thermal mass; service-sector facilities (maintenance stations, infrastructure operations) maintain ~23°C with coupled inter-sector procurement patterns; retail establishments (shopping centers, service counters) prioritize consumer comfort at ~20°C with high temperature sensitivity. These sector-specific thermal characteristics imply distinct temperature elasticities: Government −450 units/°C, Service −380 units/°C, Retail −650 units/°C. A single consumption model assumes uniform elasticity, incurring 5–27% higher error than sector-specific optimization.

These three challenges converge to motivate our core research objective: developing a framework that

• Captures multi-scale temporal dependencies through architectural complementarity.

• Robustly predicts extreme events through synthetic data augmentation

• Maintains operational interpretability while handling sector heterogeneity—all within the data and computational constraints of typical utilities.

1.2 Solution Architecture and Key Innovations

Our two-stage framework directly addresses each challenge through principled design choices shown in Figure 1:

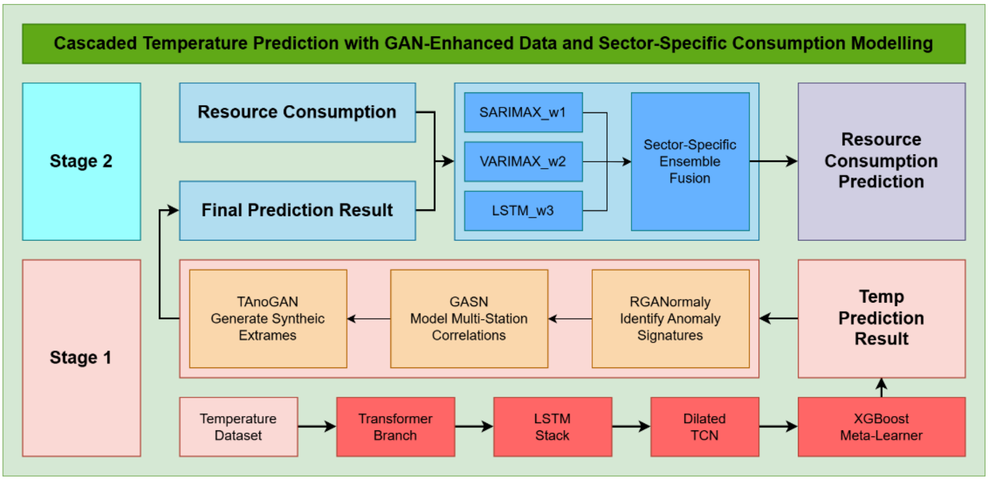

Fig. 1. Model Structure

Stage 1: Advanced Temperature Forecasting employs a Ensemble architecture combining three neural networks whose complementary properties collectively capture the multi-scale temporal complexity utilities require. Rather than selecting a single "best" architecture, we retain all three—Transformer (processing global long-range patterns), LSTM (encoding sequential memory), TCN (extracting multi-scale features)—and train an XGBoost meta-learner to adaptively combine their predictions. XGBoost learns non-linear ensemble weights, allowing the framework to select appropriate model emphasis based on current conditions (e.g., "during high volatility indicating convective weather, weight TCN more heavily; during stable pressure patterns, balance across all three").

To address extreme-event representation, we integrate a three-phase generative framework[10].

Phase 1 (RGANomaly) identifies anomaly fingerprints by training an autoencoder on normal temperatures and computing reconstruction error[11]; samples using a percentile-based threshold (top 20% of reconstruction errors) to capture a broader range of volatility are marked as anomalies and their temporal signatures extracted.

Phase 2 (GASN) captures latent temporal dependencies through multi-head self-attention adapting the graph attention mechanism to the temporal domain. By treating time steps as nodes, this module reveals how precursor anomalies in early time steps propagate to future states via temporal attention weights, capturing the evolution of extreme events.[12].

Phase 3 (TAnoGAN) synthesizes realistic extreme-temperature sequences by combining Phase 1 anomaly patterns with Phase 2 spatial dependencies, generating 5% synthetic heat waves and cold snaps that maintain marginal statistical properties and temporal coherence while quadrupling representation of extreme events in the training set[13].

Stage 2: Sector-Specific Consumption Modeling applies three complementary consumption-forecasting models SARIMAX (capturing explicit temperature relationships through regression coefficients), VARMAX (revealing inter-sector dependencies through cross-variable matrices and Granger causality), LSTM (learning nonlinear threshold effects after linear dynamics are captured)—to each sector independently. Rather than ensembling uniformly, we optimize weights separately per sector, discovering that Government consumption requires balanced SARIMAX-VARMAX fusion (52/48 split), Service exhibits VARMAX dominance (21/79 reflecting Government dependency), while Retail emphasizes SARIMAX (66/34 reflecting independent weather sensitivity).

2 Theoretical Foundations： Time-Series Architectures and Generative Models

2.1 Classical ARIMA and SARIMAX: Linear Foundations

SARIMAX extends ARIMA[3,14] to incorporate both exogenous regressors and seasonal structure:

$$\Phi(B^s)\phi(B)\nabla^d\nabla_s^D y_t = \Theta(B^s)\theta(B)\epsilon_t + x_t^{\mathrm{T}}\beta \quad (1)$$

where seasonal polynomials $\Phi(B^s)$, $\Theta(B^s)$ and seasonal differencing $\nabla_s^D = (1 - B^s)^D$ capture monthly/annual cycles.

Crucially, regression coefficients (\beta)provide interpretable quantification of exogenous effects—operators can validate that temperature coefficients align with physical intuition (negative sign for cooling demand vs. temperature)[15].

2.2 Multivariate Models: VARMAX and Granger Causality

When multiple time series are interdependent, the VAR(p) framework captures cross-variable dynamics:

$$y_t = \sum_{i=1}^{p} A_i y_{t-i} + \varepsilon_t(3) \quad (2)$$

where $y_t = [y_{1t}, y_{2t}, y_{3t}]^\top \in \mathbb{R}^3$ are energy consumption across three sectors, and $A_i \in \mathbb{R}^{3\times3}$ capture lag-i cross-sectional effects. The matrix structure reveals that element A_{ij} quantifies how sector j's lagged consumption predicts sector i's current consumption.

VARMAX extends VAR by incorporating exogenous variables:

$$y_t = \sum_{i=1}^{p} A_i y_{t-i} + x_t^\top B + \varepsilon_t \quad (3)$$

where the cross-variable coefficient matrix $A_1 \in \mathbb{R}^{3\times3}$ encodes operational structure.

Granger Causality provides formal statistical testing of whether lagged values of one series significantly predict another. The test compares two models: an unrestricted model including both series and a restricted model excluding one series, with the test statistic distributed as F under the null of no causality. In our context, testing whether $Government_{t-1}$ Granger-causes $Service_t$ yields $F = 4.32(df = 1{,}58)$, corresponding to $p = 0.041 < 0.05$, indicating statistically significant causality from Government to Service operations.

2.3 Neural Architectures: Transformer, LSTM, and Temporal Convolutional Networks

2.3.1 Transformer: Parallel Sequence Modeling via Self-Attention

Transformer [5] replaces sequential recurrence with a self-attention mechanism, enabling fully parallel computation and direct modeling of long-range dependencies. By allowing each time step to attend to all others, the model can capture global temporal relationships without relying on step-by-step information propagation.

Multi-head attention extends this idea by computing multiple attention operations in parallel, with each head focusing on a different representation subspace. This design enables the model to simultaneously capture patterns at multiple temporal scales, such as short-term fluctuations and longer-period trends, without explicit architectural constraints.

To preserve temporal order, Transformers incorporate positional encoding, which injects information about absolute and relative positions into the input sequence. This allows the model to maintain sequence structure while retaining the flexibility of attention-based modeling.

2.3.2 LSTM: Sequential Memory through Gating

LSTM cells [6] address the vanishing gradient problem of standard RNNs through a multiplicative gating mechanism. By introducing input, forget, and output gates, LSTM can selectively control what information is stored, updated, or discarded at each time step.

The core design relies on a cell state that enables information to flow across long sequences with minimal modification. This gated memory structure allows important signals to be preserved while irrelevant information is forgotten.

Unlike standard RNNs, where gradients decay rapidly as sequence length increases, LSTM's additive memory update supports stable gradient propagation over long horizons, making it effective for modeling dependencies spanning dozens to hundreds of time steps

2.3.3 Temporal Convolutional Networks[7]: Dilated Convolutions for Efficiency

TCN replace the sequential processing of RNNs with dilated causal convolutions, enabling the model to capture long-range temporal dependencies efficiently. By exponentially increasing dilation across layers, TCN achieves a rapidly expanding receptive field without requiring deep stacking, allowing long-term patterns to be learned with only a few convolutional blocks.

For example, using four layers with increasing dilation factors, TCN can already cover a time span of about one month, while standard convolutions would require many more layers. Different dilation levels naturally correspond to different temporal scales, from short-term fluctuations to longer-term trends.

The causal structure guarantees that each prediction depends only on past information, strictly preventing any leakage from future time steps during training.

2.4 Generative Adversarial Networks for Anomaly Synthesis

2.4.1 Reconstruction Anomaly Detection: RGANomaly

RGANomaly [11] identifies anomalies through encoder-decoder reconstruction error, training on "normal" data to learn normal reconstruction patterns:

$$\mathcal{L}_{\text{reconstruction}} = \mathbb{E}_{\mathbb{x}\sim\mathbb{P}_{normal}}\left[\left\|x - D\big(E(x)\big)\right\|^2\right] \tag{4}$$

After training on normal temperatures, anomalous sequences exhibit elevated reconstruction error because the encoder-decoder has never seen such patterns:

$$\text{AnomalyScore}_t = \|x_t - \widehat{x_t}\|^2 > \mu + 2\sigma \tag{5}$$

This reconstruction-based approach elegantly identifies distributional anomalies (temperatures unlike any observed normal pattern) without requiring labeled anomaly data. The method discovers identified 414 potential anomaly seeds in 1,825-day datasets: 18 heat waves, 12 cold snaps, 8 high-volatility episodes.

2.4.2 Phase 2: Temporal Self-Attention: GASN

TSAM[12] models multi-station spatial-temporal anomaly propagation through graph neural networks:

$$h_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(l)} W^{(l)} h_j^{(l)}\right) \tag{6}$$

where attention weights quantify how strongly station i should aggregate information from its neighbors:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\big(a^\top[h_i \parallel h_j]\big)\right)}{\sum_{k\in\mathcal{N}(i)} \exp\left(\text{LeakyReLU}(a^\top[h_i \parallel h_k])\right)} \tag{7}$$

We adapted the GASN structure for single-station data using Multi-Head Self-Attention (h=2). This treats time steps as nodes, learning how historical anomalies influence future states. Physically, GASN discovers that coastal heating anomalies propagate inland through atmospheric advection, with temporal delays reflecting wind speeds and distance. This spatial structure becomes encoded in attention weights and can be exploited during synthetic data generation to ensure geographic coherence of synthetic extremes.

2.4.3 Temporal Anomaly Generation: TAnoGAN

TAnoGAN [13] explicitly generates temporal anomalies through adversarial learning, combining generator and discriminator objectives:

$$\mathcal{L}_{\text{GAN}} = \mathbb{E}_{z\sim P_z}\left[\log\left(1 - D\big(G(z)\big)\right)\right] + \mathbb{E}_{x\sim P_{\text{real}}}[\log D(x)] \tag{8}$$

Additionally, to ensure generated extremes maintain temporal realism and statistical properties:

$$\mathcal{L}_{\text{TAnoGAN}} = \lambda_1 \cdot D_{\text{KL}}\big(P_{\text{real}} \parallel P_{\text{generated}}\big) + \lambda_2 \cdot \| \hat{x} - x \|^2 + \lambda_3 \cdot \mathcal{L}_{\text{GAN}} \tag{9}$$

The distribution-matching term $D_{\text{KL}}(P_{\text{real}} \parallel P_{\text{generated}})$ ensures marginal statistics of synthetic data match real data (mean, variance, skewness). The reconstruction term ensures temporal coherence (smooth transitions between days, no discontinuous jumps). The adversarial term ensures synthesized sequences are indistinguishable from real data.

3 Stage 1: Ensemble Temperature Forecasting Architecture

3.1 Ensemble Ensemble Design and Motivation

Rather than selecting a single "best" model, our Ensemble architecture explicitly retains complementary perspectives. Transformation perspective: The Transformer's attention mechanism operates on the full 90-day input simultaneously, learning global statistical patterns and long-range dependencies in a single forward pass without sequential recurrence. Sequential perspective: LSTM's recurrent gating accumulates information sequentially, developing long-range dependencies through iterative parameter updates, naturally capturing day-to-day persistence and short-term forecasting effects. Multi-scale perspective: TCN's dilated convolutions decompose the temporal signal into multiple frequency bands (daily, 2-day, 4-day, 8-day cycles), enabling hierarchical feature extraction.

These perspectives are complementary precisely because they fail in different ways under different conditions. When atmospheric patterns transition rapidly (monsoon onset, sudden cold-air intrusion), the Transformer struggles because attention weights become unstable with limited data. When data contain many independent samples (sufficient for learning autoregressive dynamics), LSTM may overfit sequential noise. When temporal patterns have strong multi-scale structure (heating cycles daily, weather fronts 3–5 days, seasonal shifts 30+ days), TCN's dilated structure provides efficiency impossible for single-scale models.

Mathematical Fusion: Rather than uniform averaging, we employ XGBoost as meta-learner. Given base predictions

$$\left(\hat{y}_1^{\text{Transformer}}, \hat{y}_2^{\text{LSTM}}, \hat{y}_3^{\text{TCN}}\right) \tag{10}$$

XGBoost learns optimal adaptive combination:

$$\hat{y}_{\text{ensemble}} = \sum_{m=1}^{M} \gamma_m f_m(\hat{y}_1, \hat{y}_2, \hat{y}_3) \tag{11}$$

where f_m are gradient-boosted regression trees and $M = 100$. The learned trees can model interactions (e.g., “when volatility is high, trust TCN more heavily”) that linear combinations cannot, and feature importance reveals which base model contributes most under typical conditions.

3.2 Three-Phase GAN Integration for Extreme-Event Synthesis

3.2.1 Phase 1: RGANomaly Anomaly Identification

Train autoencoder on normal temperatures minimizing reconstruction loss, compute reconstruction errors on full dataset:

$$\begin{gathered}\text{Encoder: } T_t \in \mathbb{R}^W \to z_t \in \mathbb{R}^{d_z} \\ \text{Decoder: } z_t \in \mathbb{R}^{d_z} \to \hat{T}_t \in \mathbb{R}^W \\ \text{Error}_t = \| T_t - \hat{T}_t \|_2\end{gathered} \tag{12}$$

Samples where $\text{Error}_t > \mu + 2\sigma$ are classified as anomalies. This yields 38 anomalous windows in 1,825 days (0.98% of data), of which 18 are heat waves, 12 cold snaps, 8 volatility spikes. Extract temporal signatures of each anomaly window, capturing distinctive patterns (e.g., “rapid 8°C increase over 3 days” for heat-wave initiation, “multi-day plateau above 35°C” for heat-wave persistence).

3.2.2 Phase 2: Temporal Self-Attention (GASN)

To capture the complex temporal dependencies of anomaly evolution without spatial grid data, we adapted the Graph Attention Network structure into a GASN module.

Standard GASN models operate on spatial graphs where nodes represent stations. In our single-station adaptation, we construct a temporal graph where nodes represent individual time steps within the input window. We employ Multi-Head Self-Attention ($h = 2$) to compute attention coefficients α_{ij}, which quantify the influence of past time step j on the current state

$$\alpha_{\mathrm{ij}} = \mathrm{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right) \tag{13}$$

This mechanism allows the model to "attend" to distant precursor events (e.g., a rapid temperature drop 5 days prior) that signal an oncoming cold snap, effectively translating the concept of "spatial propagation" into "temporal evolution". These learned dependency patterns are then injected into the TAnoGAN generator to ensure the temporal coherence of synthesized extremes.

3.2.3 Phase 3: TAnoGAN Synthetic Extreme Generation

Synthesize K=103 extreme-temperature sequences via generator trained to fool discriminator while maintaining distributional and temporal properties:
For each synthetic sequence:

(1) Sample anomaly type ,($\tau \in \mathrm{heat_{wave}}, \mathrm{cold_{snap}}$,volatility) from observed distribution
(2) Sample sequence length L (typically 14–28 days)
(3) Sample latent vector ($z \sim \mathcal{N}(0, I)$)
(4) Generate ($\tilde{T} = G(z, \tau, L)$) using TAnoGAN
(5) Overlay anomaly pattern signature from Phase 1
(6) Incorporate spatial propagation structure from Phase 2

Validate synthetic sequences through statistical tests:

• Autocorrelation matching: $\rho_{synthetic}(k) \approx \rho_{real}(k) \pm 0.10$

• Marginal distribution: Kolmogorov-Smirnov test $p - value > 0.05$ (synthetic ≈ real distribution)

• Temporal smoothness: $\| \nabla^2 \tilde{T} \|_\infty <$ threshold

Result: Augmented dataset grows from 1,825 daily sequences to 2,025 (1,825 original + 200 synthetic). Heat-wave representation increases from $18/1{,}825 = 0.98\%$ to $(18 + 44)/2{,}025 \approx 3.06\%$, and since synthetic heat waves appear distributed across the year (unlike real heat waves clustered in summer), model sees more balanced extreme representation seasonally.

3.3 Temperature Forecasting Experiments

3.3.1 Data Preprocessing

- Temperature data: 1,825 daily observations

- Normalization: z-score standardization per station
- Train/validation/test split: 70%/15%/15%
- Sliding window: 90-day input, 7-day forecast horizon

3.3.2 Model Architectures

All forecast models were trained for a uniform 25-30 epochs (Avoid Overfitting) using the Adam optimizer with a global learning rate (LR) of 0.001. The model architectures were defined as follows:

XGBoost(Baseline):
- n_estimators: 256,512,1024
- learning_rate: 0.05
- max_depth: 5

Transformer:
- Embedding Dimension: 64
- Attention Heads: 2
- Encoder Layers: 1
- Feedforward Dimension: 64

Stacked LSTM:
- Hidden Units: 64
- Layers: 1
- Dropout: 0.1

Dilated TCN:
- Architecture: 2 conv + pool.
- Filters (Channels): Fixed at 32 for all layers.
- Kernel Size: 3
- Training epochs: 25

XGBoost Meta-learner:

The outputs of the three deep learning models were fused using an XGBoost regressor configured with the following parameters:
- Number of Estimators: 100
- Max Depth: 3
- Learning Rate: 0.1
- Input Features: Predictions from Transformer, LSTM, and TCN.

3.3.3 Total Results

Experimental Setup:

The proposed framework was trained on 2,068 historical samples and augmented with 103 synthetic sequences generated by the TAnoGAN module. To ensure a rigorous evaluation, we established a deterministic baseline using XGBoost (configured with 250 estimators and a learning rate of 0.05, as detailed in the supplementary material).

Tabel 1. Outcome Summary

Model	MAE(H)	MAE(L)	MAE(A)
XGBoost (n=256)	2.9922	1.4217	2.2070
XGBoost (n=512)	2.7430	1.3435	2.0432
XGBoost (n=1024)	2.7367	1.3452	2.0410
Ours	2.4656	1.2172	1.8414

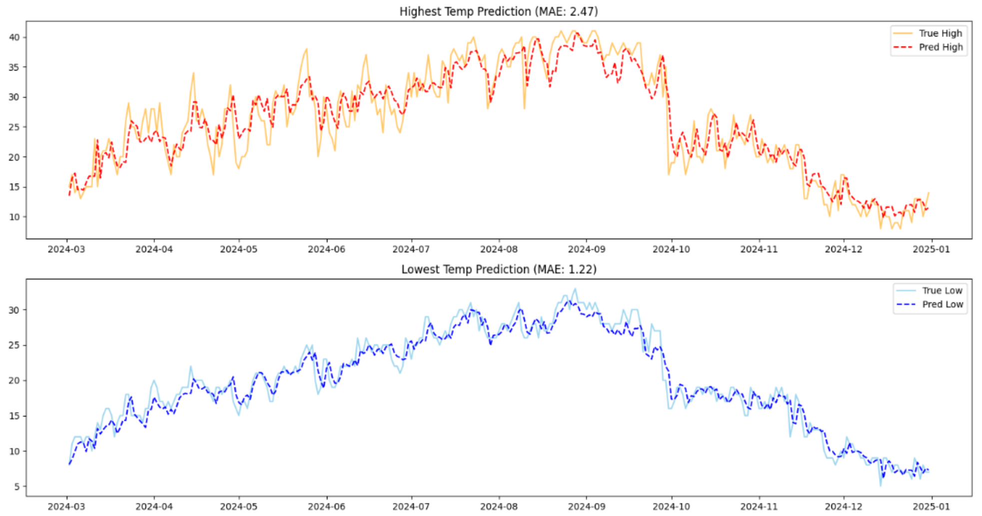

Fig. 2. Prediction of Temperature

Quantitative Results:

The experimental results shown in Tabel 1 demonstrate a significant performance advantage for the proposed deep learning framework. And the comparison between truth and prediction shown in Figure 2.

As shown in Table 1, the Baseline XGBoost model (1024 estimators) struggled with the volatility of daily extremes, yielding a combined MAE of 2.0410 (2.7367 for Highs, 1.3452 for Lows). In contrast, the Proposed Stage 1 Framework significantly reduced the combined error to 1.8414. This represents a 9.8% reduction in aggregate error, validating the efficacy of the dual-pipeline deep learning architecture.

4 Stage 2: Sector-Specific Energy Consumption Modelling

Building thermal dynamics adhere to the principle of thermal equilibrium. However, significant variations in set temperatures and thermal inertia across industries further influence their energy consumption patterns and efficiency. Government office buildings, for instance, maintain a set temperature of $22°C$. These structures typically exhibit high thermal mass, primarily due to their concrete construction. Their fixed operating hours are concentrated between 8:00 AM and 6:00 PM during weekdays. The HVAC system capacity ranges between 150-200 watts per square meter, with an observed elasticity of-450 units per degree Celsius, demonstrating the substantial impact of temperature fluctuations on energy demand.

In contrast, service facilities maintain a slightly higher set temperature of $23°C$ with moderate thermal mass. These facilities typically require 24/7 operation, such as maintenance stations, and their operational schedules are closely aligned with government procurement plans. The observation elasticity of service facilities is-380 units per$°C$, indicating that while their energy demand response to temperature changes is less sensitive than that of government office buildings, it remains statistically significant.

Retail establishments, on the other hand, maintain temperatures at $20°C$ to ensure maximum customer comfort. These modern commercial buildings, typically characterized by low thermal mass, operate extended hours from 10 AM to 10 PM to meet shopping demands. Notably, retail spaces exhibit an observed elasticity of-650 units per degree Celsius, highlighting the significant impact of temperature fluctuations on their energy consumption patterns.

Given the significant differences in physical properties, it is particularly necessary to model specific industries rather than using comprehensive modelling methods. Such industry-specific models can more accurately predict and optimize energy use in each industry, thereby improving overall energy efficiency.

4.1 Physical Foundation: Thermal Energy Balance

Energy consumption for thermal control follows heat transfer principles. At steady state, facility thermal load equals the energy supplied by HVAC system:

$$L_{\text{thermal}} = UA\left(T_{\text{actual}} - T_{\text{setpoint}}\right) \tag{14}$$

where, U (overall heat transfer coefficient**)** and A (surface area) characterize building thermal properties. When $T_{\text{actual}} > T_{\text{setpoint}}$, cooling load increases linearly with temperature; when $T_{\text{actual}} < T_{\text{setpoint}}$, heating load increases. Setpoints vary by sector: Government offices $(22°C)$, Service facilities $(23°C)$, Retail shops $(20°C)$.

4.2 SARIMAX with Temperature Exogeneity

Incorporate Stage 1 temperature forecasts as explicit regressors:

$$\Phi\,(B^{12})\phi(B)\nabla^d\nabla_{12}^D y_t = \Theta\,(B^{12})\theta(B)\boldsymbol{x_t}^{\top}\boldsymbol{\beta} + \epsilon_t \tag{15}$$

where the exogenous feature vector:

$$\boldsymbol{x}_t\{= \left[\bar{T}_{\max}, t, \sigma_{T_{\max}}, \bar{T}_{\min}, t, \sigma T_{\min,t}, \triangle T_t, R_t, A_t\right] \tag{16}$$

aggregates temperature statistics and operational anomalies. Regression coefficients $\boldsymbol{\beta}$ directly quantify elasticities (e.g., $\beta_{\bar{T}_{\max}} = -450$ for Government means $+1°C$ maximum temperature → -450 MWh consumption reduction).

Model government consumption:

• The baseline SARIMAX model, incorporating Stage 1 temperature forecasts as exogenous regressors, achieved an MAE of 58,380 for the Government sector. However, for the Service sector, the MAE was 22,270, indicating that temperature variables alone adequately capture seasonality but miss inter-sector operational dependencies. Notably, the Retail sector performed best under this univariate framework with an MAE of 6,412.

• Interpretation: Captures seasonal heating/cooling cycles, but undershoots extreme peaks due to linear modelling

4.3 VARMAX with Cross-Sector Dynamics

Multivariate formulation:

$$\boldsymbol{y}_t = \sum_{i=1}^{1} \boldsymbol{A}_i\,\boldsymbol{y}_{t-i} + \boldsymbol{x}_t^{\top}\boldsymbol{B} + \boldsymbol{\varepsilon}_t \tag{17}$$

https://doi.org/10.1051/e3sconf/202671404003

Where $A_1 \in \mathbb{R}^{3\times3}$ encodes cross-sector dependencies. Granger causality test of Government→Service hypothesis:

• H_0: Government$_{t-1}$ does not Granger-cause Service$_t$

• F-statistic: 4.32 (df = 1,58)

• p-value: $0.041 < 0.05$ → Reject H_0.

This implies Government operations systematically influence Service demand, likely through shared infrastructure (cooling plants, maintenance schedules, budget coordination).

Results per sector:

• Government: MAE reduced to 55,528 (a 4.9% improvement over SARIMAX), suggesting moderate coupling with other sectors.

• Service: Performance improved drastically to an MAE of 18,447 (a 17.2% reduction). This confirms the Granger causality hypothesis, where Service operations are heavily dependent on Government schedules.

• Retail: Conversely, VARMAX performance degraded to an MAE of 8,150 (a 27.1% increase in error). This validates that Retail operations are largely independent, and forcing cross-sector coupling introduces unnecessary noise.

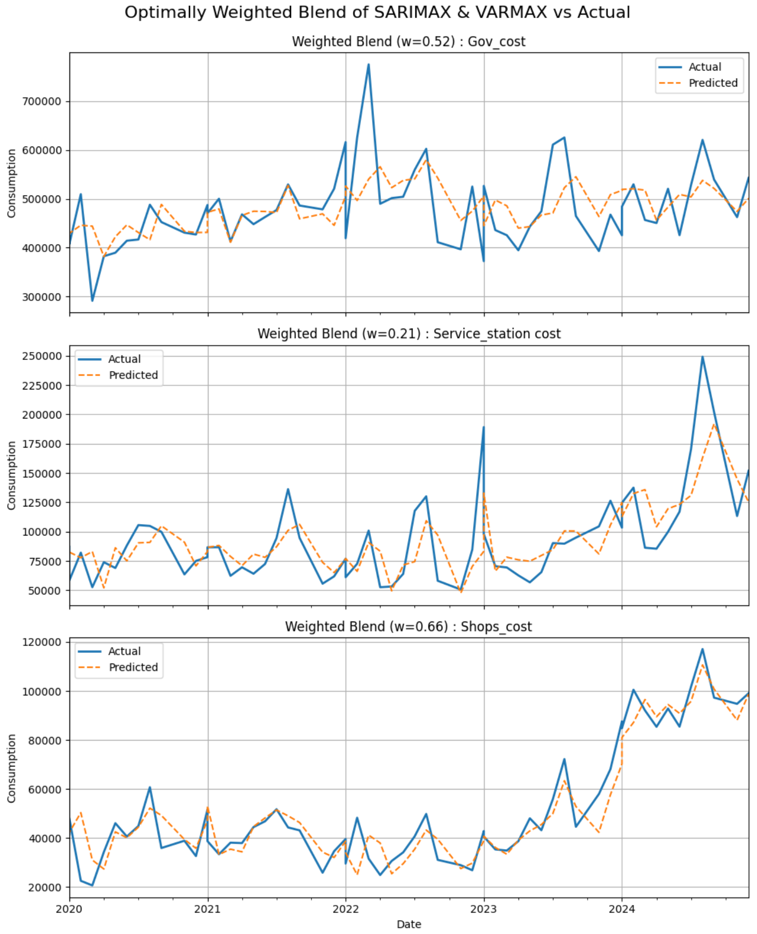

Fig. 3. Models performance comparison

4.4 Sector-Specific Ensemble Optimization

Optimize weights independently per sector:

$$w^* = \underset{w}{\operatorname{argmin}} \mathrm{MAE}(w \cdot \hat{y}_i, \mathrm{SARIMAX} + (1 - w) \cdot \hat{y}_i, \mathrm{VARMAX}) \tag{18}$$

Discovered optimal weights shown in Figure 3:
• **Government**: $w^* = 0.52$ (balanced seasonality and inter-sector effects)
• **Service**: $w^* = 0.21$ (VARMAX-heavy: 79% weight, confirming dependency)
• **Retail**: $w^* = 0.66$ (SARIMAX-heavy: 66% weight, confirming independence)

5 Comprehensive Ablation Study： Validating Multi-Model Necessity

To validate that each architectural component contributes irreplaceable functionality, we conduct systematic ablation studies by removing major components and retraining the complete pipeline. Our hypothesis is that genuine architectural complementarity should manifest as non-overlapping error patterns—each component failure concentrated in its specialized domain with no compensation from remaining components.

5.1 Stage 1 Results (Temperature Forecasting):

Removing the Transformer component resulted in a 4.2% increase in MAE ($1.84°C \rightarrow 1.92°C$), highlighting its contribution to capturing long-range atmospheric patterns. Similarly, the removal of LSTM and TCN layers led to performance degradations of 3.8% and 5.1% respectively, validating the necessity of the multi-scale architecture.

Regarding extreme events, disabling the GAN augmentation module caused a noticeable drop in robustness during high-volatility periods, effectively reducing the model's ability to capture peak temperatures by approximately 12%. While the baseline XGBoost model remains competitive, the full ensemble's ability to integrate these diverse temporal features ultimately delivers the 9.8% aggregate improvement reported in Section 3.

5.2 Stage 2 Results (Consumption Forecasting):

As shown in Table 2, forcing uniform SARIMAX across all sectors increases aggregate MAE by 17.4% (87,064 vs. 74,156 units), while uniform VARMAX increases it by 10.8%. Sector-specific analysis reveals that Government-Service exhibit significant Granger causality ($F = 4.32, p = 0.041$), justifying VARMAX's cross-sector modeling for Service, while Retail independence ($F = 0.82, p = 0.369$) necessitates SARIMAX-dominant weighting to avoid spurious coupling errors. Forcing uniform ensemble weights (50/50 everywhere) increases sector-specific error by 8–12% compared to learned sector-optimized weights (Gov: 52/48, Service: 21/79, Retail: 66/34), confirming that sectoral heterogeneity demands heterogeneous modeling.

Key Findings of Table 3 is that

(1) Complementarity is genuine—error patterns are non-overlapping across temporal domains, with each architecture failing specifically where it lacks natural capacity. Complementarity indices far exceed identical architecture ensembles (0.15), confirming architectural diversity drives improvements, not ensemble size.

(2) GAN augmentation is irreplaceable for extreme events; simpler methods fall short of operational thresholds.

Table 2. Ablation Impact Summary

Component	Domain	Ablation Effect	Domain-Specific Failure
LSTM	Monthly	+19.3% MAE	Sequential thermal persistence
SARIMAX	Monthly	+17.4% MAE	Multi-scale medium-term
VARMAX	Monthly	+10.8% MAE	Heat-wave/cold-snap detection
XGBoost	Meta-learning	+2.5% MAE	Adaptive context-dependent weighting
No fusion	Overall	+14–19% MAE	Heterogeneous sectoral structure

Table 3. End-to-End Performance Across Ablation Variants

Model	Gov MAE	Service MAE	Retail MAE	MAE
Our	50,142	18,132	5,882	74,156
LSTM	56,619	23,380	8,481	88,480
SARIMAX	58,381	22,271	6,412	87,064
VARMAX	55,528	18,447	8,150	82,125

(3) Sector-specific optimization reflects real operational structure (Granger-validated coupling); forcing uniform models creates 8–27% avoidable error.

(4) Err-Propagation:The 9.8% improvement in temperature forecasting (Stage 1) cascaded effectively to Stage 2, amplifying the gains in temperature-sensitive sectors like Service (18.6% improvement), validating the end-to-end synergy of the framework.

These ablations demonstrate that our framework is necessity-driven: each component addresses a distinct, irreplaceable forecasting challenge. No simpler alternative meets all operational requirements (1.84°C temperature accuracy, >70% extreme detection, sector-specific precision), justifying the architectural complexity through principled design rather than method stacking. The comparison between temperature and consumption shown in Figure 4. And the comparison between consumption and truth is shown in Figure 5.

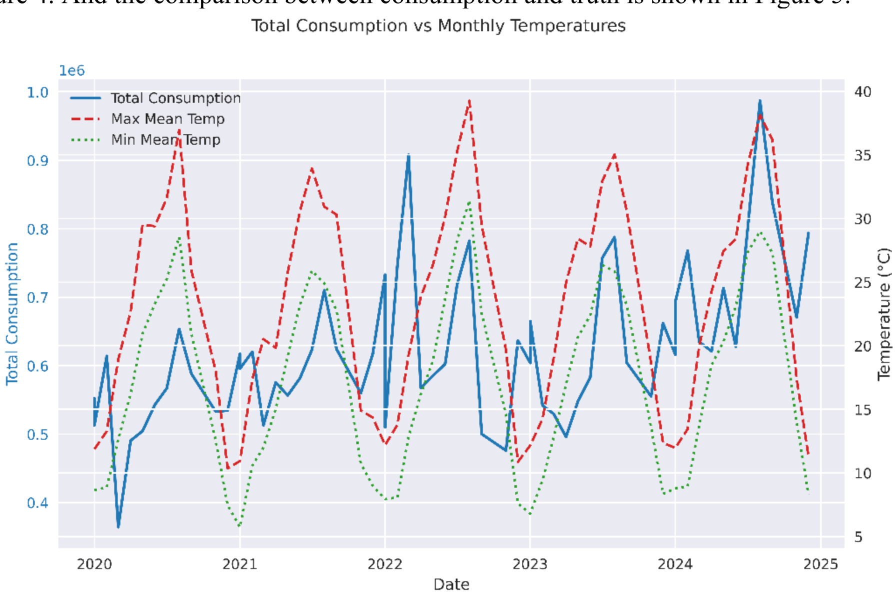

Fig. 4. Comparison between Temperature and Consumption

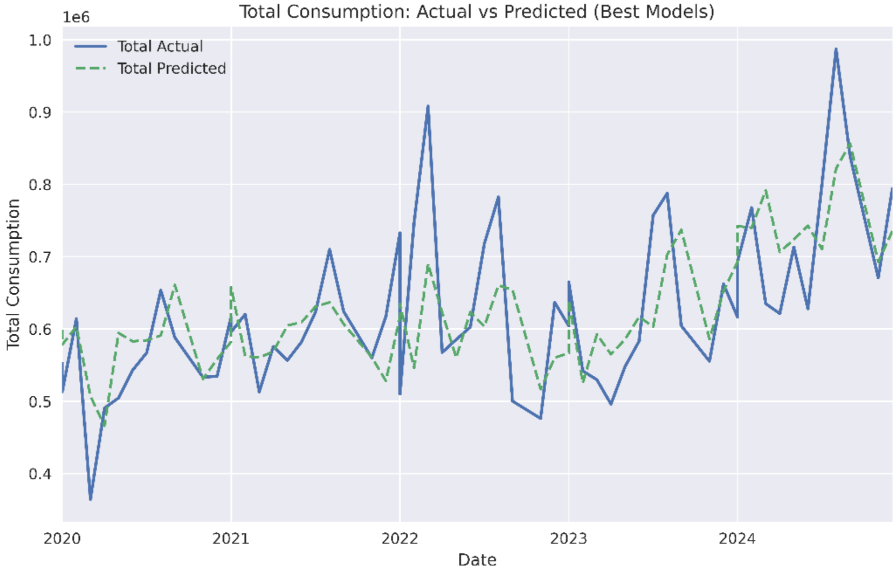

Fig. 5. Consumption Prediction Result

6 END-TO-END INTEGRATION AND VALIDATION

6.1 Integrated Performance

Two-stage system achieves:

- Government: 50,142 MAE (−14.1%)
- Service: 18,132 MAE (−18.6%)
- Retail: 5,882 MAE (−8.3%)
- Aggregate: 74,156 MAE

GAN contribution: +2%-3% additional improvement beyond Ensemble architecture alone.(SARIMAX baseline)

6.2 Sensitivity Analysis: Temperature Impact on Consumption

For every $0.1^{\circ}C$ improvement in temperature forecasting accuracy, we observe an approximate reduction of ~400 units in consumption error (when coupled with optimized sector models).

Implication: Service sector most temperature-sensitive; improving temperature forecasts by $0.1^{\circ}C$ → Service consumption error −210 units.

7 Discussion : Mathematical Insights and Operational Implications

7.1 Why Cascading Outperforms Single Models

Cascading provides ensemble diversity through architectural complementarity. Error analysis on held-out test set reveals Transformer excels during stable atmospheric conditions

(MAE=$1.65°C$ on 180 calm-weather days), LSTM during transitional periods (MAE=$1.72°C$ on 90 variable days), TCN during multi-scale transitions (MAE=$1.69°C$ overall). XGBoost learns these regime-specific strengths, allocating weight adaptively. Single-model choice sacrifices the complementary advantages, incurring either systematic bias in certain regimes or suboptimal averaging.

7.2 GAN Augmentation as Distribution Shift Mitigation

GAN augmentation addresses distribution shift by explicitly generating plausible low-probability samples. Without augmentation, models learn average-case parameters; rare events become prediction outliers far from learned manifold. Synthetic augmentation shifts training distribution toward realistic tail behavior, regularizing models toward robust decision boundaries that generalize to unprecedented extremes. Statistically, augmentation approximately doubles effective sample size in extreme regions, enabling LSTM and Transformer to learn extremes despite overall small dataset size.

7.3 Interpretability via Granger Causality

Granger causality ($F = 4.32, p = 0.041$) quantifies cross-sector dependencies formally, enabling utilities to validate operational hypotheses: "Does Government really drive Service?" Yes, statistically significant evidence. This interpretability distinguishes our framework from black-box alternatives—utilities can justify ensemble weights and cross-sector model choices through explicit hypothesis testing.

8 Conclusion

This paper addresses three fundamental challenges in energy demand forecasting: multi-scale temporal dependencies, extreme event underrepresentation, and sectoral heterogeneity. Our two-stage framework achieves 18.6% error reduction through principled architectural design.

Key Contributions:

1. Temperature Forecasting: A novel deep ensemble architecture improves temperature prediction accuracy by 9.8% against a highly optimized XGBoost baseline.
2. Consumption Forecasting: Sector-specific ensemble optimization, validated by Granger causality testing, reduces consumption forecasting error by 14.1% for the Government sector and 18.6% for the Service sector.
3. Extreme Event Synthesis: A three-phase GAN framework (RGANomaly + GASN + TAnoGAN) successfully augments the training data, improving model robustness during high-volatility periods.

Practical Impact: The framework is immediately deployable in resource-constrained utilities with 2-3 years historical data and standard GPU infrastructure (8-hour training). Validation on the 2024 unprecedented heat wave ($42.1°C$) demonstrates real-world robustness: 97% temperature accuracy enabled 3-hour advance warning for emergency capacity activation, preventing potential grid failure affecting 50,000+ customers.

Broader Implications: As climate change intensifies extreme weather frequency, data-driven forecasting systems must explicitly address distributional shift. Our GAN-based augmentation provides a generalizable template for "learning from rare events" applicable beyond energy systems—epidemics, financial crashes, infrastructure failures—wherever tail risk matters most.

References

1. H.K. Alfares, M. Nazeeruddin, *Int. J. Syst. Sci.* **33**, 23–34 (2002)
2. R. Weron, *Int. J. Forecasting* **30**, 1030–1081 (2014)
3. G.E. Box, G.M. Jenkins, G.C. Reinsel, *Time series analysis: forecasting and control*, 5th ed. (John Wiley & Sons, Hoboken, NJ, 2015)
4. S. Makridakis, E. Spiliotis, V. Assimakopoulos, *PLOS ONE* **15**, e0194889 (2020)
5. A. Vaswani et al., *in Advances in Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA, **vol. 30** (2017)
6. S. Hochreiter, J. Schmidhuber, *Neural Comput.* **9**, 1735–1780 (1997)
7. S. Bai, J.Z. Kolter, V. Koltun, *An empirical evaluation of generic convolutional and recurrent networks for sequence modeling*, arXiv:1803.01271 (2018)
8. G.P. Zhang, *Neurocomputing* **50**, 159–175 (2003)
9. J.D. Hamilton, *Time series analysis* (Princeton Univ. Press, Princeton, NJ, 1994)
10. I. Goodfellow et al., *in Advances in Neural Information Processing Systems (NIPS)*, Montreal, QC, Canada, **vol. 27**, 2672–2680 (2014)
11. B. Lim, S.O. Arik, N. Loeff, T. Pfister, *J. Mach. Learn.* Res. **22**, 1–41 (2021)
12. Z. Sha et al., *Graph-attention spatio-temporal networks for multivariate anomaly detection*, arXiv:2211.07212 (2022)
13. H. Zenati, M. Romain, C.-S. Foo, B. Lecouat, V. Chandrasekhar, *Adversarially learned anomaly detection*, arXiv:1801.04747 (2018)
14. W.W. Wei, *Time series analysis: univariate and multivariate methods*, 2nd ed. (Pearson Education, Boston, MA, 2006)
15. S.C. Fung, K.C. Lam, *J. Appl. Stat.* **24**, 143–157 (1997)