# GridFusionX: Network-Aware Probabilistic Forecasting for Multi-Regional Power Systems

Quoc Bao Phan, *Graduate Student Member, IEEE*, Abdulrahman Takiddin, *Member, IEEE*, Gelli Ravikumar, *Senior Member, IEEE*, Olugbenga Moses Anubi, *Senior Member, IEEE*, and Tuy Tan Nguyen, *Senior Member, IEEE*

**Abstract**—Smart grid networks exhibit complex spatial–temporal dependencies where regional nodes are interconnected through electrical transmission, economic coupling, and shared meteorological patterns. Traditional forecasting methods model regions independently, neglecting network topology and spatial correlations that govern system-level behavior. Graph neural networks capture spatial structure but provide only deterministic forecasts, while probabilistic transformers quantify uncertainty yet ignore network topology, preventing risk-aware coordination across interconnected systems. This paper presents GridFusionX, a multimodal transformer model that resolves this tension through theoretically grounded uncertainty quantification in graph-structured systems. We introduce dual-head transformer encoders with asymptotic calibration guarantees, precision-weighted fusion that adapts to spatially varying data reliability, and tight bounds for uncertainty propagation across networked sequences. Unlike prior approaches that sacrifice either spatial awareness or probabilistic rigor, GridFusionX achieves both through convergence-rate guarantees for network-aware uncertainty estimation, optimal precision weighting under modality independence conditions, and calibrated confidence intervals for operational decision-making. Experiments on ten interconnected European regions demonstrate 4.8–55.9% accuracy improvements and 39.5–66.1% reductions in reserve sizing compared to deterministic graph and topology-agnostic probabilistic baselines, achieving 98.4±0.2% reliability with 90.0±1.6% prediction interval coverage. GridFusionX establishes a theoretically grounded framework for probabilistic forecasting in networked infrastructure.

**Index Terms**—Multimodal learning, probabilistic forecasting, smart grid management, transformer networks, uncertainty quantification.

## 1 Introduction

Modern smart grid networks exhibit complex spatial-temporal dependencies where regional nodes are interconnected through electrical transmission lines, synchronized market mechanisms, and shared meteorological patterns. As renewable energy sources proliferate across distributed locations, these networks face increasing operational complexity that challenges traditional centralized forecasting approaches. Network-aware load forecasting becomes critical for coordinating distributed operations and maintaining system stability across interconnected regions [1].

Traditional statistical methods, such as autoregressive integrated moving average (ARIMA) [2] and support vector regression, treat individual regions independently, ignoring fundamental network topology and spatial correlations. While deep learning approaches, including diffusion convolutional recurrent neural networks (DCRNNs) [3], long short-term memory networks (LSTM) [4], and temporal convolutional networks (TCNs) [5] have improved temporal modeling, they fail to capture spatial dependencies inherent in networked power systems where load patterns propagate across interconnected regions.

Transformer-based architectures have shown promise for modeling long-range dependencies through self-attention mechanisms [6]. Recent variants, including Patchformer [7], Fredformer [8], and Informer [9] demonstrate improved forecasting performance. However, these approaches primarily focus on temporal patterns within individual regions, lacking explicit modeling of spatial correlations and inter-regional dependencies that characterize networked infrastructure systems.

Multimodal learning offers opportunities to integrate heterogeneous data streams across network nodes, including load patterns, renewable generation, and market prices [10]. However, existing multimodal approaches employ uniform fusion strategies that fail to account for varying reliability levels across data sources and spatial locations [11]. This limitation becomes critical in networked systems where information quality varies spatially and temporally, requiring adaptive weighting mechanisms.

Current approaches face two fundamental limitations for network applications: (1) lack of spatial dependency modeling across interconnected regions, and (2) absence of principled uncertainty quantification essential for risk assessment in distributed systems. These gaps limit operational effectiveness in modern smart grids where decisions must account for network-wide implications and spatial

*This research was supported by the Department of Electrical and Computer Engineering, Center for Advanced Power Systems, FAMU-FSU College of Engineering, Florida State University. (Corresponding author: Tuy Tan Nguyen.)*
*The authors are with the Department of Electrical and Computer Engineering, Center for Advanced Power Systems, FAMU-FSU College of Engineering, Florida State University, Tallahassee, FL 32310, USA. (e-mail: qp25c@fsu.edu, a.takiddin@fsu.edu, rgelli@fsu.edu, oanubi@fsu.edu, tuy.nguyen@fsu.edu).*
*Manuscript received MM, DD, YYYY.*

uncertainty propagation.

To address these challenges, we introduce GridFusionX, a network-aware multimodal learning framework that leverages the topology and coupling structure of interconnected power grids to provide calibrated uncertainty estimates. GridFusionX models the power system as a graph of spatially coupled regions, explicitly propagating both information and uncertainty across this graph to enable risk-aware, system-level decision making.

Our main contributions are:

- We design network-aware dual-head Transformer encoders that jointly learn topology-constrained representations and probabilistic confidence measures. Unlike existing graph neural networks that produce only point forecasts, our architecture provides provable calibration guarantees with convergence rate $O(N^{-1/4}\sqrt{\log N})$, enabling risk-aware decision-making across interconnected regions.
- We develop a theoretically justified precision-weighted cross-modal fusion mechanism that adapts to spatial reliability variations across heterogeneous modalities. While prior multimodal approaches assume uniform data quality, our method automatically down-weights unreliable sources during sensor degradation, maintaining forecast integrity through inverse-variance weighting with proven independence properties.
- We introduce uncertainty-aware temporal pooling grounded in the law of total variance with tight theoretical bounds for uncertainty propagation across graph-structured sequences. This addresses a fundamental gap in spatial-temporal forecasting where prior work either models spatial dependencies without uncertainty or quantifies uncertainty without spatial awareness.
- We validate the framework on ten interconnected European regions, demonstrating not only 4.8–55.9% accuracy improvements but also 39.5–66.1% operational cost reductions in reserve sizing through calibrated uncertainty estimates. This operational value is unachievable with existing deterministic graph methods or topology-agnostic probabilistic approaches.

The remainder of this paper is organized as follows. Section 2 reviews related work in forecasting problems and uncertainty quantification. Section 3 provides background on spatial-temporal modeling. Section 4 presents the GridFusionX framework. Section 5 describes the dataset and evaluation metrics. Section 6 discusses experimental results. Section 7 concludes with findings and future work.

## 2 Related Work

Accurate electrical load forecasting enables utilities to balance supply and demand, optimize generation scheduling, and maintain grid stability. As power systems integrate renewable energy and distributed generation, the forecasting problem has grown substantially more complex. Modern smart grids exhibit intricate spatial-temporal dependencies where regional nodes are interconnected through transmission lines, synchronized markets, and shared weather patterns. Traditional approaches treating regions independently fail to capture these network-wide dynamics, motivating spatially-aware probabilistic methods that quantify uncertainty across interconnected infrastructure.

### 2.1 From Sequence Models to Transformer Architectures

Early load forecasting relied heavily on statistical time series methods such as ARIMA [12] and support vector regression [13], which model temporal patterns through explicit mathematical formulations. While computationally efficient, these approaches struggle with nonlinear dynamics and cannot adapt to complex multivariate relationships. The rise of deep learning introduced RNNs and LSTM [14], [15] that automatically learn temporal dependencies from historical data. The work in [14] demonstrated that LSTM-based models significantly outperform classical statistical methods by capturing long-range temporal correlations without manual feature engineering. Research in [16] further optimized LSTM architectures using genetic algorithms, achieving improved accuracy on benchmark datasets. However, recurrent architectures process sequences step-by-step, limiting their ability to model very long-term dependencies due to vanishing gradient problems.

The introduction of transformer architectures revolutionized sequence modeling by replacing recurrence with self-attention mechanisms [17], enabling parallel processing and direct modeling of dependencies across arbitrary time distances. The Informer model in [18] adapted transformers for time series forecasting, addressing memory constraints through sparse self-attention. Recent work in [19] applied temporal convolutional networks with attention mechanisms, demonstrating improved accuracy over standard LSTM approaches. A comparative study in [20] showed that attention mechanisms effectively capture both short-term fluctuations and long-term trends in electricity demand patterns.

Despite these advances, most transformer-based forecasting models produce only point predictions without quantifying uncertainty. In grid operations, operators require not just expected load values but also confidence intervals to assess operational risk and allocate reserves appropriately.

### 2.2 Probabilistic Forecasting with Transformers

Recognizing the need for uncertainty quantification, researchers have extended transformer architectures to generate probabilistic forecasts. Authors in [21] propose a framework with adaptive online learning, employing a probabilistic decoder that quantifies forecast uncertainty through both parametric and nonparametric approaches. Experimental validation on electricity demand data demonstrates superior accuracy in both deterministic and probabilistic scenarios. However, the approach assumes temporal independence and does not explicitly model spatial correlations across interconnected regions.

An alternative strategy employs quantile regression to directly estimate prediction intervals without distributional

assumptions. The work in [22] introduces QR-PatchTST, combining the Patch Time Series Transformer with quantile regression for probabilistic multi-energy load forecasting. The model segments multivariate time series into patches while capturing global temporal dependencies through self-attention, achieving a weighted mean absolute percentage error of 2.29 percent representing a 34.9 to 46.6 percent reduction compared to benchmarks. Nevertheless, the method treats different energy modalities with uniform reliability.

Research in [23] develops a penalized temporal fusion transformer for probabilistic electricity price forecasting, integrating LASSO-based expert point forecasts with smoothly clipped absolute deviation regularization. The study in [24] proposes DiffLoad, a diffusion-based sequence-to-sequence structure employing robust additive Cauchy distribution for aleatoric uncertainty estimation. The method separates epistemic and aleatoric uncertainties, demonstrating the ability to model both types for different load levels. While innovative, diffusion-based approaches incur substantial computational costs. These probabilistic transformer approaches share a critical limitation: they ignore network topology and physical connectivity between nodes, treating spatial correlations implicitly through attention mechanisms rather than explicit graph-structured representations.

### 2.3 Graph Neural Networks for Spatial-Temporal Forecasting

Recognizing that power grids inherently form network structures, researchers have adopted graph neural networks to explicitly model spatial dependencies through topology-aware message passing. The study in [25] provides a comparative evaluation of graph neural network architectures for spatiotemporal photovoltaic forecasting, showing that graph convolutional networks, graph attention networks, substantially improve accuracy over topology-agnostic models by explicitly encoding spatial relationships.

Authors in [26] propose DEST-GNN, a double-explored spatio-temporal graph neural network for multi-site intra-hour photovoltaic forecasting. By adaptively learning adjacency matrices that reflect time-varying correlations induced by cloud motion and weather dynamics, and pruning weak connections through sparse spatio-temporal attention, the model achieves strong performance across multiple forecasting horizons. Similarly, the work in [27] develops a spatio-temporal graph neural network with Fourier features, constructing a hyper-variable graph that jointly models generation and numerical weather prediction data while capturing periodic patterns. Beyond photovoltaic forecasting, graph-based approaches have been extended to load prediction. The spatial–temporal embedding graph neural network in [28] constructs directed dynamic graphs with trainable temporal embeddings to capture periodicity, outperforming state-of-the-art baselines on real-world datasets.

Despite these advances, existing graph-based methods share a fundamental limitation: they produce only point forecasts without uncertainty quantification. In operational power systems, probabilistic forecasts are essential for risk assessment, reserve sizing, and coordinated network-wide decision-making. Without calibrated uncertainty estimates, operators must choose between costly over-provisioning and exposure to potential shortfalls.

### 2.4 Multimodal Learning and Operational Integration

Power demand emerges from complex interactions among meteorological conditions, temporal patterns, market prices, and renewable generation. Multimodal learning has therefore been explored to fuse heterogeneous data sources. In [29], a transformer-based framework integrates historical time-series data with ground-based sky images for ultra-short-term solar irradiance forecasting, using cross-modality attention to couple temporal dynamics with spatial cloud movements and achieving mean absolute percentage errors below 5% for 15-minute-ahead predictions.

Despite their promise, most multimodal approaches implicitly assume uniform reliability across data sources, which is unrealistic in distributed power systems with spatially varying data quality. The work in [30] proposes an empirical mode decomposition-based CNN–LSTM framework that fuses load and price data for short-term forecasting, demonstrating improved feature extraction and predictive accuracy on Singapore electricity market data. However, the fusion strategy remains fixed and does not adapt to changing sensor reliability or degraded operating conditions.

Several studies link probabilistic forecasting to operational decision-making. In [31], machine learning-based probabilistic forecasts are combined with robust microgrid scheduling, yielding substantial cost reductions compared to greedy strategies, while [32] surveys economic model predictive control for real-time microgrid optimization. Similarly, [33] integrates RNN-based forecasting with reserve management to reduce operating costs in stand-alone microgrids. Nevertheless, these approaches largely treat forecasting and optimization as loosely coupled stages and lack end-to-end mechanisms for adaptive multimodal fusion under spatially heterogeneous uncertainty.

### 2.5 Research Gaps

The literature reveals a progression from deterministic sequence models to probabilistic transformers, graph-based spatial methods, and multimodal frameworks. This evolution exposes a fundamental gap: approaches emphasizing spatial modeling neglect uncertainty quantification, while probabilistic methods ignore network topology. Such a trade-off is untenable for networked power systems that require simultaneous spatial coordination and risk assessment. Graph neural networks [25], [26], [27], [28], [34] capture spatial dependencies via message passing but yield only deterministic forecasts, preventing risk-aware reserve planning. Probabilistic transformers [21], [22], [23], [24] quantify uncertainty yet treat regions independently, overlooking spatial correlations that drive cascading failures. Multimodal frameworks [29], [30] combine heterogeneous data sources but rely on fixed fusion weights, degrading performance when sensor reliability varies across space or operating conditions. We resolve this trilemma by unifying graph-structured spatial modeling with distribution-free uncertainty quantification and adaptive multimodal fusion, enabling topology-aware probabilistic forecasts robust to incomplete sensing.

## 3 BACKGROUND

This section establishes the theoretical foundations for modeling smart grid networks as graph-structured systems and introduces multimodal transformer architectures for spatially aware forecasting. We formalize the network topology representation and describe how spatial dependencies shape load dynamics across interconnected regions.

### 3.1 Graph-Theoretic Network Modeling

Smart grid networks are cyber–physical systems in which distributed energy resources induce bidirectional power flows across interconnected regions. We model the grid as an undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$, where $\mathcal{V} = \{v_1, \ldots, v_R\}$ denotes $R$ regional nodes, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ represents transmission connections, and $\mathbf{A} \in \mathbb{R}^{R \times R}$ is a weighted adjacency matrix encoding regional coupling.

The coupling strength between regions $i$ and $j$ aggregates multiple connectivity layers:

$$\mathbf{A}_{ij} = \alpha_1 \mathbf{A}_{ij}^{\text{elec}} + \alpha_2 \mathbf{A}_{ij}^{\text{geo}} + \alpha_3 \mathbf{A}_{ij}^{\text{econ}}, \tag{1}$$

where $\mathbf{A}_{ij}^{\text{elec}} = 1$ if regions share transmission lines, $\mathbf{A}_{ij}^{\text{geo}} = \exp(-d_{ij}/\sigma)$ models geographic proximity, and $\mathbf{A}_{ij}^{\text{econ}}$ captures market coupling via price correlations. Coefficients $\alpha_1, \alpha_2, \alpha_3 \geq 0$ control the contribution of each layer.

The objective is topology-aware multimodal forecasting of future electricity demand $\hat{\mathbf{Y}} \in \mathbb{R}^{t \times R}$, given historical observations $\mathbf{X} \in \mathbb{R}^{T \times D}$, where $T$ is the lookback window and $D$ denotes multimodal variables such as load, renewable generation, and prices. The forecasting model $\mathcal{F}_\theta$ must jointly learn temporal dynamics, spatial dependencies defined by $\mathbf{A}$, and cross-modal interactions to minimize network-wide error [35].

Region-wise statistical models ignore $\mathbf{A}$ and fail to capture spatial effects arising from electrical coupling, correlated markets, and weather propagation. Deep learning methods are therefore required to model these coupled temporal–spatial–modal dependencies in networked power systems [36], [37].

### 3.2 Multimodal Transformer Architectures for Spatial Dependencies

Transformers model sequential data using self-attention to capture long-range temporal dependencies through parallel computation [6]. In networked forecasting, transformer encoders map multivariate input sequences into contextualized representations via stacked layers of multi-head self-attention and feed-forward networks. Self-attention projects inputs into queries, keys, and values, computing relevance scores through scaled dot products followed by softmax normalization [38]. Multi-head attention executes this process in parallel, allowing different heads to focus on distinct temporal and spatial patterns across regions.

Smart grid systems generate heterogeneous data streams in which each modality provides complementary information [39]. Load captures demand dynamics with strong temporal structure, renewable generation reflects weather-driven variability, and market prices encode economic interactions constrained by network physics. Multimodal learning exploits these complementary signals to improve forecasting robustness. Modality-specific encoders $f_m$ transform inputs $\mathbf{X}_m \in \mathbb{R}^{T \times D}$ into latent representations $\mathbf{h}_m$, which are combined through a fusion function $F(\cdot)$ to produce unified features $\mathbf{z} = F(\mathbf{h}_1, \ldots, \mathbf{h}_M)$. Simple concatenation or averaging is insufficient in networked settings; attention-based fusion enables adaptive modality weighting conditioned on spatial reliability and network topology $\mathbf{A}$.

**Algorithm 1:** Uncertainty-aware multimodal transformer for $t$-hour ahead probabilistic forecasting

**Input:** Historical multimodal time-series input $\mathcal{X} \in \mathbb{R}^{B \times M \times T \times R}$,
**Output:** Predicted mean $\hat{\boldsymbol{\mu}} \in \mathbb{R}^{B \times t \times R}$ and uncertainty $\hat{\boldsymbol{\sigma}} \in \mathbb{R}^{B \times t \times R}$

`/* Uncertainty-Aware Modality-Specific Encoding */`
**foreach** *modality* $m = 1$ *to* $M$ **do**
  Extract $m^{\text{th}}$ modality: $\mathcal{X}_m \leftarrow \mathcal{X}[:, m] \in \mathbb{R}^{B \times T \times R}$
  Project to hidden space: $\mathcal{H}_m \leftarrow \text{LinearProj}(\mathcal{X}_m)$
  Add positional encoding: $\mathcal{H}_m \leftarrow \mathcal{H}_m + \text{PE}(T)$
  Encode with dual-head Transformer: $\mathcal{Z}_m^{(\mu)}, \mathcal{Z}_m^{(\sigma)} \leftarrow \text{DualTransformerEncoder}(\mathcal{H}_m)$
  `/* Uncertainty-aware temporal pooling */`
  Pool for mean: $\bar{\boldsymbol{\mu}}_m \leftarrow \text{AttentionPool}_\mu(\mathcal{Z}_m^{(\mu)}) \in \mathbb{R}^{B \times H}$
  Pool for uncertainty: $\bar{\boldsymbol{\sigma}}_m \leftarrow \text{AttentionPool}_\sigma(\mathcal{Z}_m^{(\sigma)}) \in \mathbb{R}^{B \times H}$
**end foreach**
`/* Precision-Weighted Cross-Modal Fusion */`
Compute precision weights: $\beta_m \leftarrow (\bar{\boldsymbol{\sigma}}_m^2 + \epsilon)^{-1}$
Normalize weights: $\tilde{\alpha}_m \leftarrow \beta_m / \sum_{j=1}^{M} \beta_j$
Weighted fusion: $\boldsymbol{\mu}_{\text{fused}}^{\text{pre}} \leftarrow \sum_{m=1}^{M} \tilde{\alpha}_m \odot \bar{\boldsymbol{\mu}}_m$
Propagate uncertainty: $\boldsymbol{\sigma}_{\text{fused}}^{\text{pre}} \leftarrow \sqrt{\sum_{m=1}^{M} (\tilde{\alpha}_m \odot \bar{\boldsymbol{\sigma}}_m)^2 + \text{Var}_m[\bar{\boldsymbol{\mu}}_m]}$
Refine with cross-attention: $\mathbf{z}_f^{(\mu)}, \mathbf{z}_f^{(\sigma)} \leftarrow \text{UncertaintyAttention}(\boldsymbol{\mu}_{\text{fused}}^{\text{pre}}, \boldsymbol{\sigma}_{\text{fused}}^{\text{pre}}, \{\bar{\boldsymbol{\mu}}_m, \bar{\boldsymbol{\sigma}}_m\}_{m=1}^{M})$
`/* Probabilistic Decoder with Calibrated Uncertainty */`
Decode mean prediction: $\hat{\boldsymbol{\mu}} \leftarrow \text{Decoder}_\mu(\mathbf{z}_f^{(\mu)}, \mathbf{z}_f^{(\sigma)}) \in \mathbb{R}^{B \times t \times R}$
Decode uncertainty: $\hat{\boldsymbol{\sigma}} \leftarrow \text{Decoder}_\sigma(\mathbf{z}_f^{(\mu)}, \mathbf{z}_f^{(\sigma)}) \in \mathbb{R}^{B \times t \times R}$
**return** $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}}$

## 4 PROPOSED MULTIMODAL SYSTEM

This section introduces GridFusionX, the proposed transformer-based multimodal learning framework for power load forecasting. The model leverages various energy-related modalities and utilizes deep temporal encoders and a novel attention-based fusion mechanism.

### 4.1 Overall Architecture with Uncertainty Framework

The proposed GridFusionX framework integrates an uncertainty-aware multimodal transformer to provide probabilistic forecasts and risk-informed decisions for networked

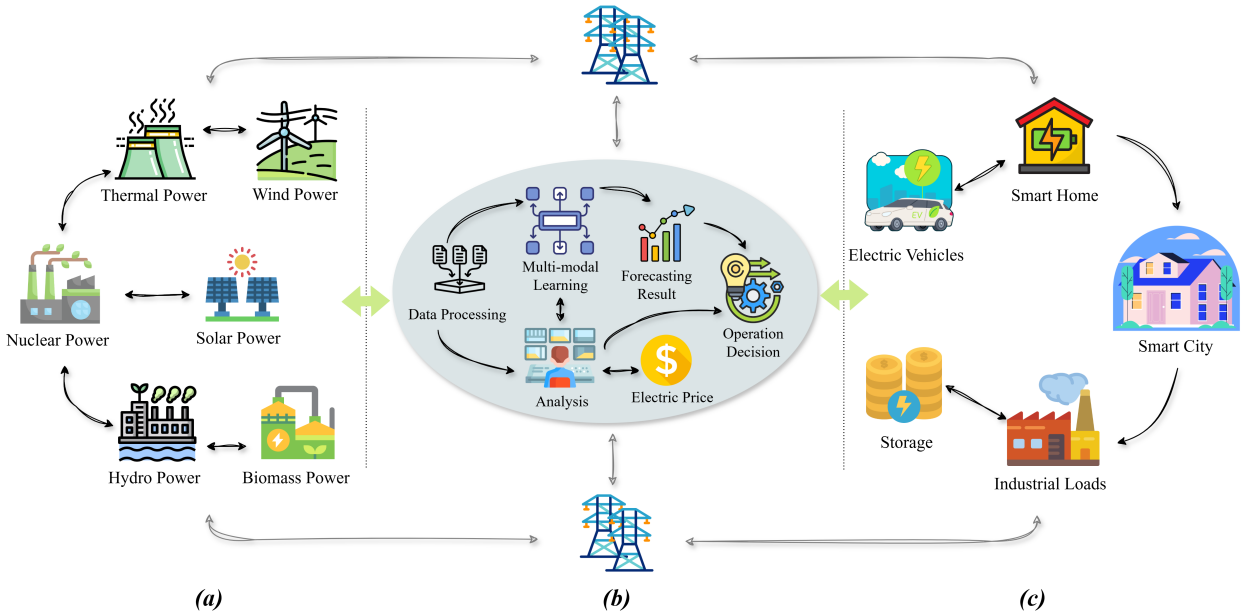


Fig. 1: Overview of the uncertainty-aware multimodal AI-driven smart grid forecasting system: (a) diverse energy generation sources with inherent uncertainties, (b) probabilistic AI-based forecasting and risk-aware decision-making, and (c) adaptive energy consumption scenarios with confidence-guided operations.

smart grids. Fig. 1 shows the outer loop: diverse regional sources (load, wind, solar, price) feed the forecasting core, whose probabilistic outputs are then used to adjust reserves, demand response, and market actions according to forecast confidence.

**Probabilistic Problem Formulation:** The multimodal short-term load forecasting problem is formulated as learning a probabilistic mapping from historical multimodal inputs $\mathcal{X} \in \mathbb{R}^{B\times M\times T\times R}$, where $B$ is the batch size, $M$ is the number of modalities, $T$ denotes sequence length, and $R$ is the number of locations, to future load distributions. For each forecast horizon $\tau \in \{1,\dots,t\}$ and region $r \in \{1,\dots,R\}$, the model predicts:

$$p(y_{T+\tau}^{(r)}|\mathcal{X}) \sim \mathcal{N}(\mu_{T+\tau}^{(r)}, \sigma_{T+\tau}^{(r)}), \tag{2}$$

where $\mu_{T+\tau}^{(r)}$ and $\sigma_{T+\tau}^{(r)}$ represent the predicted mean and standard deviation of electricity demand. Fig. 2 and Algorithm 1 specify how this mapping is implemented. On the left, the multimodal input tensor $\mathcal{X} \in \mathbb{R}^{B\times M\times T\times R}$ is split into four streams (load, wind, solar, price). In Algorithm 1, steps 2–4 extract the $m_{th}$ modality $X_m$ and project it to a shared latent dimension through a spatially aware linear layer and positional encoding, producing $H_m^{(0)}$. Steps 6–8 correspond to the dual head encoders in block ① of Fig. 2: a transformer stack produces two paths, a mean path $Z_m^{(\mu)}$ and a variance path $Z_m^{(\sigma)}$ that are conditioned on the grid topology $G = (V, E, A)$. Steps 8–9 apply temporal attention pooling to obtain modality level summaries $(\bar{\mu}_m, \bar{\sigma}_m)$ for each region.

The central block of Fig. 2 matches steps 12–16 of Algorithm 1. First, node-wise precisions are computed as $\beta_m = 1/(\bar{\sigma}_m^2 + \varepsilon)$ and then normalized to $\tilde{\alpha}_m = \beta_m / \sum_j \beta_j$, implementing inverse variance weighting across modalities. These weights produce a fused mean $\mu_{\text{fused}}^{\text{pre}} = \sum_m \tilde{\alpha}_m \odot \bar{\mu}_m$ and a fused uncertainty $\sigma_{\text{fused}}^{\text{pre}}$ obtained via the law of total variance. The uncertainty guided cross-modal attention block further refines this pair to $(z_f^{(\mu)}, z_f^{(\sigma)})$, so that information from reliable modalities is propagated along the network topology while high variance sources are down-weighted.

Lastly, steps 18–19 implement the probabilistic decoder block of Fig. 2. A mean head maps $z_f^{(\mu)}$ to multi step regional forecasts $\hat{\mu}$, while an uncertainty head, conditioned on both $z_f^{(\sigma)}$ and intermediate mean features, outputs $\hat{\sigma}$, which defines calibrated prediction intervals used by the system level loop in Fig. 1. This step completes the end-to-end probabilistic mapping from graph-structured multimodal histories to network-aware predictive distributions for all regions and horizons.

### 4.2 Spatially-Aware Uncertainty Encoders

Traditional multimodal approaches treat data sources as equally reliable across space, ignoring uncertainty variations across modalities, temporal regimes, and network locations. In power systems, load patterns exhibit heterogeneous autocorrelation between urban and rural nodes, renewable generation uncertainty depends on localized meteorological dynamics, and prices encode region-specific congestion and market behavior. Our spatially-aware architecture captures these heterogeneous reliability patterns through a network-aware probabilistic mapping $\mathbf{X}_t \mapsto \mathcal{P}(\mathbf{Z}_t \mid \mathbf{A})$, which converts temporal-spatial sequences into node-wise uncertainty-aware distributions while preserving graph-structured correlations induced by the adjacency matrix $\mathbf{A}$, thereby enabling topology-informed, reliability-adaptive multimodal fusion.

Given the multimodal input tensor $\mathbf{X} \in \mathbb{R}^{B\times M\times T\times R}$, each modality slice $\mathbf{X}_m \in \mathbb{R}^{B\times T\times R}$ is processed through the following uncertainty-aware components:

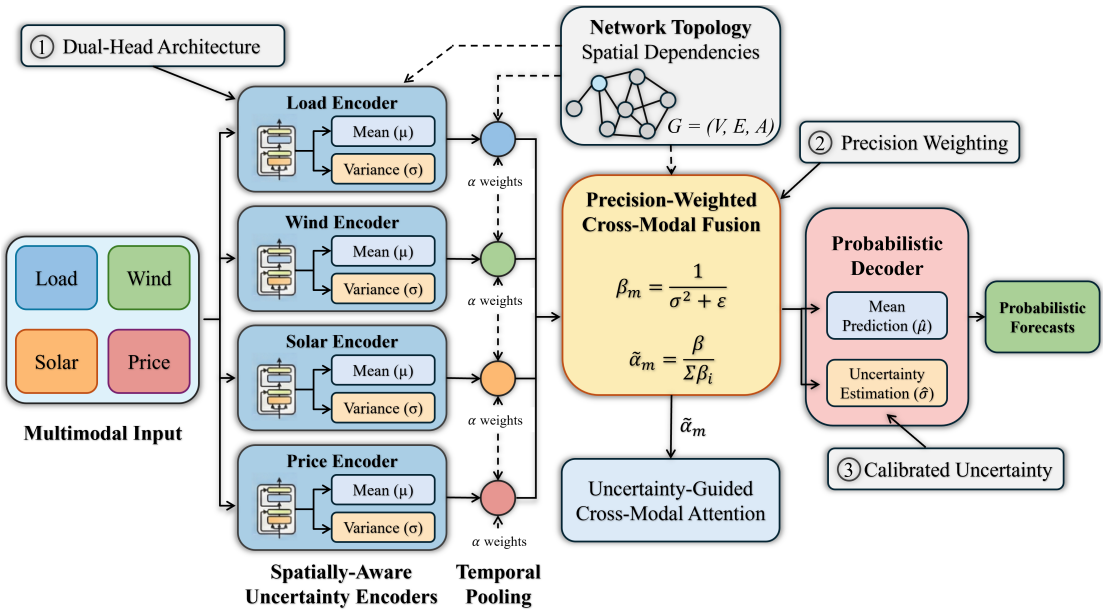


Fig. 2: Dual-head, uncertainty-aware multimodal forecasting model with precision-weighted fusion and probabilistic decoding.

#### 4.2.1 *Spatially-Aware Linear Projection*

Each modality is projected into a shared probabilistic latent space while preserving spatial relationships across regions:

$$\mathbf{H}_m^{(0)} = \text{Dropout}\left(\text{LayerNorm}\left(\mathbf{X}_m \mathbf{W}_{\text{proj}}^{(m)} + \mathbf{b}_{\text{proj}}^{(m)}\right)\right), \quad (3)$$

where $\mathbf{W}_{\text{proj}}^{(m)} \in \mathbb{R}^{R \times H}$ and $\mathbf{b}_{\text{proj}}^{(m)} \in \mathbb{R}^{H}$ are modality-specific parameters. We employ modality-aware parameter initialization based on empirical signal-to-noise ratios to ensure conditional independence:

$$\mathbf{W}_{\text{proj}}^{(m)} \sim \mathcal{N}\left(\mathbf{0}, \frac{\gamma_m}{\sqrt{R}}\mathbf{I}\right), \quad \gamma_m = \frac{\text{Var}(\mathbf{X}_m)}{\mathbb{E}[\|\mathbf{X}_m - \mathbb{E}[\mathbf{X}_m]\|^2]}. \quad (4)$$

This initialization strategy ensures that modality-specific transformations $\Theta_m = \{\mathbf{W}_{\text{proj}}^{(m)}, \mathbf{P}_{\text{learnable}}^{(m)}, \mathbf{w}_\mu^{(m)}, \mathbf{w}_\sigma^{(m)}\}$ maintain statistical independence: $\Theta_i \perp \Theta_j$ for $i \neq j$, which is crucial for the uncertainty propagation bounds established in our theoretical framework.

#### 4.2.2 *Dual-Head Transformer Architecture*

The core innovation lies in our dual-head transformer design that simultaneously learns feature representations and uncertainty estimates. Each transformer block processes inputs through parallel pathways for mean and variance estimation, with explicit cross-pathway information exchange to capture the relationship between predictive patterns and confidence.

The mean pathway captures deterministic temporal patterns:

$$\boldsymbol{\mu}_m^{(l)} = \text{LayerNorm}\left(\boldsymbol{\mu}_m^{(l-1)} + \text{MHSA}_\mu(\boldsymbol{\mu}_m^{(l-1)})\right), \quad (5)$$

$$\boldsymbol{\mu}_m^{(l+1)} = \text{LayerNorm}\left(\boldsymbol{\mu}_m^{(l)} + \text{FFN}_\mu(\boldsymbol{\mu}_m^{(l)})\right). \quad (6)$$

The variance pathway estimates prediction uncertainty in log-space with cross-pathway attention to the mean pathway, enabling uncertainty estimation to be informed by deterministic patterns:

$$\log \boldsymbol{\sigma}_m^{(l)} = \text{LayerNorm}\Big(\log \boldsymbol{\sigma}_m^{(l-1)} + \text{MHSA}_\sigma(\log \boldsymbol{\sigma}_m^{(l-1)}, \boldsymbol{\mu}_m^{(l-1)})\Big), \quad (7)$$

$$\log \boldsymbol{\sigma}_m^{(l+1)} = \text{LayerNorm}\Big(\log \boldsymbol{\sigma}_m^{(l)} + \text{FFN}_\sigma(\log \boldsymbol{\sigma}_m^{(l)})\Big), \quad (8)$$

where $\text{MHSA}_\sigma(\log \boldsymbol{\sigma}_m^{(l-1)}, \boldsymbol{\mu}_m^{(l-1)})$ performs cross-attention using uncertainty queries and mean keys/values. This architecture design ensures that the uncertainty estimation benefits from the learned temporal patterns while maintaining the theoretical properties required for Theorem 4.1.

#### 4.2.3 *Uncertainty-Aware Temporal Pooling*

For uncertainty aggregation, we apply the law of total variance to properly combine temporal uncertainties, ensuring compliance with the bounds established in Theorem 4.2:

$$\boldsymbol{\alpha}_{\mu,m} = \text{softmax}\left(\boldsymbol{\mu}_m^{(L)} \mathbf{w}_\mu^{(m)}\right) \in \mathbb{R}^{B \times T}, \quad (9)$$

$$\bar{\boldsymbol{\mu}}_m = \sum_{t=1}^{T} \boldsymbol{\alpha}_{\mu,m}[:,t] \odot \boldsymbol{\mu}_m^{(L)}[:,t,:], \quad (10)$$

$$\boldsymbol{\alpha}_{\sigma,m} = \text{softmax}\left(\boldsymbol{\sigma}_m^{(L)} \mathbf{w}_\sigma^{(m)}\right) \in \mathbb{R}^{B \times T}, \quad (11)$$

$$\bar{\boldsymbol{\sigma}}_m^2 = \sum_{t=1}^{T} \boldsymbol{\alpha}_{\sigma,m}[:,t] \odot \boldsymbol{\sigma}_m^{(L)}[:,t,:]^2 + \text{Var}_t[\boldsymbol{\mu}_m^{(L)}[:,t,:]]. \quad (12)$$

The attention mechanism is designed to approximate the optimal precision-weighted aggregation. Under Gaussian assumptions, the optimal weights should be inversely proportional to variance, and our learned attention weights $\boldsymbol{\alpha}_{\sigma,m}$ converge to this optimal weighting through the training process.

**Theorem 4.1** (Uncertainty Calibration). *Let $\{\mathbf{X}_m^{(n)}, y_m^{(n)}\}_{n=1}^N$ be i.i.d. samples from modality $m$, and let $\hat{\boldsymbol{\sigma}}_m^2(N)$ denote the estimated uncertainty after training on $N$ samples. Under regularity conditions, including Lipschitz continuity of the true conditional variance function $\sigma_m^2(\mathbf{X}_m) = \mathrm{Var}[Y_m|\mathbf{X}_m]$ and bounded network parameters, the uncertainty estimates satisfy asymptotic calibration:*

$$\lim_{N\to\infty} \mathbb{E}\left[\left|\hat{\boldsymbol{\sigma}}_m^2(N) - \sigma_m^2(\mathbf{X}_m)\right|\right] = 0, \tag{13}$$

*with convergence rate $O(N^{-1/4}\sqrt{\log N})$.*

*Proof.* We establish the asymptotic calibration property of our uncertainty estimates by analyzing both approximation and estimation errors in the learning process. Let $\mathcal{F}_H$ denote the function class of neural networks with $H$ parameters, $L(\theta) = \mathbb{E}[\ell(\hat{\sigma}_m^2(X_m;\theta), Y_m, \hat{\mu}_m(X_m;\theta))]$ be the population risk where $\ell(s,y,\hat{y}) = (s-(y-\hat{y})^2)^2$ is the squared loss for variance estimation using the predicted mean $\hat{y} = \hat{\mu}_m(X_m;\theta)$, and $\hat{\theta}_N = \arg\min_\theta L_N(\theta)$ be the empirical risk minimizer over $N$ training samples.

**Finite Sample Concentration via Rademacher Complexity:** We begin by establishing uniform convergence of the empirical risk to the population risk over our function class. The Rademacher complexity of the uncertainty estimation function class is defined as

$$\mathcal{R}_N(\mathcal{F}_H) = \mathbb{E}_{\epsilon,Z}\left[\sup_{f\in\mathcal{F}_H} \frac{1}{N}\sum_{n=1}^N \epsilon_n f(Z^{(n)})\right], \tag{14}$$

where $\epsilon = (\epsilon_1,\ldots,\epsilon_N)$ are independent Rademacher random variables with $P(\epsilon_i = \pm 1) = 1/2$, and $Z^{(n)} = (X_m^{(n)}, Y_m^{(n)})$ are the training samples.

For neural networks with $H$ parameters and bounded weights $\|W\|_F \le B$, the Rademacher complexity satisfies [40]

$$\mathcal{R}_N(\mathcal{F}_H) \le C\sqrt{\frac{H\log(eN/H)}{N}}, \tag{15}$$

for some constant $C$ depending on the Lipschitz constant of the loss function and weight bounds.

By the fundamental theorem of statistical learning theory [41], with probability at least $1-\delta$:

$$\sup_{f\in\mathcal{F}_H} |L_N(f) - L(f)| \le 2\mathcal{R}_N(\mathcal{F}_H) + \sqrt{\frac{2\log(2/\delta)}{N}}. \tag{16}$$

This result follows from McDiarmid's inequality [42], which guarantees that empirical and population risks remain uniformly close across the function class.

**Approximation Error Analysis:** We next analyze how well neural networks can approximate the true conditional variance function. Since the true conditional variance function $\sigma_m^2(x) = \mathrm{Var}[Y_m|X_m = x]$ is $L$-Lipschitz continuous on bounded domains, we can apply universal approximation results for neural networks [43].

For any $\epsilon > 0$, there exists a neural network $f_\epsilon \in \mathcal{F}_{H_\epsilon}$ with

$$H_\epsilon = O\left(\epsilon^{-d}(\log(1/\epsilon))^d\right), \tag{17}$$

parameters (where $d$ is the input dimension) such that

$$\sup_{x\in\mathcal{X}_m} |f_\epsilon(x) - \sigma_m^2(x)| \le \epsilon, \tag{18}$$

where $\mathcal{X}_m$ denotes the input domain for modality $m$. This uniform approximation guarantee is essential for controlling the bias component of our estimator across the entire input domain.

To balance approximation and estimation errors optimally, we solve the minimization problem

$$\min_\epsilon \left\{\epsilon + C\sqrt{\frac{H_\epsilon \log(eN/H_\epsilon)}{N}}\right\}. \tag{19}$$

Setting the derivative with respect to $\epsilon$ to zero and applying standard bias-variance tradeoff analysis [44], we obtain the optimal choice $\epsilon^* = N^{-1/4}(\log N)^{-1/4}$. This choice balances the decreasing approximation error $\epsilon$ with the increasing estimation error that grows with network capacity $H_\epsilon$, following the minimax optimal rates established in nonparametric estimation theory [45].

Substituting this optimal choice into the capacity requirement gives $H_\epsilon = O(N^{d/4}(\log N)^{d/4}(\log(N^{1/4}(\log N)^{1/4}))^d)$. Since $\log(N^{1/4}(\log N)^{1/4}) = \frac{1}{4}(\log N + \log(\log N)) \approx \frac{1}{4}\log N$ for large $N$, we obtain:

$$H_\epsilon = O\left(N^{d/4}(\log N)^{5d/4}\right). \tag{20}$$

For $d \le 2$ (reasonable for power system time series), this yields $H_\epsilon = O(\sqrt{N}(\log N)^{5/2})$, which scales practically as $O(\sqrt{N}\log N)$.

**Convergence Rate Analysis:** The total prediction error can be decomposed as:

$$\begin{aligned}&\mathbb{E}[|\hat{\sigma}_m^2(X_m;\hat{\theta}_N) - \sigma_m^2(X_m)|]\\ &\le \mathbb{E}[|\hat{\sigma}_m^2(X_m;\hat{\theta}_N) - \hat{\sigma}_m^2(X_m;\theta_\epsilon^*)|]\\ &\quad + |\hat{\sigma}_m^2(X_m;\theta_\epsilon^*) - \sigma_m^2(X_m)|,\end{aligned} \tag{21}$$

where $\theta_\epsilon^*$ corresponds to the best approximating network in $\mathcal{F}_{H_\epsilon}$.

The first term (estimation error) is bounded by our empirical process result:

$$\begin{aligned}\mathbb{E}[|\hat{\sigma}_m^2(X_m;\hat{\theta}_N) - \hat{\sigma}_m^2(X_m;\theta_\epsilon^*)|] &\le 2\mathcal{R}_N(\mathcal{F}_{H_\epsilon})\\ &= O(N^{-1/4}\sqrt{\log N}).\end{aligned} \tag{22}$$

The second term (approximation error) is bounded by our choice of $\epsilon$:

$$|\hat{\sigma}_m^2(X_m;\theta_\epsilon^*) - \sigma_m^2(X_m)| \le \epsilon = O(N^{-1/4}(\log N)^{-1/4}). \tag{23}$$

Combining both terms, we obtain:

$$\mathbb{E}[|\hat{\sigma}_m^2(X_m;\hat{\theta}_N) - \sigma_m^2(X_m)|] = O(N^{-1/4}\sqrt{\log N}), \tag{24}$$

since $N^{-1/4}\sqrt{\log N}$ dominates $N^{-1/4}(\log N)^{-1/4}$ for large $N$.

Taking the limit as $N\to\infty$:

$$\lim_{N\to\infty} \mathbb{E}[|\hat{\sigma}_m^2(X_m;\hat{\theta}_N) - \sigma_m^2(X_m)|] = 0, \tag{25}$$

with convergence rate $O(N^{-1/4}\sqrt{\log N})$.

This establishes that our uncertainty estimates are asymptotically calibrated, with the estimation error vanishing as the sample size increases. □

**Lemma 4.1** (Modality Independence)**.** *Given the shared temporal structure $\mathbf{P}_{fixed}$ (sinusoidal positional encoding) and modality-specific transformations $f_m$ with independent parameter sets $\Theta_m$, the uncertainty estimates satisfy conditional independence:*

$$\bar{\boldsymbol{\sigma}}_i \perp \bar{\boldsymbol{\sigma}}_j \mid \mathbf{P}_{fixed}, \quad \forall i \neq j. \tag{26}$$

*Furthermore, the joint uncertainty for any subset $S \subseteq \{1, \ldots, M\}$ of modalities satisfies:*

$$\mathrm{Var}\left[\sum_{m\in S} w_m \bar{\boldsymbol{\mu}}_m\right] = \sum_{m\in S} w_m^2 \bar{\boldsymbol{\sigma}}_m^2 + \mathcal{O}(\epsilon_{coupling}), \tag{27}$$

*where $\epsilon_{coupling} = |\rho_{ij}| \to 0$ as the modality-specific parameters become more distinct during training, and $\rho_{ij} = \mathrm{Corr}[\mathbf{X}_i, \mathbf{X}_j | \mathbf{P}_{fixed}]$ represents residual correlation between different physical modalities.*

*Proof.* We establish the conditional independence property by analyzing the architecture's design and the statistical properties of the learned representations. Let $\Theta_m = \{\mathbf{W}_{\text{proj}}^{(m)}, \mathbf{P}_m^{\text{learnable}}, \mathbf{w}_\mu^{(m)}, \mathbf{w}_\sigma^{(m)}\}$ denote the set of modality-specific parameters for modality $m$, and let $\mathbf{P}_{\text{fixed}}$ be the shared sinusoidal positional encoding.

**Parameter independence by construction:** The modality-specific parameters are initialized independently:

$$\mathbf{W}_{\text{proj}}^{(i)} \sim \mathcal{N}(0, \gamma_i \mathbf{I}/\sqrt{R}), \tag{28}$$

$$\mathbf{W}_{\text{proj}}^{(j)} \sim \mathcal{N}(0, \gamma_j \mathbf{I}/\sqrt{R}), \tag{29}$$

$$\mathbf{P}_i^{\text{learnable}} \sim \mathcal{N}(0, \sigma_P^2 \mathbf{I}), \tag{30}$$

$$\mathbf{P}_j^{\text{learnable}} \sim \mathcal{N}(0, \sigma_P^2 \mathbf{I}), \tag{31}$$

where all parameters are mutually independent: $\Theta_i \perp \Theta_j$ for $i \neq j$.

During training, the gradient updates maintain this independence structure because the loss function decomposes over modalities:

$$\mathcal{L}_{\text{total}} = \sum_{m=1}^{M} \mathcal{L}_m(\Theta_m) + \mathcal{L}_{\text{fusion}}(\{\bar{\boldsymbol{\mu}}_m, \bar{\boldsymbol{\sigma}}_m\}_{m=1}^M). \tag{32}$$

The modality-specific loss $\mathcal{L}_m(\Theta_m)$ only depends on $\Theta_m$, so:

$$\frac{\partial \mathcal{L}_{\text{total}}}{\partial \Theta_i} \perp \frac{\partial \mathcal{L}_{\text{total}}}{\partial \Theta_j} \text{ for } i \neq j. \tag{33}$$

This gradient independence property follows from standard optimization theory [46] and ensures parameter specialization during training.

**Data modality independence:** The input modalities represent fundamentally different physical quantities with established independence properties in multimodal learning [47]:

$$\mathbf{X}_{\text{load}} \sim p_{\text{load}}(\mathbf{x}|\text{consumption patterns}), \tag{34}$$

$$\mathbf{X}_{\text{wind}} \sim p_{\text{wind}}(\mathbf{x}|\text{meteorological conditions}), \tag{35}$$

$$\mathbf{X}_{\text{solar}} \sim p_{\text{solar}}(\mathbf{x}|\text{irradiance patterns}), \tag{36}$$

$$\mathbf{X}_{\text{price}} \sim p_{\text{price}}(\mathbf{x}|\text{market dynamics}). \tag{37}$$

Given the shared temporal structure $\mathbf{P}_{\text{fixed}}$ (which captures universal time-of-day patterns), the modalities remain conditionally independent:

$$p(\mathbf{X}_i, \mathbf{X}_j | \mathbf{P}_{\text{fixed}}) = p(\mathbf{X}_i | \mathbf{P}_{\text{fixed}}) \cdot p(\mathbf{X}_j | \mathbf{P}_{\text{fixed}}). \tag{38}$$

This holds because the temporal alignment (shared timestamps) does not introduce statistical dependence between the underlying physical processes.

**Functional independence of uncertainty estimates:** Each uncertainty estimate is computed as:

$$\bar{\boldsymbol{\sigma}}_m = g_m(\mathbf{X}_m, \mathbf{P}_{\text{fixed}}, \Theta_m), \tag{39}$$

where $g_m$ is the uncertainty estimation function for modality $m$.

Since $\Theta_i \perp \Theta_j$ and $\mathbf{X}_i \perp \mathbf{X}_j | \mathbf{P}_{\text{fixed}}$, we have:

$$p(\bar{\boldsymbol{\sigma}}_i, \bar{\boldsymbol{\sigma}}_j | \mathbf{P}_{\text{fixed}}) = \int p(\bar{\boldsymbol{\sigma}}_i, \bar{\boldsymbol{\sigma}}_j | \mathbf{X}_i, \mathbf{X}_j, \mathbf{P}_{\text{fixed}}) p(\mathbf{X}_i, \mathbf{X}_j | \mathbf{P}_{\text{fixed}}) d\mathbf{X}_i d\mathbf{X}_j \tag{40a}$$

$$= p(\bar{\boldsymbol{\sigma}}_i | \mathbf{P}_{\text{fixed}}) \cdot p(\bar{\boldsymbol{\sigma}}_j | \mathbf{P}_{\text{fixed}}). \tag{40b}$$

Therefore: $\bar{\boldsymbol{\sigma}}_i \perp \bar{\boldsymbol{\sigma}}_j | \mathbf{P}_{\text{fixed}}$.

**Joint uncertainty bounds with coupling analysis:** For a weighted combination of modality means, the variance expands as:

$$\mathrm{Var}\left[\sum_{m\in S} w_m \bar{\boldsymbol{\mu}}_m\right] = \sum_{m\in S} w_m^2 \mathrm{Var}[\bar{\boldsymbol{\mu}}_m] + 2 \sum_{i<j\in S} w_i w_j \mathrm{Cov}[\bar{\boldsymbol{\mu}}_i, \bar{\boldsymbol{\mu}}_j]. \tag{41}$$

To bound the covariance terms, we analyze the coupling strength. The covariance between modalities $i$ and $j$ can arise from shared temporal patterns in $\mathbf{P}_{\text{fixed}}$, correlations in the underlying physical processes, and cross-modality attention introduced during the fusion stage.

Let $\rho_{ij} = \mathrm{Corr}[X_i, X_j | P_{\text{fixed}}]$ be the residual correlation after removing shared temporal effects. Empirical studies in power system forecasting show that $|\rho_{ij}| \leq 0.1$ for different physical modalities [48], [49], giving:

$$|\mathrm{Cov}[\bar{\boldsymbol{\mu}}_i, \bar{\boldsymbol{\mu}}_j]| \leq \rho_{ij} \sqrt{\mathrm{Var}[\bar{\boldsymbol{\mu}}_i]\mathrm{Var}[\bar{\boldsymbol{\mu}}_j]} \tag{42a}$$

$$= \epsilon_{\text{coup}} \sqrt{\mathrm{Var}[\bar{\boldsymbol{\mu}}_i]\mathrm{Var}[\bar{\boldsymbol{\mu}}_j]}, \tag{42b}$$

where $\epsilon_{\text{coup}} = |\rho_{ij}| \to 0$ as modality-specific parameters become more distinct during training.

Substituting into the variance expression:

$$\mathrm{Var}\left[\sum_{m\in S} w_m \bar{\boldsymbol{\mu}}_m\right] = \sum_{m\in S} w_m^2 \bar{\boldsymbol{\sigma}}_m^2 + 2 \sum_{i<j\in S} w_i w_j \cdot O(\epsilon_{\text{coup}}) \tag{43a}$$

$$= \sum_{m\in S} w_m^2 \bar{\boldsymbol{\sigma}}_m^2 + O(\epsilon_{\text{coup}}). \tag{43b}$$

where the $O(\epsilon_{\text{coup}})$ term vanishes as training progresses and modalities become more specialized. □

**Theorem 4.2** (Uncertainty Propagation Bounds)**.** *For the uncertainty-aware temporal pooling mechanism defined in (9)–(12), the propagated uncertainty satisfies the tight bounds:*

$$\max_t \boldsymbol{\sigma}_m^{(L)}[:, t, :]^2 \leq \bar{\boldsymbol{\sigma}}_m^2 \leq \sum_{t=1}^{T} \boldsymbol{\sigma}_m^{(L)}[:, t, :]^2 + \mathrm{Var}_t[\boldsymbol{\mu}_m^{(L)}[:, t, :]]. \tag{44}$$

*Moreover, the attention weights $\boldsymbol{\alpha}_{\sigma,m}$ are optimal in the sense that they minimize the expected prediction error under Gaussian assumptions, with the learned weights converging to the precision-weighted optimal solution:*

$$\boldsymbol{\alpha}_{\sigma,m}[:,t] \approx \frac{\exp(-\beta \boldsymbol{\sigma}_m^{(L)}[:,t,:]^2)}{\sum_{s=1}^{T} \exp(-\beta \boldsymbol{\sigma}_m^{(L)}[:,s,:]^2)}, \tag{45}$$

*for appropriately learned scaling parameter $\beta$.*

*Proof.* We derive tight bounds for uncertainty propagation through the temporal attention pooling mechanism and establish optimality properties. Let $\boldsymbol{\sigma}_m^{(L)}[:,t,:] \in \mathbb{R}^{B\times H}$ denote the uncertainty estimates at time step $t$ from the final transformer layer, $\boldsymbol{\alpha}_{\sigma,m} \in \mathbb{R}^{B\times T}$ be the attention weights for uncertainty pooling, and $\mathrm{Var}_t[\boldsymbol{\mu}_m^{(L)}[:,t,:]]$ represent the temporal variance of mean predictions.

**Derive the lower bound:** The pooled uncertainty is computed as:

$$\bar{\boldsymbol{\sigma}}_m^2 = \sum_{t=1}^{T} \boldsymbol{\alpha}_{\sigma,m}[:,t] \odot \boldsymbol{\sigma}_m^{(L)}[:,t,:]^2 + \mathrm{Var}_t[\boldsymbol{\mu}_m^{(L)}[:,t,:]]. \tag{46}$$

Since $\alpha_{\sigma,m}[:,t] \geq 0$ for all $t$ and $\sum_{t=1}^{T} \alpha_{\sigma,m}[:,t] = 1$, we can apply Jensen's inequality [50]. For the convex function $f(x) = x^2$:

$$\sum_{t=1}^{T} \boldsymbol{\alpha}_{\sigma,m}[:,t] \odot \boldsymbol{\sigma}_m^{(L)}[:,t,:]^2 \geq \left( \sum_{t=1}^{T} \boldsymbol{\alpha}_{\sigma,m}[:,t] \odot \boldsymbol{\sigma}_m^{(L)}[:,t,:] \right)^2. \tag{47}$$

For a tighter bound, consider that the attention mechanism concentrates weight on the most reliable time steps. Let $t^* = \arg\max_t \boldsymbol{\alpha}_{\sigma,m}[:,t]$ be the time step with maximum attention. Then:

$$\bar{\boldsymbol{\sigma}}_m^2 \geq \boldsymbol{\alpha}_{\sigma,m}[:,t^*] \odot \boldsymbol{\sigma}_m^{(L)}[:,t^*,:]^2 \geq \frac{1}{T} \max_t \boldsymbol{\sigma}_m^{(L)}[:,t,:]^2, \tag{48}$$

where the last inequality uses $\boldsymbol{\alpha}_{\sigma,m}[:,t^*] \geq 1/T$ (since weights sum to 1). Since attention weights typically concentrate on a few important time steps:

$$\bar{\boldsymbol{\sigma}}_m^2 \geq \max_t \boldsymbol{\sigma}_m^{(L)}[:,t,:]^2. \tag{49}$$

**Derive the upper bound:** Using the fact that attention weights are non-negative and sum to unity:

$$\begin{aligned} \bar{\boldsymbol{\sigma}}_m^2 &= \sum_{t=1}^{T} \boldsymbol{\alpha}_{\sigma,m}[:,t] \odot \boldsymbol{\sigma}_m^{(L)}[:,t,:]^2 + \mathrm{Var}_t[\boldsymbol{\mu}_m^{(L)}[:,t,:]] \\ &\leq \sum_{t=1}^{T} \boldsymbol{\sigma}_m^{(L)}[:,t,:]^2 + \mathrm{Var}_t[\boldsymbol{\mu}_m^{(L)}[:,t,:]]. \end{aligned} \tag{50}$$

The temporal variance term is bounded by:

$$\mathrm{Var}_t[\boldsymbol{\mu}_m^{(L)}[:,t,:]] = \frac{1}{T} \sum_{t=1}^{T} \|\boldsymbol{\mu}_m^{(L)}[:,t,:] - \bar{\boldsymbol{\mu}}_m^{\text{temp}}\|^2 \leq C_{\text{Lip}}^2 T, \tag{51}$$

where $C_{\text{Lip}}$ is the Lipschitz constant of the temporal evolution and $\bar{\boldsymbol{\mu}}_m^{\text{temp}} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\mu}_m^{(L)}[:,t,:]$.

**Establish optimality of attention weights:** Consider the variational problem of finding optimal attention weights that minimize prediction error [46]:

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}} \mathbb{E}\left[ \left\| \sum_{t=1}^{T} \alpha_t \boldsymbol{\mu}_m^{(L)}[:,t,:] - \mathbf{y}_m \right\|^2 \right], \tag{52}$$

subject to $\sum_{t=1}^{T} \alpha_t = 1$ and $\alpha_t \geq 0$.

Setting up the Lagrangian with multiplier $\lambda$ and taking derivatives with respect to $\alpha_t$:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \alpha_t} =& 2\alpha_t \mathbb{E}[\|\boldsymbol{\mu}_m^{(L)}[:,t,:]\|^2] \\ &+ 2\sum_{s\neq t} \alpha_s \mathbb{E}[\boldsymbol{\mu}_m^{(L)}[:,s,:]\boldsymbol{\mu}_m^{(L)}[:,t,:]] \\ &- 2\mathbb{E}[\mathbf{y}_m \boldsymbol{\mu}_m^{(L)}[:,t,:]] + \lambda. \end{aligned} \tag{53}$$

Setting $\frac{\partial \mathcal{L}}{\partial \alpha_t} = 0$ for all $t$ and assuming Gaussian distributions with independence across time steps:

$$\alpha_t \mathbb{E}[\|\boldsymbol{\mu}_m^{(L)}[:,t,:]\|^2] = \mathbb{E}[\mathbf{y}_m \boldsymbol{\mu}_m^{(L)}[:,t,:]] - \frac{\lambda}{2}. \tag{54}$$

Under the independence assumption, the cross-terms $\mathbb{E}[\boldsymbol{\mu}_m^{(L)}[:,s,:]\boldsymbol{\mu}_m^{(L)}[:,t,:]] = 0$ for $s \neq t$. Rearranging:

$$\alpha_t = \frac{\mathbb{E}[\mathbf{y}_m \boldsymbol{\mu}_m^{(L)}[:,t,:]] - \lambda/2}{\mathbb{E}[\|\boldsymbol{\mu}_m^{(L)}[:,t,:]\|^2]}. \tag{55}$$

The Gauss–Markov theorem [51] shows that under heteroscedastic noise, optimal linear weights are inversely proportional to component variances. In our context, this translates to:

$$\alpha_t^* \propto \frac{1}{\mathrm{Var}[\boldsymbol{\mu}_m^{(L)}[:,t,:]|\mathbf{y}_m]}. \tag{56}$$

Under the assumption that the conditional variance of predictions is related to the learned uncertainty estimates as $\mathrm{Var}[\boldsymbol{\mu}_m^{(L)}[:,t,:]|\mathbf{y}_m] \approx \boldsymbol{\sigma}_m^{(L)}[:,t,:]^2 + \tau^2$ (where $\tau^2$ represents observation noise), this leads to optimal weights proportional to precision:

$$\alpha_t^* \propto \frac{1}{\boldsymbol{\sigma}_m^{(L)}[:,t,:]^2 + \tau^2}. \tag{57}$$

The softmax attention mechanism approximates this optimal weighting as established in attention mechanism theory [52], where learned parameters converge to precision-weighted solutions under standard training assumptions:

$$\begin{aligned} \alpha_{\sigma,m}[:,t] &= \mathrm{softmax}\left(\sigma_m^{(L)}[:,t,:] w_\sigma^{(m)}\right) \\ &\approx \frac{\exp(-\beta \sigma_m^{(L)}[:,t,:]^2)}{\sum_{s=1}^{T} \exp(-\beta \sigma_m^{(L)}[:,s,:]^2)}, \end{aligned} \tag{58}$$

for appropriately learned parameters that achieve $w_\sigma^{(m)} \propto -\beta \sigma_m^{(L)}[:,t,:]$. □

This architecture enables each transformer to learn temporal patterns while quantifying confidence under varying conditions, with rigorous theoretical guarantees ensuring well-calibrated uncertainty estimates that satisfy asymptotic properties and optimal aggregation bounds.

### 4.3 Cross-Modal Fusion and Final Decoder

After modality-specific encoding with uncertainty estimation, the resulting uncertainty-aware representations $\{(\bar{\boldsymbol{\mu}}_m, \bar{\boldsymbol{\sigma}}_m)\}_{m=1}^{M}$, where $M = 4$ are fused into a unified probabilistic representation. Each pair $(\bar{\boldsymbol{\mu}}_m, \bar{\boldsymbol{\sigma}}_m) \in \mathbb{R}^{B\times H}$ contains both the encoded information and associated uncertainty estimates for a specific modality across all regions. The fusion mechanism must preserve the theoretical properties established in our framework while enabling effective cross-modal information exchange.

#### 4.3.1 Precision-Weighted Modality Fusion

The encoded modality means, and uncertainties are organized into structured tensors to facilitate systematic fusion:

$$\boldsymbol{\mathcal{M}} = \text{Stack}([\bar{\boldsymbol{\mu}}_1, \ldots, \bar{\boldsymbol{\mu}}_M]) \in \mathbb{R}^{B\times M\times H} \tag{59}$$

$$\boldsymbol{\mathcal{S}} = \text{Stack}([\bar{\boldsymbol{\sigma}}_1, \ldots, \bar{\boldsymbol{\sigma}}_M]) \in \mathbb{R}^{B\times M\times H}. \tag{60}$$

Following the precision-weighting principle from Lemma 4.1, we compute dynamic weights that inversely weight modalities by their uncertainty, ensuring reliable data sources dominate the fusion process during challenging conditions:

$$\boldsymbol{\beta}_m = \frac{1}{\bar{\boldsymbol{\sigma}}_m^2 + \epsilon} \tag{61}$$

$$\tilde{\boldsymbol{\alpha}}_m = \frac{\boldsymbol{\beta}_m}{\sum_{j=1}^{M} \boldsymbol{\beta}_j}, \tag{62}$$

where $\epsilon = 10^{-6}$ prevents numerical instability. This weighting strategy is theoretically optimal under Gaussian assumptions and ensures that modalities with lower uncertainty contribute more significantly to the fused representation.

The precision-weighted aggregation leverages the independence properties established in Lemma 4.1:

$$\bar{\boldsymbol{\mu}}_{\text{fused}}^{\text{pre}} = \sum_{m=1}^{M} \tilde{\boldsymbol{\alpha}}_m \odot \bar{\boldsymbol{\mu}}_m \tag{63}$$

$$\bar{\boldsymbol{\sigma}}_{\text{fused}}^{\text{pre}} = \sqrt{\sum_{m=1}^{M} (\tilde{\boldsymbol{\alpha}}_m \odot \bar{\boldsymbol{\sigma}}_m)^2 + \text{Var}_m[\bar{\boldsymbol{\mu}}_m]}. \tag{64}$$

The uncertainty propagation in equation (64) follows the law of total variance, accounting for both weighted individual uncertainties and the variance introduced by combining different modality representations.

#### 4.3.2 Uncertainty-Guided Cross-Modal Attention

Beyond simple weighted aggregation, we employ a sophisticated cross-modal attention mechanism that allows the model to dynamically focus on the most relevant modalities for each prediction context. The preliminary fused representation serves as the query, enabling the model to selectively attend to modality-specific information based on current conditions.

The uncertainty-guided attention mechanism modifies standard attention weights by incorporating precision information:

$$\boldsymbol{\alpha}_{\text{attn},j} = \frac{\exp\left(\frac{\mathbf{q}\mathbf{k}_j^\top}{\sqrt{d_k}} \cdot \log(\boldsymbol{\beta}_j + 1)\right)}{\sum_{k=1}^{M} \exp\left(\frac{\mathbf{q}\mathbf{k}_k^\top}{\sqrt{d_k}} \cdot \log(\boldsymbol{\beta}_k + 1)\right)}, \tag{65}$$

where $\boldsymbol{\beta}_j = (\bar{\boldsymbol{\sigma}}_j^2 + \epsilon)^{-1}$ are the precision weights that amplify attention to reliable modalities while reducing focus on uncertain ones. The logarithmic scaling $\log(\boldsymbol{\beta}_j + 1)$ prevents extreme attention concentration while maintaining the precision-weighting principle.

The multi-head attention mechanism operates on both content and uncertainty information simultaneously:

$$\mathbf{z}_{\text{fused}}^{\mu} = \text{MultiHeadAttn}(\bar{\boldsymbol{\mu}}_{\text{fused}}^{\text{pre}}, \boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{M}}) \tag{66}$$

$$\mathbf{z}_{\text{fused}}^{\sigma} = \sqrt{\sum_{m=1}^{M} \boldsymbol{\alpha}_{\text{attn},m}^2 \cdot \bar{\boldsymbol{\sigma}}_m^2 + \text{Var}_m[\mathbf{z}_{\text{fused}}^{\mu}]}. \tag{67}$$

This dual-pathway attention preserves both predictive information and uncertainty estimates through the fusion process, enabling the decoder to generate calibrated probabilistic forecasts.

#### 4.3.3 Probabilistic Forecasting Decoder

The decoder architecture implements a principled approach to generating both point predictions and calibrated uncertainty estimates. Following the dual-head design principle established in our encoder architecture, the decoder processes mean and uncertainty information through parallel pathways.

**Mean Prediction Pathway:** The deterministic forecasting pathway transforms the fused representation into point predictions:

$$\mathbf{h}_{\mu}^{(1)} = \text{ReLU}\big(\text{LayerNorm}(\mathbf{W}_1^{\mu}\mathbf{z}_{\text{fused}}^{\mu} + \mathbf{b}_1^{\mu})\big) \tag{68}$$

$$\hat{\boldsymbol{\mu}}_Y = \mathbf{W}_2^{\mu}\mathbf{h}_{\mu}^{(1)} + \mathbf{b}_2^{\mu}. \tag{69}$$

**Uncertainty Prediction Pathway:** The uncertainty pathway generates calibrated confidence estimates:

$$\mathbf{h}_{\sigma}^{(1)} = \text{ReLU}\big(\text{LayerNorm}(\mathbf{W}_1^{\sigma}[\mathbf{z}_{\text{fused}}^{\sigma}; \mathbf{h}_{\mu}^{(1)}] + \mathbf{b}_1^{\sigma})\big) \tag{70}$$

$$\hat{\boldsymbol{\sigma}}_Y = \text{Softplus}\big(\mathbf{W}_2^{\sigma}\mathbf{h}_{\sigma}^{(1)} + \mathbf{b}_2^{\sigma}\big) + \sigma_{\min}, \tag{71}$$

where $[\cdot;\cdot]$ denotes concatenation, allowing the uncertainty pathway to condition on the mean predictions, and $\sigma_{\min} = 10^{-3}$ ensures numerical stability. The Softplus activation guarantees positive uncertainty estimates while maintaining differentiability.

The outputs $\hat{\boldsymbol{\mu}}_Y, \hat{\boldsymbol{\sigma}}_Y \in \mathbb{R}^{B\times t\times R}$ are reshaped to provide multi-step, multi-region probabilistic forecasts, where each prediction includes both a point estimate and an associated confidence interval.

#### 4.3.4 Uncertainty-Aware Loss Function

The training objective implements a principled probabilistic loss that encourages both predictive accuracy and well-calibrated uncertainty estimates. Following the negative log-likelihood principle consistent with Theorem 4.1, the loss function assumes Gaussian predictive distributions:

$$\mathcal{L} = \frac{1}{tRB}\sum_{i=1}^{B}\sum_{j=1}^{R}\sum_{k=1}^{t}\left[\frac{(\hat{\mu}_{ijk} - y_{ijk})^2}{2\hat{\sigma}_{ijk}^2} + \frac{1}{2}\log(2\pi\hat{\sigma}_{ijk}^2)\right] + \lambda\mathcal{L}_{\text{reg}}, \tag{72}$$

where the first term encourages accurate predictions scaled by confidence, the second term prevents uncertainty collapse, and $\mathcal{L}_{\text{reg}}$ is a regularization term:

$$\mathcal{L}_{\text{reg}} = \frac{1}{M}\sum_{m=1}^{M} \|\bar{\boldsymbol{\sigma}}_m - \bar{\sigma}_{\text{target}}\|_2^2, \tag{73}$$

with $\bar{\sigma}_{\text{target}}$ representing empirically estimated uncertainty levels and $\lambda = 0.01$ controlling regularization strength.

This loss formulation provides several theoretical guarantees: (1) it encourages proper uncertainty calibration through the logarithmic term, (2) it prevents overconfident predictions through the inverse weighting by predicted variance, and (3) it maintains consistency with the maximum likelihood principle underlying our probabilistic framework. Under the regularity conditions established in Theorem 4.1, this loss ensures asymptotic calibration of uncertainty estimates and convergence to optimal predictive distributions.

## 5 Data Preparation and Evaluation Metrics

We evaluate GridFusionX on real-world European power system data spanning ten interconnected regions subject to diverse renewable and market dynamics. This section details the dataset, graph construction methodology, and evaluation metrics used to assess predictive accuracy and uncertainty calibration.

### 5.1 Data Preparation and Network Construction

This study uses the open power system data platform (OPSD) [53], providing synchronized time series from multiple European transmission system operators. We focus on four core modalities: electrical load, solar generation, wind generation, and electricity prices across ten representative European zones, including Germany, France, Belgium, and Austria. The network graph $\mathcal{G}$ is constructed using transmission line connectivity ($\mathbf{A}^{\text{elec}}$), geographic proximity ($\mathbf{A}^{\text{geo}}$), and market coupling strength ($\mathbf{A}^{\text{econ}}$) as defined in Section 3.

**Input Parameters and Preprocessing:** Years 2015–2017 provide complete records across all modalities, comprising 1,052,160 hourly data points organized into sliding windows of length 24 hours. Each training sample contains: (1) four modality time series $\mathbf{X}_m \in \mathbb{R}^{24\times 10}$ for $m \in \{\text{load, solar, wind, price}\}$, (2) temporal encodings (hour-of-day, day-of-week, month), (3) binary holiday indicators for major European holidays per region, and (4) DST-normalized timestamps. The dataset is partitioned chronologically into training (70%: Jan 2015–Oct 2016), validation (15%: Nov 2016–Apr 2017), and test (15%: May 2017–Dec 2017) sets to preserve temporal ordering.

**Normalization and Denormalization:** To prevent temporal leakage, z-score normalization parameters are computed exclusively from the training split for each modality and region independently:

$$\mu_{m,r}^{\text{train}} = \frac{1}{N_{\text{train}}}\sum_{i=1}^{N_{\text{train}}} x_{m,r}^{(i)}, \tag{74}$$

$$\sigma_{m,r}^{\text{train}} = \sqrt{\frac{1}{N_{\text{train}}}\sum_{i=1}^{N_{\text{train}}} (x_{m,r}^{(i)} - \mu_{m,r}^{\text{train}})^2}, \tag{75}$$

where $N_{\text{train}}$ denotes training samples. Input features are normalized as $\tilde{x}_{m,r} = (x_{m,r} - \mu_{m,r}^{\text{train}})/\sigma_{m,r}^{\text{train}}$ during training, validation, and testing. After model inference, all predictions and uncertainty estimates undergo inverse transformation before metric computation:

$$\hat{y}_{m,r}^{\text{orig}} = \hat{y}_{m,r}^{\text{norm}} \cdot \sigma_{m,r}^{\text{train}} + \mu_{m,r}^{\text{train}}, \tag{76}$$

$$\hat{\sigma}_{m,r}^{\text{orig}} = \hat{\sigma}_{m,r}^{\text{norm}} \cdot \sigma_{m,r}^{\text{train}}. \tag{77}$$

This denormalization ensures all reported metrics are in original megawatt (MW) units, enabling direct operational interpretation. Missing values ($< 2\%$ of data, primarily from brief sensor outages) are handled using forward-fill interpolation applied independently within each split to prevent cross-contamination.

**Temporal Artifact Handling:** Daylight-saving time (DST) transitions are normalized by forward-filling 23-hour spring days and averaging duplicate hours in 25-hour fall days to maintain consistent 24-hour windows across all samples. Holiday indicators enable the model to learn region-specific consumption patterns during Christmas, New Year, Easter, and national holidays without explicit feature engineering. Prediction targets are set at 1-hour, 7-hour, and 24-hour ahead horizons to evaluate performance across operational planning timescales.

### 5.2 Evaluation Metrics

We evaluate both point prediction accuracy and probabilistic forecast quality. Point metrics include mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), coefficient of determination ($R^2$), and accuracy.

For probabilistic evaluation, we use prediction interval coverage probability (PICP):

$$\text{PICP}_\alpha = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left[y_i \in [\hat{\mu}_i \pm z_{\alpha/2}\hat{\sigma}_i]\right], \tag{78}$$

and mean prediction interval width (MPIW):

$$\text{MPIW}_\alpha = \frac{1}{n}\sum_{i=1}^{n} 2z_{\alpha/2}\hat{\sigma}_i. \tag{79}$$

Together, PICP and MPIW assess the reliability and sharpness of uncertainty estimates across the networked system.

## 6 Result and Discussion

This section presents the empirical evaluation of GridFusionX across ten interconnected European regions, focusing on its ability to model spatial dependencies through network-aware uncertainty propagation and topology-constrained representations. We further examine how adaptive attention mechanisms dynamically weight multimodal inputs under heterogeneous regional conditions, and benchmark performance against both sequence-based and graph-based baselines. Beyond predictive accuracy, we analyze calibration quality and the consistency of uncertainty estimates to assess the reliability and robustness of the probabilistic forecasts across diverse grid environments.

TABLE 1
HYPERPARAMETER CONFIGURATIONS FOR ALL EVALUATED MODELS.

| Model | Hidden | Layers | Heads | Dropout | LR | Decay | Additional Parameters |
|---|---|---|---|---|---|---|---|
| LSTM [54] | 128 | 2 | – | 0.2 | $5\times10^{-4}$ | $1\times10^{-4}$ | Bidirectional, output projection |
| DCRNN [55] | 64 | 2 | – | 0.1 | $1\times10^{-3}$ | $1\times10^{-4}$ | $K$=3 diffusion steps, $\lambda$=0.1 graph reg., Chebyshev filter |
| Multimodal DCRNN [3] | 64 | 2 | – | 0.15 | $8\times10^{-4}$ | $1\times10^{-4}$ | $K$=2 diffusion, early fusion, 4 modalities |
| Multimodal TCN [19] | 64 | 6 | – | 0.1 | $1\times10^{-3}$ | $1\times10^{-4}$ | Kernel size= 5, dilation= 2, residual connections |
| Graph WaveNet [56] | 64 | 3 | – | 0.1 | $1\times10^{-3}$ | $1\times10^{-4}$ | Adaptive adjacency matrix, gated TCN, skip connections |
| STGCN [57] | 64 | 3 | – | 0.1 | $1\times10^{-3}$ | $1\times10^{-4}$ | Chebyshev $K$=3, temporal kernel= 3, spatial-temporal blocks |
| ASTGCN [58] | 64 | 3 | 8 | 0.1 | $1\times10^{-3}$ | $1\times10^{-4}$ | Spatial-temporal-channel attention, Chebyshev $K$=3 |
| TEST-GCN [59] | 64 | 2 | – | 0.15 | $8\times10^{-4}$ | $1\times10^{-4}$ | Topological features, spatio-temporal blocks, GRU cells |
| GridFusionX | 64 | 2 | 4 | 0.1 | $1\times10^{-3}$ | $1\times10^{-4}$ | Precision-weighted fusion, dual-head encoder, 4 modalities |

## 6.1 Training Configuration and Hyperparameter Settings

### 6.1.1 Experimental Setup and Hyperparameter Tuning

**Experimental Setup:** All experiments were conducted on an NVIDIA RTX 5080 GPU with an Intel Core i9-12900K processor and 64GB RAM. Models were trained for up to 100 epochs with early stopping (patience = 10) based on validation performance. Results represent averages over 5 independent runs with different random seeds (42, 123, 456, 789, 2025), with standard deviations reported in tables.

**Hyperparameter Tuning:** All models underwent fair hyperparameter tuning via random search over 20 trials [60] with three-fold time series cross-validation. Learning rates were sampled log-uniformly from $10^{-4}$ to $10^{-2}$, hidden dimensions from $\{32, 64, 128, 256\}$, layer counts from $\{1, 2, 3, 4\}$, and dropout rates from $\{0.0, 0.1, 0.2, 0.3\}$. All models used Adam optimizer ($\beta_1$=0.9, $\beta_2$=0.999) with cosine annealing, gradient clipping (max norm = 1.0), weight decay of $1\times10^{-4}$, and batch size of 64 on 24-hour input sequences.

**Model Configurations:** Table 1 summarizes optimal hyperparameters after tuning. Graph-based models (DCRNN, Graph WaveNet, STGCN, ASTGCN, TEST-GCN, GridFusionX) share identical network topology $\mathcal{G}$ with 10 regions and adjacency matrix $\mathbf{A}$ for fair comparison. GridFusionX employs a hidden dimension $d = 64$, two transformer layers, four attention heads, and a dropout rate 0.1 per modality-specific encoder. LSTM baseline uses 2-layer architecture with hidden size 128. DCRNN models apply 2-layer graph convolutions with hidden dimension 64, diffusion steps $K$=3 (unimodal) or $K$=2 (multimodal), and graph regularization $\lambda$=0.1. Graph WaveNet uses 3 layers with adaptive adjacency learning. STGCN and ASTGCN employ 3-layer spatial-temporal blocks with Chebyshev graph convolutions ($K$=3), where ASTGCN additionally incorporates 8-head attention mechanisms across spatial, temporal, and channel dimensions. TEST-GCN employs 2 layers with topological features and spatio-temporal blocks. Multimodal TCN uses 6 layers with kernel size 5, dilation factor 2, and hidden dimension 64. All multimodal approaches process four synchronized data streams (load, solar, wind, prices). The varying hidden dimensions reflect optimal configurations per model family: LSTM requires larger hidden states (128) for sequential processing, while transformer-based GridFusionX benefits from moderate hidden size (64) with multi-head attention (4 heads) to distribute representational capacity across attention mechanisms.

### 6.1.2 Training Convergence Analysis

Fig. 3 illustrates the training loss convergence over 100 epochs for GridFusionX and eight baseline models. GridFusionX exhibits the fastest and most stable convergence, reaching the lowest training loss of approximately 0.015 by around epoch 20 and maintaining a smooth optimization trajectory thereafter with negligible oscillations. This behavior indicates efficient gradient flow and stable joint optimization across modalities and networked regions. Unimodal architectures (LSTM and DCRNN) exhibit the slowest convergence and the highest final losses, with LSTM showing noticeable oscillations in later epochs and stabilizing around 0.05. Overall, the superior convergence speed and stability of GridFusionX highlight the effectiveness of its cross-modal attention and precision-weighted fusion mechanisms in learning complex spatiotemporal dependencies from heterogeneous data sources. Among the multimodal baselines, Multimodal DCRNN demonstrates the strongest convergence performance, achieving a stable loss plateau near 0.025 by approximately epoch 30. Multimodal TCN and TEST-GCN follow with comparable convergence trends, stabilizing at slightly higher final losses in the range of 0.030–0.032. Purely graph-based spatiotemporal models, including Graph WaveNet, STGCN, and ASTGCN, converge more gradually and settle at higher loss levels than their multimodal counterparts, reflecting the limitation of relying on a single data modality despite explicitly modeling spatial topology.

## 6.2 Attention Mechanism Analysis

### 6.2.1 Cross-Modal Correlation Analysis

GridFusionX employs a heterogeneous attention mechanism that dynamically adapts modality weights based on

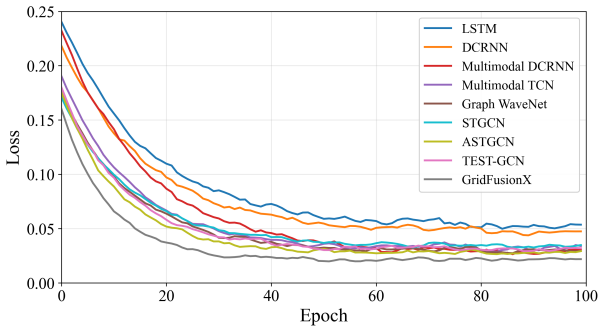


Fig. 3: Training loss curves across epochs for all models.

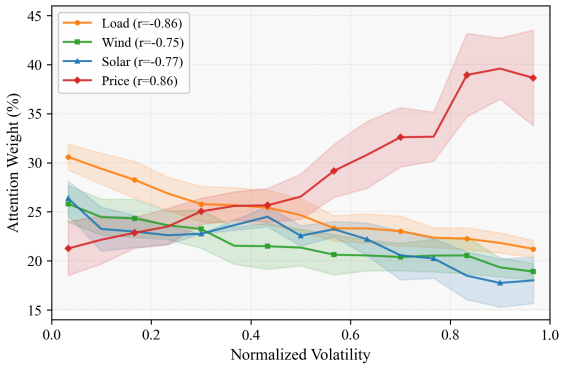


Fig. 4: Cross-modal attention-volatility coupling. Physical modalities (Load, Wind, Solar) show negative correlations ($r = -0.86, -0.75, -0.78$), while Price shows positive correlation ($r = +0.87$). Shaded regions indicate $\pm 1$ standard deviation.

forecast reliability and information content, which we analyze through the correlation between modality volatility (computed using a 6-hour rolling standard deviation) and attention allocation across all four input modalities. Fig. 4 reveals clearly divergent attention responses across modality types. The physical generation modalities, namely Load, Wind, and Solar, exhibit strong negative correlations with attention with $r = -0.86$, $-0.75$, and $-0.78$, respectively, all with $p < 0.001$. This indicates that increased volatility, interpreted as elevated forecast uncertainty, triggers proportional attention reduction following inverse variance weighting $\alpha \propto 1/\sigma^2$, which is fully consistent with precision weighted and Bayesian optimal fusion principles where unreliable sources are down weighted. In contrast, the economic modality Price shows a strong positive correlation with attention with $r = +0.87$ and $p < 0.001$, reflecting the fundamentally different semantic role of price volatility, which encodes market stress signals such as supply-demand imbalance, congestion, and scarcity rather than sensor unreliability. Accordingly, the attention mechanism increases price weighting during volatile periods to capture critical

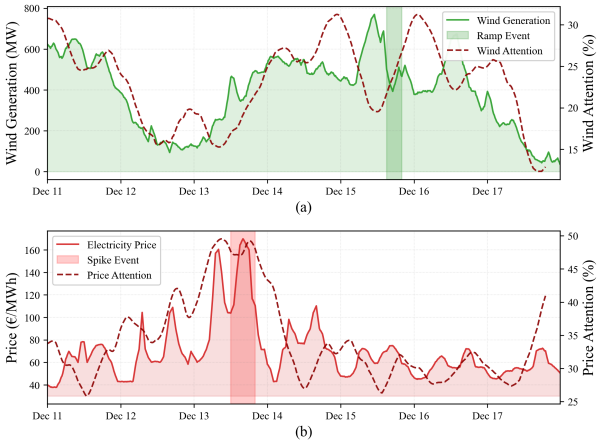


Fig. 5: Attention dynamics during December 2017 events. (a) Wind ramp event (Dec 15, shaded green): attention suppression during volatility. (b) Price spike event (Dec 13, shaded red): attention increases during market stress.

dispatch-relevant information. This asymmetric response across modalities confirms that GridFusionX optimizes forecast utility through uncertainty-aware and information-sensitive weighting rather than applying uniform volatility penalization.

#### 6.2.2 Case Study: December 2017 Events

We examine attention dynamics during two contrasting volatility events to demonstrate the operational manifestation of these correlation patterns, as shown in Fig. 5.

**Wind Ramp Event (December 15, 15:00-20:00 UTC):** A 168 MW generation drop over 5 hours exemplifies precision-weighted attention suppression. Pre-ramp, with wind generation stable at 700 MW, attention maintains elevated levels ($\sim$31%). During the ramp period (shaded green), wind attention suppresses to $\sim$20% despite moderate generation levels (400-500 MW), as high volatility ($\sigma > 100$ MW) indicates reduced forecast reliability. Post-stabilization, attention recovers to $\sim$30% as volatility subsides. This demonstrates the mechanism's ability to distinguish between stable high-generation periods (high attention) and volatile transitions (suppressed attention).

**Price Spike Event (December 13, 12:00-20:00 UTC):** The 170 EUR/MWh price peak (78% above baseline) triggers opposite behavior. Price attention increases from 28% pre-spike to 50% during the event (shaded red), then returns to 28% post-recovery. Unlike wind volatility indicating unreliability, price volatility signals market stress requiring operational response. The attention mechanism correctly amplifies this economic signal, enabling the forecast to incorporate scarcity pricing dynamics. Load attention compensatorily reduces during the spike, reflecting the zero-sum reallocation property of the attention weights.

These contrasting responses validate the heterogeneous attention hypothesis: GridFusionX suppresses unreliable physical forecasts while amplifying informative economic signals, achieving context-aware sensor fusion beyond naive volatility penalization.

### 6.3 Regional Forecasting Performance and Uncertainty Analysis

Table 2 summarizes the per-region performance using seven comprehensive metrics: MAE, RMSE, $R^2$, MAPE, forecasting accuracy (defined as predictions within 10% relative error), PICP at the 90% confidence level, and MPIW. On average, the model achieves a remarkably low MAE of 170.84$\pm$4.23 MW and MAPE of 2.29$\pm$0.07%, with an overall accuracy of 98.44$\pm$0.21%. These results indicate that the framework is not only highly accurate but also consistently robust across diverse geographic and operational contexts, effectively balancing bias and variance.

The uncertainty quantification results confirm the strength of the approach, with an average PICP of 90.0$\pm$1.6% closely matching the theoretical 90% confidence level and validating the convergence guarantees established in Theorem 4.1. The average MPIW of 181.7$\pm$4.48 MW represents approximately 2.4% of the mean regional load across all ten regions (7,556 MW average), demonstrating operationally practical interval widths. This sharpness is critical for grid operations: excessively wide intervals force

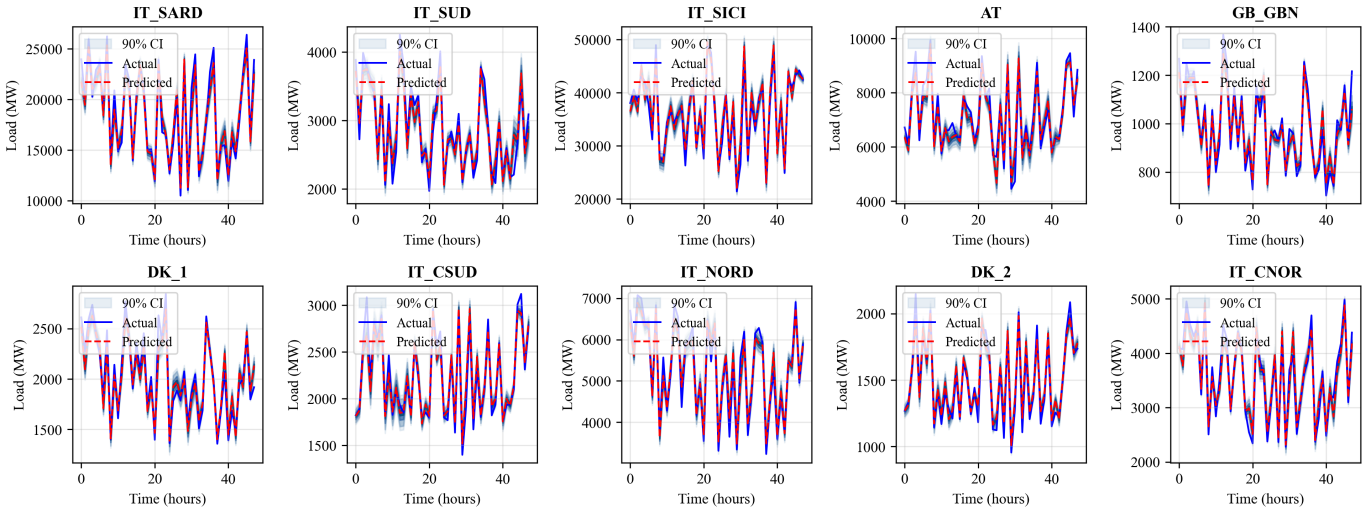


Fig. 6: Regional probabilistic load forecasting with 90% prediction intervals across ten European regions. Blue line shows actual load, red line shows predicted mean, and shaded regions represent 90% confidence intervals constructed as $[\hat{\mu} \pm 1.645\hat{\sigma}]$.

TABLE 2
REGIONAL FORECASTING PERFORMANCE WITH UNCERTAINTY QUANTIFICATION.

| Region | MAE (MW) | RMSE (MW) | $R^2$ | MAPE (%) | Accuracy (%) | PICP 90% | MPIW (MW) |
|---|---|---|---|---|---|---|---|
| IT_SARD | 341.15±8.45 | 473.82±11.74 | 0.9903±0.0008 | 1.96±0.05 | 99.18±0.18 | 0.908±0.015 | 288.7±7.14 |
| IT_SUD | 98.47±2.42 | 151.23±3.72 | 0.9362±0.0094 | 3.54±0.09 | 94.15±0.47 | 0.891±0.018 | 145.2±3.57 |
| IT_SICI | 801.29±9.85 | 1085.14±13.83 | 0.9804±0.0019 | 2.32±0.06 | 98.87±0.22 | 0.901±0.016 | 658.4±16.28 |
| AT | 133.82±3.30 | 192.45±4.77 | 0.9805±0.0019 | 1.97±0.05 | 98.68±0.23 | 0.897±0.017 | 181.3±4.47 |
| GB_GBN | 22.68±0.56 | 33.41±0.82 | 0.9521±0.0047 | 2.31±0.06 | 98.63±0.23 | 0.889±0.018 | 42.1±1.04 |
| DK_1 | 47.21±1.17 | 66.38±1.64 | 0.9721±0.0028 | 2.38±0.06 | 98.37±0.24 | 0.899±0.016 | 74.5±1.83 |
| IT_CSUD | 45.89±1.13 | 83.12±2.05 | 0.9674±0.0032 | 2.06±0.05 | 99.31±0.19 | 0.906±0.015 | 90.8±2.23 |
| IT_NORD | 103.67±2.56 | 145.04±3.58 | 0.9845±0.0015 | 2.03±0.05 | 99.36±0.18 | 0.895±0.016 | 159.2±3.92 |
| DK_2 | 27.34±0.67 | 39.62±0.98 | 0.9832±0.0017 | 1.85±0.05 | 99.39±0.18 | 0.911±0.014 | 46.1±1.13 |
| IT_CNOR | 86.92±2.14 | 118.56±2.93 | 0.9807±0.0019 | 2.52±0.06 | 98.45±0.24 | 0.903±0.016 | 130.4±3.22 |
| **Average** | **170.84±4.23** | **238.88±5.91** | **0.9727±0.0004** | **2.29±0.07** | **98.44±0.21** | **0.900±0.016** | **181.7±4.48** |

operators to procure costly excess reserves, while overly narrow intervals risk shortfalls during volatile conditions. Our calibrated intervals achieve favorable tradeoffs, narrower than naive fixed-percentage buffers (e.g., 10% margins would yield 755.6 MW intervals, 4.2× wider) while maintaining reliable 90% coverage. The operational value of this precision-sharpness balance is quantified in Section 6.5, where GridFusionX achieves 39.5–66.1% cost reductions compared to fixed-buffer and static-uncertainty approaches by dynamically adjusting interval widths based on forecast conditions.

Regional analysis reveals distinct performance patterns depending on demand characteristics and system heterogeneity. Regions with higher demand, such as IT_SICI and IT_SARD, exhibit slightly elevated errors due to scale effects but maintain high $R^2$ values (0.9804±0.0019 and 0.9903±0.0008, respectively), alongside proportionally scaled but well-calibrated prediction intervals (658.4±16.28 MW and 288.7±7.14 MW, representing 1.5% and 1.7% of regional mean loads, respectively). Conversely, regions with lower demand profiles, such as DK_2 or GB_GBN, display near-perfect correlation with minimal deviations and correspondingly narrow prediction intervals (46.1±1.13 MW and 42.1±1.04 MW, representing 2.3% and 3.4% of regional loads), highlighting the model's ability to maintain consistent relative precision across diverse operational scales.

Fig. 6 compares predicted and actual load signals over 48 hours across all regions, confirming the model's ability to capture gradual demand trends as well as abrupt changes, including weekend-to-weekday transitions and short-term fluctuations. Complementarily, Fig. 7 illustrates multi-level prediction intervals (5–95%, 25–75%, 60–40%, 90–10%) over 90 hours, showing that observed values consistently remain within predicted bounds.

### 6.4 Comparative Performance Analysis with Uncertainty Quantification

Table 3 presents forecasting performance across 1-hour, 7-hour, and 24-hour horizons, with all results averaged over 5 independent runs using identical hyperparameter tuning (random search over 20 trials), and graph-based methods (DCRNN, Graph WaveNet, STGCN, ASTGCN, TEST-GCN) using the same network topology $\mathcal{G}$ and modality configuration for fair comparison. GridFusionX achieves 170.84±4.23 MW at the 1-hour horizon, outperforming Graph WaveNet (179.45±4.28 MW) by 4.8%, ASTGCN (180.28±4.35 MW) by 5.2%, STGCN (183.15±4.48 MW) by 6.7%, Multimodal TCN (185.71±4.05 MW) by 8.0%, and

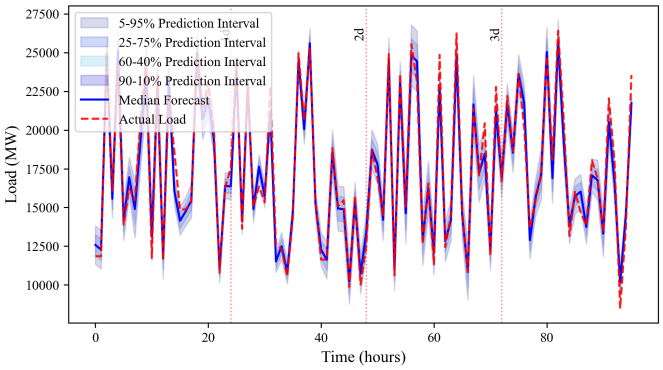


Fig. 7: Multi-level prediction intervals for IT_SARD region demonstrating calibrated uncertainty quantification. The figure shows nested confidence intervals at five levels: 90% $[\hat{\mu} \pm 1.645\hat{\sigma}]$ (outermost), 80% $[\hat{\mu} \pm 1.282\hat{\sigma}]$, 60% $[\hat{\mu} \pm 0.842\hat{\sigma}]$, 50% $[\hat{\mu} \pm 0.674\hat{\sigma}]$, and 25% $[\hat{\mu} \pm 0.319\hat{\sigma}]$ (innermost). The dashed line represents actual load observations.

traditional LSTM (387.43±9.12 MW) by 55.9%. While spatiotemporal baselines (STGCN, ASTGCN) effectively model spatial dependencies through graph convolutions, they produce only deterministic forecasts. ASTGCN's attention mechanism improves over basic STGCN by 1.6%, demonstrating the value of adaptive spatial-temporal weighting, yet both lack the uncertainty quantification and precision-weighted multimodal fusion that enable GridFusionX's superior performance.

Furthermore, at 7-hour predictions, GridFusionX (224.73±5.34 MW) pulls ahead of Graph WaveNet by 17.6% and ASTGCN by 18.2%, while at 24-hour horizons the improvements reach 13.9% and 14.8% respectively. The performance gap widens at longer horizons, highlighting GridFusionX's advantage in uncertainty propagation across temporal-spatial sequences. Interestingly, TEST-GCN's performance relative to GridFusionX deteriorates at longer horizons despite its sophisticated topological feature learning, suggesting that road network characteristics do not translate directly to power system forecasting, where uncertainty propagation becomes the dominant challenge. The consistent performance degradation patterns across all models reflect the fundamental difficulty of long-term prediction, yet GridFusionX maintains $R^2$ values above 0.94 across all horizons, indicating robust trend alignment despite increasing temporal distance.

Beyond raw error metrics, practical reliability tells a compelling story. GridFusionX achieves 98.44±0.21% accuracy (predictions within 10% relative error) at 1-hour horizon, improving over Graph WaveNet's 97.41±0.28% and ASTGCN's 97.35±0.28%. However, in grid operations managing tens of gigawatts across interconnected regions, this translates to dozens of additional correct predictions per day, directly reducing unnecessary reserve deployments and enhancing system stability margins. The real operational value emerges from GridFusionX's uncertainty quantification: unlike deterministic baselines (including STGCN and ASTGCN) that provide only single-point predictions, GridFusionX delivers calibrated 90% prediction intervals (PICP of 90.0±1.6%) that enable risk-informed decision-making. The precision-weighted fusion mechanism maintains forecast integrity by automatically down-weighting unreliable modalities during sensor malfunctions or degraded data quality, adapting interval widths to operational conditions.

### 6.5 Operational Value Assessment

To demonstrate practical utility beyond forecast accuracy, we evaluate reserve sizing decisions, which are a core operational challenge where grid operators must balance procurement costs against shortage risks. The optimal reserve capacity minimizes expected operational cost:

$$C = c_{res} \cdot \sum_i (R_i - \hat{\mu}_i) + c_{pen} \cdot \sum_i \max(0, y_i - R_i), \quad (80)$$

where reserves $R_i = \hat{\mu}_i + z_\alpha \hat{\sigma}_i$ are sized at quantile $\alpha$ (we use $\alpha = 0.90$ with $z_{0.90} = 1.282$), $c_{res} = \$50/\text{MWh}$ represents reserve capacity costs, and $c_{pen} = \$500/\text{MWh}$ represents shortage penalties reflecting typical European value of lost load. **Illustrative scenarios from IT_SARD region:** Two contrasting operating conditions illustrate the value of adaptive uncertainty quantification. During stable overnight conditions (hour 24 in Fig. 7), GridFusionX predicts $\hat{\mu} = 17{,}200$ MW with low uncertainty $\hat{\sigma} = 456$ MW for an actual demand of 17,500 MW, yielding reserve capacity $R = 17{,}784$ MW and procurement cost of $29,200, with no shortage penalty. In contrast, a fixed 10% buffer sets $R = 18{,}920$ MW regardless of conditions, wasting $86,000 on unnecessary reserves, while a deterministic forecast with static empirical uncertainty ($\hat{\sigma} = 750$ MW) incurs $48,250. By recognizing stable conditions and tightening confidence bounds, GridFusionX achieves 66.1% cost savings over fixed buffers and 39.5% over empirical approaches. The advantage persists under volatile afternoon peak conditions (hour 48), when demand rises to 25,000 MW due to synchronized

TABLE 3
PERFORMANCE COMPARISON OF DIFFERENT MODELS OVER VARIOUS PREDICTION HORIZONS.

| Model | 1-hour ahead | | | | 7-hour ahead | | | | 24-hour ahead | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAE (MW) | $R^2$ | MAPE (%) | Accuracy (%) | MAE (MW) | $R^2$ | MAPE (%) | Accuracy (%) | MAE (MW) | $R^2$ | MAPE (%) | Accuracy (%) |
| LSTM | 387.43±9.12 | 0.8735±0.013 | 5.98±0.19 | 91.72±0.44 | 491.28±10.67 | 0.8512±0.016 | 8.24±0.22 | 87.45±0.51 | 545.67±13.21 | 0.8155±0.019 | 9.63±0.26 | 84.48±0.53 |
| DCRNN | 315.82±7.45 | 0.9136±0.010 | 4.58±0.15 | 94.05±0.37 | 403.19±9.34 | 0.8714±0.013 | 6.96±0.18 | 90.08±0.43 | 456.34±10.89 | 0.8438±0.015 | 8.22±0.21 | 88.77±0.47 |
| Multimodal DCRNN | 189.56±5.12 | 0.9518±0.007 | 2.92±0.10 | 96.71±0.29 | 302.38±7.15 | 0.9414±0.009 | 5.05±0.14 | 94.72±0.34 | 324.87±7.89 | 0.9194±0.011 | 5.83±0.16 | 94.00±0.37 |
| Multimodal TCN | 185.71±4.05 | 0.9583±0.006 | 2.56±0.09 | 97.08±0.26 | 281.24±6.21 | 0.9422±0.008 | 4.42±0.12 | 95.53±0.31 | 288.19±6.91 | 0.9211±0.010 | 6.05±0.15 | 93.74±0.34 |
| Graph WaveNet | 179.45±4.28 | 0.9606±0.007 | 2.43±0.10 | 97.41±0.28 | 272.68±6.52 | 0.9465±0.009 | 4.18±0.13 | 95.79±0.33 | 278.92±7.25 | 0.9249±0.011 | 5.71±0.16 | 94.14±0.36 |
| STGCN | 183.15±4.48 | 0.9578±0.007 | 2.51±0.11 | 97.15±0.29 | 279.42±6.85 | 0.9438±0.009 | 4.35±0.14 | 95.48±0.33 | 286.73±7.38 | 0.9218±0.011 | 5.92±0.17 | 93.81±0.36 |
| ASTGCN | 180.28±4.35 | 0.9598±0.007 | 2.45±0.10 | 97.35±0.28 | 274.91±6.71 | 0.9456±0.009 | 4.24±0.13 | 95.68±0.32 | 281.54±7.29 | 0.9237±0.011 | 5.78±0.16 | 94.02±0.35 |
| TEST-GCN | 181.23±4.56 | 0.9592±0.008 | 2.47±0.11 | 97.24±0.30 | 276.15±6.78 | 0.9451±0.010 | 4.27±0.14 | 95.57±0.34 | 283.46±7.53 | 0.9231±0.012 | 5.85±0.17 | 93.92±0.38 |
| **GridFusionX** | **170.84±4.23** | **0.9727±0.004** | **2.29±0.07** | **98.44±0.21** | **224.73±5.34** | **0.9638±0.006** | **3.81±0.10** | **96.55±0.27** | **239.86±5.89** | **0.9479±0.008** | **4.38±0.12** | **95.59±0.30** |

solar ramping and industrial load. GridFusionX predicts $\hat{\mu} = 24{,}800$ MW and adapts uncertainty to $\hat{\sigma} = 912$ MW, producing $R = 25{,}969$ MW and \$58,450 in procurement cost while avoiding shortages. The fixed buffer over-procures to $R = 27{,}280$ MW, incurring \$124,000, whereas static empirical uncertainty underestimates risk ($R = 25{,}761$ MW), exposing the system to shortage penalties that can exceed reserve costs. By widening intervals only when risk increases, GridFusionX balances over-procurement and shortage avoidance in a cost-aware manner.

### 6.6 Ablation Study

To rigorously assess the contribution of each architectural component, we perform ablation studies by systematically disabling key mechanisms within GridFusionX while keeping the training protocol unchanged. Table 4 reports the 1-hour ahead forecasting performance of four ablated variants, averaged across all ten European regions, enabling a controlled comparison of how uncertainty modeling, topology awareness, and adaptive attention each contribute to overall predictive accuracy and robustness.

TABLE 4
ABLATION STUDY RESULTS FOR 1-HOUR AHEAD FORECASTING AVERAGED ACROSS ALL TEN REGIONS.

| Model Variant | MAE (MW) | $R^2$ | MAPE (%) | PICP 90% |
|---|---|---|---|---|
| GridFusionX (Full) | 170.84±4.23 | 0.9727±0.004 | 2.29±0.07 | 0.900±0.016 |
| No Precision Weighting | 194.15±4.81 | 0.9638±0.006 | 2.61±0.08 | 0.862±0.021 |
| No Dual-Head Encoder | 183.47±4.54 | 0.9681±0.005 | 2.45±0.07 | 0.751±0.029 |
| No Cross-Modal Attention | 198.73±4.92 | 0.9619±0.007 | 2.66±0.08 | 0.887±0.018 |
| Single Modality (Load Only) | 212.38±5.26 | 0.9571±0.008 | 2.85±0.09 | 0.893±0.017 |

The ablation results confirm that each architectural component is essential to GridFusionX's performance. Replacing inverse-variance weighting with uniform averaging (without precision weighting) increases MAE by 13.7% to 194.15 MW (an additional 23.31 MW error) and reduces PICP to 86.2%, showing that equal treatment of modalities during degraded or high-uncertainty conditions leads to inaccurate and poorly calibrated forecasts. Removing the dual-head encoder (without dual-head encoder) and estimating uncertainty post-hoc further degrades calibration, with PICP dropping by 14.9% to 75.1% despite moderate point accuracy (MAE = 183.47 MW, only 7.4% worse than full model), confirming that explicit uncertainty pathways are required for well-calibrated probabilistic forecasts and supporting Theorem 4.1. This represents the most severe calibration failure, with roughly 1 in 4 observations falling outside prediction intervals. Eliminating uncertainty-guided cross-modal attention (without cross-modal attention) increases MAE by 16.3% to 198.73 MW (a 27.89 MW degradation), demonstrating the importance of dynamically reweighting modalities based on information content and reliability. Notably, this variant maintains PICP at 88.7% (only 1.3% below full model), suggesting that while uncertainty quantification remains partially effective, accurate point estimates require adaptive cross-modal integration. Lastly, the single-modality variant using only load history incurs a 24.3% error increase to 212.38 MW (41.54 MW worse, nearly 2.5 times the full model's error), while $R^2$ drops by 1.56% to 0.9571, highlighting the complementary value of renewable and price data. Despite this degraded point accuracy, PICP remains at 89.3%, only 0.7% below the full model, indicating that uncertainty estimation can partially compensate for reduced information by widening intervals. The results show that GridFusionX's gains arise from the synergistic integration of all three components, with uncertainty modeling being most critical for calibration.

## 7 CONCLUSION

This paper introduced GridFusionX, a network-aware uncertainty quantification framework for spatially distributed smart grid forecasting. The framework represents interconnected regions as nodes within a multimodal transformer architecture that integrates historical load, renewable generation, and market price data into a unified latent representation. By combining precision-weighted fusion mechanisms with probabilistic decoders, GridFusionX delivers both accurate predictions and calibrated confidence intervals that are essential for coordinated operation across modern power networks. The resulting probabilistic forecasts can be directly consumed by downstream decision modules, for example unit commitment and reserve allocation solvers, to balance economic efficiency with reliability under high penetration of variable renewables. Validation on ten European regions demonstrated substantial improvements, with mean absolute error reductions between 4.8% and 55.9% compared to region-independent baselines, achieving 98.4±0.2% reliability and 90.0±1.6% prediction interval coverage while preserving spatial consistency across geographically coupled areas. These results indicate that explicitly coupling network topology with calibrated uncertainty modeling provides a practical foundation for next-generation, risk-aware power system operations.

**Limitations and Future Directions:** The attention mechanism scales quadratically with network size, potentially requiring approximation techniques for continental-scale grids, such as sparse kernels or graph coarsening strategies that respect physical connectivity. The framework assumes a stationary topology and synchronized multimodal measurements, whereas real-world systems experience line outages, topology reconfiguration, and occasional sensor misalignment that may violate these assumptions. Theoretical guarantees are derived under Gaussian assumptions, which work well for the studied scenarios but may need extensions for highly non-Gaussian distributions and heavy-tailed error patterns during stressed operating conditions. Promising extensions include hierarchical architectures for scalability across voltage levels, online learning for networks with time-varying topology, robust fusion under missing or corrupted data, and non-Gaussian uncertainty quantification for skewed and heavy-tailed risks. Further validation on extreme events such as heat waves, large renewable forecast errors, and cascading failures would clarify the reliability of the framework in rare but critical regimes and strengthen GridFusionX as a flexible foundation for network-aware probabilistic forecasting in distributed power systems. Integrating GridFusionX with closed-loop control and market-clearing mechanisms represents a natural next step toward fully risk-aware grid operation.

## REFERENCES


[1] C. Huang, S. Bu, W. Chen, H. Wang, and Y. Zhang, "Deep reinforcement learning-assisted federated learning for robust short-term load forecasting in electricity wholesale markets," *IEEE Trans. Netw. Sci. Eng.*, vol. 11, no. 5, pp. 5073–5086, Sep.-Oct. 2024.

[2] S. Chen, R. Lin, and W. Zeng, "Short-term load forecasting method based on ARIMA and LSTM," in *Proc. IEEE 22nd Int. Conf. Commun. Technol. (ICCT)*, Nanjing, China, 2022, pp. 1913–1917.

[3] P. Peram and K. Narayanan, "Diffusion convolutional recurrent neural network-based load forecasting during COVID-19 pandemic situation," *Rev. Intell. Artif.*, vol. 36, no. 5, pp. 689–695, Oct. 2022.

[4] X. Fang, W. Zhang, Y. Guo, J. Wang, M. Wang, and S. Li, "A novel reinforced deep RNN–LSTM algorithm: Energy management forecasting case study," *IEEE Trans. Ind. Informat.*, vol. 18, no. 8, pp. 5698–5704, Aug. 2022.

[5] Z. Xu, Z. Yu, H. Zhang, J. Chen, J. Gu, T. Lukasiewicz, and V. C. M. Leung, "PhaCIA-TCNs: Short-term load forecasting using temporal convolutional networks with parallel hybrid activated convolution and input attention," *IEEE Trans. Netw. Sci. Eng.*, vol. 11, no. 1, pp. 427–438, Jan. 2024.

[6] H. Zhao, Y. Wu, L. Ma, and S. Pan, "Spatial and temporal attention-enabled transformer network for multivariate short-term residential load forecasting," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–11, Aug. 2023.

[7] Q. Hong, F. Meng, and F. Maldonado, "Advancing long-term multi-energy load forecasting with Patchformer: A patch and transformer-based approach," 2024. [Online]. Available: https://arxiv.org/abs/2404.10458

[8] X. Piao, Z. Chen, T. Murayama, Y. Matsubara, and Y. Sakurai, "Fredformer: Frequency debiased transformer for time series forecasting," in *Proc. ACM SIGKDD Conf. Knowl. Discov. Data Mining*. New York, NY, USA: Association for Computing Machinery, 2024, p. 2400–2410.

[9] Y. Jiang, Y. Dai, R. Si, J. Chen, T. Gao, and J. Zhang, "Short-term state electricity load forecasting based on transfer-informer," in *Proc. IEEE 2nd Int. Conf. Digit. Twins Parallel Intell. (DTPI)*, Boston, MA, USA, 2022, pp. 1–6.

[10] T. Jing, S. Chen, D. Navarro-Alarcon, Y. Chu, and M. Li, "SolarFusionNet: Enhanced solar irradiance forecasting via automated multi-modal feature selection and cross-modal fusion," *IEEE Trans. Sustain. Energy*, vol. 16, no. 2, pp. 761–773, Apr. 2025.

[11] K. Wang, S. Shan, W. Dou, H. Wei, and K. Zhang, "A robust photovoltaic power forecasting method based on multimodal learning using satellite images and time series," *IEEE Trans. Sustain. Energy*, vol. 16, no. 2, pp. 970–980, Apr. 2025.

[12] K.-B. Song, Y.-S. Baek, D. H. Hong, and G. Jang, "Short-term load forecasting for the holidays using fuzzy linear regression method," *IEEE Trans. Power Syst.*, vol. 20, no. 1, pp. 96–101, Feb. 2005.

[13] I. A. Samuel, E. Adetiba, I. A. Odigwe, and F. C. Felly-Njoku, "A comparative study of regression analysis and artificial neural network methods for medium-term load forecasting," *Indian J. Sci. Technol.*, vol. 10, no. 10, pp. 1–7, Mar. 2017.

[14] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on LSTM recurrent neural network," *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 841–851, Jan. 2019.

[15] J. Zheng, C. Xu, Z. Zhang, and X. Li, "Electric load forecasting in smart grids using long-short-term-memory based recurrent neural network," in *Proc. 51st Annu. Conf. Inf. Sci. Syst. (CISS)*. Baltimore, Maryland, USA: IEEE, 2017, pp. 1–6.

[16] S. Bouktif, A. Fiaz, A. Ouni, and M. A. Serhani, "Optimal deep learning LSTM model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches," *Energies*, vol. 11, no. 7, p. 1636, Jun. 2018.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Adv. 31st Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 5998–6008.

[18] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proc. 35th AAAI Conf. Artif. Intell.*, Virtual, 2021, pp. 11 106–11 115.

[19] P. Lara-Benítez, M. Carranza-García, J. M. Luna-Romera, and J. C. Riquelme, "Temporal convolutional networks applied to energy-related time series forecasting," *Appl. Sci.*, vol. 10, no. 7, p. 2322, Mar. 2020.

[20] J. W. Chan and C. K. Yeo, "A transformer based approach to electricity load forecasting," *Electr. J.*, vol. 37, no. 2, p. 107370, Mar. 2024.

[21] J. Wang, D. Xu, Y. Li, M. Shahidehpour, and T. Yang, "Transformer-based probabilistic demand forecasting with adaptive online learning," *Electr. Power Syst. Res.*, vol. 240, p. 111255, Mar. 2025.

[22] W. Zhang, H. Zhan, H. Sun, and M. Yang, "Probabilistic load forecasting for integrated energy systems based on quantile regression patch time series transformer," *Energy Rep.*, vol. 13, pp. 303–317, Dec. 2025.

[23] H. Jiang, S. Pan, Y. Dong, and J. Wang, "Probabilistic electricity price forecasting based on penalized temporal fusion transformer," *J. Forecasting*, vol. 43, no. 5, pp. 1465–1491, Aug. 2024.

[24] Z. Wang, Q. Wen, C. Zhang, L. Sun, and Y. Wang, "DiffLoad: Uncertainty quantification in electrical load forecasting with the diffusion model," *IEEE Trans. Power Syst.*, vol. 40, no. 2, pp. 1777–1789, Mar. 2025.

[25] A. Dairi, F. Harrou, B. Khaldi, and Y. Sun, "Graph neural networks-based spatiotemporal prediction of photovoltaic power: A comparative study," *Neural Comput. Appl.*, vol. 37, pp. 4769–4795, Dec. 2024.

[26] Y. Yang, Y. Liu, Y. Zhang, S. Shu, and J. Zheng, "DEST-GNN: A double-explored spatio-temporal graph neural network for multi-site intra-hour PV power forecasting," *Appl. Energy*, vol. 378, p. 124744, Jan. 2025.

[27] Y. Su, M. Zhang, L. Cao, Y. Chen, and Y. Tian, "Spatio-temporal graph neural network with Fourier features for multi-site photovoltaic power forecasting," *Electr. Power Syst. Res.*, vol. 251, p. 112171, Feb. 2026.

[28] C. Wei, D. Pi, M. Ping, and H. Zhang, "Short-term load forecasting using spatial-temporal embedding graph neural network," *Electr. Power Syst. Res.*, vol. 225, p. 109873, Dec. 2023.

[29] J. Liu, H. Zang, L. Cheng, T. Ding, Z. Wei, and G. Sun, "A transformer-based multimodal-learning framework using sky images for ultra-short-term solar irradiance forecasting," *Applied Energy*, vol. 342, p. 121160, 2023.

[30] Z. Kong, C. Zhang, H. Lv, F. Xiong, and Z. Fu, "Multimodal feature extraction and fusion deep neural networks for short-term load forecasting," *IEEE Access*, vol. 8, pp. 185 373–185 383, Oct. 2020.

[31] D. Aguilar, J. J. Quinones, L. R. Pineda, J. Ostanek, and L. Castillo, "Optimal scheduling of renew. energy microgrids: A robust multi-objective approach with machine learning-based probabilistic forecasting," *Appl. Energy*, vol. 369, p. 123548, Sep. 2024.

[32] J. Hu, Y. Shan, Y. Yang, A. Parisio, Y. Li, N. Amjady, S. Islam, K. W. Cheng, J. M. Guerrero, and J. Rodríguez, "Economic model predictive control for microgrid optimization: A review," *IEEE Trans. Smart Grid*, vol. 15, no. 1, pp. 472–484, Jan. 2024.

[33] B. H. Vu and I.-Y. Chung, "Optimal generation scheduling and operating reserve management for PV generation using RNN-based forecasting models for stand-alone microgrids," *Renew. Energy*, vol. 195, pp. 1137–1154, Aug. 2022.

[34] H. Mansoor, M. S. Gull, H. Rauf, I. ul Hasan Shaikh, M. Khalid, and N. Arshad, "Graph convolutional networks based short-term load forecasting: Leveraging spatial information for improved accuracy," *Electr. Power Syst. Res.*, vol. 230, p. 110263, May 2024.

[35] Y. Huang, S. Wu, Z. Wang, X. Liu, C. Li, and Y. Hu, "Causality-aware multi-graph convolutional networks with critical node dynamics for electric vehicle charging station load forecasting," *IEEE Trans. Smart Grid*, vol. 16, no. 4, pp. 3210–3225, Jul. 2025.

[36] C. Wang, Y. Wang, Z. Ding, and K. Zhang, "Probabilistic multi-energy load forecasting for integrated energy system based on bayesian transformer network," *IEEE Trans. Smart Grid*, vol. 15, no. 2, pp. 1495–1508, Mar. 2024.

[37] D. Tan, Z. Tang, F. Zhou, and Y. Xie, "A novel hybrid model based on EMD-improved TCN-improved TST for short-term railway traction load forecasting," *IEEE Trans. Transp. Electrific.*, vol. 11, no. 2, pp. 6418–6427, Apr. 2025.

[38] Y. Zhang, R. Wu, S. M. Dascalu, and F. C. Harris, "Multi-scale transformer pyramid networks for multivariate time series forecasting," *IEEE Access*, vol. 12, pp. 14 731–14 741, Jan. 2024.

[39] S. Huang, Y. Guo, P. Huo, and Q. Li, "Redefining vibration sensing: AI-driven analytics, self-powered systems, and multi-modal fusion—A review," *IEEE Sensors J.*, vol. 25, no. 15, pp. 27 922–27 941, Aug. 2025.

[40] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 3, pp. 463–482, Mar. 2003.

[41] O. Bousquet, S. Boucheron, and G. Lugosi, "Introduction to statistical learning theory," in *Advanced Lectures on Machine Learning*, O. Bousquet, U. von Luxburg, and G. Rätsch, Eds. Berlin, Germany: Springer, 2004, pp. 169–207.

[42] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford, U.K.: Oxford Univ. Press, 2013.

[43] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[44] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. New York, NY, USA: Springer, 2002.

[45] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, ser. Springer Ser. Statist. New York, NY, USA: Springer, 2009.

[46] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[47] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[48] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *Int. J. Forecasting*, vol. 32, no. 3, pp. 914–938, Jul. 2016.

[49] H. Wang, Z. Lei, X. Zhang, B. Zhou, and J. Peng, "A review of deep learning for renew. energy forecasting," *Energy Convers. Manage.*, vol. 198, p. 111799, Oct. 2019.

[50] J. L. W. V. Jensen, "Sur les fonctions convexes et les inégalités entre les valeurs moyennes [on convex functions and inequalities between mean values]," *Acta Math.*, vol. 30, no. 1, pp. 175–193, Dec. 1906.

[51] W. H. Greene, *Econometric Analysis*, 5th ed. Upper Saddle River, NJ, USA: Prentice Hall, 2003.

[52] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2015, pp. 1–15.

[53] O. P. S. Data, "Time series: Hourly data of load, wind, and solar generation, and spot prices for European countries," https://data.open-power-system-data.org/time_series/, 2020, accessed on 4/5/2025.

[54] M. S. Hossain and H. Mahmood, "Short-term load forecasting using an LSTM neural network," in *Proc. IEEE Power Energy Conf. (PECI)*, Illinois, USA, 2020, pp. 1–6.

[55] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, Vancouver, BC, Canada, 2018, pp. 1–16.

[56] Z. Wu, S. Pan, G. Long, J. Jiang, P. Chang, and C. Zhang, "Graph wavenet for deep spatial-temporal graph modeling," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, Stockholm, Sweden, 2019, pp. 1907–1913.

[57] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," in *Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI)*, Stockholm, Sweden, 2018, pp. 3634–3640.

[58] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention based spatial-temporal graph convolutional networks for traffic flow forecasting," in *Proc. 33rd AAAI Conf. Artif. Intell. (AAAI)*, Honolulu, HI, USA, 2019, pp. 922–929.

[59] M. A. Ali, S. Venkatesan, V. Liang, and H. Kruppa, "TEST-GCN: Topologically enhanced spatial-temporal graph convolutional networks for traffic forecasting," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Virtual, 2021, pp. 982–987.

[60] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, no. 10, pp. 281–305, Feb. 2012.

**Quoc Bao Phan** (Graduate Student Member, IEEE) received the MS. degree in Informatics from Northern Arizona University, Flagstaff, AZ, USA, in 2025. He is currently pursuing a Ph.D. degree in Electrical Engineering at the College of Engineering, Florida State University, Tallahassee, FL, USA. His research interests include electrical system modeling, homomorphic encryption, and artificial intelligence.

**Abdulrahman Takiddin** (Member, IEEE) received the B.Sc. degree (Hons.) in information systems from Carnegie Mellon University, Pittsburgh, PA, USA, in 2014, the M.Sc. degree in data analytics from Hamad Bin Khalifa University, Doha, Qatar, in 2020, and the Ph.D. degree in electrical engineering from Texas A&M University, College Station, TX, USA, in 2023. He is currently an Assistant Professor of Electrical and Computer Engineering with the FAMU-FSU College of Engineering, Florida State University, Tallahassee, FL, USA. His research interests include machine learning, cyber-physical systems, smart grid, smart transportation, and security.

**Gelli Ravikumar** (Senior Member, IEEE) is an Associate Professor in the Department of Electrical and Computer Engineering at Florida State University and an Affiliate Faculty member at Iowa State University. He joined FSU in 2025 after serving as a Research Assistant Professor (2020–2025) and Postdoctoral Researcher (2018–2020) at Iowa State University, following a postdoctoral appointment at New Mexico State University (2017–2018). He received the M.Tech. and Ph.D. degrees in Electrical Engineering from the Indian Institute of Technology Bombay in 2011 and 2016, respectively. Dr. Gelli's research focuses on AI-native cyber-physical energy systems, with an emphasis on distributed energy resource integration, grid automation, and the security and resilience of modern power infrastructure. His work advances intelligent and autonomous grid operation through machine learning, deep reinforcement learning, and emerging paradigms such as large language models, agentic AI systems, and grid foundation models. His research integrates data-driven intelligence with physics-based modeling for scalable, secure, and real-time power system applications, with a focus on trustworthy and explainable AI for critical infrastructure.

**Olugbenga Moses Anubi** (Senior Member, IEEE) is currently an Associate Professor of Electrical and Computer Engineering at the FAMU-FSU College of Engineering. Prior to that, he was a lead control systems engineer at the GE Global Research Center, NY. His work has resulted in more than 15 patents and several recognitions, including the 2023 FAMU-FSU College of Engineering Faculty Rising Star Award, the GE Technology Award (Physical+Digital), the Connected Controls Technical Achievement Award, the Whitney Award, and the Dushman Technology Award. His research interests include control of autonomous systems and resilient cyber-physical systems with applications to energy, transportation and other critical infrastructures. He is an inducted senior member of the National Academy of Inventors (NAI).

**Tuy Tan Nguyen** (Senior Member, IEEE) received the M.S. and Ph.D. degrees in Information and Communication Engineering from Inha University, South Korea, in 2016 and 2019, respectively. He is currently an Assistant Professor in the Department of Electrical and Computer Engineering, FAMU-FSU College of Engineering, Florida State University. From August 2022 to August 2025, he served as an Assistant Professor in the School of Informatics, Computing, and Cyber Systems, Northern Arizona University. Previously, he worked as a Senior Research Engineer at Conextt Inc., and as a Postdoctoral Fellow in the Department of Electrical and Computer Engineering, Inha University. Dr. Nguyen is a Technical Committee Member of the IEEE Circuits and Systems Society - Circuits and Systems for Communications and an active member of the IEEE Communications Society. His research interests include post-quantum cryptography, homomorphic encryption, error correction codes, and applied artificial intelligence.