Anomaly Detection and Deterministic–Probabilistic Forecasting for Reliable Energy Time-Series Modeling

Cyrine Berrima

A Thesis

in

The Department

of

Concordia Institute for Information Systems Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science (Quality Systems Engineering) at

Concordia University

Montréal, Québec, Canada

April 2026

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Cyrine Berrima**

Entitled: **Anomaly Detection and Deterministic–Probabilistic Forecasting for Reliable Energy Time-Series Modeling**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Quality Systems Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Abdessamad Ben Hamza

_____ Examiner
Dr. Abdessamad Ben Hamza

_____ Examiner
Dr. Honghao Fu

_____ Supervisor
Dr. Manar Amayri

Approved by _____
Chun Wang, Chair
Department of Concordia Institute for Information Systems Engineering

_____ 2026

_____
Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Anomaly Detection and Deterministic–Probabilistic Forecasting for Reliable Energy Time-Series Modeling

Cyrine Berrima

Energy time series sit at the heart of monitoring, forecasting, and control in smart buildings and power systems. In real deployments, however, the data rarely match the assumptions of clean and stationary signals. Sensor faults, missing readings, and communication interruptions introduce abnormal observations, while demand patterns drift with weather, occupancy, and equipment operation. These issues can quietly erode predictive performance and, when uncertainty is not made explicit, can also mask risk in decision-making.

This thesis improves reliability through two connected contributions. First, it investigates unsupervised anomaly detection for energy consumption signals. The approach learns typical temporal behavior from clean data and flags departures without relying on labeled anomalies. Beyond detection metrics, the contribution is evaluated through practical impact on forecasting: we quantify how correcting anomalous segments changes downstream prediction errors and whether it stabilizes performance under data corruption.

The second line of work focuses on probabilistic load forecasting. Rather than producing a single point forecast, the aim is to produce a predictive distribution that accounts for variability and uncertainty. The evaluation pairs accuracy measures with an assessment of calibration and how well uncertainty estimates hold up under changing conditions.

In combination, these contributions emphasize that performance in real deployments depends on both signal integrity and uncertainty quantification. Anomaly handling reduces the impact of corrupted observations, while probabilistic forecasting better supports operation under variability and drift.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Dr. Manar Amayri, for her guidance, support, and constant encouragement throughout this research. Her expertise, careful feedback, and high standards significantly strengthened both the direction and the quality of this thesis.

I am also deeply grateful to Viet Tra for his collaboration and generosity in sharing ideas and research efforts. Our complementary work and discussions contributed to the development of several directions and insights reflected in this thesis.

Finally, I would like to thank my family and friends for their unwavering support, patience, and motivation, especially during challenging periods. This work would not have been possible without you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem statement

Smart buildings and power systems continuously generate electricity consumption time series through metering and monitoring systems. These data support performance analysis, energy efficiency assessment, and load forecasting. In real deployments, however, reliability is frequently compromised. Measurement streams may contain abnormal values arising from sensor faults, data transmission issues, equipment malfunction, or unexpected operational behaviour. Even if such events occur infrequently, they can influence how models learn normal patterns and ultimately degrade predictive performance.

Second, electricity demand is inherently uncertain and evolves over time due to weather, occupancy, operational schedules, and gradual behavioural or structural changes. As a result, forecasting systems must account not only for expected demand but also for uncertainty under non-stationary conditions.

The core problem addressed in this thesis is therefore the following: how can we build reliable monitoring and forecasting methods for univariate electricity load series when the data contain anomalies and the underlying demand patterns change over time. Current approaches often treat anomaly detection and forecasting separately, and many forecasting models focus on point predictions without providing calibrated uncertainty estimates. Moreover, deep learning models may remain sensitive to corrupted observations and can become unstable when trained on imperfect data.

This thesis addresses the problem by (i) developing an unsupervised anomaly detection approach that learns normal behaviour from clean data and identifies deviations without labeled anomalies, and (ii) proposing probabilistic forecasting methods that output predictive distributions and remain informative under changing operating conditions.

1.2 Contributions

This thesis looks at two related problems in smart building energy analysis: how to detect anomalies in electricity load data and how to forecast future consumption while taking uncertainty into account. The work focuses on univariate load series and builds models that reflect their main features, including seasonality, changing demand patterns, and occasional irregular behaviour. The aim is to make monitoring and forecasting more reliable in practical energy management settings.

- **Refined Gated WaveNet for Unsupervised Anomaly Detection**

  The first contribution designs an anomaly detector for electricity load time series by adapting WaveNet to the constraints of energy monitoring. Because the original WaveNet targets discrete audio generation, it cannot be applied directly to real-valued load signals. We therefore reformulate WaveNet as a window-based regression model that predicts the next load value from past observations, and uses the prediction error as an anomaly score. The model keeps the key WaveNet mechanisms that support multi-scale temporal learning (dilated causal convolutions, gated activations, and residual/skip connections), but replaces sample-by-sample autoregressive generation with parallel prediction over windows for efficient training and inference. To strengthen sensitivity to slow or long-range deviations, we add a compact temporal refinement module (GRU or 1D CNN). Robustness to outliers is further improved using Smooth L1 loss. Finally, performance is evaluated using a controlled synthetic anomaly injection protocol that ensures clean training and reproducible comparisons across baselines. A manuscript based on this contribution has been accepted for publication in *Sustainable Energy, Grids and Networks*.

- **Impact of Anomalies on Forecasting Robustness and Joint Monitoring**

This thesis further investigates how anomalies affect forecasting stability. The analysis shows that even limited abnormal behaviour can introduce substantial prediction bias and error propagation in multi-horizon forecasting. Integrating anomaly detection directly into the modelling pipeline improves forecasting consistency and robustness. These findings highlight that anomaly detection should be treated as an integral component of energy forecasting systems rather than a separate preprocessing stage. While anomaly handling improves robustness, forecasting performance also depends on the model's ability to represent intrinsic demand variability. This leads to the third contribution, which focuses on uncertainty-aware forecasting.

- **FiLM-Conditioned Transformer VAE for Determinsitic and Probabilistic Load Forecasting**

  The third contribution proposes a probabilistic forecasting framework combining a Transformer architecture with a variational latent component. This design enables uncertainty-aware multi-horizon load prediction from a single time series. Temporal information, such as calendar and cyclical patterns, is incorporated using FiLM based modulation of hidden representations, allowing the attention mechanism to adapt to recurring operational regimes. The variational component captures latent sources of variability and enables predictive distributions, while the same model also provides strong point forecasts. Stabilization strategies, including KL annealing and constrained variance parameterization, ensure reliable training. Experimental evaluation demonstrates improvements in both deterministic accuracy and uncertainty calibration compared to classical neural models and recent deep learning baselines.

1.3 Thesis Overview

This thesis is divided into five chapters:

- Chapter 1 sets the context using energy consumption data from smart buildings, where measurements are often affected by noise, faults, and changes in operating conditions. It ends by stating the research questions and the contributions of the thesis.

- Chapter 2 brings together the background needed for the rest of the manuscript. It reviews unsupervised anomaly detection in time series and convolution-based temporal models (in particular, dilated causal convolutions in the spirit of WaveNet), then summarizes the probabilistic forecasting concepts used later, including attention-based architectures and variational inference. The chapter also clarifies the experimental setup: datasets, preprocessing choices, train/test splits, and evaluation criteria.

- Chapter 3 is devoted to anomaly detection. It describes the proposed Refined Gated WaveNet and the way it is trained using normal data only. The chapter then details how anomaly scores are produced and how detection performance is assessed. A controlled anomaly injection protocol is also presented to support reproducible benchmarking. A manuscript based on this contribution has been accepted for publication in *Sustainable Energy, Grids and Networks*. (`ScienceDirect`).

- Chapter 4 addresses probabilistic load forecasting. It introduces the FiLM-conditioned Transformer VAE and describes how temporal information is injected through conditioning. Results are reported on multiple datasets, with emphasis on both forecasting accuracy and uncertainty quality (calibration and coverage), which are central to the forecasting contribution. A journal manuscript based on this contribution has been submitted to *Applied Intelligence* (Springer Nature).

- Chapter 5 summarizes the main findings, points out the main limitations, and suggests a few extensions, including multivariate settings and a closer connection between anomaly handling and forecasting.

Chapter 2

Theoretical Foundations and related work

2.1 Theoretical Foundations of Anomaly Detection and Probabilistic Forecasting

This section outlines the main theoretical concepts that support the methods developed in this thesis. It introduces the formulation of unsupervised anomaly detection in energy time series and then presents the modeling tools used in the proposed approaches, including dilated convolutions, attention mechanisms, and variational inference. These elements form the basis for both anomaly detection and probabilistic forecasting.

2.1.1 Unsupervised Anomaly Detection in Energy Time Series

Electricity consumption in smart buildings can be modeled as a univariate time series $\{y_t\}_{t=1}^{T}$. In unsupervised anomaly detection, the objective is to learn the normal temporal dynamics of the series and identify significant deviations without relying on labeled abnormal samples. Prediction-based approaches assume that a model trained on normal data will produce low prediction errors under regular conditions. When abnormal behavior occurs, the prediction error increases. The

anomaly score at time t is defined as:

$$e_t = |\hat{y}_t - y_t|. \tag{1}$$

where $y_t \in \mathbb{R}$ is the true observed electricity consumption at time step t, and $\hat{y}_t \in \mathbb{R}$ is the corresponding value predicted by the model. A threshold τ, derived from the distribution of training errors, is used to classify anomalies:

$$a_t = \begin{cases} 1, & \text{if } e_t > \tau \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

This framework relies on the assumption that abnormal events correspond to deviations from learned normal temporal structure.

2.1.2 Convolutional Temporal Modeling

To capture temporal dependencies efficiently, dilated causal convolutions are employed. For an input signal x_t, the dilated convolution is defined as:

$$y_t = \sum_{i=0}^{k-1} w_i \, x_{t-di}, \tag{3}$$

where d denotes the dilation factor and k the kernel size. Exponential growth of dilation across layers expands the receptive field without increasing model depth, enabling multi-scale temporal modeling while preserving causality.

2.1.3 Forecasting Objective in Energy Systems

Beyond anomaly detection, forecasting aims to estimate future load values over a horizon H. The probabilistic forecasting objective is:

$$p(y_{t+1:t+H} \mid y_{1:t}), \tag{4}$$

where the goal is to model not only expected demand but also uncertainty. Electricity consumption exhibits variability due to occupancy behavior, environmental factors, and operational changes, making deterministic point forecasts insufficient in many practical scenarios.

2.1.4 Attention Mechanisms for Long-Horizon Modeling

Self-attention enables interactions between all time steps within an input window. Given query Q, key K, and value V, attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V. \tag{5}$$

This mechanism supports modeling long-range dependencies without sequential recurrence, making it suitable for long-horizon forecasting.

2.1.5 Variational Inference for Uncertainty Quantification

To capture predictive uncertainty, a latent variable z is introduced:

$$p(y_{t+1:t+H} \mid X_t) = \int p(y_{t+1:t+H} \mid X_t, z)p(z)\, dz. \tag{6}$$

Training maximizes the Evidence Lower Bound (ELBO):

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q(z|X)}[\log p(y|X, z)] - \text{KL}(q(z|X)\|p(z)). \tag{7}$$

This framework enables calibrated predictive distributions, allowing forecasting systems to represent uncertainty explicitly.

2.2 Related works

Although anomaly detection and probabilistic forecasting are developed separately in this thesis, they are linked by a common objective: improving the reliability of energy time-series modeling. Anomaly detection deals with irregular observations that can affect model training and downstream

analysis. Probabilistic forecasting, in contrast, focuses on representing uncertainty in future demand. In this sense, the first addresses the quality of the data, while the second addresses uncertainty in prediction.

Anomaly Detection in Time-Series Data

Anomaly detection in time-series data has evolved from classical statistical residual analysis to deep generative modeling frameworks. Early approaches relied on ARIMA-based residual thresholding, where anomalies were detected as deviations from expected linear predictions. While effective under stable conditions, these methods remain sensitive to outliers and model misspecification, particularly when nonlinear dynamics or structural shifts are present. With the increasing availability of large-scale sensor data, deep learning techniques have become dominant in anomaly detection research. CNN-based and recurrent architectures have been widely adopted to capture temporal dependencies. However, many of these frameworks rely on supervised training and require labeled abnormal data, which are often scarce in smart building environments. Autoencoders and Variational Autoencoders (VAEs) detect anomalies through reconstruction error by learning compact representations of normal patterns Berahmand (2024). VAEs, in particular, introduce latent variables to model hidden variability, offering a probabilistic perspective on normal behavior. While effective, reconstruction-based approaches may struggle with subtle or gradually evolving anomalies. DGHL introduced a deep generative framework with hierarchical latent components and alternating backpropagation, mapping windowed segments to latent spaces and reconstructing them using top-down convolutional networks. Although DGHL performs well on multivariate time series with complex dynamics, its computational demands may limit deployment in real-time smart building environments. WaveNet, originally developed for raw audio synthesis Van Den Oord (2016), models long-range temporal dependencies using stacked dilated causal convolutions and gated activation units. Autoregressive WaveNet variants have been applied to anomaly detection in acoustic monitoring systems Hayashi, Komatsu, Kondo, Toda, and Takeda (2018). SW-WaveNet combined waveform and spectrogram representations to enhance detection performance Chen, Ran, Sun, and Cai (2023). Conditional WaveNet introduced auxiliary embeddings Komatsu, Hayashi, Kondo, Toda, and Takeda (2019), and CGAN-WaveNet incorporated adversarial training objectives

Zhao et al. (2018) . However, these WaveNet-based approaches are primarily tailored to dense audio signals with strong frequency-domain characteristics. Energy consumption data differ substantially in sparsity, irregularity, and regression-based anomaly formulation, limiting direct transferability of audio-focused WaveNet adaptations. rating that combining convolutional and recurrent components improves forecasting performance.

Classical and Deep Learning Approaches to Load Forecasting

Research on electricity load forecasting has followed several different paths over the years. Early studies relied heavily on classical statistical models—ARIMA, SARIMA, and a range of exponential smoothing schemes. These approaches work reasonably well when demand evolves smoothly and remains close to a linear trend Zhuang, Chen, Horata, and Sunat (2025). Their limitations become evident once the load begins shifting abruptly or displays noticeable nonlinear behaviour Maragkos and Refanidis (2025). Rotib et al. (2021) offered a clear illustration of this issue: using a PCA–ARIMA setup on household IoT data, they found that the model performed adequately only at short horizons, with accuracy declining sharply as the forecast window expanded. Similar outcomes have been noted across several studies. Deep learning techniques emerged partly to overcome these issues. In Rafi, Nahid-Al-Masood, Deeba, and Hossain (2021), a CNN–LSTM architecture was proposed to capture local patterns through convolutional layers before modelling temporal evolution with an LSTM. A related effort by used a CNN–GRU configuration Hasanat et al. (2024). Because GRUs handle temporal dependencies more effectively than CNN layers alone, the combined model outperformed either component individually. Efforts to broaden the temporal context led to the adoption of Temporal Convolutional Networks. The TCN developed by Shaikh, Nazir, Khalique, Shah, and Adhikari (2023), based on dilated convolutions and residual connections, displayed more stable behaviour than LSTM baselines and handled seasonal fluctuations more reliably. A major shift occurred with the introduction of Transformer architectures. Originally designed for language tasks Vaswani et al. (2017), Transformers became attractive for time-series applications due to their ability to relate distant events without recurrence. This property is particularly relevant for electricity load forecasting, where long-horizon structure interacts with multiple seasonal patterns. Surveys such as Chan and Yeo (2024) highlight the central role of

Transformers in recent work. Informer introduced sparsified attention to manage long sequences H. Zhou et al. (2021). Autoformer incorporated series decomposition and auto-correlation to better capture periodic behaviour Wu, Xu, Wang, and Long (2021) and FEDformer further advanced this line through Fourier-domain attention and trend–seasonal decomposition T. Zhou et al. (2022).

Probabilistic Forecasting, Variational Models, and Temporal Conditioning

As forecasting horizons increased, interest gradually shifted from point forecasts to probabilistic forecasting. System operators require not only an estimate of expected demand but also an indication of the uncertainty around it Kim, Kim, Kim, Lee, and Yoon (2025). This motivated the development of several neural probabilistic models. DeepAR Salinas, Flunkert, Gasthaus, and Januschowski (2020), for instance, uses a global autoregressive RNN together with a learned likelihood function to generate predictive distributions across many related series. Despite its contributions, the recurrent structure limits its ability to capture long-range dependencies and multimodal uncertainty. These constraints contributed to the growing interest in Variational Autoencoders. VAEs generate full predictive distributions by sampling from a latent space Bond-Taylor, Leach, Long, and Willcocks (2022), making them well suited for uncertainty modelling. They also provide a structured way to incorporate latent drivers of consumption—such as behavioural changes or weather-related influences Shrivastava, Rameshan, and Agnihotri (2024). Hybrid VAE architectures combining latent variables with recurrent or convolutional encoders Ali, Xia, Zia, Bangyal, and Iqbal (2025) have improved uncertainty modelling, although long-horizon forecasting remains challenging. Recent research has therefore examined latent-variable models integrated with self-attention. Transformer–VAE hybrids use attention layers to capture broad temporal structure while the latent layer models shorter-range variability Mentzelopoulos, Fan, Sapsis, and Triantafyllou (2024). Related hybrid designs have appeared in other fields as well. For example, Xie, Xu, Jiang, Gao, and Wang (2024) propose a self-attention VAE for anomaly detection in industrial pump systems, while TransVAE-DTA C. Zhou, Li, Song, and Xiang (2024) combines a Transformer encoder with a VAE to enhance drug–target binding affinity prediction. These studies highlight the versatility of Transformer–VAE architectures, although none have been applied to electricity load forecasting or long-horizon probabilistic prediction. Similarly, the Temporal Fusion Transformer

(TFT) Ferreira, Leite, and Salvadeo (2025) produces probabilistic forecasts using gating mechanisms, attention modules, and variable-selection networks, although it relies heavily on exogenous variables and does not directly enforce uncertainty calibration. Another line of research emphasizes the importance of explicit temporal conditioning, calendar attributes, cyclical encodings, or rolling statistics. Feature-wise Linear Modulation (FiLM) offers a simple way to incorporate such information by scaling and shifting hidden activations. Although FiLM has been applied mostly in vision and multimodal learning, it has also been used to condition Transformer encoders in other sequence-to-sequence contexts. For example, D.-H. Yang and Chang (2022) showed that FiLM can effectively modulate Transformer hidden states under noisy-sequence conditions. To date, however, FiLM has not been explored in electricity load forecasting. The FiLM Transformer VAE model proposed in this work is designed to address exactly this gap. By combining feature-wise temporal conditioning with a compact variational layer, the model preserves long-range temporal structure while producing well-calibrated uncertainty estimates. This integration enables both accurate point forecasts and reliable predictive distributions, without introducing excessive computational cost. To our knowledge, this is the first study to evaluate a Transformer VAE hybrid for long-horizon electricity load forecasting under both deterministic and probabilistic metrics, using feature-wise temporal conditioning to improve uncertainty calibration without increasing architectural complexity.

2.3 Datasets

The two methodological contributions of this thesis are evaluated on distinct datasets tailored to their respective objectives. Specifically, the AEMO dataset is used exclusively for unsupervised anomaly detection experiments, while the Office Building and PJM datasets are used exclusively for probabilistic load forecasting experiments.

2.3.1 AEMO dataset

The experimental evaluation is based on real-world electricity demand time series provided by the Australian Energy Market Operator (AEMO). The data represent aggregated electricity demand at the regional (system) level for individual Australian states, rather than consumption from specific

buildings or subsystems. The demand is measured at a 30-minute resolution, reflecting typical operational settings used in energy monitoring and forecasting applications. The dataset spans multiple consecutive months, providing sufficient temporal coverage to capture daily and weekly consumption patterns as well as longer-term dynamics. As the data are aggregated at the regional level, no specific building type is assumed. Nevertheless, the temporal structures observed in the data such as periodicity, load variability, and gradual deviations are similar to those commonly encountered in smart building energy monitoring, making the dataset a suitable benchmark for anomaly detection methods.

Data generation and preprocessing

In order to evaluate the model's ability to detect anomalies, we introduce four different kinds of anomalies Turowski (2022), each of which is intended to resemble a believable example of anomalous energy behavior:

- Type 1 — Negative Peak with Recovery: This anomaly involves a sharp, abrupt drop followed by a sequence of zero-consumption values, culminating in a compensatory spike that restores the lost energy. It simulates sudden system outages followed by delayed restoration.
- Type 2 — Consumption Drop with Drift: A gradual value decrease or zeroing, followed by an accumulation of corrections. This represents underreported energy values brought on by sensor drift or partial outages.
- Type 3 — Negative Spikes: Isolated, abrupt negative anomalies of considerable or high magnitude that mimic rapid losses in data integrity or system disruptions.
- Type 4 —Positive Spikes: Extreme to moderate high-energy bursts that represent over-reported values because of miscalibration or surges.

A diverse distribution of anomalies in both position and type is ensured by injecting each anomaly at random positions and with randomized intensity. The final configuration guarantees that approximately 10% of the test set points are anomalous, with precise lengths and positions controlled through Gaussian sampling. The expected anomaly coverage is defined at the day level using

a *day contamination rate*, and further refined through a *point-level contamination ratio* referred to as the *data contamination rate*.

The data is subjected to anomalies without any overlap, and in order to facilitate further analysis, related ground truth labels are produced simultaneously. It is crucial to maintain the unsupervised learning paradigm by keeping training data completely clean and devoid of any injected perturbations.

2.3.2 Office and PJM Dataset

We evaluated the FiLM conditioned Transformer VAE on two load datasets that represent electricity use at very different scales. The first, the PJM East Region (PJME) dataset, reports hourly power demand for a large interconnected regional system. The second, the Office Building Load dataset, provides 30-minute measurements from a single commercial building. In both cases, the series display marked daily, weekly, and seasonal structure, overlaid with irregular variations linked to weather and operating conditions.

Office Building Load Dataset (30-minute sampling)

The office data are drawn from the Industrial Park Electric Power Load Dataset (Suzhou, 2016–2021). Smart meters monitored several buildings within the park; for this study we retained the subset corresponding to one office building over the period from January 2016 to January 2018. The resulting series contains 10,801 valid observations of active power (kW) at 30-minute intervals. Basic preprocessing included removing duplicate records, correcting timestamp inconsistencies, and interpolating short gaps by linear interpolation. The cleaned load exhibits clear weekday, weekend contrasts and regular intra-day cycles associated with occupancy and HVAC operation. This dataset is used to assess short-horizon forecasting at relatively high temporal resolution.

PJM East Region Dataset (hourly sampling)

The PJM East Region (PJME) dataset corresponds to one of the largest wholesale electricity markets in North America. It reports hourly system load (MW) for a footprint covering 13 U.S. states plus the District of Columbia, from 2001 to 2018. The series reflects aggregate demand

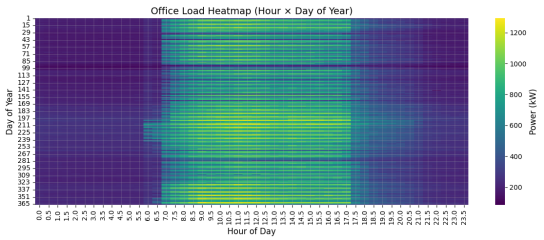
(b) Mean daily load profile — Office (30-minute resolution)

(c) Mean daily load profile — PJM (hourly resolution)

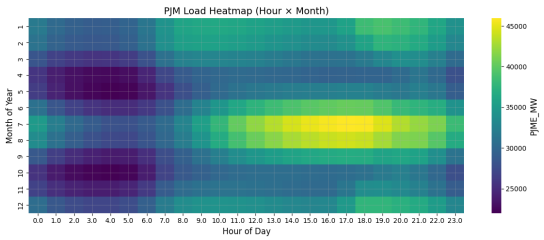
Figure 2.1: Exploratory visualizations of the load datasets (a–c).

patterns shaped by temperature, industrial activity, and broader economic conditions. As a preprocessing step, we corrected timestamp gaps, removed observations with missing values, and trimmed outliers lying more than three standard deviations from a rolling median. This dataset serves as a large-scale test case for regional load forecasting. Because of its size and temporal range, the PJME dataset is widely used as a benchmark for long-term probabilistic forecasting.

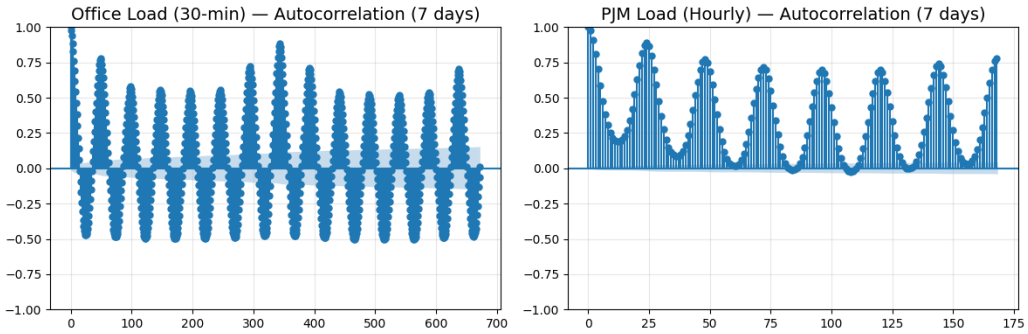
Figures 2.1 shows how both datasets behave over time. The office load has sharper peaks and quicker changes, while PJME exhibits smoother seasonal swings and stronger long-term memory. These contrasts underline the need for models able to capture both short and extended temporal patterns. Additional exploratory visualizations are provided in 2.2.



(d) Office load heatmap (hour × day of year)



(e) PJM load heatmap (hour × day of year)



(f) Autocorrelation functions of Office and PJM loads

Figure 2.2: Exploratory visualizations of the load datasets (d–f).

Chapter 3

Refined WaveNet for Robust Smart Building Monitoring

3.1 Introduction

As automation and sensor technologies continue to advance, smart buildings are becoming increasingly effective in reducing energy consumption, enhancing security, and improving occupant comfort Jia, Komeily, Wang, and Srinivasan (2019). These technological advances enable the continuous collection of high-resolution time-series energy data, which plays a pivotal role in performance analysis and operational efficiency. Recent studies highlight that data-driven strategies leveraging this information can reduce energy consumption by up to 20% Selvarajan (2021). Additionally, the widespread availability of such data facilitates the automation of smart grid functionalities, including load analysis, forecasting, and intelligent energy management. However, this rich time-series data often contains anomalies that indicate unusual energy consumption behaviors, technical malfunctions, or even security threats Diamantoulakis, Kapinas, and Karagiannidis (2015). These anomalies pose significant challenges to system stability and degrade the performance of downstream tasks such as forecasting Yan et al. (2024). Undetected, they can skew predictive models and affect efficient energy allocation, emphasizing the critical need for reliable anomaly detection. A variety of methods have been developed to address this problem. Traditional statistical models are appreciated for their simplicity and computational efficiency in modeling linear dependencies, but

they often fail to capture the non-linear or multi-scale patterns characteristic of real-world anomalies. Classical machine learning techniques, while more flexible, require careful parameters, are sensitive to noise, and often struggle to scale with large datasets Tufail, Riggs, Tariq, and Sarwat (2023). Moreover, many conventional methods rely on assumptions about data normality or specific anomaly distributions assumptions which are rarely satisfied in practice Ruff et al. (2021).In smart building environments, these limitations become particularly pronounced, as consumption patterns are shaped by occupancy cycles, operational schedules, and device-level behaviors that deep neural networks can model more effectively. In comparison, deep learning models present a robust alternative thanks to their ability to learn hierarchical representations from raw data, enabling the identification of complex, non-linear patterns. Among these, Variational Autoencoders (VAEs) are commonly employed in unsupervised anomaly detection by constructing compact latent embeddings and identifying instances with significant reconstruction errors Neloy and Turgeon (2024). Nonetheless, VAEs often lack temporal awareness and may underperform when dealing with anomalies that span several time steps. This issue is especially important in smart building applications, where abnormal energy use often appears gradually or is tied to long-term patterns. Recurrent Neural Networks (RNNs) handle sequences by modifying a hidden state at every time step, enabling them to capture short-term temporal connections. They are efficient for capturing immediate variations in energy consumption Bouktif, Fiaz, Ouni, and Serhani (2020). However, RNNs are prone to the vanishing gradient problem, which limits their ability to retain information over long durations Waqas and Humphries (2024). This reduces their effectiveness at detecting anomalies that unfold gradually over time or depend on long-range patterns. Such long-range patterns are common in smart buildings, where energy usage depends on daily routines, operational cycles, and occupancy behavior, reinforcing the need for architectures that can represent extended temporal dependencies. To overcome these limitations, Long Short-Term Memory (LSTM) networks incorporate gating mechanisms that allow for selective memory retention and forgetting Chien et al. (2021). Although they are effective at capturing extended dependencies, LSTMs are often resource-intensive, difficult to tune, and less suitable for real-time or resource-constrained settings. Recent approaches like Dual and DGHL have investigated contrastive and generative learning algorithms, respectively, but they have problems with generalization and scalability, particularly in sparse, context-independent

environments like energy time series. Together, these limitations suggest that effective anomaly detection requires models suited to the sparsity and irregular fluctuations of smart building energy data, rather than architectures originally built for well-structured domains like vision or audio. By introducing a deep generative model with alternating backpropagation and hierarchical latent structures, Deep Generative model with Hierarchical Latent Factors (DGHL) makes rich representation learning possible Challu, Jiang, Wu, and Callot (2022). However, its dependence on iterative optimization and posterior sampling increases inference latency and limits scalability. Dual, on the other hand, focuses on discriminative representations between normal and anomalous patches employing dual attention and a contrastive learning technique Y. Yang, Zhang, Zhou, Wen, and Sun (2023). It may not generalize well in sparse or noisy data circumstances, which are common in energy time series, and it is very dependent on the design of pretext tasks, despite showing promising results. Additionally, many deep learning methods for anomaly detection are supervised, requiring access to well-annotated abnormal instances. Given the rarity, diversity, and contextual specificity of anomalies, obtaining sufficient labeled samples is often infeasible Pang, Shen, Cao, and Van Den Hengel (2021). This challenge hampers model generalization, especially when adapting to new environments or detecting previously unseen faults. The class imbalance between normal and anomalous events further biases these models toward the majority class, reducing their sensitivity to critical outliers. Unsupervised learning offers a compelling alternative by modeling normal system behavior and identifying deviations without the need for labeled data. This approach enhances scalability, generalization, and adaptability to real-world deployments where labeled faults are scarce or unavailable Verma (2022).

In this study, we propose a refined adaptation of the original WaveNet architecture for unsupervised anomaly detection in univariate energy time series generated by sensors in smart building environments. The original WaveNet, originally developed for raw audio synthesis, leverages dilated causal convolutions, residual connections, and gated activation units (GAUs) to effectively model long-range temporal dependencies Van Den Oord (2016). While highly effective for dense and high-frequency audio signals, this architectural design is less suitable for energy consumption data,

which is typically characterized by low-resolution, sparsity, temporal irregularity, and limited contextual information. Furthermore, the original WaveNet relies on μ-law quantization and categorical cross-entropy loss within an autoregressive, sample by sample training regime, an approach that is computationally inefficient and less interpretable in the context of anomaly detection Van Den Oord (2016). These limitations reduce its applicability to real-world energy monitoring scenarios, where precision, scalability, and robustness to data irregularities are essential.

Our proposed Refined Gated WaveNet addresses the aforementioned limitations through a series of targeted architectural enhancements that improve its suitability for modeling anomalies in energy time series. First, we remove the μ-law quantization and the categorical output layer of the original architecture, replacing them with a continuous-valued regression framework optimized using Smooth L1 loss. This choice offers greater robustness to the heavy-tailed error distributions frequently induced by outliers in energy data. Second, while the original WaveNet employs Gated Activation Units (GAUs), we deliberately retain them in place of ReLU due to their superior ability to modulate information flow and capture gradual signal deviations characteristics essential for identifying subtle and slowly evolving anomalies in real-world energy consumption. Third, we replace the original autoregressive sampling mechanism with a window-wise parallel regression strategy, enabling efficient batch processing while preserving the temporal coherence of the input sequence. Fourth, we introduce a temporal refinement head, implemented as either a 1D convolutional layer or a gated recurrent unit (GRU), designed to capture temporal dependencies beyond the fixed receptive field, thereby enhancing the detection of long-range and multiscale anomalies. Finally, to improve training stability and generalization, particularly in sparse or noisy regimes, we integrate dropout and weight normalization within the residual and skip connections.

3.2 The proposed approach

In this section, the proposed anomaly detection framework based on the improved WaveNet architecture is introduced and detailed. The design integrates architectural improvements that are specifically suited to outlier-heavy distributions, sparse anomalies, and irregular temporal patterns found in energy consumption data. The model recognizes anomalies based on deviations from

learned temporal structures and is only trained on normal data. Through a series of architectural improvements, the suggested model expands upon the original WaveNet and improves its applicability for time series anomaly detection. In particular, it substitutes gated activation units (GAUs) for ReLU activations to allow for smoother signal transitions and adds a temporal refinement head, either a GRU or a 1D convolutional layer, after skip connection aggregation to better capture any remaining temporal patterns. For effective training and inference, it switches from autoregressive generation to parallel windowed regression. It also uses a Smooth L1 loss to increase resilience while learning from energy data that is full of outliers.

3.2.1 Original WaveNet

The original WaveNet, a deep autoregressive model developed for raw audio synthesis, achieved high-fidelity waveform generation by modeling long-range temporal dependencies without relying on recurrence. It uses stacks of dilated causal convolutions with exponentially increasing dilation rates across layers, enabling each output to incorporate a wide receptive field while maintaining a manageable network depth. Each residual block includes a gated activation mechanism, combining $\tanh$ and sigmoid outputs elementwise to better capture complex signal dynamics Van Den Oord (2016).

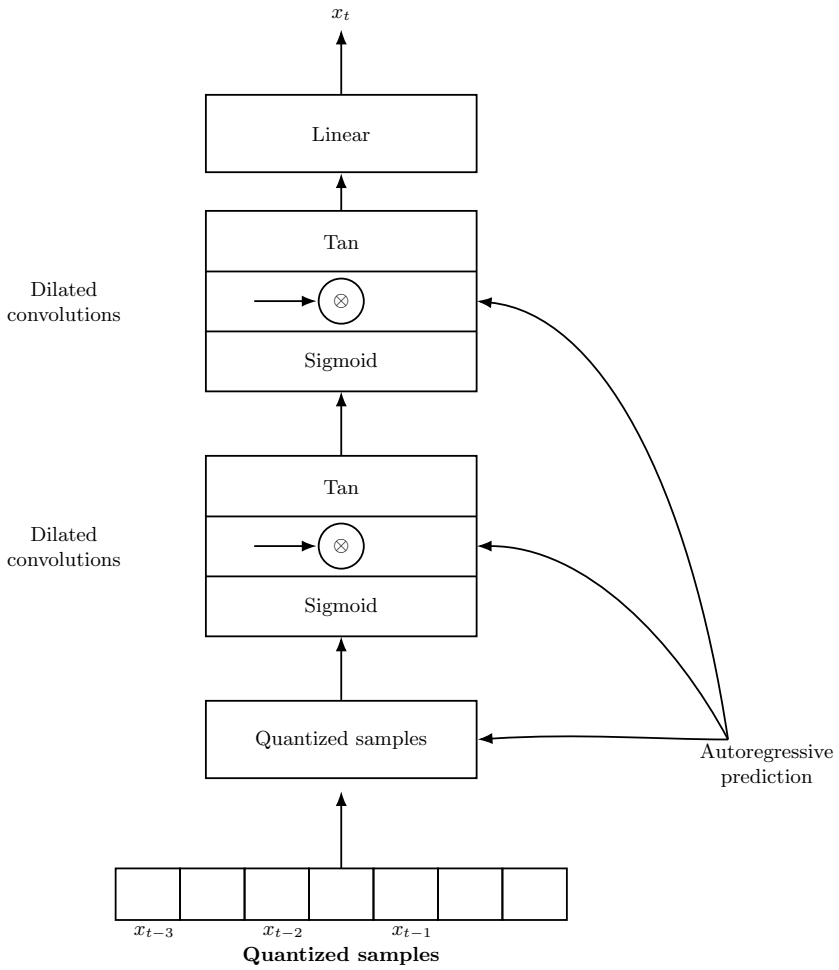


Figure 3.1: Architecture of the original WaveNet. It uses dilated causal convolutions, gated activations, and autoregressive feedback from previously quantized samples.

The model is trained in an autoregressive fashion: each quantized output sample is predicted sequentially based on preceding samples, with prior outputs recursively fed back as inputs. The targets are discretized using μ-law encoding and optimized using categorical cross-entropy loss. While this setup is ideal for dense audio signals, it imposes significant limitations for structured time series anomaly detection. These include high inference latency due to sequential generation, sensitivity to large deviations, and assumptions of smooth temporal continuity characteristics that do not hold for sparse and irregular domains such as energy consumption data. As implemented in our baseline model, this version retains the classification-based output head and uses μ-law encoded

input sequences for sample-wise prediction over 256 classes. The original WaveNet design, with its dilated causal convolutions, gated activation units, and autoregressive structure, is highlighted in Figure 3.1.

3.2.2 Refined WaveNet

A deep architecture designed for unsupervised anomaly identification in univariate energy time series is the Refined Gated WaveNet. It replaces autoregressive prediction with parallel one-step forecasting to speed up training and inference, building on fundamental WaveNet concepts such as residual routes, gated activations, and dilated causal convolutions. Its context awareness is extended by a temporal refinement head (GRU or CNN), and robust generalization is ensured by dropout and weight normalization. When combined with an error-based, the model successfully identifies anomalies in energy usage data.

Causal Dilated Convolutions

To maintain the temporal order of the signal, 1D dilated causal convolutions are used in each WaveNet residual block. The definition of the dilated convolution for a given input x is:

$$y_t = \sum_{i=0}^{k-1} w_i \cdot x_{t-d \cdot i} \tag{8}$$

where d is the dilation rate, k is the kernel size, and w_i are learnable filter weights.

Long-range temporal dependencies can be captured by the model thanks to this structure without adding more layers or processing load. Across layers (e.g., 1, 2, 4, 8,...), the dilation is increased exponentially, guaranteeing a broad receptive field appropriate for detecting temporally extended abnormalities. This design allows each convolutional layer to observe increasingly distant past inputs without increasing the model's depth, effectively expanding the memory scope of the network.

Gated Activation Units (GAUs) and Residual Pathways

Rather than ReLU, each block employs a GAU,defined by:

$$z = \tanh(W_f * x) \odot \sigma(W_g * x) \tag{9}$$

where W_f and W_g are separate convolutional filters for the $\tanh$ and sigmoid activations.

The network can adjust the flow of temporal information thanks to this gating mechanism, which increases its sensitivity to even the smallest signal changes. The integration of tanh and sigmoid results in seamless transitions, enhances resilience to slight oscillations and allows the identification of slow-drifting abnormalities.

The residual connection is formed for each residual block by adding the output z back to the input:

$$x_{\text{out}} = x + z \tag{10}$$

Additionally, skip connections $s_i = W_s * z$ are aggregated across all layers:

$$S = \sum_{i=1}^{L} s_i \tag{11}$$

These links improve the expressiveness and interpretability of the model by stabilizing gradient flow during training, enabling deeper networks, and for intermediate characteristics to contribute to the final output.

Temporal Refinement Head and Output Layer

A refinement module processes the skip outputs following their aggregation:

- GRU Refiner: A single-layer GRU receives the aggregated skip tensor S after it has been reshaped to (B, T, C). It utilizes the final hidden state h_T:

$$h_T = \text{GRU}(S)$$

Here, S is reshaped to (B, T, C) format corresponding to batch size (B), time steps (T), and channel dimension (C) to match the input requirements of the GRU layer.

- CNN Refiner: As an alternative, global average pooling is used after a 1D convolution:

$$h = \text{GlobalAvgPool}(\text{Conv1d}(S))$$

This refiner improves the model's capacity to incorporate dependencies outside of the fixed receptive field and refine latent dynamics. Additionally, it strengthens resistance to local fluctuations and transient noise spikes. A fully connected layer receives the final output representation h from the temporal refinement module and maps it to a scalar prediction $\hat{y}$, corresponding to the subsequent time step in the input sequence:

$$\hat{y} = W_{\text{out}} h + b \tag{12}$$

Here, W_{out} and b denote the learnable weights and bias of the output layer, respectively.

Regularization: Dropout and Weight Normalization

To improve generalization and stabilize training, the architecture integrates two complementary techniques: *weight normalization* and *dropout*.

Weight normalization is applied to all convolutional layers within the residual blocks. By decoupling the magnitude of weight vectors from their direction, it accelerates convergence and ensures stable training dynamics in deep configurations.

Dropout is used selectively at strategic locations in the architecture to encourage generalization and avoid overfitting. It is applied to the gated output z in the residual blocks after the GAUs. Dropout is also added at the last hidden state h_T in the GRU refinement head, but it comes after the convolutional layer and before the global average pooling in the CNN variation. When combined, these dropout techniques strengthen the model's resilience to noise, increase its capacity to identify uncommon anomalies, and help it generalize to previously unseen data.

3.2.3 Training Procedure for the Refined Gated WaveNet and Inference

Using only clean data, the Refined Gated WaveNet is trained in a fully supervised regression environment. The input segments for each training sample are $\mathbf{X}_t = [x_{t-L+1}, \ldots, x_t]$ and $y_t = x_{t+1}$, which correspond to the appropriate target value. A batch of these windowed samples is used to optimize the model in order to decrease the prediction error.

Our model conducts regression over parallel input windows, in contrast to the original autoregressive WaveNet, which predicts one sample at a time in a sequential loop. This method enables effective GPU-based batch training and significantly reduces inference latency.

The Smooth L1 loss, sometimes referred to as the Huber loss, is used to compute training error. This loss function combines the robustness of L1 with the smoothness of L2 and is well-suited for noisy time series Barron (2019). It includes a tunable hyperparameter β (set to $\beta = 1.0$ in our experiments), which controls the transition point between quadratic and linear behavior:

$$\mathcal{L}_{\text{SmoothL1}}(x, y) = \begin{cases} 0.5(x-y)^2/\beta, & \text{if } |x-y| < \beta \\ |x-y| - 0.5\beta, & \text{otherwise} \end{cases} \tag{13}$$

By reducing the average loss across the training batch, the model parameters θ are optimized:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\text{SmoothL1}}(\hat{y}_i, y_i) \tag{14}$$

The Adam optimizer is used for training, with a batch size of 64 and a fixed learning rate of 10^{-3}. To enhance generalization, dropout and weight normalization are used across the residual and skip paths. To avoid overfitting, an early stopping criterion based on the validation loss is applied. Using the same sliding window technique, the model is applied to unseen sequences after being trained on typical data. A one-step-ahead prediction $\hat{y}_t$ is produced by the model for every input segment $\mathbf{X}_t$.

The absolute prediction error is the definition of the anomaly score for every time step:

$$e_t = |\hat{y}_t - y_t| \tag{15}$$

Based on the training set's α-th percentile of prediction errors, a dynamic threshold τ is calculated:

$$\tau = \text{Percentile}_{\alpha}(e_{\text{train}}) \tag{16}$$

A time step t in the test set is considered anomalous if the prediction error is greater than the cutoff:

$$a_t = \begin{cases} 1, & \text{if } e_t > \tau \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

This unsupervised scoring method adjusts to the training data's statistical profile and without the need for labeled anomalies. It makes sure that unexpected or gradual departures from learnt normal behavior are noted as possible abnormalities. This allows the Refined Gated WaveNet to effectively identify anomalous patterns of energy use in smart building settings.

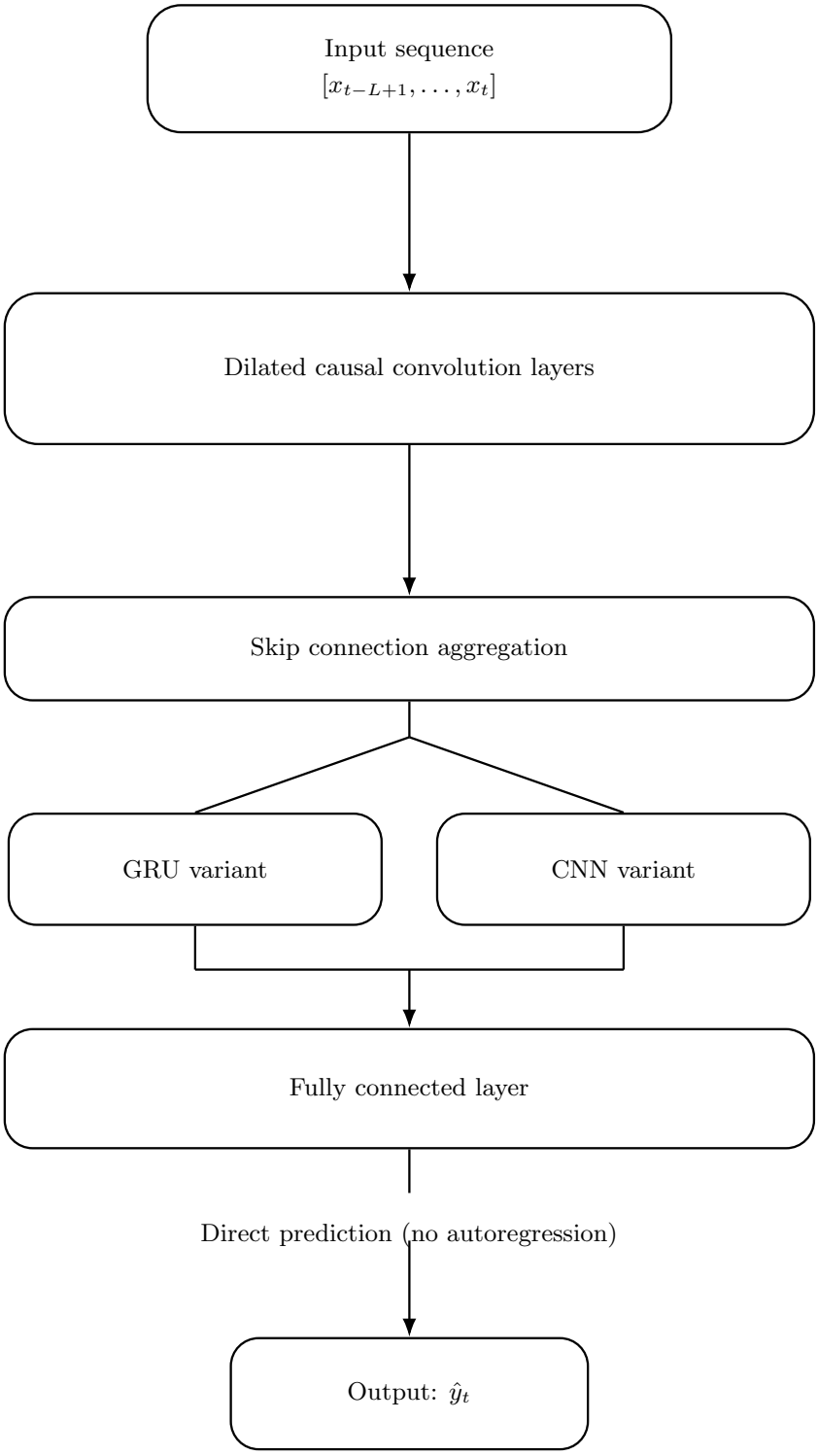


Figure 3.2: Architecture of Refined Gated Wavenet.

3.3 Experimental setup and results

3.3.1 Experimental Setup

To evaluate the effectiveness of the proposed refined WaveNet methods, we conduct a series of experiments on the AEMO dataset.

Preprocessing and Sequence Construction

The proposed anomaly detection framework follows a self-supervised learning strategy within a fully unsupervised setting, as no anomaly labels are available during training. Specifically, the model is trained using a one-step-ahead forecasting objective: given a fixed-length window of past observations, it learns to predict the subsequent value in the time series. Deviations between predicted and observed values are later exploited to derive anomaly scores during inference. First, the univariate time series is normalized using the training data's z-score normalization:

$$\hat{x}_t = \frac{x_t - \mu_{\text{train}}}{\sigma_{\text{train}}} \tag{18}$$

Next, we build training samples using a sliding window technique. The model is given an input vector made up of the prior L normalized values at every time step t:

- Input: $\mathbf{X}_t = [\hat{x}_{t-L+1}, \hat{x}_{t-L+2}, \ldots, \hat{x}_t] \in \mathbb{R}^L$
- Target: $y_t = \hat{x}_{t+1}$

The model learns a regression function $f_\theta : \mathbb{R}^L \to \mathbb{R}$ such that:

$$\hat{y}_t = f_\theta(\mathbf{X}_t) \tag{19}$$

During inference, reconstructed errors are employed to identify time steps that deviate significantly from learned normal patterns. Although this method relies on a regression-based objective, it is considered unsupervised anomaly detection because anomaly labels are not employed during training.

Evaluation Metrics

All reported results compare the proposed method with established univariate anomaly detection baselines, both in terms of anomaly detection performance evaluated against synthetically generated ground-truth labels, and in terms of their impact on downstream load forecasting accuracy after anomaly removal.

To adhere to the principles of unsupervised learning, the dataset is divided into clean and contaminated segments. The first 70% of the time series, assumed to be free of anomalies, is used for training and validation, while the remaining 30% is reserved for testing with injected synthetic anomalies (Chapter 2).

The clean portion is further split into:

- Training set: 80% of the clean segment used to train the model on typical behavior.
- Validation set: The remaining 20% is used for early stopping and generalization monitoring.

A fixed-length sliding window with stride 1 is applied across all sets. At each time step t, the model receives a window of past observations:

- Input: $\mathbf{X}_t = [\hat{x}_{t-L+1}, \ldots, \hat{x}_t]$
- Target: $y_t = \hat{x}_{t+1}$

To prevent overfitting and ensure robust generalization, we implement early stopping based on the validation loss, using the same Smooth L1 objective as during training. The model is evaluated on the validation set after each epoch, and training halts if no improvement is observed for a specified number of consecutive epochs (patience parameter). The checkpoint corresponding to the lowest validation loss is selected for final testing.

Because the validation set includes only clean sequences, an increase in loss signals poor generalization to unseen normal patterns or overfitting to transient fluctuations. This validation approach ensures that the model captures stable, broadly applicable temporal features while remaining robust to signal variability without exposure to anomalous behavior.

We assess the performance of anomaly detection using commonly used binary classification metrics, which are calculated on the test set with ground-truth labels and include Precision, Recall,

F1 Score, ROC AUC, and PR AUC. These measures offer a solid evaluation of the model's capacity to identify uncommon anomalies while accounting for the class imbalance present in time series data. Precision and recall are defined as:

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN} \tag{20}$$

where TP is the number of true positives, FP the false positives, and FN the false negatives.

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{21}$$

yields the F1 Score, which is the harmonic mean of precision and recall.

Moreover, we evaluate these metrics:

ROC AUC : The trade-off between true and false positive rates across thresholds is assessed by the Area Under the Receiver Operating Characteristic Curve.

PR AUC : A more useful metric for unbalanced datasets is the Area Under the Precision-Recall Curve, which shows how well precision and recall are balanced.

Forecasting Impact Evaluation

Even though our anomaly detection model is trained and tested in an unsupervised environment, we evaluate its influence on downstream forecasting tasks to assess its practical utility. In particular, we examine how anomalies affect the predictive performance of a basic LSTM model trained on energy data.

We evaluate the LSTM's forecasting accuracy under three different conditions:

(1) Clean Data: A dataset free of injected anomalies, used to calculate MAE and MSE.

(2) Contaminated Data: The same model is trained and tested on data with 10% synthetic anomalies.

(3) Cleaned Data: After identifying anomalous locations using each anomaly detection model separately, linear interpolation is used to replace them. Each clean dataset is then used to retrain the LSTM model, enabling evaluation of the downstream forecasting gains brought about by various

detection techniques.

We estimate performance using two regression metrics: Mean Absolute Error (MAE), which calculates the average absolute difference between expected and actual values, and Mean Squared Error (MSE), which highlights extreme deviations by amplifying larger errors through squaring. Lower values for both metrics indicate improved reliability and accuracy in forecasting.

Our Refined Gated WaveNet is evaluated against various popular univariate anomaly detection models. These comprise both reconstruction-based baselines and sequence modeling architectures:

- LSTM (Seq2One): Recurrent neural networks are trained to forecast a sequence's subsequent value. When expected and actual values differ, anomalies are identified.
- RNN: Tanh activations were used to train a vanilla recurrent neural network with the same next-step prediction goal as the LSTM.
- Variational Autoencoder (VAE): A probabilistic autoencoder that takes advantage of its latent distributional modeling to rebuild input windows and use reconstruction error as an anomaly score.
- DGHL: Alternating Back Propagation and Langevin Dynamics were used to train a deep generative model with hierarchical latent components. It uses MSE to calculate anomaly scores and reconstruct time-series windows.
- DCdetector:A contrastive learning strategy based on dual attention. By constructing permutation-invariant representations from patch-wise and in-patch views and comparing them using KL divergence, anomalies can be found without the need for reconstruction loss.

Our Refined Gated WaveNet was implemented using PyTorch and trained using a consistent configuration across all experiments to ensure fairness. The architecture consists of 10 residual blocks with exponentially increasing dilation rates (from 1 to 512), each using gated activation units and weight-normalized 1D convolutions with a kernel size of 2. A dropout rate of 0.2 is applied after each residual block to prevent overfitting. Skip connections are aggregated and passed to a temporal refinement head: a single-layer GRU with hidden size 64. The model is trained using

the Adam optimizer with a fixed learning rate of 10^{-3}, a batch size of 64, and the Smooth L1 loss function. Training runs for up to 30 epochs with early stopping based on validation loss.

Figure 3.3 shows the training and validation loss curves across epochs. The training loss drops quickly in the early stages and then levels off, suggesting that the model converges after only a few epochs. The validation loss follows a comparable pattern and stays close to the training loss, with only small variations over time. This behavior can be explained by the use of normalized inputs, regularization techniques, and non-overlapping sliding windows, which limit overfitting and promote stable generalization. The small and steady difference between the two curves suggests that the model does not overfit the training data. This point is important in an unsupervised setting, since the model is trained only on clean samples. Overall, the loss curves show that training is stable and that the model learns normal consumption behavior in a consistent way.

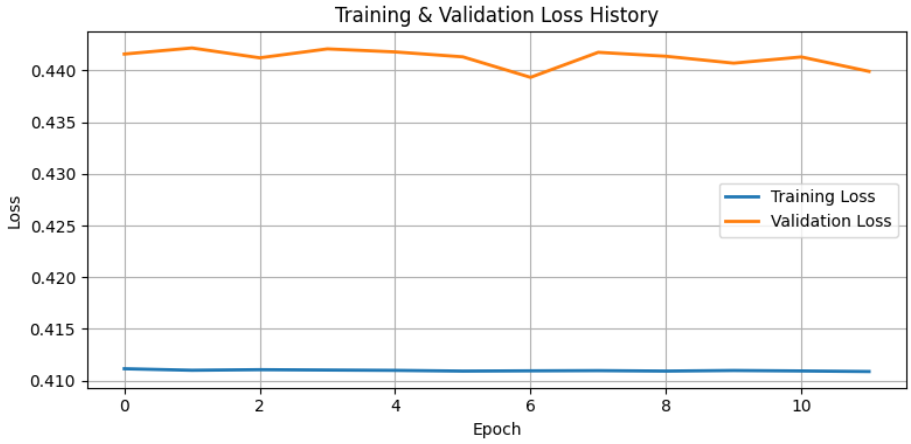


Figure 3.3: Training and validation loss curves of the Refined Gated WaveNet, showing stable convergence and good generalization.

3.3.2 Results

Impact of Anomalies on Forecasting Accuracy

Based on five widely used evaluation metrics *Precision*, *Recall*, *F1 Score*, *ROC AUC*, and *PR AUC*, Table 3.1 summarizes the detection performance of seven models. Among all contenders, the Refined Gated WaveNet delivers the strongest and most consistent performance across all metrics.

It achieves an F1 Score of 98.30%, surpassing the next-best model (VAE) by approximately 0.66%. It also leads in Precision (98.25%), Recall (98.34%), ROC AUC (99.65%), and PR AUC (99.23%), underscoring its robustness and adaptability to the class imbalance often encountered in anomaly detection.

The VAE model displays excellent sensitivity, achieving slightly higher Recall (98.64%) and PR AUC (99.33%), which reflects its strength in capturing subtle deviations. However, this comes at the cost of reduced Precision (96.67%), resulting in a slightly lower F1 Score of 97.64%. This trade-off highlights the intrinsic compromise in probabilistic reconstruction models between sensitivity and specificity.

The LSTM model demonstrates notable gains over the simplerRNN baseline, achieving an F1 Score of 88.56% versus 83.58%. This 5% improvement emphasizes the utility of gated mechanisms for modeling long-range dependencies and temporal variations in energy signals.

In contrast, the Original WaveNet and DUAL models show markedly lower performance. The Original WaveNet, which lacks the architectural refinements of our approach, achieves an F1 Score of only 41.96%, illustrating its limitations when applied to irregular, non-stationary time series. The Refined WaveNet improves upon this baseline by more than 56 percentage points, validating the impact of skip aggregation, gated activations, and temporal refinement. DUAL, with a Precision of just 42.27% and an F1 Score of 48.83%, performs the worst overall, reflecting poor adaptability to sparse and heterogeneous anomalies.

Overall, these results affirm the superiority of the Refined WaveNet, not only in quantitative terms but also in its reliability and generalization across diverse anomaly profiles. As elaborated in the following sections, this detection capability translates into significant improvements in downstream tasks such as energy forecasting.

Figure 3.4 provides a visual summary of the comparative performance, clearly showing *Refined WaveNet's* superiority across all evaluation criteria.

Table 3.1: Comparison of anomaly detection performance across models (in %).

Model	Precision	Recall	F1 Score	ROC AUC	PR AUC
LSTM Chien et al. (2021)	88.52	88.59	88.56	99.54	95.76
RNN Bouktif et al. (2020)	83.54	83.62	83.58	98.09	84.19
VAE Neloy and Turgeon (2024)	96.67	98.64	97.64	99.59	99.33
DGHL Challu et al. (2022)	63.39	97.01	76.68	81.03	35.50
DUAL Y. Yang et al. (2023)	42.27	57.80	48.83	68.73	52.14
Original WaveNet Van Den Oord (2016)	40.60	43.41	41.96	72.45	40.19
Refined WaveNet (ours)	**98.25**	**98.34**	**98.30**	**99.65**	**99.23**

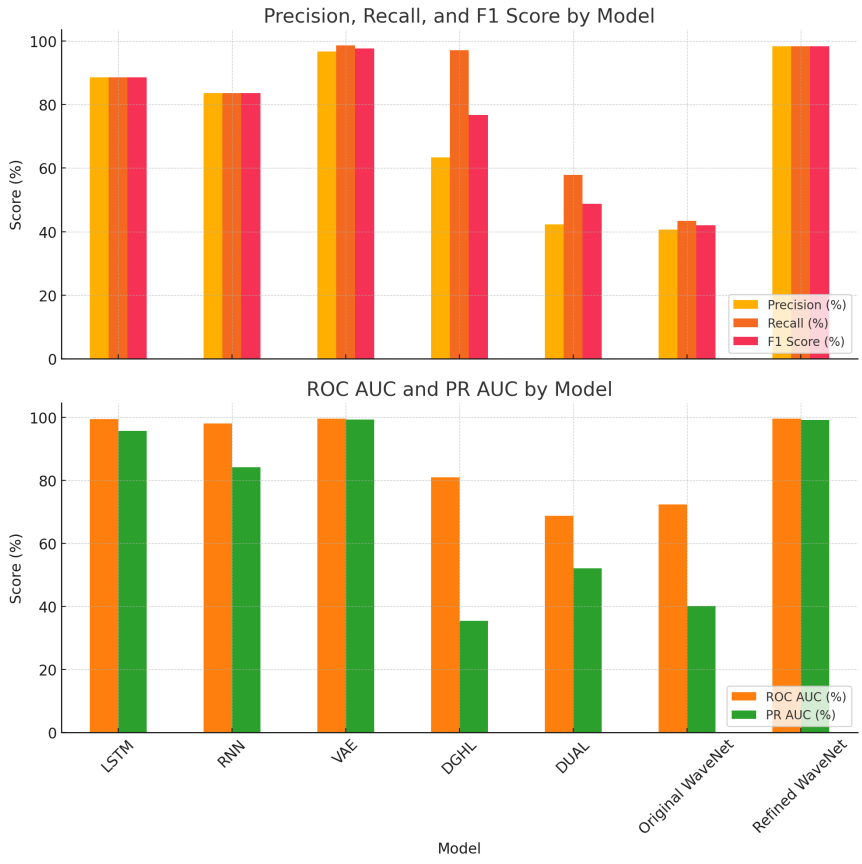


Figure 3.4: Bar chart comparison of anomaly detection performance across models using key metrics.

Impact of Anomalies on Forecasting Accuracy

By contrasting LSTM predictions on clean and corrupted data, we assessed the sensitivity of time series forecasting to anomalies. All forecasting results reported in this subsection correspond to

a one-day-ahead prediction horizon. As shown in Table 3.2, the Mean Squared Error (MSE) surged from 0.00189 to 57.69, while the Mean Absolute Error (MAE) increased markedly from 0.0311 to 2.473 with just 10% of the data artificially perturbed. These results underscore the substantial impact that even a small fraction of anomalous observations can exert on downstream predictive performance, reinforcing the importance of robust anomaly mitigation strategies within preprocessing pipelines.

Table 3.2: Forecasting performance with LSTM on clean and anomaly-injected data (one-day-ahead).

Condition	**MAE**	**MSE**	**RMSE**
Clean Data	0.0303	0.00183	0.0435
With Anomalies (10%)	2.639	57.528	7.596

Table 3.3: Forecasting performance with LSTM on clean and anomaly-injected data (two-day-ahead).

Condition	**MAE**	**MSE**	**RMSE**
Clean Data	0.0311	0.00189	0.0435
With Anomalies (10%)	2.473	57.69	7.596

Forecasting After Anomaly Removal

Using an LSTM regressor trained on data pre-cleaned by various detectors, we evaluated downstream forecasting performance one-day-ahead horizon to assess each model's corrective capability. After identifying anomalous segments, linear interpolation was used to reconstruct the affected time points, and the LSTM was retrained on the resulting series. Table 3.4 presents the resulting Mean Absolute Error (MAE) and Mean Squared Error (MSE) for each configuration.

The Refined WaveNet pipeline achieves the lowest forecasting errors, with an MAE of 0.0300 and an MSE of 0.0026, effectively restoring performance to levels observed on clean data. This result underscores the model's capacity to both accurately detect anomalies and preserve fine-grained temporal structures during denoising. As visualized in Figure 3.5, anomaly injection severely degrades forecasting accuracy. However, preprocessing with Refined Gated WaveNet successfully reduces MAE and MSE to near-baseline values, confirming the model's effectiveness in mitigating anomalies within energy time series.

The VAE also yields competitive results (MAE: 0.049, MSE: 0.005), confirming the VAE's ability to generalize across various anomaly types. However, its higher reconstruction error suggests a slight loss of local temporal precision compared to Refined WaveNet.

In contrast, traditional baselines such as DGHL and DUAL fail to restore predictive fidelity, with MSEs exceeding 27 and 35, respectively. These high errors indicate residual distortions in the reconstructed signals, limiting the effectiveness of subsequent forecasting.

Limited improvement is observed with LSTM and RNN, where the same architecture is used for both detection and prediction. Their errors remain significantly higher than the clean baseline, highlighting the limitations of low-capacity or general-purpose models in anomaly-heavy scenarios.

Overall, these results emphasize the value of integrating high-precision detection models into preprocessing pipelines. In domains such as energy forecasting, where even small anomalies can severely degrade performance, robust anomaly removal is critical to ensuring accurate and stable downstream predictions.

Table 3.4: Forecasting metrics with LSTM after anomaly removal using various detection models (one-day-ahead).

Model	**MAE**	**MSE**	**RMSE**
LSTM + LSTM	0.060	0.020	0.1414
RNN + LSTM	0.100	0.060	0.2449
VAE + LSTM	0.049	0.005	0.0707
DGHL + LSTM	3.212	27.937	5.2850
DUAL + LSTM	2.044	35.033	5.9197
Refined WaveNet + LSTM	**0.0300**	**0.0026**	**0.0510**

Table 3.5: Forecasting metrics with LSTM after anomaly removal using various detection models (two-day-ahead).

Model	**MAE**	**MSE**	**RMSE**
LSTM + LSTM	0.066	0.024	0.1549
RNN + LSTM	0.108	0.070	0.2646
VAE + LSTM	0.053	0.006	0.0775
DGHL + LSTM	3.0158	26.2908	5.1275
DUAL + LSTM	1.783	32.616	5.7137
Refined WaveNet + LSTM	**0.0300**	**0.00259**	**0.0510**

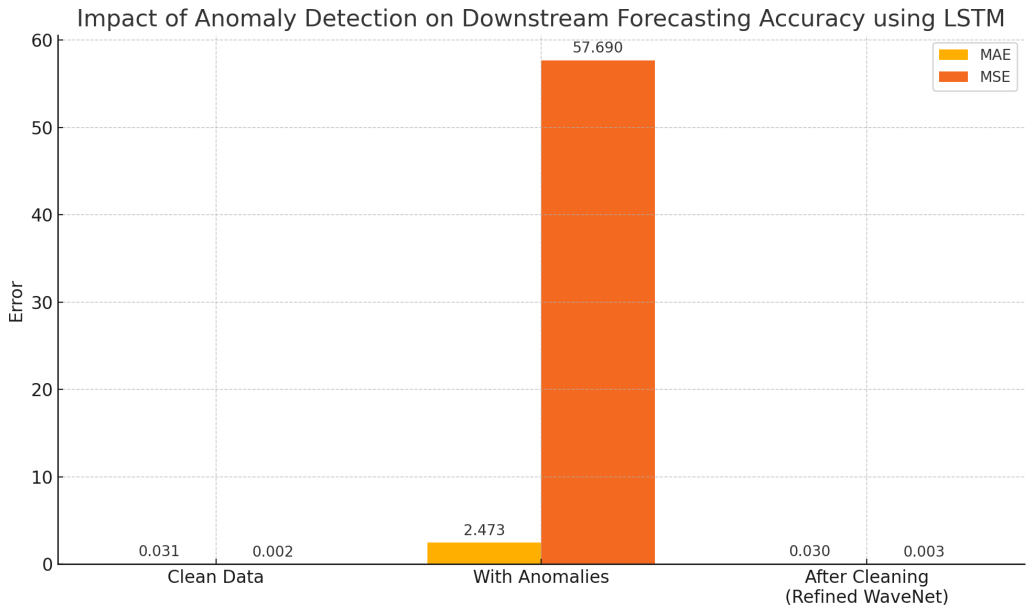


Figure 3.5: Effect of employing *Refined Gated WaveNet* for anomaly detection on the accuracy of LSTM forecasting downstream.

We also report results for a two-day-ahead forecasting horizon to examine whether the benefits of anomaly removal persist when the prediction range is extended. As shown in Table 3.5, LSTM-, RNN-, and VAE-based preprocessing leads to a small increase in error compared to the one-day-ahead case, which is expected as uncertainty accumulates over longer horizons. Nevertheless, the overall performance remains substantially better than forecasting on corrupted data. In contrast, DGHL and DUAL continue to exhibit large errors, indicating that the reconstructed signals still contain distortions that affect multi-day predictions. The Refined WaveNet maintains nearly unchanged performance across both horizons, suggesting that the cleaned time series preserves the main temporal patterns required for stable short-term forecasting.

3.3.3 Discussion

This section provides a critical analysis of the empirical results, highlighting trade-offs across models and examining their broader implications for smart building applications. We discuss generalization, robustness, deployment feasibility, and model behavior under realistic conditions.

Across both anomaly detection and downstream forecasting tasks, the Refined Gated WaveNet

consistently outperforms all baselines. This superior performance stems from targeted architectural enhancements , gated activation units, skip connections, and dilated causal convolutions which enable the model to capture long-range temporal dependencies while preserving essential signal structure. Notably, it improves the F1 Score by over 56 percentage points compared to the Original WaveNet, underscoring the impact of replacing autoregressive sampling with windowed regression and integrating temporal refinement.

Beyond energy consumption, the proposed framework can be applied to other smart building prediction tasks because it focuses on modeling temporal patterns rather than relying on application-specific assumptions. Anomaly detection is treated as a sequence modeling problem, where normal behavior is learned from historical data and deviations are detected through reconstruction or prediction errors. As a result, the same model structure and training procedure can be reused for different tasks by adjusting the input signals and the definition of normal operation, without changing the overall methodology.

The evaluation of forecasting performance further demonstrates the importance of robust anomaly detection. As observed in our experiments, introducing only 10% synthetic anomalies causes a substantial degradation in LSTM predictive accuracy, with the MSE rising from 0.00189 to 57.69. This result illustrates how even a small fraction of anomalies can distort the temporal learning process and significantly impair model performance. In practical settings, this could manifest as erroneous alerts, inefficient energy allocation, or increased operational costs due to reactive decision-making.

However, when anomalies are detected and corrected using Refined WaveNet, forecasting accuracy is effectively restored. The LSTM trained on Refined WaveNet-cleaned data achieves an MSE of 0.0026, nearly indistinguishable from the clean-data baseline. This highlights the detector's ability to preserve underlying dynamics while filtering out disruptive noise. The use of linear interpolation further facilitates the smooth reconstruction of missing values, ensuring consistency in the temporal signal used for training.

In contrast, conventional detectors such as DUAL and DGHL fail to provide meaningful improvement, resulting in residual anomalies and forecasting errors exceeding 27 and 35, respectively.

The limited performance of DGHL can be attributed to its reliance on posterior sampling and hierarchical latent structures, which require careful tuning and are highly sensitive to initialization—factors that undermine stability in outlier-heavy regimes. Meanwhile, DUAL's contrastive learning strategy depends on the design of effective pretext tasks, which is particularly challenging in unstructured, unlabeled energy data. As a result, its ability to generalize to unseen or context-dependent anomalies is significantly reduced. Similarly, low-capacity models such as LSTM and RNN used in self-detection configurations exhibit limited correction capability, emphasizing the limitations of non-specialized architectures in unsupervised anomaly detection.

From a deployment perspective, the Refined Gated WaveNet offers a favorable balance between performance and efficiency. Its lightweight design is well-suited for edge-based implementation, enabling real-time anomaly detection in smart building environments without the latency or infrastructure constraints of cloud-based systems. This makes it particularly relevant for applications such as predictive maintenance, dynamic pricing, and adaptive control systems in smart grids.

In addition, the framework could serve as a preprocessing component for control-oriented applications, such as reinforcement learning–based energy management, where reliable system state estimation is required.

These performance gains can be primarily attributed to the combination of windowed parallel regression, residual and skip connections, and gated activation units, which enable the model to capture both long-range and subtle temporal deviations while maintaining stable and interpretable training dynamics. Unlike fully autoregressive models, this design supports efficient learning across broader temporal contexts without compromising precision.

In summary, the Refined WaveNet architecture presents a scalable, modular, and resilient solution for time series anomaly detection. It not only delivers strong detection performance but also enhances the reliability of downstream forecasting models, making it a key enabler for intelligent, data-driven energy systems.

Chapter 4

A Unified Transformer VAE Architecture with FiLM Temporal Modulation for Deterministic and Probabilistic Load Forecasting

4.1 Introduction

Reliability in load forecasting is a key component of safe and effective power system operation. It supports long-term capacity development, reserve sizing, market bidding, and daily balancing Shahzad and Jasińska (2024). As heating and transportation become increasingly electrified, demand becomes more variable, and weather patterns become more unpredictable, system operators rely on forecasts that are informative over a wide range of time horizons, from multi-day scheduling to minute-ahead regulation X. Li and Jia (2024). At the same time, the growing penetration of variable renewable generation has introduced rapid, weather-driven swings in supply Borbáth and Van Hertem (2024), tightening the coupling between end-use consumption and stochastic production. When combined, these advancements highlight the need for models that can provide operating planning with reliable uncertainty estimates in addition to precise trajectories.

Forecasting electricity demand from a single series is, however, far from straightforward. The form and amplitude of a typical load profile's daily, weekly, and seasonal cycles change depending on daylight, occupancy, and operational procedures Laitsos et al. (2024). Beyond these regular patterns, structural drivers such as electrification, changes in economic conditions, and policy reforms shift the long-term level of demand and introduce non-stationarities that standard time-series models struggle to follow Wang et al. (2022). Short-lived shocks, for instance public holidays or abrupt temperature swings, further disturb the usual rhythm of electricity use and give rise to regime changes and structured residual patterns Buechler et al. (2022). Much of the behaviour behind these shifts is not directly observed in the data, so its influence has to be inferred from the way the load evolves over time Ullah et al. (2024). In practice, the forecaster only has access to a single observed series and must rely on it to detect context changes and to track evolving patterns over both short and long horizons J. Xu, Zheng, Dang, Yang, and Li (2025).

Classical statistical tools such as ARIMA, SARIMA, and exponential smoothing tend to work well for local trends and short-term seasonality, but they often fail once strong nonlinear effects or persistent structural changes appear Uzair, Shah, and Ali (2024). Recurrent neural networks, including LSTM and GRU, are meant to capture longer dependencies, yet when they are run recursively for multi-step forecasting, small errors accumulate and gradually degrade performance as the horizon grows Huang (2024). Dilated convolutional architectures widen the temporal receptive field and improve efficiency, but they are still less effective when the task requires relating distant seasonal patterns in a more explicit way F. Li, Guo, Han, Zhao, and Shen (2024). Transformers, by contrast, use self-attention to connect information over long time spans and have achieved strong performance in a range of forecasting tasks Liu (2025). Recent variants such as Informer, Autoformer, and FEDformer adapt this idea to large-scale sequence prediction Wu et al. (2021); H. Zhou et al. (2021); T. Zhou et al. (2022). Most of these models, however, are still fundamentally deterministic: they generate point forecasts and provide no direct notion of predictive uncertainty, which is problematic in risk-sensitive applications such as reserve scheduling and operational planning Dab et al. (2024).

In parallel with these developments, a line of work has turned to probabilistic forecasting, where the aim is to model full predictive distributions rather than a single future path Masood, Gantassi,

and Choi (2024). Variational Autoencoders (VAEs) fall into this category. By introducing a latent variable that can be sampled at prediction time, they provide a natural way to generate multiple plausible futures and to encode uncertainty in a compact form Leushuis (2025). On their own, however, VAEs do not encode the rich periodic and multi-scale structure that characterizes load data H. Xu, Boyaci, Lian, and Wilson (2025); time is mostly handled through the encoder rather than through a dedicated temporal inductive bias. This mismatch has motivated hybrid designs in which a Transformer backbone captures structured temporal dependencies, while the VAE component accounts for latent uncertainty. However, existing Transformer–VAE hybrids typically integrate the latent variable after representation learning, without allowing temporal context to directly influence the attention mechanism itself.

This study introduces a FiLM-conditioned Transformer–VAE architecture for univariate load forecasting with calendar, rolling, and seasonal awareness. The proposed model integrates attention-based sequence modelling, feature-wise temporal conditioning, and variational latent inference within a unified non-autoregressive framework. Rather than operating in a static embedding space, the proposed model reshapes token representations prior to self-attention, allowing temporal descriptors to directly influence similarity computation and attention weighting. At the same time, a variational latent component is injected into the encoder memory before decoding, enabling uncertainty to participate in representation learning rather than being appended post hoc. The resulting framework performs non-autoregressive multi-horizon forecasting and jointly optimizes deterministic accuracy and probabilistic calibration.

4.2 The proposed approach

The proposed FiLM Transformer VAE integrates temporal modulation, attention-based sequence modeling, and variational inference. An overview of the full architecture is provided in Fig. 4.1, while the training workflow is illustrated in Fig. 4.2.

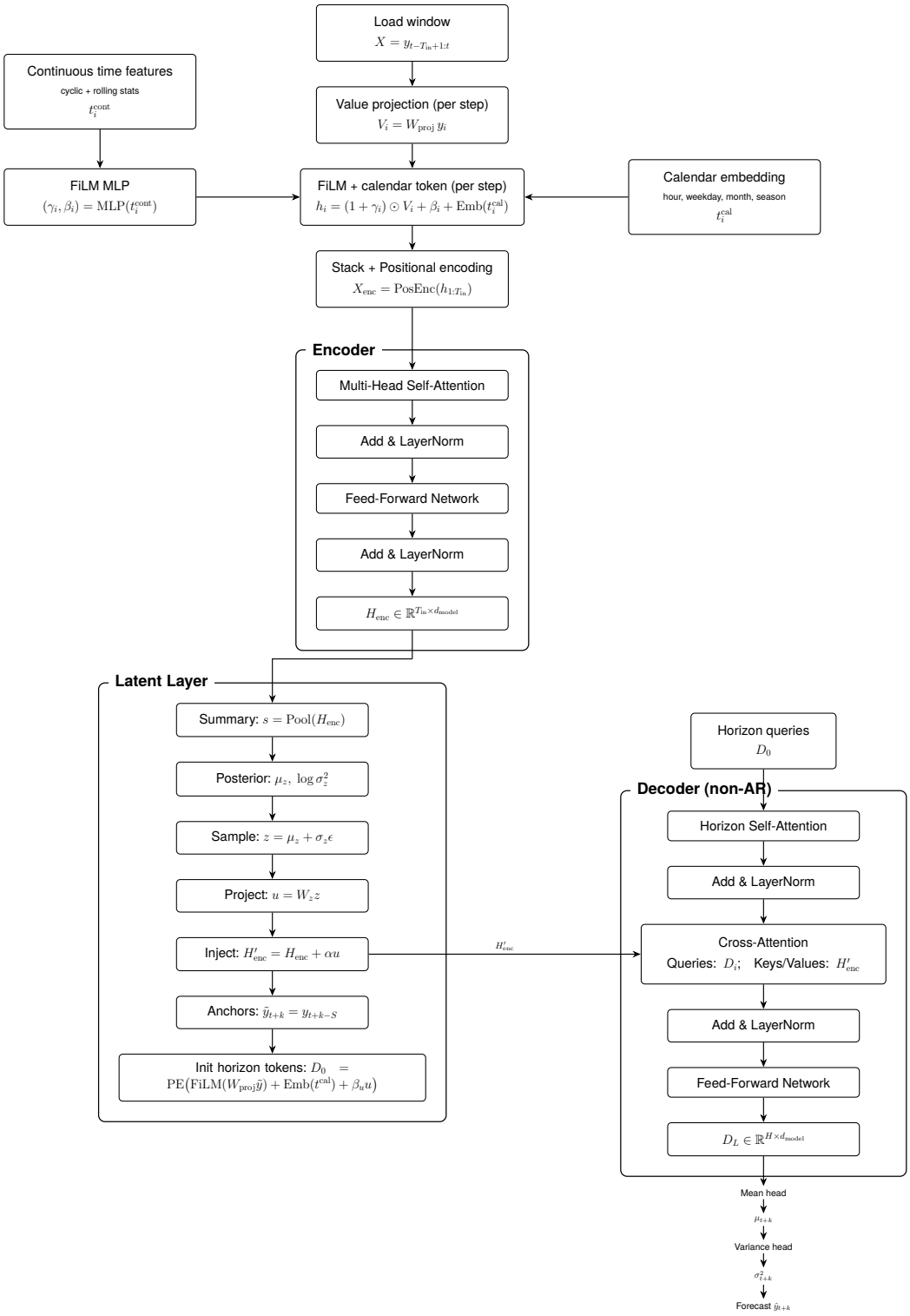


Figure 4.1: Architecture of the proposed FiLM Transformer VAE.

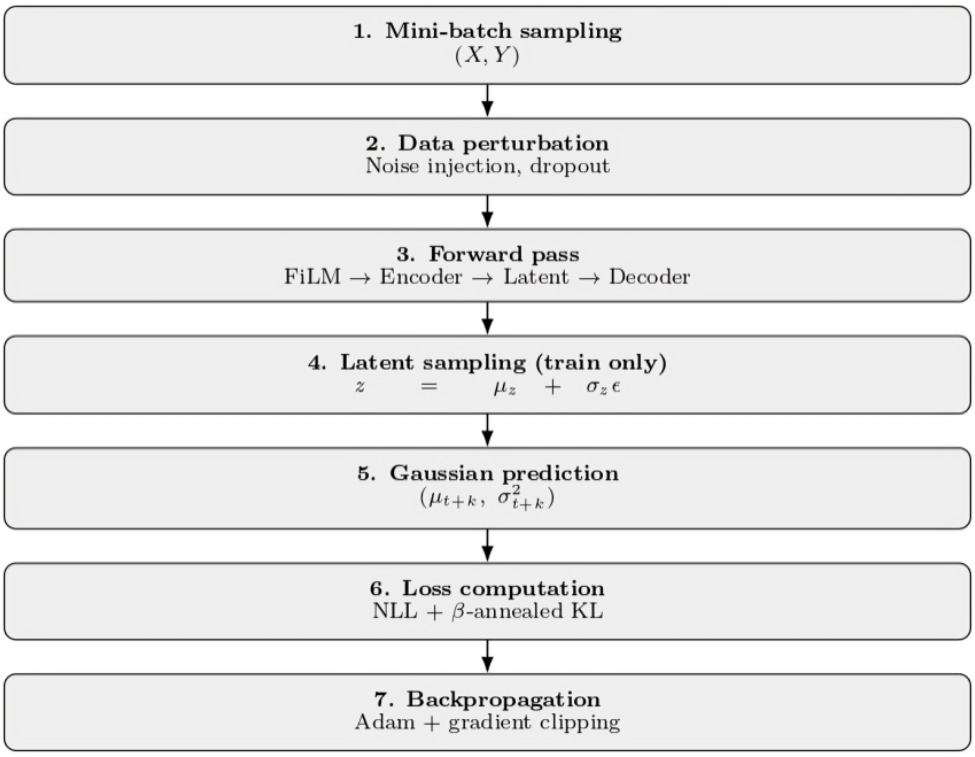


Figure 4.2: Training workflow of the FiLM Transformer VAE.

The proposed FiLM Transformer VAE integrates temporal modulation, attention-based sequence modeling, and variational inference. An overview of the full architecture is provided in Fig. 4.1, while the training workflow is illustrated in Fig. 4.2.

4.2.1 Problem Definition

Let $\{y_t\}_{t=1}^{T}$ denote a univariate electrical load series. At forecasting origin t, the model observes a historical window

$$X_t = (y_{t-T_{\text{in}}+1}, \ldots, y_t) \in \mathbb{R}^{T_{\text{in}}}. \tag{22}$$

For a mini-batch of size B, the input tensor is

$$X \in \mathbb{R}^{B \times T_{\text{in}} \times 1}.$$

The objective is to model the predictive distribution over the next H steps:

$$p_\Theta(y_{t+1:t+H} \mid X_t, c_t), \tag{23}$$

where c_t denotes temporal descriptors available at prediction time.

A latent-variable formulation is adopted:

$$p_\Theta(y_{t+1:t+H} \mid X_t, c_t) = \int p_\Theta(y_{t+1:t+H} \mid X_t, c_t, z)\, p(z)\, dz. \tag{24}$$

4.2.2 Temporal Representation and FiLM Conditioning

Each timestamp i is characterized by continuous descriptors $t_i^{\text{cont}} = \psi(i, y_{1:i})$ and categorical calendar attributes t_i^{cal}. The feature map $\psi(\cdot)$ includes cyclic encodings and rolling statistics computed without future leakage. Each scalar load value is projected to the model dimension:

$$V_i = W_{\text{proj}} y_i \in \mathbb{R}^{d_{\text{model}}}. \tag{25}$$

FiLM conditioning modulates token representations:

$$(\gamma_i, \beta_i) = \text{MLP}(t_i^{\text{cont}}), \qquad h_i^{(\text{film})} = (1 + \gamma_i) \odot V_i + \beta_i. \tag{26}$$

Calendar embeddings are added:

$$h_i' = h_i^{(\text{film})} + E_{\text{cal}}(t_i^{\text{cal}}). \tag{27}$$

Stacking tokens and adding positional encodings yields the encoder input:

$$X_{\text{enc}} = \text{PosEnc}(h_1', \ldots, h_{T_{\text{in}}}'). \tag{28}$$

4.2.3 Transformer Encoder

Let the encoder input be

$$H^{(0)} = X_{\text{enc}} \in \mathbb{R}^{B \times T_{\text{in}} \times d_{\text{model}}}. \tag{29}$$

The encoder consists of L stacked self-attention layers that model long-range temporal dependencies within the historical window. For layer $\ell = 1, \ldots, L$, the update is:

$$\tilde{H}^{(\ell)} = \mathrm{LN}\Big(H^{(\ell-1)} + \mathrm{MHSA}\Big(H^{(\ell-1)}\Big)\Big), \tag{30}$$

$$H^{(\ell)} = \mathrm{LN}\Big(\tilde{H}^{(\ell)} + \mathrm{FFN}\Big(\tilde{H}^{(\ell)}\Big)\Big). \tag{31}$$

Here, MHSA denotes multi-head self-attention applied along the temporal dimension T_{in}, while FFN is a position-wise feed-forward network. All operations are applied independently across the batch dimension B. The final encoder memory is

$$H_{\mathrm{enc}} = H^{(L)} \in \mathbb{R}^{B \times T_{\mathrm{in}} \times d_{\mathrm{model}}}. \tag{32}$$

4.2.4 Variational Latent Layer and Injection

To capture regime-level uncertainty beyond deterministic temporal patterns, a latent variable $z \in \mathbb{R}^{d_z}$ with prior $p(z) = \mathcal{N}(0, I)$ is introduced. The encoder memory is summarized using its last token:

$$s = H_{\mathrm{enc}}[:, T_{\mathrm{in}} - 1, :] \in \mathbb{R}^{B \times d_{\mathrm{model}}}. \tag{33}$$

The posterior is parameterized as

$$q_\phi(z \mid X) = \mathcal{N}(\mu_z, \mathrm{diag}(\sigma_z^2)), \tag{34}$$

with

$$\mu_z = W_\mu^z s, \qquad \log \sigma_z^2 = W_{\log v}^z s. \tag{35}$$

Sampling uses the reparameterization trick:

$$z = \mu_z + \sigma_z \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{36}$$

The latent vector is projected:

$$z_m = W_m z \in \mathbb{R}^{B \times d_{\text{model}}}. \tag{37}$$

To allow global regime modulation, the latent representation is injected additively into the encoder memory:

$$H'_{\text{enc}} = H_{\text{enc}} + \alpha \, \mathbf{1}_{T_{\text{in}}} z_m, \tag{38}$$

where $\mathbf{1}_{T_{\text{in}}} \in \mathbb{R}^{T_{\text{in}} \times 1}$.

This additive bias enables coherent adjustment of all historical representations without altering their temporal resolution.

4.2.5 Non-Autoregressive Transformer Decoder

Let the initial horizon representation be

$$D^{(0)} = D_0 \in \mathbb{R}^{B \times H \times d_{\text{model}}}, \tag{39}$$

where H is the forecasting horizon. The decoder consists of L stacked Transformer layers. For layer $\ell = 1, \ldots, L$, the update proceeds as follows.

(1) Masked Self-Attention

$$\tilde{D}^{(\ell,1)} = \text{LN}\Big(D^{(\ell-1)} + \text{MaskedMHSA}\Big(D^{(\ell-1)}\Big)\Big), \tag{40}$$

where masked multi-head self-attention operates along the horizon dimension H.

(2) Cross-Attention with Encoder Memory

$$\tilde{D}^{(\ell,2)} = \text{LN}\Big(\tilde{D}^{(\ell,1)} + \text{CrossAttn}\Big(\tilde{D}^{(\ell,1)}, \, H'_{\text{enc}}\Big)\Big), \tag{41}$$

where queries are computed from $\tilde{D}^{(\ell,1)} \in \mathbb{R}^{B\times H\times d_{\text{model}}}$, and keys/values are computed from the latent-conditioned encoder memory $H'_{\text{enc}} \in \mathbb{R}^{B\times T_{\text{in}}\times d_{\text{model}}}$.

(3) Position-wise Feed-Forward Network

$$D^{(\ell)} = \mathrm{LN}\Big(\tilde{D}^{(\ell,2)} + \mathrm{FFN}\Big(\tilde{D}^{(\ell,2)}\Big)\Big). \tag{42}$$

All operations are applied independently across the batch dimension B. After L layers, the final decoder representation is

$$D_L = D^{(L)} \in \mathbb{R}^{B\times H\times d_{\text{model}}}. \tag{43}$$

Masked self-attention captures dependencies across future steps within the horizon, while cross-attention conditions predictions on the latent-modulated historical memory.

4.2.6 Gaussian Predictive Layer

The final decoder representation

$$D_L \in \mathbb{R}^{B\times H\times d_{\text{model}}}$$

is mapped to Gaussian parameters through linear projections applied independently at each horizon step. For batch element b and horizon index k:

$$\mu_{b,k} = W_\mu^y D_L[b,k,:], \qquad a_{b,k} = W_{\log v}^y D_L[b,k,:]. \tag{44}$$

where $W_\mu, W_\sigma \in \mathbb{R}^{1\times d_{\text{model}}}$. The predictive variance is enforced positive via

$$\sigma_{b,k}^2 = \mathrm{Softplus}(\log \sigma_{b,k}^2) + \varepsilon, \qquad \varepsilon > 0. \tag{45}$$

Thus, conditioned on the latent variable z, the predictive distribution factorizes across the horizon:

$$p_\Theta(y_{t+1:t+H} \mid X, z) = \prod_{k=1}^{H} \mathcal{N}\big(y_{t+k}; \mu_k, \sigma_k^2\big)\,. \tag{46}$$

4.2.7 Training Objective

Training minimizes the negative ELBO:

$$\mathcal{L} = \mathcal{L}_{\text{NLL}} + \lambda_{\text{KL}}(t)\mathcal{L}_{\text{KL}} + \lambda \mathcal{L}_{\text{smooth}}. \tag{47}$$

The Gaussian negative log-likelihood is:

$$\mathcal{L}_{\text{NLL}} = \frac{1}{H}\sum_{k=1}^{H} \frac{1}{2}\left[\log \sigma_{t+k}^2 + \frac{(y_{t+k} - \mu_{t+k})^2}{\sigma_{t+k}^2}\right]. \tag{48}$$

The KL divergence regularizes the latent posterior toward the standard normal prior. A smoothness penalty enforces temporal coherence of predictive uncertainty.

$$\mathcal{L}_{\text{smooth}} = \frac{1}{H-1}\sum_{k=1}^{H-1} (\sigma_{t+k} - \sigma_{t+k+1})^2\,. \tag{49}$$

Optimization uses Adam with gradient clipping and early stopping.

4.2.8 Inference

Deterministic forecasts use the posterior mean:

$$\hat{y}_{t+k} = \mu_{t+k}. \tag{50}$$

Probabilistic forecasting samples latent variables:

$$z^{(s)} = \mu_z + \sigma_z \odot \epsilon^{(s)}, \tag{51}$$

and generates trajectories:

$$y_{t+1:t+H}^{(s)} = \mu^{(s)} + \sigma^{(s)} \odot \eta^{(s)}. \tag{52}$$

Repeated latent sampling yields a mixture of Gaussian trajectories approximating the full predictive distribution.

4.3 Experimental setup and results

4.3.1 Feature Construction and Data Preparation

Each timestamp t is augmented with two types of auxiliary features: (i) a continuous feature vector used for FiLM modulation and (ii) discrete calendar indices mapped to learnable embeddings.

Continuous FiLM features. We construct a context vector $t_t^{\text{cont}} \in \mathbb{R}^{d_c}$ that captures periodic structure, local statistics, and short-term spike behavior:

$$t_t^{\text{cont}} = \big[\phi_h(t),\ \phi_d(t),\ \phi_m(t),\ \mu_t^{(w)},\ \sigma_t^{(w)},\ y_{t-d},\ y_{t-\omega},\ \max_{j\in[0,r-1]} y_{t-j}\big]. \tag{53}$$

Periodic components are encoded through sinusoidal mappings. For any periodic scalar p_t with period P,

$$\phi(p_t; P) = \big[\sin(2\pi p_t/P),\ \cos(2\pi p_t/P)\big]. \tag{54}$$

Specifically, $\phi_h(t) = \phi(h_t; 24)$, $\phi_d(t) = \phi(d_t; 7)$, and $\phi_m(t) = \phi(m_t; 12)$. For 30-minute data, $h_t = \text{hour}(t) + \text{minute}(t)/60$. Rolling statistics are computed over a one-day window of length w:

$$\mu_t^{(w)} = \frac{1}{w}\sum_{j=0}^{w-1} y_{t-j}, \qquad \sigma_t^{(w)} = \sqrt{\frac{1}{w}\sum_{j=0}^{w-1}\Big(y_{t-j} - \mu_t^{(w)}\Big)^2}. \tag{55}$$

For the office dataset (30-minute sampling), we set $w = 48$ (one day), $d = 48$ (same time previous day), $\omega = 336$ (same time previous week), and $r = 12$.

Calendar embeddings. Discrete indices are extracted as

$$t_t^{\text{cal}} = (\text{hour}_t,\ \text{dow}_t,\ \text{month}_t,\ \text{season}_t), \tag{56}$$

with $\text{season}_t \in \{0, 1, 2, 3\}$ (quarter-of-year), and embedded via $\text{Emb}(\cdot)$.

Train-only normalization. Load values are min–max scaled to $[0, 1]$ using training statistics only:

$$y'_t = \frac{y_t - y^{\mathrm{tr}}_{\min}}{y^{\mathrm{tr}}_{\max} - y^{\mathrm{tr}}_{\min}}, \qquad y^{\mathrm{tr}}_{\min} = \min_{t \in \mathcal{T}_{\mathrm{tr}}} y_t, \quad y^{\mathrm{tr}}_{\max} = \max_{t \in \mathcal{T}_{\mathrm{tr}}} y_t, \tag{57}$$

and the same parameters are applied unchanged to validation and test segments. Continuous features t^{cont}_t are also min–max normalized using training-only extrema.

Windowing and splits. From $\{y'_t\}$ we build supervised pairs (X_i, Y_i):

$$X_i = y'_{i:i+T_{\mathrm{in}}-1} \in \mathbb{R}^{T_{\mathrm{in}}}, \qquad Y_i = y'_{i+T_{\mathrm{in}}:i+T_{\mathrm{in}}+H-1} \in \mathbb{R}^{H}. \tag{58}$$

For the office dataset, $T_{\mathrm{in}} = 96$ (48 hours) and $H = 48$ (24 hours). The time series is split chronologically into 70% training, 15% validation, and 15% testing. Training and validation windows are generated with stride 1. Validation windows are also generated with stride 1, with forecast targets constrained to lie fully within the validation segment while allowing historical context from the training period. Test windows use stride H to form non-overlapping evaluation windows fully contained in the test segment.

4.3.2 Training Configuration

Training and architectural hyperparameters for both datasets are summarized in Table 4.1.

All baseline models (Transformer, LSTM, VAE, and TCN variants) were trained under identical experimental conditions to ensure comparability with the proposed FiLM-conditioned Transformer–VAE. Specifically, they used the same input window T_{in} and forecast horizon H, identical chronological data splits (70% training, 15% validation, 15% testing), and the same min–max normalization computed from the training set and applied to all splits. For each baseline, hyperparameters were selected using the validation set, and training relied on the same early-stopping criterion as the FiLM model. Therefore, differences in performance across models reflect architectural design rather than discrepancies in preprocessing, or evaluation setup.

Table 4.1: Dataset-specific hyperparameters for the two experimental settings. Shared settings are listed separately.

Parameter	Office (30-min)	PJME (hourly)
(A) Dataset-specific forecasting setup		
Input length T_{in}	96	48
Forecast horizon H	48	24
Sampling resolution	30 min	1 hour
(B) Dataset-specific model capacity		
d_{model}	64	128
FFN dimension	256	512
Encoder layers	4	4
Decoder layers	4	4
Attention heads	4	4
Latent dimension z	32	32
(C) Shared training configuration (identical across datasets)		
Epochs	40	
Batch size	32	
Optimizer	Adam	
Learning rate	1×10^{-3}	
Dropout	0.1	
Gaussian noise injection σ	0.01	
Gradient clipping $\|g\|_2$	0.5	
Early stopping	Yes (validation loss)	
β-annealing (KL)	Linear warm-up to $\beta = 0.4$ over the first 8 epochs	
Normalization	Min–max (fit on train, applied to val/test)	
Split (train/val/test)	70/15/15 (chronological)	
Variance head	Softplus	

4.3.3 Evaluation Metrics and Protocol

We assessed the model in both deterministic and probabilistic forecasting modes:

- Deterministic mode: fixing $z = \mu_z$ gives the expected trajectory $\hat{y}_{t+1:t+H} = \mu_{t+1:t+H}$;
- Probabilistic mode: sampling $z^{(s)} \sim \mathcal{N}(\mu_z, \sigma_z^2)$ produces an ensemble $\{\hat{y}_t^{(s)}\}_{s=1}^{S}$, from which prediction intervals such as P10–P90 are estimated.

(a) Deterministic evaluation metrics Let y_t and $\hat{y}_t$ denote the observed and predicted values over N forecasted time steps. Deterministic accuracy is evaluated using the following metrics.

Root Mean Squared Error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2}. \tag{59}$$

Mean Absolute Error (MAE):

$$\mathrm{MAE} = \frac{1}{N}\sum_{t=1}^{N}|y_t - \hat{y}_t|. \tag{60}$$

Normalized RMSE (NRMSE):

$$\mathrm{NRMSE} = \frac{\mathrm{RMSE}}{y_{\max} - y_{\min}} \times 100, \tag{61}$$

where $y_{\max}$ and $y_{\min}$ are computed from the test segment. NRMSE provides scale-independent comparison across datasets.

Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum_{t=1}^{N}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{N}(y_t - \bar{y})^2}, \tag{62}$$

where $\bar{y}$ denotes the empirical mean of observed values. Together, these metrics quantify average deviation, relative accuracy, and the proportion of variance explained by point forecasts.

(b) Probabilistic metrics For probabilistic evaluation, $\hat{y}_t^{(s)}$ denotes the s-th Monte-Carlo sample of the predictive distribution, and $[\hat{y}_t^L, \hat{y}_t^U]$ represent the lower and upper prediction bounds (e.g., the 10th and 90th percentiles). Two standard metrics were used.

The metric Prediction Interval Coverage Probability (PICP) evaluates *coverage*, i.e., the proportion of true observations that fall inside the predicted interval:

$$\mathrm{PICP} = \frac{1}{N}\sum_{t=1}^{N}\mathbb{I}\big(\hat{y}_t^L \le y_t \le \hat{y}_t^U\big)\,, \tag{63}$$

where $\mathbb{I}(\cdot)$ is the indicator function. A well-calibrated model should yield PICP values close to the nominal coverage level of 0.90.

Table 4.2: Deterministic forecasting performance on the Office Load (kW) and PJME (MW) datasets. Reported values are rolling-origin mean ± standard deviation when applicable.

Model	RMSE	NRMSE	MAE	R^2	Ref.
Office Load dataset (kW)					
LSTM	175.49	15.82	94.90	0.70	Rafi et al. (2021)
TCN	170.49	15.37	94.23	0.72	Shaikh et al. (2023)
DenseVAE	166.77	15.03	82.73	0.73	Bond-Taylor et al. (2022)
DLinear	169.65	15.29	100.34	0.72	Toner and Darlow (2024)
Base Transformer VAE	150.89	13.59	78.85	0.78	Mentzelopoulos et al. (2024)
FiLM Transformer VAE (Ours)	41.42	3.73	27.29	0.98	Ours
Rolling-origin (mean ± std)	40.08 ± 6.41	4.05 ± 0.85	26.86 ± 2.89	0.98 ± 0.006	–
PJME dataset (MW)					
LSTM	1932.03	5.17	1380.71	0.90	Rafi et al. (2021)
TCN	1733.25	4.64	1240.31	0.92	Shaikh et al. (2023)
DenseVAE	2210.63	5.91	1639.06	0.88	Bond-Taylor et al. (2022)
DLinear	2119.89	5.68	1517.90	0.89	Toner and Darlow (2024)
Base Transformer VAE	1955.34	5.23	1402.01	0.90	Mentzelopoulos et al. (2024)
FiLM Transformer VAE (Ours)	376.59	1.00	249.78	0.99	Ours
Rolling-origin (mean ± std)	360.94 ± 101.16	1.66 ± 0.74	249.12 ± 64.60	0.99 ± 0.007	–

CRPS (continuous Ranked Probability Score) measures both sharpness (narrowness of the predicted distribution) and calibration (alignment with the true outcome): Under a Gaussian predictive distribution with mean μ_t and standard deviation σ_t, CRPS admits the closed-form expression:

$$\mathrm{CRPS} = \frac{1}{N}\sum_{t=1}^{N}\sigma_t\left[z_t\left(2\Phi(z_t)-1\right)+2\phi(z_t)-\frac{1}{\sqrt{\pi}}\right], \tag{64}$$

where

$$z_t = \frac{y_t-\mu_t}{\sigma_t}, \tag{65}$$

and $\Phi(\cdot)$ and $\phi(\cdot)$ denote the cumulative distribution function and probability density function of the standard normal distribution, respectively. Lower CRPS values indicate better probabilistic forecasts. We used a rolling-origin evaluation with five non-overlapping test windows to capture seasonal and structural differences. Final results are reported as the mean ± standard deviation across these windows.

4.4 Results and Discussion

4.4.1 Deterministic Forecasting Performance

The quantitative results summarized in Table 4.2 show that the proposed FiLM Transformer VAE achieves the highest deterministic accuracy on both datasets. Recurrent and convolutional approaches such as LSTM and TCN still exhibit relatively large errors, around 170 kW for the Office Load series and close to 1900 MW for the PJM dataset, with R^2 scores remaining below 0.92. This behaviour indicates that classical recurrent and convolutional sequence models struggle to capture long-range dependencies and scale variations present in realistic load profiles.

The proposed DLinear model decomposes the input series into trend and seasonal components and applies independent linear projections for forecasting. On the Office Load dataset, it achieves an RMSE of 169.65 kW, an MAE of 96.41 kW, an NRMSE of 15.21%, and an R^2 score of 0.71. On the PJME dataset, it records an RMSE of 2119.89 MW, an MAE of 1517.90 MW, an NRMSE of 5.68%, and an R^2 score of 0.89. These results are comparable to recurrent baselines such as LSTM, confirming that a significant portion of load variability can be explained through linear extrapolation of smoothed and residual components. However, its fully linear structure limits its capacity to learn adaptive representations or dynamically incorporate contextual information. As a result, DLinear remains less effective in capturing nonlinear demand patterns, calendar-driven effects, and regime shifts compared to the proposed FiLM Transformer VAE.

By contrast, the base Transformer VAE already represents a clear methodological step forward. Thanks to the combination of self-attention and a latent probabilistic layer, it can model a broader range of temporal dynamics and variability. On the Office dataset, it achieves an RMSE of approximately 167 kW ($R^2 = 0.73$), and on PJM around 1955 MW ($R^2 = 0.90$). Crucially, the same architecture maintains competitive performance across two datasets with significantly different temporal precision and magnitude, demonstrating its potential for use in a variety of forecasting scenarios. Adding FiLM-based temporal conditioning on top of this backbone, together with calendar embeddings, leads to a substantial further improvement in accuracy.

The resulting FiLM Transformer VAE records an RMSE of only 41.4 kW ($R^2 = 0.98$) for the Office dataset and 376.6 MW ($R^2 = 0.99$) for PJM. The rolling-origin averages in Table 4.2

Table 4.3: Probabilistic forecasting performance on the Office Load (kW) and PJME (MW) datasets. Coverage refers to prediction-interval coverage probability, and CRPS denotes the continuous ranked probability score. Rolling-origin results are reported as mean ± standard deviation.

Model	**Coverage**	**CRPS**	**Ref.**
Office Load dataset (kW)			
DeepAR	0.91	69.44	Salinas et al. (2020)
TFT	0.84	177.85	Ferreira et al. (2025)
Base Transformer–VAE	0.65	66.13	Mentzelopoulos et al. (2024)
FiLM Transformer VAE (Ours)	0.91	21.82	Ours
Rolling-origin (mean ± std)	0.917 ± 0.024	21.708 ± 1.763	–
PJME dataset (MW)			
DeepAR	0.72	1111.22	Salinas et al. (2020)
TFT	0.34	1171.78	Ferreira et al. (2025)
Base Transformer VAE	0.62	1065.60	Mentzelopoulos et al. (2024)
FiLM Transformer VAE (Ours)	0.90	206.69	Ours
Rolling-origin (mean ± std)	0.90 ± 0.052	206.33 ± 44.06	–

(40.08 ± 6.41 kW and 360.94 ± 101.16 MW, respectively) attest to the stability of these gains when assessed over several forecast periods. FiLM conditioning significantly reduces reconstruction error and improves temporal alignment, as shown in Figure 4.3, indicating the model's capacity to adjust to daily, weekly, and seasonal variations in demand. Overall, the deterministic analysis reveals a forecasting framework that goes beyond traditional recurrent or attention-only architectures while delivering high accuracy and robust performance.

4.4.2 Probabilistic Forecasting and Calibration

The capacity of the proposed FiLM Transformer VAE to provide prediction distributions that are both accurate and reasonably well calibrated is demonstrated by the probabilistic performance metrics listed in Table 4.3. The calibration quality of classical probabilistic baselines like DeepAR and TFT is remains limited. Their coverage on the Office Load dataset is at or below 0.9, and their CRPS values are relatively high (69.44 for DeepAR and 177.85 for TFT), indicating that the uncertainty estimates are either too diffuse or not adequately linked with the observed variability. TFT in particular shows very low coverage (down to 0.34) and CRPS values surpassing 1100 MW, indicating diminished robustness under large-scale demand shifts. The PJM dataset shows a similar

pattern.

An obvious improvement is already present in the base Transformer VAE. By adding a variational latent layer, it reduces CRPS to 1065.60 on PJM and 66.13 on the Office dataset, explicitly accounting for epistemic uncertainty. Nevertheless, under-coverage (0.65 on Office and 0.62 on PJM) results from this improvement, indicating that uncertainty is still not sufficiently adaptive when demand fluctuates suddenly.

The explicit temporal conditioning introduced by the FiLM approach greatly improves the model's probabilistic forecasting behavior. The FiLM Transformer VAE obtains a coverage of 0.91 with a significantly reduced CRPS of 21.82 on the Office Load dataset and a coverage of 0.90 with a CRPS of 206.69 on the PJM dataset.

The stability of these probabilistic outcomes is validated using rolling-origin evaluation. The model achieves a CRPS of 206.33 ± 44.06 on the PJM dataset. The Office Load dataset has a CRPS of 21.708 ± 1.763 and an average coverage of 0.917 ± 0.024. These results indicate steady performance over several forecast windows.

Figures 4.4 and 4.3 highlight the benefits of applying the same temporal conditioning mechanism to both probabilistic outputs and point forecasts. FiLM improves the distribution of uncertainty across time by allowing the latent distribution to adjust to both transient aberrations and recurring regimes by integrating time-related signals into intermediate representations. Prediction intervals become sharper while maintaining better calibration, which results in higher coverage and lower CRPS. The proposed framework's combination of precision, calibration, and adaptability makes it a strong fit in operational forecasting scenarios where accurate point prediction and reliable uncertainty quantification are equally critical. While the model's improved calibration and forecasting accuracy is supported by the quantitative results, a closer examination of its internal mechanisms reveals the causes of these improvements.

While the quantitative results confirm the model's superior forecasting accuracy and calibration, a closer examination of its internal components helps clarify which mechanisms drive these improvements.

In a reliability diagram, the horizontal axis shows the nominal coverage of the prediction intervals, that is, the confidence level that the model claims (for example, a 90% interval is supposed

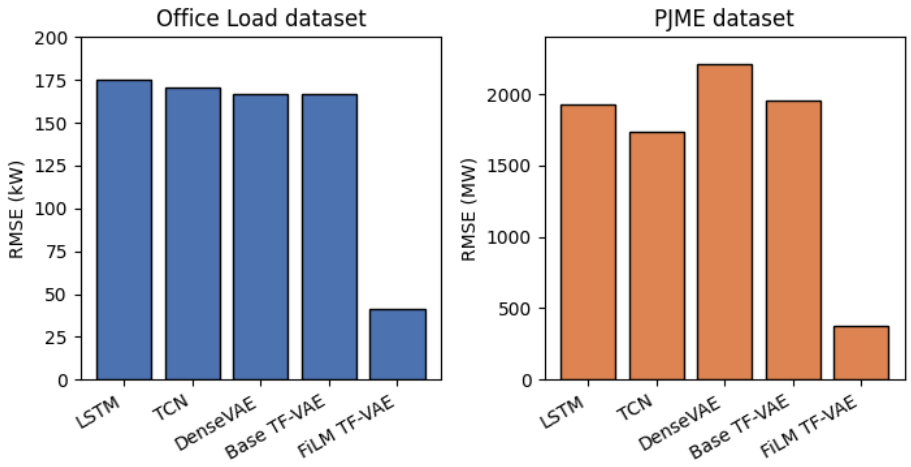


Figure 4.3: Deterministic forecasting performance of the compared models on the Office Load (kW) and PJME (MW) datasets.

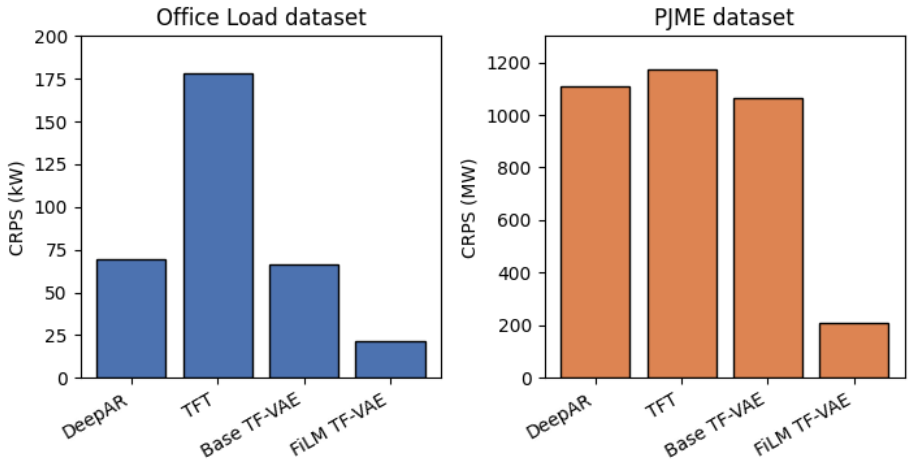


Figure 4.4: Probabilistic forecasting performance of the compared models on the Office Load (kW) and PJME (MW) datasets.

to contain the true value in 90% of the cases). The vertical axis reports the empirical coverage, computed as the fraction of test points for which the observed load actually falls inside the corresponding interval. The diagonal line $y = x$ marks perfect calibration: nominal and empirical coverage coincide at every level. If the curve of a model lies below this line, the intervals are too narrow and the forecasts are overconfident; if it lies above, the intervals are too wide and the model is overly conservative. The position of the FiLM Transformer VAE curve in Figure 4.5 indicates that its prediction intervals follow the diagonal quite closely, which suggests that the associated

Table 4.4: Ablation study on the impact of FiLM and calendar embeddings. Results are reported on the non-overlapping Office Load test set (kW scale).

Model configuration	RMSE	MAE	NRMSE	R^2	Coverage	CRPS
FiLM Transformer VAE (FiLM + calendar; full model)	41.42	27.29	3.73	0.98	0.91	21.82
FiLM disabled (no feature-wise modulation)	87.46	48.60	7.88	0.92	0.83	38.75
Calendar embeddings removed (no temporal context)	40.79	30.78	3.67	0.98	0.88	22.49

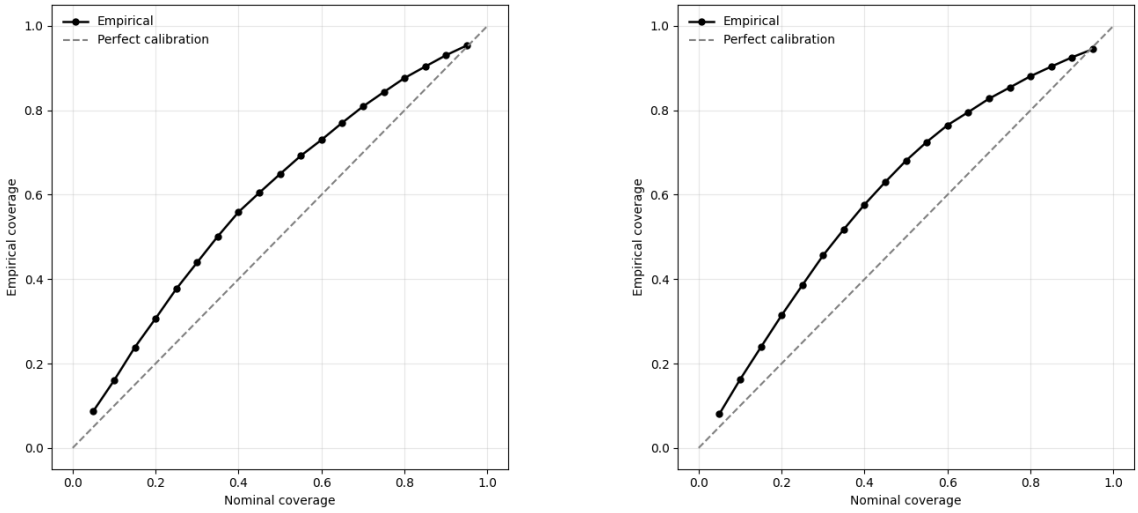


Figure 4.5: Reliability diagrams for probabilistic forecasts. Empirical coverage is plotted against nominal coverage for the Office Load (left) and PJM (right) datasets.

uncertainty estimates are reasonably well calibrated.

4.4.3 Ablation Study

Tables 4.4 and 4.5 provide a clearer picture of how each architectural component contributes to the overall performance of the model. Removing FiLM leads to a substantial and consistent degradation across both datasets. On the PJME dataset, RMSE increases from 376.59 MW to 1939.30 MW, coverage drops from 0.90 to 0.35, and CRPS rises from 206.69 to 1205.04. The Office dataset exhibits the same tendency, although at a smaller numerical scale. The magnitude of these changes indicates that FiLM is not acting as a secondary refinement but is instead deeply embedded in the representation process.

This behaviour can be understood by considering how FiLM interacts with the attention backbone. Before self-attention is computed, FiLM adjusts hidden features using temporal descriptors. As a result, the same historical signal can be represented differently depending on whether it corresponds to a peak hour, a weekend, or a seasonal transition. Since attention weights are driven by feature similarity, even moderate shifts in feature scale or offset can change which time steps influence one another most strongly. In practice, this enables the network to respond differently to recurring patterns that occur under distinct temporal conditions.

When FiLM is removed, the Transformer must rely on a single shared representation space to explain heterogeneous demand regimes. This reduces its ability to separate weekday and weekend dynamics, distinguish peak from off-peak behaviour, or accommodate seasonal amplitude variations. The sharp decline in coverage and the increase in CRPS suggest that uncertainty estimates no longer adapt properly to regime transitions, often resulting in intervals that are too narrow during structural changes in demand.

By contrast, removing calendar embeddings produces a more moderate effect. Deterministic accuracy remains close to that of the full model, but coverage decreases and CRPS increases on both datasets. Calendar information therefore appears to contribute mainly to temporal alignment and stability across repeated cycles rather than to fine-grained adaptation. Without this longer-horizon reference, the model remains flexible but becomes slightly less consistent in its uncertainty estimates.

Taken together, the ablation results suggest complementary roles for these two mechanisms. FiLM enables context-dependent shaping of internal representations, while calendar embeddings provide structural temporal anchors. Their combination allows the model to remain responsive to local variations while maintaining stable behaviour across longer-term cycles. Examining the temporal evolution of forecasts further illustrates how these two components interact under changing load regimes.

Beyond aggregated metrics, examining the temporal evolution of the forecasts provides additional insight into how FiLM conditioning and temporal embeddings interact to shape predictive behavior under different load regimes.

Table 4.5: Ablation study on the impact of FiLM and calendar embeddings. Results are reported on the non-overlapping PJME test set (MW scale).

Model configuration	RMSE	MAE	NRMSE	R^2	Coverage	CRPS
FiLM Transformer VAE (FiLM + calendar; full model)	376.59	249.78	1.008	0.99	0.90	206.69
FiLM disabled (no feature-wise modulation)	1939.30	1425.94	5.19	0.90	0.35	1205.04
Calendar embeddings removed (no temporal context)	553.65	393.47	1.48	0.99	0.79	300.1

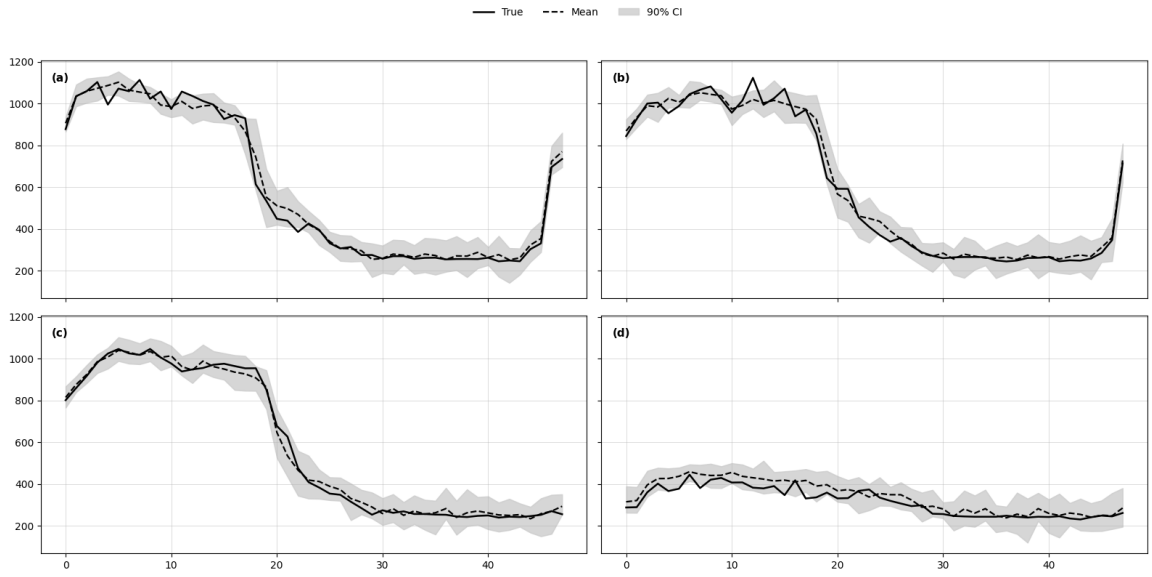


Figure 4.6: Probabilistic multi-day load forecasts on the Office Load dataset.

4.4.4 Qualitative Forecast Analysis

Figures 4.6 and 4.7 present representative multi-horizon probabilistic forecasts obtained with the proposed FiLM conditioned Transformer VAE on the Office and PJM Load datasets, respectively. Each panel reports the ground truth series, the predictive mean, and the associated 90% predictive interval.

The predictive mean closely tracks changes in power consumption over time across all shown timeframes. The model captures key events, such as sudden increases in load followed by decreases, and accurately reproduces the main cyclical swings. Long-range interdependence and short-term dynamics are both adequately captured by the lack of observable temporal lag. The PJM dataset provides a clear illustration of this pattern, as the forecasts capture both mid-horizon declines and

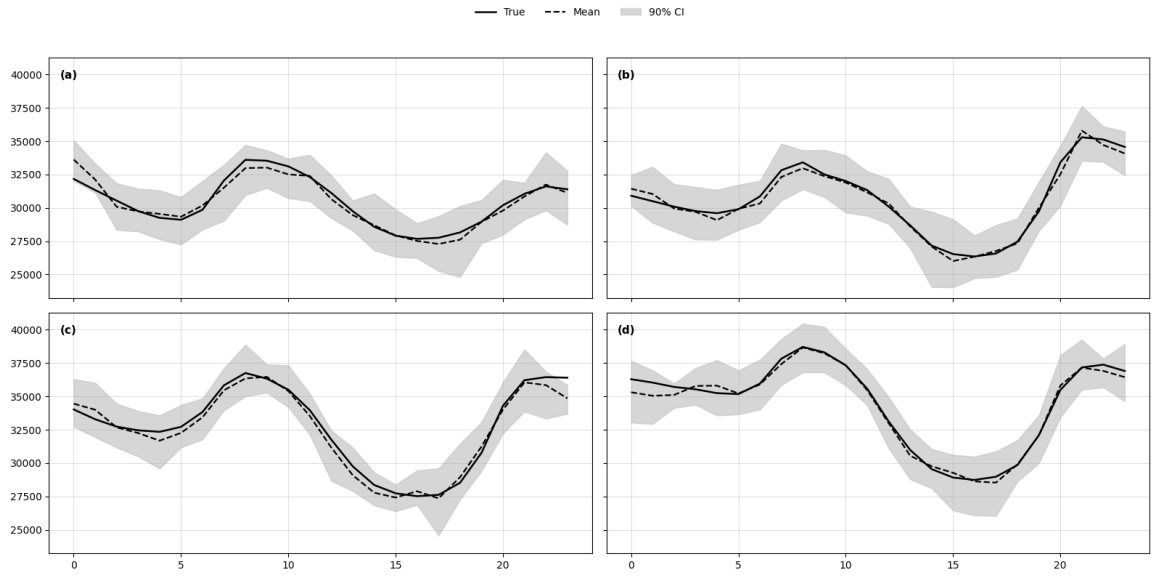


Figure 4.7: Probabilistic multi-day load forecasts on the PJM Load dataset

late-horizon recoveries.

Predictive uncertainty clearly varies over time. The width of the prediction intervals increases with rapid fluctuations in demand and decreases with more stable regime. While uncertainty expansion is particularly apparent during abrupt upward climbs in the PJM cases, wider intervals are observed during recovery phases and during transitions between high-load and low-load regimes in the Office Load examples. Long low-variability segments, in contrast, are associated with comparatively narrow intervals. This heteroscedastic pattern suggests that the model does not rely on a constant noise assumption, but instead learns input-dependent variance.

From a calibration perspective, most observed values fall within the 90% predictive intervals across the scenarios shown, with deviations occurring primarily around abrupt transitions. It is significant that the intervals do not become unduly broad during steady times, indicating that coverage is accomplished without using unduly conservative uncertainty estimates. These qualitative results are consistent with the quantitative coverage and CRPS results that were shown in the preceding section.

Taken together, the visual analysis indicates that the proposed approach yields accurate point forecasts alongside flexible, well-calibrated uncertainty estimates. Combining variational latent

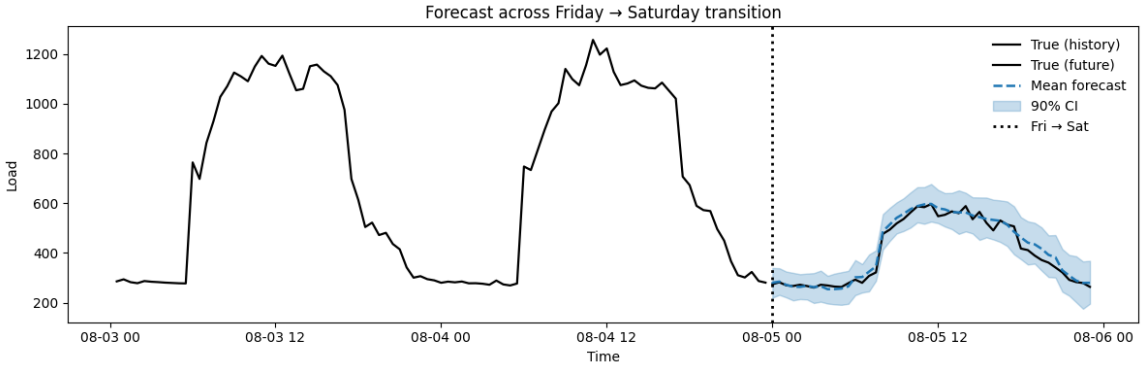


Figure 4.8: Weekday versus weekend probabilistic forecasting on the Office Load dataset.

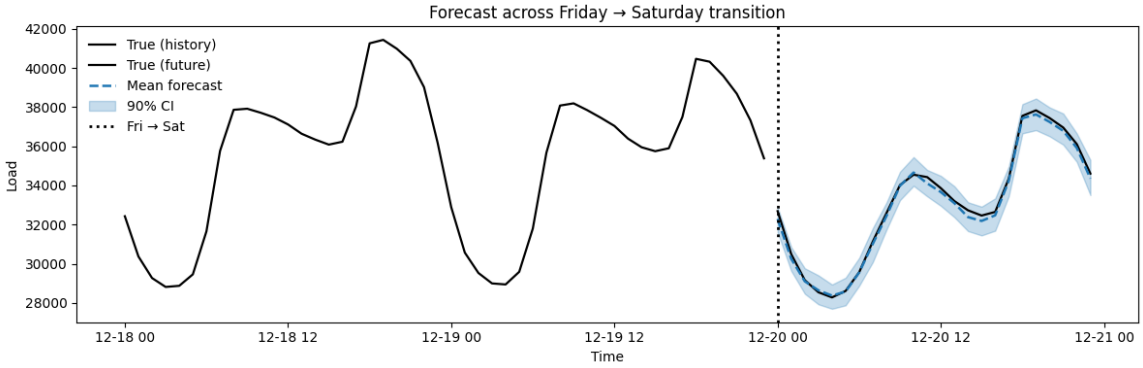


Figure 4.9: Weekday versus weekend probabilistic forecasting on the PJM Load dataset.

representations with FiLM-based temporal modulation enables the model to adjust predictive dispersion in response to regime changes, which is important for reliable energy demand forecasting at both building and grid scales.

The distinction between weekday and weekend patterns, shown in Figures 4.8 and 4.9, provides further evidence of contextual adaptation. During weekends, load curves become more irregular as occupancy declines and HVAC schedules vary. In response, predictive intervals widen to capture greater behavioral uncertainty, while mean predictions continue to follow realistic load trajectories. Once weekday operations resume, the intervals contract, indicating restored regularity. Remarkably, This behaviour emerges without exogenous weather inputs. It is driven by calendar embeddings and FiLM modulation derived from time descriptors." This implicit temporal awareness enhances interpretability and supports practical decision-making scenarios requiring reliable

uncertainty quantification.

Overall, the qualitative behavior aligns with the quantitative gains reported in Tables 4.2 and 4.3. Taken together, these findings indicate that FiLM-based temporal modulation not only enhances predictive accuracy but also introduces interpretability and behavioral awareness qualities essential for deploying probabilistic forecasting models in real-world energy systems.

4.4.5 Discussion

The combination of Transformer attention, variational inference, and FiLM-based temporal conditioning enables the model to unify deterministic and probabilistic forecasting within a single framework. Compared with deterministic baselines, it delivers more accurate and consistent load trajectories, while its probabilistic head provides well-calibrated uncertainty estimates. Unlike earlier hybrid approaches in which the variational component is appended post hoc, the proposed design integrates stochasticity directly into the attention pathway through direct latent injection into the encoder memory prior to decoding, improving both computational efficiency and interpretability.

The FiLM mechanism acts as a bridge between temporal descriptors and model activations, allowing the network to adapt continuously to periodic and contextual variations without retraining. This yields forecasts that are sharper, more stable, and sensitive to behavioral changes, particularly during regime shifts such as seasonal transitions or variations in building occupancy. Taken together, these properties give the FiLM Transformer VAE a balanced synergy between structure and stochasticity, offering a scalable solution for practical electricity-load forecasting across multiple temporal and spatial resolutions.è

Chapter 5

Conclusion

This thesis introduced the Refined Gated WaveNet, a lightweight and scalable model for unsupervised anomaly detection in univariate electricity load time series. Starting from WaveNet, which was originally developed for discrete μ-law audio generation, we recast the architecture as a continuous, window-based regression predictor so that anomaly scores can be derived naturally from prediction errors in energy monitoring settings. The proposed design preserves the key WaveNet components that support multi-scale temporal modeling, namely dilated causal convolutions and gated activation units, but adapts them to practical monitoring constraints through parallel window processing and a temporal refinement head that improves sensitivity to gradual and long-range deviations. Under a controlled anomaly injection protocol, the refined model achieved consistent gains over standard baselines, not only in detection performance but also in the robustness of downstream forecasts after anomaly removal. Overall, these results suggest that anomaly handling should be treated as an integral part of reliable monitoring pipelines rather than an optional preprocessing step.

Building on this monitoring perspective, the thesis then proposed a probabilistic forecasting approach based on a FiLM Conditioned Transformer VAE. The model combines attention-based sequence modeling with a variational latent component to produce predictive distributions, while FiLM modulation injects regime-related temporal information directly into hidden representations. Experiments on both a building-level dataset (Office Load) and a regional-scale dataset (PJM) showed that this framework improves forecasting accuracy while providing uncertainty estimates

that remain well calibrated. Reliability analyses indicated that empirical coverage stays close to nominal levels, with mean calibration errors below 7%, which means that prediction intervals are not only informative but also consistent across heterogeneous operating regimes. In operational terms, forecasts that remain both sharp and credible can support reserve planning, load balancing, and risk-aware decision making. A practical advantage is that uncertainty is inferred directly from historical load observations, without relying on external weather inputs or auxiliary forecasting modules, which is particularly useful when exogenous variables are missing, delayed, or unreliable. FiLM conditioning further contributes to stability by aligning the internal representations with recurring daily and weekly patterns and by reducing sensitivity to slow regime changes.

This work also has limitations. The anomaly detection study focused on univariate series and relied on controlled synthetic anomaly injection, which enables reproducible benchmarking but cannot fully capture the diversity and complexity of real faults in operational energy systems. Likewise, although probabilistic forecasting results were consistent across two datasets at different scales, additional validation across more buildings, climates, and grid conditions would strengthen the external validity of the conclusions.

Several directions appear promising for future research. Extending both models to multivariate and spatially correlated settings would allow the use of cross-signal structure that is not available in univariate data. Stronger coupling between anomaly detection and probabilistic forecasting, for instance via shared encoders or joint objectives, could further improve robustness under corrupted observations. Another direction is to investigate richer generative Transformer variants that separate uncertainty sources more explicitly. Finally, the proposed monitoring framework is not limited to electricity demand and could be transferred to other smart building time series such as occupancy or behavioral signals, and eventually integrated into control-oriented systems, including reinforcement learning approaches for closed-loop energy management.

References

Ali, A., Xia, Y., Zia, M. F., Bangyal, W. H., & Iqbal, M. (2025). Trustworthy load forecasting with generative ai: A dual-attention convlstm and vae-based approach. *IEEE Transactions on Consumer Electronics*.

Barron, J. T. (2019). A general and adaptive robust loss function. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (pp. 4331–4339).

Berahmand, K. e. a. (2024). Autoencoders and their applications in machine learning: a survey. *Artificial Intelligence Review*, *57*(2), 28.

Bond-Taylor, S., Leach, A., Long, Y., & Willcocks, C. G. (2022). Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(11), 7327–7347. doi: 10.1109/TPAMI.2021.3116668

Borbáth, T., & Van Hertem, D. (2024). Appropriate transmission grid representation for european resource adequacy assessments. *Applied Energy*, *355*, 122378.

Bouktif, S., Fiaz, A., Ouni, A., & Serhani, M. A. (2020). Multi-sequence lstm-rnn deep learning and metaheuristics for electric load forecasting. *Energies*, *13*(2), 391.

Buechler, E., Powell, S., Sun, T., Astier, N., Zanocco, C., Bolorinos, J., ... Rajagopal, R. (2022). Global changes in electricity consumption during COVID-19. *iScience*, *25*(1).

Challu, C. I., Jiang, P., Wu, Y. N., & Callot, L. (2022). Deep generative model with hierarchical latent factors for time series anomaly detection. In *Proceedings of the international conference on artificial intelligence and statistics (aistats)* (pp. 1643–1654).

Chan, J. W., & Yeo, C. K. (2024). A transformer based approach to electricity load

forecasting. *The Electricity Journal*, *37*(2), 107370. Retrieved from https://www.sciencedirect.com/science/article/pii/S1040619024000058 doi: 10.1016/j.tej.2024.107370

Chen, H., Ran, L., Sun, X., & Cai, C. (2023). Sw-wavenet: Learning representation from spectrogram and wavegram using wavenet for anomalous sound detection. In *Proceedings of icassp 2023 - ieee international conference on acoustics, speech and signal processing* (pp. 1–5).

Chien, H.-Y. S., Turek, J. S., Beckage, N., Vo, V. A., Honey, C. J., & Willke, T. L. (2021). Slower is better: revisiting the forgetting mechanism in lstm for slower information decay. *arXiv preprint arXiv:2105.05944*.

Dab, K., Nagarsheth, S. H., Amara, F., Henao, N., Agbossou, K., Dubé, Y., & Sansregret, S. (2024). Uncertainty quantification in load forecasting for smart grids using non-parametric statistics. *IEEE Access*, *12*, 138000–138017. Retrieved from https://doi.org/10.1109/ACCESS.2024.3465229 doi: 10.1109/ACCESS.2024.3465229

Diamantoulakis, P. D., Kapinas, V. M., & Karagiannidis, G. K. (2015). Big data analytics for dynamic energy management in smart grids. *Big Data Research*, *2*(3), 94–101.

Ferreira, A. B. A., Leite, J. B., & Salvadeo, D. H. P. (2025). Power substation load forecasting using interpretable transformer-based temporal fusion neural networks. *Electric Power Systems Research*, *238*, 111169.

Hasanat, S. M., Ullah, K., Yousaf, H., Munir, K., Abid, S., Bokhari, S. A. S., ... Ullah, Z. (2024). Enhancing short-term load forecasting with a CNN-GRU hybrid model: A comparative analysis. *IEEE Access*, *12*, 184132–184141. doi: 10.1109/ACCESS.2024.3511653

Hayashi, T., Komatsu, T., Kondo, R., Toda, T., & Takeda, K. (2018). Anomalous sound event detection based on wavenet. In *Proceedings of the 26th european signal processing conference (eusipco)* (pp. 2494–2498).

Huang, H. (2024). *Short-term wind speed forecasting model based on an attention-gated recurrent neural network and error correction strategy.* (arXiv:2404.11422)

Jia, M., Komeily, A., Wang, Y., & Srinivasan, R. S. (2019). Adopting internet of things for the development of smart buildings: A review of enabling technologies and applications. *Automation in Construction*, *101*, 111–126.

Kim, J., Kim, H., Kim, H., Lee, D., & Yoon, S. (2025). A comprehensive survey of deep learning for time series forecasting: Architectural diversity and open challenges. *Artificial Intelligence Review*, *58*(7), 1–95.

Komatsu, T., Hayashi, T., Kondo, R., Toda, T., & Takeda, K. (2019). Scene-dependent anomalous acoustic-event detection based on conditional wavenet and i-vector. In *Proceedings of icassp 2019 - ieee international conference on acoustics, speech and signal processing* (pp. 870–874).

Laitsos, V., Vontzos, G., Paraschoudis, P., Tsampasis, E., Bargiotas, D., & Tsoukalas, L. H. (2024). The state of the art electricity load and price forecasting for the modern wholesale electricity market. *Energies*, *17*(22), 5797.

Leushuis, R. M. (2025). Probabilistic forecasting with VAR-VAE: Advancing time series forecasting under uncertainty. *Information Sciences*, *713*, 122184.

Li, F., Guo, S., Han, F., Zhao, J., & Shen, F. (2024). *Multi-scale dilated convolution network for long-term time series forecasting.* (arXiv:2405.05499)

Li, X., & Jia, R. (2024). Energy-aware scheduling algorithm optimization for AI workloads in data centers based on renewable energy supply prediction. *Journal of Computing Innovations and Applications*, *2*(2), 56–65.

Liu, J. (2025). Global temporal attention-driven transformer model for video anomaly detection. In *2025 5th international conference on artificial intelligence and industrial technology applications (aiita)* (pp. 1909–1913).

Maragkos, N., & Refanidis, I. (2025). A comparative evaluation of time-series forecasting models for energy datasets. *Computers*, *14*(7), 246. doi: 10.3390/computers14070246

Masood, Z., Gantassi, R., & Choi, Y. (2024). Enhancing short-term electric load forecasting for households using quantile LSTM and clustering-based probabilistic approach. *IEEE Access*, *12*, 77257–77268.

Mentzelopoulos, A. P., Fan, D., Sapsis, T. P., & Triantafyllou, M. S. (2024). Variational autoencoders and transformers for multivariate time-series generative modeling and forecasting: Applications to vortex-induced vibrations. *Ocean Engineering*, *310*, 118639.

Neloy, A. A., & Turgeon, M. (2024). A comprehensive study of auto-encoders for anomaly detection: Efficiency and trade-offs. *Machine Learning with Applications*, 100572.

Pang, G., Shen, C., Cao, L., & Van Den Hengel, A. (2021). Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)*, *54*(2), 1–38.

Rafi, S. H., Nahid-Al-Masood, Deeba, S. R., & Hossain, E. (2021). A short-term load forecasting method using integrated CNN and LSTM network. *IEEE Access*, *9*, 32436–32448. doi: 10.1109/ACCESS.2021.3060654

Rotib, H. W., Nappu, M. B., Tahir, Z., Arief, A., Shiddiq, M. Y. A., et al. (2021). Electric load forecasting for IoT smart home using hybrid PCA and ARIMA algorithm. *International Journal of Electrical and Electronic Engineering & Telecommunications*, *10*(6), 369–376.

Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., … Müller, K.-R. (2021). A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, *109*(5), 756–795. doi: 10.1109/JPROC.2021.3052449

Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, *36*(4), 1181–1191.

Selvarajan, G. (2021). Leveraging ai-enhanced analytics for industry-specific optimization: A strategic approach to transforming data-driven decision-making. *International Journal of Enhanced Research in Science Technology & Engineering*, *10*, 78–84.

Shahzad, S., & Jasińska, E. (2024). Renewable revolution: A review of strategic flexibility in future power systems. *Sustainability*, *16*(13).

Shaikh, A. K., Nazir, A., Khalique, N., Shah, A. S., & Adhikari, N. (2023). A new approach to seasonal energy consumption forecasting using temporal convolutional networks. *Results in Engineering*, *19*, 101296. doi: 10.1016/j.rineng.2023.101296

Shrivastava, A., Rameshan, R., & Agnihotri, S. (2024). Latent space characterization of autoencoder variants. *arXiv*.

Toner, W., & Darlow, L. (2024). An analysis of linear time series forecasting models. *arXiv preprint arXiv:2403.14587*.

Tufail, S., Riggs, H., Tariq, M., & Sarwat, A. I. (2023). Advancements and challenges in machine learning: A comprehensive review of models, libraries, applications, and algorithms. *Electronics*, *12*(8), 1789.

Turowski, M. e. a. (2022). Modeling and generating synthetic anomalies for energy and power time series. In *Proceedings of the thirteenth acm international conference on future energy systems (e-energy)* (pp. 471–484).

Ullah, K., Ahsan, M., Hasanat, S. M., Haris, M., Yousaf, H., Raza, S. F., ... Ullah, Z. (2024). Short-term load forecasting: A comprehensive review and simulation study with CNN-LSTM hybrids approach. *IEEE Access*.

Uzair, M., Shah, I., & Ali, S. (2024). An adaptive strategy for wind speed forecasting under functional data horizon: A way toward enhancing clean energy. *IEEE Access*, *12*, 68730–68746.

Van Den Oord, A. e. a. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *arXiv*.

Verma, K. K. e. a. (2022). A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system. *International Journal of Information Technology*, *14*(1), 397–410.

Wang, Y., Shi, R., Zhang, C., He, Y., Jiang, H., & Kubota, J. (2022). Structural changes and trends in China's renewable electricity production in the policy evolution process. *Renewable Energy*, *182*, 879–886.

Waqas, M., & Humphries, U. W. (2024). A critical review of rnn and lstm variants in hydrological time series predictions. *MethodsX*, 102946.

Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, *34*, 22419–22430.

Xie, T., Xu, Q., Jiang, C., Gao, Z., & Wang, X. (2024). A robust anomaly detection model for pumps based on the spectral residual with self-attention variational autoencoder. *IEEE Transactions*

on Industrial Informatics, *20*(6), 9059–9069.

Xu, H., Boyaci, A., Lian, J., & Wilson, A. (2025). Explainable AI for multivariate time series pattern exploration: Latent space visual analytics with temporal fusion transformer and variational autoencoders in power grid event diagnosis. *IEEE Access*, 1–1. (Early Access) doi: 10.1109/ACCESS.2025.3602635

Xu, J., Zheng, T., Dang, Y., Yang, F., & Li, D. (2025). Distributed deep reinforcement learning for data-driven water heater model in smart grid. *IEEE Transactions on Smart Grid*.

Yan, P., Abdulkadir, A., Luley, P.-P., Rosenthal, M., Schatte, G. A., Grewe, B. F., & Stadelmann, T. (2024). A comprehensive survey of deep transfer learning for anomaly detection in industrial time series: Methods, applications, and directions. *IEEE Access*, *12*, 3768–3789. doi: 10.1109/ACCESS.2023.3349132

Yang, D.-H., & Chang, J.-H. (2022). FiLM conditioning with enhanced feature to the transformer-based end-to-end noisy speech recognition. In *Interspeech* (pp. 4098–4102).

Yang, Y., Zhang, C., Zhou, T., Wen, Q., & Sun, L. (2023). Dcdetector: Dual attention contrastive representation learning for time series anomaly detection. In *Proceedings of the 29th acm sigkdd conference on knowledge discovery and data mining* (pp. 3033–3045).

Zhao, Y., Takaki, S., Luong, H.-T., Yamagishi, J., Saito, D., & Minematsu, N. (2018). Wasserstein gan and waveform loss-based acoustic model training for multi-speaker text-to-speech synthesis systems using a wavenet vocoder. *IEEE Access*, *6*, 60478–60488.

Zhou, C., Li, Z., Song, J., & Xiang, W. (2024). TransVAE-DTA: Transformer and variational autoencoder network for drug-target binding affinity prediction. *Computer Methods and Programs in Biomedicine*, *244*, 108003. doi: 10.1016/j.cmpb.2023.108003

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient Transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, pp. 11106–11115).

Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., & Jin, R. (2022). FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th international conference on machine learning* (Vol. 162, pp. 27268–27286). PMLR.

Zhuang, F., Chen, X., Horata, P., & Sunat, K. (2025). Research on hybrid architecture neural

networks for time series prediction. *IEEE Access*.