IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 March 2026, accepted 20 April 2026, date of publication 23 April 2026, date of current version 30 April 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3686958

RESEARCH ARTICLE

Patch-Based Transformer With Mixture of Experts and Conformal Inference for Curtailment Time Series Forecasting

MIKHAIL ZIMMER HEIDRICH[1], **MARI AURORA FAVERO REIS**[1], **CRISTINA KEIKO YAMAGUCHI**[1], **LAIO ORIEL SEMAN**[2], **AND STEFANO FRIZZO STEFENON**[3]

[1]Postgraduate Program in Productive Systems in Association with UNIPLAC, UNC, UNESC, and UNIVILLE, Lages 88509-900, Brazil
[2]Department of Automation and Systems Engineering, Federal University of Santa Catarina, Florianópolis, Santa Catarina 88040-900, Brazil
[3]Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, 1959-007 Lisbon, Portugal

Corresponding author: Laio Oriel Seman (laio.seman@ufsc.br)

This work was supported in part by the Coordination for the Improvement of Higher Education Personnel (CAPES) Transformative Agreement; in part by the National Council for Scientific and Technological Development (CNPq), Brazil; and in part by the Foundation for Research and Innovation of the State of Santa Catarina (FAPESC).

ABSTRACT The growth of variable renewable energy sources has increased the occurrence of curtailment in power systems, particularly in regions affected by transmission constraints and limited operational flexibility. Reliable forecasting of curtailment time series, together with well-calibrated uncertainty estimates, is therefore essential to support operational planning, market decisions, and risk management. This paper proposes a curtailment forecasting framework that combines a high-capacity neural time series model with distribution-free conformal prediction to jointly address accuracy and uncertainty quantification. The forecasting backbone is a Transformer-based encoder-decoder architecture enhanced with reversible instance normalization to mitigate distribution shift, patch-based tokenization to reduce sequence length and improve temporal representation, and a Mixture of Experts feed-forward module to enable conditional computation and improved generalization. The proposed method is evaluated against a wide range of state-of-the-art models, including NBEATS, NBEATSx, NHITS, PatchTST, LSTM, GRU, TFT, Informer, and Autoformer, across multiple forecasting horizons. The model achieves an MSE of 0.0059 for a horizon equal to 3h, corresponding to a 13.2% reduction relative to the second-best methods. For a horizon equal to 5h, the MSE is reduced to 0.0119, representing a 29.6% improvement, while for a horizon equal to 10, the model attains an MSE of 0.0461, outperforming NHITS by 22.9%. Ablation analysis confirms the importance of patch embedding, the Mixture of Experts mechanism, and reversible instance normalization. The conformal prediction module achieves an empirical coverage of 91.27% for a nominal 90% level, with adaptive and practically useful interval widths.

INDEX TERMS Curtailment forecasting, time series prediction, transformer models, mixture of experts, conformal prediction.

I. INTRODUCTION

The Brazilian National Interconnected System (SIN) covers more than 99% of Brazil's electricity demand, integrating four main subsystems (Southeast/Midwest, South, Northeast, North) through a large network of transmission lines and substations. The system is predominantly hydroelectric, with most of the generation coming from hydroelectric plants [1], supplemented by wind, solar, and thermoelectric plants (mainly natural gas, coal, and oil) to ensure supply during periods of drought or peak demand [2].

Considering that the SIN is predominantly based on hydroelectric power and uses thermal generation when hydroelectric generation is not available, or if it plans to

The associate editor coordinating the review of this manuscript and approving it for publication was Katherine VanDenburgh.

use it in the future, the system is called hydrothermal [3]. Hydrothermal dispatch decisions are made using a chain of optimization models, each focused on different time horizons:

- **Long term (years):** The NEWAVE model uses dual random dynamic programming to optimize reservoir management, considering inflow uncertainties and minimizing expected operating costs [4];
- **Medium term (weeks)**: The DECOMP model performs weekly dispatch, balancing hydraulic and thermal generation and setting prices [5]. It takes into account the storable variability of dams, demand forecasts, and system constraints [6];
- **Short term (days/hours)**: The DESSEM model, validated since 2017, determines daily and hourly dispatch, incorporating detailed constraints from thermal plants, grid security, and real-time operating conditions [7].

These energy dispatch models (NEWAVE, DECOMP, and DESSEM) aim to minimize total system costs (fuel, deficit, losses), manage reservoir storage, and ensure reliability in different hydrological scenarios [8].

According to the Brazilian National Operator of Electric System (ONS) data from 2023 onwards, energy transmission bottlenecks, particularly in the Northeast region subsystem, have resulted in the temporary shutdown of wind and solar power plants during peak times, as the system is unable to absorb all production [9], being a problem that needs to be mitigated [10]. This results in wasted clean energy and lost revenue for investors and consumers, who could potentially pay less for electricity. This restriction or cut in energy generation, which means the waste of available and unused energy, has been given the name curtailment [11].

Limiting production by renewable power plants has become inherent in electrical systems with a high share of sources such as wind and solar, a reality observed in the SIN. Curtailment stems from the mismatch between generation variability and periods of peak demand, which can result in momentary energy surpluses. In these scenarios, the adoption of curtailment is not an arbitrary decision by the ONS, but rather an indispensable technical intervention to ensure the operational integrity and reliability of the SIN [12].

From a power system engineering perspective, curtailment is primarily driven by three fundamental factors: transmission network constraints, temporal demand variations, and weather-dependent renewable generation. Transmission congestion limits the ability of the grid to absorb available generation, particularly in regions with high renewable penetration, while demand fluctuations determine whether surplus generation must be curtailed to maintain system balance.

Based on this need, a unified framework for curtailment time series forecasting that combines a high-capacity Transformer-based neural architecture with distribution-free uncertainty quantification is proposed. The proposed approach introduces a patch-based encoder-decoder Transformer enhanced with reversible instance normalization to mitigate distribution shift and a Mixture of Experts (MoE) feed-forward module to enable conditional computation and improved generalization. The proposed model integrates a conformal prediction strategy adapted to time series, producing prediction intervals with finite sample coverage guarantees and adaptive widths.

The key innovations of the proposed method are:

- The first contribution is the introduction of a patch-based temporal representation for curtailment time series, which segments the input sequence into non-overlapping patches, reduces attention complexity, and significantly improves the extraction of local temporal patterns and forecasting accuracy.
- The second contribution is the integration of reversible instance normalization to explicitly address distribution shift between training and inference, enhancing robustness under non-stationary operating conditions while preserving the original data scale.
- The third contribution is the replacement of the standard Transformer feed-forward block with a MoE module, enabling conditional computation through learned routing, improving model capacity utilization, and enhancing generalization.
- The fourth contribution is the coupling of high-accuracy point forecasting with a distribution-free conformal prediction framework, providing valid and adaptive prediction intervals with finite sample coverage guarantees for curtailment time series.

The proposed method can directly support professionals involved in power system operation, planning, and market decision-making by providing accurate forecasts of curtailment together with reliable uncertainty estimates. By anticipating when and how much energy may be curtailed, operators can better make decisions to reduce energy waste.

The remainder of this paper is organized as follows: Section II reviews the main concepts related to curtailment in power systems and discusses existing technical and artificial intelligence-based solutions, with particular emphasis on conformal prediction methods for time series. Section III presents the proposed forecasting framework, detailing the problem formulation, model architecture, preprocessing steps, training procedure, and evaluation setup. Section IV presents the evaluation setup. Section V reports and discusses the experimental results, including comparative performance analyses, ablation studies, and uncertainty quantification using conformal prediction intervals. Finally, Section VI concludes the paper by summarizing the main findings, highlighting the practical implications for curtailment forecasting, and outlining directions for future research.

II. RELATED WORKS

Curtailment has been a prominent topic in the debate on integrating renewable sources into the electricity system, as installed wind and solar generation capacity grow at a rapid pace [13]. In Brazil, this scenario is most evident in regions with a high concentration of renewable energy use, where structural and operational limitations hinder the full flow of energy produced [14]. The review presented in this section shows the occurrence of curtailment in an integrated manner, covering its causes, consequences, and possible mitigation strategies.

A. OVERVIEW OF CURTAILMENT

A main cause of curtailment in Brazil today is transmission system constraints. A study conducted in Brotas de Macaúbas, in the state of Bahia, points out that the temporal complementarity between wind and solar generation in Northeast Brazil could significantly reduce these cuts, since the optimal combination of sources results in less overlap in generation and greater adherence to demand profiles. Campos et al. [15] demonstrate that a configuration of 40% wind power and 60% solar power increases operational efficiency and reduces the need for curtailment, reducing contracted transmission capacity by up to 11%. This shows that the main constraint for mitigating curtailment is not only the intermittency of the sources, but above all, the inability to distribute the energy generated.

The relationship between transmission limits and energy waste is similarly highlighted by a study conducted at the Federal University of Pernambuco, which shows how curtailment of wind power generation occurs when the grid cannot absorb production during periods of low demand. The researchers show that the lack of flexibility in the system forces the replacement of wind energy, usually cheaper, with other sources with higher costs, directly impacting the system's operational expenses. Furthermore, the inability to fully utilize the generated energy reinforces the need for mechanisms capable of displacing this surplus, such as storage systems [16].

Energy storage is an alternative for reducing curtailment. The use of Battery Energy Storage Systems (BESS) in critical busbars allows wasted energy to be stored and dispatched at times of peak load, helping to relieve the overload on transmission lines [17]. In addition, storage smooths the generation curve of hybrid wind-solar systems throughout the day, promoting greater operational stability and reducing the need for abrupt generation cuts. Despite this, the fact that none of the scenarios analyzed is capable of maintaining generation (in hybrid systems) close to nominal potential throughout the year indicates that storage, although important, does not eliminate curtailment and must be scaled considering its costs.

Hybridization (combining different generation sources) is still a relevant strategy for mitigating curtailment. The complementarity between wind and solar sources helps optimize the use of existing infrastructure. Veras et al. [18] demonstrate that hybrid plants are capable of increasing profits by up to 39% and improving the efficiency of transmission contracts, allowing the installation of greater generating capacities using the same physical transmission structure. Thus, from both an operational and economic point of view, hybridization is important because it reduces the probability of local overload while decreasing the risk of curtailment, even in situations where temporal complementarity (compensation made by different generating sources) is not high enough to technically justify this integration.

In addition to technical and economic strategies, there are alternatives for valuing the renewable surplus that would be lost during the curtailment process [19]. The potential of green hydrogen as a means of energy storage can be highlighted, especially in regions with high wind production, such as Ceará, where part of the energy generated is wasted due to distribution restrictions [20]. The use of this surplus for hydrogen production is a strategic alternative to overcome intermittency and reduce dependence on fossil fuels. Hydrogen should preferably be produced using only energy surpluses, since its conversion process involves losses and would not justify the use of energy intended for direct load supply. Hydrogen production appears as a complementary mechanism for mitigating curtailment and adding value to renewable energy [21].

Another technology that has been explored and considered for curtailment mitigation is floating photovoltaic plants [22]. In addition to environmental benefits, these systems could compensate for periods of low hydroelectric generation and minimize curtailment during periods of high demand, adding greater flexibility for the electric system operator [23]. In this context, the installation of distributed solar systems could replace (or reduce the need for) diesel generators, thereby reducing waste, emissions, and dependence on fossil fuels.

The international landscape shows that curtailment can still be a consequence of expansion policies that are poorly aligned with the growth of renewable energies [24]. It is important to mention the case of China, where the rapid expansion of coal capacity between 2010 and 2015 led to curtailment rates of 17% for wind and 20% for solar, despite investments in renewables. This occurred because inflexible thermal capacity reduced the operational space for variable sources, generating significant waste [25]. This example reinforces that the efficient integration of renewables depends on systemic planning and not just the addition of installed capacity.

Curtailment is a recent and complex phenomenon, influenced by transmission limitations, system operating characteristics, insufficient flexibility, expansion policies, and even the lack of mechanisms to take advantage of energy surpluses [26]. Strategies to mitigate it include hybridization, the use of storage technologies [27], such as BESS [28] and/or green hydrogen production [29], expansion of energy transmission systems [30], load management [31], and integrated planning of the operation and expansion of the electricity grid. Understanding curtailment from different perspectives allows us to identify efficient ways to increase the use of renewable energy and strengthen the security and sustainability of the electrical power system [32].

B. AI-DRIVEN SOLUTIONS

Several researchers have utilized artificial intelligence solutions to assist in the optimization of energy resources [33], power consumption [34], generation [35], or even predicting system variation [36]. To reduce wind curtailment and maximize the profitability of participation in the energy market and grid stability services, the study developed by Li et al. [37] presented a new supply strategy based on deep reinforcement learning for co-located wind-battery systems. The study concluded that the AggJointDRL strategy significantly improved the joint performance of the wind farm and BESS system in question, increasing economic viability by optimizing bids in spot markets (short-term electricity markets) and regulation frequency control markets, reducing curtailment, and outperforming reference methods in both financial return and execution efficiency.

Within this same perspective, to address the uncertainties and predictability errors and conventional failures of wind energy, the research by Li and Liu [38] proposes a method for determining the reserve capacity of power systems that include wind generation, deducing the quantitative relationship between upper reserve and load loss probability, and between lower reserve and wind power curtailment probability. This method allows for offline calculation of reserve capacity when wind power is fully absorbed, and at the same time can be used for online optimization of reserves, taking into account the impact of wind power curtailment.

In terms of electrical system planning, Schäfer et al. [39] developed a method that allows multiple future states of the grid to be predicted to assess losses or predict contingency situations when integrating renewable sources. This was achieved by comparing regression and classification methods to predict multivariate outcomes or binary classifications of time steps to identify critical loading situations. In all comparisons of the different machine learning (ML) models, the classic Multilayer Perceptron (MLP) showed the highest prediction performance.

Another example of a solution employed to improve responses to energy demands is proposed by Le et al. [40], who used power-to-gas and FlexGas as a demand response for the integrated electricity and gas system to accommodate wind energy. The proposed two-layer programming structure speeds up the calculation, with an aggregate model developed to deal with parameter heterogeneity. Experiment results showed that coordinated demand response can reduce wind curtailment and improve the efficiency of these systems.

Still dealing with hybrid generation systems, the research by Veras et al. [18] evaluated the benefits of hybridizing a power plant using a methodology based on artificial intelligence to optimize the wind-solar ratio based on the Brazilian regulatory system. To this end, a genetic algorithm was used to determine the ideal wind-solar ratio, considering transmission system costs. The algorithm achieved a monthly profit increase of over 39%, with a power cut of less than 1%, indicating economic complementarity.

In the field of microgrid optimization studies, we can mention two other studies: the first is the proposal developed by Liu et al. [41], who developed time-shifting models for participation in the energy dispatch of a Data Center Microgrid (DCM). In this solution, a dispatch optimization structure was formulated for the day following the current day, incorporating time-shifting models for multiple loads, to minimize the DCM's operating costs and renewable energy curtailment. The results confirmed the effectiveness of the proposed energy management scheme and the efficiency of time-shiftable loads in improving the operational economy of the data center microgrid and increasing the utilization rate of renewable energy.

Eberlein and Rudion [42] investigated three of the most popular droop variants (control methods for regulating voltage, frequency, and power sharing in microgrids), developing reference scenarios to evaluate and compare the applicability of the control variants, as well as their combination with virtual impedances, under practical conditions. A genetic algorithm adapted to the problem was used to optimize the control parameters for each type of controller. The authors observed that all droop variants led to high network stability when their parameters were rigorously optimized. In addition, the authors state that the observed superiority of one droop variant over the others, for some specific microgrid scenarios, could not be generalized to practical applications, since the detailed optimization of control parameters and the combination with virtual impedance are substantially more important than the choice of droop variant.

Solutions are also being considered for optimizing generation in rotating machines. The work of Karn et al. [43] proposed a technique for identifying sensitive loads and determining the amount of active and reactive power cutback required for rotor speed regulation and terminal voltage management. For this application, the MATLAB/Simulink environment was used to verify and test the feasibility of the methodology in a 10-machine, 39-bus IEEE system. Active power cutback was considered to ensure rotor angle stability, while reactive power was also reduced to maintain terminal voltage in the loads. Their method proved to be more economical in managing reactive power and voltage conditions.

Wang et al. [44] explored the potential of Integrated Demand Response (IDR) for residential users and analyzed the influence of user comfort and consumer psychology on IDR, proposing a study of optimized load aggregator operation. The analysis demonstrated that the proposed strategy is capable of offering a more reliable and targeted IDR to residential users, achieving a win-win situation for both the load aggregator and residential consumers.

Regarding integration with energy storage systems, research suggested by Lopez et al. [45] proposed an active power reduction control strategy for a stand-alone photovoltaic system feeding an electrolyzer. The controller developed has a high capacity to handle forecast errors and uncertainties, combining artificial intelligence to predict power ramps and fuzzy logic to handle imperfect forecasts. It was possible to smooth the photovoltaic power ramps to a limit of 10% of the photovoltaic plant's capacity at times of high variability, even with very inaccurate forecasts, without any type of storage and without interrupting the supply to the electrolyte.

C. CONFORMAL PREDICTION METHODS

Conformal prediction provides distribution-free uncertainty quantification with finite-sample coverage guarantees under data exchangeability [46]. The foundational inductive (split) conformal method divides data into training and calibration sets, using calibration residuals to form prediction intervals [47]. Several extensions address limitations such as inefficiency in heteroscedastic settings or lack of adaptation under distribution drift.

1) ADAPTIVE AND LOCALIZED METHODS

Localized conformal prediction improves interval efficiency by weighting or selecting calibration points near the test instance, often via k-nearest neighbors [48]. The jackknife+ method constructs predictive intervals robust to model variability, using leave-one-out residuals, and is particularly effective post-model selection [49]. Conformalized Quantile Regression combines quantile regression with conformal prediction to produce adaptive intervals that account for varying uncertainty levels [50].

2) METHODS FOR TIME SERIES AND DISTRIBUTION DRIFT

For sequential data, Ensemble Batch Prediction Intervals leverages bootstrap ensemble out-of-bag residuals to form valid intervals without data splitting, scaling well to online forecasting [51]. Adaptive Conformal Inference (ACI) monitors empirical coverage online and adjusts the quantile level to maintain validity under arbitrary distribution shifts [52]. Building on ACI, Aggregated Adaptive Conformal Inference (AgACI) aggregates multiple ACI experts with different adaptation rates for enhanced robustness in time series [53].

3) RECENT STATE-AWARE AND ATTENTION-BASED APPROACHES

More recent works incorporate explicit structure for non-stationary time series. Conformal Prediction for Time-series with Change Points uses a separate state or regime model to weight calibration residuals based on similarity in predicted states, enabling better adaptation to regime shifts [54].

Attention-Based Feature Online Conformal Prediction employs a learned attention mechanism over feature embeddings to dynamically weight historical residuals, supporting online fine-tuning for highly adaptive intervals in complex sequential data [55].

III. PROPOSED MODEL

This section presents the proposed architecture for time series forecasting. The model combines a Transformer encoder-decoder structure with instance normalization, patch-based tokenization, and a MoE feed-forward network. Fig. 1 illustrates the overall architecture.

A. PROBLEM FORMULATION

Given a multivariate time series $\mathbf{X} = \{x_1, x_2, \ldots, x_L\} \in \mathbb{R}^{L\times d}$, where L denotes the lookback window and d the number of features, the objective is to predict future values $\hat{\mathbf{Y}} \in \mathbb{R}^{T\times d_{\text{out}}}$ over a forecast horizon T. The model learns a mapping $f_\theta : \mathbb{R}^{L\times d} \to \mathbb{R}^{T\times d_{\text{out}}}$ parameterized by θ.

B. INPUT PREPROCESSING

The input sequence undergoes two preprocessing stages before entering the encoder. First, Reversible Instance Normalization (RevIN) [56] addresses distribution shift by normalizing each instance along the temporal dimension:

$$\tilde{\mathbf{X}} = \frac{\mathbf{X} - \mu}{\sigma + \epsilon} \tag{1}$$

where $\mu, \sigma \in \mathbb{R}^d$ are the per-feature mean and standard deviation. These statistics are stored and applied inversely to the decoder output to recover the original scale.

The normalized sequence is then segmented into $N = \lfloor L/P \rfloor$ non-overlapping patches of length P, following [57]. Each patch $\mathbf{x}_i^{(p)} \in \mathbb{R}^{P\times d}$ is projected to the model dimension through a learnable linear transformation:

$$\mathbf{z}_i = \mathbf{W}_p \operatorname{vec}(\mathbf{x}_i^{(p)}) + \mathbf{b}_p \tag{2}$$

where $\mathbf{W}_p \in \mathbb{R}^{d_{\text{model}}\times Pd}$. The resulting embeddings $\mathbf{Z} \in \mathbb{R}^{N\times d_{\text{model}}}$ are augmented with sinusoidal positional encodings before entering the encoder.

C. ENCODER WITH MIXTURE OF EXPERTS

The encoder consists of N_e identical layers, each following the Pre-LayerNorm configuration [58]. We adopt RMSNorm [59] instead of standard LayerNorm, which omits mean centering and normalizes using only the root mean square:

$$\text{RMSNorm}(\mathbf{h}) = \frac{\mathbf{h}}{\sqrt{\frac{1}{d}\sum_{i=1}^{d} h_i^2 + \epsilon}} \odot \gamma \tag{3}$$

where $\gamma \in \mathbb{R}^d$ is a learnable gain parameter.

Each encoder layer first applies multi-head self-attention. For input $\mathbf{H} \in \mathbb{R}^{N\times d_{\text{model}}}$, queries, keys, and values are computed through linear projections, and scaled dot-product attention is applied:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{4}$$

The outputs of h parallel attention heads are concatenated and projected back to d_{model} dimensions.

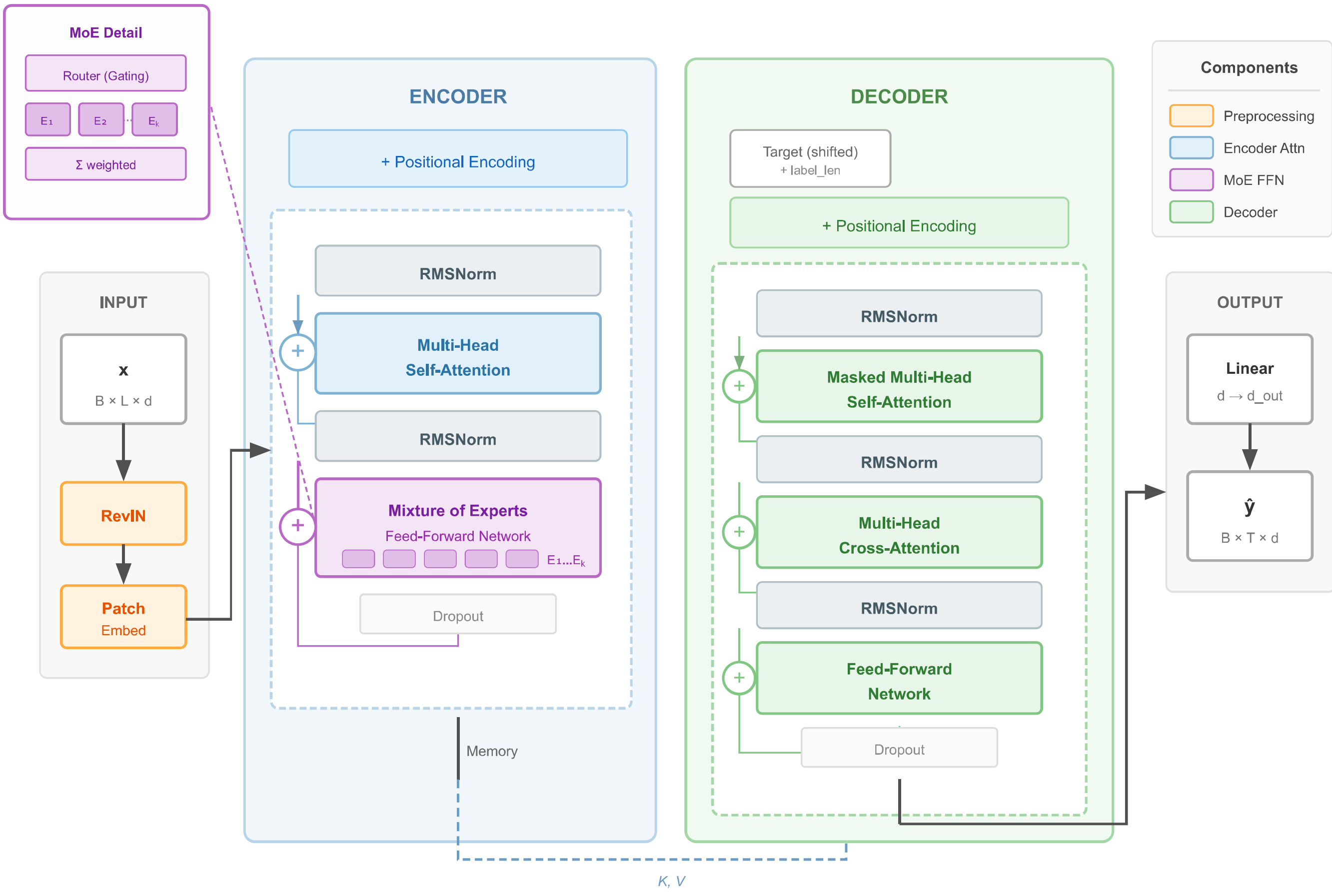


FIGURE 1. Proposed architecture. The encoder employs MoE feed-forward layers, while both encoder and decoder use Pre-LayerNorm with RMSNorm. Input preprocessing includes RevIN for distribution shift and patch embedding for sequence compression.

The feed-forward sublayer is replaced with a MoE module containing K expert networks. Each expert E_k is a two-layer feed-forward network with SwiGLU activation:

$$E_k(\mathbf{h}) = \mathbf{W}_2^{(k)}\left(\mathrm{SiLU}\left(\mathbf{W}_{1a}^{(k)}\mathbf{h}\right) \odot \left(\mathbf{W}_{1b}^{(k)}\mathbf{h}\right)\right) \tag{5}$$

A gating network computes routing weights $\mathbf{g}(\mathbf{h}) = \mathrm{softmax}(\mathbf{W}_g\mathbf{h}) \in \mathbb{R}^K$, and the MoE output is the weighted combination:

$$\mathrm{MoE}(\mathbf{h}) = \sum_{k=1}^{K} g_k(\mathbf{h}) \cdot E_k(\mathbf{h}) \tag{6}$$

To encourage balanced expert utilization, an auxiliary loss penalizes uneven routing:

$$\mathcal{L}_{\mathrm{aux}} = \alpha \cdot K \cdot \sum_{k=1}^{K} f_k \cdot \bar{g}_k \tag{7}$$

where f_k is the fraction of tokens assigned to expert k and $\bar{g}_k$ is the mean gating weight [60].

The complete encoder layer applies these components with residual connections:

$$\mathbf{H}' = \mathbf{H} + \mathrm{MHA}(\mathrm{RMSNorm}(\mathbf{H})) \tag{8}$$
$$\mathbf{H}'' = \mathbf{H}' + \mathrm{MoE}(\mathrm{RMSNorm}(\mathbf{H}')) \tag{9}$$

D. DECODER

The decoder follows an Informer-style [61] design with N_d layers. Its input consists of the last L_{label} timesteps from the encoder input (providing recent context), concatenated with placeholder tokens for the forecast horizon.

Each decoder layer contains three sublayers: masked self-attention, cross-attention, and a standard feed-forward network (FFN). Unlike the encoder, the decoder retains a conventional two-layer FFN with SwiGLU activation in place of the MoE module, as the decoder operates over a shorter sequence and does not benefit from sparse expert routing at inference time.

The masked self-attention prevents positions from attending to future timesteps by adding $-\infty$ to the corresponding attention scores before softmax. Let $\mathbf{M} \in \mathbb{R}^{N \times d_{\mathrm{model}}}$ denote the encoder memory, i.e., the final hidden states produced by the encoder after N_e layers. The cross-attention layer enables the decoder to attend to $\mathbf{M}$:

$$\mathrm{CrossAttn}(\mathbf{H}_d, \mathbf{M}) = \mathrm{Attention}(\mathbf{H}_d\mathbf{W}^Q, \mathbf{M}\mathbf{W}^K, \mathbf{M}\mathbf{W}^V) \tag{10}$$

The decoder layer transformation follows the same Pre-LayerNorm pattern:

$$\mathbf{H}_1 = \mathbf{H}_d + \mathrm{MaskedAttn}(\mathrm{RMSNorm}(\mathbf{H}_d)) \tag{11}$$

$$\mathbf{H}_2 = \mathbf{H}_1 + \text{CrossAttn}(\text{RMSNorm}(\mathbf{H}_1), \mathbf{M}) \tag{12}$$

$$\mathbf{H}_3 = \mathbf{H}_2 + \text{FFN}(\text{RMSNorm}(\mathbf{H}_2)) \tag{13}$$

The final decoder output is projected to the target dimension, and the inverse RevIN transformation restores the original data scale:

$$\hat{\mathbf{Y}} = (\mathbf{H}_{\text{out}}\mathbf{W}_{\text{proj}} + \mathbf{b}_{\text{proj}}) \cdot \sigma + \mu \tag{14}$$

E. TRAINING PROCEDURE

The model minimizes mean squared error (MSE) combined with the MoE auxiliary loss:

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T} \|\hat{y}_t - y_t\|^2 + \mathcal{L}_{\text{aux}} \tag{15}$$

During training, scheduled sampling bridges the gap between teacher forcing and autoregressive inference. The probability of using ground truth tokens decreases linearly: $p_{\text{tf}}(e) = \max(0, p_0 - e/E)$, where e is the current epoch and E is the total number of epochs.

Patch embedding reduces the effective sequence length from L to $N = L/P$, yielding a quadratic reduction in attention complexity from $\mathcal{O}(L^2)$ to $\mathcal{O}(N^2)$. The MoE layer maintains computational efficiency through sparse routing, activating only a subset of experts per token [62].

IV. EVALUATION SETUP

In this section, the considered dataset, model setup, and evaluation measures are explained.

A. DATASET

The dataset corresponds to an hourly time series for wind power generation in `LAGOA DOS VENTOS`, see Figure 2. This selection comprises information associated with the assessment of operational restrictions classified as constrained off in wind power plants operating under the regulatory modalities Type I, Type II-B, and Type II-C.

The time series data has 7,680 recorded values. The first record is January 01, 2025, at 00:00, and the last record was on December 31, 2025, at 23:00. For future comparisons, the full dataset is available at: https://dados.ons.org.br/dataset/restricao_coff_eolica_usi (accessed on January 14, 2026).

Descriptive statistics of the dataset are summarized in Table 1. The series exhibits a relatively high mean value with substantial variability, as indicated by the standard deviation. Additionally, the presence of zero values, although limited, reflects periods without curtailment events. The presented mean and standard deviation values were employed to perform the model normalization.

Before normalization, the signal was smoothed using a Savitzky–Golay filter to reduce high-frequency noise while preserving the local shape, peaks, and temporal structure of the series [63]. This filter performs a local polynomial regression within a sliding window, which is well-suited for time-series analysis where trend and short-term dynamics must be retained without excessive distortion.

The data represent the recorded conditions under which energy curtailment is imposed due to operational, electrical, or system-level constraints, reflecting the effective limitations on power injection into the grid. By focusing on this specific wind complex, the dataset ensures consistency in the analysis of constrained off events, enabling a detailed evaluation of curtailment patterns, their temporal distribution at the hourly resolution, and their relation to the regulatory framework governing these modalities.

TABLE 1. Descriptive statistics of the constrained-off time series for `LAGOA DOS VENTOS`.

Statistic	Value
Mean	698.65
Standard Deviation	464.80
Minimum	0.00
Maximum	1356.86
Percentage of Zeros	0.77%

TABLE 2. Main hyperparameters and architectural configuration of the proposed model.

Parameter	Value
Input/Output Configuration	
Input size	1
Output size	1
Sequence length	50
Prediction horizon	TARGET_LEN
Transformer Architecture	
Hidden size	32
Number of attention heads	4
Encoder layers	1
Decoder layers	1
Feedforward dimension	256
Dropout rate	0.1
Normalization	RMSNorm (Pre Norm)
Attention type	Multi-Head Attention
Decoder style	Informer-like
Label length	TARGET_LEN // 2
Mixture of Experts (MoE)	
Number of experts (E)	8
Top-k experts (K)	2
Router type	Noisy top-k
Capacity factor	1.25
Activation function	SwiGLU
Input Processing Heads	
RevIN head	Enabled (gated)
Patch embedding head	Enabled (gated)
Patch size	5
Training Configuration	
Total epochs	50
Scheduled sampling	Linear decay (0.95 $\rightarrow$ 0.0)
Forecasting strategy	Seq2Seq

B. MODEL SETUP

The model proposed in this paper was implemented using the Python language. Table 2 summarizes the main hyperparameters and architectural choices of the proposed model, which were tuned by mean of Bayesian optimization.

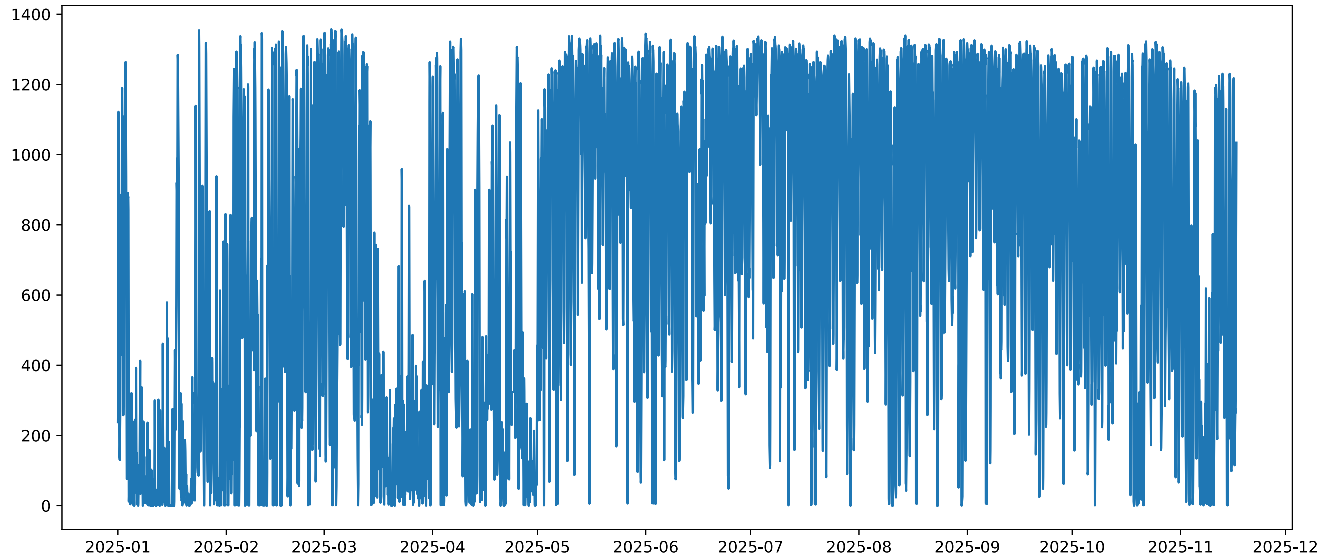


FIGURE 2. **Raw time series of constrained-off operations in** `LAGOA DOS VENTOS`**.**

C. MEASURES AND MODELS USED FOR COMPARISON

In this paper, the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). were considered for model evaluation. RMSE measures the square root of the average of the squared differences between observed and predicted values, emphasizing large errors due to the quadratic penalty, given by:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{16}$$

where y_i is the observed true value at sample index $i\hat{y}_i$ is the predicted value at sample index i and n is the total number of samples used for evaluation.

The MAE represents the arithmetic mean of the absolute deviations between observations and model forecasts, preserving the original unit of the series, given by:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|. \tag{17}$$

All computations and experiments were performed in `Google Colab`, utilizing a NVIDIA T4 GPU to accelerate the training and evaluation processes.

For benchmarking, we compare the Neural Basis Expansion Analysis for Time Series (NBEATS), NBEATSx, Neural Hierarchical Interpolation for Time Series (NHiTS) [64], Transformer for Time Series with Patching (PatchTST) [65], Gated Recurrent Unit (GRU) [66], Deep Autoregressive Forecasting (DeepAR), Long Short-Term Memory (LSTM) [67], TimeMixer [68], Temporal Fusion Transformer (TFT) [69], Temporal 2D Variation Modeling Network for Time Series (TimesNet), Informer Transformer for Long Sequence Forecasting (Informer), TimeXer, Bidirectional Temporal Convolutional Network (BiTCN), iTransformer, Vanilla-Transformer, Temporal Convolutional Network (TCN), Time series Dense Encoder (TiDE) [70], Autoformer, and Spectral Temporal Graph Neural Network (StemGNN) [36]. These models were based on `NeuralForecast` library and are available at `Nixtla` repository [71]. All models considered in the benchmarking used an equivalent lookback window size.

V. RESULTS AND DISCUSSIONS

Table 3 reports the forecasting performance for horizon $h = 3$, comparing a wide range of state-of-the-art models under the MSE and MAE metrics, with models ranked according to MSE. The proposed method (ours) achieves the best overall performance, obtaining the lowest MSE of 0.0059 and the lowest MAE of 0.0516, indicating superior accuracy and robustness in short-horizon forecasting.

Among the baseline models, NBEATSx and NBEATS present the closest performance to the proposed approach, both reaching an MSE of 0.0068 and an MAE of 0.0519. Although these results are competitive, they still exhibit a noticeable degradation when compared to the proposed method, particularly in terms of error variance captured by the MSE. NHITS and PatchTST follow next, with slightly higher errors, suggesting that while modern deep architectures can effectively model short-term temporal patterns, they are less capable of fully capturing the dynamics addressed by the proposed approach.

Recurrent models such as GRU and LSTM, as well as probabilistic methods like DeepAR, show a clear performance gap, with MSE values exceeding 0.010 and substantially higher MAE. This highlights the limitations of recurrent-based architectures in handling complex temporal dependencies at this forecasting horizon. Transformer-based and convolutional models, including TFT, TimesNet,

TABLE 3. Forecasting performance comparison for horizon $h = 3$. Models are ranked by MSE.

Model	MSE	MAE
Ours	**0.0059**	**0.0516**
NBEATSx	0.0068	0.0519
NBEATS	0.0068	0.0519
NHITS	0.0071	0.0529
PatchTST	0.0076	0.0549
GRU	0.0106	0.0768
DeepAR	0.0119	0.0776
LSTM	0.0126	0.0830
TimeMixer	0.0146	0.0838
TFT	0.0147	0.0809
TimesNet	0.0257	0.1182
Informer	0.0355	0.1387
TimeXer	0.0382	0.1549
BiTCN	0.0428	0.1525
iTransformer	0.0499	0.1591
VanillaTransformer	0.0719	0.2007
TCN	0.0772	0.1742
TiDE	0.0838	0.2197
Autoformer	0.1704	0.2876
StemGNN	0.3896	0.5045

TABLE 4. Forecasting performance comparison for horizon $h = 5$. Models are ranked by MSE.

Model	MSE	MAE
Ours	**0.0119**	**0.0721**
NBEATSx	0.0169	0.0805
NBEATS	0.0169	0.0805
NHITS	0.0169	0.0826
PatchTST	0.0209	0.0901
DeepAR	0.0275	0.1136
TimeMixer	0.0303	0.1198
TFT	0.0325	0.1147
GRU	0.0370	0.1448
TimesNet	0.0419	0.1408
TCN	0.0460	0.1672
TimeXer	0.0537	0.1704
LSTM	0.0572	0.1833
BiTCN	0.0600	0.1820
iTransformer	0.0649	0.1838
Informer	0.0685	0.1899
VanillaTransformer	0.1000	0.2375
TiDE	0.1088	0.2497
Autoformer	0.2116	0.3448
StemGNN	0.3985	0.5115

Informer, and TCN variants, perform considerably worse, with errors increasing sharply, indicating difficulties in generalization or excessive model complexity for this task.

Graph-based and decomposition-oriented models such as Autoformer and StemGNN exhibit the weakest performance, with StemGNN presenting a particularly high MSE of 0.3896 and MAE of 0.5045. Overall, the results demonstrate that the proposed method consistently outperforms all competing approaches for horizon $h = 3$, achieving both lower bias and variance, and confirming its effectiveness for accurate short-term forecasting.

Table 4 presents the forecasting results for horizon $h = 5$, with models ranked according to the MSE metric. As the prediction horizon increases, a general degradation in performance is observed across all methods. Even under this more challenging setting, the proposed approach remains clearly superior, achieving the lowest MSE of 0.0119 and the lowest MAE of 0.0721, which confirms its robustness and stability for medium short-term forecasting.

The strongest baselines remain the NBEATS family and NHITS, all reporting identical or very close MSE values of 0.0169, with MAE ranging from 0.0805 to 0.0826. Although these architectures are designed to handle longer temporal dependencies, they exhibit a significant performance gap relative to the proposed method, with an MSE increase of approximately 42%. PatchTST follows with an MSE of 0.0209 and MAE of 0.0901, indicating a further loss of accuracy as the horizon extends.

Probabilistic and hybrid models such as DeepAR, TimeMixer, and TFT show noticeably higher errors, with MSE values above 0.027. Recurrent neural networks, including GRU and LSTM deteriorate more sharply, reflecting their limited ability to propagate information over longer horizons without accumulating error. Transformer-based and convolutional architectures, including TimesNet, Informer, TCN, and their variants, consistently underperform compared to the top models, with both MSE and MAE increasing substantially.

As observed for shorter horizons, decomposition-based and graph-oriented methods, particularly Autoformer and StemGNN, yield the weakest results, with StemGNN again presenting extremely high errors. Overall, these results demonstrate that the proposed method not only maintains the best performance at horizon $h = 5$, but also exhibits a slower degradation rate than competing approaches, highlighting its effectiveness in capturing temporal dynamics under increasing forecasting difficulty.

Table 5 reports the forecasting performance for the longer horizon $h = 10$, with models ranked according to MSE. Increasing the prediction horizon leads to a substantial rise in forecasting errors for all approaches. Despite this increased difficulty, the proposed method continues to outperform all competing models by a clear margin, achieving the lowest MSE of 0.0461 and the lowest MAE of 0.1447, demonstrating strong robustness under long-horizon forecasting.

The closest competitors are NHITS and the NBEATS family, with MSE values around 0.060, which represents an increase of approximately 30% relative to the proposed approach. PatchTST follows with an MSE of 0.0642 and MAE of 0.1635, confirming its effectiveness among transformer-inspired architectures, although still significantly behind the proposed model. These results indicate that while recent deep learning based methods can partially capture long term temporal dependencies, they struggle to maintain accuracy as the horizon extends.

Models such as TimeMixer and DeepAR show a noticeable degradation, with MSE values above 0.078, suggesting limitations in handling error accumulation over longer prediction windows. Convolutional and hybrid architectures, including BiTCN and TFT, further deteriorate, while transformer-based

TABLE 5. Forecasting performance comparison for horizon $h = 10$. Models are ranked by MSE.

Model	MSE	MAE
Ours	**0.0461**	**0.1447**
NHITS	0.0598	0.1554
NBEATS	0.0603	0.1539
NBEATSx	0.0603	0.1539
PatchTST	0.0642	0.1635
TimeMixer	0.0784	0.1931
DeepAR	0.0910	0.2101
BiTCN	0.1002	0.2268
TFT	0.1087	0.2235
Informer	0.1125	0.2477
TimeXer	0.1209	0.2536
TimesNet	0.1238	0.2557
TiDE	0.1331	0.2750
GRU	0.1608	0.3030
TCN	0.1657	0.2938
iTransformer	0.1672	0.3107
VanillaTransformer	0.1809	0.3263
Autoformer	0.2005	0.3288
LSTM	0.2322	0.3619
StemGNN	0.4334	0.5333

TABLE 6. Ablation study results. Each row disables one component from our full model to evaluate its contribution.

Variant	MoE	Patch	RevIN	MSE ↓	ΔMSE
Ours (Full)	✓	✓	✓	**0.0119**	—
w/o MoE	×	✓	✓	0.0134	+12.1%
w/o Patch	✓	×	✓	0.0158	+32.6%
w/o RevIN	✓	✓	×	0.0123	+2.9%

models such as Informer, TimeXer, TimesNet, and iTransformer exhibit consistently higher errors, reflecting challenges in stability and generalization at extended horizons.

Recurrent neural networks, particularly GRU and LSTM, perform poorly at this horizon, with LSTM reaching an MSE of 0.2322 and MAE of 0.3619, highlighting the difficulty of propagating information over long time spans. As in previous horizons, Autoformer and StemGNN yield the weakest results, with StemGNN showing extremely high errors. The results for $h = 10$ reinforce the effectiveness of the proposed method, which not only achieves the best absolute performance but also exhibits a slower error growth compared to all baselines as the forecasting horizon increases.

Overall, among the baseline models, the NBEATS family (NBEATS, NBEATSx, and NHITS) consistently ranks among the top performers, followed by PatchTST. Transformer-based architectures such as Informer, Autoformer, and iTransformer exhibit higher error rates, particularly at longer horizons. This observation aligns with prior findings that standard attention mechanisms may require adaptation for time series forecasting tasks.

Table 6 presents the ablation study conducted to assess the contribution of each major component of the proposed model. The full configuration achieves the best performance, with an MSE of 0.0119, serving as the reference for all relative comparisons.

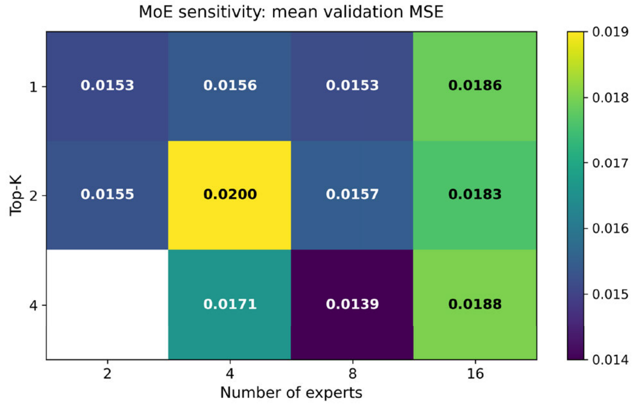


FIGURE 3. Sensitivity of MoE performance to the number of experts (K) and top-k routing. Each cell reports the mean validation MSE. Lower values indicate better performance.

Removing the MoE component leads to a noticeable degradation in performance, with the MSE increasing to 0.0134, corresponding to a relative increase of 12.1%. This result indicates that the MoE mechanism plays an important role in enhancing model expressiveness and in adapting the forecasting process to heterogeneous temporal patterns. When the patch-based representation is disabled, performance deteriorates more severely, yielding an MSE of 0.0158 and an increase of 32.6% relative to the full model. This confirms that the patching strategy is a critical element, as it enables more effective modeling of local temporal structures and reduces the burden on the global forecasting module.

In contrast, removing RevIN results in a comparatively smaller performance drop. The MSE rises slightly to 0.0123, corresponding to an increase of 2.9%, which suggests that RevIN mainly provides a stabilizing and normalization effect rather than being the primary driver of accuracy. Overall, the ablation results demonstrate that all components contribute positively to the final performance, with the patching mechanism having the largest impact, followed by the MoE, while RevIN offers a complementary but less dominant improvement.

A. MoE SENSITIVITY ANALYSIS

We conduct a sensitivity analysis to evaluate the impact of the MoE routing hyperparameters, namely the number of experts (K) and the top-k routing strategy, on forecasting performance. For that matter, we fix the overall architecture and training pipeline, enabling the MoE mechanism only in the encoder while keeping the decoder deterministic. All experiments are performed under identical conditions, including the same training schedule, optimizer configuration, and data splits. For each configuration, we vary the number of experts $K \in \{2, 4, 8, 16\}$, and the number of active experts per token $k \in \{1, 2, 4\}$, subject to $k \leq K$.

Each configuration is evaluated using the MSE on a held-out validation set. The results are averaged across multiple random seeds to reduce stochastic variability. Figure 3 presents a heatmap of the mean validation MSE as a function of the number of experts and the top-k routing parameter.

Increasing the number of experts generally improves performance up to a moderate regime. In particular, moving from $K = 2$ to $K = 8$ leads to consistent gains across most routing configurations. The best performance is achieved at $(K = 8, k = 4)$, yielding the lowest validation error.

However, further increasing the number of experts to $K = 16$ does not yield additional improvements and, in fact, slightly degrades performance. This suggests diminishing returns and possible over-parameterization, where the gating network may struggle to effectively specialize and utilize a larger pool of experts.

Second, the effect of the routing parameter k reveals a trade-off between sparsity and expressiveness. For small expert pools (e.g., $K = 2$), increasing k does not provide benefits and may introduce unnecessary redundancy. In contrast, for larger expert sets (e.g., $K = 8$), higher values of k improve performance by allowing more flexible combinations of experts. Notably, the configuration $k = 4$ significantly outperforms $k = 1$ and $k = 2$ when sufficient experts are available.

These results highlight that both K and k must be jointly tuned to balance specialization and routing efficiency. While increasing K enhances representational capacity, its benefits depend on the routing mechanism's ability to effectively select and combine experts. Similarly, increasing k improves expressiveness but reduces sparsity, leading to higher computational cost. An intermediate regime, characterized by a moderate number of experts and a small but non-trivial routing width, provides the best trade-off between accuracy and efficiency. This aligns with prior observations in sparse expert models, where controlled sparsity is crucial for achieving both scalability and strong generalization.

B. UNCERTAINTY QUANTIFICATION

Table 7 reports the conformal prediction interval performance for the proposed model at horizon $h = 5$ using the rolling conformal method. Specifically, the procedure follows Adaptive Conformal Inference (ACI) [52]: an initial split-conformal calibration is performed on a held-out calibration set to establish the conformal radius, after which the miscoverage level α_t is updated sequentially upon each newly observed labeled sample according to the ACI correction rule,

$$\alpha_{t+1} = \alpha_t + \gamma\big(\alpha - \mathbf{1}\{y_t \notin \hat{C}_t\}\big), \tag{18}$$

where α is the nominal miscoverage level, $\gamma > 0$ is a step-size hyperparameter, and $\mathbf{1}\{y_t \notin \hat{C}_t\}$ indicates whether the true value fell outside the predicted interval at time t. In parallel, a rolling buffer of the most recent w nonconformity scores (residuals) is maintained, and the conformal radius is recomputed from this buffer at each step using the current adaptive α_t. This yields an online conformal procedure that adapts to distributional shifts in the residual sequence while preserving asymptotic coverage guarantees [52].

TABLE 7. Conformal prediction interval performance.

Metric	Value
Nominal coverage	90.00
Empirical coverage	91.27
Coverage gap	+1.27
Mean interval width	0.3308
Std. interval width	0.1610
Coefficient of variation	0.49
Min interval width	0.1255
Max interval width	0.5815

This method achieves an empirical coverage of 91.27%, exceeding the nominal 90% target by 1.27 percentage points. This positive coverage gap indicates that the intervals are slightly conservative, which is preferable to under-coverage in practical applications where reliability guarantees are required.

The intervals exhibit adaptive behavior, with a coefficient of variation of 0.49. The interval widths range from 0.125 to 0.581, a ratio of approximately 4.6 : 1 between the widest and narrowest intervals. This variation reflects the model's capacity to produce tighter intervals during periods of lower predictive uncertainty and wider intervals when forecast confidence decreases. The mean interval width of 0.331 corresponds to approximately 6% of the observed data range, indicating that the uncertainty estimates remain practically useful without being overly conservative.

Figure 4 illustrates the model's predictions with conformal intervals on the test set. The top panel shows the actual time series alongside the predicted values and their associated 90% confidence intervals. The intervals consistently envelope the true values throughout the forecast period, which aligns with the empirical coverage of 91.27% reported in Table 7.

Figure 5 presents a violation heatmap depicting the temporal distribution of conformal prediction misses across the test set. Each row corresponds to a prediction window in chronological order, while columns represent forecast horizons 1 through 5. Black cells indicate instances where the true value fell outside the predicted interval, while white cells denote successful coverage. The sparse distribution of misses across the majority of the evaluation period demonstrates that the rolling conformal prediction mechanism maintains consistent coverage over time.

Notably, a concentrated band of violations appears around window indices 850 − 950, suggesting a period of heightened forecast difficulty, likely corresponding to a regime shift or anomalous behavior in the underlying time series. The rolling method responds to such episodes by updating its residual buffer and adjusting the miscoverage level α, resulting in wider intervals and a return to predominantly white cells in subsequent windows. Importantly, misses do not systematically increase with forecast horizon, indicating that the per-horizon radii adequately capture the growing uncertainty at longer lead times.

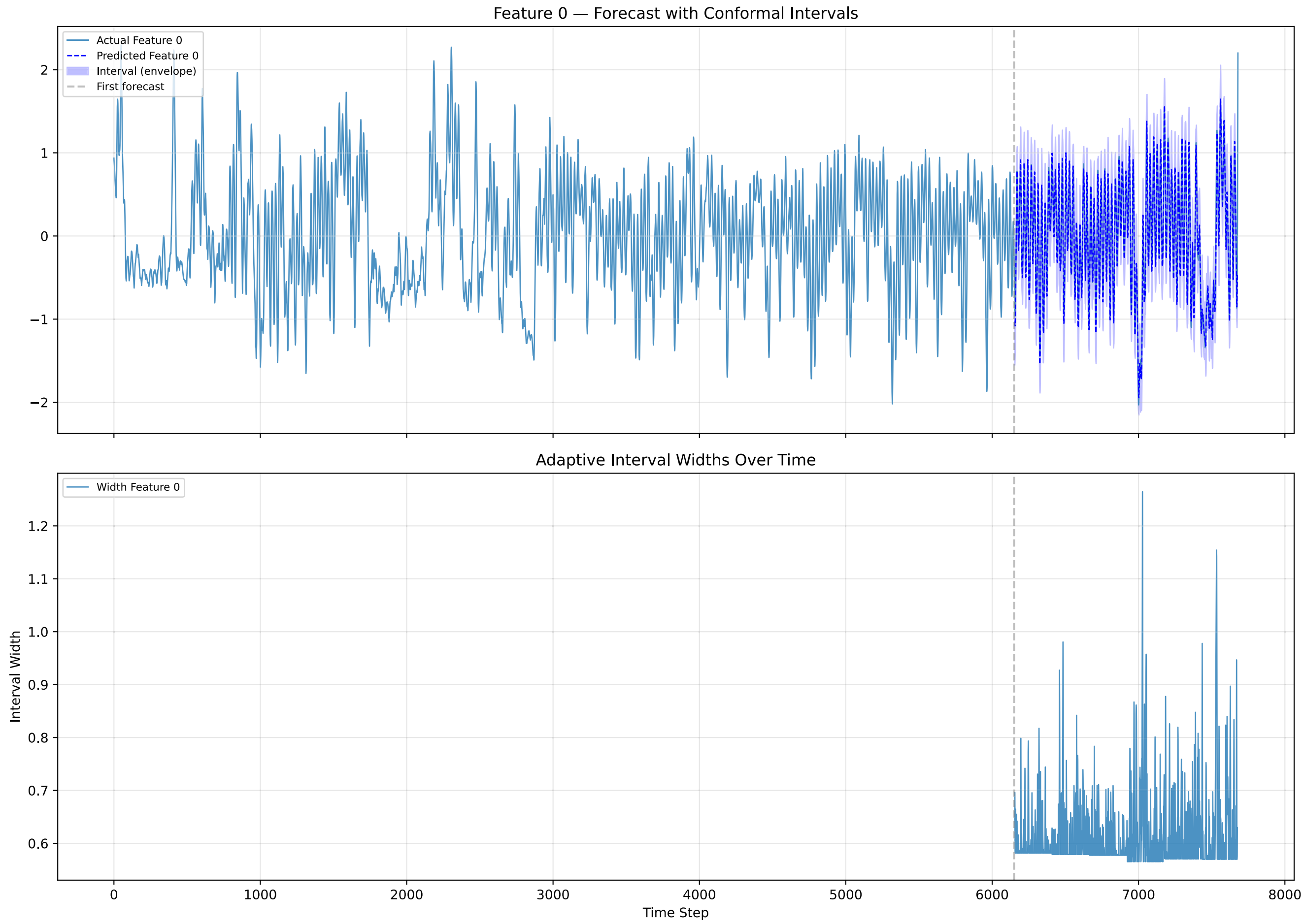


FIGURE 4. Predictions with conformal intervals on the test set. Top: actual values (solid) versus predictions (dashed) with 90% confidence intervals (shaded). Bottom: adaptive interval widths over time, showing the model's uncertainty estimates across the forecast horizon.

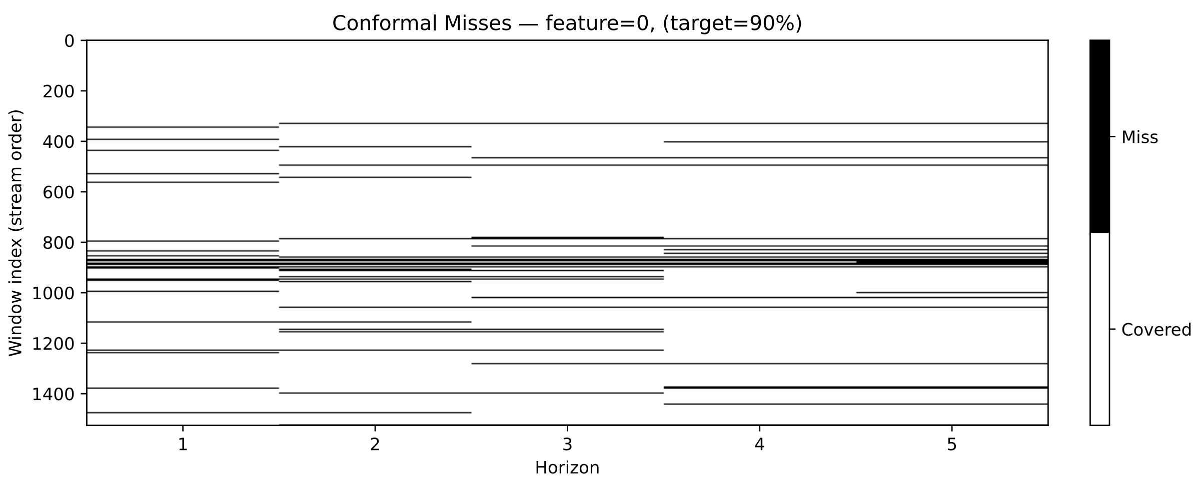


FIGURE 5. Violation heatmap for streaming conformal prediction using the rolling method. Black cells indicate prediction intervals that failed to cover the true value; white cells indicate successful coverage. The horizontal axis represents the forecast horizon, and the vertical axis shows the prediction windows in chronological order.

To further assess the generalization of the proposed network under distinct curtailment regimes, we evaluate the conformal prediction framework on two additional wind farms: Bezerra and Cumaru.

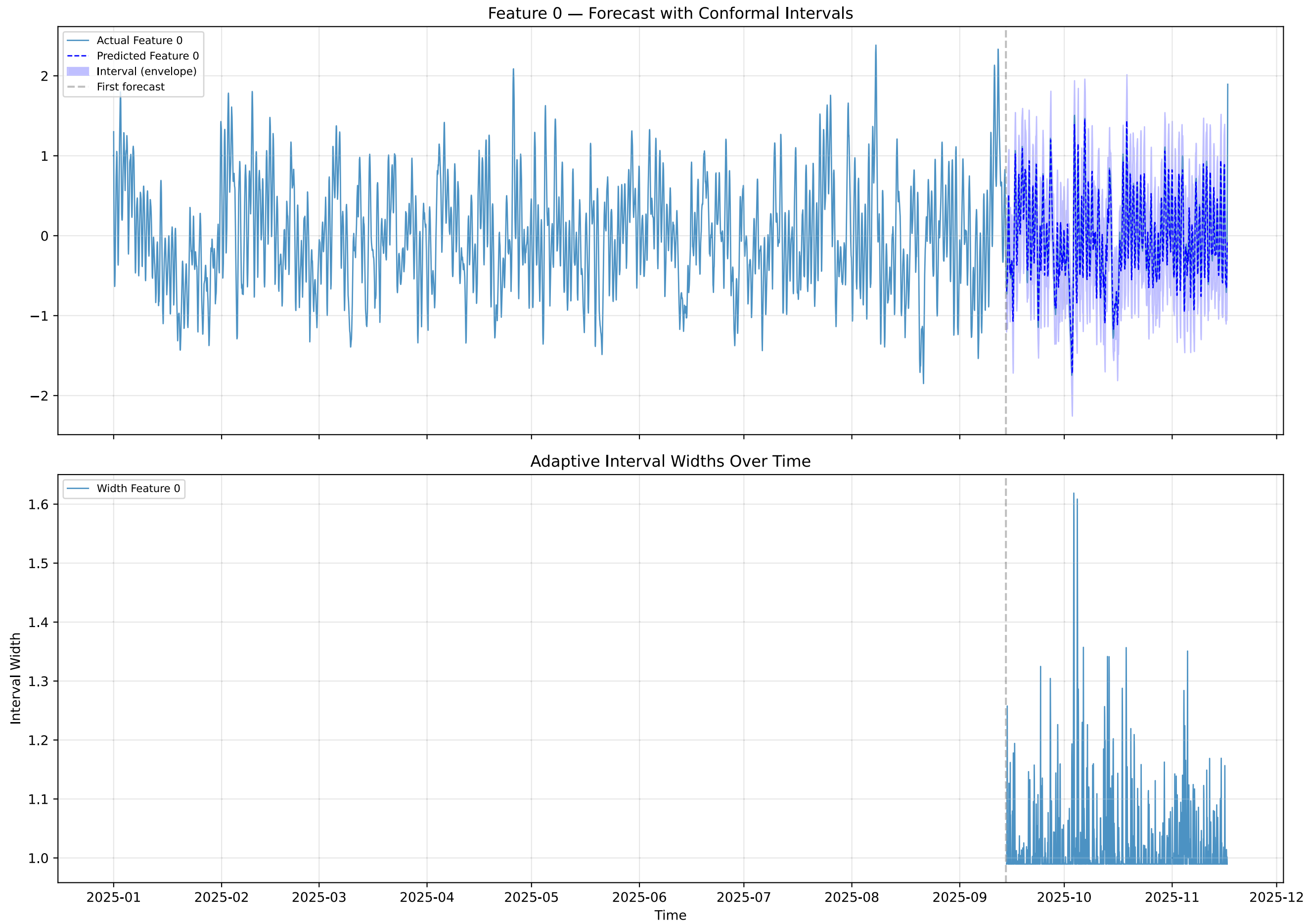


FIGURE 6. Conformal prediction results for the Bezerra wind farm.*Top:* Normalized generation time series (actual vs. predicted) with adaptive conformal envelopes; the dashed vertical line marks the start of the forecast horizon (September 2025). *Bottom:* Adaptive interval widths over the forecast period. The framework achieves a joint coverage of 91.87%against the 90% nominal target (gap: +1.87 pp), with a mean normalized interval width of 0.577 ± 0.242. Interval widths remain near their minimum of 0.27 for most of the horizon, expanding to at most 0.95 during episodes of elevated residual uncertainty.

Figure 6 presents the results for the Bezerra plant (mean generation: 102.33 MW, $\sigma = 63.44$ MW), which exhibits a sparse curtailment rate of approximately 0.60%. The adaptive intervals achieve a coverage of 91.87%, surpassing the 90% nominal target with a gap of only 1.87 percentage points, the tightest gap observed across all evaluated plants.

The mean normalized interval width of 0.577 ± 0.242 reflects compact, well-calibrated uncertainty estimates, with the adaptive mechanism producing noticeably wider bands only during the high-variability periods identified in late September and October 2025 (bottom panel of Figure 6).

Figure 7 presents the analogous results for the Cumaru plant (mean generation: 116.18 MW, $\sigma = 72.10$ MW), which has a higher curtailment prevalence of 2.89%. Here, the framework achieves a coverage of 98.30%, yielding a gap of 8.30 percentage points above the target, substantially more conservative than Bezerra. This behavior is consistent with the plant's greater output variability and higher zero-generation frequency, which drive the adaptive mechanism toward wider prediction bands (mean normalized width 0.741 ± 0.334). The larger spread of interval widths, ranging from 0.34 to 1.28, further reflects the more heterogeneous uncertainty landscape of this site. Taken together, both plants confirm that the conformal framework consistently meets or exceeds the nominal coverage guarantee while adapting interval widths to local signal characteristics.

C. COMPUTATIONAL CONSIDERATIONS

The introduction of the MoE feed-forward layer increases the total parameter count relative to lightweight baselines such as N-BEATS or recurrent models (e.g., GRU). This leads to a moderate increase in training time, attributable to the additional forward and backward passes associated with the active experts during optimization. However, because the MoE architecture relies on conditional computation, activating only a small subset of k experts per input token, the effective floating-point cost does not scale linearly with the total number of experts K, but rather with the routing width $k \ll K$, which remains small throughout our experiments.

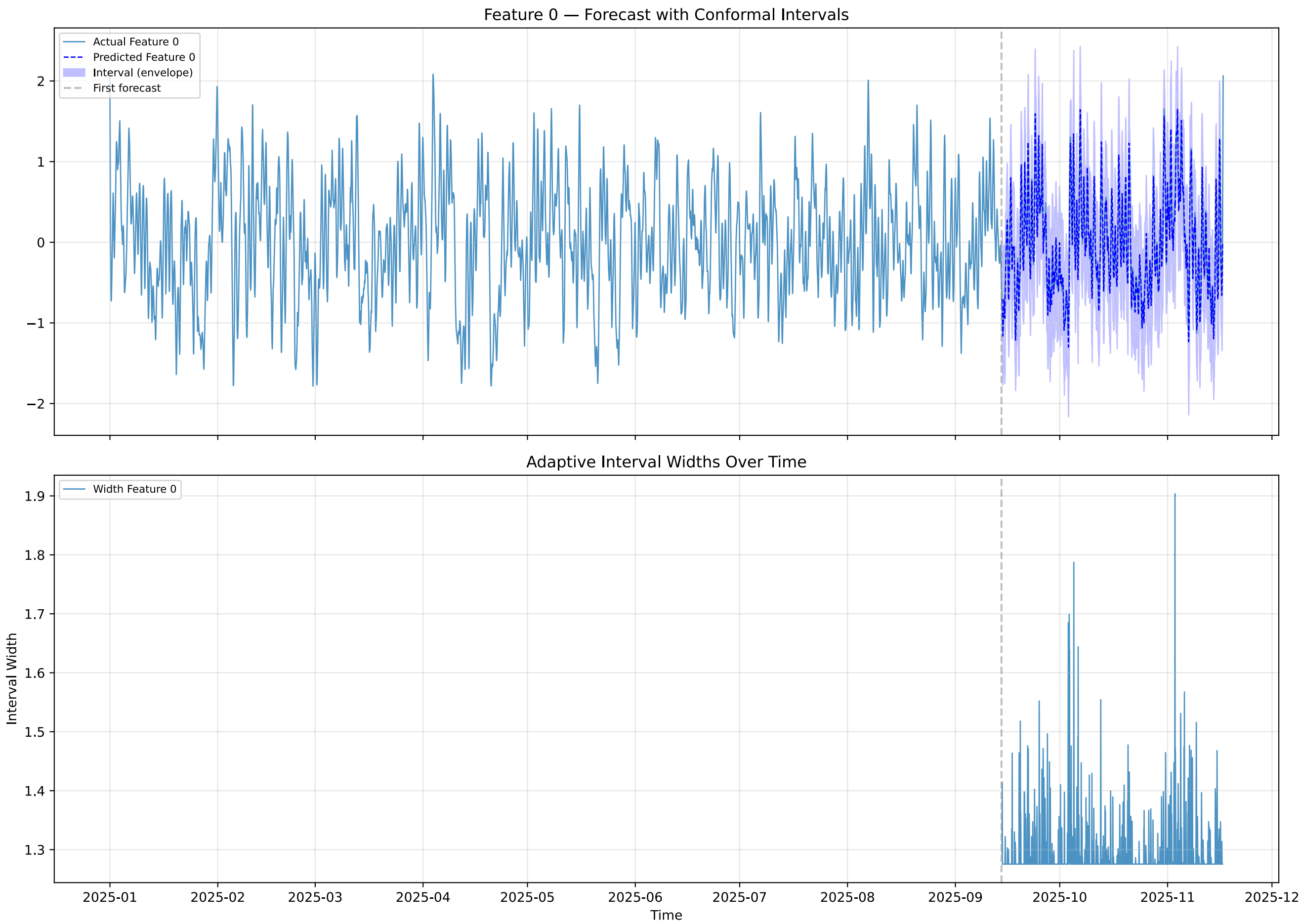


FIGURE 7. Conformal prediction results for the Cumaru wind farm.*Top:* Normalized generation time series (actual vs. predicted) with adaptive conformal envelopes; the dashed vertical line marks the start of the forecast horizon (September 2025). *Bottom:* Adaptive interval widths over the forecast period. The framework achieves a joint coverage of 98.30%against the 90% nominal target (gap: +8.30 pp), with a mean normalized interval width of 0.741 ± 0.334. Compared to Bezerra, the consistently elevated baseline width (minimum 0.34, maximum 1.28) reflects the greater heterogeneity in residual uncertainty driven by the plant's higher curtailment prevalence and larger output variance.

Regarding inference latency, the practical overhead relative to lighter baselines is negligible for the network sizes considered in this work. Two factors contribute to this: (i) only k expert networks are evaluated per forward pass, and (ii) overall model dimensions remain moderate. Furthermore, in the context of hydroelectric reservoir inflow forecasting for operational planning, predictions are issued at coarse temporal resolutions (e.g., hourly or daily horizons), where latency requirements are substantially less stringent than in real-time control applications. In such settings, the marginal increase in inference time introduced by the MoE routing mechanism is operationally inconsequential.

From a computational perspective, the model introduces a moderate increase in training cost compared to simpler architectures due to the inclusion of MoE layers and Transformer components. During inference, the cost remains practical for real-world applications, as only a few experts are evaluated and predictions are typically generated at coarse time resolutions, such as hourly or daily intervals. Therefore, the method achieves a balance between improved predictive performance and manageable computational requirements, making it suitable for operational environments.

VI. CONCLUSION

This paper addressed the problem of curtailment time-series forecasting by combining a high-capacity neural forecasting architecture with distribution-free uncertainty quantification. A Transformer-based encoder-decoder model with patch embedding, reversible instance normalization, and a MoE feed-forward block was proposed to improve predictive accuracy over multiple horizons. In parallel, a conformal prediction framework adapted to time-series data was integrated to provide valid and adaptive prediction intervals.

The experimental results showed that the proposed model consistently outperformed a wide range of strong baselines, including NBEATS, NHITS, PatchTST, and several Transformer-based variants, across short, medium, and longer forecasting horizons. The ablation study confirmed the relevance of each architectural component, with patch embedding and the MoE mechanism providing the largest performance gains, while RevIN contributed to robustness under distribution shift.

Beyond point forecasting, the conformal prediction module produced prediction intervals with empirical coverage close to or above the nominal level, while maintaining moderate and adaptive widths. The variation in interval size across time demonstrated that the method is able to reflect changes in forecast uncertainty, which is essential for operational decision-making in energy systems affected by curtailment.

The combination of accurate forecasting and reliable uncertainty quantification provides a practical tool for supporting planning and operational strategies in power systems with high penetration of renewable sources. By anticipating curtailment patterns and attaching calibrated uncertainty bounds to the forecasts, system operators, planners, and market agents can better evaluate risks, schedule resources, and design mitigation actions.

Future work will focus on extending the framework to multiscale and spatially distributed curtailment scenarios, incorporating exogenous variables such as network constraints and weather forecasts more explicitly, and evaluating online learning strategies to further improve adaptation under non-stationary operating conditions.

ACKNOWLEDGMENT

Mikhail Zimmer Heidrich thanks the Foundation for Research and Innovation of the State of Santa Catarina (FAPESC) for his scholarship to pursue the master's degree in production systems.

REFERENCES

[1] L. O. Seman, S. F. Stefenon, K.-C. Yow, L. D. S. Coelho, and V. C. Mariani, "Fourier-enhanced sequence-to-sequence latent graph neural networks for multi-node spatiotemporal forecasting in a hydroelectric reservoir," *Eng. Appl. Artif. Intell.*, vol. 167, Mar. 2026, Art. no. 113939.

[2] N. B. Amado, E. D. B. Pelegia, and I. L. Sauer, "Capacity value from wind and solar sources in systems with variable dispatchable capacity—An application in the Brazilian hydrothermal system," *Energies*, vol. 14, no. 11, p. 3196, May 2021.

[3] R. Pedrini, F. Beltrán, and E. C. Finardi, "Innovative approaches to stochastic dual dynamic programming: Two-period decomposition and parallelization in a hydrothermal scheduling case study," *J. Water Resour. Planning Manage.*, vol. 150, no. 12, Dec. 2024, Art. no. 05024014.

[4] A. Street, A. Brigatto, and D. M. Valladão, "Co-optimization of energy and ancillary services for hydrothermal operation planning under a general security criterion," *IEEE Trans. Power Syst.*, vol. 32, no. 6, pp. 4914–4923, Nov. 2017.

[5] A. L. Diniz, F. Da Serra Costa, M. E. Maceira, T. N. dos Santos, L. C. B. Dos Santos, and R. N. Cabral, "Short/mid-term hydrothermal dispatch and spot pricing for large-scale systems—The case of Brazil," in *Proc. Power Syst. Comput. Conf. (PSCC)*, Jun. 2018, pp. 1–7.

[6] T. N. Santos, A. L. Diniz, C. H. Saboia, R. N. Cabral, and L. F. Cerqueira, "Hourly pricing and day-ahead dispatch setting in Brazil: The dessem model," *Electric Power Syst. Res.*, vol. 189, Dec. 2020, Art. no. 106709.

[7] R. Santos, A. L. Diniz, and B. Borba, "Assessment of the modeling of demand response as a dispatchable resource in day-ahead hydrothermal unit commitment problems: The Brazilian case," *Energies*, vol. 15, no. 11, p. 3928, May 2022.

[8] F. J. Lachovicz, T. S. P. Fernandes, M. R. Bessa, and K. de Geus, "Optimization criteria weighing in dynamic optimal power flow performing monthly, weekly and daily hydrothermal generation dispatch," *J. Control, Autom. Electr. Syst.*, vol. 32, no. 6, pp. 1704–1717, Dec. 2021.

[9] W. N. Silva, G. G. T. Vieira, L. F. C. Simone, L. F. N. Lourenço, and M. B. C. Salles, "Data-driven analysis of wind curtailment energy imbalance and distributed generation in Brazil," in *Proc. Int. Conf. Clean Electr. Power (ICCEP)*, Jun. 2025, pp. 728–734.

[10] I. B. Renolphi, W. N. Silva, L. F. N. Lourenço, B. Z. D. Malta, T. S. Andrade, and G. T. T. Vieira, "Optimizing hybrid renewable power plants: A comparative analysis of wind-solar configurations for northeast Brazil," *Energies*, vol. 18, no. 20, p. 5329, 2025.

[11] G. T. T. Vieira, W. N. Silva, M. B. C. Salles, L. F. N. Lourenço, and R. M. Monaro, "Battery arbitrage to mitigate Brazilian curtailment due to electrical reasons," in *Proc. Int. Conf. Clean Electr. Power (ICCEP)*, Jun. 2025, pp. 455–460.

[12] G. P. D. Ponte, E. E. Rego, and F. J. G. D. S. Júnior, "Regulatory barriers to battery storage integration in Brazil: Moving beyond curtailment challenges," *Academia Green Energy*, vol. 2, no. 4, pp. 1–13, 2025.

[13] M. Laimon, "Renewable energy curtailment: A problem or an opportunity?" *Results Eng.*, vol. 26, Jun. 2025, Art. no. 104925.

[14] G. M. Longhi, C. L. T. Borges, and A. L. Diniz, "Demand response with load-block shifting in the day ahead scheduling of predominantly hydro systems: Impact on the large-scale Brazilian case," *Sustain. Energy, Grids Netw.*, vol. 43, Sep. 2025, Art. no. 101728.

[15] R. A. Campos, L. R. do Nascimento, and R. Rüther, "The complementary nature between wind and photovoltaic generation in Brazil and the role of energy storage in utility-scale hybrid power plants," *Energy Convers. Manage.*, vol. 221, Oct. 2020, Art. no. 113160.

[16] R. F. A. Menezes, G. D. Soriano, and R. R. B. de Aquino, "Locational marginal pricing and daily operation scheduling of a hydro-thermal-wind-photovoltaic power system using BESS to reduce wind power curtailment," *Energies*, vol. 14, no. 5, p. 1441, Mar. 2021.

[17] C. Root, H. Presume, D. Proudfoot, L. Willis, and R. Masiello, "Using battery energy storage to reduce renewable resource curtailment," in *Proc. IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf. (ISGT)*, Apr. 2017, pp. 1–5.

[18] R. B. S. Veras, C. B. M. Oliveira, S. L. de Lima, O. R. Saavedra, D. Q. Oliveira, F. M. Pimenta, D. C. P. Lopes, A. R. T. Junior, F. L. A. Neto, R. M. de Freitas, and A. T. Assireu, "Assessing economic complementarity in wind–solar hybrid power plants connected to the Brazilian grid," *Sustainability*, vol. 15, no. 11, p. 8862, May 2023.

[19] A. Henriot, "Economic curtailment of intermittent renewable energy sources," *Energy Econ.*, vol. 49, pp. 370–379, May 2015.

[20] A. da Silva Pontes, A. S. da Silva, N. E. Batista, A. F. Sampaio, E. M. do Sacramento, and L. C. de Lima, "Green hydrogen production from wind farm surplus in the state of ceará–Brazil," *Res., Soc. Develop.*, vol. 13, no. 10, 2024, Art. no. e110131047180.

[21] Y. Chang and H. Phoumin, "Curtailed electricity surplus from renewables for hydrogen: Economic and environmental analysis," *Yanfei Li Han Phoumin*, vol. 225, pp. 225–249, Dec. 2021.

[22] L. Migliari, D. Micheletto, and D. Cocco, "Green hydrogen generation for mitigating the impact of curtailment in floating photovoltaic–pumped hydro energy storage (FPV-PHES) systems," *J. Energy Storage*, vol. 132, Oct. 2025, Art. no. 117997.

[23] S. Sulaeman, E. Brown, R. Quispe-Abad, and N. Müller, "Floating PV system as an alternative pathway to the Amazon dam underproduction," *Renew. Sustain. Energy Rev.*, vol. 135, Jan. 2021, Art. no. 110082.

[24] H. Schermeyer, C. Vergara, and W. Fichtner, "Renewable energy curtailment: A case study on today's and tomorrow's congestion management," *Energy Policy*, vol. 112, pp. 427–436, Jan. 2018.

[25] D. Nock, "Closing the void in energy planning modeling: Integrating local realities," *Joule*, vol. 5, no. 5, pp. 1031–1033, 2021.

[26] H. Chen, J. Chen, G. Han, and Q. Cui, "Winding down the wind power curtailment in China: What made the difference?" *Renew. Sustain. Energy Rev.*, vol. 167, Oct. 2022, Art. no. 112725.

[27] C. N. Acosta Campas, M. M. Martinez, and H. F. R. Paredes, "PV power curtailment and BESS management for distribution networks: A practical approach," *IEEE Latin Amer. Trans.*, vol. 21, no. 1, pp. 133–141, Jan. 2023.

[28] L. Migliari, D. Cocco, and M. Petrollese, "Levelized cost of storage (LCOS) of battery energy storage systems (BESS) deployed for photovoltaic curtailment mitigation," *Energies*, vol. 18, no. 14, p. 3602, Jul. 2025.

[29] J. Park, K. H. Ryu, C. Kim, W. C. Cho, M. Kim, J. H. Lee, J. H. Lee, H. Cho, J. H. Lee, and J. H. Lee, "Green hydrogen to tackle the power curtailment: Meteorological data-based capacity factor and techno-economic analysis," *Appl. Energy*, vol. 340, Jun. 2023, Art. no. 121016.

[30] M. Qorbani and T. Amraee, "Long term transmission expansion planning to improve power system resilience against cascading outages," *Electric Power Syst. Res.*, vol. 192, Mar. 2021, Art. no. 106972.

[31] A. Zare Ghaleh Seyyedi, E. Akbari, S. Mahmoudi Rashid, S. A. Nejati, and M. Gitizadeh, "Application of robust optimized spatiotemporal load management of data centers for renewable curtailment mitigation," *Renew. Sustain. Energy Rev.*, vol. 204, Oct. 2024, Art. no. 114793.

[32] W. Pei, Y. Chen, K. Sheng, W. Deng, Y. Du, Z. Qi, and L. Kong, "Temporal-spatial analysis and improvement measures of Chinese power system for wind power curtailment problem," *Renew. Sustain. Energy Rev.*, vol. 49, pp. 148–168, Sep. 2015.

[33] R. N. Muniz, S. F. Stefenon, W. G. Buratto, A. Nied, R. Cardoso, C. K. Yamaguchi, and K.-C. Yow, "Time series forecasting based on multi-criteria optimization for model and filter selection applied to hydroelectric power plants," *Energy*, vol. 337, Nov. 2025, Art. no. 138688.

[34] P. Biswas, A. Rashid, A. Biswas, M. A. A. Nasim, S. Chakraborty, K. D. Gupta, and R. George, "AI-driven approaches for optimizing power consumption: A comprehensive survey," *Discover Artif. Intell.*, vol. 4, no. 1, p. 116, Dec. 2024.

[35] B. N. Rao, M. Praveen, and D. R. Babu, "A review on the role of ai in optimizing renewable energy grid management," *Int. J. Sci. Res. Eng. Manag.*, vol. 8, no. 11, pp. 1–13, 2024.

[36] L. S. Aquino, L. O. Seman, V. C. Mariani, L. D. S. Coelho, S. F. Stefenon, and G. V. González, "Spatiotemporal wind energy forecasting: A comprehensive survey and a deep equilibrium-based case study with StemGNN," *IEEE Access*, vol. 13, pp. 131461–131482, 2025.

[37] J. Li, C. Wang, and H. Wang, "Deep reinforcement learning for wind and energy storage coordination in wholesale energy and ancillary service markets," *Energy AI*, vol. 14, Oct. 2023, Art. no. 100280.

[38] Z. Li and C. Liu, "Research on calculation of spinning reserve capacity of wind power system considering multiple uncertainties," *IOP Conf., Earth Environ. Sci.*, vol. 632, 2021, Art. no. 32056.

[39] F. Schäfer, J.-H. Menke, and M. Braun, "Evaluating machine learning models for the fast identification of contingency cases," *Appl. AI Lett.*, vol. 1, no. 2, p. 19, 2020.

[40] L. Le, X. Ai, J. Fang, K. Zeng, M. Zhang, Z. Chen, P. G. Kristensen, and J. Wen, "Coordinated demand response of power-to-gas and FlexGas technologies in integrated power and gas system to accommodate wind energy," *IET Renew. Power Gener.*, vol. 14, no. 17, pp. 3300–3308, Dec. 2020.

[41] L. Liu, X. Shen, Z. Chen, Q. Sun, and R. Wennersten, "Optimal energy management of data center micro-grid considering computing workloads shift," *IEEE Access*, vol. 12, pp. 102061–102075, 2024.

[42] S. Eberlein and K. Rudion, "Optimisation, benchmark testing and comparison of droop control variants in microgrids," *IET Smart Grid*, vol. 4, no. 5, pp. 536–548, Oct. 2021.

[43] A. K. Karn, S. Hameed, M. Sarfraz, J. S. Ro, and M. R. Khalid, "Load shedding for frequency and voltage control in the multimachine system using a heuristic knowledge discovery method," *Frontiers Energy Res.*, vol. 10, Sep. 2022, Art. no. 1002064.

[44] L. Wang, C. Ma, L. Zhu, X. Wang, C. Yin, H. Cong, and T. Shi, "Optimised operation study of load aggregator based on UCCP-IDR model," *IET Gener., Transmiss. Distribution*, vol. 16, no. 12, pp. 2331–2345, Jun. 2022.

[45] V. A. Martinez Lopez, O. Isabella, M. Zeman, and H. Ziar, "Battery-less uncertainty-based control of a stand-alone PV-electrolyzer system," *J. Power Sources*, vol. 614, Sep. 2024, Art. no. 234934.

[46] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *J. Mach. Learn. Res.*, vol. 9, no. 3, pp. 371–421, 2008.

[47] H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman, "Inductive confidence machines for regression," in *Proc. Mach. Learn., ECML*, 2002, pp. 345–356.

[48] L. Guan, "Localized conformal prediction: A generalized inference framework for conformal prediction," *Biometrika*, vol. 110, no. 1, pp. 33–50, 2022.

[49] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "Predictive inference with the jackknife+," *Ann. Statist.*, vol. 49, no. 1, pp. 486–515, Feb. 2021.

[50] Y. Romano, E. Patterson, and E. J. Candès, "Conformalized quantile regression," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–19.

[51] X. Chen and Y. Xie, "Conformal prediction interval for dynamic time-series," in *Proc. 38th Int. Conf. Mach. Learn.*, 2020, pp. 11559–11569.

[52] I. Gibbs and E. J. Candès, "Adaptive conformal inference under distribution shift," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1660–1672.

[53] M. Zaffran, A. Dieuleveut, O. Féron, G. Yannig, and J. Josse, "Adaptive conformal predictions for time series," in *Proc. 39th Int. Conf. Mach. Learn.*, 2022, pp. 25834–25866.

[54] S. Sun and R. Yu, "Conformal prediction for time-series forecasting with change points," 2025, *arXiv:2509.02844*.

[55] M. Zhu, C. Guo, C. Feng, and O. Simeone, "Attention-based feature online conformal prediction for time series," 2025, *arXiv:2511.15838*.

[56] T. Kim, J. Kim, Y. Tae, C. Park, J. H. Choi, and J. Choo, "Reversible instance normalization for accurate time-series forecasting against distribution shift," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022, pp. 1–25.

[57] V. Suresh, "Benchmarking transformer variants for hour-ahead PV forecasting: PatchTST with adaptive conformal inference," in *Proc. Energies*, 2025, vol. 18, no. 18, p. 5000.

[58] C. Liu, X. Chen, P. Hu, J. Lin, J. Wang, and X. Geng, "Contrastive learning with transformer initialization and clustering prior for text representation," *Appl. Soft Comput.*, vol. 166, 2024, Art. no. 112162.

[59] Y. Kang and D. Wang, "Hardware-oriented approximations of softmax and RMSNorm for efficient transformer inference," *Micromachines*, vol. 17, no. 1, p. 84, Jan. 2026.

[60] L. O. Seman, K.-C. Yow, and S. F. Stefenon, "Sparse mixture of experts enhanced transformer architecture for short-term hydroelectric reservoir volume prediction," *Electric Power Syst. Res.*, vol. 255, Jun. 2026, Art. no. 112754.

[61] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 11106–11115.

[62] Z. Chen, Y. Deng, Q. Gu, Y. Li, and Y. Wu, "Towards understanding the mixture-of-experts layer in deep learning," in *Proc. Adv. Neural Inf. Process. Syst. 35*, 2022, pp. 23049–23062.

[63] K. Hasan, M. M. Othman, S. T. Meraj, S. Mekhilef, and A. F. B. Abidin, "Shunt active power filter based on Savitzky-Golay filter: Pragmatic modelling and performance validation," *IEEE Trans. Power Electron.*, vol. 38, no. 7, pp. 8838–8850, Jul. 2023.

[64] S. F. Stefenon, L. O. Seman, C. K. Yamaguchi, L. D. S. Coelho, V. C. Mariani, J. P. Matos-Carvalho, and V. R. Q. Leithardt, "Neural hierarchical interpolation time series (NHITS) for reservoir level multi-horizon forecasting in hydroelectric power plants," *IEEE Access*, vol. 13, pp. 54853–54865, 2025.

[65] G. Kalaiarasan, B. Premjith, and N. Madhu Manjiparambil, "A forecast-driven deep learning framework using ICEEMDAN and PatchTST for stock prediction and dynamic portfolio optimization," *IEEE Access*, vol. 13, pp. 218099–218122, 2025.

[66] Y. Shi, L. Cao, Z. Du, J. Li, S. Wang, C. Zhou, and Y. Yang, "Prediction of the low-frequency ambient noise level in the Northeast South China sea based on a gated recurrent unit network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[67] S. F. Stefenon and K.-C. Yow, "Adaptive filter-driven optimized attention-based CNN-LSTM for load forecasting in microgrids," *Results Eng.*, vol. 28, Dec. 2025, Art. no. 107470.

[68] B. Jeong and Y.-S. Jeong, "ARAScaler: Adaptive resource autoscaling scheme using ETimeMixer for efficient cloud-native computing," *IEEE Trans. Services Comput.*, vol. 18, no. 1, pp. 72–84, Jan. 2025.

[69] I. Kpereobong Friday, S. Prasanna Pati, and D. Mishra, "A multi-modal approach using a hybrid vision transformer and temporal fusion transformer model for stock price movement classification," *IEEE Access*, vol. 13, pp. 127221–127239, 2025.

[70] C. Thepprom, N. Nupairoj, and P. Vateekul, "A deep learning framework for net load forecasting with unsupervised behind-the-meter disaggregated data," *IEEE Access*, vol. 12, pp. 94958–94971, 2024.

[71] K. G. Olivares, C. Challu, F. Garza, M. M. Canseco, and A. Dubrawski, "NeuralForecast: User friendly state-of-the-art neural forecasting models," PyCon, Salt Lake City, Utah, USA, 2022. [Online]. Available: https://github.com/Nixtla/neuralforecast

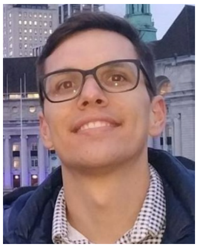
MIKHAIL ZIMMER HEIDRICH received the bachelor's degree in electrical engineering from the University of Planalto Catarinense (UNIPLAC), Brazil, in 2021. His research interests include renewable energy systems, power and storage systems, and green hydrogen technologies.

MARI AURORA FAVERO REIS received the Ph.D. degree in science and mathematics education from the Lutheran University of Brazil (ULBRA), Brazil, in 2019. She is a Professor and a Researcher at the University of Contestado (UNC). Her research interests include sustainability in energy use, applications of technologies in production systems, solar energy, and engineering education.

CRISTINA KEIKO YAMAGUCHI received the B.Sc. degree in accounting from the University of Planalto Catarinense, in 1987, the M.Sc. degree in administration from Santa Catarina State University, in 2004, and the Ph.D. degree in engineering and knowledge management from the Federal University of Santa Catarina, Brazil, in 2011.

She is the General Coordinator and a Lecturer of the Postgraduate Program in productive systems (PPGSP) in Association with UNIPLAC, UNC, UNESC, and UNIVILLE. Her research interests include knowledge management, innovation, entrepreneurship, environment, health and well-being at work, and artificial intelligence in organizations.

LAIO ORIEL SEMAN received the Ph.D. degree in electrical engineering from the Federal University of Santa Catarina, in 2017. His research interests include strategies for static and dynamic optimization, along with applications in traffic control, cyber-physical systems, and oil and gas production systems.



STEFANO FRIZZO STEFENON received the Ph.D. degree in electrical engineering from the State University of Santa Catarina (UDESC), Brazil, in 2021, and the second Ph.D. degree in computer science and artificial intelligence from the Università degli Studi di Udine (UNIUD), Italy, in 2025.

He was a Post-Doctoral Fellow Researcher with the Faculty of Engineering and Applied Science, University of Regina (UofR), SK, Canada. He is currently a Professor (Adjunct) with Lisbon School of Engineering (ISEL), Polytechnic University of Lisbon (IPL), Portugal. His research interests include electrical power systems, deep learning, computer vision, and time series forecasting.

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - ROR identifier: 00x0ma614