

EMForecaster: A Deep Learning Framework for Time Series Forecasting in Wireless Networks with Distribution-Free Uncertainty Quantification

Xavier Mootoo, Hina Tabassum, *Senior Member IEEE*, and Luca Chiaraviglio, *Senior Member IEEE*

Abstract—With the recent advancements in wireless technologies, forecasting electromagnetic field (EMF) exposure has become increasingly critical to enable proactive network spectrum and power allocation, as well as network deployment planning. In this paper, we develop a deep learning (DL)-empowered time series forecasting framework referred to as *EMForecaster*. The proposed DL architecture employs patching to process temporal patterns at multiple scales, complemented by reversible instance normalization and mixing operations along both temporal and patch dimensions for efficient feature extraction. We then augment EMForecaster with a conformal prediction mechanism, which is independent of the data distribution, to enhance the trustworthiness of model predictions through uncertainty quantification of forecasts. In particular, the conformal prediction mechanism ensures that the ground truth lies within a prediction interval with target error rate α , where $1 - \alpha$ is referred to as coverage. However, a trade-off exists, as increasing coverage often results in wider prediction intervals. To address this challenge, we propose a new metric referred to as *Trade-off Score*, that balances the trustworthiness of the forecast (i.e., coverage) and the width of prediction interval. Our empirical evaluation demonstrates that EMForecaster achieves superior performance across diverse EMF datasets, spanning both short-term and long-term prediction horizons. In point forecasting tasks, EMForecaster substantially outperforms current state-of-the-art DL approaches, showing improvements of 53.97% over the Transformer architecture and 38.44% over the average of all baseline models. In terms of conformal prediction performance, EMForecaster exhibits excellent balance between prediction interval width and coverage, as measured by the coverage-width tradeoff score. This balance is comparable to DLinear’s performance while showing marked improvements of 24.73% over the average baseline and 49.17% over the Transformer architecture. We have included all code used to produce the experiments at github.com/xmootoo/emforecaster.

Index Terms—Electromagnetic Field (EMF), time series, forecasting, conformal prediction, deep learning.

I. INTRODUCTION

Amid rapid advancements in wireless technology, concerns are growing about the potential increase in electromagnetic field (EMF) exposure due to newer radio transmission frequencies, massive radiating elements, and the dense deployment of network infrastructure in fifth-generation (5G) and beyond (B5G) [1], [2]. EMF exposure can induce thermal effects, such as heating of the exposed tissues and organs, depending

on the level of radiation absorbed. Consequently, regulatory authorities such as International Commission on Non-Ionizing Radiation Protection (ICNIRP) define rigorous limits on the maximum radio-frequency EMF exposure between 100 kHz and 300 GHz [3], which are adopted by many nations across the world. ICNIRP guidelines initially published in 1998 have been revised in 2020 to reflect recent advancements in scientific understanding of EMF exposure effects. In addition to the ICNIRP guidelines, several countries adopt more stringent EMF regulations [4].

EMF monitoring and forecasting is thus becoming critical alongside the wireless technological advancements to ensure regulatory compliance, *perform proactive network deployment, spectrum allocation, and power management*, and address public health concerns. By providing insights into long-term EMF trends, reliable EMF forecasting enables proactive risk management and informed decision-making, thereby balancing wireless network innovation with safety and public trust. EMF forecasting capabilities allow network operators to optimize infrastructure placement and power settings while maintaining EMF levels within regulatory limits, and enable public health agencies to better assess population exposure patterns and implement preventive measures where necessary.

A. Background Work

To date, a variety of research studies have focused on EMF exposure modeling and resource optimization in cellular networks using analytical or simulation-based methods. For instance, stochastic geometry tools such as Poisson Point Processes (PPP) and shot-noise processes have been considered to model the spatial distribution of EMF sources and analyze their impact on EMF exposure and coverage in cellular networks [5]–[12]. On the other hand, several studies have focused on minimizing EMF exposure in cellular systems [13], [14]. Resource management schemes, including user-scheduling [15], spectrum allocation [13], multi-cell scheduling [16], user association [17], beamforming optimization [18], [19], and cross-layer protocols [20], [21] have also been considered to minimize EMF exposure. Network-based approaches considered minimizing EMF exposure by optimizing cellular network planning, as discussed in [22]–[24].

Nevertheless, the aforementioned model-based approaches often suffer from inaccuracies due to necessary modeling assumptions, making them less applicable in complex real-world environments. Developing analytical models for such scenarios is not only intractable, but also computationally

X. Mootoo and H. Tabassum are with the Department of Electrical Engineering and Computer Science, York University, Toronto, ON, Canada (E-mail: xmootoo@my.yorku.ca, hinat@yorku.ca). L. Chiaraviglio is with the Department of Electronic Engineering, University of Rome Tor Vergata, Rome, Italy and Consorzio Nazionale Interuniversitario per le Telecomunicazioni (CNIT), Parma, Italy (E-mail: luca.chiaraviglio@uniroma2.it).

demanding, as they result in intricate mathematical expressions. Moreover, these methods are constrained by specific wireless channel assumptions, limiting their applicability to dynamic and unpredictable network conditions. Additionally, they lack the ability to learn from historical patterns or trends, preventing them from effectively performing proactive network resource management and planning. As a result, their practicality in real deployments remains a significant challenge.

Recently, deep learning (DL) approaches have gained significant attention in time series forecasting, an area historically dominated by statistical models and traditional machine learning methods [25]. Traditional techniques, such as Autoregressive Integrated Moving Average (ARIMA) [26] combines differencing, autoregression, and moving average components to model linear relationships in stationary time series data [26], and has long been a standard benchmark in time series modeling, often outperforming more modern methods in standardized testing environments.

However, the advent of sophisticated DL architectures, such as transformers, has begun to shift the paradigm [25]. DL methods offer distinct advantages over conventional forecasting methods. These models can learn complex nonlinear relationships without explicit feature engineering and adapt to dynamic changes by modeling hierarchical structures. Unlike traditional methods that require manual pre-processing such as seasonal decomposition, DL models can process raw time series data directly while handling multiple input variables and their interactions.

DL methods are increasingly being adopted for EMF pattern analysis and prediction due to their ability to handle complex, nonlinear relationships in the data. In [27], Kiouvrekis *et al.* employed hierarchical clustering to analyze EMF measurements across 205 schools in Thessaly, Greece. The findings reveal that EMF exposure patterns were independent of urban population density in the 27 MHz–3 GHz range. Bakcan *et al.* integrated static electric field measurements with artificial neural networks (ANN) in [28] to predict EMF exposure at unmeasured locations across a campus setting, validating results against Information and Communication Technology Agency (ICTA) and ICNIRP standards. Focusing on temporal prediction, Pala *et al.* in [29] analyzed EMF dataset comprised of 60 monthly mean values in the range of 1 Hz - 400 kHz using the Wavecontrol SMP2 device, and demonstrated that long short-term memory (LSTM) outperforms recurrent neural network (RNN) and traditional statistical approaches. However, the model was based on a very limited dataset for both training and testing, raising concerns about overfitting.

Moreover, the study did not consider fundamental time series parameters such as *lookback window* (how much historical data to use), *forecast horizon* (how far ahead to predict), and *rolling window stride* (how to slide the prediction window through time). Furthermore, the dataset's frequency range was significantly low to encompass modern wireless networks.

Very recently, Nguyen *et al.* in [30] demonstrated the effectiveness of Transformers [31] over the core DL architectures, including the multilayer perceptron (MLP), LSTM, and convolutional neural networks (CNNs) for short-term EMF forecasting using the data provided in [32]. The training data

was collected from real measurements taken in a city of Ordu, Turkey [32], considering multi-step input and output sequences. Building upon the pioneering application of DL to EMF time series forecasting in [30], we note that the framework is restricted to short-term prediction windows (maximum duration of 20 minutes). In addition, although the dataset was carefully collected with 24-hour recordings from 17 sites at 15-second intervals, it is relatively modest by DL standards, where larger datasets typically enable more robust performance evaluation. The limited temporal scope of the data presents challenges in capturing multi-day patterns and long-term exposure trends, as the resulting non-stationary time series limits model predictive capabilities. Furthermore, our analysis in Section V reveals high correlation in the data, suggesting limited diversity in EMF exposure scenarios, which may impact the ability to assess each model's generalizability. Moreover, the work employed classical DL architectures, not specifically designed for temporal modeling. Finally, although hyperparameter values were documented, the absence of ablation studies and sensitivity analyses makes it difficult to fully understand each model's behavior and validate design choices.

B. Contributions

None of the existing research has developed a reliable DL-empowered time series forecasting framework for wireless cellular network applications in general, nor specifically for EMF forecasting. The black-box nature of DL models makes it difficult to interpret predictions or understand the underlying decision-making process, raising concerns in applications where transparency and reliability are paramount. To this end, our contributions are listed as follows:

- We develop a novel DL-driven time series forecasting solution, referred to as *EMFForecaster*, to forecast EMF patterns considering both long-term and short-term time series, across a variety of locations with a forecast horizon of up to 50 hours. *EMFForecaster* exploits several modern DL techniques such as Reversible Instance Normalization (RevIN), which minimizes the impact of distribution shifts or non-stationarity in the EMF data, in addition to a patching and patch-embedding module. *EMFForecaster* then applies the spatiotemporal backbone (STB) to the patch-embedded data where we employ a mixing operation on the learned patch representations, enabling learning of hierarchical patterns at multiple scales.
- We augment the *EMFForecaster* with a CP framework to ensure that the ground truth lies within a certain prediction interval with *error rate* α . The performance is measured in terms of the independent coverage (IC) of each individual forecast point, the joint coverage (JC) of the forecast points across the forecasted window, and mean prediction interval width (MIW). Unlike traditional uncertainty quantification methods, which often depend on strong distributional assumptions, CP is agnostic to the original data distribution, making it compatible with any DL model and dataset.
- We propose the Trade-off Score (TOS)—a unified metric that evaluates the effectiveness of CP methods by considering two key aspects: (1) how often the true values fall within the predicted ranges (coverage) and (2) how wide these predicted

ranges are (interval width). TOS quantifies how well a method balances this fundamental trade-off. While wider prediction ranges are more likely to contain the true values, they provide less precise and thus less useful forecasts. A higher TOS indicates better performance, as the true values fall within the prediction range without requiring excessively wide intervals.

- Our experiments, covering a wide variety of environments and forecast horizons, display EMForecaster’s capabilities over the existing DL models. In point forecasting, EMForecaster demonstrates improvements of 53.97% over the Transformer architecture and 38.44% over the average of all baseline models. When compared to the DLinear architecture, EMForecaster maintains 8.01% gains. For conformal forecasting, EMForecaster provides an excellent balance between prediction interval coverage and width, as measured by the TOS. This balance is comparable to DLinear’s performance while showing marked improvements of 24.73% over the average baseline and 49.17% over the Transformer architecture.

- We conduct a comprehensive EMF data analysis by applying spectral decomposition, stationarity testing, and spatial correlation analysis on both short-term and long-term EMF exposure datasets. Stationarity is particularly crucial, which implies that time windows are reasonably independently and identically distributed (i.i.d.), therefore satisfying the exchangeability condition necessary for conformal prediction (see Section II.C for more details).

- The github repository link is given by <https://github.com/xmootoo/emforecaster>

C. Paper Organization

The remainder of this paper is organized as follows. Section II introduces fundamental concepts in time series forecasting and Section III discusses key time series forecasting architectures. Section IV presents our proposed DL model, EMForecaster, followed by conformal prediction for uncertainty quantification in Section V. Section VI details the datasets, experimental setup, and comprehensive data analysis. Section VII presents the baseline models, results, and discussion. Finally, Section VIII concludes the paper.

II. FUNDAMENTALS OF TIME SERIES FORECASTING

In this section, we discuss the fundamental concepts related to time series forecasting, such as stationarity and periodicity as well as the concepts related to building trustworthy forecasting such as exchangeability and conformal prediction.

A. General Definitions

Let $(x_t)_{t=1}^T$ be a time series of total length T . Given a subsequence (or window) $\mathbf{x} = (x_t)_{t=K}^{K+L}$ with starting index K and sequence length L , the goal of time series forecasting is to predict its continuation $\mathbf{y} = (x_t)_{t=L+K+1}^{K+L+O}$, where O is the length of the forecast horizon. We will typically view $\mathbf{x} \in \mathbb{R}^L$ and $\mathbf{y} \in \mathbb{R}^O$ as vectors. When preprocessing the initial time series $(x_t)_{t=1}^T$, we use the sliding-window technique, which samples windows $(x_t)_{t=1}^L, (x_t)_{t=2}^{L+1}, \dots, (x_t)_{t=T-L+1}^T$ by shifting the starting index K by 1 at each step. In general,

this forms our dataset, while keeping training, validation, and testing sets temporally separated, i.e., each split consists of mutually exclusive regions of the original series. Within each split (`train`, `val`, `test`), we apply the same sliding-window sampling to generate its own windows from disjoint time intervals, ensuring no window leakage across splits. We also maintain the temporal ordering $\text{training} \leq \text{validation} \leq \text{testing}$ to prevent look-ahead bias for non-IID series.

B. Stationarity of Time Series

Stationarity is an invariant property where statistical properties of a time series remain consistent over time. Stationarity can be classified as: *weak stationarity*, which only considers the covariance of a time series, and *strict stationarity*, which assumes distributions remain invariant over time. While classical models such as ARIMA [26] explicitly require stationarity to maintain their statistical properties, DL has introduced new perspectives on handling non-stationary data [33]. Traditional approaches necessitate explicit transformation of non-stationary data through techniques such as differencing or seasonal decomposition, whereas DL models may learn and adapt to certain types of non-stationarity through hierarchical feature learning capabilities. However, empirical evidence suggests that even for DL models, highly non-stationary distributions can lead to suboptimal performance when compared to traditional models. Thus, characterizing the stationarity of a time series can improve the forecasting quality [34]. One standard method to test stationarity is the *Augmented Dickey-Fuller (ADF) test*, which examines the presence of unit roots in time series using an autoregressive model estimated through ordinary least squares regression [35]:

$$x_t = c + w_1 t + w_2 x_{t-1} + \sum_{i=1}^p \phi_i \Delta x_{t-i} + \varepsilon_t \quad (1)$$

where x_t is the time series value at time t , c is the estimated drift constant, w_1 is the estimated time trend coefficient capturing deterministic trends, w_2 is the estimated process root determining persistence, ϕ_i are the estimated autoregressive coefficients capturing short-term dynamics, Δx_{t-i} represents lagged differences $(x_{t-i} - x_{t-i-1})$, and ε_t is white noise with zero mean and constant variance. The test evaluates the hypothesis $H_0 : w_2 = 1$ (non-stationary) against $H_a : w_2 < 1$ (stationary) with test statistic [35]:

$$\text{ADF} = \frac{\hat{w}_2 - 1}{\text{SE}(\hat{w}_2)} \quad (2)$$

where $\hat{w}_2$ is the estimated value of w_2 and $\text{SE}(\hat{w}_2)$ is its standard error measuring estimation uncertainty. A more negative test statistic provides stronger evidence for stationarity, with the additional terms in the regression controlling for trends and autocorrelation to ensure robust testing. In our analysis provided in Figure 6, ADF testing reveals distinct stationarity characteristics among different EMF datasets.

C. Conformal Prediction and Exchangeability in Time Series

The goal of CP is to provide a *prediction region*, Γ^α , for a given *significance level (or error rate)* $\alpha \in (0, 1)$, ensuring

that the ground truth falls within this region with a probability of at least $(1 - \alpha)$. This framework's key advantage lies in its ability to produce valid uncertainty estimates for any black-box predictor, making it particularly valuable in domains where reliable uncertainty quantification is crucial [36]. CP enables constructing prediction intervals with guaranteed finite-sample coverage under minimal assumptions on the data distribution, requiring only exchangeability of the data.

Exchangeability is a key statistical property where the order of time observations does not affect their joint probability—similar to how shuffling a well-mixed deck of cards does not change the probability of drawing any particular sequence. This property underpins CP's theoretical guarantees, but poses challenges for time series analysis. Time series data inherently violates exchangeability since future observations can depend on past states. In wireless networks, for instance, network utilization at time t can depend on its state at $t - 1$. To address this, several adaptations have been proposed, notably in [36], where the strict exchangeability assumption is relaxed through locally exchangeable windows. This approach preserves CP's validity while acknowledging the temporal nature of the data by treating similar historical patterns as locally exchangeable within defined contexts.

D. Fast-Fourier Transform and Periodicity

The Fast Fourier Transform (FFT) decomposes complex temporal patterns into their constituent frequency components, enabling efficient characterization of cyclical behaviors within the data [37]. Mathematically, the FFT efficiently computes the Discrete Fourier Transform (DFT), which decomposes a time series $\mathbf{x} = (x_t)_{t=1}^T$ into its frequency components by:

$$X_k = \sum_{t=1}^T x_t e^{-2\pi i k t / T}, \quad k = 0, \dots, T - 1 \quad (3)$$

where the amplitude $|X_k|$ reveals the dominant frequency components (cyclical patterns) in the time series. By transforming time-domain signals into the frequency domain, FFT analysis reveals both the presence and strength of periodic patterns. This spectral decomposition proves particularly valuable in scenarios where multiple overlapping cycles may exist.

III. STATE-OF-THE-ART DL MODELS FOR TIME SERIES FORECASTING: A REVIEW

In this section, we provide a review of existing DL models for time-series forecasting and summarized their benefits and drawbacks in **Table 1**. The described DL models have been applied for a variety of wireless applications to date including EMF prediction [27]–[30], network traffic prediction [38]–[42], channel prediction [42], [43], and network quality-of-service (QoS) prediction [44], [45].

A. Multilayer Perceptron (MLP)

A Multilayer Perceptron (MLP) is a type of feedforward neural network consisting of an input layer, one or more hidden layers, and an output layer, where each layer is fully connected to the next. The MLP can model complex, non-linear

relationships by applying a non-linear activation function σ at each node in the hidden and output layers. Given an input time series $\mathbf{x} \in \mathbb{R}^L$, the transformation at the k -th hidden layer is defined as

$$\mathbf{h}^{(k)} = \sigma \left(\mathbf{W}^{(k)} \mathbf{h}^{(k-1)} + \mathbf{b}^{(k)} \right),$$

where each $\mathbf{W}^{(k)}$ is a learnable weight matrix and $\mathbf{b}^{(k)}$ is a learnable bias vector for layer k , with $\mathbf{h}^{(0)} = \mathbf{x}$. The output layer computes the predicted forecast $\hat{\mathbf{y}} \in \mathbb{R}^O$ as:

$$\hat{\mathbf{y}} = \sigma_{\text{out}} \left(\mathbf{W}^{(K+1)} \mathbf{h}^{(K)} + \mathbf{b}^{(K+1)} \right),$$

where K is the number of hidden layers, and σ_{out} is an activation function appropriate for the task, for example, softmax for time series classification or the identity function for time series forecasting. By stacking multiple layers and introducing nonlinearity, MLPs are capable of approximating complex functions, making them widely applicable in various machine learning and time series forecasting contexts.

B. Convolutional Neural Network (CNN)

A 1D Convolutional Neural Network (CNN) is a type of neural network designed to process sequential data, such as time series or one-dimensional signals, by learning spatial or temporal hierarchies through convolutional operations [52], [53]. In a 1D CNN, the input signal $\mathbf{x} \in \mathbb{R}^L$ is convolved with a set of learnable filters (or kernels), producing feature maps that capture local patterns. The output of the j^{th} filter in layer k is given by:

$$\mathbf{h}_j^{(k)}[n] = \sigma \left(\sum_{m=1}^{M_k} \mathbf{w}_j^{(k)}[m] \mathbf{h}^{(k-1)}[n - m + 1] + b_j^{(k)} \right),$$

where $\mathbf{w}_j^{(k)}$ is the filter of size M_k , $b_j^{(k)}$ is the bias term, σ is a non-linear activation function, and n indexes the position in the sequence. The feature maps from one layer are passed to the next, allowing the network to progressively learn higher-level abstractions. At the output layer, these learned features can be fed into a fully connected layer or another downstream task. The use of convolutional layers makes 1D CNNs particularly effective at capturing local dependencies in sequential data while reducing the number of parameters compared to fully connected architectures.

C. Long-Short Term Memory (LSTM) Networks

The *Long Short-Term Memory* (LSTM) network is a specialized recurrent neural network (RNN) designed to model long-term dependencies in sequential data by maintaining a memory cell $\mathbf{c}_t$ at each time step [54]. This structure mitigates vanishing and exploding gradient problems, making LSTMs effective for univariate time series forecasting and classification. At each time step t , a hidden state $\mathbf{a}_t \in \mathbb{R}^d$ serves as a summary of the information processed by the model up to time t where d is the hidden state dimension. The goal of $\mathbf{a}_t$ is to capture both short-term patterns and contextual signals needed for immediate predictions. The memory cell $\mathbf{c}_t \in \mathbb{R}^d$, on the other hand, provides a more stable, long-term storage,

TABLE I: Comparison of Deep Learning Models for Time Series Forecasting in Wireless Networks

Model	Pros	Cons
MLP [27], [28], [30], [38], [46]	Simple architecture; Fewer parameters; Fast training/inference; Low computational cost; Adaptable to different inputs	Limited temporal modeling; Poor with long-range dependencies; Requires feature engineering; Overfits on complex EMF patterns; Cannot capture periodic patterns
CNN [30], [42], [46]–[48]	Captures local patterns effectively; Efficient parameter sharing; Detects signal hierarchies; Noise-robust; Parallel processing	Limited receptive field; Issues with irregular sampling; Weak global context modeling; Fixed kernel limitations; Requires careful design
LSTM [29], [30], [39], [41], [43], [45], [46]	Designed for long-term dependencies; Memory cell preserves patterns; Effective gating mechanisms; Handles variable sequences; Models complex dynamics	Sequential processing limits parallelization; Computationally expensive for long sequences; Gradient issues; Many parameters; Difficult interpretation
Transformer [30], [31], [44], [49]	Models dependencies at any distance; Parallel computation; Global context modeling; Multi-head attention; Scales to long sequences	Quadratic complexity; High memory usage; Requires positional encoding; No inherent sequential understanding; Weak with fine-grained patterns
PatchTST [40], [50]	Captures semantic information through patches; Channel independence reduces overfitting; Efficient multi-channel processing; Better noise handling; Superior time series performance	Sensitive to patch parameters; May lose fine-grained details; Parameter tuning challenges; Higher complexity than linear models; More complex implementation
DLinear [40], [51]	Effective seasonal-trend decomposition; Computationally efficient; Outperforms many Transformer models; Resistant to overfitting; Simpler architecture	Limited non-linear modeling; Fixed decomposition constraints; Lower expressiveness; Critical hyperparameter selection; Struggles with non-stationary data
EMForecaster (Proposed)	Patching with spatiotemporal mixing operations; Enhanced feature extraction with reversible normalization; Superior performance across diverse datasets; Balanced prediction intervals for conformal prediction; Effective for both short/long-term forecasting; Low computational cost	More complex architectural design; Requires deeper understanding of the model; More challenging initial implementation; Requires adequate volume of data to generalize

selectively preserving information that might be important for future time steps. Together, the hidden state and memory cell allow the model to balance between remembering essential past information and responding to new inputs. Each gate (update, forget, and output) produces a vector $\Gamma \in \mathbb{R}^d$ that modulates the flow of information, ensuring that the model learns to focus on relevant parts of the sequence while discarding irrelevant or redundant details. This formulation enables the LSTM to handle both short-term dependencies (through $\mathbf{a}_t$) and long-term dependencies (through $\mathbf{c}_t$) effectively, which is crucial for univariate time series forecasting.

Candidate memory cell and gates. The LSTM computes a *candidate memory cell* $\tilde{\mathbf{c}}_t$ as follows:

$$\tilde{\mathbf{c}}_t = \tanh(\mathbf{W}_c[\mathbf{a}_{t-1}, x_t] + \mathbf{b}_c) \quad (1)$$

Here, $[\mathbf{a}_{t-1}, x_t] \in \mathbb{R}^{d+1}$ denotes the concatenation of the previous hidden state $\mathbf{a}_{t-1}$ and the current scalar input x_t . The weight matrix $\mathbf{W}_c \in \mathbb{R}^{d \times (d+1)}$ maps the concatenated vector to the hidden size d , and $\mathbf{b}_c \in \mathbb{R}^d$ is the bias vector.

The gates are computed as:

- **Update Gate:** Controls how much of the candidate memory cell to incorporate into the current memory cell.

$$\Gamma_u = \sigma(\mathbf{W}_u[\mathbf{a}_{t-1}, x_t] + \mathbf{b}_u) \quad (4)$$

where $\mathbf{W}_u \in \mathbb{R}^{d \times (d+1)}$ and $\mathbf{b}_u \in \mathbb{R}^d$.

- **Forget Gate:** Determines how much of the previous memory cell to retain.

$$\Gamma_f = \sigma(\mathbf{W}_f[\mathbf{a}_{t-1}, x_t] + \mathbf{b}_f) \quad (5)$$

where $\mathbf{W}_f \in \mathbb{R}^{d \times (d+1)}$ and $\mathbf{b}_f \in \mathbb{R}^d$.

- **Output Gate:** Controls the prominence of the new activation.

$$\Gamma_o = \sigma(\mathbf{W}_o[\mathbf{a}_{t-1}, x_t] + \mathbf{b}_o) \quad (6)$$

where $\mathbf{W}_o \in \mathbb{R}^{d \times (d+1)}$ and $\mathbf{b}_o \in \mathbb{R}^d$.

Memory and activation updates. The memory cell update and new hidden state (activation) are computed as:

$$\mathbf{c}_t = \Gamma_u \odot \tilde{\mathbf{c}}_t + \Gamma_f \odot \mathbf{c}_{t-1} \quad (7)$$

$$\mathbf{a}_t = \Gamma_o \odot \tanh(\mathbf{c}_t) \quad (8)$$

Here, $\odot$ denotes element-wise multiplication. Given a time series $\mathbf{x} = (x_t)_{t=1}^L$, we first obtain the entire sequence of hidden states $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_L\}$ recursively. A common approach is to use the final hidden state:

$$\mathbf{a}_{\text{final}} = \mathbf{a}_L \in \mathbb{R}^d \quad (9)$$

This representation is then passed through a fully connected layer to generate the forecast:

$$\hat{y} = \mathbf{w}_{\text{out}}^T \mathbf{a}_{\text{final}} + b_{\text{out}} \quad (10)$$

where $\mathbf{w}_{\text{out}} \in \mathbb{R}^d$ and $b_{\text{out}} \in \mathbb{R}$ produce a scalar output $\hat{y} \in \mathbb{R}$ for the one-step-ahead forecast. For multi-step forecasting in the univariate case, we can either: (1) use a single output neuron and apply the model recursively, feeding each prediction back as input; or (2) use multiple output neurons to directly predict the entire forecast horizon:

$$\hat{\mathbf{y}} = \mathbf{W}_{\text{out}} \mathbf{a}_{\text{final}} + \mathbf{b}_{\text{out}} \quad (11)$$

where $\mathbf{W}_{\text{out}} \in \mathbb{R}^{O \times d}$ and $\mathbf{b}_{\text{out}} \in \mathbb{R}^O$ produce an output $\hat{\mathbf{y}} \in \mathbb{R}^O$, with O being the length of the forecast horizon.

D. The Transformer

The Transformer architecture. The *Transformer* is a versatile model architecture that uses self-attention mechanisms to capture dependencies across all time steps in a sequence, regardless of their temporal distance [31]. This makes it particularly suitable for time series forecasting, where both short-term and long-term patterns can be critical. Unlike recurrent models such as LSTMs, which process sequences step-by-step, the Transformer architecture processes the entire input sequence in parallel, enabling direct interactions between all positions through its *self-attention* mechanism. In this section, we focus on the *encoder-only* Transformer architecture for univariate time series forecasting.

Input and positional encoding. For a univariate time series with input sequence $\mathbf{x} \in \mathbb{R}^L$, where L is the input sequence length, we first project each scalar value to the model's hidden dimension using a learned projection matrix $\mathbf{W}_{\text{in}} \in \mathbb{R}^{d \times 1}$. Since the Transformer lacks an inherent notion of temporal order, a *positional encoding* vector $\mathbf{p}_t = (p_{t,1}, \dots, p_{t,d})^T \in \mathbb{R}^d$ is then added to each projected value, with each entry defined by:

$$p_{t,i} = \begin{cases} \cos(t/10000^{2i/d}) & \text{if } i \text{ is even} \\ \sin(t/10000^{2(i-1)/d}) & \text{if } i \text{ is odd} \end{cases} \quad (12)$$

The full projection transformation for each time point t is given by:

$$\mathbf{z}_t = \mathbf{W}_{\text{in}} x_t + \mathbf{p}_t, \quad (13)$$

This produces an initial sequence representation $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_L]^T \in \mathbb{R}^{L \times d}$.

The self-attention mechanism. The core of the Transformer encoder is the self-attention mechanism, which allows each time step to attend to all other time steps in the sequence. For each embedded input $\mathbf{z}_t$, the model computes three vectors: the query $\mathbf{q}_t$, key $\mathbf{k}_t$, and value $\mathbf{v}_t$:

$$\mathbf{q}_t = \mathbf{W}_Q \mathbf{z}_t, \quad \mathbf{k}_t = \mathbf{W}_K \mathbf{z}_t, \quad \mathbf{v}_t = \mathbf{W}_V \mathbf{z}_t, \quad (14)$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d}$ are learned projection matrices. The attention score between time steps t and s is computed as the scaled dot product:

$$\alpha_{ts} = \frac{\mathbf{q}_t^T \mathbf{k}_s}{\sqrt{d}}. \quad (15)$$

These scores are normalized across all time steps using the softmax function:

$$\alpha_{t1}, \alpha_{t2}, \dots, \alpha_{tL} = \text{softmax} \left(\frac{\mathbf{q}_t^T \mathbf{K}^T}{\sqrt{d}} \right), \quad (16)$$

where $\mathbf{K} = [\mathbf{k}_1, \mathbf{k}_2, \dots, \mathbf{k}_L]^T \in \mathbb{R}^{L \times d}$. The output of the attention mechanism for time step t is the weighted sum:

$$\text{Attention}(\mathbf{q}_t, \mathbf{K}, \mathbf{V}) = \sum_{s=1}^L \alpha_{ts} \mathbf{v}_s, \quad (17)$$

where $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_L]^T \in \mathbb{R}^{L \times d}$.

Encoder block design. The Transformer encoder consists of N identical blocks stacked sequentially. Each encoder block contains two main components: a multi-head self-attention layer and a position-wise feed-forward network, both wrapped with residual connections and layer normalization. The multi-head attention component enhances the model's capacity by allowing multiple attention mechanisms to operate in parallel:

$$\text{MultiHead}(\mathbf{Z}) = \text{concat}(\mathbf{H}_1, \mathbf{H}_2, \dots, \mathbf{H}_h) \mathbf{W}_O, \quad (18)$$

where h is the number of attention heads, $\mathbf{H}_i \in \mathbb{R}^{L \times d_h}$ is the output of the i -th head, $d_h = d/h$ is the dimensionality of each head, and $\mathbf{W}_O \in \mathbb{R}^{d \times d}$ is a learned projection matrix. Each attention head $\mathbf{H}_i$ is computed independently:

$$\mathbf{H}_i = \text{Attention}(\mathbf{Z} \mathbf{W}_Q^{(i)}, \mathbf{Z} \mathbf{W}_K^{(i)}, \mathbf{Z} \mathbf{W}_V^{(i)}), \quad (19)$$

where $\mathbf{W}_Q^{(i)}, \mathbf{W}_K^{(i)}, \mathbf{W}_V^{(i)} \in \mathbb{R}^{d \times d_h}$ are the learned projection matrices specific to the i -th head.

Following multi-head attention, a position-wise *feedforward network* (FFN) is applied independently at each time step:

$$\text{FFN}(\mathbf{z}_t) = \mathbf{W}_2(\sigma(\mathbf{W}_1 \mathbf{z}_t + \mathbf{b}_1)) + \mathbf{b}_2, \quad (20)$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_{\text{ff}} \times d}$, $\mathbf{W}_2 \in \mathbb{R}^{d \times d_{\text{ff}}}$, $\mathbf{b}_1 \in \mathbb{R}^{d_{\text{ff}}}$, $\mathbf{b}_2 \in \mathbb{R}^d$, and d_{ff} is the hidden dimension of the FFN. The Transformer encoder block also includes *layer normalization* and residual connections:

$$\mathbf{Z}' = \text{LayerNorm}(\mathbf{Z} + \text{MultiHead}(\mathbf{Z})), \quad (21)$$

$$\mathbf{Z}'' = \text{LayerNorm}(\mathbf{Z}' + \text{FFN}(\mathbf{Z}')). \quad (22)$$

Forecasting with Transformers. For univariate time series forecasting with forecast horizon O , the final hidden representations $\mathbf{Z}_{\text{final}} = [\mathbf{z}_1'', \mathbf{z}_2'', \dots, \mathbf{z}_L'']^T \in \mathbb{R}^{L \times d}$ from the last encoder block capture both local and global temporal dependencies across the input sequence. To generate the forecast sequence, we use all time points by first flattening the final hidden representations and then applying a linear transformation:

$$\mathbf{z}_{\text{flat}} = \text{flatten}(\mathbf{Z}_{\text{final}}) \in \mathbb{R}^{Ld} \quad (23)$$

$$\hat{\mathbf{y}} = \mathbf{z}_{\text{flat}} \mathbf{W}_{\text{out}} + \mathbf{b}_{\text{out}}, \quad (24)$$

where $\mathbf{W}_{\text{out}} \in \mathbb{R}^{Ld \times O}$ and $\mathbf{b}_{\text{out}} \in \mathbb{R}^O$. This produces the forecast sequence $\hat{\mathbf{y}} = (\hat{y}_1, \dots, \hat{y}_O)^T \in \mathbb{R}^O$, which represents the predicted values for the next O time steps following the input sequence.

E. PatchTST

Patching. Time series forecasting attempts to model the relationships between data points from different time steps. A single time point does not equivocate to the level of semantic meaning of words within a sentence, similar to how individual letters do not carry high semantic information by themselves. In contrast, subseries-level patches enhances the model and captures higher semantic information, similar to modeling the relationship of words or subwords in a sentence [50]. However, most Transformer-based methods for time series do not rely on patching and patch-wise attention, and instead treat individual time points as token. Instead of using individual time points as tokens, *PatchTST* treats patches as tokens to be fed into

the standard Transformer architecture which greatly improves forecasting performance [31], [50].

F. DLinear

DLinear is linear neural network that borrows ideas from statistical time series modelling, and applies seasonal-trend decomposition to the input time series $\mathbf{x}$ before processing it through two distinct linear transformations [51]. While Transformer-based methods have been the most prevalent architecture for time series forecasting, they can often struggle with the temporal information loss due to their permutation-invariant self-attention mechanism. DLinear, however, has demonstrated superior performance over many state-of-the-art Transformer-based models across standardized benchmarks, questioning of the reliability of Transformer-based approaches and highlighting the effectiveness of more simpler deep learning approaches for time series forecasting. The method for DLinear is as follows. Given a time series $\mathbf{x}$, we first pad $\mathbf{x}$ with m values on both sides of the series using its first and last values, extending its length to $L + 2m$. We then define the moving average μ_t as:

$$\mu_t = \frac{1}{2m+1} \sum_{k=t-m}^{t+m} \mathbf{x}_k \quad (25)$$

for each $0 \leq t \leq L$, where $m \in \mathbb{N}$ is a hyperparameter. The local averaging window m smooths the input time series to form $(\mu_t)_{t=1}^L$, which in practice should be tuned to balance trend detection and noise reduction. To obtain the seasonal component we subtract the mean at each timestep, i.e. $\mathbf{x}_{\text{season}} = (\mathbf{x}_t - \mu_t)_{t=1}^L$, whereas the trend component is simply the moving average itself $\mathbf{x}_{\text{trend}} = (\mu_t)_{t=1}^L$. Viewing both $\mathbf{x}_{\text{season}}, \mathbf{x}_{\text{trend}} \in \mathbb{R}^{L \times M}$ as matrices with L time steps and M channels, we apply the transformation:

$$\hat{\mathbf{y}} = \mathbf{W}_1 \mathbf{x}_{\text{trend}} + \mathbf{W}_2 \mathbf{x}_{\text{season}} \quad (26)$$

where $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{O \times L}$ are learnable matrices and $\hat{\mathbf{y}} \in \mathbb{R}^{O \times M}$ is the forecast. Note that operations from Equations (25)–(26) are both linear, thus the transformation $\hat{\mathbf{y}} = f_\theta(\mathbf{x})$ is linear where f_θ is the network, making it a unique case among neural networks.

IV. EMFORECASTER: PROPOSED DL-EMPOWERED EMF FORECASTING ARCHITECTURE

Given a univariate time series representing EMF exposure over time, we detail the pre-processing and the architecture of the proposed EMForecaster, shown in Figure 1. EMForecaster incorporates RevIN, hierarchical patching, and a mixing operation to process temporal patterns at multiple scales. By decomposing the input signal into localized patches, the architecture can efficiently capture both fine-grained temporal dynamics and long-range dependencies, enabling effective modeling compared to traditional approaches that process the entire sequence at once. In following subsections, we describe each component of the EMForecaster following the sequence of the pipeline shown in Figure 1.

A. EMF Time Series Preprocessing

Consider a time series $\mathbf{x} = (x_t)_{t=1}^T$, where $x_t \in \mathbb{R}, \forall t$. To eliminate outliers, we set a carefully selected threshold $\delta > 0$, where x_t is replaced with a linear interpolation of its neighbors, x_{t-1} and x_{t+1} . We then partition $(x_t)_{t=1}^T$ into training, validation, and test datasets. We obtain the mean and standard deviation from the training set, and compute z -score to normalize the training, validation, and test datasets. Next, we apply the sliding window method with a stride of 1, to obtain window pairs of size L and O , representing the input and target, respectively. Thus, we obtain our dataset of n input-output pairs, $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ where $\mathbf{x}_i \in \mathbb{R}^L$ is the lookback window (input), and $\mathbf{y}_i \in \mathbb{R}^O$ is its continuation (target), for any sequence i .

B. Reversible Instance Normalization (RevIN)

We first apply RevIN to the input time series window $\mathbf{x}$ before feeding it into any temporal modules [55]. RevIN has been proven to be effective for reducing the impact of distribution shifts and non-stationarity in time series by using a symmetric structure to adjust and restore statistical information, leading to enhanced performance forecasting [55]. Given a window $\mathbf{x} \in \mathbb{R}^L$, we first normalize by:

$$\mathbf{x}^{(r)} = \text{RevIN}(\mathbf{x}) = \gamma \left(\frac{\mathbf{x} \ominus \mu}{\sigma} \right) \oplus \delta, \quad (27)$$

where $\gamma, \delta \in \mathbb{R}$ are learnable affine parameters, and $\mu = \frac{1}{L} \sum_{i=1}^L x_i$ and $\sigma = \sqrt{\frac{1}{L-1} \sum_{i=1}^L (x_i - \mu)^2}$ are the mean and standard deviation, respectively of $\mathbf{x} = (x_1, \dots, x_L)^T$. Here, $\oplus$ and $\ominus$ denote element-wise scalar-vector addition and subtraction, respectively, e.g., $\mathbf{v} \oplus \alpha = (v_1 + \alpha, \dots, v_k + \alpha)^T$ for any $\mathbf{v} \in \mathbb{R}^k$ and $\alpha \in \mathbb{R}$, and similarly for $\ominus$. Once the *candidate forecast*, $\hat{\mathbf{y}}_r$, is computed—the output of the spatiotemporal backbone defined in Section IV-D—we perform the inverse operation (RevIN^{-1}) of Equation (27) by adding back the mean and multiplying the standard deviation into the channel output:

$$\hat{\mathbf{y}} = \text{RevIN}^{-1}(\hat{\mathbf{y}}^{(r)}) = \sigma \left(\frac{\hat{\mathbf{y}}^{(r)} \ominus \delta}{\gamma} \right) \oplus \mu, \quad (28)$$

where δ and γ are the learnable affine parameters, and μ and σ are the mean and standard deviation of $\mathbf{x}$ from Equation (27). Note that: $\text{RevIN} \circ \text{RevIN}^{-1} = \text{RevIN}^{-1} \circ \text{RevIN} = \mathbb{I}$, where $\mathbb{I}$ is the identity function.

C. Patching and Embedding

Inspired from the PatchTST and Vision Transformer (ViT) methods [50], [56], we consider a patching mechanism along with a transformer encoder backbone. Given the demonstrated effectiveness of patching in time series forecasting compared to non-patching methods, we customize this mechanism for our architecture. After passing the input time series $\mathbf{x}$ to obtain $\mathbf{x}^{(r)} = \text{RevIN}(\mathbf{x})$, we apply a patching module to split up the sequence into equally-sized contiguous subsequences. Given a pre-determined patch dimension P , and patch stride S —similar

to a convolutional kernel—we obtain $N = \lfloor \frac{L-P}{S} \rfloor + 1$ patches in total:

$$\mathbf{x}^{(p)} = \text{Patcher}(\mathbf{x}^{(r)}), \quad (29)$$

where $\mathbf{x}^{(p)} \in \mathbb{R}^{N \times P}$. Before applying patching, we pad $\mathbf{x}^{(r)}$ with zeros at the end of the time series to ensure there are enough positions to extract all N complete patches of size P . After patching, we perform patch embedding defined by:

$$\mathbf{x}^{(d)} = \text{PatchEmbed}(\mathbf{x}^{(p)}) \quad (30)$$

$$= [\mathbf{W}_d(\mathbf{x}_1^{(p)}) \ \mathbf{W}_d(\mathbf{x}_2^{(p)}) \cdots \mathbf{W}_d(\mathbf{x}_N^{(p)})]^T \quad (31)$$

$$= \mathbf{x}^{(p)} \mathbf{W}_d^T, \quad (32)$$

where $\mathbf{x}_i^{(p)} \in \mathbb{R}^P$ is the i^{th} row of $\mathbf{x}^{(p)}$, $\mathbf{x}^{(d)} \in \mathbb{R}^{N \times D}$ is the resultant embedding, and $\mathbf{W}_d \in \mathbb{R}^{D \times P}$ is a learnable weight matrix which embeds each patch of size P to a (typically larger) dimension of size D , which we refer to as the patch embedding dimension. This is in contrast to several other embedding mechanisms found in other DL time series models, which may apply an embedding along the sequence dimension with $\mathbf{x}^{(d)} = \mathbf{W}_d \mathbf{x}^{(p)}$ with $\mathbf{W}_d \in \mathbb{R}^{D \times N}$. The patch-wise embedding approach learns local temporal patterns while maintaining parameter efficiency, as the embedding weights are shared across all patches and independent of the sequence length. Selecting an appropriate patch dimension P and patch embedding dimension D for the task is critical, as P controls the temporal receptive field of local pattern learning, while D determines the richness of the learned representations and the model’s capacity to capture complex temporal behavior.

D. Spatiotemporal Backbone (STB)

MLPMixer, a model proposed initially for computer vision, processes information through two distinct MLP operations: one that mixes information across tokens and another that mixes features within each token [57]. MLPMixer demonstrated that MLP-based architectures could achieve state-of-the-art performance in computer vision, even when compared to several modern Transformer variants [56], challenging the fact that convolutions or self-attention mechanisms are necessary for visual processing. TSMixer adapts the MLPMixer architecture for time series by utilizing two distinct MLP operations per block: one that mixes information across temporal locations and another that mixes features within each time step. For a layer l , the temporal MLP $f_{\theta^{(l)}} : \mathbb{R}^{N \times D} \rightarrow \mathbb{R}^{N \times D}$ is defined as:

$$f_{\theta^{(l)}}(\mathbf{u}) = \mathbf{W}_{\theta_2^{(l)}}(\sigma(\mathbf{W}_{\theta_1^{(l)}} \mathbf{u})), \quad (33)$$

for any $\mathbf{u} \in \mathbb{R}^{N \times D}$, where $\mathbf{W}_{\theta_1^{(l)}} \in \mathbb{R}^{D_h \times N}$ and $\mathbf{W}_{\theta_2^{(l)}} \in \mathbb{R}^{N \times D_h}$ are learnable matrices with parameters $\theta_1^{(l)}$ and $\theta_2^{(l)}$ respectively, σ is a nonlinearity (e.g., ReLU), and D_h is the hidden dimension of the MLP. Similarly, the patch MLP $g_{\phi^{(l)}} : \mathbb{R}^{N \times D} \rightarrow \mathbb{R}^{N \times D}$ is defined as:

$$g_{\phi^{(l)}}(\mathbf{u}) = \left(\mathbf{W}_{\phi_2^{(l)}}(\sigma(\mathbf{W}_{\phi_1^{(l)}} \mathbf{u}^T)) \right)^T, \quad (34)$$

for any $\mathbf{x} \in \mathbb{R}^{N \times D}$, where $\mathbf{W}_{\phi^{(l)}} \in \mathbb{R}^{D_h \times D}$ and $\mathbf{W}_{\phi_2^{(l)}} \in \mathbb{R}^{D \times D_h}$ are learnable matrices with parameters $\phi_1^{(l)}$ and $\phi_2^{(l)}$, respectively. The hidden dimension D_h and nonlinearity σ are the same as in the temporal MLP.

We provide the patch-embedded time series $\mathbf{x}^{(d)} \in \mathbb{R}^{N \times D}$ as an input to STB composed of K blocks:

$$\mathbf{x}^{(b)} = \text{STB}(\mathbf{x}^{(d)}) = \left(\text{Block}^{(K)} \circ \cdots \circ \text{Block}^{(1)} \right) (\mathbf{x}^{(d)}). \quad (35)$$

where the l^{th} block is modeled as:

$$\text{Block}^{(l)}(\mathbf{u}) = \mathbf{u} + f_{\theta^{(l)}}(\mathbf{u}) + g_{\phi^{(l)}}(\mathbf{u} + f_{\theta^{(l)}}(\mathbf{u})). \quad (36)$$

The above is the explicit form of composing $f_{\theta^{(l)}}$ and $g_{\phi^{(l)}}$ with two residual connections, which can be written equivalently as follows:

$$\mathbf{u}' = \mathbf{u} + f_{\theta^{(l)}}(\mathbf{u}), \quad (37)$$

$$\text{Block}^{(l)}(\mathbf{x}) = \mathbf{u}' + g_{\phi^{(l)}}(\mathbf{u}'). \quad (38)$$

where, $f_{\theta^{(l)}}$ “mixes” the temporal dimension N and $g_{\phi^{(l)}}$ “mixes” the embedding dimension D .

Applying the STB to patch-embedded data represents a novel fusion where the mixer operates on learned patch representations rather than raw temporal slices, enabling the capture of hierarchical patterns at multiple scales. The combination of patch embeddings and mixing operations allows the model to learn relationships between these higher-level temporal abstractions while maintaining the architectural simplicity of MLP-based mixing. We then pass the output of the STB through a nonlinear activation σ and a layer normalization module [58], i.e., $\mathbf{x}^{(a)} = \text{LayerNorm}(\sigma(\mathbf{x}^{(b)}))$, and flatten $\mathbf{x}^{(a)}$ to $\mathbf{x}^{(f)} = \text{Flatten}(\mathbf{x}^{(a)})$, where $\mathbf{x}^{(f)} \in \mathbb{R}^{(ND)}$, in which we obtain the candidate forecast by:

$$\hat{\mathbf{y}}^{(r)} = \text{Head}(\mathbf{x}^{(f)}) = \mathbf{W}_{\text{head}} \mathbf{x}^{(f)}, \quad (39)$$

$$\text{where } \mathbf{x}^{(f)} = \text{Flatten}(\text{LayerNorm}(\sigma(\mathbf{x}^{(b)}))), \quad (40)$$

where $\mathbf{W}_{\text{head}} \in \mathbb{R}^{O \times (ND)}$ is a learnable weight matrix, and $\hat{\mathbf{y}}^{(r)} \in \mathbb{R}^O$ is the candidate forecast, which is refined to obtain the final forecast by $\hat{\mathbf{y}} = \text{RevIN}^{-1}(\hat{\mathbf{y}}^{(r)})$ given by (28).

V. EMFORECASTER WITH CONFORMAL PREDICTION

CP is a powerful framework for uncertainty quantification that provides rigorous statistical guarantees [59]. The goal of CP is to provide a *prediction region*, Γ^α , for a given *significance level (or error rate)* $\alpha \in (0, 1)$, ensuring that the true outcome falls within this region with a probability of at least $(1 - \alpha)$. Most CP approaches rely on an *inductive approach*, where predictions are generated using a combination of an *underlying model* f and an additional *calibration set*—a technique referred to as *inductive conformal prediction (ICP)* [60]. Rather than providing point predictions, such as a single numerical value in regression, models calibrated with ICP generate a continuous interval in which ground truth is theoretically guaranteed to be contained with probability $1 - \alpha$.

In this section, we first describe CP in the context of regression with functions of the form $f : \mathbb{R}^L \rightarrow \mathbb{R}$. We then

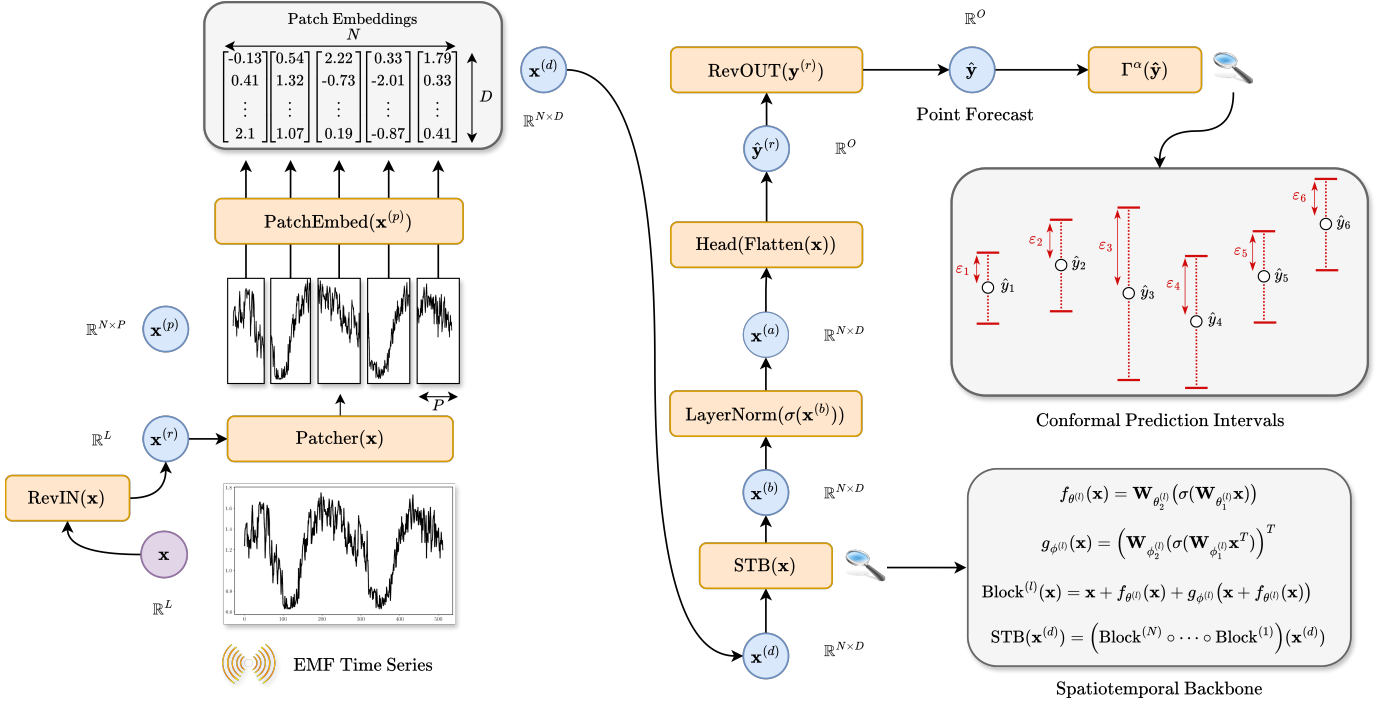


Fig. 1: The EMForecaster with Conformal Prediction: An EMF input time series $\mathbf{x} \in \mathbb{R}^L$ is processed by Reversible Instance Normalization (RevIN), before being patched to obtain $\mathbf{x}^{(p)} \in \mathbb{R}^{N \times P}$, where each patch of length P is embedded independently to obtain $\mathbf{x}^{(d)} \in \mathbb{R}^{N \times D}$. $\mathbf{x}^{(d)}$ is then fed through the Spatiotemporal Backbone (STB), which applies separate MLPs to the temporal domain and spatial domain of the patch-embedded input. The STB outputs a representation $\mathbf{x}^{(b)} \in \mathbb{R}^{N \times D}$, where we apply a nonlinear activation σ and layer normalization, before flattening the representation to a vector and applying a linear head to obtain the output $\hat{\mathbf{y}}^{(r)} \in \mathbb{R}^O$. The output $\hat{\mathbf{y}}^{(r)}$ is further refined by RevOUT (the inverse of RevIN) which adjusts $\hat{\mathbf{y}}^{(r)}$ to the appropriate distribution to obtain the point forecast $\hat{\mathbf{y}}$. The conformal predictor γ^α then outputs a conformal interval for each of the O forecasted points in $\hat{\mathbf{y}} \in \mathbb{R}^O$.

discuss the extension of CP to regression functions of the form $f : \mathbb{R}^L \rightarrow \mathbb{R}^O$ to predict multiple future time points simultaneously. Next, we discuss the evaluation metrics and the proposed metric referred to as Trade-off Score (TOS).

A. CP: Single Time-Step Forecast

While CP has been widely studied, its application to time series forecasting was limited due to the *exchangeability* condition which requires that any reordering of the dataset is equiprobable.

Definition 1 (Exchangeability): Given a dataset with n observations $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^n \subseteq \mathbb{R}^T \times \mathbb{R}$, we say that it is *exchangeable* if any of its $n!$ permutations are equiprobable. Note that independent and identically distributed (i.i.d.) observations satisfy exchangeability.

Given the exchangeability, the trustworthiness or reliability of a conformal predictor is described by the conformal coverage guarantee defined below.

Property 1 (Conformal coverage guarantee): Under the exchangeability assumption, any conformal predictor will return the prediction region $\Gamma^\alpha(\mathbf{x}^{(i)})$ such that the probability of error $y^{(l+1)} \notin \Gamma^\alpha(\mathbf{x}^{(l+1)})$ is not greater than α , that is:

$$\mathbb{P}[y^{(l+1)} \in \Gamma^\alpha(\mathbf{x}^{(l+1)}) \mid \mathcal{D}] \geq 1 - \alpha. \quad (41)$$

Note that the exchangeability assumptions and validity properties are *distribution-free*; that is, there are no required

distributional assumptions on the underlying data $\mathcal{D}$, and applies to any underlying predictive model so long as the exchangeability assumption is satisfied on $\mathcal{D}$.

Time series data, however, exhibits temporal dependencies, which inherently violating the exchangeability condition in most scenarios. As a result, directly applying CP to forecast intervals in time series data may not provide theoretical guarantees. Recent advancements have demonstrated that the exchangeability assumption can be relaxed through problem reframing this to the exchangeability of windows, allowing CP to be effectively applied to time series tasks [36].

The inductive variant of CP divides the dataset into two subsets: a proper training set and *calibration set*, such that $\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{cal}}$. Denote $n = |\mathcal{D}|$ and $m = |\mathcal{D}_{\text{cal}}|$. The training set is used to fit the underlying model f , while the calibration set is utilized to compute the *nonconformity scores* $\eta(\mathcal{D}, (\mathbf{x}^{(i)}, y^{(i)}))$ for all $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}_{\text{cal}}$, which measures the distribution errors on the underlying model. CP guarantees validity for any choice of nonconformity score, even random ones, but in regression settings, a commonly used score is defined as:

$$R_i = \eta(\mathcal{D}, (\mathbf{x}^{(i)}, y^{(i)})) = d(f(\mathbf{x}^{(i)} \mid \mathcal{D}), y^{(i)}), \quad (42)$$

where $d : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ is a distance metric. While the choice of d is flexible, it should reflect the problem's objectives with respect to the dataset and task. The residuals $\{R_i\}_{i=1}^m$, computed from the calibration set, form an empirical distribution. The *critical*

nonconformity score, $\hat{\epsilon}$, is selected as the $(1-\alpha)$ -quantile of this distribution. To account for finite-sample effects, a correction is applied by considering the $\lceil(m+1)(1-\alpha)\rceil$ -th smallest residual. For any new input $\mathbf{x}^{(n+1)}$, the prediction interval is defined by:

$$\Gamma^\alpha(\mathbf{x}^{(n+1)}) = [\hat{y}^{(n+1)} - \hat{\epsilon}, \hat{y}^{(n+1)} + \hat{\epsilon}], \quad (43)$$

where $\hat{y}^{(n+1)} = f(\mathbf{x}^{(n+1)})$ is the model prediction.

Theorem 1 (Vovk, Gammernan, and Saunders [61]): Suppose Γ^α is defined as in Equation (43), then Γ^α is a conformal predictor and Property (1) is satisfied.

B. CP: Multiple Time-Step Forecast

In univariate time series forecasting with horizon $O > 1$ and underlying model $f: \mathbb{R}^L \rightarrow \mathbb{R}^O$, our goal is to generate prediction intervals for each future time point. Specifically, for a model that produces predictions $f(\mathbf{x}^{(n+1)}) = \hat{\mathbf{y}} = (\hat{y}_1, \dots, \hat{y}_O)^T$, we must construct intervals as follows:

$$[\hat{y}_t^{(n+1)} - \hat{\epsilon}_t, \hat{y}_t^{(n+1)} + \hat{\epsilon}_t], \quad 1 \leq t \leq O.$$

For any new observation $(\mathbf{x}^{(n+1)}, \mathbf{y}^{(n+1)})$ where $\mathbf{y}^{(n+1)} = (y_1^{(n+1)}, \dots, y_O^{(n+1)})^T$, and any time step $1 \leq t \leq O$, these intervals should satisfy the following coverage guarantee similar to Property (1):

$$\mathbb{P}[y_t^{(n+1)} \in [\hat{y}_t^{(n+1)} - \hat{\epsilon}_t, \hat{y}_t^{(n+1)} + \hat{\epsilon}_t]] \geq 1 - \alpha \quad (44)$$

For any example $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$, the nonconformity score can be generalized to:

$$R_i = [|y_1^{(i)} - \hat{y}_1^{(i)}|, \dots, |y_O^{(i)} - \hat{y}_O^{(i)}|] \quad (45)$$

In [36], the authors indicate that the only restriction needed is that the underlying model must map to all O predictions simultaneously, rather than recursively, such as in Bayesian forecasting methods. Given this constraint, as all O predicted values are obtained from the same representation, we employ a Bonferroni correction when obtaining the critical nonconformity scores, which are obtained by considering the $\lceil(m+1)(1-\alpha/O)\rceil$ -th smallest residuals $\hat{\epsilon}_1, \dots, \hat{\epsilon}_O$ from each score distribution. By defining:

$$\Gamma_t^\alpha(\hat{y}_t) = [\hat{y}_t - \hat{\epsilon}_t, \hat{y}_t + \hat{\epsilon}_t] \quad (46)$$

for all $1 \leq t \leq O$, we obtain our valid conformal predictor.

Theorem 2: Let $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^n$ be the dataset of exchangeable time series windows, where $\mathbf{y}^{(i)} \in \mathbb{R}^O$ is the continuation of time series window $\mathbf{x}^{(i)} \in \mathbb{R}^L$, for every $1 \leq i \leq n$, generated from the same underlying distribution. Let $f: \mathbb{R}^L \rightarrow \mathbb{R}^O$ be a forecasting model which maps to all forecasted values simultaneously. Then for any $\alpha \in (0, 1)$, using the method in Equation (46) we have that:

$$\mathbb{P}[y_t \in \Gamma_t^\alpha(\hat{y}_t), \text{ for } 1 \leq t \leq O] \geq 1 - \alpha \quad (47)$$

C. Evaluation Metrics

For a set of predictions $\hat{\mathbf{y}}_1, \dots, \hat{\mathbf{y}}_n$ with corresponding target sequences $\mathbf{y}_1, \dots, \mathbf{y}_n$, the *independent coverage* (IC) and *joint*

coverage (JC) are defined by:

$$\text{IC} = \frac{1}{nO} \sum_{i=1}^n \sum_{t=1}^O \mathbb{1}[y_t^{(i)} \in \Gamma_t^\alpha(\hat{y}_t^{(i)})] \quad (48)$$

$$\text{JC} = \frac{1}{n} \sum_{i=1}^n \mathbb{1} \left\{ \bigwedge_{t=1}^O y_t^{(i)} \in \Gamma_t^\alpha(\hat{y}_t^{(i)}) \right\} \quad (49)$$

where $\mathbb{1}$ is the indicator function and $\bigwedge$ denotes the logical AND operation. IC measures the proportion of examples where the true value falls within the prediction interval at each horizon time t separately, while JC measures the proportion of examples where the true values fall within their respective prediction intervals across the full horizon time points simultaneously. While the empirical IC and JC typically increase with lower significance levels α , the *mean interval width* (MIW) of the prediction given below:

$$\text{MIW} = \frac{1}{O} \sum_{t=1}^O 2\hat{\epsilon}_t, \quad (50)$$

typically increases as well, which provides the average measure of uncertainty for model predictions. Therefore, a fundamental trade-off exists between coverage and interval width: models achieving higher JC often do so at the cost of wider prediction intervals.

The core challenge is therefore to minimize the MIW while maximizing JC and IC across a chosen significance level α .

D. The Trade-off Score (TOS)

Given the fundamental trade-off between achieving optimal coverage while maintaining efficient prediction intervals, we propose a unified scoring metric that synthesizes JC, IC, and MIW into a single measure bounded in $[0, 1]$ referred to as weighted-average coverage (WAC) defined below:

$$\text{WAC} = \frac{\beta \text{JC} + (1 - \beta) \text{IC}}{2} \quad (51)$$

where $\beta \in [0, 1]$ is a hyperparameter controlling the relative importance of joint coverage versus independent coverage. While the choice of β should reflect the practitioner's objectives, we recommend $\beta > \frac{1}{2}$ as joint coverage typically presents a more important and stricter criterion in CP. To incorporate the width of prediction regions, consider a collection of k conformal predictors derived from k models with corresponding MIWs $m_1, \dots, m_k \in \mathbb{R}^+$. We standardize these measures using z -score normalization $z_i = \frac{\mu - m_i}{\sigma}$, where μ and σ denote the sample mean and standard deviation of $\{m_1, \dots, m_k\}$, respectively. Finally, we introduce our proposed trade-off score (TOS), a metric that combines both coverage validity and prediction width. For each conformal predictor i , we define:

$$\text{TOS}(i) = \lambda \text{WAC} + (1 - \lambda) \frac{1}{1 + e^{z_i}} \quad (52)$$

where $\lambda \in [0, 1]$ controls the balance between WAC and normalized MIW. The reflected sigmoid transformation of the z -scores ensures the second term remains bounded in $[0, 1]$, resulting in $\text{TOS}(i) \in [0, 1]$ for all conformal predictors. This unified metric acknowledges that an ideal conformal predictor should not only

TABLE II: Long-term EMF datasets from Italy with reported the duration of measurement, year of recording, and reported sampling rate (Δt).

City	Location	Duration	Year	Δt
Rome	Dept. of Electronic Engineering (DEE 22')	190 days	2022	6 min
Rome	Dept. of Electronic Engineering (DEE 23')	491 days	2023	6 min
Rome	University Hospital (UH 22')	139 days	2022	6 min
Rome	University Hospital (UH 23')	592 days	2023	6 min
Turin	Polytechnic University of Turin (PT)	168 days	2020	6 min
Turin	Train Station (TS)	29 days	2022	6 min

achieve the desired coverage levels, but do so with reasonably tight prediction intervals. Without such a combined metric, we might favor methods that achieve perfect coverage at the cost of excessively wide intervals, or methods that produce deceptively narrow intervals but fail to maintain reliable coverage. The TOS enables practitioners to make these trade-offs explicit through the tuning parameters β and λ , while ensuring fair comparison across different methods through appropriate normalization. For meaningful comparisons, it is recommended to evaluate against a diverse set of methods that span different architectural families and methodological approaches. This helps ensure robust normalization and provides comprehensive performance context.

VI. EXPERIMENTAL SET-UP AND DATA ANALYSIS

In this section, we describe each of the selected datasets along with our experimental configuration for point forecasting and conformal forecasting. Our benchmarks include both *short-term* and *long-term* EMF exposure series. We define short-term to be any range within 24 hours, and long-term to exceed 24 hours. We analyze the characteristics of EMF exposure in several datasets through spectral decomposition, stationarity testing, and spatial correlation analysis.

A. EMF Datasets

1) *Long-term EMF Series (Italy)*: Adda *et al.* [62] presented long-term EMF measurements from four locations within Rome and Turin, Italy. Table II describes all measurement sites and their respective dataset specifications such as the year of recording, total length of each time series (i.e. duration of measurement), and reported sampling rate (Δt). Measurement sites include the Department of Electronic Engineering (DEE) at the University of Rome Tor Vergata, a University Hospital (UH) in Rome, Polytechnic University of Turin (PT), and the Porta Nuova Train Station (TS).

Although each location exhibits different characteristics in their EMF distribution, each site displays similar daily exposure patterns with cyclical behaviour, as shown in Figure 2. Note that Δt represents the 6 min sampling interval (pre-processed from the data provider), where the true sampling rate of the device is 3 sec. For each location within Rome, narrow-band monitoring was performed using an Anritsu MS27102A spectrum monitor (9 kHz – 6 GHz) connected to a Keysight N6850A omnidirectional antenna (20 MHz – 6 GHz) and the required EMF aggregated over all frequencies is then obtained by computing the root sum squared values of the EMF measured

at narrow-band frequencies. In Turin, a Narda 8059 wideband monitor with an electric field sensor (100 kHz – 7 GHz) was used. For more information regarding each dataset, we refer the readers to [62].

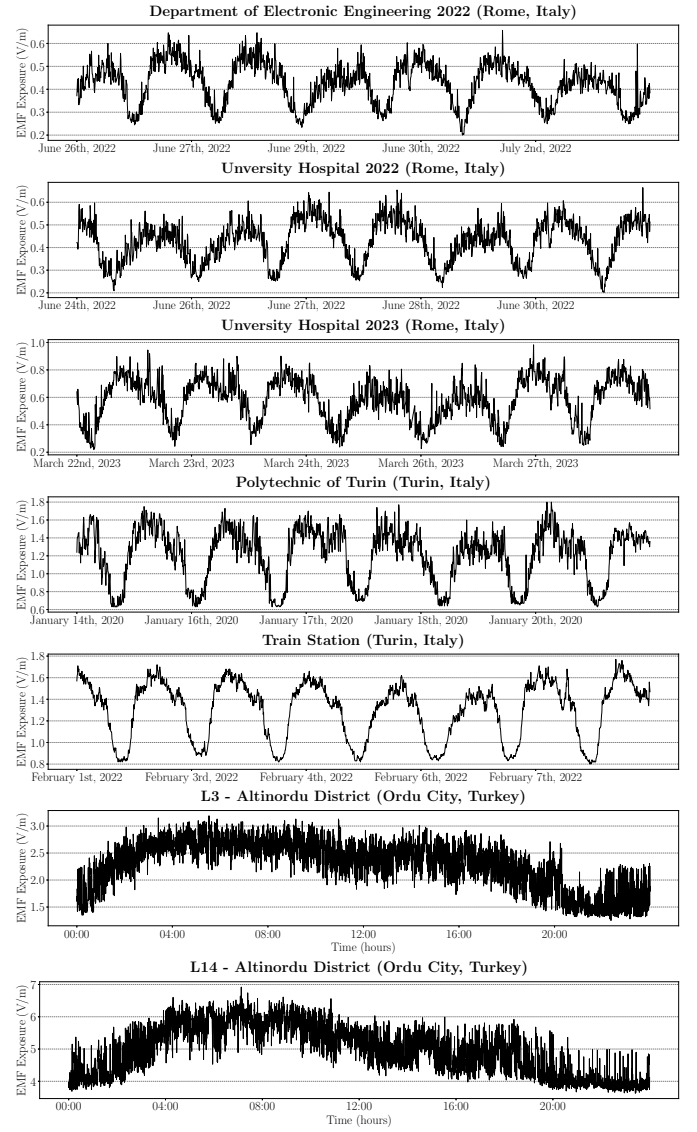


Fig. 2: Visualization of the EMF exposure (V/m) over time from five locations in the Italy (DEE 22', UH 22', UH 23', PT, TS) and two locations from the Turkey dataset (L3, L14).

2) *Short-term EMF Series (Turkey)*: In [32], Kurnaz *et al.* collected EMF exposure data in the Altinordu District of Ordu City, Turkey. Short-term (≤ 24 hr) broadband EMF measurements were conducted at 17 locations along the main streets of Altinordu, covering a frequency range of 100 kHz to 3 GHz, with a sampling rate of 15 seconds and duration of 24 hours. The authors used a PMM-8053 EMF meter equipped with an EP-330 electric field isotropic probe for the wideband measurements.

B. Data Analysis & Visualization

1) *Long-Term vs. Short-Term Forecasting*: The datasets from Italy and Turkey exhibit fundamental differences in their

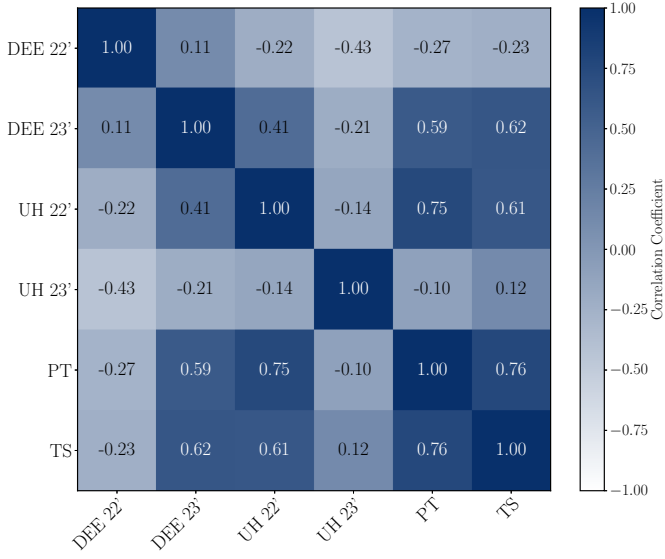


Fig. 3: Correlation heatmap of EMF exposure over a one week period for all 4 locations in Italy.

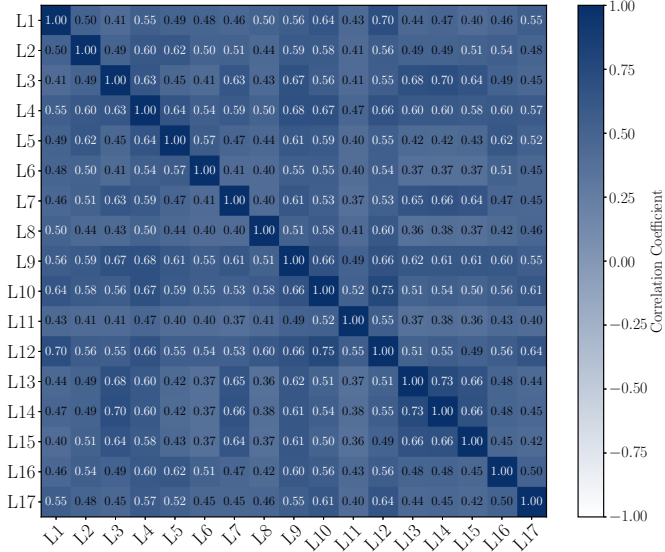
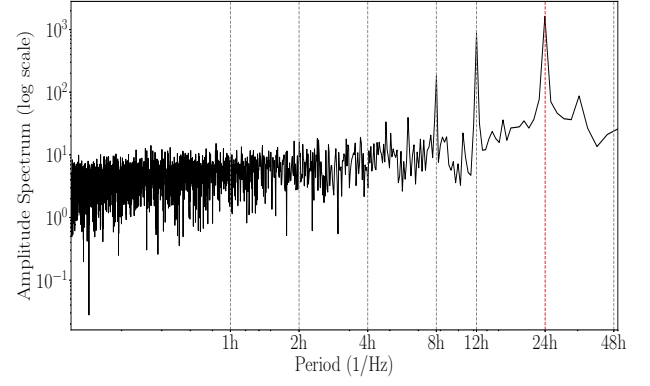
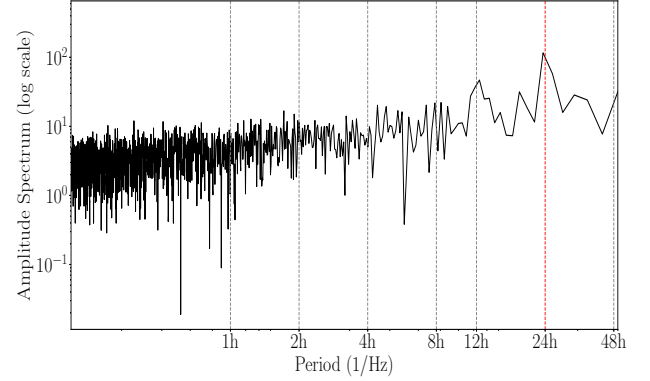


Fig. 4: Correlation heatmap of EMF over a 24 hour (total) period for all 17 locations within the dataset from Turkey.

exposure measurements that significantly impact our analysis and potential model forecasting capabilities. The datasets from Italy contain extensive recordings spanning up to 592 days across six different locations, showing clear and consistent daily cyclical patterns throughout the extended monitoring period. On the other hand, the dataset from Turkey covers 24-hours periods for each of the 17 locations, but comes with more frequent measurements at 15-sec intervals compared to Italy's 6 min reported sampling rate. These contrasting characteristics indicate that our forecasting approach for the datasets from Italy are able to effectively model both the prominent daily cycles and their deviations, whereas the datasets from Turkey presents us with a more challenging prediction scenario due to its shorter observation window and less apparent periodicity,



(a) FFT magnitude (DEE 22') with respect to the period (1/Hz).



(b) FFT magnitude (UH 22') with respect to the period (1/Hz).

Fig. 5: FFT magnitude plots of DEE 22' and UH 22'.

despite its higher temporal resolution. Our decision to maintain the 15-sec sampling rate for the datasets from Turkey rather than downsampling to match the 6 min intervals of datasets from Italy are driven by data quantity considerations. Averaging to 6 min intervals would have drastically reduced our already limited amount of data available from the 24-hours recordings, which would be insufficient for a fair comparison between our proposed DL model and baselines.

2) *Correlation of Measurement Sites:* Figures 3 and 4 illustrate the correlation between measurement sites for Italy and Turkey through heatmap visualizations, respectively. While datasets from Italy exhibit selective high correlations between specific measurement pairs, such as PT and TS, the majority of sites demonstrate weak inter-site correlations. In contrast, Turkey reveals consistent positive correlations across all measurement site pairs. This uniform correlation pattern in the Turkish data may be attributed to the environmental conditions during data collection and geographical proximity, in addition to the limited 24-hours sampling period, which could potentially inflate the observed correlation coefficients due to the shorter temporal window.

3) *Cyclical Pattern Analysis:* We observe that all time series within the dataset from Italy exhibit cyclical patterns, as shown in Figure 2. Using the FFT, we analyze these periodicities in the frequency domain, as illustrated in Figures 5a and 5b, which display the magnitude of Fourier coefficients as a function of period (1/Hz) rather than frequency (Hz). This representation, where a red vertical line marks the 24-hours

period, clearly reveals dominant peaks at both 12-hours and 24-hours duration, demonstrating strong daily periodic components in the EMF exposure signals across both DEE '22 and UH '22 environments. Similar observations were noted for other datasets from Italy.

4) *Stationarity of EMF Exposure:* We apply the ADF test to assess stationarity, as shown in Figure 6, which displays the ADF test statistics and statistical significance (p) for the datasets considering two scenarios: the original raw data and first-order differenced data (indicated by dark borders). The results reveal that all time series from Italy are stationary ($p < 0.05$), while many measurement sites in Turkey exhibit non-stationary behavior. This distinction can be attributed to the limited 24-hours recording period in Turkey, which captures only a single daily cycle and thus appears as a non-stationary time series. In contrast, the Italy measurements, as demonstrated in the previous section, display cyclical patterns centered around a mean value without persistent trends, confirming their stationary nature. Upon applying first-order differencing, we observe improved (more negative) ADF statistics across both datasets, with all time series from Italy becoming increasingly stationary with significantly reduced ADF statistics.¹

C. Training Configuration for DL Models

We partition $(x_t)_{t=1}^T$ into training, validation, and test datasets, with proportions of 70%, 10%, and 20%, respectively. For all considered DL models, including EMForecaster and baselines, we conduct training using the Adam optimizer across 100 epochs with a batch size of 2048, employing early stopping with a patience of 20 epochs on the validation Mean Square Error (MSE) [63]. We use the MSE as our objective function and main evaluation metric on the test set, given by:

$$\text{MSE}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{O} \sum_{i=1}^O (y_i - \hat{y}_i)^2, \quad (53)$$

for a prediction $\hat{\mathbf{y}} \in \mathbb{R}^O$ and label $\mathbf{y} \in \mathbb{R}^O$. We performed comprehensive hyperparameter tuning through grid search, optimizing separately for each model, dataset, and prediction length O . Model training was executed on a single NVIDIA RTX 6000 Ada Generation GPU with 48GB of memory. The reported performance metrics on the test set represent averages across five random seeds, using the model configuration that achieved the lowest validation MSE (for the specific dataset and prediction length).

VII. BASELINES, RESULTS, AND DISCUSSIONS

In this section, we describe the considered baselines and analyze the performance of the proposed EMForecaster for both *point forecasting*, which provides deterministic predictions, and *conformal forecasting*, which generates prediction intervals with

¹The stationarity of a time series, i.e., maintaining consistent statistical properties over time, suggests that windows drawn from it are more likely to be i.i.d., and thus exchangeable. This connection to exchangeability is particularly relevant for CP-based forecasting. This observation aligns with our empirical results, where we achieved state-of-the-art performance in both point forecasting and CP tasks on the stationary Italy datasets, while experiencing modest results on the non-stationary Turkey dataset.

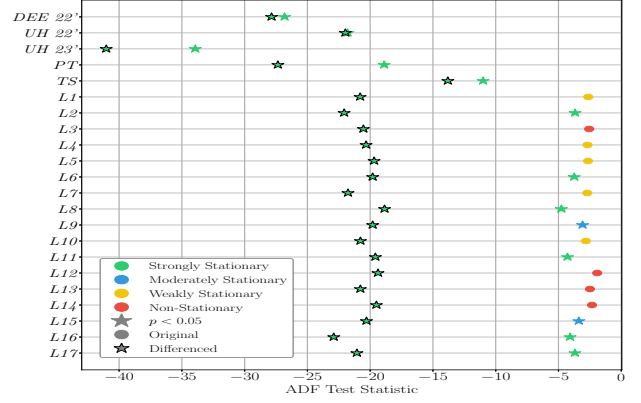


Fig. 6: ADF Test to characterize stationarity of the time series datasets from Turkey (17 locations) and Italy (4 locations). ADF test statistics and their statistical significance (p) are shown.

statistical guarantees. The code to reproduce all experiments is contained in github.com/xmootoo/emforecaster.

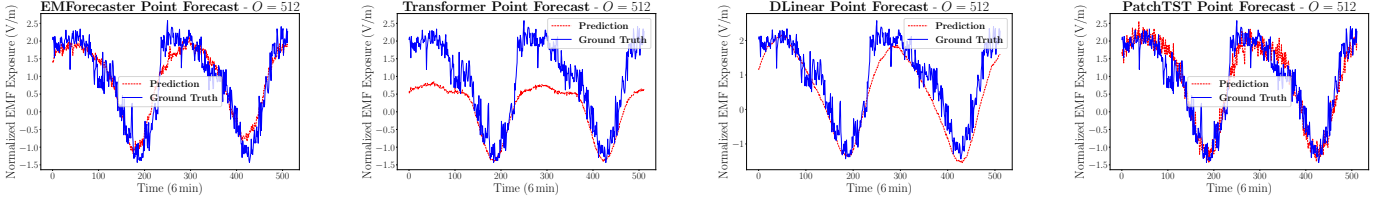
A. Baselines

The considered baselines are listed as follows:

- **MLP** processes the input time series through multiple fully-connected layers, capturing non-linear relationships between historical and future values through a direct mapping approach.
- **LSTM** uses gating mechanisms and recurrent connections to model temporal dependencies, enabling selective retention of relevant historical information across the series [54].
- **CNNs** utilize 1D convolutions sliding over the input sequence to extract local temporal patterns, effectively capturing hierarchical features through shared parameters.
- **Transformer** employs self-attention mechanisms to model relationships among all time steps simultaneously, allowing it to capture both long-range dependencies and local patterns. The parallel processing nature of attention makes it particularly effective at modeling complex temporal interactions [31]. In this work, we use the standard encoder-only architecture.
- **DLinear** is a modern neural network that has consistently outperformed many contemporary Transformer-based forecasting models. This is done by explicitly decomposing each input time-series into trend and seasonal components, processing them through separate linear projection layers. The trend extractor uses a moving-average filter and linear layer to capture long-term dependencies and the seasonal extractor passes the high-frequency residual through another linear layer to model periodic patterns—and then recombining these outputs to form an interpretable final forecast [51].
- **TSMixer** is an MLP-based time series model, inspired by MLP-mixer [57], [64]. TSMixer processes time series through parallel MLPs that separately handle temporal patterns (across time steps) and feature interactions (across measurements).
- **PatchTST** is a patched time-series Transformer that segments each series into fixed-length patches, feeds them into a standard Transformer encoder to capture local dynamics via patch embeddings and global dependencies via self-attention.

TABLE III: EMF point forecasting in terms of MSE with lookback window $L = 336$ and prediction lengths $O \in \{96, 192, 336, 512\}$.

		EMForecaster (2025)	DLinear (2023)	TSMixer (2023)	PatchTST (2023)	Transformer (2017)	LSTM (1997)	CNN (1988)	MLP (1986)	ARIMA (1970)
DEE 22'	96	0.3019	0.3488	0.3119	0.3710	0.7744	0.3409	0.3217	0.3217	1.2371
	192	0.3461	0.4055	0.3561	0.4151	1.1312	0.4033	0.3794	0.3794	1.4484
	336	0.3811	0.4791	0.3911	0.4664	1.1417	0.4426	0.4314	0.4314	1.4193
	512	0.4436	0.5420	0.4536	0.5336	1.3673	0.4854	0.4806	0.4806	1.4333
UH 22'	96	0.2404	0.2841	0.2685	0.2992	1.2874	0.3540	0.3599	0.4059	1.6604
	192	0.2726	0.3159	0.3026	0.3345	1.2291	0.4863	0.4185	0.4175	1.9352
	336	0.3148	0.3735	0.3580	0.3801	0.8394	0.5407	0.4733	0.4733	1.9328
	512	0.3535	0.4165	0.4045	0.3973	0.6675	0.6049	0.5463	0.5463	1.9635
UH 23'	96	0.2999	0.3251	0.3003	0.3490	1.0483	0.3180	0.3133	0.3133	1.0896
	192	0.3158	0.3417	0.3158	0.4017	0.7665	0.3453	0.3391	0.3393	1.1964
	336	0.3346	0.3716	0.3326	0.4083	0.6530	0.3517	0.3600	0.3600	1.2197
	512	0.3434	0.3945	0.3429	0.4422	0.7776	0.3670	0.3787	0.3787	1.2328
PT	96	0.1383	0.1423	0.1481	0.1460	0.3441	0.2027	0.1554	0.1554	0.2151
	192	0.1476	0.1483	0.1564	0.1509	0.3970	0.2536	0.1616	0.1616	0.2931
	336	0.1571	0.1592	0.1697	0.1632	0.3977	0.2546	0.1723	0.1723	0.3348
	512	0.1656	0.1670	0.1756	0.1675	0.3314	0.2571	0.1787	0.1787	0.3574
Turkey	96	0.4174	0.4054	0.4396	0.4822	0.5349	0.7714	0.7660	0.8014	0.8014
	192	0.4564	0.4422	0.4793	0.5348	0.5804	0.8221	0.7758	0.8143	0.8143
	336	0.4885	0.4818	0.5106	0.6050	0.6007	0.7917	0.7827	0.8296	0.8296
	512	0.5230	0.5101	0.5463	0.6612	0.6241	0.8018	0.7848	0.8356	0.8356

Fig. 7: Point forecasting plot of model predictions (red) versus ground truth labels (blue) with lookback $L = 336$ and prediction length $O = 512$ on the UH 22' dataset.

PatchTST has empirically outperformed leading variants across multiple benchmarks while gaining broad adoption in diverse domains [50].

- **ARIMA** is a traditional statistical learning model, which combines differencing, autoregression, and moving average components to model linear relationships in stationary time series data [26].

B. Results and Discussions: Point Forecasting

1) Comparative Analysis of EMForecaster with Baselines:

We evaluate the performance of point forecasting, i.e., forecasting individual future time points in the traditional sense on all datasets. Window sampling is performed on all 17 time series from Turkey, before taking the union of all windows to produce a full, multi-measurement site dataset. In contrast, we considered each dataset from Italy separately (DEE 22', UH 22', UH 23', PT) due to their distributional differences and weaker correlation compared to Turkey, as shown in Fig. 4.

For each dataset, we evaluate four prediction lengths $O \in \{96, 192, 336, 512\}$ independently, with a fixed lookback

window of $L = 336$. The reported sampling rate is 6 min for Italy, thus the forecast ranges from 9.6 hrs to 51.2 hrs with a lookback window of 33.6 hrs. For the dataset from Turkey, the lookback window is fixed at 84 min, with prediction lengths ranging from 24 min to 128 min. Therefore, although the units for L and O are identical, the forecasting lengths differ as the reported sampling rate is different in the two datasets.

2) *Point Forecasting Plots:* Figure 7 compares point forecasting results within the UH 22' dataset for EMForecaster alongside the Transformer, DLinear, and PatchTST with a lookback window $L = 336$ and prediction length $O = 512$. EMForecaster strikes a desirable middle ground: it captures the key daily EMF exposure cycle and its finer microstructure without capturing spurious noise. In contrast, DLinear produces an overly smooth trajectory that closely follows the dominant cycle, but systematically lags and ignores short-term fluctuations, leading to substantial phase errors at the beginning and end of the forecast window. PatchTST, while highly responsive to transient perturbations, overfits fine-scale noise and yields excessive variance around the true

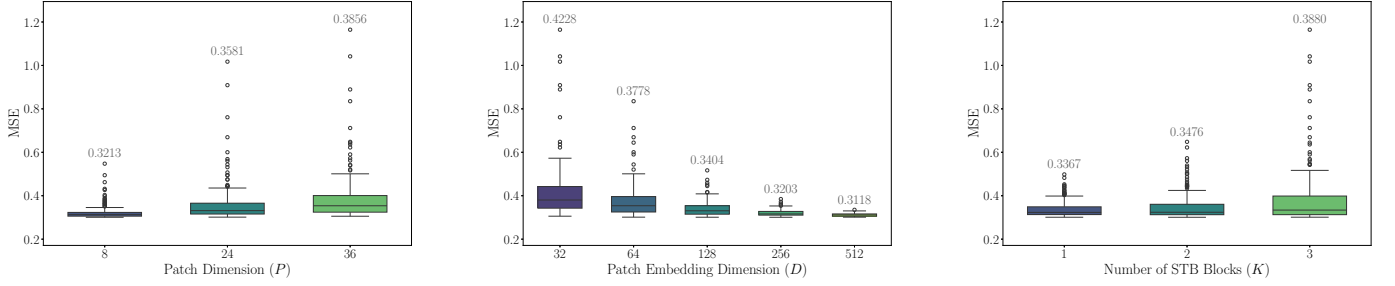


Fig. 8: MSE boxplots for the (a) patch dimension P , (b) patch embedding dimension D , and (c) the number of STB blocks K , over multiple hyperparameter configurations on the DEE 22' dataset with lookback window $L = 336$ and prediction length $O = 96$. Average MSE is reported at the top of each boxplot.

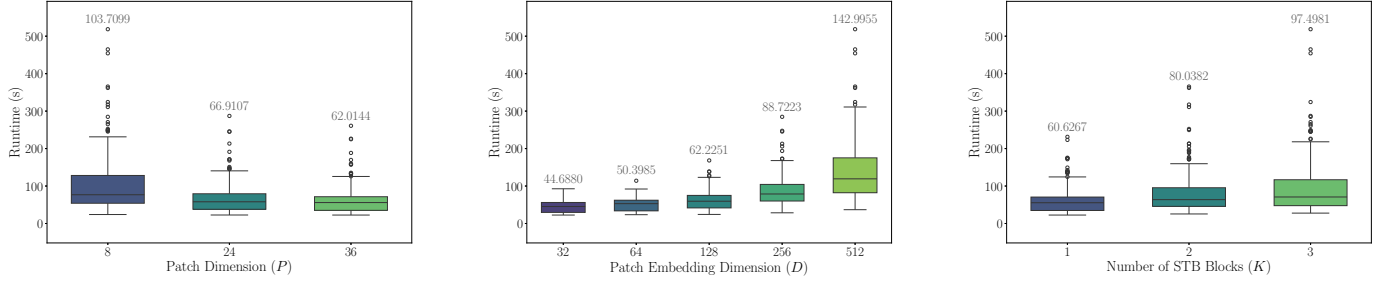


Fig. 9: Runtime boxplots for the patch dimension P , patch embedding dimension D , and the number of STB blocks K , over multiple hyperparameter configurations on the DEE 22' dataset with lookback window $L = 336$ and prediction length $O = 96$. Average Runtime is reported at the top of each boxplot.

signal. The standard Transformer underfits the cycle amplitude, consistently underestimating trough depths and failing to align with the highest peaks. Overall, EMForecaster maintains stable adherence to the underlying exposure pattern, while reproducing its microstructure—avoiding both the under- and overfitting exhibited by the other models.

Table III displays point forecasting performance across all baselines. EMForecaster achieves consistently superior MSE performance compared to modern DL baselines (PatchTST, TSMixer, DLinear), while traditional approaches (Transformer, LSTM, CNN, MLP, and ARIMA) show higher MSE on average. While DLinear and TSMixer achieve comparable performance on the PT and Turkish datasets, their performance degrades significantly on other datasets, particularly with increasing forecast horizons O . EMForecaster demonstrates robust performance across the Italian datasets, though it shows slightly lower performance on the Turkish dataset due to inherent non-stationarity and limited data availability. These results indicate that in stationary environments, EMForecaster exhibits strong generalization capabilities across diverse settings and forecast horizons.

3) *Analysis of Patch Dimension and Patch Embedding Dimension:* Figures 8 (MSE) and 9 (Runtime) together illustrate three key trade-offs in EMForecaster:

- **Patch size (P):** As P grows from $8 \rightarrow 36$, MSE increases from $0.32 \rightarrow 0.39$, while runtime drops from $104s \rightarrow 62s$ indicating that smaller patches capture fine-grained dynamics more accurately at the expense of longer computation.
- **Embedding dimension (D):** Increasing D from $32 \rightarrow 512$ reduces MSE from $0.42 \rightarrow 0.31$, but raises runtime

from $45s \rightarrow 143s$ showing that richer embeddings enhance representational power (lower error) but incur higher computing time.

- **Number of STB blocks (K):** Moving from $1 \rightarrow 3$ blocks drives MSE up from $0.34 \rightarrow 0.38$ and runtime from $61s \rightarrow 97s$ revealing that deeper stacks begin to overfit and slow down inference rather than improve generalization.

The inverse relationship between patch size (P) and performance, together with the strong gains from larger embedding dimensions (D), highlights a synergistic interaction. That is, using smaller patches concentrates the model's attention on fine-grained, localized signal fluctuations, and then projecting each patch into a high-dimensional embedding space amplifies its representational richness. In practice, a small P ensures that each embedding encodes tightly scoped temporal features, while a large D provides sufficient capacity for the patch MLP and time MLP to disentangle and recombine these features into expressive, global patterns. Consequently, configurations with $D \gg P$ maximize the information extracted per patch—validating our design choice to pair fine-resolution inputs with high-capacity embeddings.

C. Results and Discussions: Conformal Forecasting

We conduct our conformal forecasting experiments on the same datasets used for point forecasting for the prediction length $O = 96$ with lookback window $L = 336$ using the framework in section V-B on EMForecaster and the baselines.

1) Comparative Analysis of EMForecaster with CP:

Fig. 10 displays the results across all datasets for each

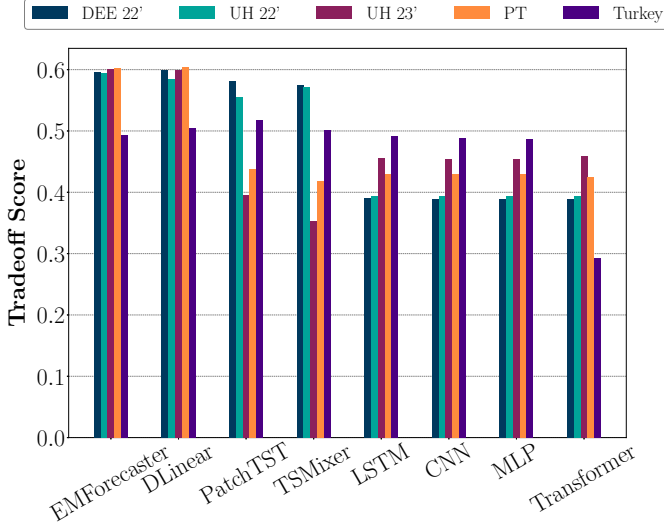


Fig. 10: Performance comparison of different forecasting models across datasets using the Tradeoff Score (TOS) metric using $\alpha = 0.01$, $\beta = \frac{2}{3}$, and $\lambda = \frac{1}{2}$.

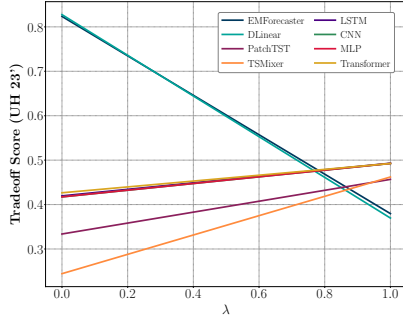


Fig. 11: Tradeoff Score (TOS) as a function of λ with $\alpha = 0.01$ and $\beta = \frac{2}{3}$ for UH 23' dataset.

TABLE IV: Performance evaluation of EMForecaster, examining the effect of sampling rates Δt across several datasets, with lookback window $L = 336$ and prediction length $O = 96$.

	Δt	MSE	JC	IC	MIW
DEE 22'	6min	0.3008	82.89	98.98	4.99
	30min	0.2104	84.96	99.49	6.93
UH 22'	6min	0.2467	73.70	97.95	4.11
	30min	0.2133	51.19	96.07	5.94
PT	6min	0.1372	62.41	98.58	2.73
	30min	0.1181	65.04	98.81	3.09

model with respect to the TOS. We see that EMForecaster achieves comparable performance to DLinear while achieving superior performance compared to all other baselines. For this experiment, we used $\beta = \frac{2}{3}$ and $\lambda = \frac{1}{2}$, as suggested in the discussion from section V-D. $\beta = \frac{2}{3}$ skews the WAC to the joint coverage as it is a harder metric to optimize, while still optimizing partly for the IC, while $\lambda = \frac{1}{2}$ provides an equal importance between coverage and prediction interval width

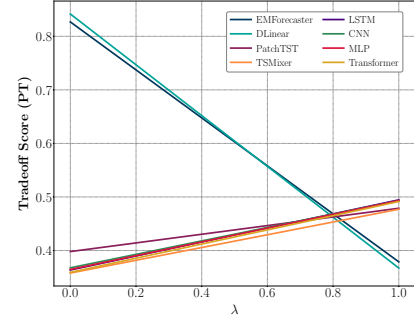


Fig. 12: Tradeoff Score (TOS) as a function of λ with $\alpha = 0.01$ and $\beta = \frac{2}{3}$ for PT dataset.

when computing the TOS.

Table V results reveal EMForecaster achieving strong JC while maintaining narrowest MIW, striking the best balance between JC and MIW among the three methods as seen in the results of the Tradeoff Score. Although DLinear runs a very close tradeoff between MIW and JC, EMForecaster slightly outperforms it on both the JC/MIW front and in point-forecast accuracy (lower MSE). The Transformer attains the highest JC, but does so with excessively wide MIW, rendering its output to be far less practical. This delicate balance highlights EMForecaster's ability to deliver reliable conformal intervals without sacrificing predictive precision.

2) *Choice of λ in Conformal Forecasting:* Fig. 11 and Fig. 12 examine the impact of λ on the TOS for the UH 23' and PT datasets respectively. Both EMForecaster and DLinear outperform all other baselines up to $\lambda \approx 0.8$. Beyond this point, the TOS becomes heavily influenced by coverage. Thus, while many baselines can achieve good coverage, it requires a very large λ to achieve an optimal TOS, indicating a poor balance between obtaining coverage and interval widths for most baselines. In contrast, EMForecaster and DLinear strike an optimal balance, avoiding excessively wide intervals while maintaining strong coverage.

D. Analysis of Differencing and Temporal Resolution

Table IV presents EMForecaster's performance across different temporal resolutions. We compare results at the original sampling rate of $\Delta t = 6$ min and a down-sampled rate of $\Delta t = 30$ min, where the latter is obtained by averaging five consecutive 6 min measurements. While independent coverage remains consistent across both sampling rates, we observe a notable decrease in joint coverage for the UH '22 dataset at $\Delta t = 30$ min. The coarser temporal resolution ($\Delta t = 30$ min) generally yields lower MSE due to the smoothing effect that reduces high-frequency noise. Interestingly, despite higher MSE values, the $\Delta t = 6$ min resolution achieves narrower mean prediction intervals, suggesting superior conformal forecasting performance. This demonstrates a trade-off between point forecast accuracy and the precision of uncertainty quantification across different temporal resolutions.

TABLE V: Conformal time series forecasting performance for EMForecaster and select baselines on the UH 23', PT and Turkey datasets (lookback $L = 336$, horizon $O = 96$).

	UH 23'				PT				Turkey			
	IC	JC	MIW	MSE	IC	JC	MIW	MSE	IC	JC	MIW	MSE
EMForecaster	98.11	64.83	4.645	0.3005	98.98	67.23	3.123	0.1202	99.20	61.92	4.978	0.4174
DLinear	98.22	61.77	4.620	0.3689	98.50	60.93	2.680	0.1420	99.31	63.47	5.026	0.4080
Transformer	99.97	97.74	6.343	1.0020	99.97	97.46	4.416	0.3539	100.0	100.0	8.477	0.7714

E. Computational Complexity

For DET 22' with lookback window $L = 336$, forecast horizon $O = 96$, and $\alpha = 0.01$, the combined point forecasting and conformal forecasting training and evaluation durations were 38.51 sec for EMForecaster, 27.19 sec for DLinear, 10.98 min for TSMixer, 10.87 min for PatchTST, 1.25 min for the Transformer, 1.04 min for the LSTM, 52.68 sec for the MLP, and 54.99 sec for the CNN. Our computational results demonstrate that EMForecaster is the second fastest model, with only DLinear achieving marginally faster execution times. Modern architectures such as TSMixer and PatchTST exhibit significantly longer training times, approximately 17 times that of EMForecaster, indicating less efficient parameter utilization. In addition, EMForecaster outperforms traditional DL architectures including LSTM and Transformer in terms of computational efficiency, executing 1.6-2x faster, which makes it particularly suitable for applications requiring rapid training or deployment under computational constraints.

VIII. CONCLUSION

In this work, we have introduced EMForecaster, a DL-empowered time series forecasting framework designed to predict EMF trends across diverse locations and varying forecast horizons beyond 50 hours. To improve the reliability of EMF predictions, we integrated a distribution-free uncertainty quantification framework, using CP, ensuring the ground truth falls within a specified prediction interval with a user-specified error rate. To enhance the evaluation of conformal forecasting, we introduce a novel TOS evaluation metric which balances prediction interval width and empirical coverage, allowing for objective comparisons across different conformal predictors. Our extensive experiments demonstrate that EMForecaster significantly outperforms existing DL models in both point forecasting and conformal forecasting. Future research directions include extending EMForecaster to Beyond 5G (B5G) data and expanding the forecast horizon to multiple years.

IX. ACKNOWLEDGMENTS

This research was supported by the Alliance International Catalyst and the Canada Graduate Scholarship—Master's (CGS-M) funded by the Natural Sciences and Engineering Research Council (NSERC) of Canada; the Vector Scholarship in Artificial Intelligence, provided through the Vector Institute, Canada; and the Ontario Graduate Scholarship (OGS) granted by the provincial government of Ontario, Canada.

This work was partially supported by European Union through Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of

the Future" (program "RESearch and innovation on future Telecommunications systems and networks (RESTART)" under Grant PE00000001).

We also sincerely appreciate Dr. Sara Adda from Agenzia Regionale Protezione Ambiente Piemonte (ARPA Piemonte), Ivrea, Italy, and Dr. Cetin Kurnaz from the Department of Electrical and Electronics Engineering, Ondokuz Mayıs University, Turkey, for generously providing valuable datasets that contributed to this study. While the majority of our results are based on datasets from the Department of Electronic Engineering at the University of Rome Tor Vergata and University Hospital in Rome, the datasets from Dr. Adda published in [62] and Dr. Kurnaz published in [32] were used for benchmarking purposes.

REFERENCES

- [1] M. A. Saeidi, H. Tabassum, and M. Alizadeh, "Molecular absorption-aware user assignment, spectrum, and power allocation in dense THz networks with multi-connectivity," *IEEE Trans. on Wireless Commun.*, pp. 1–1, 2024.
- [2] L. Chiaraviglio, A. Elzanaty, and M.-S. Alouini, "Health risks associated with 5G exposure: A view from the communications engineering perspective," *IEEE Open Jnl. of the Commun. Society*, vol. 2, pp. 2131–2179, 2021.
- [3] (ICNIRP), "Guidelines for limiting exposure to electromagnetic fields (100 kHz to 300 GHz)," *Health Physics*, 2020. [Online]. Available: <https://www.icnirp.org/cms/upload/publications/ICNIRPrfgdl2020.pdf>
- [4] R. Stam, "Comparison of international policies on electromagnetic fields:(power frequency and radiofrequency fields)," 2018.
- [5] M. Al Hajj, S. Wang, L. Thanh Tu, S. Azzi, and J. Wiert, "A statistical estimation of 5G massive MIMO networks' exposure using stochastic geometry in mmwave bands," *Applied Sciences*, vol. 10, no. 23, 2020. [Online]. Available: <https://www.mdpi.com/2076-3417/10/23/8753>
- [6] Q. Gontier, C. Wiame, J. Wiert, F. Horlin, C. Tsigros, C. Oestges, and P. D. Doncker, "On the impact of dynamic beamforming on EMF exposure and network coverage: A stochastic geometry perspective," 2024. [Online]. Available: <https://arxiv.org/abs/2405.01190>
- [7] N. A. Muhammad, N. Seman, N. I. A. Apandi, C. T. Han, Y. Li, and O. Elijah, "Stochastic geometry analysis of electromagnetic field exposure in coexisting sub-6 GHz and millimeter wave networks," *IEEE Access*, vol. 9, pp. 112 780–112 791, 2021.
- [8] Q. Gontier, C. Wiame, J. Wiert, F. Horlin, C. Tsigros, C. Oestges, and P. D. Doncker, "On the uplink and downlink EMF exposure and coverage in dense cellular networks: A stochastic geometry approach," 2024. [Online]. Available: <https://arxiv.org/abs/2312.08978>
- [9] Y. Qin, M. A. Kishk, A. Elzanaty, L. Chiaraviglio, and M.-S. Alouini, "Unveiling passive and active EMF exposure in large-scale cellular networks," *IEEE Open Jnl. of the Commun. Society*, vol. 5, pp. 2991–3006, 2024.
- [10] L. Chen, A. Elzanaty, M. A. Kishk, L. Chiaraviglio, and M.-S. Alouini, "Joint uplink and downlink EMF exposure: Performance analysis and design insights," *IEEE Trans. on Wireless Commun.*, vol. 22, no. 10, pp. 6474–6488, 2023.
- [11] Q. Gontier, L. Petrillo, F. Rottenberg, F. Horlin, J. Wiert, C. Oestges, and P. De Doncker, "A stochastic geometry approach to EMF exposure modeling," *IEEE Access*, vol. 9, pp. 91 777–91 787, 2021.

- [12] C. Wiame, S. Demey, L. Vandendorpe, P. De Doncker, and C. Oestges, "Joint data rate and EMF exposure analysis in manhattan environments: Stochastic geometry and ray tracing approaches," *IEEE Trans. on Veh. Tech.*, vol. 73, no. 1, pp. 894–908, 2024.
- [13] Y. A. Sambo, M. Al-Imari, F. Hélot, and M. A. Imran, "Electromagnetic emission-aware schedulers for the uplink of OFDM wireless communication systems," *IEEE Trans. on Veh. Tech.*, vol. 66, no. 2, pp. 1313–1323, 2016.
- [14] H. Jiang, L. You, A. Elzanaty, J. Wang, W. Wang, X. Gao, and M.-S. Alouini, "Rate-splitting multiple access for uplink massive MIMO with electromagnetic exposure constraints," *IEEE Jnl. on Sel. areas in Commun.*, vol. 41, no. 5, pp. 1383–1397, 2023.
- [15] Y. A. Sambo, F. Hélot, and M. A. Imran, "A user scheduling scheme for reducing electromagnetic emission in the uplink of mobile communication systems," in *IEEE Online Conf. on Green Commun. (OnlineGreenComm)*, 2014, pp. 1–5.
- [16] Y. A. Sambo, F. Hélot, and M. A. Imran, "Electromagnetic emission-aware scheduling for the uplink of multicell OFDM wireless systems," *IEEE Trans. on Veh. Tech.*, vol. 66, no. 9, pp. 8212–8222, 2017.
- [17] M. Matalatala, M. Deruyck, E. Tanghe, S. Goudos, L. Martens, and W. Joseph, "Joint optimization towards power consumption and electromagnetic exposure for massive MIMO 5G networks," in *IEEE 29th Annual Intl. Symposium on Personal, Indoor and Mobile Radio Commun. (PIMRC)*, 2018, pp. 1208–1214.
- [18] M. Wang, L. Lin, J. Chen, D. Jackson, W. Kainz, Y. Qi, and P. Jarmuszewski, "Evaluation and optimization of the specific absorption rate for multi-antenna systems," *IEEE Trans. on electromagnetic compatibility*, vol. 53, no. 3, pp. 628–637, 2011.
- [19] D. Ying, D. J. Love, and B. M. Hochwald, "Beamformer optimization with a constraint on user electromagnetic radiation exposure," in *47th Annual Conf. on Information Sciences and Sys.(CISS)*, 2013, pp. 1–6.
- [20] J. Penhoat, R. Agüero, F. Hélot, and M. Tesanovic, "Enabling low electromagnetic exposure multimedia sessions on an LTE network with an IP multimedia subsystem control plane," in *Mobile Networks and Management: 6th Intl. Conf., MONAMI*. Springer, 2015, pp. 207–216.
- [21] L. Diez, R. Aguero, and J. Penhoat, "Reducing the electromagnetic exposure over LTE networks by means of an adaptive retransmission scheme: A use case based on a video service," in *IEEE 81st Veh. Tech. Conf. (VTC Spring)*, 2015, pp. 1–5.
- [22] L. Chiaraviglio, A. S. Cacciapuoti, G. Di Martino, M. Fiore, M. Montesano, D. Trucchi, and N. B. Melazzi, "Planning 5G networks under EMF constraints: State of the art and vision," *IEEE Access*, vol. 6, pp. 51 021–51 037, 2018.
- [23] International Telecommunication Union, "Guidance on complying with limits for human exposure to electromagnetic fields," <https://www.itu.int/rec/T-REC-K.Sup14-201909-I>, 2019, accessed: 2025-01-02.
- [24] M. Matalatala, M. Deruyck, E. Tanghe, L. Martens, and W. Joseph, "Optimal low-power design of a multicell multiuser massive MIMO system at 3.7 GHz for 5G wireless networks," *Wireless Commun. and Mobile Computing*, vol. 2018, no. 1, p. 9796784, 2018.
- [25] J. A. Miller, M. Aldosari, F. Saeed, N. H. Barna, S. Rana, I. B. Arpinar, and N. Liu, "A survey of deep learning and foundation models for time series forecasting," 2024. [Online]. Available: <https://arxiv.org/abs/2401.13912>
- [26] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.
- [27] Y. Kiourekis, I. Givisis, T. Panagiotakopoulos, I. Tsilikas, A. Ploussi, E. Spyrou, and E. P. Efsthopoulos, "A comparative analysis of explainable artificial intelligence models for electric field strength prediction over eight european cities," *Sensors*, vol. 25, no. 1, p. 53, 2024.
- [28] M. R. Bakkan *et al.*, "Measurement and prediction of electromagnetic radiation exposure level in a university campus," *Tehnivski vjesnik*, vol. 29, no. 2, pp. 449–455, 2022.
- [29] Z. Pala, "Examining EMF time series using prediction algorithms with R," *IEEE Canadian Jnl. of Electrical and Computer Engineering*, vol. 44, no. 2, pp. 223–227, 2021.
- [30] C. Nguyen, A. A. Cheema, C. Kurnaz, A. Rahimian, C. Brennan, and T. Q. Duong, "Deep learning models for time-series forecasting of RF-EMF in wireless networks," *IEEE Open Jnl. of the Commun. Soc.*, 2024.
- [31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [32] C. Kurnaz and M. Mutlu, "Comprehensive radiofrequency electromagnetic field measurements and assessments: a city center example," *Environmental Monitoring and Assessment*, vol. 192, pp. 1–14, 2020.
- [33] Y. Liu, H. Wu, J. Wang, and M. Long, "Non-stationary transformers: Exploring the stationarity in time series forecasting," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9881–9893, 2022.
- [34] D. Santoro, T. Ciano, and M. Ferrara, "A comparison between machine and deep learning models on high stationarity data," *Scientific Reports*, vol. 14, no. 1, p. 19409, 2024.
- [35] S. E. Said and D. A. Dickey, "Testing for unit roots in autoregressive-moving average models of unknown order," *Biometrika*, vol. 71, no. 3, pp. 599–607, 1984.
- [36] K. Stankeviciute, A. M. Alaa, and M. van der Schaar, "Conformal time-series forecasting," *Advances in neural information processing systems*, vol. 34, pp. 6216–6228, 2021.
- [37] P. Bloomfield, *Fourier analysis of time series: an introduction*. John Wiley & Sons, 2004.
- [38] A. Y. Nikravesh, S. A. Ajila, C.-H. Lung, and W. Ding, "Mobile network traffic prediction using MLP, MLPWD, and SVM," in *IEEE Intl. congress on big data (BigData Congress)*, 2016, pp. 402–409.
- [39] A. Dalgakis, M. Louta, and G. T. Karetsos, "Traffic forecasting in cellular networks using the LSTM RNN," in *Proc. of the 22nd Pan-Hellenic Conf. on informatics*, 2018, pp. 28–33.
- [40] P. F. Perez, C. Fiandrino, M. Fiore, and J. Widmer, "Dissecting advanced time series forecasting models with aichronolens," in *IEEE INFOCOM 2024-IEEE Conf. on Computer Commun. Workshops (INFOCOM WK-SHPS)*. IEEE, 2024, pp. 01–02.
- [41] D. P. Isravel, S. Silas, J. W. Kathrine, E. B. Rajsingh, and J. Andrew, "Multivariate forecasting of network traffic in SDN based ubiquitous healthcare system," *IEEE Open Jnl. of the Communications Society*, 2024.
- [42] S. P. Sone, J. J. Lehtomäki, and Z. Khan, "Wireless traffic usage forecasting using real enterprise network data: Analysis and methods," *IEEE Open Jnl. of the Commun. Society*, vol. 1, pp. 777–797, 2020.
- [43] Y. Zhang, Y. Wu, A. Liu, X. Xia, T. Pan, and X. Liu, "Deep learning-based channel prediction for leo satellite massive mimo communication system," *IEEE Wireless Commun. Letters*, vol. 10, no. 8, pp. 1835–1839, 2021.
- [44] A. Hameed, J. Violos, A. Leivadreas, N. Santi, R. Grünblatt, and N. Mitton, "Toward QoS prediction based on temporal transformers for IoT applications," *IEEE Trans. on Network and Service Management*, vol. 19, no. 4, pp. 4010–4027, 2022.
- [45] A. G. Colpitts and B. R. Petersen, "Short-term multivariate KPI forecasting in rural fixed wireless LTE networks," *IEEE Networking Letters*, vol. 5, no. 1, pp. 11–15, 2023.
- [46] M. Di Mauro, G. Galatro, F. Postiglione, W. Song, and A. Liotta, "Multivariate time series characterization and forecasting of VoIP traffic in real mobile networks," *IEEE Trans. on Network and Service Management*, vol. 21, no. 1, pp. 851–865, 2023.
- [47] E. Ak and B. Canberk, "Forecasting quality of service for next-generation data-driven wifi6 campus networks," *IEEE Trans. on Network and Service Management*, vol. 18, no. 4, pp. 4744–4755, 2021.
- [48] T. Zhou, H. Zhang, B. Ai, C. Xue, and L. Liu, "Deep-learning-based spatial-temporal channel prediction for smart high-speed railway communication networks," *IEEE Trans. on Wireless Commun.*, vol. 21, no. 7, pp. 5333–5345, 2022.
- [49] M. A. Habib, P. E. I. Rivera, Y. Ozcan, M. Elsayed, M. Bavand, R. Gaigalas, and M. Erol-Kantarci, "Transformer-based wireless traffic prediction and network optimization in O-RAN," in *2024 IEEE Intl. Conf. on Commun. Workshops (ICC Workshops)*. IEEE, 2024, pp. 1–6.
- [50] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," in *Intl. Conf. on Learning Representations*, 2023.
- [51] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are transformers effective for time series forecasting?" in *Proc. of the AAAI Conf. on Artificial Intelligence*, 2023.
- [52] K. Fukushima, "Neocognitron: A hierarchical neural network capable of visual pattern recognition," *Neural networks*, vol. 1, no. 2, pp. 119–130, 1988.
- [53] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [54] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [55] 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 *Intl. Conf. on Learning Representations*, 2021.
- [56] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16

words: Transformers for image recognition at scale,” in *Intl. Conf. on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=YicbFdNTTy>

- [57] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit *et al.*, “MLP-mixer: An all-MLP architecture for vision,” *Advances in neural information processing systems*, vol. 34, pp. 24 261–24 272, 2021.
- [58] J. Lei Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” *ArXiv e-prints*, pp. arXiv-1607, 2016.
- [59] A. N. Angelopoulos and S. Bates, “A gentle introduction to conformal prediction and distribution-free uncertainty quantification,” *arXiv preprint arXiv:2107.07511*, 2021.
- [60] H. Papadopoulos, “Inductive conformal prediction: Theory and application to neural networks,” in *Tools in artificial intelligence*. Citeseer, 2008.
- [61] V. Vovk, A. Gammernan, and C. Saunders, “Machine-learning applications of algorithmic randomness,” *PMLR*, 1999.
- [62] S. Adda, L. Chiaraviglio, D. Franci, C. Lodovisi, N. Pasquino, S. Pavoncello, C. Pedroli, and R. Pelosini, “High noon for mobile networks: Short-time EMF measurements to capture daily exposure,” *IEEE Trans. on Instrumentation and Measurement*, vol. 72, pp. 1–10, 2023.
- [63] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [64] S.-A. Chen, C.-L. Li, N. Yoder, S. O. Arik, and T. Pfister, “TSMixer: An all-MLP architecture for time series forecasting,” *arXiv preprint arXiv:2303.06053*, 2023.



Xavier Mootoo (Student Member, IEEE) received the Specialized Honours B.Sc. degree in Mathematics from York University, Toronto, Canada, in 2023, and is currently an M.Sc. candidate in Computer Science at York, graduating in 2025. His research centres on modern deep learning methods for time series analysis, including variable-length sequence modeling and conformal prediction, spanning applications in neuroscience, healthcare, finance, and wireless networks. His collaborations include the Krembil Research Institute at Toronto Western Hospital and

the Temerty Centre for AI Research in Education in Medicine (T-CAIREM) within the Temerty Faculty of Medicine at the University of Toronto, yielding state-of-the-art seizure localization techniques presented at the NeurIPS 2024 Workshop on Time Series in the Age of Large Models.



Hina Tabassum (Senior Member, IEEE) (M’12-SM’18) received the Ph.D. degree from the King Abdullah University of Science and Technology (KAUST). She is currently an Associate Professor with the Lassonde School of Engineering, York University, Canada, where she joined as an Assistant Professor, in 2018. She is also appointed as a Visiting Faculty at University of Toronto in 2024 and the York Research Chair of 5G/6G-enabled mobility and sensing applications in 2023, for five years. She has been selected as the IEEE ComSoc Distinguished

Lecturer for the term 2025-2026. She is listed in the Stanford’s list of the World’s Top Two-Percent Researchers in 2021-2024. She received the Lassonde Innovation Early-Career Researcher Award in 2023 and the N2Women: Rising Stars in Computer Networking and Communications in 2022. She has co-authored over 120 refereed articles in well-reputed IEEE journals, magazines, and conferences. She was the Founding Chair of the Special Interest Group on THz communications in IEEE Communications Society (ComSoc)-Radio Communications Committee (RCC). She served as an Associate Editor for IEEE COMML (2019–2023), IEEE OJCOMS (2019–2023), and IEEE TGCN (2020–2023). Currently, she is also serving as an Area Editor for IEEE OJCOMS and an Associate Editor for IEEE Transactions on Communications, IEEE Transactions on Mobile Computing, IEEE Transactions on Wireless Communications, and IEEE Communications Surveys and Tutorials. Her current research interests include multi-band 6G wireless communications and sensing networks, connected and autonomous systems, and AI-enabled network mobility and resource management solutions.



Luca Chiaraviglio (Senior Member, IEEE) (M’09-SM’16) received the Ph.D. degree in Telecommunication and Electronics Engineering from the Politecnico di Torino, Italy. He is currently an Associate Professor with the University of Rome “Tor Vergata”, Italy. He has co-authored more than 180 articles published in international journals, books, and national/international conferences. His current research topics cover 5G and B5G networks, optimization applied to telecommunication networks, electromagnetic fields, and health risks assessment of 5G and B5G communications. Prof. Chiaraviglio received the Best Paper Award at the IEEE Vehicular Technology Conference (VTC)-Spring 2020, the IEEE VTC-Spring 2016, and the Conference on Innovation in Clouds, Internet and Networks (ICIN) 2018, all of them appearing as the first author. Some of his papers are listed as the Best Readings on Green Communications by the IEEE. Moreover, he has been recognized as an author in the top 2% world scientists according to the 2021-2024 updates of the science-wide author databases of standardized citation indicators. He is currently the principal investigator for the RESTART programme at the University of Rome Tor Vergata, managing a total budget of more than 9.8 MEUR.