

Large-Scale Forecasting of Electric Vehicle Charging Demand Using Global Time Series Modeling

Tijmen van Etten¹^a, Victoria Degeler¹^b and Ding Luo²^c

¹*University of Amsterdam, Science Park, Amsterdam, the Netherlands*

²*Shell Information Technology International B.V., Amsterdam, the Netherlands*
tijmenettenvan@gmail.com, v.o.degeler@uva.nl, ding.luo@shell.com

Keywords: Time Series, Deep Learning, Multiple Time Series, E-Mobility, Electric Vehicles, Intelligent Transportation, Forecasting, Energy Demand

Abstract: Electric Vehicle (EV) charging demand forecasting holds paramount significance in advancing sustainable transportation systems, particularly as electric vehicle adoption surges globally. Accurate predictions of charging demand are instrumental for optimizing charging infrastructure, energy management, and grid stability. By forecasting the demand for charging, stakeholders can effectively distribute resources, plan ahead for peak usage times, and lay out blueprints for the growth of infrastructure. Furthermore, precise forecasting enables the seamless integration of renewable energy sources into transportation, promoting a cleaner and greener future. In this work, challenges in EV charging demand forecasting are addressed, and an innovative framework tailored for large-scale prediction is proposed. The methodology involves generating individual forecasts for multiple charging stations, enabling a comprehensive evaluation of forecasting models across diverse contexts. The potential of global deep learning models to enhance prediction accuracy by capturing shared patterns across time series is explored. These models exhibit remarkable generalization capabilities, proving effective even in forecasting demand at previously unobserved charging stations. The contributions of this research encompass both methodologies and insights, enriching the realm of accurate EV charging demand forecasting. This work bears significance in fostering the integration of electric vehicles into transportation systems, aligning with the trajectory towards sustainable energy solutions.

1 INTRODUCTION


Electric Vehicle (EV) charging demand forecasting is crucial for ensuring sustainable transportation systems. As EV adoption increases, accurate predictions become critical for optimizing infrastructure, managing energy efficiently, and maintaining grid stability. This enables resource allocation, integration of renewable energy sources, and cleaner transportation, ultimately facilitating the widespread adoption of EVs and a greener transportation ecosystem.


Existing research primarily focuses on predicting single demand curves, which may not generalize well to diverse geographical areas, time periods, and demographic segments. To address this limitation, a framework for large-scale EV charging demand forecasting is presented. This framework involves gen-


erating forecasts for individual charging stations and collectively evaluating their accuracy.

This framework offers a more nuanced and realistic evaluation of forecasting models by considering multiple individual time series instead of a single aggregated one. It aims to reduce bias and potential inaccuracies associated with focusing on a single time series, thereby advancing EV charging demand forecasting for practical applications. Additionally, the potential of deep learning-based models to discern patterns across diverse time series is explored, addressing the complexity of forecasting at new charging station locations.

This research addresses two key questions: how can global deep learning models enhance demand forecasting by extracting and sharing patterns across time series? How well do these global models generalize to predict charging demand at new, unseen charging station locations? These questions aim to overcome current limitations in the literature and contribute to the development of robust, scalable, and ef-

^a  <https://orcid.org/0009-0006-5659-3046>

^b  <https://orcid.org/0000-0001-7054-3770>

^c  <https://orcid.org/0000-0003-2661-0926>

fective demand forecasting models for the EV charging industry.

This work contributes to the field of EV charging forecasting by proposing a novel framework for large-scale demand forecasting across multiple charging station locations. A robust solution for accurately estimating forecasting model performance using historical data is offered. Additionally, the applicability of global deep learning in EV charging demand forecasting is demonstrated, showcasing superior performance while reducing operational complexity. The research validates the effectiveness of global deep learning models in predicting charging demand for previously unseen stations, emphasizing their capacity to generalize and adapt to new situations.

2 RELATED WORK

2.1 Electric Vehicle Charging Load Forecasting

In recent years, there has been a growing interest in forecasting EV charging demand, leading to numerous studies in the field. However, the literature on this topic is characterized by a significant level of fragmentation and divergence (Amara-Ouali et al., 2021). This division can be attributed to the wide variety of datasets used and the diverse range of forecasting applications considered, each with its own corresponding geographical and temporal scales. As a result, various forecasting techniques have been studied, with many different techniques appropriate depending on the task at hand. In this section, the different approaches and models used in EV forecasting literature will be described, as well as the different geographical and temporal scales on which forecasts are generally made.

2.1.1 Approaches and Models

Originally due to the lack of real-world EV charging data, studies have been conducted using simulations methods. However, these approaches often use proxies for electricity consumption such as road traffic data (Su et al., 2017; Andrenacci et al., 2016; Xydas et al., 2013) or individual EV charging profiles (Gerossier et al., 2019; Yan et al., 2020; Huber et al., 2020). These methods are therefore often relied upon upon strong assumptions, such as the replacement of the current car fleet by electric vehicles (Kim and Kim, 2021).

More recently, charging demand data has become increasingly available due to the development of new

charging infrastructure (Amara-Ouali et al., 2021), facilitating the potential to leverage statistical and machine learning methods for supervised learning. These methods can be broadly classified into three categories: statistical, classical machine learning, and deep learning methods.

Simple statistical methods have been proven to provide highly competitive results for charging load forecasting. The autoregressive integrated moving average (ARIMA) (Kim and Kim, 2021; Ren et al., 2022) model, for example, is commonly implemented and used as a basis for more advanced models due to its ease-of-use and interpretability. Extensions of the ARIMA model include the Seasonal Autoregressive Integrated Moving Average (SARIMA) (Louie, 2017)

Machine learning techniques, such as support vector machines (SVM) (Xydas et al., 2013), random forests (RF) (Buzna et al., 2019), gradient boosting regression tree (GBRT) (Buzna et al., 2019), and eXtreme Gradient Boosting (XGBoost) (Yi et al., 2022), have been effective in load forecasting. On the other hand, the rise of deep learning, especially models like artificial neural networks (ANNs), convolutional neural networks (CNNs) (Zhu et al., 2019), and recurrent neural networks (RNNs) (Zhu et al., 2019; Yi et al., 2022; Moon et al., 2018), has enabled sophisticated charging demand forecasting due to their prowess in handling sequential data and learning non-linear relationships. Notably, the Long Short-Term Memory (LSTM) model, and its variations, have emerged as solutions to handle datasets with long dependencies (Yi et al., 2022; Koohfar et al., 2023; Eddine and Shen, 2022). One standout hybrid approach is the SARIMA-LSTM model (Ren et al., 2022), which combines linear and non-linear components for more precise forecasting.

While RNN-based architectures such as LSTM's have shown effective on a wide variety of tasks, more recently attention-based mechanisms have been shown to outperform these approaches on tasks such as Natural Language Processing (NLP). Koohfar et al. (2023) attempts to fill the gap in the EV forecasting literature by applying Transformer-based models to the task of forecasting charging demand. Their research shows that these models can offer superior performance compared to both statistical and other deep learning-based approaches.

A different type of network that is becoming increasingly popular is the Graph-Neural Network. As mentioned previously, one study successfully modeled the dependencies between charging stations (Hüttel et al., 2021). In the paper they propose their novel Temporal Graph Convolution Model, demonstrating outperformance of their model on both short

and long-term forecasting compared to other forecasting methods.

2.1.2 Geographical Scales

As previously mentioned, studies use a wide range of different geographical resolutions on which energy load predictions are made depending on the type of application.

Studies have attempted to predict charging load for small-scale power consumption types such as several EVs (Gerossier et al., 2019) or a single road (Wang et al., 2018), while other studies attempt to forecast the charging load for an individual charging station (Kim and Kim, 2021; Koohfar et al., 2023; Eddine and Shen, 2022; Ren et al., 2022).

Yi et al. (2022) uses clustering to group a number of charging stations together into regions to forecast the aggregated charging demand for a number of regions in the U.S. state of Utah and the city of Los Angeles. This approach significantly reduces the variance of the aggregated load curve, leading to more stable results. However, this aggregated approach sacrifices the granularity of forecasting demand for individual charging stations.

Furthermore, as stated in Amara-Ouali et al. (2021), the intricate spatial and temporal dependencies between charging stations is one of the difficulties in predicting the demand for EV charging. While the forecasting of charging load for a charging station has been relatively well studied, few account for the dependencies between individual sites. Instead, the charging demand for each electric vehicle charging station (EVCS) is more commonly aggregated and forecasted as a single time series (Louie, 2017). Hüttel et al. (2021) proposed a solution that combines the charging data of multiple charging stations in Palo Alto using a spatio-temporal graph-based modeling approach to account for these spatial-temporal correlations between individual stations. Other studies have been conducted that attempt to predict the charging demand of a city (Kim and Kim, 2021; Yi et al., 2022) or province (Buzna et al., 2019). Lastly, country-level forecasting attempts have been made to predict the total load demand for a total of 1,916 charging stations in Korea (Kim and Kim, 2021) and similarly for the country of China (Eddine and Shen, 2022).

2.1.3 Forecasting Horizons

Besides different geographical scales, approaches in the existing literature use a wide range of forecasting horizons. Forecasting horizons can generally be divided into three different categories: short-

medium, and long-term forecasting. Short-term forecasts, ranging from minutes (Hu et al., 2021), up to hours (Ren et al., 2022), can aid energy suppliers to plan and optimize their short-term energy production to efficiently satisfy energy

demand. Medium-term forecasts, ranging from days up to several weeks (Ren et al., 2022; Eddine and Shen, 2022; Hüttel et al., 2021), can be used by EVCS operators to make informed decisions about capacity planning, load management, and maintenance planning. Lastly, long-term forecast horizons can further be used to assist long-term investment planning and allocation of resources for charging infrastructures.

2.1.4 Limitations in Charging Load Forecasting Literature

While research in the field of EV charging demand forecasting has been extensive across various geographical scales, significant limitation arises in how the accuracy of forecasting methods is both assessed and reported, often focusing on just a singular time series. This narrow focus on individual time series restricts the scalability and applicability of proposed models in diverse settings. Although some papers, such as Kim and Kim (2021), have explored multiple geographical scales, it is important to note that each corresponding scale is typically still investigated solely based on a single aggregated demand curve.

To address this limitation and improve the forecasting models, the need to move beyond analyzing just one time series is emphasized. In this work, by studying multiple individual time series from a specific geographical area, a more complete evaluation is aimed for. This method is intended to enhance the accuracy and versatility of the models.

Another prevalent limitation found in the literature is the lack of proper model validation in the evaluation of the presented forecasting models. A common approach is to make a forecast with a given horizon for only a single window in a held-out test set. This approach, as seen in studies such as Koohfar et al. (2023) and Hüttel et al. (2021), often involves evaluating the model's performance using only the first consecutive data points. Restricting the evaluation to a single forecast window introduces a notable bias in the reporting of results, which can potentially lead to an overestimation or underestimation of the model's actual performance.

To address this limitation, a rolling-window historical forecasting approach is incorporated. With this approach, models undergo a more realistic evaluation in terms of forecasting performance. This methodology allows for testing the models on a diverse and representative set of historical data windows, offering

a comprehensive assessment of their predictive capabilities and generalization across different time periods.

2.2 Global Time Series Modeling

Training machine learning models on multiple related time series data –also known as “cross-learning”– has gained substantial attention in recent years due to its potential to enhance forecasting accuracy and capture interdependencies among variables. This approach involves leveraging data from multiple time series that measure the same phenomenon or variable, aiming to exploit the relationships and patterns among them. The motivation behind training on multiple related time series stems from the recognition that individual time series often exhibit inherent dependencies that can be better understood and harnessed when considered together.

Notable methodologies in this space include DeepAR (Salinas et al., 2020), which melds the capabilities of LSTM-based recurrent networks and Bayesian probabilistic models. This framework has shown promise in predicting intricate time-dependent patterns based on a multitude of related time series.

Similarly, N-BEATS (Oreshkin et al., 2020) offers a unique deep learning architecture designed for univariate time series point forecasting. Its capability to outperform the previous M4 competition winner, ES-RNN, underlines its efficacy in capturing complex temporal sequences through a combination of deep stacks and residual connections.

N-HiTS (Neural Hierarchical Time Series), proposed in Challu et al. (2022), extends N-BEATS’ capability for long-horizon forecasting with hierarchical interpolation & multi-rate data sampling techniques. It shows a 20% improvement in accuracy over the state-of-the-art while reducing computational time by 50 times, highlighting its efficiency.

Yet another noteworthy approach is the probabilistic forecasting framework based on Temporal Convolutional Neural Networks (TCNs) (Chen et al., 2020). It leverages stacked dilated causal convolutional networks to grasp complex temporal dependencies, significantly improving forecast accuracy even when historical data is sparse.

3 METHODOLOGY

To enable effective large-scale forecasting, several key aspects of the time series forecasting lifecycle need to be reconsidered, specifically focusing on the model, training setup, and evaluation methodologies.

This section delves into these facets, providing an in-depth understanding of their enhancements.

3.1 Task Definition

Given a time series dataset $\mathcal{D} = \{\mathcal{T}_1, \dots, \mathcal{T}_N\}$ comprising N time series, where each time series \mathcal{T}_n is represented by a sequence of values: $(y_1^n, y_2^n, \dots, y_L^n)$ of length L , the objective is to construct a forecasting model F that accurately predicts future values for each time series sequence.

Mathematically, the forecasting model F can be represented as a function mapping historical observations within each time series \mathcal{T}_n up to time step t to predicted values for subsequent time steps $t + 1$ to H points in advance. Therefore, for each time series \mathcal{T}_n , the forecasting process can be formalized as follows:

$$\hat{y}_{t+1}^n, \hat{y}_{t+2}^n, \dots, \hat{y}_{t+H}^n = F(y_1^n, y_2^n, \dots, y_t^n) \quad (1)$$

Where:

- F is the forecasting model.
- $\hat{y}_{t+1}^n, \hat{y}_{t+2}^n, \dots, \hat{y}_L^n$ are the predicted values for time steps $t + 1$ to L for time series \mathcal{T}_n .
- $y_1^n, y_2^n, \dots, y_t^n$ represent the historical observations up to time step t for time series \mathcal{T}_n .

3.2 Model

This methodology utilizes the N-HiTS model for time series forecasting. Similar to the N-BEATS architecture, the N-HiTS architecture follows a hierarchical structure composed of stacks, each consisting of blocks. With each block, the model learns to accurately approximate a specific segment of the input signal while delegating the remaining portions to be approximated by subsequent blocks in the model through a process called doubly residual stacking. For a more detailed description of the model architecture, the reader is referred to the original N-BEATS and N-HiTS papers.

3.3 Splitting Multiple Time Series Data

For our research objectives, two different types of splits of time series data are utilized.

Temporal Partitioning Unlike conventional machine learning procedures, which often assume the supervised data follows an independent and identically distributed (i.i.d.) pattern, time series data has distinct characteristics. Given the inherent sequential nature of time series data, careful consideration is required

when partitioning it into training and testing sets, necessitating specialized methods.

The foremost and widely employed approach that is utilized in this work, involves splitting time series across time intervals. Let L denote the total number of data points in each time series \mathcal{Y} , with l_{train} indicating the allocated training duration. Consequently, $L - l_{\text{train}}$ data points are left for testing. This partitioning method can be formulated as follows:

$$\mathcal{T}_n^{\text{train}} = (y_1^n, \dots, y_{l_{\text{train}}}^n) \quad (2)$$

$$\mathcal{T}_n^{\text{test}} = (y_{l_{\text{train}}+1}^n, \dots, y_L^n) \quad (3)$$

Series-Wise Partitioning The next strategy employed involves partitioning the dataset across individual series. Considering the same dataset \mathcal{D} comprising N time series. For effective implementation of series-wise partitioning, a subset of these time series is designated for training, while the remaining ones are allocated for testing. This allocation can be expressed mathematically as:

$$\mathcal{D}^{\text{train}} = \{\mathcal{T}_1, \dots, \mathcal{T}_{n_{\text{train}}}\} \quad (4)$$

$$\mathcal{D}^{\text{test}} = \{\mathcal{T}_{n_{\text{train}}+1}, \dots, \mathcal{T}_N\} \quad (5)$$

Unlike temporal partitioning, which maintains chronological order, series-wise partitioning can be achieved through random shuffling as it does not depend on the order of the time series.

3.4 Training on Multiple Time Series

Before feeding the data into the N-HITS model, the data is processed into consecutive pairs of input and output sub-series, each of which has a length defined by the combined input chunk length and output chunk length. The input sequences within these pairs serve as the neural network's inputs, while the output sequences are used to calculate the training loss. The processing of this dataset can be defined using the following mathematical notation:

$$\mathcal{D}_{\text{input}}^{\text{train}} = \{(\mathcal{X}_1^{\text{train}}, \mathcal{Y}_1^{\text{train}}), \dots, (\mathcal{X}_m^{\text{train}}, \mathcal{Y}_m^{\text{train}})\} \quad (6)$$

Each element in the set, $(\mathcal{X}_i^{\text{train}}, \mathcal{Y}_i^{\text{train}})$, represents a consecutive sub-series of a time series \mathcal{T} with an input chunk length of $|\mathcal{X}|$ and an output chunk length of $|\mathcal{Y}|$. Here, m is the total number of consecutive input/output pairs that could be generated from all time series $\mathcal{T} \in \mathcal{D}_{\text{train}}$. By combining these pairs from different datasets, the model can effectively learn from multiple time series, capturing diverse patterns and

dependencies in the data, which enhances its forecasting capabilities and generalization across various contexts.

3.5 Historical Forecasting

The evaluation of time series forecasting models is a critical aspect in assessing their predictive accuracy. Traditionally, studies in EV charging demand forecasting (Kooohfar et al., 2023; Hüttel et al., 2021; Kim and Kim, 2021) have predominantly utilized the multi-step forecasting approach, a common practice involves setting aside a fixed test set with a length corresponding to the forecast horizon H , following the training data. Making predictions for the evaluation of the forecasting model on the held-out test set can be mathematically formulated as:

$$\hat{\mathcal{Y}}_n^{\text{test}} = (\hat{y}_{t+1}^n, \hat{y}_{t+2}^n, \dots, \hat{y}_{t+H}^n) = F(y_1^n, y_2^n, \dots, y_t^n) \quad (7)$$

Where the set $(y_1^n, y_2^n, \dots, y_t^n)$ represents the input data up to time t , $\hat{\mathcal{Y}}_n^{\text{test}} = (\hat{y}_{t+1}^n, \hat{y}_{t+2}^n, \dots, \hat{y}_{t+H}^n)$ are the predictions of the test values and F is the forecasting model.

However, this conventional approach faces two main challenges. Firstly, the dedicated test set is restricted in size, limiting the generalizability of the evaluation and potentially leading to overfitting. Secondly, all but the last forecasted point within this approach fall within a timeframe less than H steps ahead, failing to assess the model's performance at the full forecast horizon and biasing the evaluation towards shorter-term predictions.

To overcome the constraints imposed by the evaluation of a restricted number of data points, historical forecasting, commonly known as backtesting, is utilized. This systematic methodology provides an in-depth approach to assess the effectiveness of time series forecasting models. Unlike conventional single-window forecasting, historical forecasting entails predicting past values within a time series through the utilization of a sliding window technique. By adopting this approach, a more comprehensive evaluation of the model's performance can be achieved, shedding light on its reliability and stability across diverse time segments within the time series.

Secondly, to more realistically capture the accuracy of forecasting with a specific forecast horizon, a distinct approach is proposed that offers enhanced insights into the quality of predictions over extended time periods. Forecasting H days ahead in time is advocated. This shift from the conventional multi-step forecasting technique to the proposed approach provides a more comprehensive understanding of the

model’s capability to predict specific days well in advance.

To clarify the approach, the existing mathematical representation can be extended to account for a rolling window of forecasting where each forecast is made H days in advance. Let’s designate S as the sliding window such that for every data point y_t^n in the test set $\mathcal{Y}_n^{\text{test}}$, a H -day ahead forecast is made using all the preceding data points.

$$S_t = (y_1^n, y_2^n, \dots, y_{t-H}^n) \quad (8)$$

With window S_t , the H -day ahead forecast \hat{y}_{t+H}^n , can be generated.

$$\hat{y}_{t+H}^n = F(S_t) \quad (9)$$

Over the test set, the collection of forecasts would be:

$$\mathcal{T}_n^{\text{test}} = (\hat{y}_{H+1}^n, \hat{y}_{H+2}^n, \dots, \hat{y}_L^n) \quad (10)$$

Where T is the end of the test set.

4 EXPERIMENTAL SETUP

We explore the impact of global training using the N-HiTS Model and make a comparative analysis against using local training and various well-established models frequently used in the EV charging demand forecasting literature. We run the experiments for four datasets separately, to evaluate the applicability on a wide range of datasets each with a variable number of time series and different characteristics.

4.1 Datasets

In this study, four datasets, are employed each providing insights into EV charging station energy consumption in kilowatt-hours (kWh) across different geographical locations. Three of these datasets are publicly available.

London The proprietary London dataset contains charging session data of 113 charging stations from in and around the greater London area in the United Kingdom. With 476,639 records over the time span of January 2020 to October 2022 it is the most comprehensive dataset out of four. It contains information related to the session data regarding the location information, driver information, charging fee, power type and session duration.

Palo Alto The Palo Alto dataset (of Palo Alto, 2021) is a public dataset containing data from electric vehicle charging activities across 22 locations in Palo Alto, California. This dataset provides the longest range of EV charging data, spanning from 2011 to 2020. It also includes a range of attributes for each charging session, such as station information, location information (including address and postal code), charging time, gasoline and greenhouse-gas savings, power type, charging fee, as well as driver information.

Perth Another publicly available dataset is the Perth dataset (Council, 2019), encompassing session data originating from Perth & Kinross, a region located in Scotland. Covering the period from January 2016 to December 2019, this dataset encompasses data from 22 distinct charging station locations. Its attributes include location information, charging time, and connector type.

Boulder The last public dataset is the Boulder dataset (of Boulder, 2020), which contains EV charging session data from 32 distinct charging station locations from the city of Boulder in the U.S. state of Colorado. Encompassing data from January 2018 to March 2023, this dataset enables the observation of EV charging trends over a significant timeframe. Similar to the Palo Alto dataset, it includes essential attributes like station information, location information, charging time, power type, and metadata on gasoline and greenhouse-gas savings.

In this study, one specific attribute is utilized: energy consumption (measured in kWh) per transaction. We made this choice because it can be easily calculated across all datasets, making our study relevant and adaptable to various scenarios.

4.2 Pre-processing

To process the raw session data for our purposes, the session data is aggregated to represent the total daily energy delivered in kWh per charging station. As an additional preprocessing step, negative values in the data detected as outliers are removed. To balance the trade-off between the number of charging days and the number of time series, charging station time series that have at least 690 days of data are selected, specifically focusing on the most recent 690 days of data points. The series that contain over 10% missing values are discarded, and the missing daily values for the remaining time series are filled using linear interpolation. This preprocessing approach results in a total of 34 time series for the London dataset, 8 time

series for the Palo Alto dataset, 5 time series for the Boulder dataset and 8 time series for the Perth dataset. We found that the relatively small number of remaining Time Series in the Palo Alto, Perth and Boulder datasets can largely be attributed to a large number of missing values. A description of the processed datasets can be found in Table 1.

Dataset	Number of EV Stations	Total Data points	Start Date	End Date
London	34	23460	29 July 2020	31 Oct. 2022
Palo Alto	8	5520	22 Nov. 2018	31 Dec. 2020
Perth	8	5520	11 Oct. 2017	8 Dec. 2019
Boulder	5	3450	11 May 2021	31 Mar. 2023

Table 1: Overview of Processed EV Charging Session Datasets. This table delineates each dataset’s number of EV station time series, total data points, and the date range of the selected aggregated time series.

For training, a temporal split on each time series is employed, allocating 600 consecutive days for training and 90 days for testing. The training data is further split into a 70/30 ratio for training and validation, respectively, balancing data usage for effective model learning and robust validation while considering dataset limitations.

4.3 Trend Analysis

To get a general impression of the trend and demand pattern over time in each dataset, we visualized the aggregated the daily average delivered energy demand across time series. The time series plot for each dataset curve is depicted in Figure 1. The London demand curve exhibits a clear upward trajectory, indicating a steadily increasing demand for EV charging. Also, we notice that the magnitude of EV charging demand varies significantly between the minimum and maximum daily delivered energy over the dataset’s span. This indicates large variability in demand curves over individual time series. The aggregated time series for Palo Alto demonstrates a slight upward trend up until early 2020, followed by a steep decline. This decline in early 2020 are attributed to the effects of the COVID-19 pandemic. After this period, a gradual resurgence in demand can be observed. Notably, the range of demand scale remains relatively narrow throughout. The Perth dataset showcases a trajectory that bears resemblance to London’s, albeit with a bit more fluctuation due to the smaller number of time series. The Boulder dataset also exhibits considerable fluctuation. Due to its limited number of time series data, it appears especially susceptible to noise in individual time series, leading to this pronounced variability.

Hyperparameter	Values
Number of Stacks	[1, 2, 4, 8, 16]
Number of Blocks	[1, 2, 3, 4, 5]
Number of Layers	[1, 2, 3, 4, 5]
Layer Widths	[32, 64, 128, 256, 512]
Dropout Rate	[0, 0.1, 0.2]

Table 2: Explored hyperparameter values for the N-HiTS model during tuning.

4.4 Training & Hyperparameter Tuning

We configure an N-HiTS model with an input chunk length of 30 and an output chunk length of 7, to balance optimization for different forecasting horizons. Furthermore, the model is configured to encode the weekdays as covariates using a one-hot encoding.

Training is done using the train split of each time series. Before feeding the data into the model, Min-Max scaling is applied to each series independently, ensuring that the unique characteristics of each series are preserved and allowing for a fair comparison between the local and global training processes. During training Early Stopping is employed with a patience of 5 and minimum delta of 0.05. The N-HiTS model is trained using the Adam optimizer with an MSE loss function. The batch size is configured at 32, and an initial learning rate of 1e-3 is set.

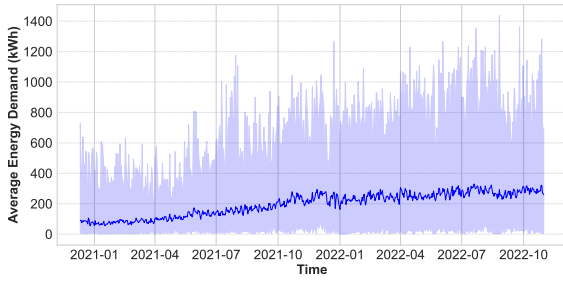
We employ a comprehensive exploration of hyperparameters to fine-tune the N-HiTS-architecture for optimal performance. The hyperparameter space includes choices for the number of blocks, stacks, layer widths, and dropout rates. The details of the hyperparameter ranges are presented in Table 2.

Using Ray Tune for Hyperparameter Tuning (Liaw et al., 2018), performance is measured based on MSE on the validation set. We incorporate an Asynchronous Successive Halving Algorithm (ASHA) scheduler, executing 20 iterations. The configuration that yields the minimum validation loss is deemed optimal for predictions on the test set. The detailed hyperparameters for the N-HiTS models are presented in Table 2.

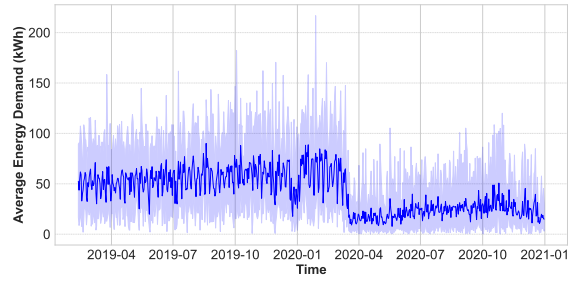
4.5 Benchmarks

We conduct a comprehensive comparison of the N-HiTS_{global} model with the following four distinct approaches:

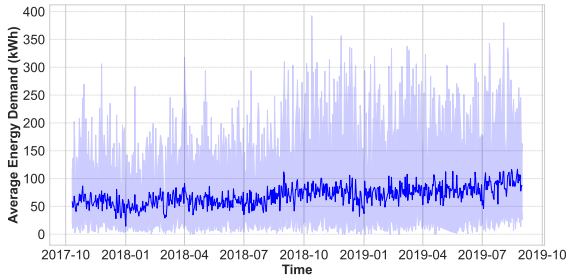
- **Naive:** This model serves as a simple baseline in time series forecasting, assuming that future values will equal the mean of historical values. Termed “naive,” this model overlooks any underlying patterns, trends, or seasonality in the data.



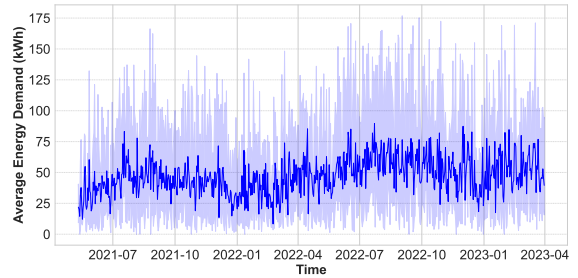
(a) London



(b) Palo Alto



(c) Perth



(d) Boulder

Figure 1: Overview of the Average Daily EV Charging Demand across charging stations. The dark blue line shows the daily average delivered energy demand across time series for different datasets. The light blue area represents the minimum and maximum values of each day. This gives insights into the general trend and demand pattern over time, as well as the distribution of scales of the time series present in each dataset.

- **ARIMA:** A well-established statistical method, the ARIMA model is characterized by three parameters: p , d , and q . For this study, these are set to $p = 30$, $d = 0$, and $q = 30$, aligning with the input chunk length of 30 used in the N-HiTS Model.
- **Transformer:** The Transformer model used in our study follows the architectural setup as outlined in Koohfar et al. (2023). This state-of-the-art architecture offers powerful sequence modeling capabilities. Details of the implementation, including specific hyperparameters, can be found in Table 3.
- **N-HiTS-Local:** We employ the $N\text{-HiTS}_{\text{local}}$ model, which contrasts the $N\text{-HiTS}_{\text{global}}$ by initializing and training a distinct model for each time series. This comparison sheds light on the differences between training the N-HiTS Model on multiple time series simultaneously versus a separate model for each series.

4.6 Evaluation

Each model’s performance is assessed using historical forecasting on the held-out test set of each time series, which consists of 90 days of data for each time series.

The process of historical forecasting, as described in subsection 3.5, uses forecast horizons of 1, 7, and 30 days, considering the various forecasting applications. The accuracy is reported for each dataset separately using the mean average error (MAE), root mean squared error (RMSE) and mean average percentage error (MAPE) metrics. Additionally, the $N\text{-HiTS}\text{-London}_{\text{global}}$ model is employed to make predictions on the Boulder, Palo Alto, and Perth datasets, using the same metrics and forecast horizons. This enables a comparison of the $N\text{-HiTS}\text{-London}_{\text{global}}$ model’s performance with models specifically trained on each individual dataset.

4.7 Implementation Details

For the implementation of all our experiments, the Darts library (Herzen et al., 2022) is utilized. Darts is an open-source Python library specifically tailored for time series forecasting tasks. Darts provides a unified framework integrating statistical models from the statsmodels library and deep learning models implemented in PyTorch.

Hyperparameter	N-HiTS-Boulder	N-HiTS-PaloAlto	N-HiTS-London	N-HiTS-Perth	Transformer	N-HiTS-London _{train}
Input Chunk Length	30	30	30	30	30	30
Output Chunk Length	7	7	7	7	7	7
Batch Size	32	32	32	32	32	32
Number of Stacks	8	4	2	2	N/A	2
Number of Blocks	2	1	3	5	N/A	2
Number of Layers	5	4	5	3	1*	2
Layer Widths	64	256	256	32	128	32
Dropout Rate	0	0.1	0.1	0	0.1	0
Learning Rate	1e-3	1e-3	1e-3	1e-3	1e-3	1e-3
Optimizer	Adam	Adam	Adam	Adam	Adam	Adam
Activation Function	ReLU	ReLU	ReLU	ReLU	ReLU	ReLU
Max Pooling	True	True	True	True	True	True

Table 3: Table showing the optimal hyperparameters selected for the N-HiTS model across different datasets, the Transformer model, and the N-HiTS-London_{train} model.* indicates one encoder and one decoder layer.

5 RESULTS

Comparing Model Forecasting Accuracy The results from Table 4 unequivocally highlight the N-HiTS_{global} model’s dominance in terms of forecasting accuracy. This superiority is evident across the majority of datasets and forecast horizons, with the N-HiTS_{global} consistently outshining its benchmark counterparts.

In the London dataset, the accuracy difference is most pronounced. Notably, at the 1-day forecasting horizon, the N-HiTS_{global} model significantly outperforms all other models. While it remains superior at the 7 and 30-day horizons, the margin of its dominance decreases, pointing to the intricacies of extended forecasts.

The Palo Alto dataset presents a tighter competition. While the N-HiTS_{global} model retains its lead, especially in metrics like MAE and RMSE, its MAPE performance closely mirrors that of the Naive baseline across all forecast durations.

For the Perth dataset, the scenario is more mixed. The N-HiTS_{global} model shows modest advantages at shorter forecasting horizons. Intriguingly, at the 30-day mark, the Naive model takes the lead, emphasizing the inherent challenges of long-range forecasts.

The Boulder dataset displays the N-HiTS_{global} model’s consistent strengths in time series forecasting, with it regularly outpacing the Naive model. However, when juxtaposed with other benchmarks, the performance differences appear to be minimal, suggesting a balanced competitive landscape for this dataset.

Generalization to Unseen Stations The outcomes from the experiment are particularly striking when examining the London-trained N-HiTS model’s performance on the Perth dataset. Its impressive accuracy on this external dataset indicates that the London data

harbors valuable patterns and insights, enabling enhanced knowledge transfer to different geographical contexts. This underlines the model’s robust capacity to generalize and its adaptability to varied infrastructural scenarios.

Additionally, for other external datasets, such as Boulder and Palo Alto, the N-HiTS model, once again trained on the London data, demonstrates commendable generalization capabilities. Despite some minimal accuracy reductions, the consistent performance showcases the model’s resilience and versatility.

Collectively, the evidence strongly advocates for the utility and robustness of the N-HiTS model, especially its ability to perform reliably across diverse charging station datasets. This solidifies the case for the broader adoption of global deep learning models in the realm of EV charging demand forecasting.

6 DISCUSSION

The forecasting of EV charging demand using historical data at the level of individual charging stations remains challenging. The presence of substantial volatility within the demand curve of single charging stations, alongside the limited availability of high-quality time series within each dataset, as demonstrated by our data analysis, continues to pose a hurdle. Despite these challenges, our study provides a foundational framework for understanding the dynamics of EV charging demand forecasting and offers insights into the potential of global deep learning models in tackling this complex task.

The benchmark models did not perform as well as expected, often falling short of even the baseline results. This highlights the challenges of achieving accurate forecasting using pre-configured models, especially when applied to specific time series. It becomes apparent that fine-tuning hyperparameters for each in-

Model	Horizon=1			Horizon=7			Horizon=30			
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	
London	N-HiTS-London _{global}	88.79	70.91	30.16	103.52	83.56	34.86	121.20	99.90	42.86
	N-HiTS-London _{local}	165.90	146.58	61.49	156.81	135.52	158.08	187.02	168.21	125.36
	Transformer	200.72	180.94	190.82	199.18	178.72	285.39	206.32	188.44	267.16
	ARIMA	113.45	93.40	75.74	134.41	112.11	239.57	155.56	134.73	323.80
	Naive	139.48	119.15	66.98	140.38	120.20	67.97	140.22	120.99	69.82
Palo Alto	N-HiTS-PaloAlto _{global}	19.33	15.41	51.84	19.96	16.04	51.69	20.95	17.47	51.56
	N-HiTS-PaloAlto _{local}	27.52	24.07	51.18	27.01	24.01	52.28	27.53	24.53	52.41
	N-HiTS-London _{global}	23.67	20.29	50.48	24.22	20.70	51.56	30.89	27.23	55.00
	Transformer	28.24	25.16	52.36	32.75	29.63	55.00	34.55	31.48	56.83
	ARIMA	35.41	31.85	57.37	31.30	27.64	55.85	26.94	23.74	52.80
	Naive	25.09	21.95	50.84	25.10	21.94	50.80	25.19	22.18	51.34
Perth	N-HiTS-Perth _{global}	40.21	31.97	49.76	41.99	33.17	55.28	50.71	39.65	90.78
	N-HiTS-Perth _{local}	47.46	37.41	67.41	49.02	39.01	72.86	55.24	44.39	94.44
	N-HiTS-London _{global}	38.84	31.77	37.79	39.94	32.23	39.37	44.84	35.88	42.41
	Transformer	49.45	39.18	81.50	47.03	36.98	65.25	48.12	37.71	66.58
	ARIMA	44.46	35.12	74.49	49.11	39.10	79.96	53.63	42.80	80.78
	Naive	44.71	34.76	58.32	44.86	34.89	58.78	46.78	36.40	61.17
Boulder	N-HiTS-Boulder _{global}	24.96	20.44	42.39	24.83	19.70	49.27	23.68	18.79	47.69
	N-HiTS-Boulder _{local}	28.01	22.13	67.26	28.64	23.11	83.71	27.04	21.31	71.15
	N-HiTS-London _{global}	31.70	26.65	42.03	35.00	30.04	43.62	51.98	47.19	51.29
	Transformer	29.56	23.74	72.69	25.86	20.82	54.54	24.48	19.49	51.92
	ARIMA	34.80	28.34	208.56	35.95	30.21	324.58	25.58	20.30	61.20
	Naive	25.99	20.78	53.76	25.67	20.53	53.26	24.05	19.10	49.82

Table 4: Comparison of forecasting accuracy of investigated models at various forecast horizons across datasets. The N-HiTS_{global} model exhibits superior performance across most metrics and datasets, showcasing its effectiveness for time series forecasting.

dividual model is crucial for success. For instance, the Local N-HiTS model, although designed for a fair comparison, emphasizes the need for customizing models to suit particular time series data. The Transformer-based approach, despite being touted as state-of-the-art in previous work, couldn't be entirely replicated in our setup, possibly due to differences in data handling and potential overfitting. The ARIMA model, while adjusted to align with N-HiTS parameters, might have benefited from a wider range of hyperparameter exploration, particularly for complex, non-stationary time series like those in the London dataset. Conversely, the global modeling approach stands out by simplifying the modeling process, saving valuable time and computational resources, making it especially advantageous when dealing with a large number of time series.

In assessing our forecasting models, RMSE, MAE, and MAPE are used to evaluate their performance comprehensively. The proximity between reported MAE and RMSE values might be due to the dataset's limited outlier presence, minimizing the impact of RMSE's outlier-penalizing nature. MAPE, although scale-invariant, demonstrated high sensitivity to the low signal-to-noise ratio in our context, leading to exaggerated errors. The elevated MAPE values likely stem from the challenges posed by this low signal-to-noise ratio, causing the models to struggle with accurate predictions amidst noise and outliers.

7 CONCLUSION

In this study, a novel framework is presented, tailored to tackle the complexities of forecasting EV charging demand at multiple charging stations over longer periods of time. By considering several time series, a clearer understanding of demand variations and trends is gained. Moreover, by evaluating these models over extended periods, the aim is to ensure their durability and adaptability, reflecting the actual dynamics observed on the ground and providing dependable insights over different periods.

Through a series of experiments, the efficacy of global deep learning models in enhancing the accuracy and reliability of demand forecasting for EV charging demand is demonstrated. The applied framework not only assesses performance across varied charging station sites but also leverages the strengths of these models. Specifically, the N-HiTS model's capability to discern intricate patterns via global training distinguishes it from conventional benchmarks, emphasizing its utility for real-world applications that necessitate precise and robust time se-

ries predictions.

In exploring the capacity of global deep learning models to predict demand at newly established charging stations, which were previously unobserved, the adaptability of the N-HiTS Model to unfamiliar data from various stations is examined. The results emphasize its consistent ability to generalize across diverse datasets, showcasing its reliability in delivering accurate forecasts for a wide range of datasets.

Lastly, additional analysis of the N-HiTS model's generalization performance is provided. By exploring the effects of varying training lengths on the model, a deeper understanding of its strengths and limitations is gained. The experiment highlights the superior robustness of global learning while shedding light on the intricate and sometimes unpredictable behavior of local learning. These insights provide valuable guidelines for the implementation of global deep learning models across diverse contexts and requirements.

The forecasting of EV charging demand using historical data at the level of individual charging stations remains challenging. The presence of substantial noise within the demand curve of single charging stations, alongside the limited availability of high-quality time series within each dataset, as demonstrated by the data analysis, continues to pose a hurdle. Despite these challenges, the study provides a foundational framework for understanding the dynamics of EV charging demand forecasting and offers insights into the potential of global deep learning models in tackling this complex task.

Future Research As previously mentioned, one of the pivotal challenges encountered in this study is the volatile nature of data. One potential strategy to alleviate these concerns is to expand the forecasting framework to encompass broader geographical and temporal dimensions. This could aid in dampening the inherent noise seen within individual demand curves, enabling more reliable analysis of cross-series learning by global deep learning models.

Drawing from the insights provided by Oreshkin et al. (2020), there is growing interest surrounding the application of zero-shot learning for time series forecasting. Leveraging pre-trained models across disparate time series could open new horizons in terms of forecast accuracy and model adaptability.

Lastly, inspired by the methodology presented by Yi et al. (2022), clustering time series based on common attributes offers an intriguing prospect. This method holds the potential to enhance cross-learning capabilities among models, thereby fortifying their generalization capabilities.

REFERENCES

- Amara-Ouali, Y., Goude, Y., Massart, P., Poggi, J.-M., and Yan, H. (2021). A review of electric vehicle load open data and models. *Energies*.
- Andrenacci, N., Ragona, R., and Valenti, G. (2016). A demand-side approach to the optimal deployment of electric vehicle charging stations in metropolitan areas. *Applied Energy*, 182:39–46.
- Buzna, L., De Falco, P., Khormali, S., Proto, D., and Straka, M. (2019). Electric vehicle load forecasting: A comparison between time series and machine learning approaches. In *2019 1st International Conference on Energy Transition in the Mediterranean Area (SyNERGY MED)*, pages 1–5.
- Challu, C., Olivares, K. G., Oreshkin, B. N., Garza, F., Mergenthaler-Canseco, M., and Dubrawski, A. (2022). N-hits: Neural hierarchical interpolation for time series forecasting.
- Chen, Y., Kang, Y., Chen, Y., and Wang, Z. (2020). Probabilistic forecasting with temporal convolutional neural network. *Neurocomputing*, 399.
- Council, P. K. (2019). Perth & kinross council ev charging station data. Datasets for Perth & Kinross Council’s EV charging stations under the ChargePlace Scotland scheme. Period from 2016 to 2019.
- Eddine, M. D. and Shen, Y. (2022). A deep learning based approach for predicting the demand of electric vehicle charge. *J. Supercomput.*, 78(12):14072–14095.
- Gerossier, A., Girard, R., and Kariniotakis, G. (2019). Modeling and forecasting electric vehicle consumption profiles. *Energies*, 12:1341.
- Herzen, J., Lässig, F., Piazzetta, S. G., Neuer, T., Tafti, L. T., Raille, G., Pottelbergh, T. V., Pasięka, M., Skrodzki, A., Huguenin, N., Dumonal, M., Kościsz, J., Bader, D., Gusset, F., Benheddi, M., Williamson, C., Kosinski, M., Petrik, M., and Grosch, G. (2022). Darts: User-friendly modern machine learning for time series. *Journal of Machine Learning Research*, 23(124):1–6.
- Hu, T., Liu, K., and Ma, H. (2021). Probabilistic electric vehicle charging demand forecast based on deep learning and machine theory of mind. In *2021 IEEE Transportation Electrification Conference & Expo (ITEC)*, pages 795–799.
- Huber, J., Dann, D., and Weinhardt, C. (2020). Probabilistic forecasts of time and energy flexibility in battery electric vehicle charging. *Applied Energy*, 262:114525.
- Hüttel, F. B., Peled, I., Rodrigues, F., and Pereira, F. C. (2021). Deep spatio-temporal forecasting of electrical vehicle charging demand.
- Kim, Y. and Kim, S. (2021). Forecasting charging demand of electric vehicles using time-series models. *Energies*, 14(5).
- Koohfar, S., Woldemariam, W., and Kumar, A. (2023). Prediction of electric vehicles charging demand: A transformer-based deep learning approach. *Sustainability*, 15(3).
- Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J. E., and Stoica, I. (2018). Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*.
- Louie, H. M. (2017). Time-series modeling of aggregated electric vehicle charging station load. *Electric Power Components and Systems*, 45(14):1498–1511.
- Moon, H., Park, S. Y., Jeong, C., and Lee, J. (2018). Forecasting electricity demand of electric vehicles by analyzing consumers’ charging patterns. *Transportation Research Part D: Transport and Environment*, 62:64–79.
- of Boulder, C. (2020). Electric vehicle charging station energy consumption. Dataset containing energy consumption data for electric vehicle charging stations in Boulder, Colorado.
- of Palo Alto, C. (2021). Electric vehicle charging station usage (july 2011 - december 2020). Open data provided by the City of Palo Alto containing electric vehicle charging station usage data from July 2011 to December 2020.
- Oreshkin, B. N., Carпов, D., Chapados, N., and Bengio, Y. (2020). Meta-learning framework with applications to zero-shot time-series forecasting.
- Ren, F., Tian, C., Zhang, G., Li, C., and Zhai, Y. (2022). A hybrid method for power demand prediction of electric vehicles based on sarima and deep learning with integration of periodic features. *Energy*, 250:123738.
- Salinas, D., Flunkert, V., Gasthaus, J., and Januschowski, T. (2020). Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191.
- Su, S., Zhao, H., Zhang, H., Lin, X., Yang, F., and Li, Z. (2017). Forecast of electric vehicle charging demand based on traffic flow model and optimal path planning. In *2017 19th International Conference on Intelligent System Application to Power Systems (ISAP)*, pages 1–6.
- Wang, S., Xue, G., Ping, C., Wang, D., You, F., and Jiang, T. (2018). The application of forecasting algorithms on electric vehicle power load. In *2018 IEEE International Conference on Mechatronics and Automation (ICMA)*, pages 1371–1375.
- Xydas, E. S., Marmaras, C. E., Cipcigan, L. M., Hassan, A. S., and Jenkins, N. (2013). Forecasting electric vehicle charging demand using support vector machines. In *2013 48th International Universities’ Power Engineering Conference (UPEC)*, pages 1–6.
- Yan, J., Zhang, J., Liu, Y., Lv, G., Han, S., and Alfonzo, I. E. G. (2020). Ev charging load simulation and forecasting considering traffic jam and weather to support the integration of renewables and evs. *Renewable Energy*, 159:623–641.
- Yi, Z., Liu, X. C., Wei, R., Chen, X., and Dai, J. (2022). Electric vehicle charging demand forecasting using deep learning model. *Journal of Intelligent Transportation Systems*, 26(6):690–703.
- Zhu, J., Yang, Z., Mourshed, M., Guo, Y., Zhou, Y., Chang, Y., Wei, Y., and Feng, S. (2019). Electric vehicle charging load forecasting: A comparative study of deep learning approaches. *Energies*, 12(14).