CarbonX: An Open-Source Tool for Computational Decarbonization Using Time Series Foundation Models

Diptyaroop Maji*
University of Massachusetts, Amherst
Amherst, USA
dmaji@cs.umass.edu

Kang Yang*
University of California, Los Angeles
Los Angeles, USA
kyang73@g.ucla.edu

Prashant Shenoy
University of Massachusetts Amherst
Amherst, USA
shenoy@cs.umass.edu

Ramesh K. Sitaraman
University of Massachusetts Amherst
and Akamai Tech
Amherst, USA
ramesh@cs.umass.edu

Mani Srivastava†
University of California, Los Angeles
& Amazon AWS
Los Angeles, USA
mbs@ucla.edu

Abstract

Computational decarbonization aims to reduce carbon emissions in computing and societal systems such as data centers, transportation, and built environments. This requires accurate, fine-grained carbon intensity forecasts, yet existing tools have several key limitations: (i) they require grid-specific electricity mix data, restricting use where such information is unavailable; (ii) they depend on separate grid-specific models that make it challenging to provide global coverage; and (iii) they provide forecasts without uncertainty estimates, limiting reliability for downstream carbon-aware applications.

In this paper, we present `Carbon`*X*, an open-source tool that leverages Time Series Foundation Models (TSFMs) for diverse decarbonization tasks. `Carbon`*X* utilizes TSFMs to provide strong performance in carbon intensity forecasting and imputation across multiple and heterogeneous grids. Using only historical carbon intensity data and a single general model, our tool achieves a zero-shot forecasting Mean Absolute Percentage Error (MAPE) of 15.82% across 214 grids worldwide. Across 13 benchmark grids, `Carbon`*X* performance is comparable with the current state-of-the-art, with an average MAPE of 9.59% and tail forecasting MAPE of 16.54%, while also providing prediction intervals with 95% coverage. `Carbon`*X* can provide forecasts for up to 21 days with minimal accuracy degradation. Further, when fully fine-tuned, `Carbon`*X* outperforms the statistical baselines by 1.2—3.9× on the imputation task. Overall, these results demonstrate that `Carbon`*X* can be used easily on any grid with limited data and still deliver strong performance, making it a practical tool for global-scale decarbonization.

*Both authors contributed equally to this research.
†Mani Srivastava holds concurrent appointments as a Professor of ECE and CS (joint) at the University of California, Los Angeles, and as an Amazon Scholar at Amazon. This paper describes work performed at UCLA and is not associated with Amazon.

E-Energy '26, Banff, AB, Canada

ACM ISBN 979-8-4007-2011-6/26/06
https://doi.org/10.1145/3744255.3798120

CCS Concepts

• **Social and professional topics** → **Sustainability**; • **Computing methodologies** → *Neural networks*; • **Applied computing** → *Forecasting*.

Keywords

Computational Decarbonization, Time Series Foundation Models, Carbon Intensity Forecasting, Imputation

ACM Reference Format:
Diptyaroop Maji, Kang Yang, Prashant Shenoy, Ramesh K. Sitaraman, and Mani Srivastava. 2026. CarbonX: An Open-Source Tool for Computational Decarbonization Using Time Series Foundation Models. In *The 17th ACM International Conference on Future and Sustainable Energy Systems (E-Energy '26), June 22–25, 2026, Banff, AB, Canada.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3744255.3798120

1 Introduction

Societal infrastructure, including electricity grids, buildings, transportation, and computing systems, contributes over 80% of global greenhouse gas (GHG) emissions, underscoring the need for scalable decarbonization [39]. To reduce such GHG emissions, a new field of computational decarbonization is gaining attention in recent years [18], which aims to reduce both operational and embodied carbon emissions associated with complex computing and societal infrastructure systems by leveraging sensing, optimization, and Artificial Intelligence (AI)-driven control.

A key enabler for computational decarbonization is accurate, fine-grained estimates and forecasts of the electricity grid's *carbon intensity*, which quantifies the emissions associated with per unit electricity usage over a period (in gCO_2eq/kWh). Carbon intensity depends on several factors such as the electricity (energy) mix, weather, and demand patterns [25], and therefore, varies significantly both spatially and temporally. For example, Figure 1 shows the carbon intensities in two locations: California exhibits a "duck curve" with midday solar dips and evening gas spikes, while Germany shows a higher and more volatile pattern driven by fossil fuels and intermittent wind or solar generation. Systems can take advantage of such variability to reduce their carbon emissions without compromising performance [1, 18, 29, 37, 41]. For instance, flexible demands such as electric vehicle charging or large-scale Machine Learning (ML) model training can be scheduled during low-carbon

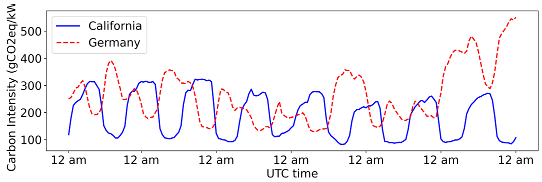


Figure 1: ***Hourly carbon intensity*** (gCO_2eq/kWh) ***in the California and Germany grids over one week.***

periods to reduce the associated emissions. Thus, fine-grained carbon intensity estimates and accurate forecasts, when accessible at a global scale, can help in a significant reduction of carbon emissions and subsequently advance the decarbonization goals.

Limitations of Prior Work. In recent years, several ML-based specialized tools have emerged that can provide accurate short-term grid carbon intensity forecasts [25, 45, 47]. Although these tools show considerable promise, these tools face three key limitations when availability at a global scale is considered:

(i) Dependence on data that are not generally available. Existing tools rely on information about a grid's source mix to generate accurate carbon intensity forecasts. While some grid operators [5, 8] and services [9, 40] have begun releasing historical carbon intensity data, reporting source mix remains non-mandatory [42] and is thus unavailable for many grids, limiting the applicability of existing models. For example, there are entire continents, such as Africa, where electricity mix data is either not available or only available at an annual scale [17]; hence, prior tools cannot be used.

(ii) Dependence on grid-specific models. Prior tools rely on grid-specific models trained separately for each individual grid, and on some occasions, even individual electricity sources [25]. Although this design effectively captures grid-specific carbon dynamics, it is highly challenging to scale globally across grids. For example, at the transmission system operator level, there are more than 350 grids worldwide [11]. With each grid having multiple electricity sources, prior works need to train and maintain thousands of models. At the distribution level, there are more than 2500 operators alone in Europe [13], making the per-grid design even more challenging.

(iii) No quantification of forecast uncertainty. Similar to other forecasting systems (*e.g.*, weather forecasts), carbon intensity prediction involves inherent uncertainty. Most prior works generate point forecasts without capturing possible variations. Such forecasts can deviate significantly from the ground truth, especially over multi-day horizons [22], reducing the reliability of downstream applications that perform optimizations by relying on these forecasts.

Our Goal. These gaps motivate the need for a tool that (i) operates without relying on auxiliary information such as energy mix, (ii) generalizes across global electricity grids using a unified architecture, (iii) produces uncertainty-aware forecasts suitable for real-world deployment, and is also easy to use and maintain at a global scale. In addition, tasks such as carbon intensity imputation remain underexplored despite their importance for carbon-aware optimizations. For example, imputation can recover missing data or enhance temporal resolution in regions with sparse carbon intensity reporting [9], increasing the effectiveness of downstream carbon-aware applications. Therefore, there is also a need for a tool that can perform other tasks, such as carbon intensity imputation, in addition to forecasting.

Motivation for Using TSFMs. Recent advances in large-scale ML have introduced Time Series Foundation Models (TSFMs) [2, 6, 16, 24, 36], which demonstrate strong forecasting and imputation performance across domains such as electricity demand and weather. These models are pre-trained on massive, heterogeneous data, enabling general temporal representations that support reliable zero- and few-shot forecasting. This capability directly addresses key limitations of existing tools. TSFMs can operate using only univariate historical signals, eliminating the need for electricity-mix data that is unavailable in many regions. Their strong zero-shot generalization also enables a single unified model to serve hundreds of grids, avoiding per-grid model development and maintenance. Furthermore, zero-shot deployment removes the need for per-grid data collection and retraining, significantly reducing operational cost at scale. Leveraging TSFMs, therefore, provides a practical pathway toward scalable, globally deployable, and data-efficient tools that advance computational decarbonization.

Our Research Contributions. In this paper, we present Carbon*X*, a tool built using TSFMs for tasks such as carbon intensity forecasting and imputation. Using only carbon intensity series and a single general model, Carbon*X* achieves an average Mean Absolute Percentage Error (MAPE) of 15.82% across 214 grids worldwide for the forecasting task. Although Carbon*X* has slightly higher forecasting errors than existing specialized tools in some grids, it delivers similar tail performance (see Section 6.1) using only historical carbon intensity as input. Further, it offers significantly more flexibility and versatility (see Table 1), such as one model for any grid and tasks beyond forecasting. These attributes make Carbon*X* more practical than any similar tools existing today. In summary, this paper makes the following contributions:

(1) *Application of TSFMs on Decarbonization Tasks.* We present Carbon*X*, a tool for global-scale decarbonization. Our work is the *first* to look at application of TSFMs on decarbonization tasks such as carbon intensity forecasting and imputation. Carbon*X* has a modular design that leverages the versatility of TSFMs to provide comparable accuracy with the current state-of-the-art using only historical carbon intensity data as input. On the 13 benchmark grids used in prior works, Carbon*X* attains 9.59% mean MAPE and 16.54% tail MAPE, compared to 7.75% and 15.61% for the current state-of-the-art.
(2) *Open-Source Tool.* We release Carbon*X* as an open-source tool for researchers and practitioners to facilitate and accelerate decarbonization research. Carbon*X* currently supports carbon intensity forecasting and imputation on a global scale, and the saved TSFM models, curated datasets, and an API suite is made available at **https://github.com/codecexp/CarbonX**.
(3) *Global Coverage with Strong Forecasting Accuracy.* Carbon*X* provides multi-day carbon intensity forecasts across 214 grids globally. Due to its zero-shot capability, which requires no additional training, support for any new grids can also be added easily. Across all 214 grids, the mean and tail MAPEs are 15.82% and 29.61%, respectively, showcasing Carbon*X*'s strong forecasting accuracy. Carbon*X* is the first research work to provide and evaluate carbon intensity forecasting on a global scale.
(4) *Extended Forecasting Horizon.* Carbon*X* can forecast carbon intensity for an extended horizon with minimal performance loss. Our evaluation across 13 grids shows that Carbon*X* can forecast

up to 21 days ahead with only a 5.40% average MAPE degradation. To our knowledge, it is the first work to support carbon intensity forecasting beyond a 96-hour horizon.

(5) *Uncertainty Quantification.* Carbon*X* provides calibrated Prediction Intervals (PIs) to make the outputs reliable for downstream tasks. For forecasting, it achieves 95% average coverage with a 54.2% normalized interval width across 13 grids, demonstrating effective uncertainty quantification.

(6) *Imputation Support.* Carbon*X* also supports carbon intensity imputation, offering versatility beyond forecasting. Our fully fine-tuned implementation achieves an average normalized RMSE of 0.24 across 50 grids even with 75% missing values, outperforming the statistical baselines by 1.2—3.9×.

2 Background

2.1 Electricity Grids and Carbon Intensity

The electricity grid combines generation from renewable and non-renewable sources. Electricity is transmitted through high-voltage lines and distributed to end users. Each regional grid is managed by an operator who balances supply and demand. As demand varies, operators monitor the grid and weather conditions in real time, and dispatch generators accordingly to meet current demand.

Carbon Intensity (CI). The average carbon intensity of a grid (in gCO_2eq/kWh) is the generation-weighted average of emissions from each electricity source at a certain time:

$$\text{Carbon Intensity (CI)} = \frac{\sum (E_i \cdot CEF_i)}{\sum E_i}, \quad (1)$$

where E_i is the electricity generated by source i (*e.g.*, coal, solar, gas) in kWh, and CEF_i is its corresponding carbon emission factor (CEF, in gCO_2eq/kWh), a standardized value for each source [7, 33].

2.2 Time Series Foundation Models (TSFMs)

TSFMs are large-scale ML models pre-trained on diverse time-series data [2, 16]. By learning universal temporal representations, it enables zero- or few-shot adaptation across domains. TSFMs have been applied to multiple domains ranging from electricity load forecasting to healthcare [3, 15, 20, 23, 28]. These applications demonstrate TSFMs' ability to generalize across domains and encourage application of new or existing TSFMs on carbon intensity data.

2.2.1 *Supported TSFMs.*

Carbon*X* currently supports the following five TSFMS. These models differ in architecture (*e.g.*, encoder-decoder) and training strategy (*e.g.*, supervised, self-supervised).

MOMENT [16]: A transformer-based model, pre-trained using a masked time series. It segments input sequences into fixed-length patches and encodes them using a modified T5-style [30] transformer. A reconstruction head then enables patch-level prediction.

Chronos [2]: A transformer-based forecasting model that repurposes language models for time series. It tokenizes time series via scaling and quantization into a fixed-size vocabulary, enabling training with off-the-shelf encoder-decoder [30]. It uses cross-entropy loss to model the categorical distribution of tokenized values.

TimesFM [6]: A decoder-only model that processes time series as sequences of non-overlapping patches, which are embedded using residual multilayer perceptron blocks and passed through stacked transformer layers with causal self-attention.

Sundial [24]: A Transformer-based model that pre-trains directly on continuous time series without tokenization, leveraging a flow-matching TimeFlow loss to predict next-patch distributions, mitigate mode collapse, and generate arbitrary-length forecasts.

Time-MoE [36]: A decoder-only Transformer with a sparse Mixture-of-Experts (MoE) design that activates only a few experts per step, reducing computation while preserving model capacity. Similar to others, it supports variable forecasting horizons.

Table 10 in Appendix B lists the TSFM variants currently supported, and their parameter counts. We select five representative TSFMs for their architectural diversity, training strategies, strong cross-domain performance, and open-source availability. While other TSFMs exist (e.g., TimeGPT [14], LagLlama [31], and TimeLLM [19]), we exclude them as: (i) LagLlama shows weaker performance relative to the selected models [26]; (ii) TimeLLM follows an architecture similar to Chronos by adapting language models for time series; and (iii) TimeGPT is closed-source with limited accessibility. Importantly, Carbon*X* 's modular design allows seamless integration of existing and future TSFMs for decarbonization tasks.

2.2.2 *TSFM Fine-Tuning.*

TSFMs show improved performance under a few-shot setting, when fine-tuned with data from the same domain on which they are applied [16]. Fine-tuning strategies can be categorized as *model-oriented* or *data-oriented*.

In *model-oriented fine-tuning*, either a subset or the entire model is updated using domain-specific data. Fine-tuning only part of the model, *e.g.*, the task head layers, is referred to as *lightweight* fine-tuning. Updating all model parameters constitutes *full* fine-tuning.

Data-oriented fine-tuning for grid data falls into two types. In *domain-specific* fine-tuning, the model is trained on data from multiple grids and evaluated on an unseen grid, testing inter-grid generalization. In *grid-specific* fine-tuning, the model is fine-tuned and evaluated on data from the same grid, testing intra-grid accuracy.

3 Related Work

Table 1 summarizes the key differences and advantages of our tool relative to previous approaches.

(1) Carbon Intensity Forecasting. Short-term carbon intensity forecasting is well-studied due to its importance for decarbonization. Traditional forecasting methods, such as EWMA [35] and ARIMA [4, 21, 34], have been applied in several prior works because of their simplicity and effectiveness in capturing temporal patterns. However, such statistical models struggle to handle high variability and are less effective for multi-day forecasting horizons.

(a) Domain-Specific ML Models. To address the limitations of traditional statistical methods, ML models have emerged as state-of-the-art solutions [25, 45, 47]. CarbonCast [25] employs a two-tier ML architecture for 96-hour carbon intensity forecasting. The first tier predicts electricity generation for each energy source. The second tier takes these predictions along with historical carbon intensity and weather forecasts to estimate 96-hour carbon intensity forecasts. More recently, EnsembleCI [45] applies ensemble learning across three base models to produce 96-hour forecasts. Each base model predicts the carbon intensity of a given grid, and a stacking-based strategy is used to generate the final forecast. Despite their effectiveness, these models rely on extensive grid-specific electricity mix data, frequent retraining, and grid-specific model development

	Energy mix not required	Training data volume & training frequency	Transferrable across grids	Prediction intervals	Tasks beyond forecasting	Forecasting accuracy
EWMA [35]	✓	High	✗	✗	✗	Moderate
CarbonCast [25]	✗	High	✗	✗	✗	High
CarbonCast w/ PI [22]	✗	High	✗	✓	✗	High
EnsembleCI [45]	✗	High	✗	✗	✗	High
CarbonX	✓	**Low**	✓	✓	✓	**Moderate-High**

Table 1: ***Comparison between* CarbonX *and existing state-of-the-art models.***

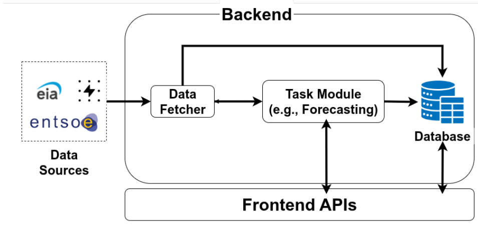


Figure 2: **CarbonX** ***end-to-end pipeline. The frontend communicates with the task module and the database during execution.***

to maintain accuracy. In contrast, CarbonX requires only historical carbon intensity and achieves similar accuracy in many grids.
(b) Proprietary Forecasting Services. Electricity Maps [9] and WattTime [40] offer global carbon intensity forecasts. However, these forecasts are proprietary, and their models and forecasting accuracy are not publicly available. In contrast, our tool is fully open-sourced to support reproducibility for researchers and practitioners.
(c) Forecast Uncertainty. Most prior works provide only point-value forecasts and do not quantify uncertainty. Recently, Li *et al.* [22] have extended CarbonCast [25], where they apply a conformal prediction method [43] to generate prediction intervals. However, underlying limitations remain, such as its reliance on electricity mix data and the per-grid model. In contrast, CarbonX requires only historical carbon intensity data to generate intervals.
(2) Carbon Intensity Imputation. Carbon intensity imputation enables recovery of missing data and generation of higher-resolution estimates beyond currently available granularity. For example, hourly carbon intensity data can be imputed to produce sub-hourly values. Currently, only Electricity Maps performs such imputation using proprietary statistical and ML techniques [10]. Although statistical methods are widely used for imputation, to the best of our knowledge, CarbonX is among the first to treat carbon intensity imputation as a standalone task.

4 CarbonX Design

CarbonX is a generalizable, versatile tool that performs carbon domain tasks across electricity grids with high accuracy, using univariate carbon intensity series. It currently supports multi-day forecasting and imputation, and can be extended to other decarbonization tasks, such as anomaly detection, with minimal changes.

4.1 End-to-end Pipeline and Usage

We begin by briefly describing the end-to-end pipeline and illustrating how to use CarbonX. Figure 2 illustrates the CarbonX pipeline, including the frontend and backend, where the frontend interacts with the backend's task module and database during execution.

4.1.1 ***CarbonX Backend.*** CarbonX backend comprises a data fetcher module, a task module, and a database to store both fetched and generated data.
Data Fetcher. This module periodically retrieves carbon intensity data from various public data sources. Currently, CarbonX fetches hourly carbon intensity data from Electricity Maps [11] for 214 grids. We plan to include other sources (*e.g.*, EIA [8], ENTSOE [12]) in the future. The module currently runs once daily, but the periodicity can be adjusted as per requirements. Once fetched, the data is forwarded to the task module as well as updated to the database.
Task Module. The task module receives data from the data fetcher and performs the desired task. Regardless of the task, it first runs imputation (using MOMENT) so that any missing data is filled.

Figure 3 illustrates the CarbonX task module for carbon intensity forecasting. The module can use any new or existing TSFM to forecast $\hat{y}_{t+1:t+H}$ from historical input $y_{t-L+1:t}$, where L is the input length and H is the forecast horizon. The output is then passed through a conformal prediction layer, which produces a prediction interval with coverage guarantees. Once the forecasts and prediction intervals are generated, we update the database. We describe the task module in more detail in Section 4.2.
Database. The database is currently a list of files containing the actual carbon intensity data, carbon intensity forecasts, and prediction intervals for all the supported grids. The database is updated daily after the data fetcher and task modules run.

4.1.2 ***CarbonX Frontend.*** The frontend provides a suite of APIs that users or downstream services can use to easily access any historical carbon intensity data, as well as forecasts from any prior or current date for any supported grid. For example, suppose a data center in Texas requires carbon intensity forecasts for the next day in order to optimally schedule a batch job. Calling `get_ci_forecasts(ERCOT, <date>, 24, True)` will retrieve day-ahead carbon intensity forecasts from CarbonX for the Texas grid (ERCOT) along with prediction intervals.

Additionally, CarbonX users can set any preferred TSFM and the mode of operation (zero-shot or fine-tuned), query which grids are supported, and even generate finer resolution carbon intensity data as per requirements. Table 9 in Appendix A provides a description of the APIs currently supported by CarbonX, the required parameters, and the values returned. A detailed documentation can be found in **https://github.com/codecexp/CarbonX**.

4.2 CarbonX Task Module

The CarbonX task module takes only a univariate carbon intensity time series as input and performs forecasting or imputation. The task module consists of three sub-modules:

(i) The input first passes through a TSFM for carbon intensity imputation, ensuring missing values are estimated before further

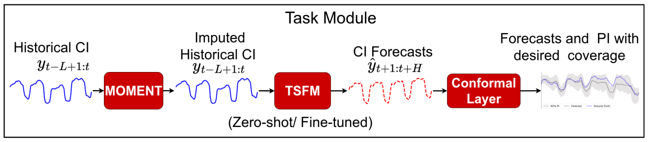


Figure 3: CarbonX *task module comprises TSFMs followed by a conformal layer. This figure shows the forecasting task: the TSFM takes L timesteps of historical carbon intensity and forecasts the next H timesteps. A conformal layer provides reliable prediction intervals with the desired coverage.*

processing. The TSFM can operate in zero-shot or fine-tuned mode; currently, Carbon*X* uses MOMENT for imputation. If the desired task is imputation, the next sub-module is skipped, and the data is sent to the conformal layer.

(ii) For forecasting, the data is input to a second TSFM. This sub-module is model-agnostic and supports any TSFM suited to the task, enabling evaluation of both new and existing models across domains. For instance, we plug in five representative TSFMs to assess their forecasting performance. The TSFMs can operate in either zero-shot or fine-tuned modes. The output is passed to the conformal layer.

(iii) The final sub-module is a conformal layer that produces calibrated prediction intervals with desired coverage, defined as the fraction of times ground truth values fall within the interval. Tasks such as multi-day forecasting introduce increasing uncertainty as the horizon grows. While many TSFMs generate probabilistic forecasts, such intervals do not guarantee coverage. To address this, Carbon*X* applies conformal prediction and generates calibrated intervals that *cover* the ground truth 95% of the time.

Our modular design allows seamless integration of different TSFMs and conformal techniques with minimal changes, improving reliability for downstream decarbonization applications that require robust carbon intensity estimates.

Next, we describe the zero-shot and fine-tuned modes of operation in detail, followed by our uncertainty quantification approach.

4.2.1 ***Zero-Shot Mode.***

Applications face three challenges when estimating or forecasting carbon intensity to assess emissions. First, many grids provide only historical carbon intensity data. Second, available data may contain large gaps due to reporting errors or outages. Third, maintaining separate models per grid becomes difficult as applications scale across regions.

The *zero-shot* mode addresses these challenges by applying TSFMs pre-trained on other domains directly to unseen carbon-related tasks. A single TSFM is deployed across all grids without fine-tuning. This "plug-and-play" mode is well-suited for short-term tasks where forecasts are only required for a limited amount of time, or grids lacking sufficient data for fine-tuning. It is also practical for large-scale deployments requiring global forecasts — *e.g.*,, spatial demand shifting across data centers to reduce emissions. In such settings, zero-shot mode will be beneficial, while maintaining separate models per grid may be infeasible.

(i) Zero-shot Forecasting. For zero-shot forecasting, Carbon*X* input consists solely of univariate historical carbon intensity data from electricity grids. At present, Carbon*X* does not include any other information that may be available (*e.g.*, weather forecasts). However, covariate support can be added with minimal changes. As mentioned, we evaluate five TSFMs in this mode (see Table 10).

During zero-shot forecasting, we also vary the input length (*e.g.*, 24 to 480 hours) and output length (*e.g.*, 1 to 21 days) to study how historical context and forecasting horizon shape model behavior. Short inputs may capture recent trends and support short-term responsiveness, while longer inputs may reveal seasonal patterns but risk introducing noise. On the other hand, TSFMs may be able to generalize over a longer time due to their large model size, contrary to existing domain-specific ML models [25, 45, 47] that typically support only up to 4-day forecasts due to limited model capacity.

(ii) Zero-shot Imputation. At present, Carbon*X* uses MOMENT for imputation, since only MOMENT supports imputation among the explored TSFMs. The input for the imputation task is a carbon intensity time series with portions intentionally masked. This setup mirrors real-world conditions where data gaps arise from reporting delays or power outages [38]. Similar to forecasting, a single model is applied uniformly across grids to estimate missing values using the temporal patterns that the model learnt during pre-training.

4.2.2 ***Fine-Tuned Mode.***

Many grids publish long-term historical carbon intensity data [11]. Since fine-tuning on target-domain data typically improves TSFM performance [16], Carbon*X* supports fine-tuning with historical data to enhance forecast accuracy. This mode is particularly beneficial for long-term deployments or settings with a small number of grids, where it can better capture grid-specific temporal patterns than zero-shot inference. In our experiments, we fine-tune all TSFMs except Sundial, which currently does not support fine-tuning [24].

(i) Fine-Tuned Forecasting. We perform two types of fine-tuning for forecasting. First, we employ a *lightweight, grid-specific* strategy to improve the forecasting accuracy. For example, we fine-tune only the linear forecasting head of MOMENT using carbon intensity data (y) from a grid while keeping the transformer backbone fixed. At each timestep t, the model looks back at the last L time steps ($y_{t-L+1:t}$), and is trained to forecast the next H steps ($\hat{y}_{t+1:t+H}$). The model is then evaluated on unseen data from the same grid.

Unlike domain-specific tools such as CarbonCast [25] or EnsembleCI [45], which train separate models for each grid, we use a shared pre-trained backbone across all grids and update only a few model layers. This enables efficient few-shot adaptation and inter-grid generalization.

Since the above strategy still needs some grid-specific modifications, we also perform *full, domain-specific* fine-tuning to build a scalable, general model, where all model parameters are updated using combined training data from over 200 grids. This mode falls between zero-shot and grid-specific fine-tuning, and is designed to capture diverse carbon intensity patterns without requiring per-grid customization.

Among TSFMs with fine-tuning capabilities, MOMENT and TimesFM gain primarily from *grid-specific* fine-tuning of their forecasting heads and are less suitable for this large-scale strategy, while Chronos and Time-MoE show little benefit (see Section 6.1). Hence, we focus only on Chronos and Time-MoE for the this strategy.

(ii) Fine-Tuned Imputation. We explore both *light-weight* and *full, grid-specific* fine-tuning strategies for the imputation task. The objective is to reconstruct missing carbon intensity values, given a

time series. Formally, let y = $\{y_{1:T}\}$ denote the complete time series over T time steps, and $m = \{m_{1:T}\}$ be a binary mask vector, where:

$$m_t = 0 \text{ if } y_t \text{ is missing (masked), else } 1 \tag{2}$$

The observed data is denoted by $y^{\text{obs}} = \{y_t \mid m_t = 1\}$, and the model is trained to impute the missing values $\hat{y}_t \ \ \forall t \mid m_t = 0$.

While the *light-weight* approach emphasizes scalability and generalization, *full fine-tuning* provides high reconstruction accuracy that outperforms the statistical baselines (see Section 7). We note that *full fine-tuning* results in a separate model for each grid and hence may not always be feasible. However, it achieves strong performance using only historical carbon intensity data and few-shot learning. Hence, we present the results from both strategies to provide a comprehensive comparison, leaving the desired strategy to the downstream application.

4.2.3 ***Uncertainty Quantification.*** Carbon intensity estimates and forecasts involve considerable uncertainties due to external factors such as weather and changing grid conditions. Given a desired coverage level $\alpha \in (0, 1)$, CarbonX's conformal layer generates prediction intervals that cover the ground truth at least $100 \times \alpha\%$ of the time to make the output more robust to uncertainties. For example, $\alpha = 0.95$ corresponds to a 95% coverage. While some TSFMs, such as Chronos or Sundial, can provide prediction intervals, we observe that naive quantile regression used by these models often fails to provide the desired coverage when forecasting horizons are long (see Sec. 6.4). Consequently, we propose a data-driven conformal prediction method that provides prediction intervals with the desired coverage by looking at the most recent historical residuals. Our approach to ensure the desired coverage is as follows:

Since CarbonX provides 96-hour forecasts ($\hat{y}$) on a daily basis, historical residuals for hours 25—96 are available after a delay. Suppose $H_{1..96}$ is the forecasting horizon for a day d. Then, the most recent historical residuals for $H_{25..48}$ are available from day $d-1$. Similarly, the most recent historical residuals for $H_{73..96}$ are available from day $d-3$. This makes quantifying uncertainty particularly challenging. To quantify effectively and prevent leakage of any future data, we break down the 96-hour forecasting horizon into four forecasting days and compute separate lists of available historical residuals. The conformal layer then looks back at the residuals from the previous available N days to calibrate the current interval. We denote the previous N days as the calibration window, which is a hyperparameter that can be tuned. In our experiments, setting N as 75 provides the best results.

We employ a horizon-specific strategy to obtain the prediction intervals. Specifically, for each hour H of the forecasting window, we look at the residual distribution for that particular hour in the calibration window to get the lower ($q^H_{lo} = \alpha/2$) and upper ($q^H_{hi} = 1-\alpha/2$) quantiles for that hour. Once we obtain the quantiles, we generate prediction intervals $[L^H, U^H]$ for that hour, where:

$$L^H = \hat{y^H} + q^H_{lo}, \quad U^H = \hat{y^H} + q^H_{hi}$$

Thus, the final prediction interval for the whole horizon becomes $[L, U]$, where $L = \hat{y} + q_{lo},\ U = \hat{y} + q_{hi}$.

Since we employ a rolling calibration window for each evaluated day, q_{lo} and q_{hi} reflect the tail behavior of the most recent residuals and provide a data-driven estimate of the uncertainty. Calibrating with the most recent residuals ensures that the intervals adapt to local error patterns and achieve the desired coverage on average.

5 Experimental Methodology

5.1 Datasets and Data Sources

Benchmark Grids. We compare CarbonX against the state-of-the-art on 13 grids spanning the US, Europe, and Australia used in prior studies [25]. Data come from the CarbonCast repository [25] and include two years of hourly carbon intensity measurements from 2020 to 2021. Following prior work, we evaluate forecasting accuracy on the final six months of 2021 for fair comparison.

Global Carbon Intensity Dataset. Electricity Maps recently released hourly carbon intensity data for 350 grids worldwide [11]. For global evaluation, we use data from 214 of these grids covering most major geographical areas, with the number of grids per area listed in Table 5. Each grid provides four years of hourly data from 2021 to 2024, split into training and testing sets with a 70–30 ratio.

Uncertainty Quantification Dataset. We evaluate uncertainty quantification on the same 13 benchmark grids using the two-year datasets from CarbonCast. We provide prediction intervals for the last 6 months in each dataset.

Imputation Dataset. We use 5 minute carbon intensity data from Electricity Maps [11] covering 50 grids, 30 in the United States and 20 in Europe. Each dataset spans 2021 to 2024 and is split 70 to 30 for training and testing. To simulate missing data, 12.5% to 75% of each time series is randomly masked for reconstruction.

5.2 Performance Metrics

Forecasting. We use MAPE to assess forecasting accuracy. Lower values indicate better performance. In addition to the average MAPE over the test period, we report the 90th-percentile MAPE to capture tail forecasting errors of the models.

Uncertainty Quantification. We assess the quality of prediction intervals using *Coverage*, *Normalized Interval Width (NIW)*, and *Continuous Ranked Probability Score (CRPS)*.

Coverage measures the percentage of times the ground truth (y_t) falls within the prediction intervals ($[L_t, U_t]$):

$$\text{Coverage (\%)} = \frac{100}{T} \sum_{t=1}^{T} \mathbf{I}\left[y_t \in [L_t, U_t]\right], \tag{3}$$

where $\mathbf{I}[\cdot]$ denotes the indicator function.

NIW at each time step is the interval width relative to the ground truth: $\text{NIW}_t = \frac{U_t - L_t}{y_t}$. Lower NIW indicates a narrower interval.

CRPS at each timestep is a generalization of the Mean Absolute Error (MAE) for probabilistic forecasting, evaluating the accuracy of an entire predicted probability distribution against the observed value. When the forecast distribution is represented by quantile functions $q(\tau)$ for quantile levels $\tau \in [0, 1]$, the *CRPS* is:

$$\text{CRPS} = 2 \int_0^1 \rho_\tau\big(y - q(\tau)\big)\, d\tau \tag{4}$$

where $\rho_\tau(u)$ is the quantile loss function. We computed CRPS on normalized time series to capture the effects of our approach across the grids. A lower CRPS indicates better calibration.

Imputation. We use Root Mean Square Error (RMSE) on normalized time series to assess imputation accuracy.

	Baselines												CarbonX Zero-Shot Mode									
	EWMA [35]		ARIMA [34]		CC-uni[1]		ECI-uni[1]		CarbonCast [25][2]		EnsembleCI [45][2]		MOMENT [16]		Chronos [2]		TimesFM [6]		Sundial [24]		Time-MoE [36]	
	Mean	90th	Mean	90th	Mean	90th	Mean	90th	Mean	90th	Mean	90th	Mean	90th	Mean	90th	Mean	90th	Mean	90th	Mean	90th
CISO	12.93	28.15	28.83	56.01	12.90	24.08	12.43	23.39	11.45	22.21	**9.17**	18.38	23.38	48.49	24.80	50.04	24.31	48.41	10.79	**17.25**	11.84	24.79
PJM	6.18	13.01	10.24	17.82	5.37	10.81	4.64	7.08	5.29	8.74	**3.41**	**7.17**	6.97	14.73	9.40	17.83	9.36	17.80	4.68	7.98	4.56	9.86
ERCOT	17.33	37.45	29.55	53.21	15.37	28.98	16.04	30.31	11.14	**21.44**	**10.61**	24.69	17.14	37.15	18.59	33.62	18.40	32.69	14.64	25.46	15.45	35.82
ISO-NE	7.40	15.91	11.04	20.67	7.21	13.72	6.97	11.41	6.41	**10.96**	**5.74**	12.98	8.35	17.54	12.20	22.17	12.27	22.24	6.89	11.52	6.79	14.99
BPAT	10.08	20.90	19.33	32.79	12.33	22.33	22.09	34.57	11.22	18.23	10.95	22.42	12.47	25.53	24.17	42.16	24.32	42.30	**9.92**	**14.24**	10.01	20.43
FPL	3.28	7.35	6.66	12.35	3.53	6.63	4.77	7.23	3.15	**5.24**	**2.78**	6.02	4.96	9.91	6.93	13.72	6.85	13.64	2.86	5.31	2.85	6.39
NYISO	8.81	18.08	13.83	25.28	9.67	22.27	9.54	28.32	9.09	28.09	8.75	30.85	9.91	21.01	17.45	35.01	17.64	35.33	8.59	16.19	**7.86**	**15.97**
SE	7.43	16.46	14.09	26.32	6.87	13.25	6.50	10.71	5.78	**9.47**	**5.42**	11.63	7.07	15.84	8.83	17.71	8.74	17.69	6.22	10.37	6.37	13.62
DE	26.25	57.89	42.39	95.86	21.26	43.02	23.27	39.75	**11.72**	**20.90**	12.09	26.82	27.05	58.59	26.20	41.66	25.71	40.67	21.06	40.59	20.46	43.23
PL	8.43	19.27	16.26	33.22	6.81	12.98	6.56	11.94	4.37	**7.49**	**4.23**	9.34	8.72	19.65	7.79	14.78	7.62	13.68	6.48	10.91	6.42	13.26
ES	25.95	52.46	34.69	69.24	21.20	40.35	19.38	30.32	**16.65**	**30.07**	16.96	35.39	25.53	53.99	23.75	43.53	23.54	42.98	20.42	34.99	19.68	39.73
NL	11.30	24.74	22.43	40.87	9.13	16.29	9.57	15.71	8.25	**13.43**	**6.96**	13.72	10.49	23.42	10.48	18.12	10.36	17.92	8.69	15.51	8.78	17.69
AUS-QLD	4.04	9.67	23.79	46.67	3.93	6.09	4.21	5.66	4.46	6.72	3.72	8.99	15.47	33.56	15.76	39.42	15.46	39.30	**3.48**	**4.69**	3.56	8.49
Average	11.49	24.72	21.00	40.80	10.43	20.06	11.23	19.72	8.38	**15.61**	**7.75**	17.57	13.63	29.18	15.87	29.98	15.73	29.59	9.59	16.54	9.59	20.25

(1) CC-uni (resp. ECI-uni) have similar model architecture and capacity as CarbonCast (resp. EnsembleCI), but only uses historical data as input.
(2) CarbonCast and EnsembleCI require energy mix information and train grid-specific models to forecast carbon intensity.

Table 2: *4-day MAPE comparison (best results in bold). TSFMs provide competitive performance in some grids. Sundial (with a 30-day lookback) achieves the highest accuracy in the zero-shot mode, having 9.59% mean and 16.54% tail forecasting MAPE, compared to state-of-the-art mean MAPE of 7.75% (EnsembleCI) and tail MAPE of 15.61% (CarbonCast).*

	CarbonX Grid-Specific Fine-Tuning							
	MOMENT [16]		Chronos [2]		TimesFM [6]		Time-MoE [36]	
	Mean	90th	Mean	90th	Mean	90th	Mean	90th
CISO	12.27	25.07	23.51	43.57	12.47	26.79	12.03	25.54
PJM	5.24	10.84	9.50	17.60	5.21	11.46	5.20	11.32
ERCOT	16.06	33.66	18.90	35.64	15.59	33.10	17.35	37.50
ISO-NE	6.99	14.26	12.69	22.17	7.23	15.43	7.21	15.02
BPAT	10.45	21.68	24.85	41.91	10.45	21.67	10.05	20.61
FPL	3.43	7.00	7.98	14.23	3.03	6.44	3.30	7.14
NYISO	7.95	15.77	16.50	38.02	10.14	24.35	8.39	17.24
SE	6.59	14.85	10.31	18.53	6.19	14.12	7.07	15.31
DE	22.07	41.45	26.27	42.64	22.21	47.46	23.48	53.72
PL	6.92	13.74	8.35	15.73	7.01	15.21	7.55	17.29
ES	20.85	42.12	24.99	44.14	22.84	46.84	23.45	45.49
NL	9.38	18.16	11.47	18.63	8.96	19.59	9.88	20.45
AUS-QLD	4.83	11.82	16.80	38.61	3.69	8.53	3.72	8.73
Average	10.23	20.80	16.32	30.11	10.39	22.38	10.66	22.72

Table 3: CarbonX *accuracy (MAPE %) after grid-specific fine-tuning (Sundial is excluded from evaluation).*

	CarbonX Domain-Specific Fine-Tuning			
	Chronos [2]		Time-MoE [36]	
	Mean	90th	Mean	90th
CISO	22.39	40.64	12.41	27.07
PJM	9.04	17.28	4.97	10.87
ERCOT	19.11	38.57	16.57	35.66
ISO-NE	11.70	21.45	7.09	15.50
BPAT	24.61	41.67	9.83	20.29
FPL	7.15	13.80	3.02	6.66
NYISO	16.20	38.35	8.98	18.56
SE	9.24	18.11	6.71	14.78
DE	25.80	40.76	23.20	49.65
PL	7.70	15.14	7.25	15.58
ES	24.15	43.84	23.42	45.80
NL	10.41	18.16	9.50	19.79
AUS-QLD	16.23	39.71	3.98	9.30
Average	15.67	29.80	10.53	22.27

Table 4: CarbonX *accuracy (MAPE %) after domain-specific fine-tuning (only Chronos and Time-MoE are evaluated).*

5.3 Baselines

Forecasting. We compare CarbonX with two state-of-the-art tools: CarbonCast [25] and EnsembleCI [45]. While EnsembleCI performs better on average, CarbonCast has better tail accuracy (Table 1). So, we include both as baselines. Since both use auxiliary inputs, we also include two univariate variants with similar model architecture and capacity — CC-uni and ECI-uni — that rely only on historical carbon intensity for a fair comparison. Finally, we implement two statistical baselines: EWMA [35] and ARIMA [34].

We could not compare CarbonX with proprietary services such as Electricity Maps [9] and Watttime [40], as their forecasts are not publicly available.

Uncertainty Quantification. Li *et al.* [22] provide prediction intervals for carbon intensity forecasts. Although both our and their works achieve target coverage, directly comparing NIW is difficult since the interval width depends on the point forecast accuracy. Because Li *et al.* extends CarbonCast — which is more accurate than CarbonX— its NIW may be narrower. We keep this comparison for future work. Importantly, CarbonX's modular design allows integration of improved interval methods to produce narrower intervals with the available forecasts.

Imputation. We compare CarbonX's imputation performance against three statistical baselines: (1) Naive interpolation, which fills a missing value using the same hour from the nearest available day, (2) Linear interpolation, and (3) Cubic spline interpolation. Linear and cubic spline interpolation are standard statistical baselines for time series, while the naive approach reflects a common operational heuristic.

5.4 Fine-Tuning Strategies

During forecasting, we fine-tune the TSFMs for one epoch for both grid- and domain-specific strategies. During imputation, we employ both light-weight and full fine-tuning for five epochs. For imputation, we only opt for grid-specific fine-tuning.

6 Forecasting Task Evaluation

We evaluate CarbonX across multiple dimensions. First, we compare its accuracy on 13 benchmark grids and show that CarbonX matches the tail performance of the current state-of-the-art. Next, we show that its forecasts remain reliable for up to 21 days. We then demonstrate that CarbonX can provide accurate forecasts globally, by evaluating across 214 grids. Finally, we assess uncertainty quantification, confirming well-calibrated prediction intervals.

6.1 Comparison against Baselines

Tables 2, 3, and 4 present detailed forecasting accuracy of CarbonX across 13 grids in both zero-shot and fine-tuned modes.

6.1.1 ***Zero-Shot TSFM Performance.*** Statistical models achieve only moderate accuracy. EWMA achieves an average mean MAPE of 11.49%, outperforming ARIMA (21.00%), and performs well on stable grids such as FPL, where it records 3.28% mean MAPE but degrades on volatile grids such as DE (26.25% MAPE). CC-uni (resp. ECI-uni), the univariate variant of CarbonCast (resp. EnsembleCI), improves over the moving average baselines, having mean and tail MAPE of 10.43% and 20.06% (resp. 11.23% and 19.72%) across the 13 grids. Similar to the previous baselines, their accuracy remains high on stable grids but degrades on volatile ones such as DE and ERCOT. CarbonCast and EnsembleCI remain the strongest domain-specific models, leveraging energy mix features and grid-specific architecture and training to achieve mean and 90th-percentile MAPE of 8.38% and 15.61%, and 7.75% and 17.57%, respectively.

Among pre-trained TSFMs, Sundial with a 30-day lookback leads with an average mean and 90th-percentile MAPE of 9.59% and 16.54%. It performs strongly on stable grids such as PJM and FPL while maintaining robustness on higher-variance grids such as DE and ES, where it even outperforms EnsembleCI in tail forecasting accuracy. While Sundial has higher errors than CarbonCast or EnsembleCI in some grids, it consistently outperforms both CC-uni and ECI-uni even without grid-specific training, highlighting the benefits of TSFMs when auxiliary inputs are unavailable.

Among other TSFMs, Time-MoE achieves similar zero-shot mean accuracy (9.59%) but slightly higher tail errors (20.25%). MOMENT follows with 13.63% mean and 29.18% 90th-percentile MAPE, benefiting from strong performance on North American grids but dropping sharply on European ones. Chronos records 15.87% mean and 29.98% 90th-percentile MAPE, and TimesFM records 15.73% mean and 29.59% 90th-percentile MAPE, reflecting weaker adaptation to the heterogeneous temporal dynamics of carbon intensity data.

From the results, we have two additional observations. First, even with univariate inputs, grid-specific models sometimes outperform TSFMs despite smaller capacity. This is because smaller models trained specifically with data from a particular grid may adapt better to that grid and hence have better accuracy in some cases. Second, *newer* TSFMs such as Sundial perform better than *older* TSFMs such as MOMENT even with fewer parameters. One reason for this accuracy difference can be that Sundial is trained on much more data (1T time points) than MOMENT (1.2B time points), and hence has seen more diverse data patterns. However, understanding these differences requires further explainability analysis, which we leave for future work.

We also evaluate other variants of Sundial and Time-MoE to assess the effects of model size on accuracy. Accuracy increases with model size for most grids, showing that larger models can typically capture the patterns better. However, in some grids, smaller models perform better — especially for Time-MoE. The results are available in Appendix B, Table 12. These results underscore the need for more in-depth analysis of model size, architecture, and performance.

		CarbonX (w/ Sundial)		CarbonX (w/ Time-MoE)	
Area	# of grids	Mean	90th	Mean	90th
Africa	28	4.72	7.96	4.76	10.79
Asia	39	3.87	6.91	4.35	9.28
Central America	9	16.78	33.05	16.39	39.18
Europe	39	21.44	39.10	22.92	52.33
North America	73	16.10	28.86	16.90	35.36
Oceania	11	73.57	153.59	82.02	244.53
South America	15	8.64	15.07	10.70	25.45
Total/Average	214	15.82	29.61	17.01	33.37

Table 5: *Zero-shot forecasting accuracy (MAPE %) of* CarbonX *aggregated over 214 grids worldwide. With Sundial,* CarbonX *has an average MAPE of 15.82%.*

Summary. CarbonX can deliver competitive accuracy across diverse grids using *just one model*, only historical carbon intensity data, and no grid-specific modifications. This highlights the strength and versatility of CarbonX, making it a strong candidate for global carbon intensity forecasting. Among the TSFMs, Sundial achieves the highest accuracy, providing robust tail performance that *outperforms EnsembleCI*.

6.1.2 ***Fine-Tuned TSFM performance.*** To enhance performance beyond the zero-shot setting, we evaluate both grid- and domain-specific fine-tuning of the TSFMs.

Grid-specific Fine-Tuning. Table 3 reports the 4-day MAPE when each model is trained on the historical carbon intensity of a single grid and tested only on that grid's hold-out data. Sundial is excluded because it currently lacks a fine-tuning interface.

Grid-specific fine-tuning yields clear gains for some models. MOMENT's average mean MAPE drops from 13.63% to 10.23%, and its 90th-percentile error falls from 29.18% to 20.80%. TimesFM improves in a similar way, with mean MAPE decreasing from 15.73% to 10.39% and the 90th-percentile from 29.59% to 22.38%.

In contrast, Chronos and Time-MoE show little benefit and even slight degradation on average. Chronos rises from 15.87% to 16.32% mean MAPE, while Time-MoE moves from 9.59% to 10.66%.

Domain-Specific Fine-Tuning. In this mode, we train single instances of Chronos and Time-MoE on the combined training data from 214 grids and evaluate their accuracy on the 13 benchmark grids. Table 4 presents the mean and tail MAPE of both models.

Chronos shows marginal improvements on average, with mean MAPE moving from 15.87% in the zero-shot mode to 15.67%. However, the tail MAPE shifts from 29.18% to 29.80%. The larger, more heterogeneous dataset offers small gains on average but introduces variability that limits gains in the tail. On the other hand, although Time-MoE's accuracy improves over its grid-specific version, it still falls behind its zero-shot accuracy. The mean (resp. tail) MAPE is 10.53% (resp. 22.27%), which is lower than 10.66% (resp. 22.72%) achieved with grid-specific fine-tuning, but is higher than 9.59% (resp. 20.25%) achieved in the zero-shot mode.

Interpretation of Fine-Tuning Results. Our experiments show that fine-tuning improves performance for some TSFMs but not others, and in a few cases can slightly degrade accuracy (see Tables 2, 3, and 4). This behavior aligns with architectural and training

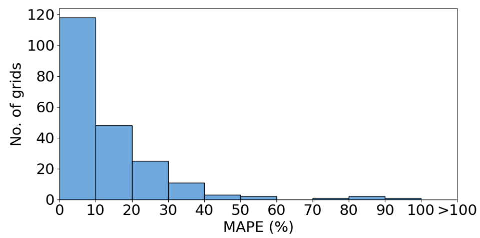


Figure 4: ***MAPE distribution of*** **CarbonX** ***across 214 grids using Sundial as the TSFM. 166 grids have < 20% MAPE on average.***

differences across models, and the limited amount of grid-specific data available for fine-tuning.

For MOMENT and TimesFM, we fine-tune only the final linear forecasting head while freezing the pre-trained backbone. This setup is well-suited for fine-tuning. The backbone captures general temporal patterns from large-scale pre-training, while the lightweight head adjusts these representations to the carbon intensity domain using relatively few parameters. Consequently, even limited grid-specific data provides a strong learning signal, leading to consistent accuracy improvements.

In contrast, Chronos and Time-MoE require *full* fine-tuning in our current implementation. The amount of grid-specific data — both for grid- and domain-specific mode — is negligible compared to the pre-training data, resulting in gradients that are relatively weak and noisy. Chronos relies on quantized token representations, where small shifts in token distributions or transformer parameters can introduce noise rather than meaningful adaptation under limited data. Time-MoE also amplifies this effect due to sparse mixture-of-experts routing, where minor, unbalanced parameter updates can disrupt learned expert assignments without sufficient data to re-stabilize them. Thus, these models often show negligible improvement or slight performance degradation on some grids.

Overall, these results indicate that our simple fine-tuning strategies are not uniformly effective across TSFMs. Architectures supporting lightweight, head-only adaptation perform better in low-data regimes typical of individual grids. In contrast, models relying on complex tokenization or sparse expert routing may require more advanced strategies, such as adapter layers, partial-layer freezing, or larger domain-specific datasets. We keep a detailed study of effective fine-tuning strategies as an important future work.

Summary. Grid-specific fine-tuning enables some TSFMs, such as MOMENT and TimesFM, to close the accuracy gap with CarbonCast and EnsembleCI. Domain-specific fine-tuning provides marginal improvements over grid-specific fine-tuning for some TSFMs such as Chronos. However, a more detailed analysis is required to fully realize the potential of this approach. Interestingly, zero-shot Sundial still outperforms other fine-tuned TSFMs.

6.2 Worldwide Evaluation

Next, we evaluate Carbon*X* across 214 grids spanning all continents except Antarctica, using a generalized zero-shot setting where a single model forecasts all grids. Since Sundial and Time-MoE perform best in zero-shot mode on the 13 benchmark grids, we evaluate the accuracy by plugging both these TSFMs into our task module.

	D1	D4	Drop_4	D21	Drop_{21}
CISO	8.03	10.96	2.93	12.38	4.35
PJM	3.12	4.76	1.64	5.89	2.77
ERCOT	11.63	15.00	3.37	15.84	4.21
ISO-NE	4.91	6.99	2.08	9.90	4.99
BPAT	6.93	9.43	2.50	12.97	6.04
FPL	2.09	2.97	0.88	4.19	2.10
NYISO	5.18	8.96	3.78	15.46	10.28
SE	4.26	6.28	2.02	7.91	3.65
DE	14.83	21.15	6.32	24.79	9.96
PL	4.31	6.58	2.27	7.34	3.03
ES	11.84	21.47	9.63	23.87	12.03
NL	6.06	8.88	2.82	9.98	3.92
AUS-QLD	2.81	3.59	0.78	4.05	1.24
Average	6.62	9.77	3.16	11.89	5.27

Table 6: ***Extended forecasting accuracy (MAPE) and degradation (absolute difference in MAPE) compared to D1 forecasts. Sundial with a 15-day lookback provides the best results.***

Table 5 summarizes the number of grids evaluated in each geographical area and the forecasting accuracy of both models, reporting mean and 90th-percentile MAPE. Sundial delivers strong performance globally, with 15.82% mean MAPE and 29.61% tail MAPE across 214 grids. Time-MoE follows Sundial performance closely, with 17.01% mean and 33.37% tail MAPE.

Figure 4 and Table 5 further show the global error distribution for Carbon*X* using Sundial as the TSFM. Most grids have a low forecasting error, with 166 out of 214 grids having a mean MAPE below 20%. Grids in Asia and Africa — often dominated by fossil fuels and thus more stable carbon intensity profiles — have highly accurate forecasts, with an average MAPE of less than 5%. In contrast, grids in the Americas and Europe, with higher renewable penetration and greater volatility, show higher errors. Nevertheless, 96-hour forecasting errors remain below 20% for most grids. The only exception is Oceania, where Carbon*X* has a very high forecasting error. This high error rate is primarily due to three grids in Tasmania and its neighboring islands, where frequent large swings in renewable share make the grids highly volatile and, hence, forecasting significantly difficult. If these three grids are excluded, the mean and tail MAPE in Oceania decrease to 20.69% and 36.20%, respectively, similar to other geographical areas. Consequently, the global mean and tail MAPE across 211 grids also decrease to 12.99% and 23.40%, respectively, further underscoring Carbon*X*'s strength.

We observe similar trends when using Time-MoE as the TSFM.

Carbon*X* achieves strong global zero-shot forecasting using only historical carbon intensity data: 77.5% of the 214 grids attain mean MAPE below 20%. Notably, over 35% of these grids lack electricity-mix data, making prior works infeasible and Carbon*X* the *only viable option*. Even when such data exists, earlier methods would require hundreds of grid-specific models, whereas Carbon*X* delivers global coverage with *one generalized model*. This makes Carbon*X* both scalable and practical for computational decarbonization.

6.3 Extended Forecasting

TSFMs have shown strong long-horizon forecasting across various domains. We therefore hypothesize that TSFMs can forecast carbon intensity accurately for up to several weeks. In contrast, existing carbon forecasting tools are typically designed and evaluated for horizons up to 96 hours. Their accuracy beyond this horizon is

	Sundial		CarbonX		
	MAPE	Coverage (%)	Coverage (%)	NIW (%)	CRPS (norm.)
CISO	10.79	57.6	96.3	68.5	0.28
PJM	4.68	68.6	96.4	29.0	0.27
ERCOT	14.64	59.6	95.0	67.7	0.39
ISO-NE	6.89	62.2	96.9	41.6	0.34
BPAT	9.92	58.8	96.1	53.1	0.31
FPL	2.86	64.3	95.3	21.6	0.23
NYISO	8.59	59.3	96.1	70.1	0.34
SE	6.22	60.7	95.3	36.0	0.36
DE	21.06	62.1	96.7	103.6	0.42
PL	6.48	62.6	96.6	36.9	0.43
ES	20.42	60.3	94.8	96.3	0.43
NL	8.69	63.8	97.3	49.6	0.45
AUS-QLD	3.48	52.2	98.1	30.8	0.13
Average	9.59	60.9	96.2	54.2	0.34

Table 7: *Uncertainty quantification:* CarbonX *provides more than the desired 95% coverage with moderately wide intervals.*

expected to decrease substantially due to their limited model capacity. Further, these tools rely on auxiliary information such as weather forecasts, which are only available for *up to 16 days* in many regions [27]. This creates a practical bottleneck for existing tools to provide extended carbon intensity forecasts.

In this section, we assess CarbonX's extended forecasting capabilities by evaluating its zero-shot accuracy for up to 3 weeks (21 days). Since Sundial achieves the highest accuracy while providing zero-shot 4-day forecasts, we use Sundial as the TSFM for this experiment. We vary the lookback window from 1 to 30 days in the past, and extend the forecasting horizon from 1 to 21 days. While a 30-day lookback achieves the highest accuracy for 4-day forecasts, a 15-day lookback yields the lowest accuracy degradation for 21-day forecasts across the grids. We posit that this is because while shorter lookbacks cannot capture the longer temporal patterns, longer lookbacks may incorporate noise or infrequent events when forecasting longer horizons. A 15-day lookback achieves the best balance, capturing the patterns without amplifying noise. More detailed ablations of lookback length and forecast horizon are provided in Appendix C, Table 11.

Table 6 summarizes the results across the 13 benchmark grids. Sundial's has a high day-ahead accuracy, with an average MAPE of 6.62%. The error growth over time is modest: the 4-day horizon averages 9.77% MAPE, only 3.16% points above day-ahead, and the 21-day horizon averages 11.89%, just 5.27 points higher. This limited increase underscores CarbonX 's ability to maintain reliable performance even for longer forecasts.

The degradation depends on the carbon intensity volatility. Stable grids (*e.g.*, PJM, FPL) show exceptional accuracy, having less than 3% degradation and MAPE below 6% even at 21 days. More volatile grids (*e.g.*, Germany (DE), Spain (ES)) show larger degradation: DE rising from 14.83% to 24.79% and ES from 11.84% to 23.87%. Still, the degradation remains manageable over three weeks. Notably, most of the degradation occurs in the first four days; then the degradation rate decreases. For example, in DE, accuracy decreases from 14.83% on day one to 21.15% by day 4. However, the accuracy only degrades by 3.64% over the remaining forecasting horizon.

Summary. CarbonX delivers strong zero-shot accuracy with *minimal extended-horizon degradation* (5.27% accuracy drop across 21 days). Given an appropriate historical context, CarbonX provides reliable forecasts for applications that require horizons well beyond conventional day-ahead planning.

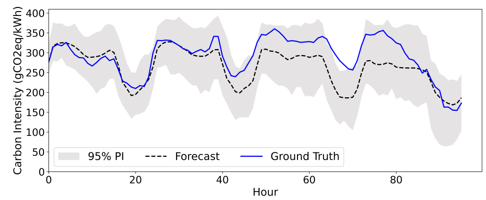


Figure 5: CarbonX *forecasts and prediction intervals during a 96-hour forecasting period in the CISO grid. The intervals provide desired coverage around the ground truth.*

6.4 Uncertainty Quantification

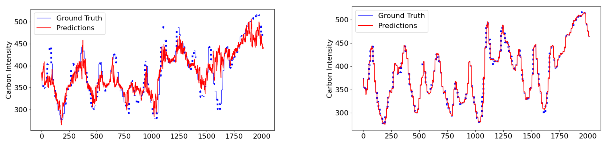
We now evaluate how CarbonX quantifies forecast uncertainty using its lightweight conformal layer. Figure 5 shows 96-hour prediction intervals for the California (CISO) grid. The intervals are constructed from the most recent historical residuals, which ensures that the intervals closely follow the ground truth trajectory and provide the desired coverage (95%). This highlights the robustness of the conformal layer and the reliability CarbonX provides to downstream optimization tasks.

Table 7 reports the empirical coverage provided by Sundial and CarbonX across the 13 benchmark grids. Sundial provides an average coverage of 60.9% (ranging between 52.2% and 68.6%), and consistently fails to meet the desired 95% coverage. In contrast, CarbonX attains over 95% coverage both day-wise and across the full 96-hour horizon when averaged across grids. At an individual grid level, CarbonX provides an average coverage above 95% in all grids except Spain (ES), where the coverage falls slightly below 95% due to increased uncertainty in forecasting beyond the 24-hour horizon (see Appendix D, Table 13).

Table 7 also reports the Normalized Interval Widths (NIWs) and the Continuous Ranked Probability Scores (CRPSs). NIW varies from 21.6% in Florida (FPL) to 103.6% in Germany (DE), with an average of 54.2% across grids, indicating that interval widths are slightly more than half the magnitude of ground truth values on average. NIW generally correlates positively with forecast error: grids with lower point forecast accuracy tend to require wider intervals. This is expected, as greater uncertainty and hence residual variance require wider intervals to maintain coverage guarantees.

CRPS varies from 0.13 to 0.45 across grids, with an average of 0.34, indicating that prediction intervals are, on average, 0.34 standard deviations away from ground truth. Grids with lower forecast errors typically achieve lower CRPS scores, reflecting narrower intervals. However, some grids, such as Poland (PL), show higher CRPS despite low MAPE due to sudden spikes and drops in carbon intensity, which require wider intervals to maintain target coverage.

Overall, CarbonX produces well-calibrated intervals across most grids. However, intervals are relatively wide in certain cases (e.g., Germany), limiting usability in those grids. Reducing interval width while preserving coverage would require more advanced conformal methods [43]. Notably, our modular design enables seamless integration of such advanced techniques with minimal changes, and we keep that as future work.



(a) *Zero-shot (ERCOT).* (b) *Fine-tuned (ERCOT).*

Figure 6: CarbonX *provides good estimates of missing data.*

Summary. Carbon$X$'s conformal layer effectively quantifies forecast uncertainty by producing adaptive, well-calibrated prediction intervals in most grids while providing the desired 95% coverage. These results enable Carbon$X$ to be a practical tool with reliable and robust forecasts for downstream applications that require carbon-aware decision-making under uncertainty.

7 Imputation Task Evaluation

Finally, we evaluate carbon intensity imputation by masking random fixed-length patches of the data to simulate scenarios where some of the data is missing. A masking percentage of 50% means that we simulate a scenario where half of the data is missing. We vary this masking percentage from 12.5% to 75% during our experiments, and evaluate for 50 grids (30 in the US and 20 in Europe).

Figure 6 visualizes Carbon$X$'s imputation performance for the Texas grid (ERCOT), both in zero-shot and fine-tuned modes. The blue segments indicate the ground truth, with gaps denoting missing data. Carbon$X$ estimates follow the ground truth in both modes, with fine-tuned estimates being more accurate.

Table 8 details the imputation accuracy of Carbon$X$ in different modes and compares Carbon$X$ against three statistical baselines. Since we average the RMSE across multiple grids that have different carbon intensity magnitudes, we first normalize the values for fair comparison. The naive and cubic spline baselines perform modestly, with RMSE between 0.83 and 0.94 across the masking percentages. Zero-shot Carbon$X$ is substantially better, having 0.18 RMSE (resp. 0.52) with 12.5% (resp. 75%) missing data. With lightweight fine-tuning, the accuracy improves further: RMSE reduces to 0.41 for 75% missing data. Interestingly, we observed that the linear interpolation method often performs better in terms of RMSE. We observed that while Carbon$X$ may provide better estimates when data is missing for longer periods, linear interpolation outperforms Carbon$X$ for shorter gaps or when the missing data has a monotonic trend (see Appendix E for details). However, the linear method connects endpoints with straight lines and can miss temporal patterns. Thus, in pattern-sensitive applications, lightly fine-tuned Carbon$X$ may be preferable over linear interpolation.

To further improve the accuracy, we also fully fine-tune Carbon$X$ per grid. This mode outperforms all baselines, achieving an RMSE of 0.07 (resp. 0.24) with a 12.5% (resp. 75%) masking percentage, a 1.2—3.9× improvement. We note that full fine-tuning yields grid-specific models, and hence, may be challenging to deploy at scale; however, we report the values so that users can decide the mode per their requirements.

Summary. Carbon$X$ provides moderate to high imputation accuracy for missing data. While linear interpolation can provide better RMSE than lightweight fine-tuning, Carbon$X$ preserves the patterns better on average. Fully fine-tuned version achieves the best results overall, outperforming all baselines by 1.2—3.9×.

	12.5%	25%	37.5%	50%	62.5%	75%
FT (full)	**0.07**	**0.08**	**0.11**	**0.14**	**0.18**	**0.24**
FT (head)	0.19	0.21	0.24	0.28	0.34	0.41
ZS	0.18	0.23	0.30	0.36	0.44	0.52
Naive	0.83	0.85	0.86	0.89	0.91	0.94
Cubic-Spline	0.86	0.86	0.86	0.86	0.87	0.87
Linear	0.13	0.14	0.15	0.17	0.21	0.29

Table 8: Normalized imputation RMSE for different methods at various masking percentages.

8 Discussions and Future Work

Effectiveness of carbon intensity for grid decarbonization. Carbon-aware optimizations have leveraged grid carbon intensity to reduce emissions by shifting computational demand spatially or temporally [1, 29, 37, 41, 44]. In particular, optimizations that shift demand temporally often require forecasts beyond 96 hours (*e.g.*, scheduling cloud data backups) and are sensitive to forecast accuracy, as mispredicted low-carbon periods can increase emissions [22]. Accordingly, Carbon$X$ is designed to provide accurate, global, and long-horizon carbon intensity forecasts, along with prediction intervals that capture uncertainty in point forecasts.

Recent optimizations [32, 46, 48] are also beginning to incorporate additional signals (*e.g.*, curtailment), alongside carbon intensity. Carbon$X$'s flexible design enables extension to such signals, and we plan to support forecasting of these signals in the future.

Covariate support. Many grids (*e.g.*, Africa) still do not publish electricity mix data, limiting the applicability of existing tools. Carbon$X$ is designed as a general tool that can operate in such environments. While grid mix may be unavailable, other information, such as weather forecasts, is more prevalent and can improve forecasting accuracy [25, 45]. However, we currently restrict Carbon$X$ to univariate inputs since the supported TSFMs do not yet handle covariates. Including covariate support is an important future direction. As TSFMs evolve, we plan to integrate newer TSFMs with covariate support. Due to its modular architecture, Carbon$X$ can seamlessly integrate such TSFMs, further improving its accuracy.

9 Conclusion

We present Carbon$X$, an open-sourced tool that leverages the versatility and strong performance of TSFMs to perform carbon intensity forecasting and imputation globally. Carbon$X$ achieves strong zero-shot performance across 214 grids globally using only historical carbon data, with a mean MAPE of 15.62%. It supports forecasting up to 21 days, provides well-calibrated prediction intervals with desired coverage, and performs accurate imputation, making it a practical and easy-to-use tool for carbon-aware applications. We release code and datasets to foster future sustainability research.

Acknowledgments

This work is supported in part by NSF grants CNS-2105494, CNS-2325956, 23091241, 2213636, 2211302, and 2211888, DOE grant DEEE0010143, and a grant from VMware. Mani Srivastava was also partially supported by the Mukund Padmanabhan Term Chair at UCLA.

References

[1] Bilge Acun, Benjamin Lee, Fiodar Kazhamiaka, Kiwan Maeng, Udit Gupta, Manoj Chakkaravarthy, David Brooks, and Carole-Jean Wu. 2023. Carbon Explorer: A Holistic Framework for Designing Carbon Aware Datacenters. In *ACM ASPLOS*.

[2] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. 2024. Chronos: Learning the Language of Time Series. *arXiv preprint arXiv:2403.07815* (2024).

[3] Hamzeh Asgharnezhad, Pegah Tabarisaadi, Abbas Khosravi, and Hao Quan. 2024. Time Series Foundation Models for Load Forecasting: A Comprehensive Analysis. https://doi.org/10.2139/ssrn.5067419.

[4] Neeraj Dhanraj Bokde, Bo Tranberg, and Gorm Bruun Andresen. 2021. Short-term CO2 emissions forecasting based on decomposition approaches and its impact on electricity market scheduling. *Applied Energy* 281 (2021), 116061.

[5] CAISO. 2025. California ISO. https://www.caiso.com/.

[6] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. 2024. A decoder-only foundation model for time-series forecasting. In *ICML*.

[7] Department of Business, Energy and Industrial Strategy. 2021. Greenhouse Gas Reporting: Conversion Factors 2021. https://www.gov.uk/government/publications/greenhouse-gas-reporting-conversion-factors-2021.

[8] EIA. 2025. Hourly Electric Grid Monitor. https://www.eia.gov/electricity/gridmonitor/dashboard/electric_overview/US48/US48.

[9] Electricity Maps. 2025. Electricity Maps. https://www.electricitymaps.com/.

[10] Electricity Maps. 2025. Estimation Methods. https://github.com/electricitymaps/electricitymaps-contrib/wiki/Estimation-methods.

[11] Electricity Maps. 2025. Granular Historical Electricity Data. https://portal.electricitymaps.com/datasets.

[12] ENTSO-E. 2025. ENTSO-E Dashboard. https://transparency.entsoe.eu/dashboard/show.

[13] European Court of Auditors. 2025. Making the EU Electricity Grid Fit for Net-Zero Emissions. https://www.eca.europa.eu/ECAPublications/RV-2025-01/RV-2025-01_EN.pdf.

[14] Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. 2023. TimeGPT-1. *arXiv preprint arXiv:2310.03589* (2023).

[15] Anubha Goel, Puneet Pasricha, Martin Magris, and Juho Kanniainen. 2025. Foundation Time-Series AI Model for Realized Volatility Forecasting. *arXiv preprint arXiv:2505.11163* (2025).

[16] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. 2024. MOMENT: A Family of Open Time-series Foundation Models. In *ICML*.

[17] IEA. [n. d.]. Africa. https://www.iea.org/regions/africa/electricity.

[18] David Irwin, Prashant Shenoy, Mohammad Hajiesmaili, Walid A Hanafy, Jimi Oke, Ramesh Sitaraman, Yuvraj Agarwal, Geoff Gordon, Zico Kolter, Deepak Rajagopal, et al. 2025. A Vision for Computational Decarbonization of Societal Infrastructure. *IEEE Internet Computing* (2025).

[19] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. 2023. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. *arXiv preprint arXiv:2310.01728* (2023).

[20] Suprabhath Kalahasti, Benjamin Faucher, Boxuan Wang, Claudio Ascione, Ricardo Carbajal, Maxime Enault, Christophe Vincent Cassis, Titouan Launay, Caroline Guerrisi, Pierre-Yves Boëlle, et al. 2025. Foundation time series models for forecasting and policy evaluation in infectious disease epidemics. *medRxiv* (2025), 2025–02.

[21] Kenneth Leerbeck, Peder Bacher, Rune Grønborg Junker, Goran Goranović, Olivier Corradi, Razgar Ebrahimy, Anna Tveit, and Henrik Madsen. 2020. Short-term forecasting of CO2 emission intensity in power grids by machine learning. *Applied Energy* 277 (2020), 115527.

[22] Amy Li, Sihang Liu, and Yi Ding. 2024. Uncertainty-aware decarbonization for datacenters. *ACM SIGENERGY Energy Informatics Review* 4, 5 (2024), 141–147.

[23] Rui Liang, Yang Deng, Donghua Xie, Fang He, and Dan Wang. 2024. Enabling Time-series Foundation Model for Building Energy Forecasting via Contrastive Curriculum Learning. *arXiv preprint arXiv:2412.17285* (2024).

[24] Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. 2025. Sundial: A family of highly capable time series foundation models. *arXiv preprint arXiv:2502.00816* (2025).

[25] Diptyaroop Maji, Prashant Shenoy, and Ramesh K Sitaraman. 2022. CarbonCast: Multi-Day Forecasting of Grid Carbon Intensity. In *ACM BuildSys*. 198–207.

[26] Ozan Baris Mulayim, Pengrui Quan, Liying Han, Xiaomin Ouyang, Dezhi Hong, Mario Bergés, and Mani Srivastava. 2024. Are Time Series Foundation Models Ready to Revolutionize Predictive Building Analytics?. In *ACM BuildSys*.

[27] National Weather Service. 2025. Forecast Models. https://www.weather.gov/hnx/models.

[28] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. 2023. ClimaX: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343* (2023).

[29] Ana Radovanović, Ross Koningstein, Ian Schneider, Bokan Chen, Alexandre Duarte, Binz Roy, Diyue Xiao, Maya Haridasan, Patrick Hung, Nick Care, Saurav Talukdar, Eric Mullen, Kendal Smith, MariEllen Cottman, and Walfredo Cirne. 2023. Carbon-Aware Computing for Datacenters. *IEEE Transactions on Power Systems* 38, 2 (2023), 1270–1280.

[30] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.

[31] Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Hena Ghonia, Rishika Bhagwatkar, Arian Khorasani, Mohammad Javad Darvishi Bayazi, George Adamopoulos, Roland Riachi, Nadhir Hassen, et al. 2023. Lag-Llama: Towards Foundation Models for Probabilistic Time Series Forecasting. *arXiv preprint arXiv:2310.08278* (2023).

[32] Tella Rajashekhar Reddy, Rohan Gandhi, Anjaly Parayil, Chaojie Zhang, Mike Shepperd, Liangcheng Yu, Jayashree Mohan, Srinivasan Iyengar, Shivkumar Kalyanaraman, Debopam Bhattacherjee, et al. 2025. AI Greenferencing: Routing AI Inferencing to Green Modular Data Centers with Heron. *arXiv preprint arXiv:2505.09989* (2025).

[33] Steffen Schlömer, Thomas Bruckner, Lew Fulton, Edgar Hertwich, Alan McKinnon, Daniel Perczyk, Joyashree Roy, Roberto Schaeffer, Ralph Sims, Pete Smith, and Ryan Wiser. 2014. Annex III: Technology-Specific Cost and Performance Parameters. https://archive.ipcc.ch/pdf/assessment-report/ar5/wg3/ipcc_wg3_ar5_annex-iii.pdf.

[34] Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and Statistical Modeling with Python. *SciPy* 7, 1 (2010), 92–96.

[35] Mihir Shenoy and Walid A Hanafy. 2024. Home Decarbonizer: Greening Household Energy Consumption Using Temporal Shifting. In *IEEE URTC*. IEEE.

[36] Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. 2024. Time-moe: Billion-scale time series foundation models with mixture of experts. *arXiv preprint arXiv:2409.16040* (2024).

[37] John Thiede, Noman Bashir, David Irwin, and Prashant Shenoy. 2023. Carbon Containers: A System-level Facility for Managing Application-level Carbon Emissions. In *ACM SoCC*.

[38] US Energy Information Administration. 2022. EIA statement on the Hourly Electric Grid Monitor. https://www.eia.gov/pressroom/releases/press514.php.

[39] U.S. Environmental Protection Agency. 2025. Sources of Greenhouse Gas Emissions. https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions.

[40] Watttime. 2025. Watttime Official Website. https://www.watttime.org/.

[41] Philipp Wiesner, Ilja Behnke, Dominik Scheinert, Kordian Gontarska, and Lauritz Thamsen. 2021. Let's wait awhile: how temporal workload shifting can reduce carbon emissions in the cloud. In *ACM Middleware*.

[42] World Resources Institute. 2015. GHG Protocol Scope 2 Guidance: Executive Summary. https://ghgprotocol.org/sites/default/files/2022-12/Scope2_ExecSum_Final.pdf.

[43] Chen Xu and Yao Xie. 2023. Sequential predictive conformal inference for time series. In *ICML*. PMLR.

[44] Kaiqiang Xu, Decang Sun, Han Tian, Junxue Zhang, and Kai Chen. 2025. {GREEN}: Carbon-efficient Resource Scheduling for Machine Learning Clusters. In *22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI 25)*. 999–1014.

[45] Leyi Yan, Linda Wang, Sihang Liu, and Yi Ding. 2025. EnsembleCI: Ensemble Learning for Carbon Intensity Forecasting. In *ACM E-Energy*.

[46] Fan Yang and Andrew A Chien. 2016. ZCCloud: Exploring wasted green power for high-performance computing. In *2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 1051–1060.

[47] Xiaoyang Zhang and Dan Wang. 2023. A GNN-based Day Ahead Carbon Intensity Forecasting Model for Cross-Border Power Grids. In *ACM e-Energy*.

[48] Jiajia Zheng, Andrew A Chien, and Sangwon Suh. 2020. Mitigating curtailment and carbon emissions through load migration between data centers. *Joule* 4, 10 (2020), 2208–2222.

A `Carbon`*X* API Suite

Table 9 lists the APIs currently supported by `Carbon`*X* that other applications can use to obtain carbon intensity data and forecasts. `Carbon`*X* has a concise suite of APIs for easy integration. In the current version, downstream applications can get carbon intensity data (`get_ci_historical()`) and forecasts (`get_ci_forecasts()`) for 214 grids between the years 2021 and 2024. We are updating the implementation to provide support for fetching data and forecasts in real time. Users can also query the accuracy of forecasts from any previous date (`get_forecast_accuracy()`), and choose the model with the best accuracy or task requirements (`set_model()`).

Function	Parameters	Returns	Description
`get_ci_historical()`	grid, date	Hourly CI	Get historical carbon intensity (CI) for the specified grid and date.
`get_ci_forecasts()`	grid, date, horizon, PI	Hourly CI forecasts	Get CI forecasts for the given horizon hours (max 96) on the specified date. Returns prediction intervals if `PI` is True.
`get_forecast_accuracy()`	grid, date, horizon	MAPE (%)	Get CI forecasting accuracy for the given horizon hours (max 96) on the specified date. Returns error if ground truth is not available.
`get_missing_ci_data()`	CI series, mask	Imputed CI	Get imputed CI data from the specified series. A mask value of 0 indicates missing data.
`get_supported_grids()`	–	Grid list	List of currently supported grids.
`set_model()`	model_name, mode	Success/ Failure	Specify the TSFM model and mode (ZS, FT).

Table 9: CarbonX API Functions to access worldwide carbon intensity data and forecasts. (ZS: Zero-Shot; FT: Fine-Tuned)

TSFM	Variant	Parameters
MOMENT [16]	Large (*AutonLab/MOMENT-1-large*)	385M
Chronos [2]	Large (*amazon/chronos-t5-large*)	710M
TimesFM [6]	*TimesFM-2.0-500M*	500M
Sundial [24]	*thuml/sundial-base-128m*	128M
Time-MoE [36]	*Maple728/TimeMoE-50M*	50M

Table 10: *Supported TSFMs with variants and parameter sizes.*

Finally, users can also use CarbonX to estimate any missing carbon intensity data, or get data at a finer temporal resolution than the hourly granularity if required (`get_missing_ci_data()`).

B Zero-Shot Variant Analysis

Table 10 shows the various TSFMs currently supported by CarbonX, their variants we used in our evaluation, and their respective parameter counts. The number of parameters ranges from 50M in Time-MoE to 710M in Chronos. Since Sundial and Time-MoE have the highest zero-shot accuracy, we further examine CarbonX zero-shot performance using alternative Sundial and Time-MoE variants with different parameter sizes. This analysis highlights how model scale influences forecasting accuracy. Table 12 reports 4-day MAPE for the 13 benchmark grids along with overall mean and 90th-percentile results.

Table 2 in the main text evaluates the Sundial-Base-128M and Time-MoE-50M models. In contrast, Table 12 presents results for the Sundial-Timer-84M and Time-MoE-200M variants.

For Sundial, the 84M variant records a mean MAPE of 9.97% and a 90th-percentile of 21.62%. This is slightly higher than the 128M base model in Table 2, showing that additional parameters improve the modeling of fine-grained dynamics. Performance remains strong on stable grids such as FPL with 3.33% MAPE, but the gap widens on more variable grids such as ERCOT with 16.19%, where larger models capture variability more effectively. The Timer configuration, while efficient and competitive, highlights the trade-off between model size and forecasting accuracy.

For the Time-MoE model, the larger 200M variant achieves a mean MAPE of 9.54% and a 90th-percentile MAPE of 20.02%, showing slight improvement on average over the smaller 50M version in Table 2. The additional parameters allow the model to capture more complex temporal patterns in carbon intensity, producing more accurate zero-shot forecasts in some grids such as California (CISO) and Washington (BPAT). However, in some other grids, such as Spain (ES), accuracy decreases slightly, underscoring the need for more detailed analysis and ablation studies.

Overall, we find that scaling model size enhances zero-shot accuracy on average, with gains more pronounced for the Sundial model. These findings suggest that while larger configurations increase the model footprint, they can typically reduce forecasting errors. Thus, practitioners should decide on the right model variant based on their requirements and resource availability.

C Extended Forecasting Details

Table 11 provides a detailed view of Sundial's mean MAPE as a function of both the input (lookback) window and the forecast horizon across four representative grids: California (CISO), Texas (ERCOT), Germany (DE), and Sweden (SE).

Longer lookback windows provide better accuracy at shorter horizons. In every grid, a 30-day input consistently yields the lowest error for day-ahead (D1) forecasts. For example, CISO records 8.01% MAPE at D1 with a 30-day history compared with 10.33% when the input is just the past day. Similar patterns appear in ERCOT at 11.55% versus 13.39%, DE at 14.17% versus 18.00%, and SE at 4.17% versus 5.03%, showing that long histories capture richer context for short-term prediction.

As the horizon extends, accuracy decreases regardless of the lookback window length. Similar to the case of shorter horizons, longer lookbacks achieve lower errors. However, interestingly, for most grids, a 15-day lookback window starts achieving better accuracy than a 30-day window from day 7 onwards, as evident in CISO, DE, and SE. For example, while day four forecasting error in CISO is 10.96% with a 15-day lookback compared to 10.79% with a 30-day lookback, day 8 errors are the same, and at day 21, a 15-day lookback has a 12.38% error compared to 12.70% in the other case. These results confirm that, while additional historical context helps capture slow seasonal patterns needed for multi-week forecasts, excessively longer windows may amplify the noise in the data.

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20	D21
California (CISO) Grid																					
1*24	10.33	12.03	13.14	14.00	14.60	14.98	15.26	15.34	15.49	15.45	15.42	15.38	15.33	15.24	15.33	15.31	15.42	15.53	15.65	15.74	15.78
2*24	9.52	11.21	12.39	13.30	13.82	14.13	14.33	14.44	14.46	14.51	14.37	14.31	14.24	14.17	14.23	14.31	14.43	14.45	14.62	14.68	14.82
4*24	8.98	10.94	12.05	12.81	13.19	13.40	13.51	13.56	13.55	13.40	13.37	13.28	13.16	13.21	13.31	13.39	13.45	13.57	13.67	13.80	13.85
7*24	8.78	10.44	11.36	11.93	12.28	12.34	12.52	12.58	12.45	12.38	12.37	12.38	12.42	12.42	12.52	12.61	12.69	12.80	12.91	13.02	13.00
15*24	8.03	9.49	10.37	10.96	11.34	11.48	11.63	11.82	11.79	11.83	11.87	11.93	11.98	11.97	12.00	12.01	12.10	12.24	12.20	12.29	12.38
30*24	8.01	9.49	10.27	10.79	11.16	11.44	11.62	11.82	11.88	11.95	11.89	12.03	12.06	12.09	12.24	12.25	12.35	12.46	12.55	12.58	12.70
Texas (ERCOT) Grid																					
1*24	13.39	15.47	16.62	17.49	18.20	18.79	19.21	19.34	19.35	19.24	19.13	19.14	19.16	19.12	19.17	19.21	19.26	19.25	19.29	19.31	19.39
2*24	13.55	15.24	16.41	17.30	17.94	18.44	18.61	18.78	18.73	18.63	18.51	18.55	18.48	18.47	18.54	18.53	18.59	18.65	18.70	18.75	18.80
4*24	12.75	14.86	16.07	16.79	17.43	17.72	17.81	17.83	17.79	17.71	17.62	17.57	17.61	17.53	17.56	17.62	17.60	17.64	17.67	17.73	17.78
7*24	12.48	14.58	15.48	16.17	16.33	16.71	16.82	16.80	16.79	16.66	16.54	16.61	16.61	16.65	16.65	16.70	16.67	16.71	16.79	16.69	16.74
15*24	11.63	13.49	14.60	15.00	15.33	15.68	15.75	15.74	15.70	15.74	15.73	15.74	15.77	15.66	15.69	15.76	15.71	15.74	15.71	15.81	15.84
30*24	11.55	13.32	14.23	14.64	15.11	15.22	15.39	15.45	15.49	15.34	15.38	15.48	15.42	15.54	15.49	15.57	15.67	15.63	15.72	15.65	15.70
Germany (DE) Grid																					
1*24	18.00	22.14	24.27	25.89	26.77	27.27	27.14	27.34	27.43	27.51	27.29	27.34	27.55	27.84	28.03	28.25	28.42	28.56	28.61	28.83	28.96
2*24	16.79	21.50	23.66	25.10	25.72	25.97	25.92	26.03	26.13	25.98	26.04	26.15	26.42	26.64	26.83	26.97	27.16	27.23	27.41	27.58	27.78
4*24	16.02	20.69	22.30	23.47	23.78	23.88	24.05	24.02	23.89	23.99	24.17	24.35	24.60	24.88	25.08	25.18	25.23	25.50	25.67	25.80	26.08
7*24	15.04	19.13	20.75	21.49	22.12	22.25	22.41	22.57	22.56	22.97	23.12	23.40	23.60	23.72	24.00	24.05	24.17	24.50	24.68	24.80	25.02
15*24	14.83	18.26	20.43	21.15	21.67	21.94	22.19	22.33	22.53	22.74	23.21	23.29	23.35	23.76	23.84	24.06	24.24	24.34	24.50	24.64	24.79
30*24	14.17	18.20	20.07	21.06	21.82	21.84	22.24	22.47	22.69	22.89	22.93	23.42	23.50	23.83	24.02	24.25	24.36	24.41	24.70	24.79	25.07
Sweden (SE) Grid																					
1*24	5.03	6.30	6.94	7.26	7.55	7.72	7.86	7.97	8.13	8.22	8.29	8.35	8.43	8.49	8.59	8.66	8.76	8.86	8.92	8.99	9.00
2*24	4.89	6.10	6.65	7.01	7.23	7.41	7.53	7.69	7.80	7.85	7.96	8.02	8.11	8.16	8.24	8.31	8.45	8.55	8.62	8.65	8.63
4*24	4.55	5.70	6.19	6.47	6.66	6.83	7.00	7.16	7.24	7.42	7.50	7.56	7.61	7.72	7.84	7.95	8.06	8.14	8.18	8.22	8.28
7*24	4.41	5.45	6.09	6.31	6.56	6.69	6.81	6.97	7.06	7.23	7.36	7.41	7.46	7.59	7.75	7.83	7.91	7.98	8.00	8.05	8.04
15*24	4.26	5.39	5.97	6.28	6.56	6.70	6.89	7.02	7.12	7.22	7.32	7.43	7.52	7.56	7.67	7.74	7.84	7.86	7.91	7.92	7.91
30*24	4.17	5.24	5.85	6.22	6.52	6.74	6.93	7.05	7.22	7.30	7.36	7.46	7.60	7.65	7.71	7.82	7.88	7.96	7.95	8.03	7.98

Table 11: *Mean MAPE of* CarbonX *(Sundial) for varying input lengths (rows) and prediction horizons (D1 to D21, columns) across four representative grids.*

	Carbon*X* Zero-Shot Mode			
	Sundial [24] (84M)		Time-MoE [36] (200M)	
	Mean	90th	Mean	90th
CISO	11.35	24.14	11.12	23.35
PJM	4.87	10.41	4.69	10.42
ERCOT	16.19	35.08	15.53	35.82
ISO-NE	6.96	14.81	6.75	14.24
BPAT	9.88	20.11	9.64	19.75
FPL	3.33	7.72	2.77	6.26
NYISO	9.02	19.55	7.64	15.30
SE	6.29	14.67	6.52	14.17
DE	20.71	47.99	20.44	41.23
PL	6.90	14.87	6.32	13.36
ES	21.26	44.53	20.13	40.03
NL	9.19	18.72	8.93	17.97
AUS-QLD	3.68	8.44	3.53	8.35
Average	9.97	21.62	9.54	20.02

Table 12: *4-day zero-shot MAPE for* CarbonX *using Sundial and Time-MoE with different variants.*

D Uncertainty Quantification Details

Table 13 provides a detailed report of day-wise coverage for prediction intervals available using the vanilla quantile regression method in Sundial and our conformal approach in Carbon*X*. Sundial coverage across the different forecasting days ranges from 59.0% to 61.8%, and fails to achieve the desired 95% coverage on any forecasting day and across the 96-hour forecasting horizon. On the other hand, our conformal approach in Carbon*X* provides more than 95% coverage on average both daywise and across the 96-hour forecasting horizon, showcasing its reliability.

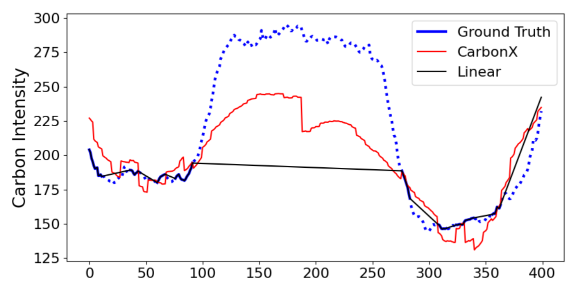


Figure 7: *Comparison of imputation results between* CarbonX *and linear interpolation.*

E Interpretation of Imputation Results

Figure 7 compares the imputation results between Carbon*X* and the linear interpolation baseline. Carbon*X* uses MOMENT in its task module, and we fine-tuned the linear head of MOMENT for this experiment while keeping the transformer backbone frozen. Each time step on the X-axis represents a 5-minute period. The blue line represents the actual carbon intensity values, with the solid parts referring to the values that are available and the dotted parts referring to the missing (masked) values.

Figure 7 highlights scenarios where Carbon*X* provides better estimates than the linear interpolation method, as well as scenarios where linear interpolation is more effective. We see that when the missing data has a monotonic trend, or if the period of missing data is smaller (as evident from time steps 340 to 356), linear interpolation may estimate the values better than Carbon*X*, which has some intrinsic model errors. On the other hand, if data is missing for larger periods, or when the values follow a more volatile pattern (as

	Sundial						**CarbonX**				
	MAPE (%)	**Coverage (%)**					**Coverage (%)**				
		D1	D2	D3	D4	Overall	D1	D2	D3	D4	Overall
CISO	10.79	53.7	58.7	59.8	58.0	57.6	95.6	95.9	96.7	96.9	96.3
PJM	4.68	66.1	71.0	69.5	67.8	68.6	96.4	96.1	96.6	96.4	96.4
ERCOT	14.64	55.7	62.6	59.9	60.2	59.6	95.6	94.7	95.1	94.6	95.0
ISO-NE	6.89	55.9	64.2	64.4	64.5	62.2	97.6	97.3	96.5	96.0	96.9
BPAT	9.92	58.4	59.6	58.9	58.2	58.8	96.9	95.9	95.8	95.7	96.1
FPL	2.86	64.4	64.5	64.8	63.4	64.3	96.0	95.3	95.2	94.5	95.3
NYISO	8.59	57.6	60.8	59.5	59.6	59.3	97.4	96.7	95.7	94.7	96.1
SE	6.22	60.4	63.0	60.1	59.3	60.7	96.5	95.6	94.7	94.3	95.3
DE	21.06	61.5	61.2	64.1	61.7	62.1	96.7	97.7	97.2	95.3	96.7
PL	6.48	62.4	61.3	64.0	62.8	62.6	97.5	96.6	96.4	96.0	96.6
ES	20.42	60.8	62.0	58.7	59.9	60.3	96.5	94.3	94.6	93.7	94.8
NL	8.69	64.5	64.1	63.3	63.2	63.8	97.4	97.4	97.3	96.9	97.3
AUS-QLD	3.48	46.6	51.1	53.8	57.4	52.2	97.6	97.5	98.8	98.5	98.1
Average	9.59	59.0	61.8	61.6	61.2	60.9	96.7	96.2	96.2	95.6	96.2

Table 13: ***Uncertainty quantification:*** **`CarbonX`** ***provides more than the desired 95% coverage with moderately wide intervals. In contrast, vanilla quantile regression provided by Sundial fails to reach 95% coverage in any of the grids.***

evident from time steps 92 to 276), `Carbon`X may provide estimates that are closer to the actual value and hence have lower estimation errors as well as preserve the pattern better.