Public review for

Multi-day Forecasting of Electric Grid Carbon Intensity using Machine Learning

Diptyaroop Maji, Prashant Shenoy, Ramesh K. Sitaraman

This paper aims to develop multi-day forecasts for the carbon intensity of the electric grid using a neural network based machine learning approach. This specific task of multi-day carbon intensity prediction is generally challenging as we have the limited availability and uncertainty in the data collection for renewable energy production, weather conditions, and other socioeconomic considerations. The previous research efforts focused on shorter time horizons (e.g., 24-hour) for their prediction tasks by utilizing detailed and continuous data. On the other hand, the authors in this paper developed a hierarchical method to provide two models which sperate the prediction task in 1) energy source generation and 2) the contribution to the carbon intensity by such energy source generation. Based on their simulation-based evaluation results, the proposed approach shows low MAPE prediction score over a 96-hour time window. Both weather and renewable energy production are the most sensitive factors in this prediction. The authors also tested their approach in 13 different geographically distributed locations, where they achieved all reasonable performance rates. The research motivation is very important as we must target the decarbonization of our electric grid in near future. We absolutely acknowledge the clarity and novelty of the method as well as their thorough evaluation process. Their approach is publicly available on the author's Github repository which will contribute to the community by sharing their knowledge and approach for future benchmarking tasks. Overall, this research is well conducted and critical to the community and society.

Public review written by

June Young Park

University of Texas at Arlington, USA

Multi-day Forecasting of Electric Grid Carbon Intensity using Machine Learning

DIPTYAROOP MAJI, University of Massachusetts Amherst, USA
PRASHANT SHENOY, University of Massachusetts Amherst, USA
RAMESH K. SITARAMAN, University of Massachusetts Amherst, USA

The ever-increasing demand for energy is resulting in considerable carbon emissions from the electricity grid. In recent years, there has been growing attention on demand-side optimizations to reduce carbon emissions from electricity usage. A vital component of these optimizations is short-term forecasting of the carbon intensity of the grid-supplied electricity. Many recent forecasting techniques focus on day-ahead forecasts, but obtaining such forecasts for longer periods, such as multiple days, while useful, has not gotten much attention. In this paper, we present CarbonCast, a machine-learning-based hierarchical approach that provides multi-day forecasts of the grid's carbon intensity. CarbonCast uses neural networks to first generate production forecasts for all the electricity-generating sources. It then uses a hybrid CNN-LSTM approach to combine these first-tier forecasts with historical carbon intensity data and weather forecasts to generate a carbon intensity forecast for up to four days. Our results show that such a hierarchical design improves the robustness of the predictions against the uncertainty associated with a longer multi-day forecasting period. We analyze which factors most influence the carbon intensity forecasts of any region with a specific mixture of electricity-generating sources and also show that accurate source production forecasts are vital in obtaining precise carbon intensity forecasts. CarbonCast's 4-day forecasts have a MAPE of 3.42–19.95% across 13 geographically distributed regions while outperforming state-of-the-art methods. Importantly, CarbonCast is the first open-sourced tool for multi-day carbon intensity forecasts where the code and data are freely available to the research community.

CCS Concepts: • **Social and professional topics** → **Sustainability**; • **Computing methodologies** → *Neural networks.*

Additional Key Words and Phrases: grid carbon intensity, multi-day forecasting, hierarchical design, source production forecasts, machine learning

Availability of Data and Material:
The data and code used in this paper are available at **https://github.com/UMass-LIDS/CarbonCast** (commit as of paper submission: f4c751b).

1 INTRODUCTION

Modern society depends on the electric grid to power many aspects of our daily lives, such as lighting, heating, and cooling, to name a few. According to the US Energy Information Administration (EIA), electricity consumption in the US was around 3.8 trillion kWh in 2020 [34], and the total energy demand is slated to rise by nearly 50% by 2050 [32]. The grid's energy demand exhibits temporal variations over a day and across seasons. A region's electricity grid uses various sources, ranging from conventional sources such as coal, oil, and natural gas to renewable sources such as hydro, solar, and wind, to generate sufficient electricity to meet the demand. However, electricity generation emits a significant amount of greenhouse gases and is one of the major contributors to greenhouse gas emissions in many regions worldwide [21, 36, 37]. These emissions, which depend on the generation source, vary over time as the mix of sources itself changes. The rising deployment of renewable sources such as solar and wind also introduces substantial variations in the grid's carbon emissions due to their intermittent nature.

As part of the ongoing energy transition in line with the United Nations' climate goals [29], there is an emergence of carbon reduction and trading policies, as well as an increasing interest in developing techniques to reduce the carbon emissions from the electricity grid. From an energy supply perspective, increasing the fraction of clean, renewable sources and masking the intermittent nature of renewable sources through energy storage are key approaches for reducing grid emissions. From a demand-side perspective, techniques to shift energy demand from periods when the carbon intensity of energy is high to periods when it is low have started gaining attention. Future knowledge of the grid's carbon intensity is an essential requirement for both complying with carbon trading policies and reducing carbon emissions using demand-side optimization techniques. Given carbon forecasts of the electric grid's electricity generation, demand-side techniques can leverage this knowledge to decide how much load to shift, where to shift, and to what hours, in accordance with some existing policy. For example, suppose "Baseline and Credit" [20] policy is in effect in a region. Then, residential charging of electric vehicles in that region can be scheduled intelligently based on future knowledge of when grid emissions are lower [13], keeping the total carbon emissions under the baseline and earning carbon credits. In the context of buildings, flexible loads (e.g., laundry) can be deferred to low-carbon periods. Such techniques are also being employed in other sectors, such as cloud computing, driven by aggressive goals of major cloud providers to reduce their carbon footprint [11, 26]. Since computing loads exhibit substantial temporal elasticity (e.g., batch workloads, interruptible machine learning), researchers have also begun to develop techniques for shifting loads to low-carbon hours [11, 18].

Short-term forecasts of grid carbon intensity are key for building carbon-aware systems and applications in order to reduce their carbon footprint. The carbon intensity of electricity is defined as the average carbon per unit of electricity generated and is expressed in the units of $grams/kWh$. Recently, grid operators have begun releasing real-time data about the carbon intensity of their supplied electricity [3, 25], and third-party services such as Watttime [38] and ElectricityMap [9] have begun to aggregate such data and expose real-time carbon intensity via cloud interfaces. In addition to exposing the real-time carbon intensity of electricity, there has also been work on short-term forecasting of carbon intensity using historical data [16, 22]. Both Watttime and ElectricityMap have also begun to provide such forecasts as part of their commercial

Authors' addresses: Diptyaroop Maji, dmaji@cs.umass.edu, University of Massachusetts Amherst, Massachusetts, USA, 01002; Prashant Shenoy, shenoy@cs.umass.edu, University of Massachusetts Amherst, Massachusetts, USA, 01002; Ramesh K. Sitaraman, ramesh@cs.umass.edu, University of Massachusetts Amherst, Massachusetts, USA, 01002.

service. Much of the work on near-term forecasting has emphasized day-ahead forecasts, which provide carbon intensity predictions for the next 24 hours. For example, many recent efforts [9, 12, 14, 22] provide day-ahead forecasts of carbon intensity, while some like Bokde et al. [16] provide 48-hour forecasts. While such forecasts are useful for various types of demand-side carbon optimizations, some techniques that operate over multiple days (e.g., intelligent battery charging, scheduling long-running cloud jobs) require forecasts for periods longer than 24 hours. Similar to how weather forecasts provide predictions for the next day as well as several days into the future, the design of techniques for multi-day carbon intensity forecasts is a problem of considerable importance but one that has not received much attention.

Research contributions. In this paper, we present CarbonCast[1], which is a system based on machine learning to forecast multi-day (up to 96 hours) grid carbon intensity. Extending day-ahead forecast methods to multiple days is challenging since the factors influencing carbon intensity are more unpredictable and have greater variability over longer time horizons. To reduce prediction error for multi-day forecasting, CarbonCast considers both historical data of the sources used for electricity generation, as well as other factors like weather forecasts and the electricity generation forecasts for each source. We show that judiciously using such forecast information improves prediction accuracy over longer time periods. For example, for a given energy demand, a day with a forecast for high winds is likely to have a lower carbon intensity due to more electricity generation from wind. We make the following specific contributions.

(i) Hierarchical design. CarbonCast uses a two-tiered hierarchical approach, each based on machine learning. The first tier uses neural network models to provide individual source electricity generation forecasts. The second tier, based on a hybrid CNN-LSTM combination, uses these forecasts with weather forecasts and historical carbon intensity data to generate 96-hour carbon intensity forecasts. Importantly, our hierarchical approach makes CarbonCast robust to noisy or partially-missing inputs and so is suitable for multi-day forecasts. Our two-tiered approach also provides a modular design where each tier can be independently improved. For instance, if we can obtain improved wind energy production forecasts from any method, that improved forecast can be incorporated directly into our system to improve the carbon intensity forecasts.

(ii) Multi-day forecasts. We provide 96-hour forecasts for the grid carbon intensity of 13 regions across the US, Europe, and Australia. We show that CarbonCast can be used in different regions of the world with minimal changes to get good multi-day carbon intensity forecasts. We also provide forecasts based on both lifecycle (operational and infrastructural) and direct (only operational) emission factors. Thus, our system can be incorporated by both scope 2 [30] and scope 3 [31] carbon emission optimization solutions.

(iii) Feature importance for carbon intensity prediction. We analyze which features are important for predicting the carbon intensity of the electricity grid in a given region. For example, we show that in California, solar production forecast is the most important feature for predicting the carbon intensity. In contrast, wind forecast plays a bigger role in predicting the carbon intensity in Texas.

(iv) CarbonCast error analysis. We analyze why CarbonCast performs better and has lower forecasting errors in certain regions compared to others. We show how to improve the results in regions with higher errors. In particular, we provide guidance on which source forecasts to improve to enhance the precision of the carbon intensity forecast.

(v) Improving the state-of-the-art. We compare CarbonCast to state-of-the-art methods, as well as other baselines. When averaged over 96 hours, CarbonCast using direct (resp. lifecycle) emission factors has a MAPE of 9.78% (resp. 8.38%) across all the regions. CarbonCast also reduces the forecasting MAPE by 9.96% (resp. 8.91%) over the current state-of-the-art across all the considered regions.

(vi) Open source tool. Energy research in diverse areas, from buildings to data centers, requires longer-range location-dependent predictions of carbon intensity. Both the code and data of CarbonCast are available to the public, and it is the first open-source tool[2] for multi-day predictions of carbon intensity. We hope that our tool will be used by the community in energy research projects that rely on such predictions.

Roadmap. The rest of this paper is as follows: Section 2 discusses the background. Section 3 explains the CarbonCast system design. Section 4 experimentally evaluates the accuracy, robustness, and runtime of our approach and also analyzes which factors are important in predicting the grid carbon intensity. Section 5 talks about the compatible nature of CarbonCast and justifies why CarbonCast provides *average* carbon intensity forecasts. Section 6 discusses the related work, and Section 7 concludes the paper.

2 BACKGROUND

In this section, we provide background on regional electricity grids, types of sources generating electricity, carbon emission factors of each source, carbon intensity associated with electricity generation, and how it varies across regions and with time.

2.1 Regional grids and electricity sources

The electricity grid in each region performs three functions: generation, transmission, and distribution [33]. Electricity is generated by power plants of various types, transmitted over a network of transmission lines and finally distributed to end customers via stations and substations. Typically, electricity is generated from a mix of renewable and non-renewable sources. Since supply should match demand, the grid uses a set of dispatchable generators that can be turned on or off to match a time-varying demand. Sources such as renewable solar and wind tend to be intermittent and are assumed to be uncontrolled; other non-renewable sources are then used to meet the remaining demand. Thus, the fraction of electricity generated by each source varies over time and across different regions. Factors like locational marginal price [28] and imported electricity also govern the current source mix of a particular region.

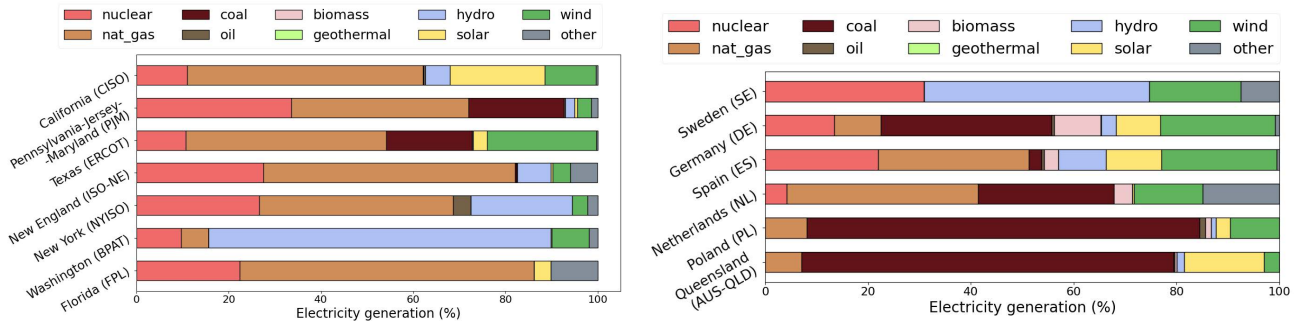
In this paper, we consider electricity grids in several regions across the US, Europe (EU), and Australia (AUS).

[1]This paper is an extended version of an earlier paper that was published in ACM BuildSys 2022.

[2]https://github.com/UMass-LIDS/CarbonCast, commit as of paper submission: f4c751b

Emission factors	Coal	Oil	Natural gas	Nuclear	Solar	Wind	Hydro	Other	Biomass	Geothermal
Lifecycle	820	650	490	12	45	11	24	700	230	38
Direct	760	406	370	0	0	0	0	575	0	0

Table 1. Median lifecycle and direct carbon-emission factors (g/kWh) for different renewable and non-renewable sources.



(a) Regions in the US.

(b) Regions in Europe and Australia.

Fig. 1. Average electricity production by source during the period Jan 2020 — Dec 2021 showing wide variations across regions.

- In the US, we consider the following seven regions: California, US (CISO), Pennsylvania-Jersey-Maryland Interconnection (PJM), Texas (ERCOT), New England (ISO-NE), Washington (BPAT), Florida (FPL), and New York (NYISO).
- In Europe, we consider five regions: Sweden (SE), Germany (DE), Poland (PL), Spain (ES), and Netherlands (NL).
- In Australia, we consider Queensland (AUS-QLD).

When aggregated over all the regions, the sources include non-renewable sources like *natural gas, coal, oil, and nuclear*, and renewable sources like *solar, wind, hydro, geothermal, and biomass*. Note that the exact mix of sources used to generate electricity is not always known accurately. If some generation source is not reported for a region, it is assumed to be non-renewable and is listed as "other" with an approximate carbon intensity. In other cases, some sources may be missing in the data reported by the grid operator. These factors introduce noise and uncertainty when estimating current and future carbon intensity, especially over longer time horizons.

2.2 Carbon emission factor (CEF)

We define the carbon emission factor (CEF, in g/kWh) of a source as the amount of carbon emitted into the atmosphere per unit of electricity generated by that source. CEFs of non-renewable sources are usually much higher than that of renewable sources. Based on the type of accounting (scope 2 [30] or scope 3 [31] emissions), there can be two types of CEFs for a source:

- **Direct emission factors:** These are the operational emissions when a source is converted into electricity and are used when accounting for scope 2 [30] emissions.
- **Lifecycle emission factors:** These include operational as well as infrastructural emissions up the supply chain and are considered when accounting for scope 3 [31] emissions.

CEF values for a source may vary across power plants in different regions. For example, power plants burning black coal (anthracite/bituminous coal) to generate electricity will emit more carbon than those burning brown coal (lignite). Determining CEFs is a separate problem. Instead, our work considers CEFs as input to CarbonCast. In this paper, we simplify the sources (referring to both black and brown coal as coal) and use standardized median values of carbon emission factors for each source [6, 7], as shown in Table 1. For our modelling and forecasting purposes, we assume that the CEF of "other" sources is the same across all regions. We provide forecasts using both types of emission factors and leave it up to the practitioners to choose which forecast to use.

2.3 Average carbon intensity

The average carbon intensity per unit of electricity generated in a region is the weighted average of carbon emitted by each source due to the electricity generated by them. Mathematically, the average carbon intensity (in g/kWh) of a region at any time is as follows:

$$(Carbon\ Intensity)_{avg} = \frac{\sum (E_i * CEF_i)}{\sum E_i} \quad (1)$$

where E_i is the electricity generated (MW) by a Source i & CEF_i is the CEF (g/kWh) of that source.

Electricity grids often exchange electricity with neighbouring grids to meet the demand. Hence, when calculating the average carbon intensity of any region, we should also consider the carbon intensity of any imported electricity. However, any grid exporting electricity may import electricity from other neighbouring grids. We need to find the origin source of electricity to calculate the carbon intensity of imported electricity, which is not straightforward. So, we only consider *the average carbon intensity of electricity generated in a region, ignoring any imports/exports* in this paper, for simplicity. We discuss the effects of imports and exports on CarbonCast in detail in Section 5.2.

2.3.1 ***Spatial variability of average carbon intensity.*** Fig. 1 shows the average fraction of electricity generated by each source for 2020 – 21 in all the regions. We see that the fraction of each

renewable or non-renewable energy source varies across regions. For example, PJM depends heavily on fossil fuels, whereas Sweden relies heavily on renewables. Typically, the carbon intensity of a region is proportional to the fraction of electricity generated by fossil fuels in that region. The greener the source mix, the lower the value of average carbon intensity.

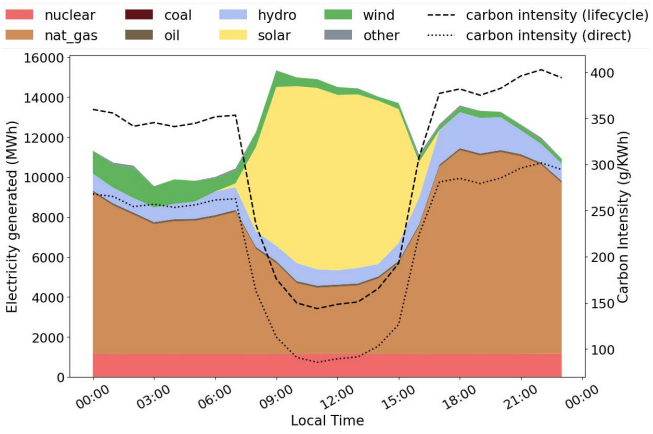


Fig. 2. The sources of electricity vary temporally in California. This results in a temporal variation in the average carbon intensity, with lower values during day when solar production is high.

2.3.2 ***Temporal variability of average carbon intensity.*** The source mix also varies with time. Renewable sources generating electricity in a region depend on weather and are highly volatile. Additionally, since electricity demand is ever-changing and supply must always match the demand, additional dispatchable generators may need to be turned on during peak load and turned off during low demand. The grid's dispatch schedule and types of generators used during periods of high load depend on the price of generating electricity using a particular source at that time. As a result, average carbon intensity also varies with time. Fig. 2 shows how the source fractions change throughout a specific day in California and how it affects the average carbon intensity. Note that the temporal pattern of average carbon intensity is similar for both lifecycle and direct emission factors.

3 CARBONCAST DESIGN

We present the design of our CarbonCast approach in this section.

3.1 Overview

The goal of CarbonCast is to take historical data of the source mix used for electricity generation in a region, the carbon-emission factors (lifecycle or direct) of each source, and weather forecasts, to produce a multi-day hourly forecast of the carbon intensity of electricity in that region. CarbonCast currently produces a 4-day (i.e., 96-hour) forecast, and we believe it can be enhanced further in the future to produce 7-day to 10-day forecasts.

CarbonCast uses a hierarchical two-tiered forecasting approach based on machine learning, as shown in Fig. 3. The first tier uses a set of models, one for each generation source, to predict the electricity production from that source for the next 96 hours. The second tier takes these first-tier predictions along with weather forecasts to predict the hourly carbon intensity of electricity in that region for the next 4 days. Several challenges need to be addressed when making multi-day forecasts, which we discuss next.

First, the amount of renewable sources in electricity generation varies by region. In regions with significant penetration, their intermittent nature can complicate carbon intensity forecasting, especially since intermittent generation causes the carbon intensity of the grid to vary noticeably over time. Our CarbonCast uses weather forecasts, in addition to historical production data, to accurately predict future generation from renewables.

Second, if accurate source production forecasts for all sources were available, we could use Eq. 1 to calculate the overall carbon intensity. This is the approach used by DACF [12]. However, source production data may often be unavailable during some time periods, or some sources may be unknown and listed as "other". In such cases, Eq. 1 can produce higher errors or may be infeasible to use.

Third, tier-1 forecasts get progressively worse with the increasing time horizon—partly due to limitations of the models generating these forecasts and partly due to weather forecasts becoming less accurate as we go further into the future. In such cases, using the equation may not be optimal, as it always assigns a fixed weight to each source, and so cannot adjust to the inaccuracies in the inputs.

Consequently, adding a second tier of deep-learning model architecture, which can accommodate noisy/missing inputs by readjusting the weights assigned to each input feature, can be a better approach to forecast multi-day carbon intensity over using Eq. 1.

Finally, the carbon intensity of a region may have noticeable seasonal and daily patterns (e.g., carbon intensity in California is low during the day due to high solar generation, and solar generation is higher in summer). Hence, CarbonCast considers these patterns when using the historical source mix and carbon intensity data.

The above observations motivate CarbonCast's two-tier approach to achieve its goals of multi-day carbon intensity forecasting. Similar to DACF [12], we refer to the hourly electricity produced (in *MW*) by a source (past 24 hours) as the *historical source production*, whereas the hourly predicted electricity production by a source (in *MW*) is referred to as the *source production forecast*. Specifically, the first tier takes in hourly electricity generated by the individual sources and outputs individual source production forecasts. These forecasts are then fed into the second tier along with other features like historical carbon intensity and weather forecasts. The second tier then computes the carbon intensity forecasts using a machine learning model which uses a combination of CNNs and LSTMs.

We now elaborate on each tier of CarbonCast and discuss how our approach can be applied to any region with minimal changes.

3.2 First-tier design

The goal of the first tier is to estimate the hourly electricity generated by each production source present in a region over the next 96 hours. In some cases, day-ahead forecasts for renewable sources, such as solar and wind, are available from the grid operator, but such forecasts are rarely available for all types of generation sources. Also, when available, those are limited to 24 hours rather than 96 hours. Consequently, CarbonCast uses ANN models, one per production

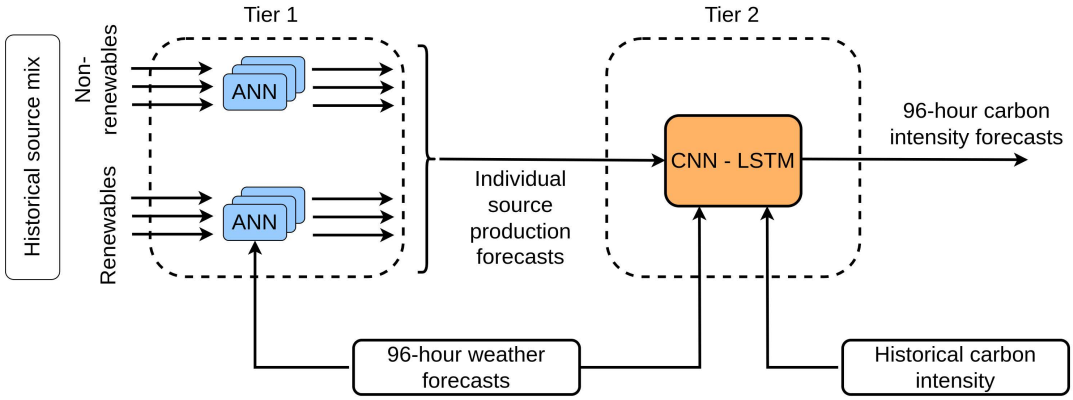


Fig. 3. CarbonCast architecture. Historical source mix and carbon intensity values for the past 24 hours are used as inputs.

source, at the first tier to predict the source production forecasts. This first-tier forecasting approach is inspired by other methods, such as DACF [12], which use similar models.

Fig. 4 shows our first-tier architecture. For each region, we consider all the sources producing electricity in that region. We have a separate ANN model for each source that takes in the source's historical electricity production as input. Moreover, we include features like hour-of-day and hour-of-year as input to the ANN model to capture diurnal or seasonal trends. We also consider whether the current day is a weekday or a weekend since electricity demand and consequently, production varies across weekdays and weekends.

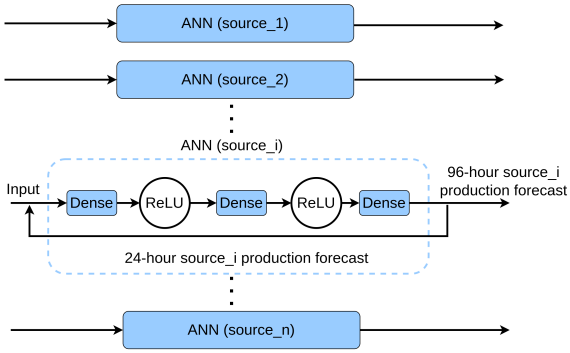


Fig. 4. CarbonCast uses neural network models, one per source, to predict future generation.

For renewable sources like solar, wind and hydro, we also consider weather forecasts as additional inputs. This is because weather affects renewable energy production (e.g., more precipitation correlates with more electricity production from hydro; solar energy production is lower during winter months with shorter days).

3.3 Second-tier design

The second tier aims to take the hourly generation forecasts from each source, as produced by the first tier, and produces an aggregate carbon intensity estimate. As discussed earlier, our second tier relies on a deep learning model to compute carbon intensity rather than an analytic equation to deal with the impact of noisy or missing data, which makes the multi-day forecasting problem more challenging. To do so, we use a combination of CNN and LSTM models. Fig. 5 shows our CNN-LSTM design. Forecasting grid carbon intensity is essentially a time-series forecasting problem, and LSTM is a state-of-the-art technique used in such problems. Since the inputs to the second tier are multiple time series data, we add two 1-D CNN layers to extract high-level "short-term" temporal features from those inputs and feed them to the LSTM layer, which can learn "longer-term" temporal patterns in the time series.

In addition to the 96-hour source production forecasts obtained from the first tier, we use several other features as input to the second tier so that CarbonCast can learn more effectively and counter the errors in source production forecasts. We include historical average carbon intensity data of that region (past 24 hours), calculated from the historical source mix using Eq. 1. We also add date-time related features (e.g., hour-of-day, hour-of-year etc.). Finally, we add 96-hour weather forecasts as input to this tier as well. Specifically, we use the following weather variables in both tiers: *u- and v-component of wind* (in m/s) at $10m$ height above sea level from which we derive the wind speed (in m/s), *temperature* (in K) and *dewpoint temperature* (in K) at $2m$ height above sea level, and *downward short-wave radiation flux (DSWRF)* (in W/m^2) and *total precipitation* (in kg/m^2) at the surface level.

This second tier of deep-learning model with additional features enables CarbonCast to accommodate noisy or even missing source production forecast data by re-adjusting the weights of each input feature. Consequently, this makes CarbonCast more suitable for multi-day forecasts, as it adds robustness even if the tier-1 forecasts get progressively worse with an increasing forecasting period.

3.4 CarbonCast implementation

CarbonCast is implemented using Keras [5] on Tensorflow [15]. The first-tier ANN models have three fully connected (dense) layers. The first dense layer has 50 hidden units, followed by the second layer with 34 hidden units. The final layer has 24 units, which outputs the source production forecast. We use Rectified Linear Unit (ReLU) activation between the layers. In the second tier, there are two CNN layers with a max-pooling layer in between. The first CNN layer

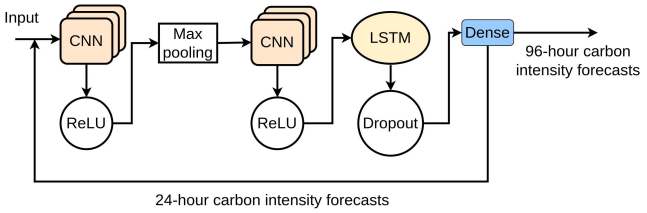


Fig. 5. CarbonCast second tier model architecture.

has four 4 x 4 filters and necessary padding to keep the output dimensions the same as the input dimensions. The second CNN layer has 16 4 x 4 filters. The CNN layers are followed by a LSTM layer with 24 units and a dropout layer with rate 0.1. Finally, there is a dense layer having 24 output units. In both the tiers, we obtain the 96-hour forecasts one day at a time and treat the forecasted values for i^{th} day as historical data for forecasting the $(i+1)^{th}$ day. For example, we use historical solar production data to compute the day-ahead solar production forecasts. Then, we replace the historical data with the forecasted data to get solar production forecasts for the next 24 hours. We continue this process till all 96 hours of forecasts are obtained. We obtain carbon intensity forecasts in a similar way.

We build our datasets from openly available data (refer Table 2). Our code and datasets are available at ***https://github.com/UMass-LIDS/CarbonCast*** for practitioners and researchers to incorporate into their carbon optimization and accounting-related solutions. Since we provide forecasts based on both lifecycle and direct emission factors of sources, CarbonCast can be used in solutions aiming to reduce both scope 2 [30] and scope 3 [31] emissions.

Both our design and its implementation are modular. The first tier of CarbonCast individually forecasts each electricity-producing source in the electric grid of a particular region and is decoupled from the second tier. This modular architecture allows CarbonCast to seamlessly integrate any new model in either tier if it improves the overall forecasting performance without changing the remaining components. It also enables CarbonCast to use the same framework in any region of the world, regardless of significant differences in the source mix across the electric grids. CarbonCast needs training data from a region to produce a two-tier model specific to that region. However, our evaluation shows that we can use the same set of features and model hyperparameters across various regions.

4 EXPERIMENTAL EVALUATION

In this section, we evaluate our design choices and CarbonCast performance. First, we show the advantages of our hierarchical two-tier approach compared with a non-hierarchical approach. Then, we show how CarbonCast performs across electric grids in 13 different regions across the US, Europe, and Australia and analyze which features are important in a particular region. We then analyze the reason behind forecasting errors in CarbonCast and why some regions have higher errors than others. We also compare CarbonCast with recent carbon intensity forecasting methods and show that CarbonCast provides better multi-day forecasts than the current state-of-the-art. Finally, we show that it is practical to run CarbonCast daily, if required, by evaluating its runtime overheads.

4.1 Experimental methodology

Data sources. Table 2 lists the data sources used in this paper. For any region, the 96-hour weather forecasts provided by [24] need to be aggregated over the whole region. For that, we refer to [8] to get the bounding boxes for all the regions we have considered in this paper. Then, following the weighted average procedure suggested in [23], we aggregate the weather data over a particular region. These forecasts are given at three-hour granularity, while our carbon and electricity production data are at hourly intervals. For the sake of simplicity, we assume that weather variables have the same values across the three hours. If day-ahead solar and wind forecasts are available for a region, we directly use that and compute the day-2, 3 and 4 forecasts. For other sources or regions, we compute the full 96-hour source production forecasts using our first tier.

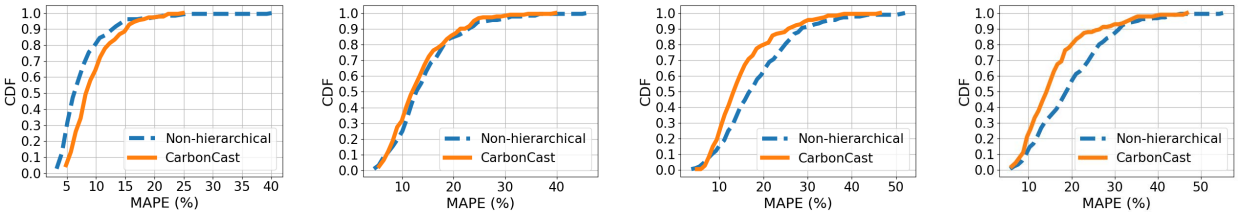
Type of data	Regions		
	US	**AUS**	**EU**
Historical source mix	EIA [1]	OpenNEM [2]	ENTSOE [19]
Day-ahead solar/wind forecasts	OASIS [4] for CISO, N/A for others	N/A	
96-hour weather forecasts	NCEP GFS ds084.1 [24]		

Table 2. Publicly available sources used to build our datasets.

Although we consider only hourly granularity in this paper, many electric grids and carbon optimization solutions operate at sub-hourly intervals. CarbonCast can work with data of any time granularity without any design changes. However, additional experiments would be required to evaluate CarbonCast's performance for the finer time granularities and is a possible direction for future work.

CarbonCast training and testing. We consider data from 2019 to 2021 for training CarbonCast and predicting average grid carbon intensity. The first tier of ANN models uses hourly historical source production data from January 1 to December 31, 2019, for training, and we predict individual source production forecasts for the remaining period. As the prediction period of 2 years is long, we only predict six months at a time and update and re-train the ANN models with new data every six months to increase the prediction accuracy, similar to DACF [12]. The input data to the second tier is from January 1, 2020, to December 31, 2021, in hourly granularity. The train-validation-test split is 50%–25%–25%.

We use the sliding-window technique for training all the models. At each time step (~ 1 hour), the model looks at the most recent 24 hours as input data and values for the next 24 hours as labels. We use Root Mean Square Error (RMSE) as the loss function that is minimized during training. During testing, for the i^{th} day, we predict the next 96 hours' average carbon intensity at 00:00 hours, 24 hours at a time. We use the actual data from the $(i-1)^{st}$ day



(a) Day 1 forecasts (0-24 hr). (b) Day 2 forecasts (24-48 hr). (c) Day 3 forecasts (48-72 hr). (d) Day 4 forecasts (72-96 hr).

Fig. 6. Comparison between forecast CDF plots of Hierarchical CarbonCast vs non-hierarchical equation based approach. CarbonCast forecasts are resilient to noises in the input and CarbonCast can still give good predictions even when source production forecasts get progressively worse with an increasing forecasting period.

to predict the first 24 hours. Then, we take this forecasted data to predict the next 24 hours and continue this process till 96 hours. We evaluate the forecasting performance in terms of Mean Absolute Percentage Error (MAPE). Further, since CarbonCast uses stochastic methods, we take an average of three runs whenever we report the values.

4.2 Benefits of a hierarchical design

First, we justify the need for our hierarchical design. Since our first tier generates all the source production forecasts, we can directly use Eq. 1 to get average carbon intensity forecasts by replacing the fraction of electricity produced by a source with the *forecasted fraction* of electricity produced by that source. However, adding a second-tier model instead of using Eq. 1 has the following benefits:

4.2.1 ***Resiliency against missing data.*** A major challenge in forecasting carbon intensity using data-driven methods is the availability of good-quality electricity production data. Any system that calculates carbon intensity using Eq. 1 needs a consistent flow of electricity production data by each source. However, such data is unavailable in sufficient detail and granularity outside the US, Europe, and Australia. Even these regions may suffer from outages [35], which may result in electricity generated from one or multiple sources not being reported for an extended time. In these cases, carbon intensity forecasts cannot be calculated using the analytical equation approach since we do not have sufficient data. If we calculate by adding zero values for the missing source, forecast quality deteriorates heavily. However, since CarbonCast uses an additional tier of machine learning model, it is resilient against such missing data and can perform well even if some source is unavailable.

We design a simple experiment to prove our claim. Fig. 1 shows the fraction of electricity produced by each source in California during the 2020 – 21 period. Solar is one of California's most important sources of electricity, contributing to about 20% of the total electricity generation. To prove our hypothesis, we remove *solar production forecast* from both the input to our second tier and the equation to calculate the carbon intensity forecast, to simulate a scenario where this data is unavailable.

We see that CarbonCast performance has negligible effect as other features compensate for the missing data. In this particular case, the model assigned more weights to historical carbon intensity and solar irradiance (DSWRF) to cope with the missing solar production forecast. In contrast, the non-hierarchical performance in California with missing solar data degrades by 2 – 5x times (refer Table 3).

CISO	Hierarchical		Non-hierarchical	
	All sources	No solar data	All sources	No solar data
Day 1	9.40	9.53	7.51	33.54
Day 2	13.23	12.24	14.15	34.05
Day 3	15.09	13.99	18.34	36.31
Day 4	15.75	14.98	19.76	37.27

Table 3. CarbonCast performance (in terms of MAPE) is comparable even when some source data is missing, whereas performance of the non-hierarchical approach degrades heavily.

While this is an extreme case and techniques like replacing missing data with data from an earlier time period can offset such performance degradation to some extent, there may still be high forecasting errors as past data may not always be similar to the missing data. We show in Section 4.2.2 that CarbonCast works better than non-hierarchical methods even in these cases.

4.2.2 ***Resiliency against noise in input.*** We can replace missing data using various imputation techniques. Since the substitute data is an estimate, we can treat it as noise in the input data. The amount of noise depends on the type of data missing, but CarbonCast is more robust than a non-hierarchical method in all such cases. To show this, we carry out two experiments. First, we assume solar data is missing for the test period (June – December 2021). Next, we assume natural gas data is missing for the same time period. We choose natural gas because it generates the most fraction of California's electricity and has a high CEF. Hence, noisy natural gas data is expected to have the most effect on carbon intensity forecasts. In both cases, we substitute the missing data with data from the previous year (June – December 2020) and forecast carbon intensity using both approaches. Table 4 lists the results. We see that although the non-hierarchical method can cope with the previous year's solar data, its performance degrades by 1.3 – 2x times when the previous year's natural gas data is used. In contrast, CarbonCast performance is similar in all such cases.

Additionally, even when data is available, forecast accuracy generally deteriorates as the forecasting period increases. Fig. 7 shows

CISO	Hierarchical			Non-hierarchical		
	All sources	Solar data imputed	Nat_gas data imputed	All sources	Solar data imputed	Nat_gas data imputed
Day 1	9.40	9.49	9.42	7.51	8.84	14.76
Day 2	13.23	12.64	13.34	14.15	14.33	19.80
Day 3	15.09	14.44	15.72	18.34	18.49	22.74
Day 4	15.75	15.08	17.24	19.76	19.67	24.18

Table 4. When original data is missing, it can be replaced via techniques like using data from an earlier time period. Non-hierarchical approach performance can still degrade, whereas CarbonCast is more resilient to the noise introduced by such techniques.

how wind and natural gas production forecast errors increase in California with the forecasting period. Our hierarchical design is also resilient to such noisy source production forecasts. Even for short forecast periods, since equation-based approaches assign a fixed weight to each source production forecast (where the weight is the CEF of that source), the accuracy of such methods depends heavily on the accuracy of these forecasts. However, adding another tier of learning enables CarbonCast to re-weight the input features and adjust for the noise. To prove this, we set up a simple experiment. We start with perfect source production forecasts in California and then gradually add Gaussian noise to the natural gas production forecast and observe its effect on day-ahead carbon intensity forecasts. Since natural gas has the highest electricity-generation fraction in California and also has a high CEF, if all inputs to the system are perfect and only one is varied, a noisy natural gas forecast is expected to have the most effect on carbon intensity forecasts. Fig. 8 shows that in California, the equation-based approach worsens linearly as we add noise to the natural gas production forecast, while CarbonCast is more robust to the noise.

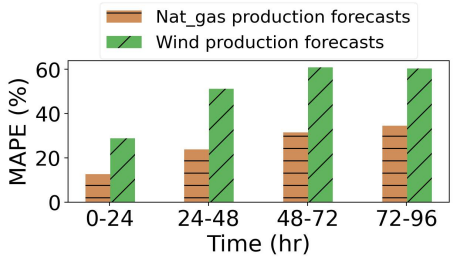


Fig. 7. Source production forecasts get more erroneous with increasing forecasting period.

Fig. 6 shows that the CDF plots for CarbonCast become better than that of an equation-based approach from day two onwards. The $90th$ percentile MAPE of CarbonCast becomes better by 18.39% (resp. 15.83%), 23.47% (resp. 22.54%), and 25.57% (resp. 21.64%) with direct (resp. lifecycle) emission factors on average across the regions when forecasting days 2, 3 and 4, respectively. In this paper, we only show the CDF plots for California considering direct emission factors, but this effect is visible for other regions as well as lifecycle emission factors. Ideally, CarbonCast should be able to learn optimal weights and match the performance of the equation-based approach even on day one. However, any neural-network model has intrinsic errors

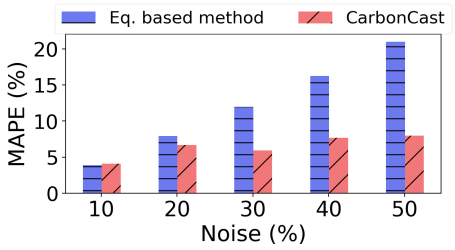


Fig. 8. CarbonCast is resilient to noise in the input, in contrast to equation based approaches.

while learning, and due to this, CarbonCast performance is slightly worse than the equation-based approach. This is evident from Fig. 12. Even with perfect source production forecasts, CarbonCast forecasts still have errors, whereas the equation-based approach would give perfect forecasts in this case.

In general, we see that CarbonCast performs on par or better as the forecasting period increases. Thus, we conclude that our hierarchical design is more suitable for multi-day forecasts due to the ability to re-weight the input features.

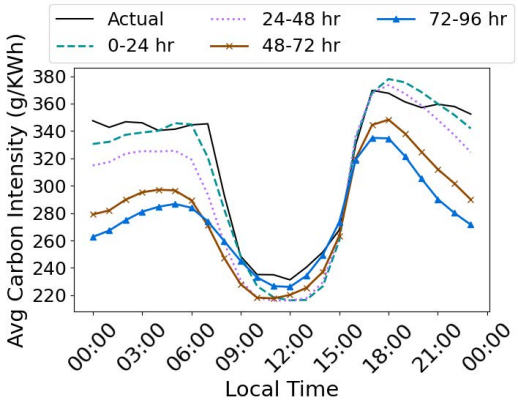


Fig. 9. CarbonCast forecasts generally match actual values, but get progressively worse with larger forecasting periods.

4.3 CarbonCast forecasting performance

We now evaluate how CarbonCast performs in 13 regions across the US, Europe, and Australia. Fig. 9 shows an hourly time series

averaged over a week for the actual and forecasted carbon intensities of the electricity grid in California. Table 5 (resp. Table 6) lists CarbonCast 96-hour forecasting performance across all the regions (in terms of MAPE) when direct (resp. lifecycle) emission factors are considered while calculating the average carbon intensity forecast. CarbonCast MAPE ranges from 3.42 – 19.95% when aggregated over 96 hours. When averaged over all the regions, CarbonCast has a MAPE of 9.78% (resp. 8.38%) across the regions when direct (resp. lifecycle) emission factors are considered. The day-wise MAPE in CarbonCast ranges from 2.93 – 24.70% (resp. 2.52 – 21.12%) across the regions with direct (resp. lifecycle) emissions. In general, we see that CarbonCast errors increase with the forecasting period.

Region	Mean	Median	90th percentile	95th percentile
CISO	13.37	11.96	22.21	25.99
PJM	4.80	4.04	8.10	9.70
ERCOT	11.13	8.76	21.25	27.17
ISO-NE	6.46	5.55	11.33	13.50
NYISO	9.52	5.72	31.64	35.82
BPAT	14.71	13.49	23.73	27.86
FPL	3.54	2.93	5.93	7.45
SE	10.07	8.54	17.78	20.39
DE	13.93	11.30	24.34	29.99
PL	4.58	4.07	7.22	9.07
ES	19.95	17.25	35.47	40.74
NL	9.68	9.00	15.72	17.66
AUS-QLD	5.35	5.07	7.85	8.54

Table 5. CarbonCast 96-hour forecast performance (using direct emission factors, in terms of MAPE)

Region	Mean	Median	90th percentile	95th percentile
CISO	11.45	9.91	18.58	24.27
PJM	5.29	4.51	8.74	10.17
ERCOT	11.14	8.40	21.44	28.85
ISO-NE	6.41	5.50	10.96	13.29
NYISO	9.09	5.86	28.09	32.24
BPAT	11.22	10.51	18.23	21.16
FPL	3.15	2.64	5.24	6.44
SE	5.78	5.12	9.47	11.36
DE	11.72	9.16	20.90	27.60
PL	4.37	3.76	7.49	9.31
ES	16.65	14.26	30.07	34.26
NL	8.25	7.67	13.43	14.88
AUS-QLD	4.46	4.18	6.72	7.64

Table 6. CarbonCast 96-hour forecast performance (using lifecycle emission factors, in terms of MAPE)

The regions where we have evaluated CarbonCast are diverse in terms of location, energy sources, and renewable production. For example, Sweden has a high hydro and wind (renewable) generation. In contrast, other regions have a high percentage of non-renewable fossil fuels (e.g., PJM with natural gas and coal). Our technique performs well and is robust enough to be re-trained and used in these representative regions. Given that CarbonCast is able to work in diverse regions, we conclude that it is an effective system for forecasting grid carbon intensity in most regions across the world.

4.4 CarbonCast error analysis

We analyze why CarbonCast has higher forecasting errors when aggregated over 96 hours for some regions. While we show the analysis for direct emission factors, the results are similar for lifecycle emission factors. We posit that regions with a higher fraction of

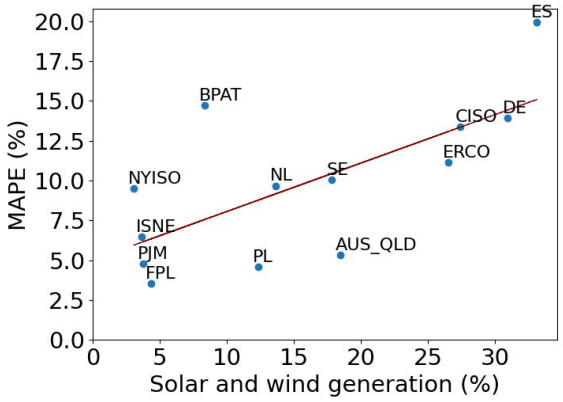


Fig. 10. CarbonCast errors are generally higher in regions that have a larger contribution from volatile sources such as solar and wind.

electricity generated from volatile sources have higher forecasting errors. Since solar and wind are typically the volatile sources in most of the regions, we plot the correlation between the fraction of electricity generated by solar/wind in a region over the testing period (July — December 2021) and the forecasting errors in terms of MAPE. Fig. 10 shows a positive linear correlation between solar and wind generation and the forecasting errors. One notable outlier in Fig. 10 is Washington (BPAT), which has a high MAPE despite having a relatively lower fraction of electricity generated from solar and wind. This is because electricity in BPAT is mostly generated from hydro (~ 70%), which meets any electricity requirement in the region when volatile solar and wind are unavailable. This is shown in Fig. 11, which shows electricity generation by each source over a week during the test period, where hydro compensates for solar and wind. Consequently, hydro acts as another volatile electricity-generating source, which increases the forecasting error.

To further confirm that errors in the source production forecasts of volatile solar and wind are partly the reasons for higher errors, we take three regions with high forecasting errors and replace the solar and wind production forecasts with actual values, simulating a zero-error forecast scenario. We keep all other inputs the same. Fig. 12 shows the MAPE of these regions with ideal solar and wind forecasts. The MAPE decreases by 22.29%, 20.53%, and 11.58% in California (CISO), Germany (DE), and Spain (ES), respectively, confirming that

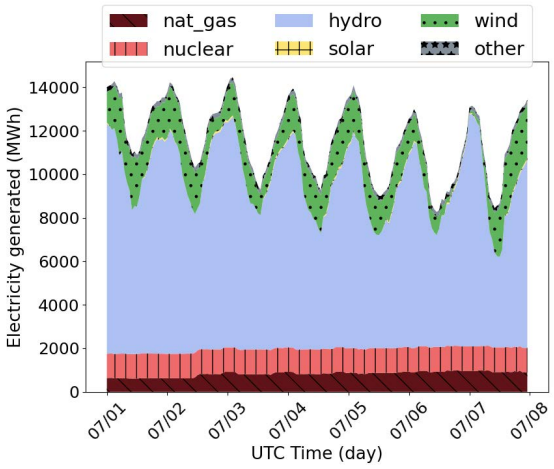


Fig. 11. Washington electricity generation by source over a week. Hydro compensates for the volatile nature of solar and wind in the region.

errors in production forecasts of volatile sources are partly why CarbonCast has higher forecasting errors in some regions.

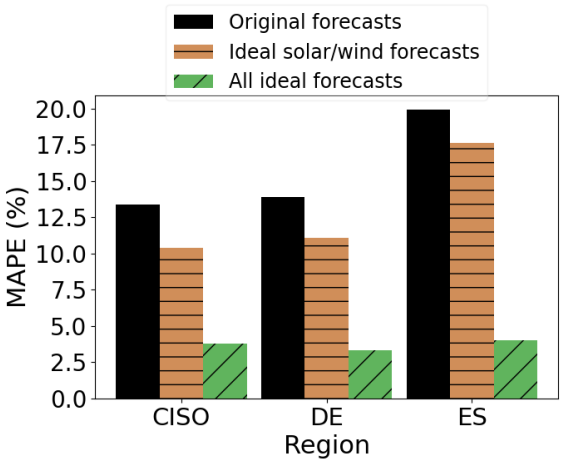


Fig. 12. CarbonCast performance (in terms of MAPE) improves with better source production forecasts.

Next, we extend this idea and claim that for any region, the forecasting errors in CarbonCast are mainly due to errors in source production forecasts. We show this by replacing all source production forecasts with their actual values (represented as "All ideal forecasts" in Fig. 12. With ideal forecasts, the forecasting errors (MAPE) in these regions decrease to 3.78%, 3.32%, and 3.99%, respectively. In this case, the MAPE obtained represents the error intrinsic to our prediction method independent of the source production forecast errors. This confirms our hypothesis that CarbonCast forecasting errors are mainly due to errors in the source production forecasts.

Thus, from this section and Section 4.2.2, we conclude that although the hierarchical nature of CarbonCast adds robustness to the forecasts against noisy inputs like erroneous source production forecasts, the performance can be further improved by bettering the source production forecasts themselves. In the next section, we show which source productions to focus on for a particular region to get the maximum improvement in performance.

4.5 Feature importance

We evaluate which features are deemed important by CarbonCast for a particular region while forecasting the grid carbon intensity. Fig. 13 shows the top 10 features while predicting the carbon intensity of California using direct emission factors. A higher absolute value means more weightage has been assigned to that particular feature, and it is more important. We see that the carbon intensity forecasts for California rely heavily on solar production forecasts. In general, historical carbon intensity is a strong indicator of future carbon intensity for any region because carbon intensity has both daily and seasonal patterns. For example, carbon intensity is generally lower during the day in California due to solar production. The ordering of features varies across the regions. For example, in Texas, wind speed and wind production forecasts have a high weightage, while coal production forecast is one of the most important features in PJM.

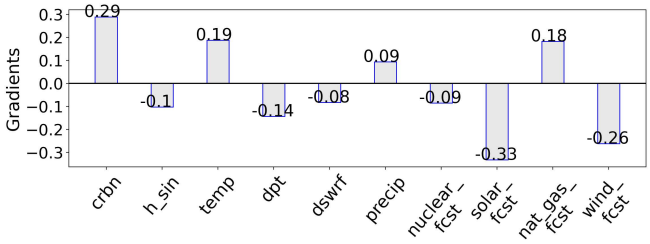


Fig. 13. Features (L to R): Historical carbon intensity, hour of day (h_sin); temperature forecast, dewpoint forecast, solar irradiance (DSWRF) forecast, average precipitation forecast; forecasts of nuclear production, solar production, natural gas production, and wind production.

We inferred from Section 4.4 that even though the hierarchical nature of CarbonCast provides robustness against noise in inputs, the performance can be further improved with better source production forecasts. Our feature importance results tell us which feature to focus on first to most improve CarbonCast's performance. Suppose we want to invest in improving a feature in California. Although natural gas produces the highest fraction of net electricity in California (refer Fig. 1), we should first invest in improving the solar production forecast to get maximum improvement.

4.6 Comparison with state-of-the-art

We now show how CarbonCast compares with the state-of-the-art. Recent works mostly forecast only day-ahead [12, 14, 22] or 48-hour carbon intensity [16], and extending many of those approaches to forecast over a 96-hour period is not straightforward. Some methods provide multi-day forecasts [25] but are proprietary, and we do not have access to their model/data for comparison. Consequently, we compare CarbonCast with the following two recent works, which we could extend to give 96-hour forecasts:

Region	Day 1 forecast			Day 2 forecast			Day 3 forecast			Day 4 forecast		
	$SOTA_1$	$SOTA_2$	CC	$SOTA_1$	$SOTA_2$	CC	$SOTA_1$	$SOTA_2$	CC	$SOTA_1$	$SOTA_2$	CC
CISO	10.71	**6.45**	8.08	18.99	12.26	**11.19**	25.24	16.02	**12.93**	31.64	17.22	**13.62**
PJM	4.27	**3.08**	3.69	7.11	5.51	**4.93**	8.90	7.06	**5.87**	9.77	8.15	**6.67**
ERCOT	14.09	**7.87**	9.78	20.86	12.74	**10.93**	24.46	14.94	**11.61**	26.30	16.21	**12.23**
ISO-NE	5.54	**4.32**	5.10	8.10	9.23	**6.33**	10.07	10.69	**6.97**	11.26	11.57	**7.25**
NYISO	**6.03**	6.84	6.91	10.48	10.56	**9.06**	13.92	13.48	**9.95**	16.29	15.74	**10.42**
BPAT	8.40	**6.99**	7.81	12.42	**10.44**	10.61	14.62	13.18	**12.44**	16.57	15.65	**14.00**
FPL	2.90	**2.39**	2.52	4.38	3.11	**3.01**	5.63	3.65	**3.41**	6.58	4.00	**3.68**
SE	5.10	**3.28**	4.29	7.96	5.92	**5.64**	9.59	6.79	**6.43**	10.17	7.33	**6.74**
DE	15.54	**7.21**	7.81	31.56	11.82	**10.69**	42.16	13.95	**12.80**	50.85	16.57	**15.55**
PL	5.58	**2.37**	3.12	10.29	4.16	**4.14**	12.93	4.85	**4.72**	14.47	5.57	**5.50**
ES	13.45	10.82	**10.12**	26.66	17.57	**16.00**	34.47	20.44	**19.37**	40.08	21.41	**21.12**
NL	8.07	**5.02**	6.06	13.43	**7.52**	7.87	16.34	9.24	**9.08**	18.81	10.64	**9.99**
AUS-QLD	4.43	**3.91**	3.93	7.67	5.06	**3.98**	10.77	5.76	**4.06**	12.99	6.27	**5.87**

Table 7. Daywise MAPE comparison of CarbonCast (CC) versus state-of-the-art methods (lifecycle emission factors).

Region	Day 1 forecast			Day 2 forecast			Day 3 forecast			Day 4 forecast		
	$SOTA_1$	$SOTA_2$	CC	$SOTA_1$	$SOTA_2$	CC	$SOTA_1$	$SOTA_2$	CC	$SOTA_1$	$SOTA_2$	CC
CISO	12.40	**7.51**	9.40	21.92	14.15	**13.23**	29.14	18.34	**15.09**	36.64	19.76	**15.75**
PJM	4.25	**3.32**	3.44	7.08	6.03	**4.56**	8.89	7.97	**5.25**	9.78	9.35	**5.94**
ERCOT	13.69	**7.95**	9.65	20.04	12.36	**10.93**	23.29	14.53	**11.64**	25.02	15.80	**12.29**
ISO-NE	5.69	**4.43**	5.22	8.25	9.46	**6.49**	10.18	10.96	**6.96**	11.34	11.89	**7.15**
NYISO	**6.55**	7.40	7.45	11.29	11.42	**9.64**	15.00	14.59	**10.42**	17.55	16.99	**10.59**
BPAT	10.95	**9.37**	10.25	16.25	**14.08**	14.10	19.44	17.85	**16.39**	22.21	21.16	**18.10**
FPL	3.06	**2.60**	2.93	4.57	**3.36**	3.46	5.86	3.94	**3.77**	6.83	4.32	**4.01**
SE	7.07	**6.52**	7.95	10.87	10.85	**9.79**	13.11	12.28	**11.00**	14.00	13.21	**11.55**
DE	17.59	**8.67**	9.46	35.65	**12.69**	13.46	47.80	**14.80**	15.28	58.13	19.56	**17.52**
PL	5.84	**2.59**	3.50	10.76	4.47	**4.24**	13.58	5.16	**4.85**	15.23	5.90	**5.73**
ES	15.92	12.98	**12.89**	31.49	20.92	**19.39**	41.07	24.23	**22.81**	48.08	25.35	**24.70**
NL	8.56	**5.85**	6.38	14.11	9.24	**8.99**	17.17	11.66	**10.98**	19.84	13.42	**12.38**
AUS-QLD	4.72	**4.20**	4.95	8.14	5.47	**5.41**	11.45	6.24	**5.57**	13.81	6.80	**5.48**

Table 8. Daywise MAPE comparison of CarbonCast (CC) versus state-of-the-art methods (direct emission factors).

1) $SOTA_1$: Bokde et al. [16] decompose the univariate carbon intensity time series into seasonality, trend and noise components. Then, they forecast each component separately using techniques like Feed Forward Neural Network (FFNN) and ARIMA and recombine to get 48-hour carbon intensity forecasts. For the purpose of comparison, we implement a representative univariate Seasonal-ARIMA (SARIMA) model with a 96-hour forecasting period as an example of such a method and compare it with CarbonCast.

2) $SOTA_2$: DACF [12] forecasts each source individually and then uses Eq. 1 to get day-ahead average carbon intensity forecasts. Their approach is similar to our first tier, and their code is publicly available. So, we extended their approach to forecasting carbon intensity for 96 hours. They used direct emission factors, but converting to lifecycle emission factors for our comparison is straightforward.

Other works on forecasting short-term carbon intensities either mostly use a combination of the methods mentioned above [22] or have limitations in extending the forecast period [14]. Thus, we claim that the above two implementations fairly represent any state-of-the-art multi-day carbon forecasting systems.

Table 7 provides a day-wise comparison of all the approaches in terms of MAPE, with the best-performing method highlighted in bold. We see that CarbonCast almost always outperforms $SOTA_1$ [16]. When compared with $SOTA_2$ *(DACF)* [12], there is a clear distinction wherein $SOTA_2$ gives better forecasts initially, but as the

input data (weather forecasts, source production forecasts etc.) get progressively worse with an increasing forecasting period, CarbonCast starts to perform better as it is more robust to input noise. For most of the regions, CarbonCast is better than $SOTA_2$ from day 2 onwards. In some regions, $SOTA_2$ has a better MAPE till day 2 or even day 3. However, CDF plots of both approaches in these cases show that their results are still comparable, and the above conclusion still holds. The results are similar with both lifecycle and direct emission factors.

When aggregated over 96 hours for the whole test period, CarbonCast using direct (resp. lifecycle) emission factors has an average MAPE reduction of 9.96% (resp. 8.91%) across the regions.

4.7 CarbonCast runtime overheads

Finally, we break down the time taken by CarbonCast to generate 96-hour forecasts using commodity hardware. The first tier takes 2 secs per epoch to train. We limited the number of epochs to 100. So, the first tier takes at most 200 secs to train. We forecast six months at a time, which takes $\sim$ 30 secs (forecasting a single 96-hour period takes 0.15 secs on average). The time taken to generate such forecasts is similar for all the sources, and since the source production forecasts can be generated in parallel, we say that the first tier runtime upper bound is $\sim$ 4 mins. The second tier can only be run after the first tier forecasts are obtained, and takes 9 secs on average per epoch during training. Hence, it takes at most 15 mins to train (max. 100 epochs). After training, forecasting a 4-day period takes 0.5 secs on average.

We see that CarbonCast takes ~20 mins, with most of the time taken during training. Usually, the forecast accuracy of CarbonCast decreases as the forecast period increases. To counter the inaccuracies due to a longer forecasting period of 96 hours, we can train CarbonCast periodically (say, every n^{th} day) and generate 96-hour forecasts daily. This method will result in CarbonCast taking ~20 mins on day one and taking $<$ 1 sec on days $2 - n$, and enable practitioners to update their carbon optimization decisions daily with an updated and more accurate forecast if required.

5 DISCUSSION

We now discuss aspects of CarbonCast that pertain to its architecture and use.

5.1 CarbonCast compatibility and flexibility

From the results in Sections 4.2 and 4.6, we conclude that CarbonCast is a better system for multi-day forecasts than the current state-of-the-art. Additionally, CarbonCast's flexibility gives practitioners the option to use CarbonCast along with approaches like DACF [12]. Both approaches can be run in parallel, and DACF [12] can be used for shorter forecasting periods, while CarbonCast can replace DACF [12] as its performance starts to degrade. For example, carbon optimization decisions for $0 - 24$ hours can be based on DACF [12] forecasts, with the following hours' forecasts coming from CarbonCast. Another benefit of CarbonCast's hierarchical approach is that if some source production forecast is highly erroneous and degrades the overall performance, it can be removed since there are other features that can compensate for the removed input, which increases robustness.

5.2 Effects of electricity imports and exports on CarbonCast

When considering electricity imports, the average carbon intensity of a region can be considered as a weighted average of the electricity generated in that region and the carbon intensities of each importing region. Thus, for any region, if the percentage of electricity imported from other regions is small in comparison with the amount of electricity generated or if the importing regions have a similar average carbon intensity to that region, the average carbon intensity of that region will not differ significantly with or without imports.

If that is not the case, the carbon intensity of imported electricity may be an important factor in calculating the average carbon intensity of a region. In this case, more experiments are needed to understand the effects of imported carbon intensity. However, calculating the amount of electricity imported from a particular region is complex, and we treat it as an important future work. Having said that, CarbonCast framework will still work if we have imported carbon intensity data without any design changes. If we have such data, we can treat that as another electricity-generating source in the region with carbon emission factor equal to the carbon intensity value, and add it as an input to CarbonCast for getting carbon intensity forecasts.

5.3 Average versus marginal carbon intensity forecasting

In this work, we focused on average carbon intensity, which is defined to be the weighted average of the carbon emission factors (CEFs) of all sources in a particular region. However, not all sources contribute to each new unit of electricity generated in a region. Typically, the generators of only a subset of the sources are ramped up to produce the incremental amount of energy needed to satisfy an incremental amount of new demand. We call these sources to be on the margin. The marginal carbon intensity per unit of electricity generated in a region is the weighted average of carbon emission factors of *only those sources that are on the margin*.

Both average and marginal carbon intensities are relevant metrics that are applicable in different contexts. For instance, consider the use of carbon intensity forecasts for building a carbon-aware load balancer. If we wish to redirect a small incremental load to data centers in greener regions in real-time, and we wish to know the incremental carbon impact of that action, *marginal carbon intensity* is a better metric for quantifying that impact. However, if we are delaying the execution of workloads to greener times, or if the workload is big enough to change the electricity sources in the margin, *average carbon intensity* may be a better metric to use.

In the longer term, we would like CarbonCast to provide forecasts of both types of carbon intensities and leave it to the application developers to decide what types of forecasts to use. We started by providing average carbon intensity forecasts since it is much easier to obtain the ground truth and compute the MAPE for this metric. Ground truth data for marginal carbon intensity is seldom available. We mentioned in Section 4.2.1 that hourly electricity generation

data is mainly available in the US, Europe, and Australia. However, even within the aforementioned regions, grid operators generally only publish how much electricity is generated by each source in that region but no information about which electricity sources are on the margin at a specific time, rendering the calculation of marginal carbon intensity difficult. Although Watttime [38] publishes marginal carbon intensity forecasts, they do not disclose their data sources, and hence it cannot be used by other forecasting solutions. To the best of our knowledge, only PJM has recently started publishing marginal emissions data [27]. As more grid operators start publishing marginal emissions data, CarbonCast can use that to provide multi-day marginal carbon intensity forecasts, and this is listed as future work.

6 RELATED WORK

Predicting carbon intensity is becoming increasingly popular for solving problems like smart charging of electric vehicles [13], reducing carbon footprints in residential heating [17], and other carbon reduction optimizations. Lowry [22] provides grid carbon intensity forecasts for heating, ventilation and air-conditioning (HVAC) systems using only historical data. Leerbeck et al. [14] forecast grid carbon intensity for Denmark using linear regression and ARIMA. These works have a flat design, whereas CarbonCast has a hierarchical architecture. CarbonCast also considers source production forecasts for all electricity-generating sources in a region, unlike [14, 22]. DACF [12] uses an approach similar to CarbonCast's first tier, but CarbonCast has an additional tier of deep learning models, making it more robust to inaccuracies in the inputs. Tomorrow's ElectricityMap [9] provides carbon intensity forecasts for many regions but is a proprietary service: its models are not public, and the data is available at a cost. Additionally, all these techniques provide only day-ahead carbon intensity forecasts, while CarbonCast can forecast up to 96 hours. Among the techniques providing multi-day forecasts, National Grid ESO [25] provides freely accessible APIs [10], but they are constrained to the UK region since neither their data nor models are available publicly. Watttime [38] provides up to 72-hour marginal carbon intensity forecasts, whereas CarbonCast provides 96-hour average carbon intensity forecasts. Besides, Watttime [38] also have the same problems as [9]. Bokde et al. [16] use decomposition techniques and statistical methods to get 48-hour forecasts. However, similar to [22], they also use only historical data and hence suffer from high forecasting errors, whereas CarbonCast uses future knowledge to get more precise forecasts.

7 CONCLUSIONS

In this paper, we presented CarbonCast, an open-source two-tiered hierarchical modelling framework to provide grid carbon intensity forecasts for up to 96 hours. CarbonCast obtains source production forecasts from its first tier and then combines all source forecasts with weather forecasts and historical data in the second tier to compute a carbon intensity forecast. Our results show that our hierarchical design makes CarbonCast robust against the uncertainty associated with a longer forecasting period. CarbonCast has a MAPE of 9.78% (resp. 8.38%) across the regions using direct (resp. lifecycle) emission factors. It achieves an average decrease of 9.96% (resp. 8.91%) in MAPE over the 96-hour forecasting period compared to the state-of-the-art approaches. We also show which source production forecasts are crucial to obtaining precise carbon intensity forecasts in a particular region. Further, its plug-and-play framework provides the flexibility to choose the best-performing model for each region while also providing a general approach that works well across various geographically distributed electric grids.

CarbonCast is the first open-source tool for multi-day forecasting, with both code and data freely available for researchers. We hope that CarbonCast will enable more research in carbon-aware systems that require carbon intensity forecasts. As future work, we plan to extend CarbonCast to even more regions, provide marginal carbon intensity forecasts, incorporate the impact of energy exchange between electric grids, generate sub-hourly forecasts and also increase its forecasting period further.

ACKNOWLEDGMENTS

This work is supported in part by NSF grants 2105494, 2021693, and 2020888, and a grant from VMware.

REFERENCES

[1] US Energy Information Administration. 2018. Real-time Operating Grid. Retrieved July 28, 2022 from https://www.eia.gov/electricity/gridmonitor/dashboard/electric_overview/US48/US48
[2] An Open Platform for National Electricity Market Data. 2022. OpenNEM. Retrieved December 30, 2022 from https://opennem.org.au/energy/nem/?range=7d&interval=30m
[3] CAISO. 2022. California Independent System Operator. Retrieved July 28, 2022 from http://www.caiso.com/Pages/default.aspx
[4] California ISO. 2005-2022. Open Access Same-time Information System (OASIS). Retrieved July 28, 2022 from http://oasis.caiso.com/mrioasis/logon.do
[5] François Chollet et al. 2015. Keras. https://keras.io.
[6] Climate Change. 2014. Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]. IPCC, Geneva, Switzerland, 151 pp. Retrieved July 28, 2022 from https://archive.ipcc.ch/pdf/assessment-report/ar5/wg3/ipcc_wg3_ar5_annex-iii.pdf#page=7
[7] Department of Business, Energy and Industrial Startegy. 2021. Greenhouse gas reporting: conversion factors 2021. Retrieved September 29, 2022 from https://www.gov.uk/government/publications/greenhouse-gas-reporting-conversion-factors-2021
[8] ElectricityMap. 2020. Zone bounding boxes. Retrieved July 28, 2022 from https://github.com/electricityMap/electricitymap-contrib/blob/master/config/zones.json
[9] ElectricityMap. 2022. Retrieved July 28, 2022 from https://electricitymap.org/
[10] A. Bruce et al. 2021. Carbon intensity forecast methodology. *National Grid ESO: Warwick, UK*. 20 (2021). https://github.com/carbon-intensity/methodology/
[11] A. Radovanovic et al. 2021. Carbon-aware computing for datacenters. *arXiv preprint arXiv:2106.11750* (2021).
[12] D. Maji et al. 2022. DACF: Day-Ahead Carbon Intensity Forecasting of Power Grids Using Machine Learning. In *Proceedings of the Thirteenth ACM International Conference on Future Energy Systems (e-Energy '22)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3538637.3538849
[13] J. Huber et al. 2021. Carbon efficient smart charging using forecasts of marginal emission factors. *Journal of Cleaner Production* 284 (2021), 124766.
[14] K. Leerbeck et al. 2020. Short-term forecasting of CO2 emission intensity in power grids by machine learning. *Applied Energy* 277 (2020), 115527.
[15] M. Abadi et al. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. http://tensorflow.org/ Software available from tensorflow.org.
[16] N. D. Bokde et al. 2021. Short-term CO2 emissions forecasting based on decomposition approaches and its impact on electricity market scheduling. *Applied Energy* 281 (2021), 116061.
[17] P. J. C. Vogler-Finck et al. 2018. Reducing the carbon footprint of house heating through model predictive control–A simulation study in Danish conditions. *Sustainable Cities and Society* 42 (2018), 558–573.
[18] P. Wiesner et al. 2021. Let's wait awhile: how temporal workload shifting can reduce carbon emissions in the cloud. In *Proceedings of the 22nd International*

Middleware Conference. 260–272.

[19] European association for the cooperation of transmission system operators. 2008. ENTSOE transparency platform. Retrieved July 28, 2022 from https://transparency.entsoe.eu/

[20] Organisation for Economic Co-operation and Development. 2023. Emission trading systems. Retrieved March 9, 2023 from https://www.oecd.org/env/tools-evaluation/emissiontradingsystems.htm#:~:text=Under%20a%20baseline%2Dand%2Dcredit,regulations%20they%20are%20subject%20to.

[21] International Energy Agency. 2019. Global Energy & CO2 Status Report 2019: Emissions. Retrieved July 28, 2022 from https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data

[22] Gordon Lowry. 2018. Day-ahead forecasting of grid carbon intensity in support of heating, ventilation and air-conditioning plant demand response decision-making to reduce carbon emissions. *Building Services Engineering Research and Technology* 39, 6 (2018), 749–760.

[23] Luke George. 2021. The Correct Way to Average the Globe. Retrieved July 28, 2022 from https://towardsdatascience.com/the-correct-way-to-average-the-globe-92ceecd172b7

[24] National Centers for Environmental Prediction, National Weather Service, NOAA, U.S. Department of Commerce. 2015. NCEP GFS 0.25 Degree Global Forecast Grids Historical Archive. Retrieved July 28, 2022 from https://doi.org/10.5065/D65D8PWK

[25] National Grid ESO. 2022. National Grid ESO. Retrieved July 28, 2022 from https://www.nationalgrideso.com/

[26] Sundar Pichai. 2020. Our third decade of climate action: Realizing a carbon-free future. *The Keyword* (2020). Retrieved July 28, 2022 from https://blog.google/outreach-initiatives/sustainability/our-third-decade-climate-action-realizing-carbon-free-future/

[27] PJM Data Miner. 2017. Five Minute Marginal Emission Rates. Retrieved December 30, 2022 from https://dataminer2.pjm.com/feed/fivemin_marginal_emissions/definition

[28] PJM Inside Lines. 2020. Locational Marginal Pricing Explained. Retrieved September 29, 2022 from https://insidelines.pjm.com/locational-marginal-pricing-explained/

[29] United Nations. 2020. Carbon neutrality by 2050: the world's most urgent mission. Retrieved July 28, 2022 from https://www.un.org/sg/en/content/sg/articles/2020-12-11/carbon-neutrality-2050-the-world%E2%80%99s-most-urgent-mission

[30] United States Environmental Protection Agency. 2022. Scope 1 and Scope 2 Inventory Guidance. Retrieved July 28, 2022 from https://www.epa.gov/climateleadership/scope-1-and-scope-2-inventory-guidance

[31] United States Environmental Protection Agency. 2022. Scope 3 Inventory Guidance. Retrieved July 28, 2022 from https://www.epa.gov/climateleadership/scope-3-inventory-guidance

[32] US Energy Information Administration. 2021. EIA projects nearly 50% increase in world energy use by 2050, led by growth in renewables. Retrieved July 28, 2022 from https://www.eia.gov/todayinenergy/detail.php?id=49876#

[33] US Energy Information Administration. 2021. Electricity explained: How electricity is delivered to consumers. Retrieved July 28, 2022 from https://www.eia.gov/energyexplained/electricity/delivery-to-consumers.php

[34] US Energy Information Administration. 2021. Electricity explained: Use of electricity. Retrieved July 28, 2022 from https://www.eia.gov/energyexplained/electricity/use-of-electricity.php

[35] US Energy Information Administration. 2022. EIA tweet: Our Hourly Electric Grid Monitor remains offline. Retrieved July 28, 2022 from https://twitter.com/EIAgov/status/1549826047008546822

[36] US Environmental Protection Agency. 2021. Global Greenhouse Gas Emissions Data. Retrieved July 28, 2022 from https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data

[37] US Environmental Protection Agency. 2021. Sources of Greenhouse Gas Emissions. Retrieved July 28, 2022 from https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions#:~:text=Larger%20image%20to%20save%20or%20print%20The%20Electricity%20sector%20involves,2O)%20are%20also%20emitted.

[38] Watttime. 2022. Retrieved July 28, 2022 from https://www.watttime.org/