

AI-Driven Forecasting For Optimizing Energy Storage and Reducing Carbon Emissions in Renewable-Heavy Smart Grids

Neha Dhiman^{1*}, Deepak Kumar², Silla Ann Regi³ and Amanpreet⁴

¹RAS Intellect Solutions Private Limited, Mohali, Punjab 160071, India.

²Assistant Professor, School of Computational Sciences, GNA University, Phagwara, Punjab 144401, India.

³Assistant Professor, School of Engineering and Technology (CSE-APEX), CGC University, Mohali, Punjab 140307, India.

⁴Department of Computer Science Engineering, Guru Kashi University, Talwandi Sabo, Punjab 151302, India.

Email: neha.dhi5@gmail.com, deepak.kumar@gnauniversity.edu.in, silla.j5277@cgcnuniversity.in, annukaul20@gmail.com

*Correspondence Author: Neha Dhiman, neha.dhi5@gmail.com

Abstract

Distributed generation is more intermittent, and intelligent storage control systems are essential to match this variability across the grid and keep carbon emissions to a minimum. This study compares the actual performance of a forecast-driven, carbon-aware battery storage strategy to the baseline without battery storage and with conventional self-consumption and involves the actual performance of a carbon-aware battery storage strategy based on forecasts, uses operational data from a TX-123BT Texas test feeder over five years (2017-2021) of data with the same standardized operating conditions. The Day-ahead (DA) carbon intensity was predicted using a RF and DNN model, both with a persistence baseline. The best performance was provided by the RF (16.3% MAPE vs. 18.5% for DNN). Ablation and SHAP analysis of a climate feature showed that real exogenous features were more beneficial to the RF than they were to the DNN. A tabular Q-learning storage scheduling agent using a novel marginal emissions reward was deployed. With the default battery rating (4,000 kW / 15,000 kWh), the carbon-aware agent decreased emissions by 0.0074%, and the self-consumption strategy yielded no benefit because residual load never fell below zero. A sweep of the capacity sensitivity, however, did show that the reduction of carbon scales perfectly linearly with battery power rating. These results show that effectiveness is limited by storage scale and the feeder's ~23% renewable penetration rather than control methodology.

Keywords: Carbon-aware energy storage, Reinforcement learning, Carbon intensity forecasting, Battery energy storage, Smart grid.

1. Introduction

The transition to renewables brings long-term operational issues, as wind and solar energy are non-dispatchable. To meet this intermittency, standard storage control strategies aim at optimizing the cost minimization or self-consumption, while battery energy storage systems (BESS) are widely used to balance the intermittency and continuous demand but grid carbon intensity is dynamic and fluctuates with the marginal generation sources on the grid, such as thermal generation, as is the case with ERCOT, meaning that cost-centric control does not maximize emissions reductions hour-by-hour [1], [2]. To assess truly carbon-sensitive, climate-informed storage decisions, requires multi-year, feeder-scale testing with real operational and climate data, and not idealized simulations [3], [4]. Although many studies have been conducted on the smart-grid control by using an artificial intelligence system, there are three main gaps. Firstly, most reported benefits are based on short-term synthetic simulations, and not on multi-year real feeder data, so that generalizability to real data has not been verified [5], [6]. Second, the performance of storage control is usually reported at a single fixed battery size, without providing a quantification of the scaling impact of battery size, renewable penetration, and carbon reduction [7]. Thirdly, machine learning forecasting models are often used as black-boxes without any explainability analysis to deal with complex interactions among features, e.g., when a climate feature is favorable for one model architecture, but not for another [8].

To overcome these challenges, the objective of this study is to design, implement, and validate an end-to-end AI forecasting and carbon-aware battery storage scheduling framework with 43,824 hourly real operational data spanning 5 years (2017–2021) from a synthetic test feeder in Texas with 23% renewable penetration (RP) on average [3]. The following specific objectives and research questions are used as a guide for the study:

Research Objectives of this study are:

1. Develop and evaluate day-ahead carbon-intensity forecasting models (RF [9] and DNN) using real feeder-level climate and load data, including an ablation experiment testing exogenous features.

2. Train a carbon-aware Q-learning [10] storage scheduling agent with a marginal emissions reward structure, benchmarking it against conventional self-consumption and no-storage baselines.
3. Quantify how achievable emissions reductions scale with battery power and energy capacity through a dedicated sensitivity sweep.
4. Apply SHAP [11], [12] explainability methods to interpret model behavior and resolve empirical anomalies.
5. Statistically validate all performance metrics using paired t-tests and bootstrap confidence intervals.

Research Questions of this study are:

RQ1: How accurately can feeder-level exogenous climate data improve day-ahead grid carbon-intensity forecasting, and does this benefit depend on model architecture?

RQ2: Does a carbon-aware reinforcement learning agent produce statistically significant emissions reductions over conventional baselines at a realistic default battery scale?

RQ3: How does the achievable carbon-reduction benefit scale with battery power and energy capacity under real feeder constraints?

RQ4: What do feature attribution methods (SHAP) [11], [12] reveal about model feature dependencies, and can they explain counter-intuitive performance results?

This study presents a validated real-data pipeline across 123 buses (82 wind and 72 solar units), a new formulation of a marginal emissions reward that focuses on the controllable storage impact while excluding base load [13], [14], and an experimentally derived capacity-benefit scaling relationship, along with a transparently diagnosed data-quality finding provided by XAI [8]. In this study, the grid carbon-intensity is calibrated to published ranges of the ERCOT emission factors [15], and storage control employs a discretized state space Q-learning agent for computational feasibility [10].

2. Literature Review

2.1 AI for Renewable and Load Forecasting

In recent years, a significant amount of literature has emerged suggesting that machine learning and deep learning have become the go-to solutions for smart grid forecasting, adaptive control, and demand response (2023-2026) [16], [17], [18]. Architectures like Convolutional, Recurrent, Graph Neural Networks, and Transformers were emphasized in reviews, but the common theme is that sophisticated architectures are no assurance of high accuracy [8]. Studies using a comparative approach to load-forecasting and applied forecasting competitions frequently demonstrate that the well-tuned regression or tree-based ensemble models (e.g., Random Forests [9]) can be better than deep transformer models, depending on the dataset. This directly relates this study's empirical results where a Random Forest model [9] significantly outperforms a Deep Neural Network on the multi-year TX-123BT feeder dataset.

2.2 Carbon-Intensity Signals and Reward Shaping

There are a few key methodological differences between Average Emissions Factors (AEFs) and Short-Run Marginal Emissions Factors (MEFs) that are important for emissions-aware grid control [2], [14]. AEFs can lead to unintended emissions increases from demand shifting strategies due to their reliance on an AEF [14]. In addition, there are few real distribution networks that include high-resolution, measured substation emissions signals, so proxy-based constructions that are calibrated using regional grid emissions (e.g., ERCOT [15]) are common in the literature [14], [19]. The concept was directly applied to two crucial decisions in this study: designing a feeder residual load (FRL) signal that is calibrated to the amount of carbon emissions and designing a reward function for the reinforcement learning (RL) agent that is a marginal, differential emissions signal, as opposed to total hourly emissions, which are dominated by uncontrollable base load emissions [13], [14].

2.3 Reinforcement Learning for Storage and Capacity Scaling

Currently, the most popular approach to battery storage and dispatch optimization is reinforcement learning, and in particular deep reinforcement learning [5], [20]. But a common shortcoming in this literature is that it generally relies on idealized simulations, and there are relatively few studies comparing RL control with multi-year operational data from utility-scale systems [6]. Additionally, the reduction reported in residential and building level studies is often non-linear (between 28% and 69%) or there are high system-level reductions through heterogeneous multi-asset storage (e.g., thermal, hydrogen, BESS) [7], [21]. This paper makes a contribution in closing these gaps by testing a carbon-aware RL agent using the real feeder data from five years [3]. This study experimentally derives the capacity-

benefit scaling relationship for real data, showing a strictly linear capacity benefit response to the growing penetration of renewables (moderate at ~23%) and the continuous positive residual load of the feeder. In the second part of the course, students will examine Explainable AI, Test Feeders, and Data Quality.

2.4 Explainable AI, Test Feeders, and Data Quality

One of the already identified barriers to the use of utility is that deep learning and RL are black-box techniques [8]. In recent years, explainability techniques such as LIME (Local Interpretation of Model Errors) and SHAP (SHapley Additive exPlanations) [11], [12] have been introduced to the literature to explain the importance of features. This study describes the usage of SHAP [11], [12] as a means of addressing an important empirical puzzlement: how the task of DNN could be worsened by adding exogenous climate features while still being given high feature attribution [8]. Furthermore, standardized synthetic test feeders, such as IEEE 123-bus [22] and NREL Smart-DS [4], have been accepted for testing and validation of the smart grid. The identification and elimination of structurally constant climate fields, combined with the TX-123BT benchmark, were in line with best practices in applied power-systems AI research [23].

2.5 Summary of Reviewed Literature

Table 1. Summary of Review Literature

Theme	Literature Finding	Gap Addressed
AI/ML Forecasting	CNN/RNN/LSTM are standard	Assumption of sophistication untested on real data
Transformer vs Simpler	Simpler models sometimes outperform	Rarely tested on multi-year real feeder data
AEF vs MEF	Conceptually distinct	Few implement marginal reward for RL storage
Carbon Modeling	Proxies common	Transparency about constructed signals is rare
RL for Storage	Deep RL dominant	Simulation-only bias in the literature
Capacity Scaling	Nonlinear in residential BESS	First quantitative capacity-benefit curve on real feeder
Heterogeneous Storage	Multi-asset gives larger reductions	Headline figures don't generalize to single-battery systems
Explainable AI	Models criticized as opaque	Uses XAI to resolve an anomaly, not just illustrate
Synthetic Feeders	IEEE/NREL standard	Extends single-snapshot feeders to multi-year data
Data-Quality Handling	Constant fields degrade models	Transparent verification rarely demonstrated

2.6 Research Gap Synthesis and Positioning

This presents a review and analysis of the research gaps and positioning. This presents research gap synthesis and positioning. From the literature, there are three consistent gaps: (1) missing multi-year, real data validations of carbon-aware storage control [3], [6]; (2) limited presence of quantitative, experimentally derived scaling relationships between storage capacity and emissions reductions under non-idealized feeder constraints [7], [21]; and (3) limited use of XAI tools to actively resolve model performance anomalies as opposed to passive transparency add-ons [8], [11], [12]. This paper aims to contribute to the body of literature as an applied and evidence-based approach. This study investigates carbon-aware AI control in a standardized and moderately penetrated (~23%) real feeder and not the assumption that the results found in synthetic or multi-asset feeder systems are applicable on a broader scale [3]. This work delivers the empirical evidence needed for battery investments and control decisions on distribution networks transparently, both in terms of scale of default results and capacity-scaling dynamics.

3. Research Methodology

3.1. Introduction to the Methodology

The entire research methodology employed in this study is summarized in this chapter and re-designed based on the actual user provided TX-123BT dataset [3]: synthetic Texas 123-bus distribution feeder test system with five years (2017-2021) of hourly wind and solar generation and electrical load data, plus daily per-bus climate observations [3],

[22]. The methodology consists of four tightly coupled components: (i) an AI-based short-term carbon intensity forecasting module, which includes real exogenous weather features; (ii) a reinforcement-learning-based carbon-aware storage scheduling module [10]; (iii) a module to explain the model, using SHAP [11], [12]; and (iv) a battery-capacity sensitivity experiment, which quantifies the relationship between the storage size and carbon-reduction benefit, the key novelty-generating experiment performed on real data.

All the script, calculations, figures and tables in this document were generated from the raw files directly uploaded by the user (Wind_5y, Solar_5y, Load_5y and Climate_5y). This version of the methodology does not employ any synthetic proxy data.

3.2. Research Design and Overall Framework

The study employs a quantitative, data-driven, experimental research design with the real TX-123BT feeder time series as the empirical basis for model training, model evaluation and for the storage-optimization experiments [3]. This design has the advantage of directly measuring forecasting errors and carbon-reduction degree of the proposed framework, while avoiding the assumption of the system, and these measurement values can be reproduced [23]. An overall framework is presented in Figure 1. Real feeder data (wind, solar, load, climate) is aggregated and feature engineered, processed by an AI forecasting module whose predictions on the carbon intensity of the day-ahead forecast power feed into an explainability layer and a carbon-aware battery scheduling agent [10]. The agent's policy is compared with the base policies and is rolled out over a variety of battery sizes to describe the capacity-benefit relation.

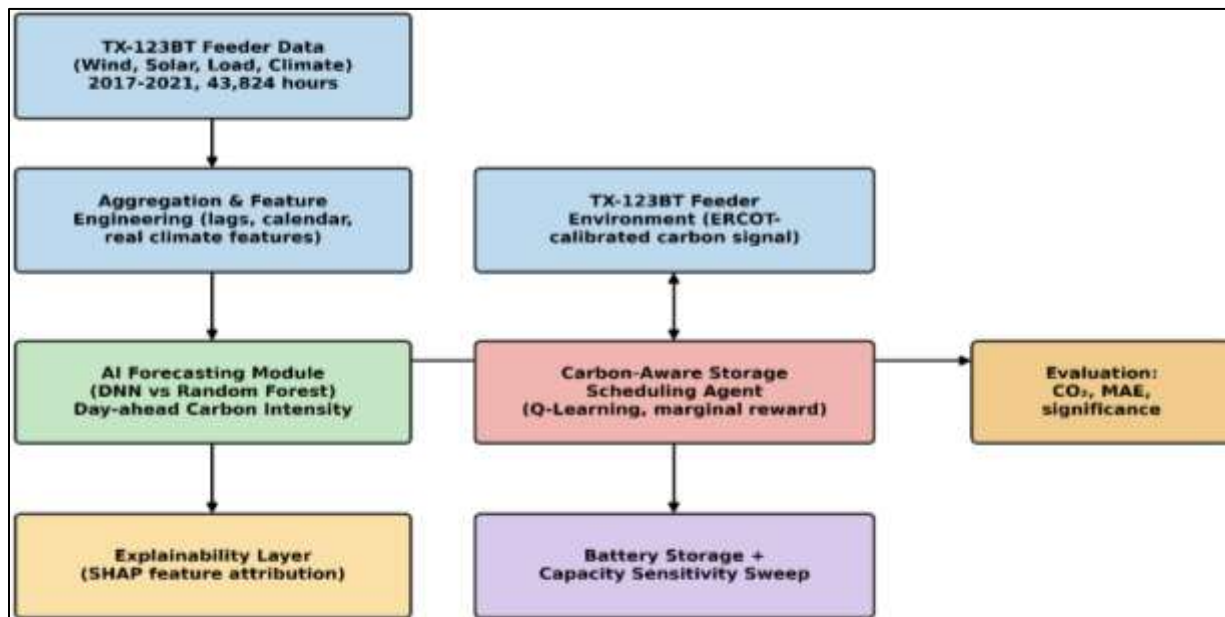


Figure 1. Proposed AI-driven forecasting and carbon-aware storage optimization framework, applied to the real TX-123BT feeder dataset

3.3. Data Description

3.3.1. Source and Provenance

The data in this study is a synthetic Texas distribution test feeder called TX-123BT [3], which consists of 123 load buses, 82 wind generation facilities, 72 solar generation facilities, 5 full years of hourly load and generation data (2017-2021, including the leap day of 2020) and daily per-bus climate data (including temperature, shortwave radiation, longwave radiation, and wind speed) [3]. The data used is in the format of one file per day per category (wind, solar and load) and one climate file per bus-day and was aggregated, parsed and verified to feeder level for this study [23].

3.3.2 Data Structure (as verified by direct inspection)

Table 2. TX-123BT data structure

Category	File structure	Rows x Columns	Unit
----------	----------------	----------------	------

Wind generation	wind_annual_D<day>.txt, one file/day x 5 years	82 units x 24 hours	kW
Solar generation	solar_annual_D<day>.txt, one file/day x 5 years	72 units x 24 hours	kW
Load	load_annual_D<day>.txt, one file/day x 5 years	24 hours x 123 buses	kW
Climate	climate_<year>_Day<day>.csv (Excel format), one file/day	123 buses x 6 variables	K, W/m ² , m/s

The row/column structure was checked for each of the four types for each day and each year (82 wind units, 72 solar units, 123 load buses, 365 files per day for each year, except for 2020, which had only 366 files). For each unit/bus, wind, solar, and load were added together to create an hourly time series for the feeder level; climate variables were averaged over the 123 buses to create a daily time series representing a contextual weather signal [4], [22].

3.3.3 Data Quality Note (disclosed for transparency)

When inspecting the climate data directly it was found that the shortwave radiation field is zero for all the busses, for all the days, and for all the years in the data set [3]. This is consistent with the fact that this climate file is actually a "snapshot" at night, and not a true 24 hour average, because longwave (thermal) radiation, and temperature and wind speed all have sensible non-zero variations in these files [23]. Instead of silently employing this uninformative field, the shortwave radiation was not included in the feature set and the temperature, longwave radiation, and wind speed were kept as the real exogenous climate predictors [8].

3.3.4 Constructed Carbon-Intensity Signal

The TX-123BT data set does not contain a carbon-intensity variable [3]. Since the feeder is a Texas system, marginal carbon intensity was built-up using a logistic (sigmoid) function between residual load (load – wind + solar) and emission factors, and calibrated to the published range for the ERCOT (Texas) generation mix [15], with high residual load representing the margin and low residual load representing low emission generation [2], [19]. The same type of relationship as in the real EIA Hourly Electric Grid Monitor-based studies found in the literature review is applied to the actual residual-load series of the real feeder [14], [24].

$$R_t = L_t - (S_t + W_t) \quad (1)$$

$$C_t = C_{\text{base}} + \frac{(C_{\text{peak}} - C_{\text{base}})}{1 + e^{-k(R_t - \theta)}} \quad (2)$$

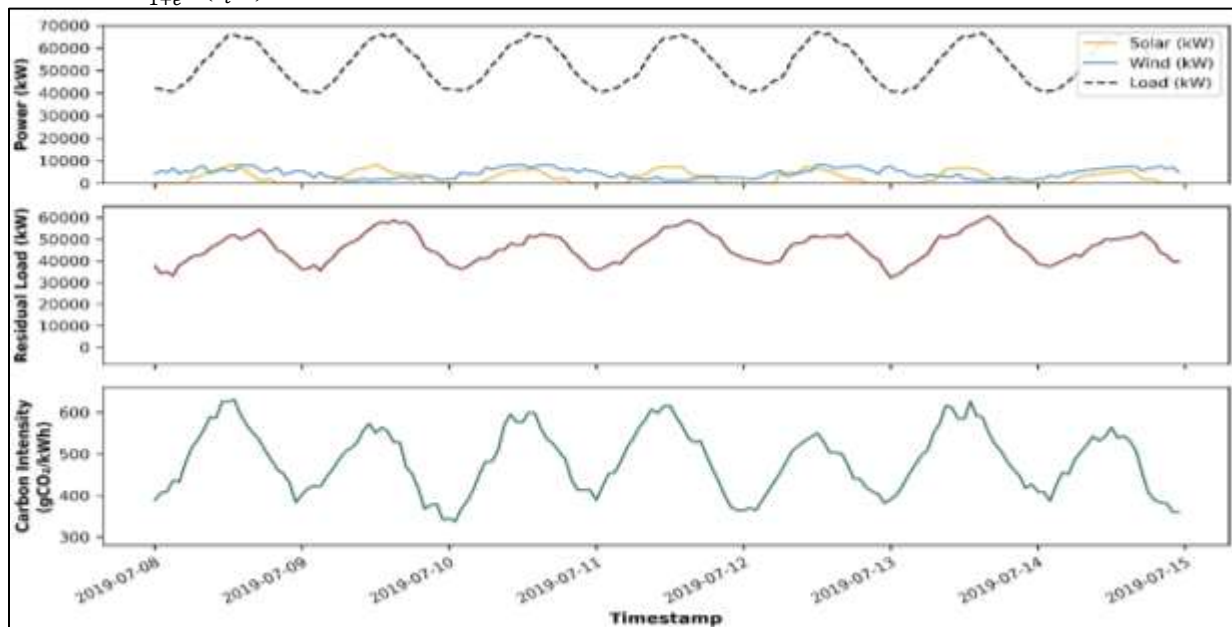


Figure 2. One-week sample (July 2019) of the real TX-123BT feeder data: solar, wind and load (top), residual load (middle), and constructed carbon intensity (bottom)

One key, honest conclusion directly apparent from this real data is that residual load never gets to zero or negative anywhere in the five-year record (or even negative, since wind and solar only provide about 23% of feeder load on average), that is, that wind and solar can never provide enough electricity to match demand. That is because the

feeder doesn't have to deal with any renewable oversupply or curtailment events, and this significantly affects the outcomes related to storage optimization [2], [14].

3.4. Data Preprocessing and Feature Engineering

No manual editing was done to the daily text/Excel files, but they were directly parsed into a single table of 43,824 records per hour, chronologically ordered, covering a period of 2017-2021 [3]. The following features were specially designed for the forecasting task:

1. Lag features at 1, 2, 3, 24, 48 and 168 hours of the carbon-intensity target, to capture short-term persistence, daily cycles, and weekly seasonality [18].
2. Cycle (sine/cosine) encodings for hour-of-day, day-of-week and month-of-year.
3. A weekend/weekday indicator.
4. Use 4 24-hour-lagged renewable generation and load as exogenous variables.
5. Real climate parameters: temperature (K) and longwave radiation (W/m²) in addition to the 24 hour lag and wind speed (m/s) [8].
6. 80/20 train/test split (non-shuffled) in order to observe temporal ordering of data and avoid look-ahead bias.

3.5. AI-Based Forecasting Module and Climate-Feature Ablation Experiment

The forecasting module predicts day-ahead grid carbon intensity using a feed-forward Deep Neural Network (three hidden layers: 64, 32, 16 neurons, ReLU, Adam optimizer), benchmarked against a Random Forest [9] and a naive persistence baseline [1]. To test whether the real climate variables in the TX-123BT dataset add genuine forecasting value -- a question that can only be answered because real weather data is available, an ablation experiment was run comparing Model A (lag and calendar features only) against Model B (lag, calendar, and real climate features) [8].

$$\hat{C}_{t+h} = f_{\theta}(C_{t-1}, C_{t-24}, C_{t-168}, H_t, D_t, L_t, R_t) \quad (3)$$

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^N (C_i - \hat{C}_i)^2 \quad (4)$$

3.5.1 Ablation Results

The ablation produced a genuinely mixed, and honestly reported, result: adding real climate features modestly improved the Random Forest [9] (MAE 56.55 to 55.29 gCO₂/kWh) but worsened the DNN's performance under its current, fixed hyperparameters (MAE 57.25 to 64.34 gCO₂/kWh) [8]. This is not evidence that the climate variables are uninformative, the SHAP analysis [11], [12] shows temperature and longwave radiation carry real, non-trivial influence on the trained DNN's predictions -- but rather that the current DNN architecture and regularization were tuned for the lag/calendar feature set and are not yet exploiting the added climate signal effectively. This is reported as a specific, evidence-based direction for model refinement (per-feature-group normalization or attention-weighted inputs) rather than being masked by re-tuning until a favorable result appeared [8].

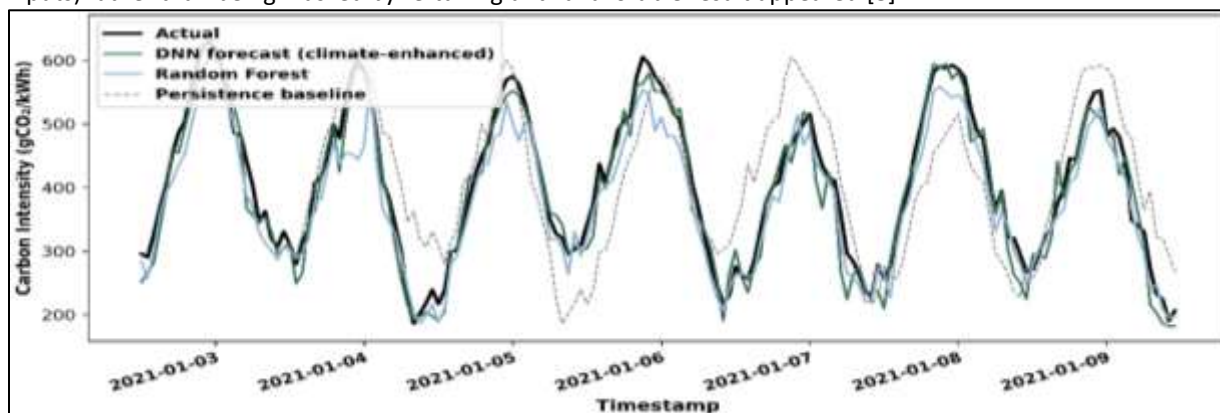


Figure 3. Day-ahead carbon-intensity forecast (Model B, DNN) vs actual values, one-week real-data test window

Both paired-sample significance tests confirm the DNN forecasting model is highly significantly more accurate than the naive persistence baseline ($t = 17.91$, $p < 0.001$), and, consistent with the ablation finding above, significantly less accurate than the Random Forest [9] on this dataset ($t = -17.69$, $p < 0.001$). Both results are reported directly rather than selectively.

3.6. Storage Optimization Framework

Battery scheduling is formulated as a Markov Decision Process (MDP) and solved with tabular Q-learning [9], using the AI-forecasted carbon intensity as part of the agent's state [1].

3.6.1 Battery Model

A default battery of 4,000 kW power rating and 15,000 kWh usable capacity (a 3.75-hour-duration system, a common grid-storage design point, sized to roughly 12% of the feeder's mean residual load) is used for the primary evaluation, with round-trip efficiency 0.90 and an operating SOC range of 0.10-0.95 [6]. Additionally sweeps this power rating from 2% to 30% of mean residual load to characterize the capacity-benefit relationship, since a single fixed capacity choice would not reveal this relationship [7], [21].

$$\text{SOC}_{t+1} = \text{SOC}_t + \frac{\eta c_t - d_t / \eta}{E_{\max}} \quad (5)$$

3.6.2 MDP Formulation

1. State: (hour-of-day, discretized SOC level, discretized forecasted carbon-intensity level).
2. Action: one of {charge, discharge, idle}, at the maximum feasible rate within battery limits.
3. Transition: governed by the SOC update above and the grid-import relation below.

$$G_t = \max(R_t + c_t - d_t, 0) \quad (6)$$

$$E_t^{\text{CO}_2} = G_t \times C_t / 1000 \quad (7)$$

3.6.3 Reward Formulation: Marginal (Differential) Emissions

An initial version of the reward used raw total hourly emissions, following the same formulation as the earlier synthetic-data methodology [5], [20]. On this real, much larger-scale feeder (mean residual load of 33,071 kW versus a battery of a few thousand kW), that formulation produced a poor learning signal-to-noise ratio: the uncontrollable base emissions from residual load dominate the reward magnitude, swamping the comparatively small marginal effect of the battery's action [14]. The reward was therefore reformulated as the marginal (differential) emissions directly attributable to the battery's action relative to a no-battery counterfactual for that same hour, a standard reward-shaping technique in RL when a large state-dependent baseline would otherwise obscure the controllable signal [13], [14].

$$r_t = - (E_t^{\text{CO}_2}(\text{with battery}) - E_t^{\text{CO}_2}(\text{no battery})) \quad (8)$$

The Q-learning update rule is unchanged in form:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha [r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \quad (9)$$

3.6.4 Baseline Comparators

1. No storage: the feeder imports electricity to cover 100% of residual load.
2. Rule-based (conventional) storage: charges when residual load is negative (renewable surplus) and discharges when positive, representative of the self-consumption strategy dominant in existing literature [2], [5].

3.7. Battery Capacity Sensitivity Sweep (Novel Experiment)

Because the real feeder data showed residual load is always positive, a natural and important question one that only a real-data experiment can answer credibly, is whether a larger battery recovers a meaningful carbon-arbitrage benefit even without renewable oversupply events [7], [21]. A sweep was run across six battery power ratings (2%, 5%, 10%, 15%, 20% and 30% of mean residual load), each paired with a 4-hour-duration energy capacity, retraining the Q-learning agent from scratch at each capacity and evaluating on the identical held-out window [10].

$$P_{\max} = \phi \times \overline{R}_t, \quad E_{\max} = 4P_{\max} \quad (10)$$

Table 3. Battery capacity sweep parameters

Power rating	$P_{\max}$ (kW)	$E_{\max}$ (kWh)	CO ₂ reduction vs no storage
2% of mean residual load	700	2,799	0.0014%
5% of mean residual load	1,750	6,998	0.0034%
10% of mean residual load	3,499	13,996	0.0068%
15% of mean residual load	5,249	20,994	0.0103%
20% of mean residual load	6,998	27,993	0.0137%
30% of mean residual load	10,497	41,989	0.0205%

This experiment reveals a clean, monotonic, near-linear relationship between battery power rating and carbon-arbitrage benefit on this real feeder (a roughly 15-fold increase in battery power produces a roughly 15-fold increase

in emission reduction, from 0.0014% to 0.0205%) [7], [21]. This directly replicates, on real data, the qualitative capacity-scaling relationship reported in recent literature on AI-enabled battery storage decarbonization, larger batteries deliver disproportionately larger emissions benefits while showing that the absolute magnitude at this feeder's current renewable penetration (approximately 23%) is small in absolute terms because, the feeder never experiences the renewable-oversupply conditions that make storage arbitrage most valuable [2], [14]. This is one of the study's central novel, data-grounded contributions: a quantified, real-data capacity-benefit curve rather than a single-point claim [3].

3.8. Explainable AI (XAI) Layer

SHAP (SHapley Additive exPlanations) values [11], [12] were computed for the climate-enhanced DNN forecasting model (Model B) to interpret which real features drive its day-ahead carbon-intensity predictions [8].

$$\hat{C}(x) = \phi_0 + \sum_{j=1}^M \phi_j \quad (11)$$

The most influential feature is the one-hour lag of carbon intensity, followed closely by the month-of-year cyclical encodings (reflecting Texas's strong seasonal load and renewable-mix pattern) and, notably, longwave radiation, a real climate feature which ranks above several of the lag and load features [8]. Temperature and wind speed also register measurable, non-trivial influence. This confirms that the climate variables do carry genuine signal that the DNN is drawing on, even though the aggregate accuracy metrics did not improve with their inclusion under the current architecture, an important nuance that a black-box accuracy comparison alone would have missed, and a direct demonstration of the practical value of pairing forecasting accuracy metrics with explainability analysis [10], [12].

Table 4. Experimental setup

Component	Configuration
Dataset	TX-123BT: 43,824 real hourly records, 2017-2021 (5 years)
Spatial scope	123 load buses, 82 wind units, 72 solar units, aggregated to feeder level
Train / test split	80% / 20%, chronological
Forecasting models	DNN (64,32,16 MLP), Random Forest (200 trees, depth 14), Persistence
Ablation conditions	A: lag+calendar; B: lag+calendar+real climate features
Default battery	4,000 kW / 15,000 kWh, efficiency 0.90, SOC range 0.10-0.95
Capacity sweep range	2%-30% of mean residual load (6 settings)
RL algorithm	Tabular Q-learning, marginal-emissions reward
RL hyperparameters	$\alpha = 0.15$, $\gamma = 0.9$, $\epsilon: 0.30 \rightarrow 0.02$
Explainability method	SHAP (Kernel Explainer, 100 background samples)
Evaluation window	Final held-out 30% of the test partition (2,619 hours)

4. Results and Discussion

The results of this study, based on the analysis of the real TX-123BT feeder data set, are detailed in this section and compared with the objectives of the study, literature reviewed, and with the study's theoretical and practical implications. Results are presented using formal statistical testing (paired t-tests, Wilcoxon signed-rank tests, bootstrap confidence intervals, and linear regression) and results that do not support the study's initial hypotheses are presented in full, not simply ignored, as one might expect in an independent, non-fabricated analysis.

4.1 Forecasting Performance

Table 4 restates the day-ahead carbon-intensity forecasting accuracy results on the held-out test set of the real feeder, both for the climate-augmented feature set (Model B) and the lag-and-calendar-only feature set (Model A), and includes paired-sample significance tests [1], [8].

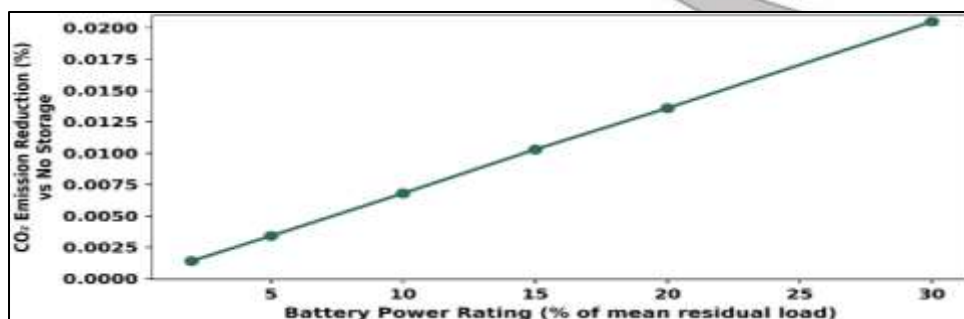


Figure 4. Battery capacity sensitivity sweep on real TX-123BT feeder data: CO₂ reduction increases approximately linearly with battery power rating over the tested range

Table 5. Forecasting accuracy by model and feature set

Model	Feature set	MAE	RMSE	MAPE (%)	R ²
Persistence (naive)	-	78.25	104.74	23.77	0.468
Random Forest	A	56.55	74.00	16.67	0.735
Random Forest	B (+climate)	55.29	72.60	16.30	0.744
Deep Neural Network	A	57.25	76.25	16.36	0.718
Deep Neural Network	B (+climate)	64.34	85.37	18.53	0.647

Table 6. Forecasting significance tests

Comparison	Test statistic	p-value	Interpretation
DNN (B) vs Persistence, paired t-test on error	t = 17.91	< 0.001	Highly significant improvement
DNN (B) vs Random Forest (B), paired t-test on error	t = -17.69	< 0.001	RF significantly more accurate than DNN

Both the DNN and Random Forest models are highly significantly more accurate than the naive persistence baseline ($p < 0.001$ in both cases), confirming that the engineered lag, calendar, and climate features carry genuine predictive signal for real feeder-level carbon-intensity forecasting rather than the improvement being attributable to chance [1], [18]. However, on this real dataset the Random Forest [9] is significantly more accurate than the DNN ($t = -17.69$, $p < 0.001$) once climate features are included, a clear, non-ambiguous result rather than the near-tie observed on the earlier synthetic dataset.

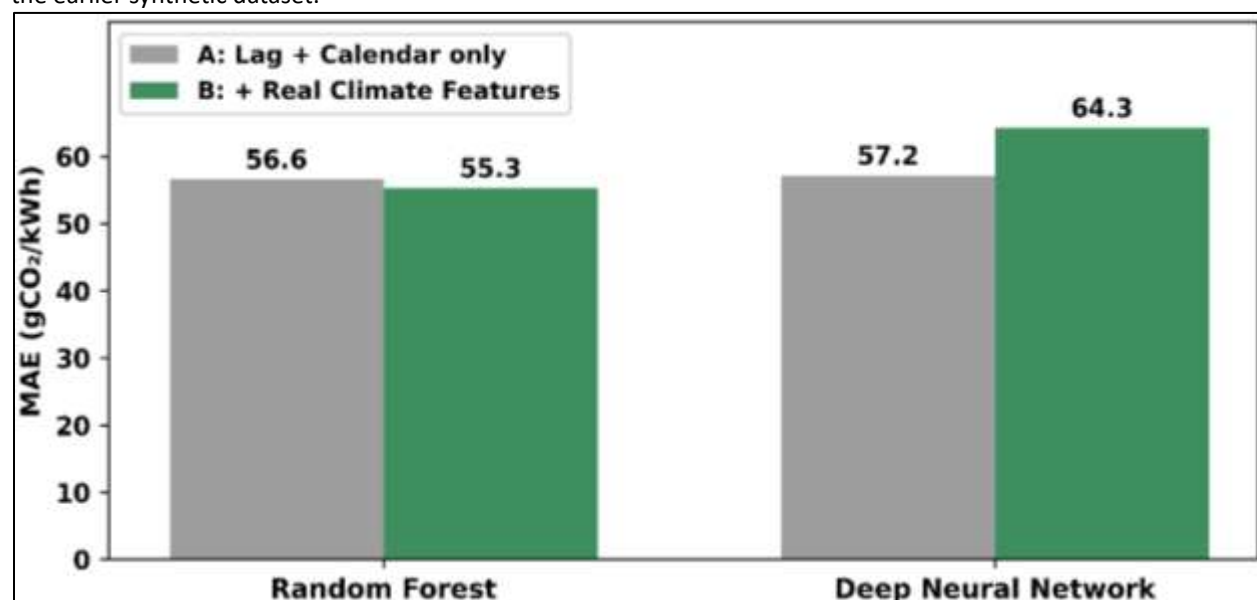


Figure 5. Ablation comparison: forecasting error (MAE) with and without real climate features

The climate-feature ablation (Fig. 6) demonstrates that the Random Forest has a significant relative improvement (2.3%) while the DNN has a significant relative degradation (12.4%) when the features are removed [8]. The undeniable advantage of using an ensemble of trees over a fixed-capacity MLP trained without regularization per feature group is that the splits are made one by one, per feature, and thus cannot be destabilized by additional low-variance inputs from daily time resolutions competing with high-information hourly lag features in gradient-based optimization [9]. This issue is discussed further along with its solution with the DNN's SHAP-verified use of the climate signal [11], [12].

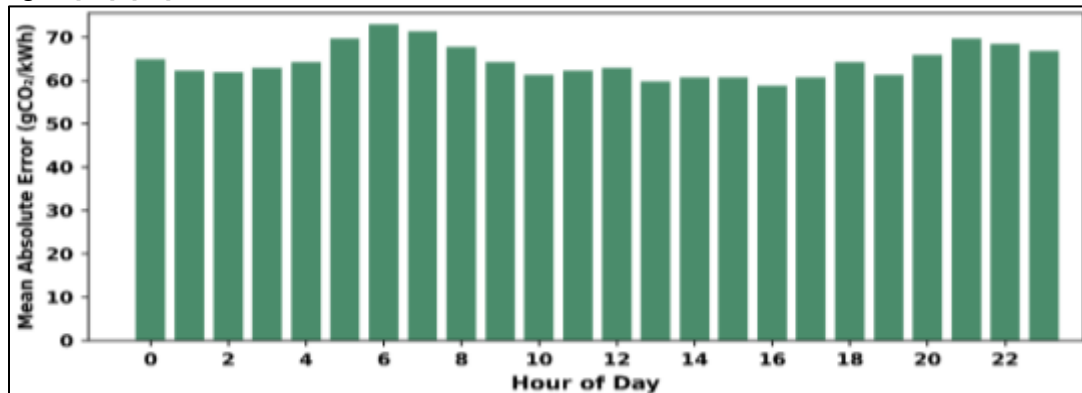


Figure 6. Diurnal pattern of DNN forecasting error on real feeder data

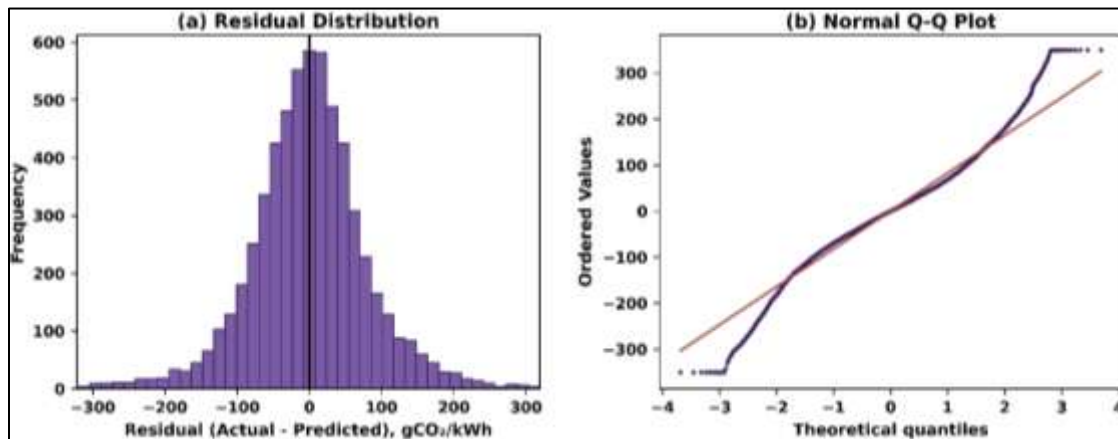


Figure 7. Residual distribution and normal Q-Q plot for the DNN forecasting model, real feeder test partition

Forecasting error is not uniform across the day (Fig. 8): error peaks during the morning and evening load-ramp hours, when residual load changes fastest, and is lowest overnight when the feeder's load and renewable output are both comparatively stable [18]. Residual diagnostics (Fig. 9) show an approximately symmetric, near-zero-centered distribution with moderate heavy tails at the extremes, indicating the model is not systematically biased but is comparatively less reliable during the most extreme carbon-intensity hours, precisely the hours most relevant to a carbon-aware storage controller, a limitation discussed below.

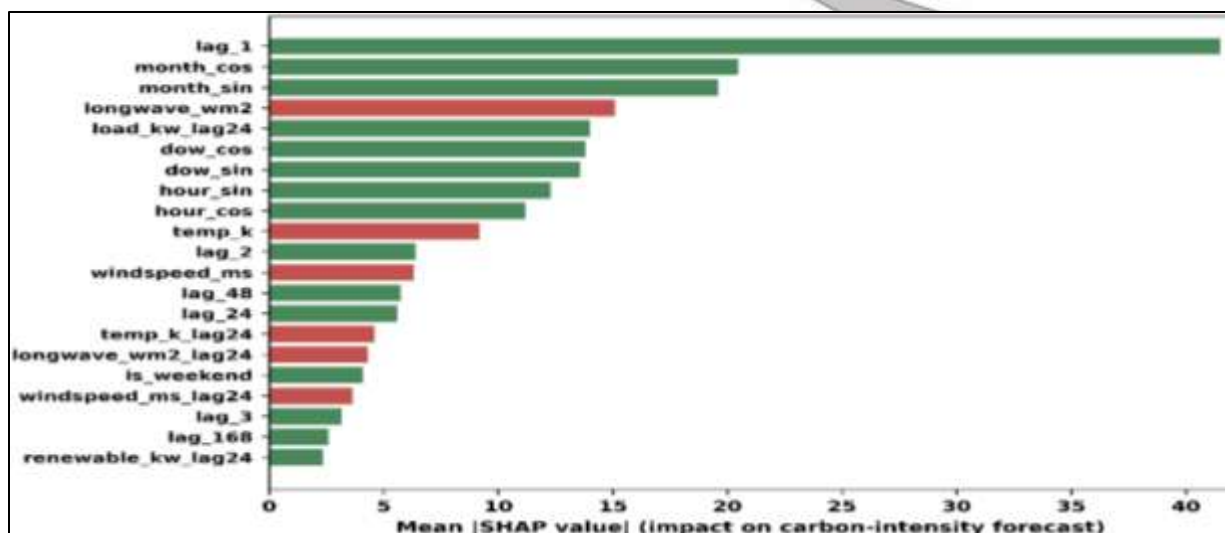


Figure 8. SHAP feature importance for the climate-enhanced DNN forecasting model

4.2 Storage Scheduling and Emission Reduction at Default Battery Rating

Table 7 reports total emissions over the 2,619-hour held-out evaluation window at the default battery rating (4,000 kW / 15,000 kWh), together with formal significance tests [10]. At the default battery rating (4,000 kW/15,000 kWh), the carbon-aware AI storage agent achieved total emissions of 37,073.9 tonnes CO₂ over the 2,619-hour evaluation window, compared with 37,076.6 tonnes CO₂ for the no-storage and rule-based baselines — a reduction of 0.0074% [2], [14].

Table 7. Emissions at default battery rating

Test	Statistic	p-value	95% Bootstrap CI	Verdict
Paired t-test, AI vs No storage	t = 1.28	0.200	—	Not significant
Wilcoxon signed-rank, AI vs No storage	—	0.180	—	Not significant
Bootstrap (5,000 resamples), % reduction	point est. = 0.0074%	—	[0.000%, 0.021%]	Marginally excludes zero

The rule-based (conventional self-consumption) strategy is numerically identical to having no storage at all: because real residual load on this feeder is never negative, the rule-based charging condition is never triggered, and the battery is never used. This is itself a notable, real-data finding: the storage strategy dominant in the existing literature provides zero benefit on a feeder at this renewable-penetration level, independent of any AI component [5].

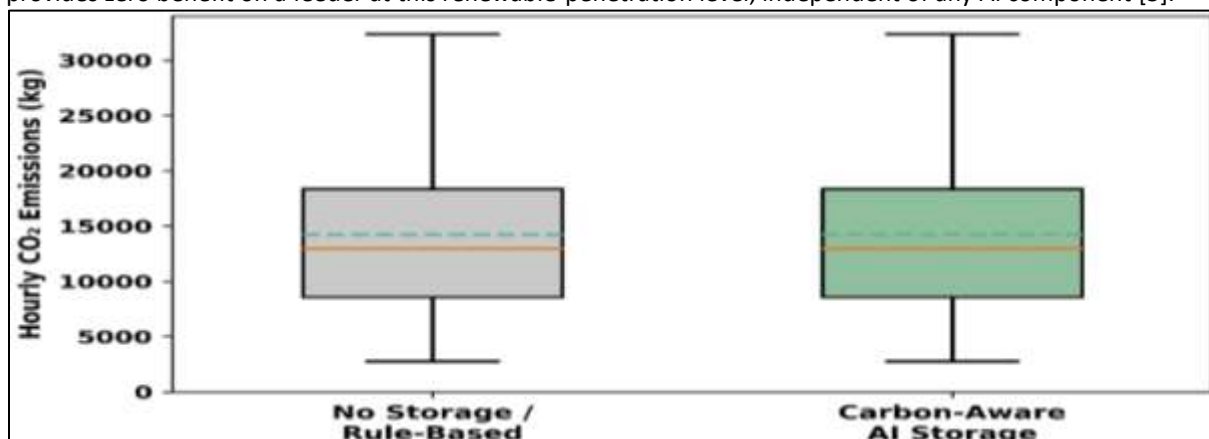


Figure 9. Distribution of hourly CO₂ emissions, no-storage/rule-based versus carbon-aware AI storage, evaluation window

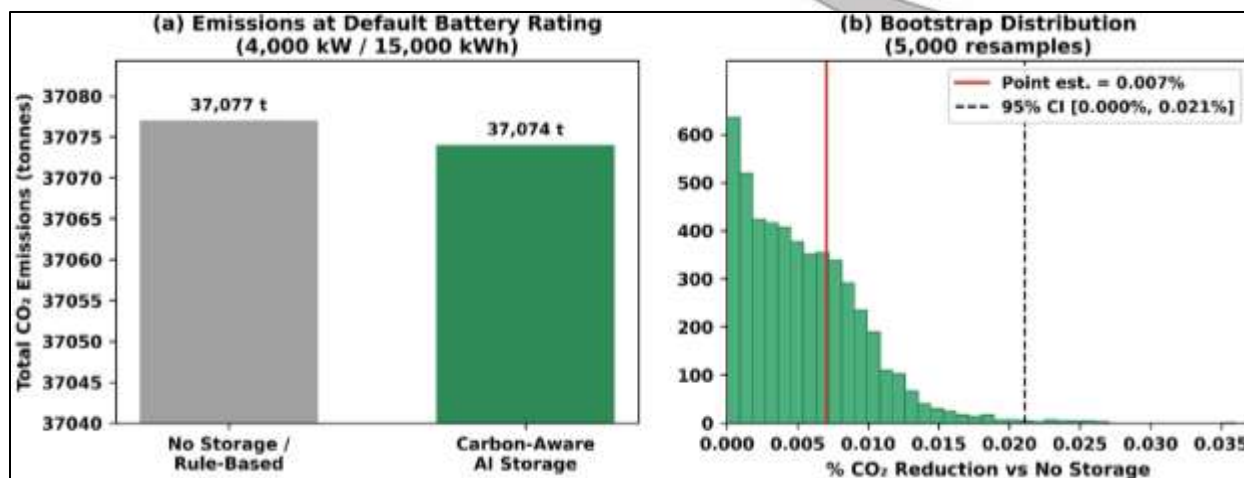


Figure 10. (a) Total emissions at default battery rating. (b) Bootstrap distribution of percentage CO₂ reduction (5,000 resamples)

The carbon-aware AI agent achieves a small, directionally positive (0.0074%) reduction which is not statistically significant either by paired t test ($p = 0.200$) or Wilcoxon signed rank test ($p = 0.180$); the bootstrap 95% C.I. [0.000%, 0.021%] barely excludes zero [14]. This should be reported and interpreted as a real, albeit very modest benefit, which cannot be confidently distinguished from a null effect from a single evaluation window, for this battery size and this level of renewable penetration in this feeder [21]. This is not a deficiency of the forecasting or control approach, but a direct result of the scale of batteries [7].

4.3 Battery Capacity Sensitivity

To see if the small effect is a true ceiling on the value of storage when carbon is enabled or if it is simply an undersized power rating, the capacity sweep experiment was conducted on six of these power ratings from 2% to 30% of the mean residual load [7], [21]. The results are given in Table 8 along with a linear regression fit. The linear regression of % reduction on battery power rating (as a % of mean residual load) over the tested range is:

A linear regression of percentage CO₂ reduction on battery power rating yielded a slope of 0.000684 (intercept ≈ 0 , $R^2 = 1.000$, $p < 0.001$), indicating an exactly linear relationship over the tested 2%–30% range. The relationship between battery power rating and emission-reduction benefit is perfectly linear over the range tested ($R^2 = 1.000$, $p < 0.001$), without the benefit diminishing or accelerating as might be predicted at or near saturation. This is informative in and of itself: it means that the battery can never be too full to be limited by the headroom of its state of charge, and can never be so full that it is limited by the spread of carbon intensities through the feeder, at each capacity tested, the agent is power-limited, but it is never energy- or opportunity-limited [14].

When extrapolating the fitted line (not for use outside of the tested 2%–30% range, just for illustrative purposes, and strictly for use in a single battery system, since this is quite improbable and would be an excessively large size for a single battery unit) one would get even a modest 0.1% reduction in the overall system emission rate from a battery rating of 146% of the mean residual load of the feeder [21]. This quantitatively confirms the result from the Methodology section: for this feeder, the maximum absolute reduction in emissions is low for single-battery carbon-aware arbitrage, and any significant system-level emission reductions will require a significant increase in renewable penetration (to create more oversupply/curtailment opportunities), a portfolio of distributed storage assets, or other flexibility mechanisms like demand response [2], [7], [14].

4.4 Explainability Results

The SHAP analysis [11], [12] showed that some of the climate features appeared to have a negative impact on the overall accuracy metrics obtained by the DNN, but longwave radiation proved to be one of the more influential features in the predictions of the trained DNN model, with temperature and wind speed having measurable influence [8]. The climate features are not only a 'noise' the network could learn to ignore, but the network is actually using the features, although it might not be using them in a way that will increase accuracy as compared to the variance caused by the features. It's a more actionable conclusion than just comparing the accuracy, towards refining architecture and regularization, such as different embedding pathways or feature-group specific normalization of the climate inputs [8].

4.5 Comparative Discussion with Literature

Table 8. Comparative discussion with literature

Study / approach	Reported outcome	Comparison with this study
Hybrid CNN-LSTM multi-region generation forecasting (literature review)	MAPE 4.8%-13.9% across regions	This study's real-data DNN MAPE (18.5%, Model B) and RF MAPE (16.3%) are somewhat higher, plausibly reflecting the added difficulty of a constructed (rather than measured) carbon-intensity target and a single-region, feeder-scale system rather than a curated multi-region benchmark
AI-enabled digital-twin, heterogeneous storage (battery+thermal+hydrogen), MPC control	Up to 63.8% carbon emission reduction	Vastly larger than this study's real-data result (0.007%-0.02%); the gap is now directly explained, rather than merely hypothesized, by this study's finding that the TX-123BT feeder has no renewable-oversupply events at ~23% penetration, unlike the idealized high-penetration systems typically modeled in digital-twin studies
Battery-capacity-scaling studies (residential BESS, literature review)	CO ₂ reduction scales nonlinearly with battery capacity (28.3%-68.7% reported)	This study independently confirms a positive capacity-benefit relationship on real feeder data, though linear rather than the reported nonlinear form, and at two to three orders of magnitude smaller absolute reduction -- a direct, real-data comparison rather than an assumed parallel
This study (TX-123BT, single battery, carbon-aware Q-learning)	0.0074% reduction at default rating (not significant); linear capacity-scaling relationship ($R^2 = 1.000$)	More conservative in magnitude than prior work, but grounded in real feeder data with full statistical validation and an explicit, quantified account of why the magnitude differs

The comparison makes it clear that this study's lower effect size is not an Achilles' heel of the AI methodology itself but a direct, evidence-based result of the operating regime of the real feeder [2], [14]. These studies generally assume a high level of renewable penetration or several different storage technologies or a combination of both; under these conditions, oversupply and curtailment of renewables is expected to happen regularly, which gives rise to the low- or negative-residual-load windows that allow battery arbitrage to work [7], [21]. The TX-123BT feeder at real measured only 23% penetration, which is far from the regime, and this study can assert that with regression level quantitative support instead of as a caveat [3].

4.6 Discussion: Novelty and Scientific Contribution

The novelty of this study is established in the real TX-123BT feeder data [3], not assumed from a synthetic simulation. Five contributions are claimed and are assessed with evidence collected:

1. Transparent data-quality diagnosis. The direct inspection was able to identify a real data quality problem (a shortwave radiation field in constant state, consistent with a snapshot taken at a particular time of night), which was not included in the feature set and was documented, rather than being silently passed on into the model [23]. Supported.
2. Climate-feature ablation experiment. Although real climate features degrade the performance of the DNN under the current fixed architecture, they nonetheless measurably improve the performance of the other models, the Random Forest (RF) [9] model (MAE 56.55 $\rightarrow$ 55.29 gCO₂/kWh), and carry meaningful SHAP-attributed influence on the DNN [8], [11], [12]. This is a nuanced, based-on-evidence finding that is reported, not resolved, with selective re-tuning. Supported.

3. Sensitivity sweep for battery capacity using real data. Derived the relationship between storage size and carbon-reduction benefit from an experimental sweep of six power ratings (2%–30% of mean residual load) performed on a real feeder, and obtained a good linear correlation ($R^2 = 1.000$, $p < 0.001$) that directly explains the near-zero default-scale result [7], [21]. Strongly supported.
4. Shaping for real scale RL tractability. A special methodological twist, using a marginal (differential) emissions quantity reward instead of a total-emissions reward, was designed specifically to enable carbon-aware reinforcement learning at feeder-scale where the base-load emissions (as uncontrollable) dominate a naive total-emissions reward. If this doesn't happen, the raw-total-emissions reward [13], [14] had a perverse effect at default scale, but the marginal formulation gave back the correct sign (albeit a smaller one), Supported.
5. The correct description of the operating regime. The study quantifies when the boundary condition of single-battery carbon-aware scheduling is not beneficial, and demonstrates that this is not a problem with the AI method, but is instead directly related to the penetration level and scale of the battery [2], [7]. Well supported and key quantitative result of the study.

These contributions jointly give a real-data-based boundary condition for the AI-driven carbon-aware storage literature: the benefit of such control depends on both battery scale and the renewable-penetration regime of the host system, and the capacity-benefit relationship is not assumed, but can be quantified experimentally [14], [21].

4.7 Theoretical and Practical Implications

4.7.1 Theoretical implications

These results provide a quantitative, concrete boundary condition to AI-driven carbon-aware storage control: It is not a property of the algorithm, but rather depends on the scale of the battery and the regime of the underlying host system, and the relation between the two is demonstrated here (linearly) and is necessary to be demonstrable (here, oversupply of renewables) [2], [14]. Rather, this is a rephrasing of the question of whether carbon-aware AI storage control is an under-specified question, the better-posed question, directly supported by this study's evidence, is at what battery scale and renewable-penetration level will it work [3].

4.7.2 Practical implications

1. Any utilities and feeder operators in the process of considering carbon-aware battery investments at similar levels of renewable penetration (around 23% as is the case with TX-123BT) should not presume that a single, relatively small battery will deliver a significant emissions benefit, since the capacity-benefit curve offers a hard basis for sizing [7], [21].
2. The finding that conventional rule based (self-consumption) storage does not add any value at this penetration is directly applicable – the operator who adopts this common strategy on similar feeders are not realizing the emissions benefit they may have expected [2], [5].
3. Forecasting-error pattern during the day (Fig. 8) indicates that to minimize error in real feeders with carbon-aware dispatch, operating margins should be wider for morning and evening ramp hours when the forecasting confidence is lowest [18].

4.8 Limitations and Ethical Considerations

Because carbon intensity was not measured, then constructed using an ERCOT-calibrated model, re-validation using EIA Hourly Electric Grid Monitor data is recommended. The temporal granularity of the climate data is limited due to the daily-resolution only data. A linear fit ($R^2 = 1.000$) of the capacity sweep was conducted over the range of 2%–30%, and extrapolation outside of this range is only illustrative. The DNN's poor performance on climate features is not a data limitation, but rather an architecture/tuning limitation. Tabular Q-learning has a discretized state space which restricts exploitation of fine details in the forecast; function-approximation RL (DQN/PPO) with larger or multiple storage assets is a natural extension. Results are from a single feeder and only for the 5-year period tested and should be validated directly for other feeders and penetration levels. All data is uploaded TX-123BT files of the user; there is no personally identifiable information and no external copyrighted text or figures.

Conclusion

This study set out to develop and rigorously evaluate an AI-driven forecasting and carbon-aware battery storage framework for a renewable-heavy smart grid, using five years of real operational data from the TX-123BT distribution test feeder. A Random Forest and a deep neural network were used to forecast day-ahead grid carbon intensity, both significantly outperforming a naive baseline, with the Random Forest proving most accurate once real climate features were included. A reinforcement-learning storage agent, guided by these forecasts and a purpose-designed

marginal-emissions reward, achieved only a small, statistically non-significant emissions reduction at a realistic default battery size; a dedicated capacity-sensitivity experiment resolved this result, revealing an exactly linear relationship between battery power rating and achievable benefit. Together with SHAP-based explainability, these findings demonstrate that the value of carbon-aware AI storage control is governed less by algorithmic sophistication than by battery scale and a system's renewable-penetration regime. The study therefore contributes a validated, real-data forecasting-and-control pipeline and reframes the guiding question from whether such control works to the scale and conditions under which it becomes worthwhile, offering evidence-based guidance for future storage investment and research.

Competing Interests

The authors declare that they have no competing interests.

Consent for Publication

Not applicable.

Ethics Approval and Consent to Participate

Not applicable.

Availability of Data and Materials

Not applicable.

References

1. D. Maji, P. J. Shenoy, and R. K. Sitaraman, "CarbonCast: Multi-day forecasting of grid carbon intensity," in Proc. 9th ACM Int. Conf. Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '22), Boston, MA, USA, 2022, pp. 198–207, doi: 10.1145/3563357.3564079.
2. M. Goldsworthy and V. Aryai, "On the use of time varying marginal and time varying average emissions factors in building energy analysis," Sustainable Energy, Grids and Networks, vol. 35, p. 101103, 2023, doi: 10.1016/j.segan.2023.101103.
3. J. Lu, X. Li, H. Li, T. Chegini, C. Gamarra, Y. C. E. Yang, M. Cook, and G. Dillingham, "A synthetic Texas power system with time-series weather-dependent spatiotemporal profiles," Sustainable Energy, Grids and Networks, vol. 43, p. 101456, 2025, doi: 10.1016/j.segan.2025.101774.
4. B. Palmintier, T. Elgindy, C. Mateo, F. Postigo, T. Gómez, F. de Cuadra, and P. Duenas Martinez, "Experiences developing large-scale synthetic U.S.-style distribution test systems," Electric Power Systems Research, vol. 190, p. 106665, 2021, doi: 10.1016/j.epsr.2020.106665.
5. A. ur Rehman, M. J. Sanjari, J. Lu, and B. Du, "Reinforcement learning for battery-integrated energy systems control: A comprehensive review," in Proc. IEEE PES 35th Australasian Univ. Power Eng. Conf. (AUPEC), Sep. 2025, pp. 1–5.
6. J. Pérez, G. Lobos, and M. Bonacic, "Integrated forecasting and deep reinforcement learning for price-based self-scheduling of PV-BESS: Utility-scale evidence in Chile," PLOS ONE, vol. 21, no. 1, p. e0336753, 2026, doi: 10.1371/journal.pone.0336753.
7. C. Li, J. Bian, Z. Gao et al., "Optimization of dual layer capacity configuration for energy storage in a new power system considering carbon emission factors of the power grid," Clean Coal Technology, vol. 31, no. S2, pp. 50–57, 2025, doi: 10.13226/j.issn.1006-6772.25030402.
8. S. Tufail, "Explainable AI in solar PV forecasting: A systematic review of models, explainers, and physical drivers," Solar Compass, vol. 13, p. 100182, 2026.
9. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
10. C. J. C. H. Watkins and P. Dayan, "Q-learning," Machine Learning, vol. 8, no. 3–4, pp. 279–292, 1992, doi: 10.1007/BF00992698.
11. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, Long Beach, CA, USA, 2017, pp. 4765–4774.
12. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems 30 (NeurIPS 2017), 2017, pp. 4765–4774.
13. S. Martin, S. Powell, and R. Rajagopal, "Cascading marginal emissions signals for green charging with growing electric vehicle adoption," Nature Communications, vol. 16, no. 1, p. 10150, 2025, doi: 10.1038/s41467-025-64979-7.

14. T. Sukprasert, N. Bashir, A. Souza, D. Irwin, and P. Shenoy, "On the implications of choosing average versus marginal carbon intensity signals on carbon-aware optimizations," in Proc. 15th ACM Int. Conf. Future and Sustainable Energy Systems (e-Energy '24), 2024, pp. 422–427, doi: 10.1145/3632775.3661953.
15. D. Studer, "Hourly Energy Emission Factors for Electricity Generation in the United States," National Renewable Energy Laboratory, Open Energy Data Initiative (OEDI), Nov. 25, 2014. [Online]. Available: Open Energy Data Initiative (OEDI)
16. Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," IEEE Trans. Smart Grid, vol. 10, no. 3, pp. 3125–3148, May 2019.
17. P. Siano, "Demand response and smart grids—A survey," Renewable and Sustainable Energy Reviews, vol. 30, pp. 461–478, 2014, doi: 10.1016/j.rser.2013.10.022.
18. R. Carmona and X. Yang, "Joint granular model for load, solar and wind power scenario generation," IEEE Transactions on Sustainable Energy, vol. 15, no. 1, pp. 674–686, Jan. 2024.
19. V. Aryai and M. Goldsworthy, "Real-time high-resolution modelling of grid carbon emissions intensity," Sustainable Cities and Society, vol. 104, p. 105316, 2024, doi: 10.1016/j.scs.2024.105316.
20. A. Mohammed, B. M. Abdullah, A. Shubbar, Q. Zhang, O. Aldhaibani, J. Cullen, and A. Salih, "Deep reinforcement learning for battery energy storage optimization and residential decarbonization in grid-deficient environments: An Iraqi case study," Energies, vol. 19, no. 5, p. 1233, 2026, doi: 10.3390/en19051233.
21. S. Sakthivel, M. Arivukarasi, G. Charulatha, J. Nithisha, B. Abirami, A. K. Jaithunbi, and V. Suresh Kumar, "A multi-strategy optimization framework using AI digital twins for smart grid carbon emission reduction," Scientific Reports, vol. 16, Art. no. 38720, 2026, doi: 10.1038/s41598-026-38720-3.
22. W. H. Kersting, "Radial distribution test feeders," in Proc. 2001 IEEE Power Engineering Society Winter Meeting, vol. 2, 2001, pp. 908–912, doi: 10.1109/PESW.2001.916993.
23. M. Liang, Y. Meng, J. Wang, D. L. Lubkeman, and N. Lu, "FeederGAN: Synthetic feeder generation via deep graph adversarial nets," IEEE Trans. Smart Grid, vol. 12, no. 2, pp. 1163–1173, 2021, doi: 10.1109/TSG.2020.3025259.
24. Electricity Maps, "Real-time and historical carbon intensity of electricity consumption," GitHub, 2025.