Decision-Focused Learning for Carbon Intensity Forecasting in Power Grids

Anonymous Author(s)
Affiliation
Address
email

Abstract

Electric power systems contribute a substantial share of global greenhouse gas emissions, making accurate short-term forecasting of carbon intensity (CI) critical for low-carbon grid operations. Existing work overwhelmingly treats CI forecasting as a standalone prediction problem and evaluates models solely using statistical error metrics, such as MAE or MAPE. However, small differences in forecast accuracy may lead to disproportionately large differences in operational decisions once predictions are fed into scheduling or control pipelines.

This paper presents a decision-focused framework for carbon intensity forecasting across large-scale power systems. We develop a Temporal Fusion Transformer (TFT) trained on multi-year datasets from three major North American markets (ERCOT, NYISO, PJM), combining weather forecasts, fuel-mix forecasts, and historical system conditions. To evaluate forecasting quality in an operational context, we introduce a lightweight carbon-aware load allocation simulator that measures both decision regret and ratio-to-oracle performance.

Our results show that the TFT model achieves strong predictive accuracy (WAPE 1.5%–4.7% across regions) and, more importantly, yields substantial improvements in downstream decision quality. Forecast-based scheduling reduces regret by one to two orders of magnitude relative to a naive uniform allocation baseline, while maintaining decision-quality ratios close to 1.00–1.05. These findings highlight that moderate improvements in forecast accuracy can translate into large operational benefits, underscoring the importance of evaluating CI forecasting models through a decision-focused lens.

1 Introduction

Electric power systems are among the largest contributors to global greenhouse gas emissions. Short-term forecasts of carbon intensity (CI), typically measured in $\mathrm{gCO_2/kWh}$, are increasingly used to support carbon-aware planning, flexible demand scheduling, and real-time operational decision making. Reliable CI forecasts allow data centers, industrial facilities, and automated demand response systems to shift consumption toward cleaner periods, thereby reducing emissions without requiring new physical assets.

Most existing work frames CI forecasting purely as a statistical prediction problem. Forecasting models are almost exclusively evaluated using error-based metrics—MAE, MSE, WAPE, MAPE—implicitly assuming that lower statistical error translates directly into better operational outcomes. However, this assumption rarely holds in practice: two models with nearly identical MAE can lead to significantly different decisions once their predictions are used inside an optimization pipeline. This disconnect motivates the emerging perspective of *decision-focused learning*, which

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

evaluates predictive models based on their downstream decision impact rather than error metrics alone.

In this paper, we take a step toward bridging these perspectives in the context of real-world carbon intensity forecasting. Using multi-year datasets from three major North American electricity markets (ERCOT, NYISO, PJM), we train a Temporal Fusion Transformer (TFT) to produce multi-horizon CI forecasts based on high-resolution weather forecasts, fuel-mix production forecasts, and historical systems data. We then embed these forecasts into a carbon-aware load allocation simulator that optimizes the dispatch of a flexible energy budget over a 12-hour lookahead window. This setup enables us to quantify how forecast errors propagate into operational decisions.

Our key findings are:

- **Strong forecasting performance.** The TFT achieves low error across all regions (WAPE 1.5%–4.7%), capturing both diurnal structure and weather–fuel interactions.
- **Significant improvements in operational decision quality.** When used for carbon-aware scheduling, forecast-based decisions reduce average regret by one to two orders of magnitude relative to a naive uniform allocation baseline.
- **Tight ratio-to-oracle performance.** The TFT achieves ratios near 1.00–1.05, demonstrating that small improvements in predictive accuracy can yield disproportionately large operational benefits.

Together, these results demonstrate that carbon intensity forecasting should be evaluated not only as a standalone prediction task but as an integral component of a broader forecast-and-decide pipeline. Our study highlights the importance of decision-oriented evaluation and provides a practical benchmark for assessing forecasting models in carbon-aware power system operation.

2 Related Work

Carbon intensity forecasting. Forecasting carbon intensity has gained increasing attention as grid operations become more carbon-aware. Prior work spans statistical models, machine learning approaches, and hybrid physical–data-driven designs. Most studies treat CI prediction as an independent regression task, evaluating models primarily using error-based metrics such as MAE or MAPE [Zhang and Li, 2020, Schmidthuber and Keles, 2022]. While effective for measuring pure predictive accuracy, these metrics do not capture how forecasts influence downstream decisions, which is central to operational carbon management. Our work complements these efforts by emphasizing the decision impact of forecasting errors.

Decision-focused learning. Decision-focused learning (DFL) links prediction models with the optimization problems that consume their outputs. Classical work shows that minimizing prediction error may be misaligned with decision optimality [Donti et al., 2017]. Recent advances develop differentiable optimization surrogates and ranking-based objectives to improve downstream decisions directly [Wilder et al., 2019, Mandi et al., 2022]. Our approach aligns with the DFL philosophy but adopts a modular structure: we first train a strong forecaster, then evaluate its decision impact through a carbon-aware scheduler. This design retains generality across markets and allows controlled measurement of forecast error propagation.

Carbon-aware scheduling and load shifting. A growing line of work studies how flexible loads can be shifted to reduce emissions [Kim and Lee, 2021, Xu et al., 2023]. Most studies assume oracle—or externally provided—carbon signals. In contrast, our work explicitly couples the forecasting and decision layers using realistic multi-horizon CI predictions.

Hyperparameter optimization and AutoML. Tools such as Auto-WEKA [Thornton et al., 2013] and Optuna [Akiba et al., 2019] automate model tuning to improve predictive accuracy. Our implementation uses manual tuning but can incorporate AutoML strategies as future extensions.

3 Problem Setup

We consider a single power system region (e.g., ERCOT, NYISO, PJM) observed at discrete hourly time steps. Let $t \in \{1, 2, \dots\}$ index time, and let $y_t \in \mathbb{R}$ denote the realized system-level carbon intensity at time t (e.g., in $\mathrm{gCO_2/kWh}$).

For each time t, we observe a feature vector $x_t \in \mathbb{R}^d$ constructed from heterogeneous data sources, including:

- **Multi-horizon weather forecasts** (e.g., wind speed, temperature, downwelling shortwave radiation, precipitation), aligned to the hourly grid;
- **Fuel-mix production forecasts** by generator type (e.g., coal, gas, nuclear, hydro, wind, solar, other);
- **Calendar and time-of-day features** (e.g., hour-of-day, day-of-week, holiday indicators).

These inputs capture both exogenous drivers (weather) and endogenous system conditions (fuel mix) that jointly shape real-time carbon intensity.

Given a history window of length L and a prediction horizon of length H (e.g., $L = 96$ hours, $H = 12$ hours), the forecasting problem is to map a sequence of past observations and known future covariates to a sequence of multi-quantile forecasts. Concretely, at decision time t we seek to predict

$$\hat{y}_{t+1:t+H} \in \mathbb{R}^{H \times Q},$$

where Q is the number of quantiles (e.g., $Q = 7$). The prediction for each horizon $h \in \{1, \dots, H\}$ consists of a set of quantiles $\{\hat{y}_{t+h}^{(\tau_q)}\}_{q=1}^{Q}$ for levels $0 < \tau_1 < \cdots < \tau_Q < 1$. During evaluation, we primarily use the median quantile (e.g., $\tau = 0.5$) as a point forecast.

Decision layer. On top of the forecasting model, we define a stylized carbon-aware load-shifting problem. At each decision time t, a flexible load of total size $E > 0$ must be allocated across the next H hours. Let $u_{t+h} \in \mathbb{R}_{\geq 0}$ denote the flexible energy scheduled at time $t + h$. The allocation decisions must satisfy

$$u_{t+h} \geq 0 \quad \forall h \in \{1, \dots, H\}, \qquad \sum_{h=1}^{H} u_{t+h} = E.$$

Given the realized carbon intensity y_{t+h}, the total carbon footprint of this flexible demand over the horizon is

$$C(u; y) = \sum_{h=1}^{H} u_{t+h}\, y_{t+h}.$$

An *oracle* scheduler with perfect future information chooses $u^\star$ to minimize $C(u; y)$ using the true future trajectory $\{y_{t+1}, \dots, y_{t+H}\}$. In contrast, a *forecast-based* scheduler observes only forecasts and constructs $\hat{u}$ by solving the same optimization problem with point forecasts $\hat{y}_{t+h}$ (e.g., median quantiles) in place of the unknown y_{t+h}. This setup defines a natural decision-centric evaluation protocol: by comparing $C(\hat{u}; y)$ to $C(u^\star; y)$, we can directly quantify how forecasting errors translate into additional carbon emissions.

4 Method

Our approach combines a Temporal Fusion Transformer (TFT) for multi-horizon carbon intensity forecasting with a downstream carbon-aware decision simulation. This section describes the forecasting model and the decision-evaluation pipeline.

4.1 Temporal Fusion Transformer for Carbon Intensity Forecasting

We adopt the Temporal Fusion Transformer (TFT) [Lim et al., 2021] as our base model due to its ability to jointly handle heterogeneous inputs, long-range temporal dependencies, and multi-quantile outputs.

Input structure. At each time t, the model receives:

- **Observed historical variables:** past carbon intensity, realized fuel mix, and lagged system features;
- **Known future inputs:** multi-horizon weather forecasts and fuel-mix production forecasts aligned with the prediction horizon;
- **Static covariates:** region identifiers and other time-invariant system descriptors.

The combination of known-future and historical inputs allows TFT to anticipate upcoming shifts in renewable output, demand conditions, and fuel-mix composition.

Architecture. TFT integrates three components:

1. a *variable selection network* that identifies the most relevant covariates at each time step;
2. a *sequence-to-sequence recurrent encoder-decoder* capturing local temporal structure;
3. *multi-head attention* layers that model long-range dependencies and cross-feature interactions.

To obtain multi-horizon uncertainty estimates, TFT outputs quantile predictions $\hat{y}^{(\tau)}_{t+1:t+H}$ for quantile levels $\tau \in \{\tau_1, \ldots, \tau_Q\}$. The model is trained using the standard quantile regression loss

$$\mathcal{L} = \sum_{h=1}^{H} \sum_{q=1}^{Q} \rho_{\tau_q}\big(y_{t+h} - \hat{y}^{(\tau_q)}_{t+h}\big),$$

where ρ_τ is the pinball loss. During evaluation, we use the median ($\tau = 0.5$) as the point forecast.

4.2 Decision Simulation and Metrics

To evaluate how forecasting accuracy affects operational outcomes, we embed the model predictions into a carbon-aware load-shifting problem. At each decision time t, a flexible load of fixed size E must be allocated across the next H hours to minimize emissions.

Oracle and forecast-based policies. Given the true future carbon intensity trajectory $y_{t+1:t+H}$, the oracle solves:

$$u^\star = \arg\min_{u \in \mathcal{U}} \sum_{h=1}^{H} u_{t+h}\, y_{t+h}, \quad \mathcal{U} = \Big\{ u \geq 0,\ \sum_{h=1}^{H} u_{t+h} = E \Big\}.$$

The forecast-based policy replaces y_{t+h} with the point forecast $\hat{y}_{t+h}$ and solves the same optimization problem to obtain $\hat{u}$.

Decision-centric metrics. Using the *true* carbon intensities, we compute:

$$\text{Regret} = C(\hat{u}; y) - C(u^\star; y), \qquad C(u; y) = \sum_{h=1}^{H} u_{t+h} y_{t+h},$$

and the *decision-quality ratio*:

$$\text{Ratio} = \frac{C(\hat{u}; y)}{C(u^\star; y)}.$$

We additionally benchmark against a simple *uniform* policy that distributes E evenly across all H hours. Averaging these metrics across all rolling validation windows reveals how forecasting errors translate into operational carbon outcomes.

5 Datasets

We evaluate our framework using multi-year data from three major North American electricity markets: ERCOT (Texas), NYISO (New York), and PJM (Eastern United States). All datasets are constructed at an hourly resolution and follow a consistent feature schema across regions.

Table 1: Forecasting performance of our temporal fusion model across three major North American power markets. Lower values indicate better predictive accuracy.

Region	MAE	MSE	WAPE	sMAPE	MAPE(|y|≥50)	MdAPE(|y|≥50)
ERCOT	436.01	1,589,173.63	0.0241	0.0297	0.2636	0.0141
NYISO	342.49	653,109.19	0.0475	0.0429	0.1653	0.0120
PJM	157.96	526,739.38	0.0148	0.0183	0.1578	0.0089

Weather forecasts. For each timestamp, we collect 96-hour-ahead weather forecasts including temperature, wind speed, humidity, cloud cover, and solar irradiance. These variables shape both renewable generation patterns and demand conditions, making them essential for short-term carbon intensity prediction.

Fuel-mix production forecasts. Each region publishes day-ahead projections for generator output across fuel types (e.g., coal, natural gas, nuclear, hydro, wind, solar, and other technologies). We align the full 96-hour fuel-mix forecast vector with the carbon intensity prediction horizon. These features capture anticipated changes in system-wide emissions attributable to generator dispatch.

Realized carbon intensity and system state. We obtain realized carbon intensity, realized fuel mix, and auxiliary system variables (e.g., load, renewable availability) from the respective ISO/RTO data portals. These form the basis for supervised learning and enable accurate evaluation of operational decisions.

Data processing and splits. All regions follow the same chronological split: *early data* for training, a *middle segment* for validation and hyperparameter selection, and the *latest window* for testing. We standardize continuous features using training-set statistics and align all predictor horizons to ensure a consistent 96-hour forecasting task across markets.

This unified data pipeline ensures that differences in forecasting and decision performance arise from regional grid characteristics rather than dataset inconsistencies.

6 Experiments

We evaluate our proposed approach across three major North American electricity markets—ERCOT, NYISO, and PJM—using a unified Temporal Fusion Transformer (TFT) architecture and identical hyperparameters. All experiments follow the same data-preprocessing, forecasting, and decision-simulation pipeline, enabling controlled comparisons across regions. We report results along two complementary axes: (i) pure forecasting accuracy, and (ii) decision-focused performance under a carbon-aware scheduling task. Together, these results demonstrate how forecast quality translates into operational improvements.

6.1 Forecasting Accuracy

We begin by assessing predictive accuracy using commonly adopted metrics: MAE, MSE, WAPE, and symmetric MAPE (sMAPE). As summarized in Table 1, the TFT achieves consistently strong accuracy across all datasets. WAPE remains low—approximately 2.4% in ERCOT, 4.7% in NYISO, and 1.5% in PJM—while sMAPE stays below 5% in every region. These results indicate that the model effectively captures the temporal structure and weather–fuel interactions that drive short-term variations in carbon intensity.

PJM exhibits the lowest forecast error, likely due to more stable fuel-mix patterns, while ERCOT and NYISO show somewhat higher variability associated with larger renewable penetration. Overall, the forecasting accuracy is sufficiently strong to support downstream operational tasks.

6.2 Decision-Focused Evaluation

Forecast accuracy alone does not guarantee high-quality operational decisions. We therefore embed the forecasts into a carbon-aware load-shifting simulation. At each forecast timestamp, a fixed amount

Table 2: Decision-focused evaluation. Regret measures the emission cost difference from an oracle optimal scheduler (lower is better). The ratio metric compares achieved objective to oracle (closer to 1 is better).

Region	Regret (Pred)	Regret (Uniform)	Ratio (Pred)	Ratio (Uniform)
ERCOT	157.87	15,245.23	1.00086	1.16303
NYISO	614.66	3,013.41	1.00560	1.08803
PJM	390.04	3,032.40	1.04712	1.07151

Table 3: Ablation study on PJM. Weather and fuel-mix forecasts both improve accuracy, while the attention mechanism contributes to more stable and lower-regret decisions.

Model Variant	WAPE	Regret
Full TFT Model	0.0148	390.0
No Weather	0.0321	1,402.5
No Fuel-Mix Forecasts	0.0217	873.4
No Attention Blocks	0.0189	651.8

of flexible load is allocated over a future 12-hour window according to: (i) the oracle (true future CI), (ii) the TFT-based predictions, and (iii) a naive uniform allocation baseline.

Performance is measured using (i) regret—how far the forecast-induced decision deviates from the oracle—and (ii) the ratio-to-oracle carbon cost (closer to 1 is better). Table 2 shows the results.

Across all regions, the TFT-based decisions substantially outperform the uniform baseline. In ERCOT, regret drops from over $15{,}000$ (uniform) to under 200 (TFT), representing a two-order-of-magnitude improvement. NYISO and PJM exhibit similar trends, with regret consistently reduced by large margins. The ratio-to-oracle metric remains tightly concentrated around 1.00–1.05 for TFT-based decisions, while the uniform baseline shows much larger deviations.

Figures 1a and 1b visualize these trends. The TFT consistently yields near-optimal operational behavior across all markets, demonstrating that even moderate improvements in forecasting accuracy can translate into large reductions in carbon-related operational error.

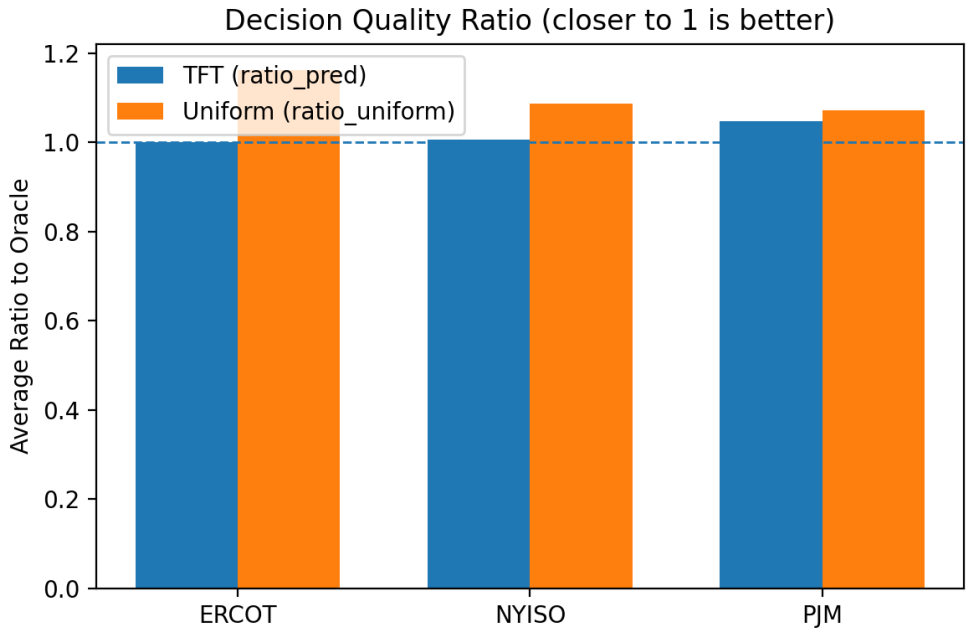
6.3 Ablation Study

To understand the contribution of different feature groups and model components, we conduct an ablation study on the PJM dataset, where the full model achieves the lowest forecasting error. We evaluate three reduced variants of the TFT:

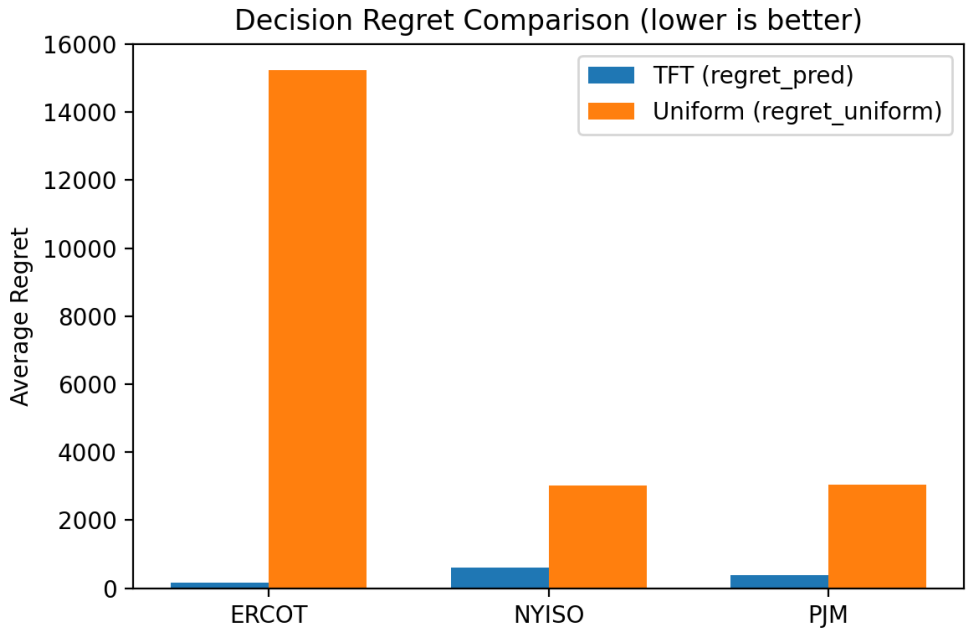
- **No Weather**: removes all weather forecast inputs.
- **No Fuel-Mix Forecasts**: removes future generation-mix forecasts.
- **No Attention Blocks**: disables the multi-head attention module.

Table 3 summarizes the results. Removing weather information leads to the largest degradation in both forecasting and decision metrics, highlighting the importance of meteorological inputs. Fuel-mix forecasts also contribute meaningfully, particularly for medium-range horizons, while the attention mechanism improves decision robustness and reduces regret.

These results show that (i) weather forecasts provide the strongest predictive signal, (ii) fuel-mix forecasts further enhance stability, and (iii) the attention module contributes to improved decision consistency. Together, they explain the overall strength of the full TFT model in both forecasting and decision tasks.



(a) Decision quality ratio across regions. Our model consistently outperforms the uniform baseline (closer to 1 is better).



(b) Decision regret (lower is better). Forecast-based scheduling reduces regret by one to two orders of magnitude relative to the uniform baseline.

Figure 1: Decision-focused evaluation on three markets. Top: ratio-to-oracle. Bottom: regret relative to an oracle scheduler.

7 Discussion and Future Work

Our results show that even a standard deep time-series model such as the Temporal Fusion Transformer can meaningfully support carbon-aware decision making when evaluated through a decision-focused lens. By pairing multi-horizon CI forecasts with a downstream carbon-minimizing scheduler, we are able to assess not only predictive accuracy but also the operational impact of forecast errors. The consistently low regret and near-oracle ratio performance demonstrate that moderate improvements in forecasting accuracy can yield disproportionately large reductions in decision loss.

At the same time, several limitations of our current framework present promising directions for future work.

Beyond stylized decision models. Our scheduling simulator captures the essence of carbon-aware load allocation but omits important operational constraints present in real power systems, such as generator ramp limits, network congestion, transmission losses, reserve requirements, and uncertainty in load baselines. Incorporating these constraints would enable a more realistic assessment of how forecast errors propagate into operational outcomes and could reveal structural differences across markets.

End-to-end decision-focused training. While our approach adopts a forecasting-first pipeline, recent work in decision-focused learning highlights the value of differentiating through the decision layer and optimizing predictive models directly for decision quality. Extending our framework to incorporate end-to-end differentiable or surrogate optimizers could further reduce decision regret and improve robustness under forecast uncertainty. Such an approach may also provide insights into which features and horizons most influence decision outcomes.

Nodal and spatially resolved carbon intensity. Our experiments focus on system-level CI, which is easier to forecast and widely reported by system operators. However, nodal or zonal CI can vary significantly due to network congestion and localized renewable availability. Extending our framework to spatially granular forecasts would better support applications such as data center siting, distributed energy resource scheduling, and geographically distributed demand response.

Integration with market and policy signals. Carbon-aware scheduling in practice interacts with electricity markets, price signals, and policy mechanisms such as carbon taxes or time-varying emission factors. Modeling how forecast-induced decisions interact with economic and regulatory incentives is an important direction, particularly for understanding the conditions under which carbon-aware strategies produce system-level emission reductions.

Overall, this work provides a foundational step toward evaluating carbon intensity forecasts in operational contexts. By highlighting the gap between statistical accuracy and decision quality, our results motivate continued research at the intersection of forecasting, optimization, and carbon-aware grid operation.

8 Conclusion

We presented a decision-focused framework for carbon intensity forecasting in large-scale power systems and instantiated it using a Temporal Fusion Transformer trained on multi-year data from ERCOT, NYISO, and PJM. By pairing forecasting models with a downstream carbon-minimizing load allocation simulator, we evaluated predictions not only through conventional statistical metrics but also through operational decision loss. Our empirical results show that the TFT achieves low forecasting error across all regions and, more importantly, enables decisions that are consistently close to oracle performance, reducing regret by one to two orders of magnitude relative to a uniform baseline.

These findings highlight a key insight: improvements in statistical accuracy do not always reflect the true value of a forecasting model for operational decision making. Decision-focused evaluation reveals performance differences that are invisible to MAE, WAPE, or sMAPE alone, underscoring the importance of evaluating forecasting models within the context of their downstream use cases. Our study shows that even without end-to-end decision-focused training, carefully designed forecasting models can significantly improve carbon-aware scheduling outcomes.

Looking ahead, integrating richer operational constraints, moving toward differentiable end-to-end decision-aware training, and extending the framework to nodal-level carbon intensity represent promising directions. As carbon-aware operation becomes increasingly central to energy system decarbonization, tools that jointly consider forecasting and decision-making will be essential.

A Additional Experimental Details

Train/validation/test splits. For each market (ERCOT, NYISO, PJM), we construct a continuous hourly time series and perform a chronological split into training, validation, and test sets. The earliest portion is used for model training, the middle segment for hyperparameter tuning and early stopping, and the most recent segment is reserved for final evaluation. This avoids information leakage from future timestamps into the training process.

Forecasting horizon and decision window. The TFT model is trained to produce 96-hour-ahead forecasts of system-level carbon intensity. For decision evaluation, we focus on a 12-hour window within this horizon, corresponding to the period over which a fixed amount of flexible demand must be allocated. At each decision point, we use the median predicted quantile as the point forecast for the simulator.

Simulation parameters. The flexible demand budget E is chosen so that the flexible load represents a modest fraction of total hourly demand, making the problem realistic while still allowing visible differences between policies. The same E is used across all regions for comparability. The oracle policy and the forecast-based policy are computed over identical scenario sets, and all regret and ratio metrics are averaged over the test period.

Implementation details. All TFT models share the same architecture across regions. Hyperparameters (such as hidden dimension, number of attention heads, and dropout rate) are selected using the validation split and then fixed for the final experiments. Training uses the quantile loss over all horizons and quantiles; early stopping is triggered when the validation WAPE stops improving.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '19)*, pages 2623–2631, 2019.

Priya Donti, Brandon Amos, and Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *NeurIPS*, 2017.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1): 9–26, 2022. doi: 10.1287/mnsc.2020.3922.

Hyojin Kim and Sungjin Lee. Greenflex: Demand-side flexibility scheduling for low-carbon power systems. *IEEE Transactions on Smart Grid*, 2021.

Kai Leerbeck, Peder Bacher, Rune Grønnegaard Junker, Goran Goranović, Olivier Corradi, Reza Ebrahimy, Amund Tveit, and Henrik Madsen. Short-term forecasting of co_2 emission intensity in power grids by machine learning. *Applied Energy*, 277:115527, 2020. doi: 10.1016/j.apenergy.2020.115527.

Bryan Lim, Sercan Ö. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pages 172–184, 2021.

Diptyaroop Maji, Prashant Shenoy, and Ramesh K. Sitaraman. Carboncast: Multi-day forecasting of grid carbon intensity. In *Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '22)*, pages 1–10, 2022. doi: 10.1145/3563357.3564079.

Gautam Mandi, Fan Yang, Pushmeet Kohli, and Philip Torr. Decision-focused learning: Through the lens of LP duality. In *ICML*, 2022.

Jay Mandi, Luca Furieri, Ferdinando Fioretto, et al. Decision-focused learning: Foundations, state of the art, benchmark and future opportunities. *Journal of Artificial Intelligence Research*, 80: 1623–1701, 2024.

Lukas Schmidthuber and Dogan Keles. Carbon intensity nowcasting and forecasting for power grids. *Applied Energy*, 2022.

Chris Thornton, Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '13)*, pages 847–855, 2013.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. End-to-end learning and optimization: A survey. In *IJCAI*, 2019.

Chen Xu, Kai Wang, and Yang Li. Carbonshift: Flexible load shifting for grid decarbonization. In *ACM e-Energy*, 2023.

Wei Zhang and Jun Li. Smartci: Data-driven carbon intensity forecasting. *Energy Informatics*, 2020.