advances in carbon neutrality

MDPI

Article

Carbon-Aware Scheduling of Flexible Electricity Loads via Fuel-Mix-Derived Grid Carbon Intensity: An ISO New England Case Study

Rajan Saha [1,2], Allan Yu [2,3], Rashanjot Kaur [4] and Eugene Pinsky [4,*]

[1] Conestoga High School, 200 Irish Road, Berwyn, PA 19312, USA; rajsaha@bu.edu
[2] RISE (Research in Science and Engineering) Program, Boston University, Boston, MA 02215, USA; allanyu@bu.edu
[3] Fairview High School, 1515 Greenbriar Boulevard, Boulder, CO 80305, USA
[4] Department of Computer Science, Metropolitan College, Boston University, Boston, MA 02215, USA; rashan@bu.edu
* Correspondence: epinsky@bu.edu

Abstract

Grid carbon intensity varies hour to hour, so a load that must finish by a deadline but not at a specific hour, such as electric-vehicle charging, heat-pump operation, or batch computation, can be shifted to a cleaner time at no cost to delivered energy. We present a transparent, fully reproducible pipeline that derives hourly, lifecycle-based grid carbon intensity for ISO New England from public fuel-mix data and documented emission factors, and that computes the exact minimum-carbon schedule for real deferrable loads: mixed-integer optimization is used where a shared per-hour allocation budget binds, and a provably optimal greedy placement where it does not. Over the full 2019 ISO New England year, hourly carbon intensity ranges from 116.7 to 389.4 gCO_2/kWh; in an idealized re-timing benchmark that holds delivered electrical energy fixed and does not model building thermal dynamics or comfort bounds, carbon-aware shifting of a residential heat-pump load avoids 8.29% of its average-basis emissions under a shifted-load allocation budget of three times its median active load (12.70% without one; 12.19% when, instead, total hourly draw is held within the building's observed peak). The attainable saving is set primarily by two operational quantities we characterize, deadline slack and the shared allocation budget, and it is strongly seasonal, ranging from roughly 10% in the milder shoulder seasons to about 3.1% in January, where the limit is the allocation budget rather than a shortage of clean hours: January's ceiling without an allocation budget is the highest of any month. It is robust across buildings and to the least certain emission factor (between 5.57% and 7.70%). In our experiments, a weather-blind climatology forecast retains about 88% of the savings without an allocation budget and 91.5% of the savings under the allocation budget.

Keywords: grid carbon intensity; carbon-aware scheduling; demand flexibility; load shifting; emission factors; IPCC AR5; ISO New England; decarbonization; low-carbon computing; sustainable operations

Check for updates

Academic Editor: John Bell

Received: 6 September 2026
Revised: 29 September 2026
Accepted: 15 October 2026
Published: 21 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The carbon intensity of electricity is not a constant. On any real power system, it rises and falls hour by hour as the generation mix shifts between fossil and low-carbon sources: when demand is met by gas- and oil-fired plants, the grid is dirty; when wind,

solar, hydro, and nuclear supply a larger share, it is clean. The same kilowatt-hour of demand, therefore, carries several times more carbon dioxide, depending solely on the hour at which it is drawn. This temporal structure is well documented and is the physical basis of every carbon-aware operating strategy: the grid supplies a time-varying carbon price that a consumer pays implicitly, whether or not they act on it [1–3].

Many electricity loads are indifferent to *when*, within a window, they run. Electric-vehicle charging must be completed before the vehicle departs, but it need not draw power continuously from the moment it is plugged in; a heat pump maintaining a comfort band can pre-heat or pre-cool and coast; a batch computation must finish by a service deadline, but it does not care which hours it occupies in between. For electric-vehicle and batch loads, this deadline-bounded flexibility is a directly observed or directly assigned property of the job. For a heat pump it is an idealization: the latitude to run an hour's consumption earlier or later exists only insofar as the building's thermal mass keeps the indoor temperature within its comfort band, and we do not model that thermal state. We model acting on flexibility as re-timing alone: a job is moved whole, at its metered power, so that the schedule changes only *when* energy is drawn and not *how much*, and any saving is attributable to timing rather than to curtailment. For the heat-pump load the results are therefore an idealized scheduling benchmark rather than a demonstrated, physically attainable saving: the formulation omits indoor-temperature dynamics, comfort bounds, the energy penalty of pre-conditioning, and the dependence of heat-pump efficiency on outdoor temperature (Section 6.2). The engineering question is not whether the headroom exists but how much carbon it can remove and on what that amount depends [4].

The gap this paper addresses is one of transparent, reproducible measurement of the *ceiling* rather than of algorithm design. Production systems already schedule against carbon [5], open pipelines already derive grid carbon intensity from public data [2], and the deferrable-load literature has long shifted thermostatic and charging loads for cost and emissions [6–8]. What is comparatively scarce is an end-to-end, auditable estimate of the *best* saving attainable by re-timing alone, under explicitly stated modeling idealizations, on a specific grid, built entirely from public inputs, reported as a function of the operational quantities that govern it, and honest about how modest the number is. Recent work has argued directly that the returns to carbon-aware temporal shifting are limited on the grounds that carbon intensity varies little at many locations and that, across regions, the locations where it varies least tend to be those with the highest absolute emissions, so the lever is weakest where it would matter most [9]. Measuring the ceiling exactly, rather than quoting a heuristic scheduler's achieved saving, is precisely what allows that claim to be tested rather than assumed.

1.1. Research Motivation

Acting on the grid's time-varying carbon requires a carbon-intensity signal an operator can audit. The signals used in practice are often proprietary, report only combustion-based marginal rates, or lack a documented provenance for every emission factor behind them. A decision to defer a load is only as defensible as the number that justified it, and for independent verification, that number must be reconstructible from public data, with each coefficient traceable to a named source. Equally, a single headline percentage is of little use to an operator: the achievable saving depends on how much deadline slack a load actually has and on how much relocated load an operator permits to be concentrated into any single clean hour. That allocation budget is a policy parameter; for example, a limit on the rebound peak that synchronized deferral creates; it is not an equipment rating, and Section 3.3 treats a physical limit on total demand separately. A measurement that collapses these dependencies into one figure hides exactly the levers a practitioner

controls. Our motivation is therefore to produce a carbon-accounting and scheduling pipeline that is transparent in its inputs, exact in its optimization, and explicit about the surface Savings $= f(\text{slack}, \text{budget})$ rather than a point estimate.

1.2. Research Questions

This study investigates the following questions:

1. How much does the hourly, lifecycle-basis carbon intensity of the Independent System Operator New England (ISO New England) grid vary over a full year, when it is reconstructed from public fuel-mix data and documented emission factors?
2. What is the ceiling on carbon savings that temporal shifting of real deferrable loads (electric-vehicle charging, residential heat-pump operation, and batch computation) can attain on this grid, computed exactly rather than heuristically, under an idealized re-timing model?
3. How does that ceiling depend on the two operational quantities an operator actually controls, namely the deadline slack available to each load and the per-hour allocation budget shared among loads?
4. How robust is the measured saving to the choices embedded in it: the least-certain emission factor, the particular building or site, the season of the year, and the substitution of a realistic forecast for perfect foresight of the carbon signal?

1.3. Contributions

The contributions of this paper are:

- **A transparent, lifecycle-based carbon-intensity pipeline.** We derive hourly grid carbon intensity for ISO New England from public U.S. Energy Information Administration (EIA) fuel-mix data and a tabulated set of eight emission factors, each reported with its provenance, including the two coefficients that are *not* IPCC AR5 lifecycle medians and are flagged as such. The pipeline is reproducible end-to-end from public inputs, and because the signal is a deterministic identity over published data rather than a learned forecast, any hour of it can be re-derived by hand.
- **An exact ceiling on temporal-shifting savings for non-datacenter deferrable loads under a shared allocation budget.** Using real electric-vehicle sessions, a real residential heat-pump load, and a real batch-computation trace, we compute the minimum-carbon feasible schedule exactly (by mixed-integer linear programming where a shared allocation budget binds and by a greedy placement that provably attains the optimum where it does not), so that the reported figure is the exact optimum of the stated scheduling model rather than one heuristic's achievement; for the heat-pump load, that model is an idealized benchmark that omits thermal dynamics and comfort constraints.
- **A savings surface rather than a headline number.** We characterize how the ceiling moves with deadline slack and with a shared per-hour allocation budget, quantifying the penalty a simple greedy rule pays exactly when the budget binds, rather than assuming savings scale monotonically with flexibility.
- **An honest robustness account.** We report the sensitivity of the headline saving to the least-certain emission factor, to the choice of building, to the season, and to replacing oracle foresight of the carbon signal with a weather-blind climatology forecast, and we engage directly with the argument that such savings are inherently limited on a demand-correlated grid.

We deliberately avoid several claims that the literature would not support. We do not claim to be first, we do not claim that schedulers ignore carbon, and we do not present the scaling of savings with slack as a novel finding; each of these is established prior

work [2,5,10]. Our position is narrower and, we believe, defensible: a transparent, lifecycle-based carbon accounting and scheduling pipeline for spatially fixed deferrable loads under a shared allocation budget, built end to end from public data so that the avoided average-basis emissions can be independently verified.

The remainder of the paper is organized as follows. Section 2 reviews related work on carbon-aware datacenter computing, average versus marginal emissions accounting, and the scheduling of deferrable loads, ending with the specific gap this paper fills. Section 3 sets out the carbon-intensity identity and its emission factors, the three load models, the scheduling optimization, and the data provenance. Section 4 reports the carbon signal, the per-load and slack- and budget-resolved savings, the seasonal pattern, and the robustness analyses. Section 5 discusses what the results mean, Section 6 states the limitations, and Section 7 concludes.

2. Related Work

2.1. Carbon-Aware Datacenter Computing

Temporal shifting of computational load to reduce operational carbon is an established idea with a substantial body of prior work [5,10]. Wiesner et al. demonstrated that deferring cloud workloads into low-carbon hours reduces emissions and characterized how the achievable reduction grows with the scheduling flexibility available; that savings increase with deadline slack is therefore not a new observation but a starting point [10,11], and Section 4.3 reports it as a measured property of this grid. Radovanović et al. describe a carbon-aware computing system operated in production at Google, which shifts flexible datacenter load in response to day-ahead carbon-intensity forecasts, establishing that large operators already act on exactly the signal studied here [5]. Other work has examined the movement of jobs between data centers across different regions [12]. Our study keeps loads in one region and shifts them only in time. Acun et al. present Carbon Explorer, a framework for evaluating operational and embodied carbon-reduction strategies for data centers, including temporal load shifting, renewable procurement, and storage [13]. Sukprasert et al. provide an important counterweight: they analyze the limitations of both temporal and spatial workload shifting and find the realistically attainable savings from temporal shifting alone to be modest, because carbon intensity varies little at many locations, and, across regions, the locations with the least variability tend to be those with the highest absolute emissions [9]. Their anti-correlation is spatial (between a location's carbon variability and its carbon level) and is drawn across regions; we measure a temporal analog of it within a single region, across seasons (Section 4.5). Our work is positioned against this literature not as a better scheduler but as an exact, publicly reproducible measurement of the ceiling for non-datacenter loads on a specific grid, and our seasonal results speak directly to the Sukprasert argument.

2.2. Average Versus Marginal Emissions Accounting

A carbon-intensity signal can be built on either of two accounting bases, and the choice materially changes the decision [14,15]. An *average* signal attributes to each kilowatt-hour the mean emissions of all generators in that hour; a *marginal* signal attributes to it the emissions of the generator that would actually respond to a marginal change in demand. A third basis is also in use: consumption-based accounting traces power flows from producer to consumer and charges an importing region for the mix it actually draws, which, for a net importer such as ISO New England, can diverge materially from the production-basis mix we compute here [3]. Callaway, Fowlie, and McCormick set out the economic and environmental consequences of this distinction, showing that average and marginal emission factors can rank locations differently and that the marginal basis is the one aligned

with the causal effect of a consumption decision [1]. Graff Zivin et al. show that the basis can also reverse the ranking of an *intervention*: on marginal rates that vary by region and by hour, a plug-in electric vehicle charged at the recommended overnight hours emits more per mile than a comparable hybrid across the entire Eastern interconnection, which contains ISO New England [16]. The signals published by grid operators and by commercial data providers differ on precisely this axis, and much of the datacenter literature is implicitly average-based, while some deferrable-load work is explicitly marginal [6]. That choice of basis is contested rather than neutral: Huber et al. argue that average factors should not be used to approximate marginal ones and can mislead a scheduler outright [8].

A distinct line of work produces the carbon signal itself rather than consuming it. CarbonCast forecasts average grid carbon intensity up to 96 h ahead across thirteen regions, combining per-source generation forecasts from neural networks with historical intensity and weather forecasts in a hybrid CNN–LSTM, and reports four-day mean absolute percentage errors spanning 3.42–19.95% depending on region [2]. The difference is not in public code availability but lies in the object produced and in what can be audited: CarbonCast outputs a learned prediction of a future signal, whereas the signal used here is a deterministic identity (Equation (1)) over published fuel mix and a tabulated factor set whose provenance is stated coefficient by coefficient, so any hour can be re-derived by hand. The two compose rather than compete. CarbonCast never schedules a load; this study measures what an exact scheduler could attain and then prices the cost of replacing oracle foresight with a weather-blind climatology (Section 4.9). A forecaster of CarbonCast's quality could narrow the 1.46 percentage-point gap reported there; we have not tested this. We compute our carbon signal on a lifecycle-average basis, for which every emission factor is publicly documented. Section 4.10 then measures what that choice costs, scoring an average-optimized schedule against a marginal reference rather than assuming the average basis is decision-relevant. Consumption-based accounting is noted here for completeness but not pursued: the public fuel-mix series this study rests on does not resolve ISO New England's import mix.

2.3. Scheduling Deferrable Loads: Thermostatic Loads and EV Charging

Outside the datacenter, the power-systems literature has studied the scheduling of deferrable loads for decades. Thermostatically controlled loads (water heaters, air conditioners, heat pumps) store energy thermally and can be advanced or delayed within a comfort band, and a large body of work treats them as a dispatchable, aggregable resource for demand response and, more recently, for emissions reduction [7,17,18]. Electric-vehicle charging is the other canonical deferrable load: a session specifies an energy requirement and a departure deadline, leaving the intervening hours free to be chosen. Li et al. optimize EV charging against marginal carbon emission factors and time-of-use prices, reporting meaningful emissions reductions from timing alone and illustrating the marginal-basis approach in practice [6]. Closest to our setting, Yang et al. schedule thermostatically controlled loads against a carbon signal using a bilevel distributionally robust chance-constrained formulation, hedging over a Wasserstein ambiguity set against forecast errors in *ambient temperature and base load* while respecting comfort bounds [19]. Their contribution is a controller that stays feasible and near-optimal when the weather and the non-thermostatic demand are uncertain; the carbon signal itself is taken as given. Ours is a measurement of what is attainable under perfect foresight *of that signal*, against which such a controller can be benchmarked, together with the transparency to reconstruct the signal from public inputs. Uncertainty in the carbon signal is hedged by none of the work above: Wiesner et al. perturb it with uniform random noise and name that a limitation [10], Radovanović et al. consume a vendor forecast and report its error without hedging against it [5], and Yang et al.

randomize the weather rather than the signal. Section 4.9 addresses that gap directly, replacing the oracle with a causal climatology forecast and pricing what it costs. This literature also motivates the allocation budget central to our study: when many deferrable loads respond to the same signal, they move into the same clean hours, so an operator relocating them must decide how much relocated load any one hour may absorb. Section 3.3 formalizes that allocation budget as M, together with a separate physical bound on total draw, and Section 4.4 measures what each costs. That budget, rather than the existence of flexibility, is what makes the scheduling problem nontrivial.

2.4. Research Gap

The prior work above establishes that carbon-aware shifting works, that it is deployed in production, that its savings grow with slack, that the accounting basis matters, that deferrable loads are well studied, and that thermostatic loads in particular have been scheduled against carbon signals by controllers robust to weather and base-load uncertainty [19]. Wiesner et al. are the closest prior work and provide public code and data while analyzing regional variation, scheduling flexibility, strategy, and forecast error; public availability and the flexibility–savings relationship are therefore not gaps [10]. Their carbon data are empirical, but their two evaluated workload scenarios are simulated; their scheduler assumes no shared allocation budget; and their forecast-error analysis adds uniform random noise, with the latter two explicitly identified as limitations. Consequently, that study does not evaluate three different workload classes using public ResStock (version 3.4.0), Adaptive Charging Network (ACN-Data), and Alibaba inputs against one common grid signal, address the shared per-hour allocation budget, or quantify a heuristic's optimality gap; compare average- and marginal-emissions accounting; or report savings robustness across buildings, seasons, emission-factor choices, and bootstrap resamples. The other studies differ in scope as opposed to quality. Radovanović et al.'s work is unauditable because its carbon feed is proprietary [5]; no load is scheduled against CarbonCast's predictions [2]; Yang et al. model what we do not—thermal dynamics, comfort bounds, and robustness to weather uncertainty—but treat the carbon signal as given, without provenance, accounting basis, or sensitivity analysis signal [19]; Sukprasert et al., similar to our work, do bound attainable saving, but only for cloud workloads, using published intensity regional intensity rates, and with no shared allocation budget, so their bound says nothing about non-datacenter loads competing for the same clean hours [9]. What remains comparatively underdeveloped is a single artifact that (i) builds a lifecycle-average carbon signal from public fuel-mix data with every emission factor's provenance stated, including those that are not standard medians; (ii) computes the *exact* minimum-carbon schedule under perfect foresight, so that the reported figure is a model ceiling against which thermally constrained controllers such as that of Yang et al. [19] can be benchmarked, rather than the achievement of any particular controller, and quantifies what a simple rule forfeits against it; and (iii) reports that ceiling as an explicit surface over deadline slack and the allocation budget, together with its sensitivity to the least-certain factor, to the site, to the season, and to forecast error. This paper is that artifact for the ISO New England grid. It is a measurement contribution, and its value lies in being auditable end-to-end rather than in proposing a new scheduling algorithm.

3. Materials and Methods

The pipeline has three stages: a carbon signal $CI(t)$ derived from public fuel-mix data (Section 3.1); a set of deferrable jobs derived from real load datasets (Section 3.2); and a scheduling optimization that places each job in its lowest-carbon feasible window subject to its deadline and to a shared per-hour allocation budget (Section 3.3). Figure 1 shows

the end-to-end workflow. Because the total energy delivered is fixed by the scheduler's constraints, every gram of carbon dioxide avoided is attributable solely to timing, not to any reduction in energy use. Section 3.4 tabulates the provenance of every input. Three choices in what follows are inherited from Section 2 rather than made freely. The signal is built on the lifecycle-average basis because every coefficient it requires is publicly documented (Section 2.2), where a marginal signal for this system and year is not; Section 4.10 measures what that costs. Flexibility is modeled as re-timing alone—a job moves whole at its metered power—so that delivered energy is fixed and any avoided carbon is attributable to timing, at the price of omitting the thermal state and comfort bounds that controllers such as Yang et al.'s carry [19]. And the scheduler is constrained by a shared allocation budget because the relocated load that any one hour absorbs must be bounded (Section 2.3); without that budget the problem separates into independent placements reached greedily, which is why Section 4.4 treats the budget rather than slack as the operationally decisive quantity.

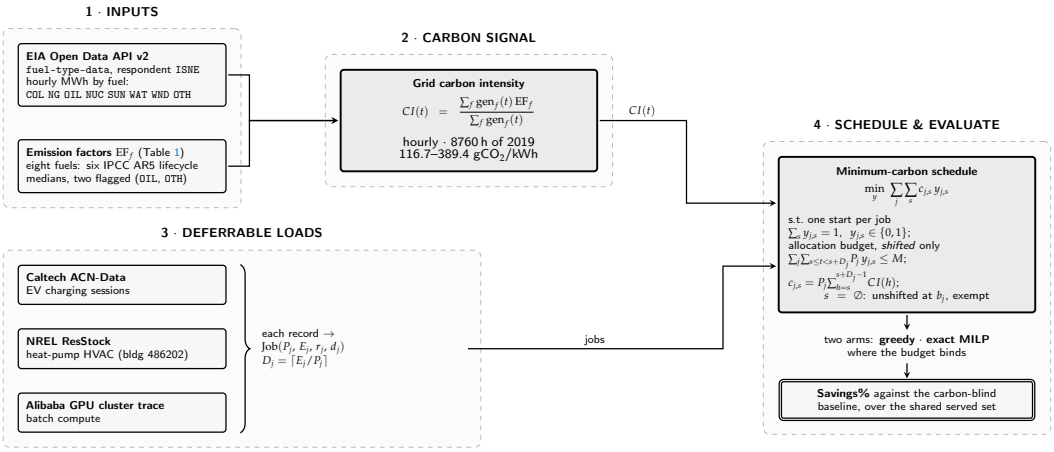


Figure 1. Methodological workflow of the carbon-aware scheduling pipeline: public hourly fuel-mix data and documented emission factors are combined into an hourly grid carbon-intensity signal $CI(t)$; real deferrable loads (electric-vehicle sessions, a residential heat-pump load, and a batch-computation trace) are converted into deadline-bounded jobs; and a scheduler places each job in its lowest-carbon feasible window subject to its deadline and a shared per-hour allocation budget, with avoided carbon evaluated against a carbon-blind baseline over the set of jobs both arms serve.

3.1. Grid Carbon Intensity from Fuel Mix

The carbon intensity of the grid in hour t is the generation-weighted average of the emission factors of the fuels generated in that hour. Let $\mathrm{gen}_f(t)$ be the electricity generated by fuel f in hour t (in MWh) and EF_f the lifecycle emission factor of fuel f (in $\mathrm{gCO_2/kWh}$). Then

$$CI(t) = \frac{\sum_f \mathrm{gen}_f(t)\,\mathrm{EF}_f}{\sum_f \mathrm{gen}_f(t)} \qquad [\mathrm{gCO_2/kWh}]. \tag{1}$$

Equation (1) is a weighted mean: it is bounded below by the cleanest fuel generating in hour t and above by the dirtiest, and it moves purely as the mix reweights between them from one hour to the next. It is a lifecycle-*average* signal by construction, since each kilowatt-hour is charged the mean emissions of the hour's generation, and we return to the average-versus-marginal distinction in Sections 2.2 and 4.10. Throughout, *avoided emissions* means the reduction in emissions attributed to a load on this average basis. It is not an estimate of the causal change in system emissions, which would require a marginal signal; Section 4.10 evaluates the schedules against an observed one.

The hourly fuel mix is taken from the U.S. Energy Information Administration (EIA) Open Data API v2 `fuel-type-data` endpoint for respondent `ISNE` (ISO New England), which reports hourly generation by fuel type in eight categories: coal (`COL`), natural gas (`NG`), petroleum/oil (`OIL`), nuclear (`NUC`), solar (`SUN`), hydro (`WAT`), wind (`WND`), and other (`OTH`) [20,21]. The eight emission factors and their provenance are given in Table 1. Six of the eight are lifecycle medians from the IPCC Fifth Assessment Report (AR5), Working Group III, Annex III [22]. The remaining two are flagged explicitly because they are not AR5 medians and because they carry more uncertainty than the rest: the petroleum/oil factor of 650 gCO_2/kWh is taken from the datacenter carbon-accounting literature (Carbon Explorer and CarbonCast) rather than from AR5 [2,13], and the `OTH` factor of 230 gCO_2/kWh is a biomass-median proxy that we adopt for the heterogeneous refuse, biomass, and landfill-gas blend that EIA reports under `OTH`. We report the sensitivity of every headline result to this proxy in Section 4.8.

Table 1. The eight emission factors used to compute grid carbon intensity via Equation (1), with provenance. Six are lifecycle medians from IPCC AR5 WG-III Annex III [22]. Two are flagged: the petroleum/oil factor is credited to the datacenter carbon-accounting literature (Carbon Explorer/CarbonCast), *not* to IPCC SRREN; and the `OTH` factor is a biomass-median proxy this work adopts for the refuse/biomass/landfill-gas blend that EIA groups under a single "other" code. The proxy is subjected to a dedicated sensitivity analysis in Section 4.8.

Code	Fuel	EF (gCO_2/kWh)	Provenance
COL	Coal	820	IPCC AR5 lifecycle median [22]
NG	Natural gas	490	IPCC AR5 lifecycle median [22]
OIL	Petroleum/oil	650	Carbon Explorer/CarbonCast [2,13] (*not* AR5)
NUC	Nuclear	12	IPCC AR5 lifecycle median [22]
SUN	Solar PV	48	IPCC AR5 lifecycle median [22]
WAT	Hydro	24	IPCC AR5 lifecycle median [22]
WND	Wind	11	IPCC AR5 lifecycle median [22]
OTH	Other (refuse/biomass)	230	Biomass-median proxy, this work (*flagged*)

3.2. Deferrable Load Models

We study three archetypal deferrable loads. Each real record is converted into a job j with a power rating P_j (kW), an energy requirement E_j (kWh), an earliest start r_j, and a deadline d_j, from which a whole-hour duration $D_j = \lceil E_j/P_j \rceil$ and a slack $s_j = (d_j - r_j) - D_j$ follow. The scheduler bills energy in whole-hour blocks at full power (a convention discussed in Section 6); both arms share the bill. For the heat-pump load, it cancels out exactly because each hourly block's power equals that hour's energy, and E_j/P_j is therefore identically 1; no job is rounded. The electric-vehicle load is evaluated in continuous time (below) and is not subject to this convention. For the batch load, whose durations are not whole hours, it does not cancel exactly: rounding can change which window is cheapest, and it shifts that percentage by less than 0.2 percentage points.

Electric-vehicle charging. We use the Caltech ACN-Data dataset of real workplace and campus charging sessions [23]. For each session, the energy requirement is the delivered energy, $E_j = \texttt{kWhDelivered}$; the charging power is inferred from the active charging duration, $P_j = E_j/\tau_j$ with $\tau_j = \texttt{doneCharging} - \texttt{connect}$, and the session is scheduled in continuous time: it charges one contiguous block of length τ_j at P_j, starting at any real time $s \in [\texttt{connect}, \texttt{disconnect} - \tau_j]$. Availability is therefore exactly the observed plug-in interval, with no charging before arrival or after departure, and the energy is exactly

E_j, with no rounding of duration. Because $CI(t)$ is constant within each hour, the cost $C_j(s) = P_j \int_s^{s+\tau_j} CI(t)\,dt$ is continuous and piecewise linear in s, with breakpoints only where s or $s + \tau_j$ crosses an hour boundary. Its minimum over the interval is therefore attained at an endpoint or at one of those breakpoints, and evaluating that finite candidate set gives the exact continuous-time optimum. Without a budget the sessions do not interact, so these per-session minima are the global optimum. The carbon-blind baseline is the observed charge, [`connect`, `doneCharging`]. A session delivering 24.5 kWh over a 3.5 h active charge within a nine-hour connection window; for example, has $P_j = 7.0$ kW, $\tau_j = 3.5$ h, and 5.5 h of slack. Figure 2 shows the resulting slack distribution across the whole session set.

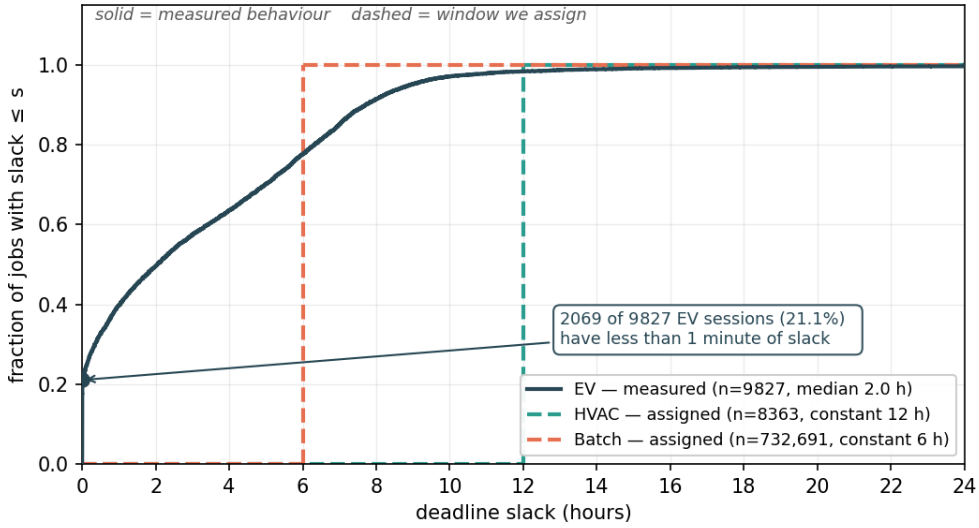


Figure 2. Deadline slack of the 9827 Caltech ACN-Data charging sessions (2019), measured in continuous time as `disconnect` − `doneCharging`. The median is 2.0 h against a median dwell of 5.9 h, and 21.05% of sessions have less than one minute of slack. Under our earlier whole-hour discretization, 2495 sessions (29%) appeared to have zero slack, but only 979 of them have less than one minute of real slack; the rest were artifacts of flooring and rounding. 33 sessions with slack above 24 h are omitted from the axis; the tail runs to 139 h. The heat-pump and batch loads appear as dashed vertical lines because their slack is assigned rather than measured: a constant 12 h ($2 \times$ `flex_hours`) and 6 h, respectively.

Residential heat-pump HVAC. We use hourly electric-HVAC energy for a Massachusetts heat-pump home from the NREL ResStock dataset [24]. Each hour whose summed electric-HVAC consumption exceeds 0.05 kWh becomes a one-hour job whose energy is that consumption and whose power is therefore that energy over one hour ($D_j = 1$ h). The threshold excludes hours when the heat pump is effectively idle and there is nothing to defer: of the 8738 h with nonzero consumption, 375 fall below it, leaving the 8363 jobs reported throughout. The job is assigned a deadline window of $\pm$ `flex_hours` about the hour in which the load actually ran, i.e., $r_j =$ `run` $-$ `flex` and $d_j =$ `run` $+ 1 +$ `flex`, so that a flexible thermostat may advance or delay that hour's operation by up to `flex` hours. This window is an assumed scheduling latitude, not a property derived from the building. No indoor-temperature state is tracked, so we do not verify that a schedule keeps the dwelling within a comfort band, and the electrical energy of a relocated hour is held at its metered value even though a heat pump's coefficient of performance falls as the source–sink temperature difference grows [25]. Results in Section 4 report the saving for windows from $\pm$2 h to $\pm$8 h so that a reader can select a latitude consistent with a particular building's thermal time constant. The carbon-blind baseline prices each such job at the real hour in which it ran, not at its earliest feasible hour, so the saving measures

displacement from actual operation. The ResStock profile is generated on actual 2018 meteorology, whereas the carbon signal is for 2019, so the load timestamps are advanced by 365 days. Both years are non-leap, making this an exact day-for-day mapping that preserves the calendar date and therefore the season, at the cost of a one-day weekday drift. We preserve the date rather than the weekday because the seasonal component dominates the weekly one in both signals: over 2019, the carbon intensity varies by 93.1 gCO_2/kWh between its cleanest and dirtiest months but only 21.4 gCO_2/kWh between weekdays and weekends, and the heat-pump load varies by 1.139 kW seasonally against 0.023 kW between weekdays and weekends. All four figures are means taken over every hour of the year, including idle hours: the seasonal term is the range across the twelve monthly means, and the weekly term is the difference between the weekday and weekend means.

Batch (AI) computation. We use the Alibaba GPU cluster trace (v2020) [26]. Each batch task becomes a job whose power is the product of its GPU count and an assumed per-GPU power draw, $P_j = \texttt{num_gpus} \times \texttt{gpu_power_kw}$; its energy is that power times the task duration; and it is assigned a deadline window. Because the available flexibility is not directly observed, we sweep over `flex_hours` and report the batch result as a function of `flex_hours`. The assumed per-GPU power scales the baseline and the scheduled emissions equally, so it cancels out in the reported percentage and affects only the absolute tonnage; we confirm this across `gpu_power_kw` values of 0.3, 0.4, and 0.5, which yield savings identical to four decimal places. The Alibaba trace records task timing as offsets in seconds from an unspecified epoch rather than as calendar dates, so it must be anchored to a carbon year to price it. Alibaba reports the trace as spanning July and August 2020, and we therefore anchor the trace's second zero to 1 July 2019, mapping the workload to the same calendar dates one year earlier; the tasks then span 7 July 2019 to 13 September 2019. The placement is season-faithful by construction rather than arbitrary. Re-anchoring to January, April, or October gives flex-6 savings of 4.49%, 5.34%, and 5.03%, respectively, so the reported July figure is the most conservative of the four placements tested.

3.3. The Scheduling Optimization

Let $D_j = \lceil E_j/P_j \rceil$ be job j's whole-hour duration and $\mathcal{W}_j = \{ s : r_j \le s \le d_j - D_j \}$ the contiguous start hours available to it. Let $y_{j,s} \in \{0,1\}$ indicate that j starts in hour s, and $y_{j,\varnothing} \in \{0,1\}$ that it runs unshifted at its metered hour b_j. Writing $c_{j,s} = P_j \sum_{h=s}^{s+D_j-1} CI(h)$ for the cost of running j from s, the carbon-minimizing schedule solves

$$\min_{y} \quad \sum_{j} \Big(\sum_{s \in \mathcal{W}_j} c_{j,s}\, y_{j,s} + c_{j,b_j}\, y_{j,\varnothing} \Big) \tag{2}$$

$$\text{s.t.} \quad \sum_{s \in \mathcal{W}_j} y_{j,s} + y_{j,\varnothing} = 1 \qquad \forall j \quad \text{(run exactly once)} \tag{3}$$

$$\sum_{j} \sum_{\substack{s \in \mathcal{W}_j \\ s \le t < s + D_j}} P_j\, y_{j,s} \le M \qquad \forall t \quad \text{(shifted-load allocation budget)} \tag{4}$$

$$y_{j,s},\ y_{j,\varnothing} \in \{0,1\} \qquad \text{(discrete placement).} \tag{5}$$

Constraint (3) makes every job run exactly once, either in a window drawn from $\mathcal{W}_j$ or unshifted at b_j, so the objective (2) rewards timing alone; (5) makes that choice discrete rather than a divisible allocation of power; and constraint (4) bounds the power the scheduler may *relocate* into any single hour. We read M as a shifted-load allocation budget rather than as a physical service limit: it is the headroom an operator commits to deferred load, not a rating the site's total draw must respect. Jobs the optimizer cannot place within budget, and jobs for which no feasible slot improves on doing nothing, run at their original

metered hour and are not counted against M: the set of relocated jobs is $\mathcal{S} = \{\, j : y_{j,\varnothing} = 0 \,\}$, and (4) sums only over shift variables, so a job running unshifted contributes nothing to the budget. The baseline floor is implicit rather than a separate constraint: because the objective already carries c_{j,b_j}, the cost of running unshifted, the optimizer selects $y_{j,\varnothing}$ whenever no feasible window improves on the metered hour, and no job is ever placed at a cost above doing nothing. At the headline operating point ($M = 1.269$ kW, three times the median active load), the relocated load in the exact capped optimum peaks at M and reaches it in 32 h; the site's overall hourly peak, including unshifted jobs, reaches 6.667 kW and exceeds M in 1845 h, and 1944 of 8363 jobs run at their metered hour (1731 in the greedy schedule). Because M excludes unshifted load, it does not enforce any equipment rating. To address physical capacity directly, we also solve the stricter problem in which total hourly draw, shifted plus unshifted, may not exceed a rating $\bar{P}$:

$$\sum_j P_j \Big(\sum_{\substack{s \in \mathcal{W}_j \\ s \le t < s + D_j}} y_{j,s} \;+\; 1[\, b_j \le t < b_j + D_j \,]\, y_{j,\varnothing} \Big) \le \bar{P} \qquad \forall t. \tag{6}$$

For the heat-pump home we set $\bar{P}$ to the building's observed peak hourly HVAC draw, 6.667 kW. This is a conservative stand-in for the equipment rating: the unit is observed to draw it, nothing above it is assumed available, and the do-nothing schedule satisfies (6), so the problem is always feasible. It prevents the optimizer from stacking several hours of heat-pump operation into one hour beyond what the unit has been observed to draw. The value of M itself is not derived from a physical rating: $k = 3$ is one operating point on the swept surface discussed in more detail in Section 4 (k = 2, 3, 4, 6 and the observed peak), and the conclusions of Section 4.4 are stated across that sweep rather than at $k = 3$ alone. Scheduling deferrable loads described by an arrival time, an energy requirement, and a flexibility window is a well-studied problem; Shao et al. pose the same object as a constrained finite-horizon dynamic program and solve it by fitted value iteration [27]. The substantive difference from our formulation is preemption: their charging processes may be interrupted and resumed, whereas ours are not. Constraint (3), together with the block cost $c_{j,s}$, places each job as a single contiguous whole-hour run, which is what makes (4) a coupling constraint over start variables rather than over divisible per-hour power.

Two regimes matter. When the budget constraint (4) does not bind, the jobs do not interact: each is independently placed in its own cheapest feasible window, and a greedy rule that assigns every job its minimum-cost window is therefore globally optimal. In this uncapped regime, the greedy schedule *equals* the exact optimum, which we verify with regression tests in the code repository. When the budget binds, jobs compete for the same clean hours, and the problem becomes a genuine combinatorial optimization; we solve it exactly as a mixed-integer linear program (MILP) over the contiguous-placement decision variables, and we quantify the gap between the greedy heuristic and the MILP optimum as the price of using a simple rule under contention. The MILP is solved with HiGHS [28] through `scipy.optimize.milp` (SciPy 1.18.0 [29]), with the HiGHS default relative optimality tolerance of 10^{-4}. *Exact* therefore means optimal to within that relative gap. The capacity sweep solved all terminated with status Optimal. Section 4 lists the achieved relative gap and the dual bound of each, and the dual bound brackets the saving from above. All savings are evaluated over the set of jobs that *both* the carbon-aware and the baseline arm serve feasibly, so that a scheduler cannot manufacture an apparent saving by stranding difficult jobs.

The carbon-blind control differs by load, and the distinction is not cosmetic. For electric-vehicle sessions the baseline is the observed charging interval, which begins at the earliest feasible start, so a first-come-first-served baseline and an unshifted baseline coincide

by construction. For the heat-pump load, the window is $[r_j, d_j) = [\texttt{run} - \texttt{flex}, \texttt{run} + 1 + \texttt{flex})$, so placing every block at its earliest feasible start would advance the entire building's operation by exactly `flex` hours. That is itself a load-shifting intervention, pre-heating or pre-cooling the whole dwelling, whose magnitude is a function of the very parameter under study; nor is it conservative, since the displaced hours are measurably cleaner. We therefore price each heat-pump job at the rate for the hour in which it actually ran. We define the saving of a carbon-aware schedule with total carbon C_{aware} against this carbon-blind baseline with total carbon C_{base} as

$$\text{Savings\%} = \frac{C_{\text{base}} - C_{\text{aware}}}{C_{\text{base}}} \times 100\%, \tag{7}$$

and we distinguish the greedy saving ($C_{\text{aware}} = C_{\text{greedy}}$), the exact capped optimum ($C_{\text{aware}} = C_{\text{opt}}$), and the uncapped ceiling (the exact optimum with $M \to \infty$), which satisfy $C_{\text{opt}} \leq C_{\text{greedy}} \leq C_{\text{base}}$ when the budget binds. We call a schedule *capped* when it is computed with M in force and *uncapped* when $M \to \infty$.

3.4. Data and Provenance

Table 2 records the source, resolution, and role of every dataset and constant entering the pipeline so that each reported quantity can be traced to a public input. The carbon signal relies on two public sources (EIA fuel mix and the emission factors in Table 1); the loads rely on three public datasets; and the scheduling constants (the allocation budget M and the per-job slack) are parameters we sweep rather than fixed values.

Table 2. Provenance of every input to the pipeline. All primary sources are public. "Swept" indicates a parameter varied across a range in the results rather than fixed to a single value.

Input	Source	Role/Resolution
Hourly fuel mix	EIA Open Data API v2, respondent `ISNE` [20]	Hourly MWh by fuel; drives $CI(t)$ over 8760 h of 2019
Emission factors	IPCC AR5 + flagged sources (Table 1) [2,13,22]	Fixed gCO_2/kWh per fuel
EV sessions	Caltech ACN-Data [23]	Per session energy + connect/done/disconnect times
HVAC load	NREL ResStock (Massachusetts heat-pump home) [24]	Hourly electric-HVAC kWh; ±flex deadline
Batch tasks	Alibaba GPU cluster trace v2020 [26]	GPU count + duration; assigned deadline
Allocation budget M	This work	Swept; per-hour limit on relocated load
Total-draw rating $\bar{P}$	Observed peak HVAC draw (this work)	Fixed at 6.667 kW; per-hour limit on total draw
Deadline slack	This work	Swept; per-load flexibility in hours

4. Results

Unless stated otherwise, all results use the hourly ISO New England fuel mix for the full calendar year 2019 (8760 h, respondent `ISNE`), the emission factors of Table 1, and the HVAC headline configuration of a single Massachusetts heat-pump home (ResStock building 486,202) with six hours of deadline flexibility.

4.1. The Carbon Signal

Over the 8760 h of 2019, the reconstructed ISO New England carbon intensity of Equation (1) has a minimum of 116.7, a median of 260.4, and a maximum of 389.4 gCO_2/kWh. The signal is reported over calendar 2019; the fetched series extends by two days to ensure jobs near the year boundary retain priceable flexibility windows.

The cleanest hour is thus roughly one-third as carbon-intensive as the dirtiest, a spread of about 273 gCO_2/kWh, and it is precisely this spread that temporal shifting can, in principle, exploit. The range is modest by the standards of coal- or hydro-dominated grids: ISO New England is a comparatively gas-heavy, comparatively clean system, so its carbon signal never reaches the extremes of a system that swings between coal and renewables. This bounds the achievable savings before any scheduling is attempted.

Figure 3 shows the carbon signal as a month-by-hour heatmap, exposing the two structures a scheduler exploits: a diurnal pattern and a seasonal pattern across the rows. The diurnal shape is a single overnight trough rather than the twin-trough profile of a solar-heavy grid: mean intensity bottoms out at 221.5 gCO_2/kWh at 03:00 standard time (UTC−5), climbs steadily through the morning, and then holds a broad plateau from late morning to a 20:00 peak of 283.9 gCO_2/kWh before falling sharply after 21:00. Midday belongs to that plateau rather than forming a second trough: at 276.5 gCO_2/kWh it lies within 7.4 of the daily maximum, 88% of the way from trough to peak, and the same ordering holds in every season: ISO New England is gas-heavy, with too little solar to depress midday intensity. The salient feature relevant to Section 4.5 is that the overnight trough persists in every month, so a cleaner hour is reachable year-round; what changes is the level around it. In July and August, even the cleanest hour of the average day (287.5 and 268.9 gCO_2/kWh) is dirtier than the 261.0 gCO_2/kWh mean of the month-by-hour cells of Figure 3, and they are the only two months for which that holds; in January, the corresponding figure is 208.4.

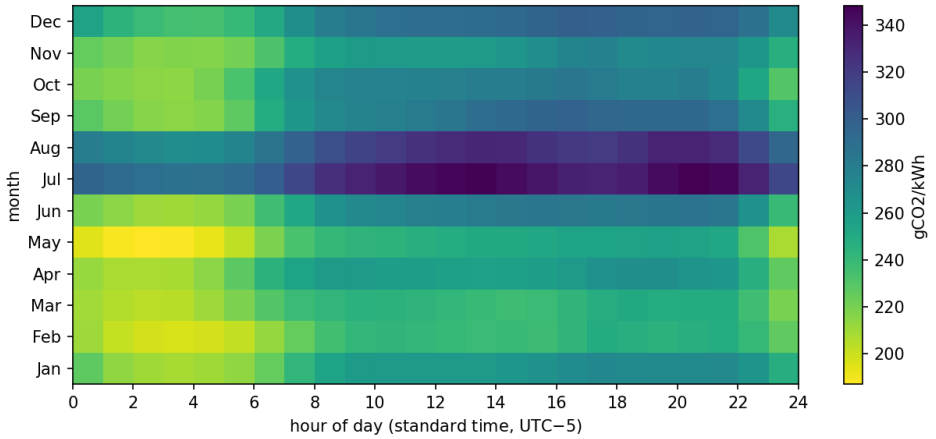


Figure 3. Hourly ISO New England grid carbon intensity for 2019, computed from public fuel-mix data via Equation (1), shown as a month (rows) by hour-of-day (columns) heatmap in gCO_2/kWh. Hours are in ISO New England standard time (UTC−5); a single fixed offset is used because a month-by-month climatology cannot represent daylight-saving time transitions without smearing the columns. Over the full year, the signal ranges from 116.7 (cleanest hour) to 389.4 (dirtiest hour) with a median of 260.4. The dirtiest cells cluster in the evening demand peak, but the overnight trough is present in every month (a within-day range of 46 to 82 gCO_2/kWh), so what varies seasonally is the level rather than the availability of a trough: in July and August the trough itself sits above the 261.0 gCO_2/kWh mean of the cells plotted here, which is what limits the summer ceiling.

4.2. Savings by Load Type

Figures 4 and 5 compare the carbon savings attainable for the three load archetypes. The residential heat-pump load is our headline case because both its data and its flexibility model are fully specified; the electric-vehicle and batch cases are reported alongside it, with the batch case swept over flexibility as described in Section 3.2 and shown in Figure 4. For the EV workload, the Caltech ACN-Data 2019 pull contains 10,617 charging sessions. Of these, 790 are invalid records and are excluded: 0 with missing or non-positive delivered energy, 466 with a missing timestamp, 7 whose charging end does not follow plug-in,

and 317 whose charging end is recorded after unplugging. The remaining 9827 sessions are all schedulable in continuous time. In our previous hourly formulation a further 1479 physically feasible sessions were rejected only because flooring the window and rounding the duration up left negative whole-hour slack; the continuous-time evaluation retains them. Sessions with little slack are retained: they are feasible but nearly immovable, and their large share is part of why the EV saving is small. Table 3 collects the per-load headline saving alongside the measured slack that produces it.

Table 3. Carbon savings by load type, with the deadline slack that produces each figure, measured for the electric-vehicle sessions and assigned for the heat-pump and batch loads, on the 2019 ISO New England signal. Figures are the exact uncapped optimum, placed on a common basis so the loads compare like for like; the budget-constrained HVAC headline of 8.29% appears in Section 4.4. The electric-vehicle row uses the real 2019 Caltech (workplace) ACN-Data sessions, whose short 2.0-h median slack leaves little room to shift and therefore yields a much smaller saving than the heat-pump load, directly illustrating the dependence on slack. The electric-vehicle row is evaluated in continuous time (Section 3.2). The batch row uses all 732,691 completed tasks. Table 4 lists the configuration behind every reported saving.

Load	Schedulable Jobs	Savings (%)	Slack
HVAC (heat pump)	8363	12.70 (exact, uncapped)	12 h (±6 h flex)
EV charging	9827	1.81 (exact, uncapped)	2.0 h (median slack)
Batch (AI)	732,691	3.97 (exact, uncapped)	6 h (assigned flex)

Table 4. Configuration behind each reported saving (Reviewer 1, Comment 6). All savings are computed over the set of jobs that both the carbon-aware and the baseline arm price.

Result	Job Set and Filter	Jobs	Flex	Budget	Solver
HVAC 12.70%	ResStock 486202, hours > 0.05 kWh	8363	±6 h	none	greedy (=MILP)
HVAC 8.29/7.28%	same	8363	±6 h	$M = 1.269$ kW	MILP/greedy
HVAC 12.19%	same	8363	±6 h	total draw ≤ 6.667 kW	MILP
EV 1.81%	ACN-Data 2019, valid records	9827	observed	none	exact, continuous time
Batch 3.97%	Alibaba v2020, Terminated tasks with usable duration and GPU count	732,691	6 h	none	greedy (=optimum)
Batch, Forecast 11.24/7.58%	same as HVAC	732,691 8363	6 h ±6 h	k = 2–6 none/$M = 1.269$ kW	greedy greedy/MILP on forecast

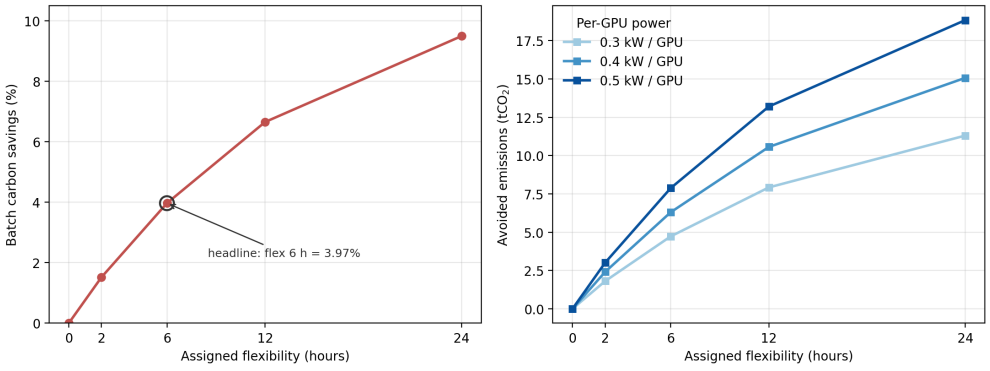


Figure 4. Batch (AI) savings on the 2019 ISO New England signal as a function of assigned flexibility, over 732,691 completed tasks from the Alibaba GPU cluster trace. (**Left**): the saving rises from 0% with no flexibility to 1.52% at two hours, 3.97% at six, 6.65% at twelve, and 9.49% at twenty-four. The percentage is identical for every assumed per-GPU power because power scales the baseline and the schedule equally, so they cancel in the ratio. (**Right**): the absolute avoided emissions do scale with that assumption, reaching 11.3, 15.1, and 18.8 tCO_2 at twenty-four hours for 0.3, 0.4, and 0.5 kW per GPU, respectively.

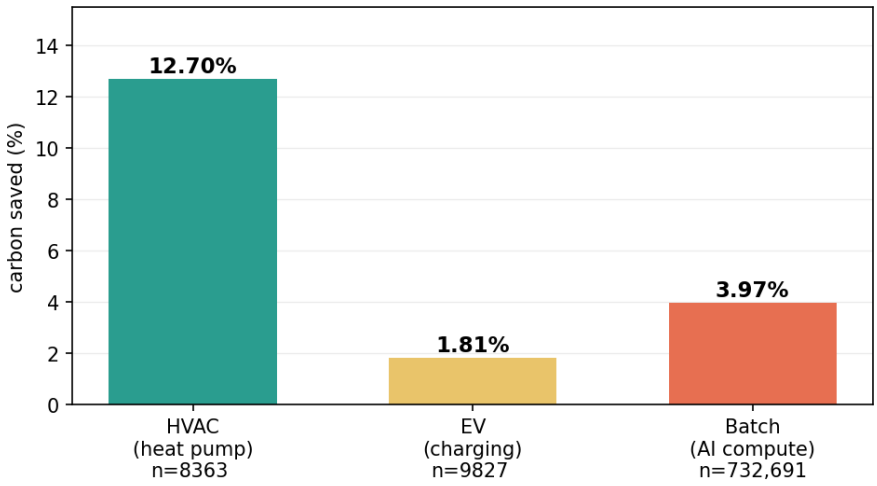


Figure 5. Carbon savings by load type on the 2019 ISO New England signal. Savings differ across archetypes because their deadline slack and power profiles differ: the heat-pump job is assigned 6 h of flexibility about its metered hour (a symmetric ±6 h window), whereas the median EV session leaves only 2.0 h between the end of charging and departure. The batch trace records only relative timestamps, which we anchor to 1 July 2019; the batch bar therefore covers roughly ten weeks of the 2019 signal rather than the full year, so its percentage is valid within-load, but its absolute total is not commensurable with the other two.

4.3. Savings Versus Deadline Slack

Figure 6 shows how the HVAC saving grows as the deadline flexibility `flex_hours` is increased. The saving rises steeply from little or no flexibility because a job confined to its actual hours cannot move at all, and continues to increase across the swept range, from 5.01% at two hours of flexibility to 15.53% at eight hours. Flattening is expected once the window spans the diurnal cycle, since the additional hours a wider window then reaches are no cleaner than those already available, but within the range swept here, the relationship remains close to linear. This monotonic relationship is consistent with the flexibility–savings relationship established in prior work [10]; we reproduce it here as a measured property of this grid rather than as a novel finding. This is shown in Figure 6.

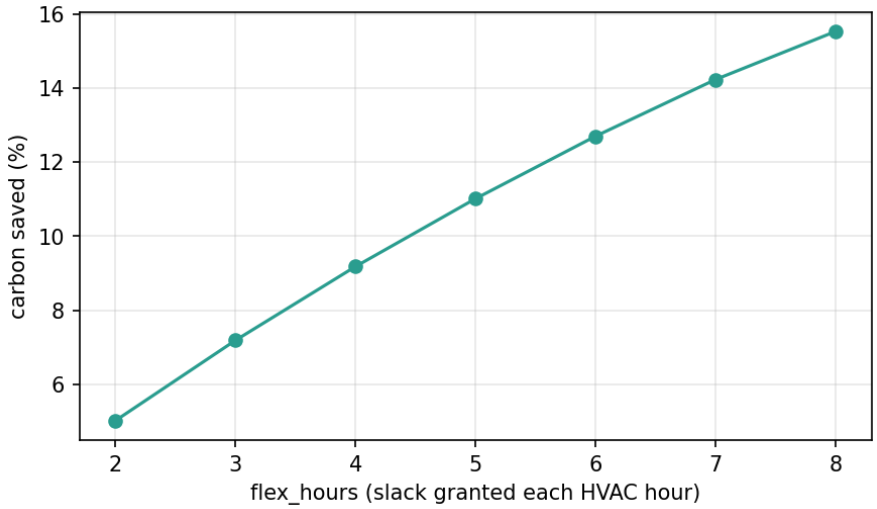


Figure 6. HVAC carbon savings as a function of deadline flexibility (`flex_hours`) on the 2019 ISO New England signal. Savings increase monotonically with slack across the swept range, from 5.01% at two hours to 15.53% at eight. Diminishing returns are expected once the window spans the diurnal cycle, because the additional hours a wider window then reaches are no cleaner than those already within reach.

4.4. Savings Versus the Allocation Budget

The shared per-hour allocation budget M is the operational lever that turns the scheduling problem from a set of independent placements into a contention problem, and it is where a simple rule and the exact optimum diverge. Figure 7 shows the HVAC and batch savings as M is tightened, both plotted against k, the budget expressed as a multiple of each load's own median hourly aggregate draw. When the budget is generous, the jobs do not compete, and the greedy schedule equals the optimum; as the budget tightens, jobs are forced out of the cleanest shared hours into dirtier ones, and the savings fall toward zero.

At the headline operating point (a budget set to three times the median active HVAC load, with six hours of flexibility over the full year 2019 for building 486202), the exact capped optimum saves 8.29%, while the greedy heuristic saves 7.28%. The greedy rule, therefore, forfeits 1.01 percentage points precisely because it commits jobs to locally cheapest hours without anticipating the contention the budget creates, whereas the MILP resolves that contention globally. For reference, removing the budget entirely raises the ceiling to 12.70%: roughly a third of the model's uncapped saving on this load is lost to the allocation budget, which is why we regard the allocation-budget dimension, rather than slack alone, as the operationally decisive one.

The allocation budget does not bound total draw (Section 3.3). Enforcing the physical limit of Equation (6) instead, with total hourly draw held within the building's observed peak of 6.667 kW, gives 12.19%, and enforcing both gives 8.29%, unchanged, because at $M = 3\times$ median the schedule's total draw already stays within 6.667 kW (Table 5). The uncapped 12.70% is therefore not attainable by a single heat pump whose hourly draw is limited to its observed peak; with that limit, the model's ceiling is 12.19%.

The two loads exhaust that lever at very different budgets, and the shared k axis is what makes the difference visible: the batch arm reaches its uncapped ceiling of 3.97% at k = 6 and gains nothing beyond it, whereas the heat pump is still climbing at $k = 15.8$ (the ratio of its observed peak to its median hourly draw) and has not reached 12.70% even there. The mechanism is the concentration of demand in time rather than its magnitude. A per-hour budget binds only insofar as the carbon-optimal schedule seeks to stack load into a few clean hours, so what determines where the budget stops mattering is the load's own peak-to-median ratio, not its absolute size. The heat pump is a single dwelling contributing about one block per hour, and its hourly draw peaks at 15.8 times its own median, so its carbon-optimal schedule has ample room to concentrate, and a budget at any modest multiple of the median bites hard. The batch fleet is the opposite: across the 1642 h, its trace spans roughly 991 jobs concurrently at a nearly flat aggregate draw, peaking at only 2.2 times its median, so there is little concentration for a budget to prevent, and one set above 6 times the median ceases to bind at all. One caveat bounds the comparison: the batch trace spans only 7 July 2019 to 13 September 2019, so its 3.97% ceiling is a summer-window figure, and the like-for-like heat-pump ceiling over those same 69 days is 9.57% rather than the full-year 12.70%, so the two ceilings differ by a factor of about 2.4 rather than 3.2. All batch results use one population: the 732,691 completed (`Terminated`) tasks with a usable duration and GPU count, out of 1,261,050 trace rows (885,073 `Terminated`). An earlier version of the capacity sweep additionally dropped tasks at or below 0.1 kWh, leaving 227,529 jobs and an uncapped saving of 3.21%; because that threshold was applied after multiplying by the assumed per-GPU power, it made the job set depend on that assumption, and we have removed it. This is shown in Figure 7.

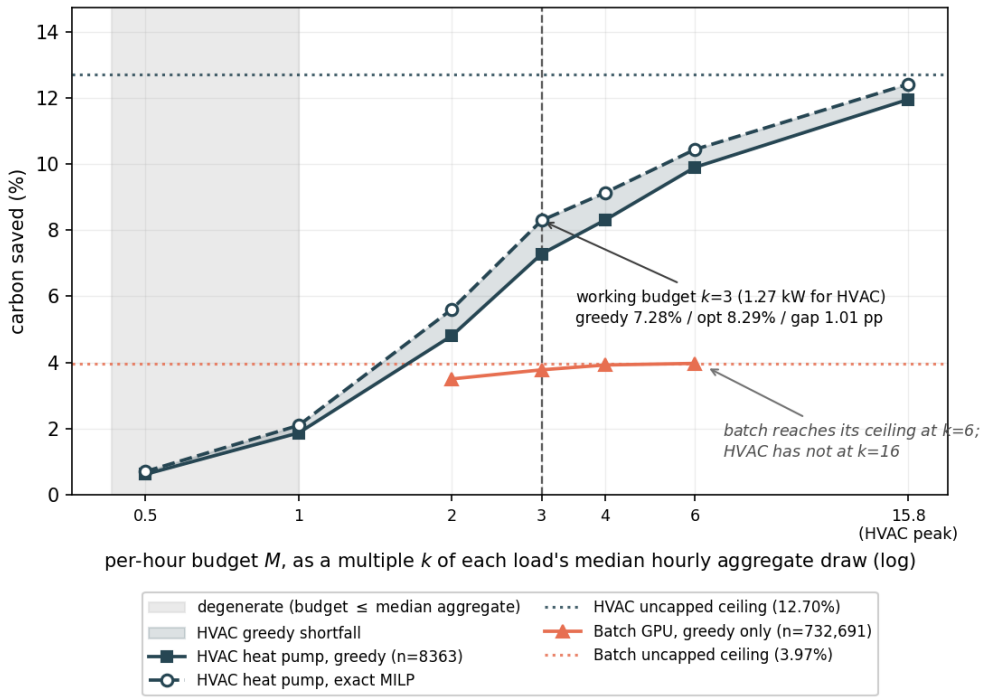


Figure 7. Carbon savings as a function of the shared per-hour allocation budget M on the 2019 ISO New England signal for the residential heat-pump load (building 486202, six hours of flexibility) and the batch GPU load. Both arms are swept on a common dimensionless axis, $M = k\times$ that load's own median hourly aggregate draw; the two differ by roughly 600× in absolute kilowatts, so only the normalized budget makes them comparable. As the budget tightens, jobs are displaced from the cleanest shared hours, and the savings fall. At the headline budget $k = 3$, the exact optimum for the heat pump saves 8.29% and the greedy heuristic 7.28% (a 1.01 percentage-point gap that appears only when the budget binds); with no budget, the ceiling is 12.70%, though the two arms do not cover the same span: the heat pump runs over all of 2019, while the Alibaba trace covers 69 summer days, over which the heat-pump ceiling is 9.57% (Section 4.4). No exact optimum is reported for the batch arm (732,691 jobs, greedy only): the dense constraint matrices of our original implementation exhausted memory above about 32,000 jobs.

MILP termination data is summarized in Table 5.

Table 5. MILP termination for every exact heat-pump result (building 486202, ±6 h). Relative gap = |incumbent − dual bound|/|incumbent|; the upper-bound column converts the dual bound into the largest saving any feasible schedule could reach. Savings for the M rows are those plotted in Figure 7.

Constraint	Status	Relative Gap	Saving (%)	Upper Bound (%)
$M = 2\times$ median	Optimal	8.4×10^{-5}	5.61	5.615
$M = 3\times$ median	Optimal	6.8×10^{-5}	8.29	8.299
$M = 4\times$ median	Optimal	8.5×10^{-5}	9.13	9.140
$M = 6\times$ median	Optimal	7.7×10^{-5}	10.44	10.444
M = observed peak (6.667 kW)	Optimal	3.0×10^{-5}	12.41	12.417
Total draw $\leq$ 6.667 kW, no M	Optimal	5.7×10^{-5}	12.19	12.194
$M = 3\times$ median and total $\leq$ 6.667 kW	Optimal	9.2×10^{-5}	8.29	8.298

4.5. Seasonal Variation

The seasonal pattern is the study's most consequential finding, and it speaks directly to the argument that carbon-aware shifting yields limited savings [9]. Figure 8 shows the HVAC savings by month. Savings are highest in the transition seasons (about 10.8% in fall

and 9.3% in spring) and lowest in the demand-peak seasons (about 5.5% in winter and 6.6% in summer), with January the worst month at 3.1%. Season figures are energy-weighted, as are the monthly figures in Figure 8. This is the within-region temporal analog of the cross-region spatial anti-correlation Sukprasert et al. report: they find that the *locations* with the least carbon variability are those with the highest absolute emissions, and we find the same in summer: July and August combine below-average window spread with the year's largest heat-pump loads after January's. The correspondence does not extend to winter, where variability is above the annual mean and the binding constraint is the allocation budget rather than the signal (Section 4.6). Sukprasert et al.'s is inferred across published regional intensity series; ours is measured within one region, on one signal, against one metered load. The basis matters here: on the lifecycle-*average* signal used throughout, the summer demand peak raises the day's overall intensity, so the ceiling itself is lowest then: a cleaner hour still exists within every window, but the improvement it offers is smaller. The January minimum has a different cause, which Section 4.6 separates. On a *marginal* basis the seasonal ordering need not hold, because the marginal unit's emission rate does not track the hour's average mix [8,16]; Section 4.10 scores the schedules on an observed marginal reference. Flexibility is worth less in the winter cold snap and summer heat wave, exactly when a heat pump's load is largest and the operator would most want to abate it—but for different reasons: in summer because the signal offers least, in winter because the budget binds hardest against the largest loads. This anti-correlation between when abatement is wanted and when it is attainable is the honest limit on temporal shifting, and it is visible here as a direct measurement rather than an assertion.

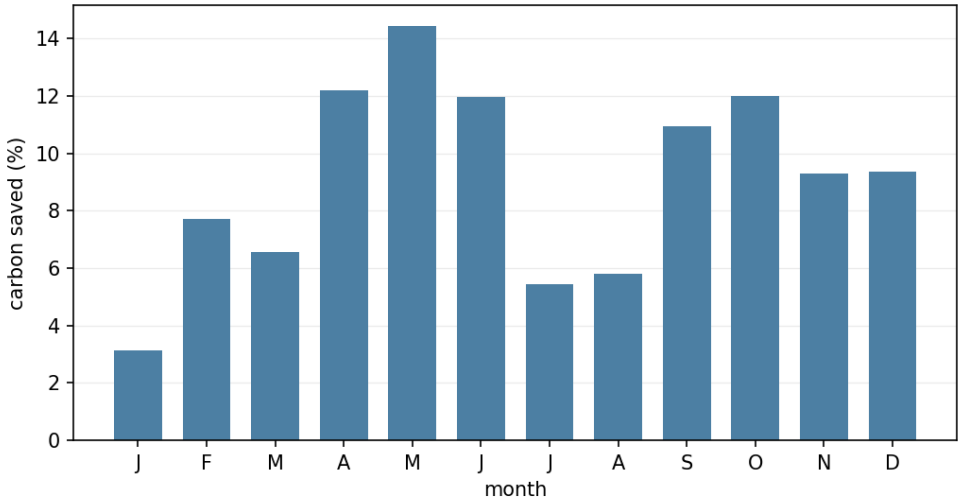


Figure 8. HVAC carbon savings by month on the 2019 ISO New England signal (building 486202). Savings peak in the mild shoulder seasons (fall ≈ 10.8%, spring ≈ 9.3%) and fall in the demand-peak seasons (winter ≈ 5.5%, summer ≈ 6.6%), reaching a minimum of 3.1% in January; all figures are energy-weighted. The two minima have different causes: in summer the ceiling itself is lowest, because the whole day is carbon-intensive and the available improvement is small, whereas in January clean hours remain available and the allocation budget is what binds (Section 4.6).

4.6. What Determines Savings

The seasonal pattern above is a result rather than an explanation. Because each HVAC job occupies a single hour, the uncapped saving decomposes exactly, with no fitted model. Write $\mathcal{W}_j$ for job j's feasible window and b_j for the hour in which it actually ran. With

$$\text{spread}_j = \max_{h \in \mathcal{W}_j} CI(h) - \min_{h \in \mathcal{W}_j} CI(h), \qquad \text{position}_j = \frac{CI(b_j) - \min_{h \in \mathcal{W}_j} CI(h)}{\text{spread}_j}, \tag{8}$$

the saving attained by job j is

$$\text{Savings}_j = \frac{\text{position}_j \times \text{spread}_j}{CI(b_j)}. \tag{9}$$

These are the three quantities at issue. spread_j is the carbon volatility the grid offers inside the window; $\text{position}_j \in [0,1]$ is the original load timing, zero when the job already run in the cleanest hour it could reach and one when it ran in the dirtiest; and deadline slack sets $\mathcal{W}_j$, and hence both. The three are therefore not independent, which is why a regression over them would read correlated inputs. Equation (9) is an identity rather than a fit: summed over the 8363 jobs and weighted by energy, it reproduces the uncapped ceiling of 12.70% exactly, with a maximum per-job residual of zero.

Slack acts on both terms, and unequally. Widening the window from one to eight hours raises the mean spread from 14.1 to 69.4 gCO_2/kWh while moving the mean position only from 0.511 to 0.605: most of what slack buys is access to a wider carbon range, not a worse starting point. The ceiling rises from 2.68% to 15.53% accordingly, reproducing the sweep of Section 4.3.

Applied by month, the decomposition separates two mechanisms that Section 4.5 reports as one. In July and August, the signal is the binding constraint: mean intensity over the jobs' baseline hours, weighted by job energy as throughout this section, is 322.5 and 309.2 gCO_2/kWh, the highest of the year, and the uncapped ceiling falls to 8.44% and 9.87%. Across the twelve months, the correlation between mean baseline intensity and the uncapped ceiling is -0.79. That is the mechanism Sukprasert et al. describe, measured directly [9]. January is not that case. It means baseline intensity of 245.0 gCO_2/kWh is below the annual energy-weighted mean over baseline hours of 267.5, its mean window spread of 63.4 gCO_2/kWh is above the corresponding annual mean spread of 58.9, and its uncapped ceiling of 15.55% is the highest of any month: clean hours are amply available in January. Its capped savings of 3.1% are a content effect instead. January forfeits 12.42 percentage points to the allocation budget, compared with 6.38 for the next-worst month, and across the year, the correlation between a month's heat-pump energy and the budget loss it incurs is 0.85. The arithmetic is direct: the mean January job draws 1.32 kW against a budget of 1.269 kW, so a single average job exhausts the entire allocation.

The seasonal minimum therefore has two distinct causes, and only one is a property of the carbon signal. In summer, the ceiling itself is low because the grid is uniformly dirty. In winter, the ceiling is the highest of the year, and the budget prevents reaching it. The distinction is operationally consequential, because the two admit different remedies: the summer ceiling moves only with deadline slack, and slowly widening the window from six to eight hours lifts the July ceiling from 8.44% to 10.57%—whereas the winter loss is recovered by enlarging the allocation budget, which Section 4.4 shows returns the saving to 12.70%.

4.7. Robustness Across Buildings

To test whether the headline saving is an artifact of a single home, we repeat the HVAC analysis on two further Massachusetts heat-pump buildings under the same configuration. They yield savings of 5.64% (95% confidence interval: 5.04–6.30) and 6.89% (6.06–7.78), against the headline building's greedy figure of 7.28% (6.72–7.87). Intervals are 1000 paired bootstrap resamples blocked by the whole day so that a day's heat-pump jobs are resampled together, and intra-day correlation is preserved, with a fixed seed per building; they describe within-year daily variability for a given building and not uncertainty across the Massachusetts building population. The three buildings therefore agree to within about 1.6 percentage points, indicating that the result is a property of the grid signal and the load class rather than of one particular dwelling. The residual spread reflects genuine

differences in each home's load timing: a building whose HVAC demand already falls in cleaner hours has less to gain from shifting.

4.8. Sensitivity to the OTH Emission Factor

Because the `OTH` emission factor of 230 gCO_2/kWh is the least-certain input, a proxy rather than a documented median (Table 1), we sweep it across a plausible range and re-measure the headline saving. As Figure 9 shows, the greedy HVAC saving moves between 5.57% and 7.70%, sensitive to the factor but never leaving the single-digit-percent range, so no plausible value overturns the conclusion.

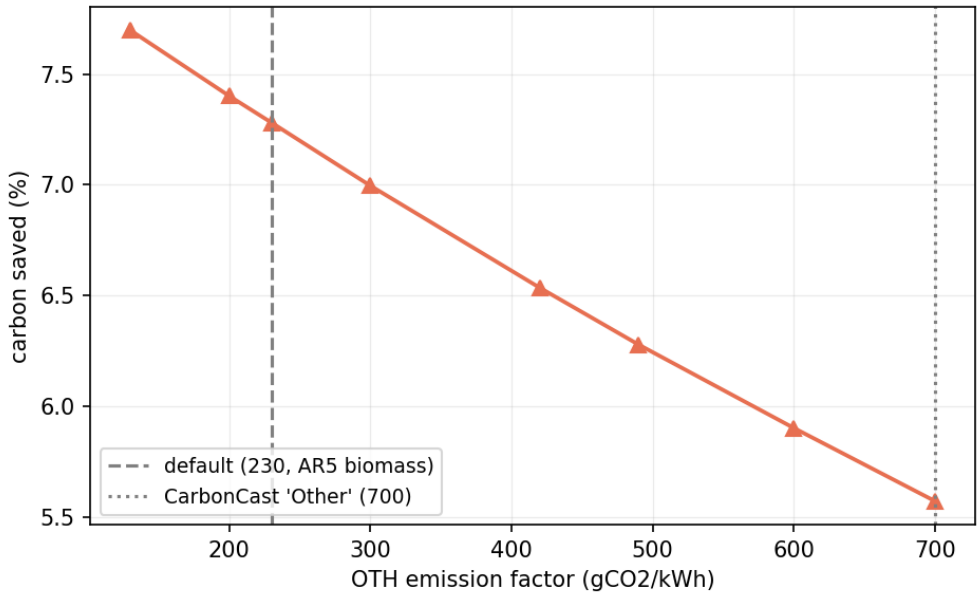


Figure 9. Sensitivity of the greedy HVAC carbon saving to the `OTH` emission factor, which is a biomass-median proxy rather than a documented median (Table 1). Across the swept range, the saving moves between 5.57% and 7.70%, so the result depends on but is not dominated by this least-certain factor: no plausible value moves the saving out of the single-digit-percent range. The solid orange line with triangle markers traces the greedy HVAC saving computed at each swept `OTH` value; the two vertical reference lines mark the default proxy (230 gCO_2/kWh, the AR5 biomass median; dashed) and the CarbonCast "Other" value (700 gCO_2/kWh; dotted).

4.9. Forecast Penalty

Every result above assumes perfect foresight of the carbon signal (an "oracle"), which no operator has. To measure the cost of realistic foresight, we replace the oracle with a weather-blind *climatology* forecast. It is causal: each hour is predicted from the mean of *strictly earlier* hours sharing its month and hour-of-day, so no hour contributes to its own prediction, and no later observation is used. Where that cell has no history yet (the first occurrence of each month-and-hour pair), the forecast falls back to earlier hours of the same hour-of-day, and failing that to the running mean of all earlier hours; only the first hour of the series has no history of any kind. No weather information is used at any point.

Issue time and information set. The forecast used to schedule job j is issued at its release time r_j, the metered hour minus `flex`. For each hour h in the job's window, the primary predictor and its first fallback use only observations at the same hour of day on earlier days ($h - 24$ h, $h - 48$ h, …). Because the window spans at most $2\,\texttt{flex} + 1 = 13$ h, every such observation is at least 12 h older than r_j and was available at issue time. The last fallback, the running mean of all earlier hours, is used only on 1 January, the first day of the series; it can draw on observations after r_j, and it enters the windows of 25 of the 8363 jobs. Each schedule is computed on the forecast and then priced on the observed signal.

Figure 10 shows that under this forecast, the saving falls from an oracle of 12.70% to 11.24%, a penalty of 1.46 percentage points in the uncapped case. This oracle figure coincides with the uncapped ceiling of Section 4.4, as it must: both price the same unconstrained schedule against the same baseline. Repeated with the same allocation budget ($M = 1.269$ kW) and the same 8363 jobs as the 8.29% headline, scheduling on the forecast with the exact MILP gives 7.58% against the oracle 8.29%, a penalty of 0.71 percentage points, so 91.5% of the capped saving is retained; with the greedy rule the figures are 6.60% against 7.28%. Under the budget, 639 jobs are placed in hours that are dirtier on the observed signal than their metered hour, because the forecast misranks them. The penalty is not uniform across the year: it is 2.63 percentage points in January, when almost no history has accumulated, against 0.60 in July, once six months are available. We report 1.46 percentage points as the penalty observed in this experiment. A climatology trained on prior years would begin January with a full set of month-and-hour cells, which may reduce the January penalty, but we have not tested this and do not claim that 1.46 percentage points bounds the penalty of other forecasts or years. In this experiment the climatology forecast retains about 88% of the uncapped saving and 91.5% of the capped saving, which is an encouraging result for practicability: the diurnal and seasonal structure of the carbon signal is regular enough that most of the benefit survives even a crude, weather-agnostic forecast. A more accurate forecast would be expected to narrow the gap, although forecast accuracy and scheduling value are not the same thing: what matters is whether the forecast ranks the hours within each job's window correctly.

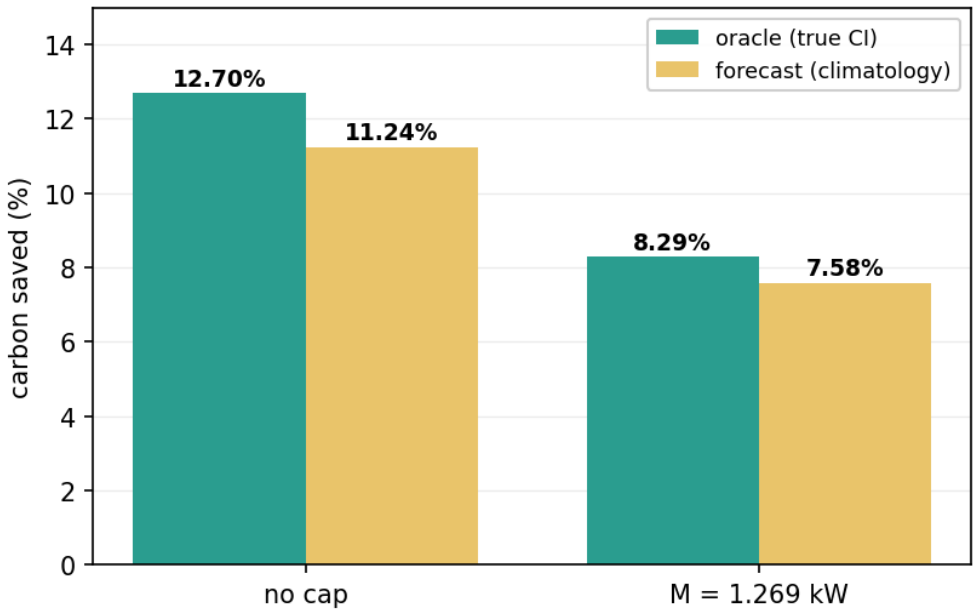


Figure 10. Carbon savings under perfect foresight of the carbon signal (oracle) versus under a weather-blind climatology forecast that predicts each hour from earlier hours sharing its hour-of-day and month, with each schedule computed on the forecast and priced on the observed signal. Uncapped: 12.70% versus 11.24%, about 88% retained. With the allocation budget $M = 1.269$ kW and the same 8363 jobs: 8.29% versus 7.58%, 91.5% retained.

4.10. Average Versus Marginal Accounting

All results above use a lifecycle-*average* carbon signal (Equation (1)), in which each kilowatt-hour is charged the mean emissions of the hour's generation. As Section 2.2 discusses, a *marginal* signal, charging each kilowatt-hour the emissions of the responding generator, is the basis aligned with the causal effect of a shifting decision, and it can produce a different schedule and a different saving. The marginal signal is a two-valued

step function over the hours of the same fuel-mix series. Writing $\mathrm{gen}_{\mathrm{OIL}}(t)$ for oil generation in hour t and τ for a generation threshold,

$$CI_{\mathrm{mar}}(t) = \begin{cases} 650\ \mathrm{gCO_2/kWh} & \text{if } \mathrm{gen}_{\mathrm{OIL}}(t) > \tau, \\ 490\ \mathrm{gCO_2/kWh} & \text{otherwise,} \end{cases} \tag{10}$$

with $\tau = 0$ in the headline configuration, so that any petroleum generation at all places oil on the margin. The two values are the oil and natural-gas factors from Table 1. The rule is the coarse merit-order approximation for ISO New England: gas combined-cycle sets the margin in the large majority of hours, while oil steam and combustion-turbine units are dispatched only under peak conditions. The rule classifies oil as marginal in 1945 of the 8760 h of 2019 (22.2%), concentrated in the summer peak (330 h in July and 293 in August against 142 in January). On this system, the proxy identifies a summer phenomenon as opposed to heating-dominated winter. The classification is consistent with the independently computed average signal: median average-basis intensity is 297.9 $\mathrm{gCO_2/kWh}$ in the hours the rule calls oil-marginal, against 251.9 in the remainder.

Unlike Equation (1), no generation weighting is applied: marginality is a property of the hour as opposed to a fuel, and passing a marginal vector through the generation-weighted mean would re-average it across nuclear, wind, hydro and solar—none of which set the margin in this system, because they are must-run or zero-marginal-cost and already fully dispatched. The resulting signal is correspondingly coarser than the average one: it spans 160 $\mathrm{gCO_2/kWh}$ against 272.7, a maximum-to-minimum ratio of 1.33 against 3.34, and correlates with the average signal at only $r = 0.375$.

Because $\tau = 0$ counts even minimal oil generation as setting the margin, we sweep it. Raising τ from 0 to 100 MWh reduces the oil-marginal hour count from 1945 to 338 and the marginal-basis saving from 4.44% to 0.69%. The marginal figure is therefore strongly dependent on the threshold, and $\tau = 0$ is its most favorable setting. The mechanism is the coarseness of a two-valued signal: at $\tau = 0$ the marginal carbon intensity is already constant across the entire feasible window of 46.1% of jobs, leaving no gradient for the scheduler to exploit, and raising τ flattens further windows. We therefore report the marginal comparison as a bounded sensitivity analysis spanning 0.69–4.44% rather than as a point estimate. Both arms use the capped greedy scheduler at the headline allocation budget, so the comparison isolates the accounting basis.

The do-nothing baseline is 1.24 $\mathrm{tCO_2}$ on the average signal and 2.46 $\mathrm{tCO_2}$ on the marginal one, so the difference of the two percentages measures how much headroom each signal exposes rather than the cost of the accounting choice. We therefore score all three schedules—the unchanged baseline, the average-optimized schedule, and the marginal-optimized schedule—against the proxy as a single reference (Table 6). An operator running the average-basis pipeline attains 2.49% of the 4.44% available on the proxy's marginal ruler for the headline building, a basis-mismatch cost of 1.94 percentage points. That cost is stable across the allocation-budget regime, at 1.95 percentage points uncapped (4.31% against 6.25%), indicating that it reflects the divergence of the two signals rather than contention for clean hours.

The estimated benefit of temporal shifting is therefore sensitive to the accounting basis, and that sensitivity is not uniform in sign. Péan et al. report larger savings on a marginal basis than on an average one for a Spanish residential heat pump, and Hawkes documents cases running both ways [7,14]. The same reversal appears within our year: aggregated over 2019 the marginal basis lowers the attainable saving for all three buildings, but in July and August it raises it, 5.44% against 6.16% and 5.81% against 6.29%, respectively the two months in which oil is on the margin most often. The sign is therefore a property of the

season as well as of the system, and should not be read as a general result. These figures are computed on the proxy's own ruler; the validation below tests whether that ruler tracks the observed marginal signal.

Table 6. Average- and marginal-basis schedules for the three Massachusetts heat-pump buildings at the headline operating point (allocation budget $M = 3\times$ that building's own median active load, six hours of flexibility, 2019 ISO New England signal). The first numeric column is the average-basis saving measured on the average signal, as reported in Section 4.7. The remaining columns are measured against a single marginal reference: what the average-optimized schedule actually attains, what a schedule optimized directly on the marginal signal attains, and the former as a percentage of the latter. Scoring each basis against its own signal would compare ratios with different denominators—the do-nothing baseline is 1.24 tCO_2 on the average signal and 2.46 tCO_2 on the marginal one—so only the common-ruler columns are mutually comparable. Ratios are computed before rounding. The marginal reference is the oil-or-gas proxy of Equation (10) at $\tau = 0$, whose threshold sensitivity is reported below.

	Average Ruler	**Proxy Marginal Ruler**		
Building	**Avg.-Opt.**	**Avg.-Opt.**	**Marginal-Opt.**	**Retained**
486202 (headline)	7.28%	2.49%	4.44%	56.2%
274807	6.89%	2.36%	4.22%	55.8%
286081	5.64%	1.89%	3.22%	58.5%

Validation against an independent marginal reference. ISO New England publishes, for 2019, hourly marginal CO_2 emission rates by fuel type, derived from the marginal units it identifies in each five-minute real-time dispatch interval, on both a load-weighted and a time-weighted basis [30]. We use the load-weighted series, $CI_{\mathrm{ISO}}(t) = \sum_f \pi_f(t)\, r_f(t)$, where $\pi_f(t)$ is the share of hour t for which fuel f was marginal and $r_f(t)$ its CO_2 rate, converted from lb/MWh to gCO_2/kWh. It is a combustion (stack) basis rather than a lifecycle one, so we use it to test the proxy's ordering of hours and to score schedules, not to test its level. The reference does not support the proxy's classification: ISO New England reports that oil was the marginal fuel about 0.1% of the time in 2019 [31], whereas Equation (10) labels 22.2% of hours oil-marginal, most likely because the EIA `OIL` category is non-zero whenever any oil unit generates, including units that are not on the margin. Correspondingly, the proxy correlates with CI_{ISO} at $r = 0.074$ (the average signal at $r = 0.189$), and mean CI_{ISO} is 344.2 gCO_2/kWh in the hours the proxy calls oil-marginal against 320.7 in the remainder.

Table 7 scores four schedules on this independent reference: the unchanged baseline, the average-optimized schedule, the proxy-optimized schedule, and a schedule optimized directly on CI_{ISO}, which shows what the same scheduler attains with perfect knowledge of the ISO New England marginal signal (without a budget, the greedy placement is exact, so the uncapped column is the true optimum). At the headline budget the average-basis pipeline attains 12.52% of that. Part of the ISO-optimized saving (15.7% of it capped, 30.0% uncapped) comes from moving load into the 2.60% of 2019 h in which only zero-rate units (pumped storage, hydro, wind, demand response and imports) are marginal; most of the rest comes from the wide hour-to-hour spread of nonzero marginal rates. For energy-limited units such as hydro and pumped storage, a zero short-run rate overstates what shifting into those hours avoids, because their energy is drawn from other hours, so the ISO-optimized column is in part an optimistic reference. What the comparison establishes is that the average-basis schedule lowers marginal emissions measured on observed data by 3.62% at the headline budget and 7.05% without one. These figures evaluate schedules against an observed marginal-unit reference; they are not estimates of causally avoided emissions, because marginal-unit identification is itself an attribution method, and a load

large enough to change dispatch would change the marginal unit. The proxy results of Table 6 are retained as a stress scenario in which oil sets the margin whenever it generates.

Table 7. Heat-pump schedules (building 486202, ±6 h, the same capped greedy scheduler as Table 6) scored on ISO New England's observed 2019 hourly marginal CO_2 rates [30]. Every entry is a saving against the same do-nothing baseline priced on the same reference, so all entries are mutually comparable. Jobs whose window reaches an hour without an ISO New England value (10, all at the year boundary) are excluded from every schedule, leaving 8353 jobs.

Schedule Optimized on	$M = 3 \times$ Median	Uncapped
Nothing (unchanged baseline)	0	0
Average signal, Equation (1)	3.62%	7.05%
Oil-or-gas proxy, Equation (10)	0.67%	1.17%
ISO New England marginal rates	28.93%	65.77%

5. Discussion

The measurements converge on a single, deliberately unglamorous conclusion: on a comparatively clean grid such as ISO New England, temporal shifting of a deferrable load saves carbon in the single-digit-percent range, and the exact ceiling depends far more on operational constraints than on the scheduling algorithm. The headline heat-pump saving is 8.29% under an allocation budget of three times its median active load and 12.70% without one, and the seasonal, budget, and forecast results all explain why the figure is where it is rather than higher.

The seasonal result is the one we regard as most informative, because it carries the mechanism Sukprasert et al. describe across from the spatial axis to the temporal one and measures it rather than asserting it [9]. Savings are largest in the mild shoulder months and smallest in the demand peaks, with January the worst month at 3.1%. The two peaks fail for different reasons. In summer the grid is uniformly carbon-intensive, and the ceiling itself is low, which is a property of the grid. In January the ceiling is the highest of any month, and the allocation budget prevents reaching it, which is a property of the operating constraint rather than the signal (Section 4.6). Flexibility is therefore worth it least when there is the most load to make flexible—in summer because the ceiling itself is lowest, and in winter because the budget leaves no room to reach a ceiling that is the year's highest. Any claim that carbon-aware shifting is a large lever must contend with both, and our results support the more modest reading of the opportunity.

The allocation-budget result carries the clearest operational message. Removing the shared budget raises the ceiling from 8.29% to 12.70%, so roughly a third of the model's uncapped saving is lost once the relocated load in any hour is bounded, and holding total draw within the observed equipment peak alone costs 0.51 percentage points. Moreover, the gap between the greedy heuristic and the exact optimum, 1.01 percentage points at the headline budget, appears *only* when the budget binds; uncapped, the greedy rule is already optimal. An operator who models flexibility but ignores the allocation budget will both overestimate the available savings and, if they then deploy a naive greedy controller, forfeit a further tenth of it to avoidable contention. This is why we present the result as a surface over slack and the allocation budget rather than as a single number.

The forecast result is encouraging. That a weather-blind climatology forecast retains about 88% of the oracle saving without a budget (11.24% against 12.70%) and 91.5% with one (7.58% against 8.29%) suggests that the diurnal and seasonal regularity of the carbon signal is strong enough for the strategy to be practicable without sophisticated forecasting infrastructure. In these experiments the forecast penalty is smaller than the loss to the allocation budget, so the larger barrier to acting on grid carbon here is not forecast quality

but the modest size of the underlying opportunity and the allocation-budget constraints that erode it.

Three caveats bound how far these numbers should be carried. First, they are computed on a lifecycle-*average* basis. The average signal is the one whose every coefficient is publicly documented, but the marginal basis is the one aligned with the causal emissions effect of a shifting decision [1], and the two can diverge; we treat the delta as a measured quantity (Section 4.10) rather than assuming the average basis is decision-relevant. On ISO New England's observed marginal rates, the average-optimized schedule does lower marginal emissions by 3.62% at the headline budget, but it attains only about an eighth of what a schedule optimized on those rates achieves (Section 4.10). Second, the saving is sensitive to the least certain emission factor: sweeping the `OTH` proxy shifts the headline between 5.57% and 7.70%. Third, the heat-pump figures come from an idealized re-timing model without thermal state or comfort bounds (Section 6.2); a thermally constrained controller would be expected to attain less. None of these overturns the average-basis finding, but all three are reasons to read the specific percentages as careful estimates on a stated basis rather than as universal constants.

6. Limitations

We separate limitations that affect the precision of the reported numbers from those that affect whether a conclusion holds at all, because the two carry different consequences for a reader weighing a result.

6.1. Limitations Affecting Numerical Precision

The batch (AI) load model relies on an assumed per-GPU power draw, since the Alibaba trace reports GPU counts and durations but not power. We sweep this parameter rather than fix it, and the reported percentage is unaffected; only the absolute tonnage inherits its uncertainty. The `OTH` emission factor of 230 gCO_2/kWh is a biomass-median proxy for a heterogeneous refuse/biomass/landfill-gas blend, and Section 4.8 shows it moves the headline HVAC saving across a roughly two-percentage-point band. The scheduler bills energy in whole-hour blocks at full power, i.e., $\lceil E_j/P_j \rceil$ hours, so a job whose energy is not a whole multiple of its hourly draw over-counts its final partial hour; both arms share this convention. For the heat-pump load, where $E_j/P_j = 1$ for every job, the terms cancel exactly, and the headline figures are unaffected. For the batch load it does not: 9.4% of jobs select a different window under exact-energy costing, moving that percentage by under 0.2 percentage points. The electric-vehicle load is evaluated in continuous time with exact energy and exact availability (Section 3.2), so the convention does not apply to it. It inflates the absolute gram totals of the batch load, which is why we do not headline them. These items affect how many significant figures a reported saving supports, not its direction.

6.2. Limitations Affecting the Validity of Conclusions

The deadline windows for the HVAC and batch loads are *assigned* rather than observed: the heat-pump job is given a symmetric $\pm$flex window around its actual run hour, and the batch job is given an assigned deadline, because neither dataset records the true latitude the operator had. A load with less real flexibility than we assign would save less. Relatedly, the heat-pump job is re-timed as a whole-hour block at metered power, with no thermal state and no comfort constraint, so delivered energy is conserved by construction. Three physical effects are therefore outside the model. First, indoor temperature: moving an hour of heating or cooling by up to six hours is feasible only if the envelope's thermal capacity keeps the dwelling inside its comfort band, which we do not check. Second, efficiency: an air-source heat pump's coefficient of performance falls as the source–sink

temperature difference grows [25], so the same heat delivered at a different hour requires a different amount of electricity. On this grid the cleanest hours cluster overnight (the mean trough is at 03:00, Section 4.1), when outdoor temperature is typically lower than in the afternoon, so in the heating season this effect would tend to reduce the saving; we have not quantified it. Third, pre-heating raises standby losses, and the recovery that follows a demand-response event consumes extra energy [32]. Quantifying these requires building thermal dynamics and explicit comfort bounds, which our formulation omits [19,33]. A controller that respects all three, such as that of Yang et al. [19], is the appropriate tool for establishing physically attainable savings; the heat-pump figures reported here are an idealized benchmark for such controllers, not a substitute for them. All results are computed for a single grid, ISO New England, which is comparatively clean and gas-heavy; a coal-and-renewables grid with a wider carbon swing would likely show larger savings, and the specific percentages here should not be transferred to other balancing authorities without re-running the pipeline. The carbon signal is computed on a lifecycle-average basis; the marginal basis, which is the one aligned with the causal effect of a shifting decision, is evaluated in Section 4.10 against ISO New England's observed marginal emission rates: on that reference the average-optimized schedule saves 3.62% at the headline budget, against 28.93% for a schedule optimized on those rates. No public dataset of New England workplace charging exists, so the transferability of Californian charging behavior to this grid is argued rather than verified: the reported percentage is algebraically independent of session energy and power, and we bound the effect of timing differences, but we cannot demonstrate that New England dwell and arrival distributions fall inside that bound. The heat-pump building is in Barnstable County, the mildest climate-zone-5A county in Massachusetts, selected for a balanced heating and cooling load rather than for maximum attainable saving. The load profile is generated on 2018 meteorology and priced against the 2019 carbon signal, a cross-year pairing that the 365-day re-date of Section 3.2 makes exact in date but not in weather year. Finally, the headline load is an electrification-forward heat-pump home, which represents only about 1.4% of the current Massachusetts housing stock; it is the relevant case for a decarbonizing future but is not representative of the average home today, and the savings should be read as characteristic of that forward-looking load class rather than of present-day residential demand in aggregate.

7. Conclusions

We have presented a transparent, lifecycle-based carbon accounting and scheduling pipeline for spatially fixed, deferrable loads under a shared allocation budget, built end-to-end from public data so that the avoided average-basis emissions can be independently verified. The pipeline derives hourly ISO New England grid carbon intensity from public EIA fuel-mix data and eight documented emission factors (two of them flagged as non-standard) and computes the exact minimum-carbon schedule for real electric-vehicle, heat-pump, and batch-computation loads.

In the 2019 ISO New England signal, hourly carbon intensity ranges from 116.7 to 389.4 gCO_2/kWh, with a median of 260.4. For the headline residential heat-pump load, temporal shifting in the idealized benchmark saves 8.29% under an allocation budget of three times its median active load (7.28% with a simple greedy rule, which forfeits 1.01 percentage points to contention created by the budget) and 12.70% with no budget. Holding total hourly draw within the building's observed peak, a physical limit that the allocation budget does not enforce, gives 12.19%. The savings are highest in the mild shoulder seasons (about 10.8% in fall and 9.3% in spring) and lowest in the demand-peak seasons (about 5.5% in winter, 6.6% in summer, and 3.1% in January) on an energy-weighted basis. The two minima have different causes: in summer the ceiling itself is lowest, because the

grid is carbon-intensive across the whole day, whereas in January the ceiling is the highest of any month and the allocation budget is what prevents reaching it. The result is robust across buildings (agreeing to within about 1.6 percentage points), depends on but is not dominated by the least-certain emission factor (5.57% to 7.70%), and in our experiments a weather-blind climatology forecast retains about 88% of the uncapped saving (11.24% against an oracle 12.70%) and 91.5% of the capped saving (7.58% against 8.29%).

The honest reading is that, on a lifecycle-average basis, temporal shifting on a clean grid is a real but modest lever, governed more by shared-budget constraints and by the seasonal anti-correlation between load and clean hours than by the scheduling algorithm. Reporting the ceiling exactly from public data, is what allows that conclusion to be stated as a measurement rather than an assertion. For the heat-pump load, these figures are an idealized benchmark: they state what re-timing alone achieves under the model, and a thermally constrained controller is needed to establish how much of it is physically attainable. On a marginal basis the lever may be larger: scored on ISO New England's observed marginal rates, a schedule optimized on those rates saves 28.93% at the headline budget, although part of that figure reflects zero-rate, energy-limited units (Section 4.10); measuring the marginal-basis ceiling on a validated signal is the natural next step.

Author Contributions: Conceptualization, R.S. and E.P.; methodology, R.S. and R.K.; software, R.S. and A.Y.; validation, R.S., A.Y. and R.K.; formal analysis, R.S. and A.Y.; investigation, R.S. and A.Y.; resources, E.P.; data curation, R.S.; writing, original draft preparation, R.S. and A.Y.; writing, review and editing, R.S., A.Y., R.K. and E.P.; visualization, R.S. and A.Y.; supervision, E.P.; project administration, R.K. and E.P.; funding acquisition, E.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All primary data are public: hourly ISO New England fuel mix from the EIA Open Data API v2 (respondent `ISNE`); electric-vehicle sessions from the Caltech ACN-Data dataset; residential heat-pump load from NREL ResStock; and batch-computation tasks from the Alibaba GPU cluster trace (v2020). The emission factors are from IPCC AR5 and the sources are tabulated in Table 1. The code that reproduces the carbon signal, the schedulers, the figures, and every reported number is available at https://github.com/anacodicAI-labs/carbon-aware-load-scheduling (last accessed on 21 August 2026). The heat-pump load profiles are included in that repository, so the heat-pump results reproduce from a clone given an EIA key alone. The carbon, electric-vehicle and batch inputs are left to their original providers rather than redistributed, so reproducing those figures additionally requires an EIA Open Data API key, credentials for the Caltech ACN-Data API, and a download of the Alibaba GPU cluster trace.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Callaway, D.S.; Fowlie, M.; McCormick, G. Location, Location, Location: The Variable Value of Renewable Energy and Demand-Side Efficiency Resources. *J. Assoc. Environ. Resour. Econ.* **2018**, *5*, 39–75. [CrossRef] [PubMed]
2. Maji, D.; Shenoy, P.; Sitaraman, R.K. 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)*; ACM: New York, NY, USA, 2022; pp. 198–207. [CrossRef]
3. Tranberg, B.; Corradi, O.; Lajoie, B.; Gibon, T.; Staffell, I.; Andresen, G.B. Real-Time Carbon Accounting Method for the European Electricity Markets. *Energy Strategy Rev.* **2019**, *26*, 100367. [CrossRef]

4. D'Ettorre, F.; Banaei, M.; Ebrahimy, R.; Pourmousavi, S.A.; Blomgren, E.M.V.; Kowalski, J.; Bohdanowicz, Z.; Łopaciuk-Gonczaryk, B.; Biele, C.; Madsen, H. Exploiting Demand-Side Flexibility: State-of-the-Art, Open Issues and Social Perspective. *Renew. Sustain. Energy Rev.* **2022**, *165*, 112605. [CrossRef]
5. Radovanović, A.; Koningstein, R.; Schneider, I.; Chen, B.; Duarte, A.; Roy, B.; Xiao, D.; Haridasan, M.; Hung, P.; Care, N.; et al. Carbon-Aware Computing for Datacenters. *IEEE Trans. Power Syst.* **2023**, *38*, 1270–1280. [CrossRef]
6. Li, J.; Wang, G.; Wang, X.; Du, Y. Smart Charging Strategy for Electric Vehicles Based on Marginal Carbon Emission Factors and Time-of-Use Price. *Sustain. Cities Soc.* **2023**, *96*, 104708. [CrossRef]
7. Péan, T.Q.; Costa-Castelló, R.; Salom, J. Price and Carbon-Based Energy Flexibility of Residential Heating and Cooling Loads Using Model Predictive Control. *Sustain. Cities Soc.* **2019**, *50*, 101579. [CrossRef]
8. Huber, J.; Lohmann, K.; Schmidt, M.; Weinhardt, C. Carbon Efficient Smart Charging Using Forecasts of Marginal Emission Factors. *J. Clean. Prod.* **2021**, *284*, 124766. [CrossRef]
9. Sukprasert, T.; Souza, A.; Bashir, N.; Irwin, D.; Shenoy, P. On the Limitations of Carbon-Aware Temporal and Spatial Workload Shifting in the Cloud. In *Proceedings of the Nineteenth European Conference on Computer Systems (EuroSys '24)*; ACM: New York, NY, USA, 2024; pp. 924–941. [CrossRef]
10. Wiesner, P.; Behnke, I.; Scheinert, D.; Gontarska, K.; Thamsen, L. Let's Wait Awhile: How Temporal Workload Shifting Can Reduce Carbon Emissions in the Cloud. In *Proceedings of the 22nd International Middleware Conference (Middleware '21)*; ACM: New York, NY, USA, 2021; pp. 260–272. [CrossRef]
11. Grange, L.; da Costa, G.; Stolf, P. Green IT Scheduling for Data Center Powered with Renewable Energy. *Future Gener. Comput. Syst.* **2018**, *86*, 99–120. [CrossRef]
12. Mahmud, A.H.; Iyengar, S.S. A Distributed Framework for Carbon and Cost Aware Geographical Job Scheduling in a Hybrid Data Center Infrastructure. In Proceedings of the 2016 IEEE International Conference on Autonomic Computing, Wuerzburg, Germany, 17–22 July 2016; pp. 75–84. [CrossRef]
13. Acun, B.; Lee, B.; Kazhamiaka, F.; Maeng, K.; Gupta, U.; Chakkaravarthy, M.; Brooks, D.; Wu, C.J. Carbon Explorer: A Holistic Framework for Designing Carbon Aware Datacenters. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '23), Volume 2*; ACM: New York, NY, USA, 2023; pp. 118–132. [CrossRef]
14. Hawkes, A.D. Estimating Marginal CO2 Emissions Rates for National Electricity Systems. *Energy Policy* **2010**, *38*, 5977–5987. [CrossRef]
15. Leerbeck, K.; Bacher, P.; Junker, R.G.; Goranović, G.; Corradi, O.; Ebrahimy, R.; Tveit, A.; Madsen, H. Short-Term Forecasting of CO_2 Emission Intensity in Power Grids by Machine Learning. *Appl. Energy* **2020**, *277*, 115527. [CrossRef]
16. Graff Zivin, J.S.; Kotchen, M.J.; Mansur, E.T. Spatial and Temporal Heterogeneity of Marginal Emissions: Implications for Electric Cars and Other Electricity-Shifting Policies. *J. Econ. Behav. Organ.* **2014**, *107*, 248–268. [CrossRef]
17. Callaway, D.S.; Hiskens, I.A. Achieving Controllability of Electric Loads. *Proc. IEEE* **2011**, *99*, 184–199. [CrossRef]
18. Kohlhepp, P.; Harb, H.; Wolisz, H.; Waczowicz, S.; Müller, D.; Hagenmeyer, V. Large-Scale Grid Integration of Residential Thermal Energy Storages as Demand-Side Flexibility Resource: A Review of International Field Studies. *Renew. Sustain. Energy Rev.* **2019**, *101*, 527–547. [CrossRef]
19. Yang, Y.; Shi, J.; Wang, D.; Wu, C.; Han, Z. Carbon-Aware Scheduling of Thermostatically Controlled Loads: A Bilevel DRCC Approach. *IEEE Trans. Smart Grid* **2025**, *16*, 1233–1247. [CrossRef]
20. U.S. Energy Information Administration. Open Data API v2: Hourly Electric Grid Monitor, Fuel-Type Generation (Fuel-Type-Data Endpoint). Respondent ISNE (ISO New England). 2024. Available online: https://www.eia.gov/opendata/ (accessed on 21 August 2026).
21. ISO New England. ISO New England: System Operator for the New England Regional Power Grid. Balancing Authority for the ISO-NE/ISNE Region. 2024. Available online: https://www.iso-ne.com/ (accessed on 15 September 2026).
22. IPCC. Climate Change 2014: Mitigation of Climate Change. In *Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*; Annex III: Technology-Specific Cost and Performance Parameters; Technical Report; Cambridge University Press: Cambridge, UK, 2014.
23. Lee, Z.J.; Li, T.; Low, S.H. ACN-Data: Analysis and Applications of an Open EV Charging Dataset. In *Proceedings of the Tenth ACM International Conference on Future Energy Systems (e-Energy '19)*; ACM: New York, NY, USA, 2019; pp. 139–149. [CrossRef]
24. National Renewable Energy Laboratory. End-Use Load Profiles for the U.S. Building Stock (Generated Using the ResStock Model), `resstock_amy2018_release_2`; Massachusetts Single-Family Heat-Pump Home, Hourly Electric-HVAC End-Use Consumption. 2024. Available online: https://data.openei.org/submissions/4520 (accessed on 15 September 2026).
25. Staffell, I.; Brett, D.; Brandon, N.; Hawkes, A. A Review of Domestic Heat Pumps. *Energy Environ. Sci.* **2012**, *5*, 9291–9306. [CrossRef]
26. Alibaba Group. Alibaba Cluster Trace Program: GPU Cluster Trace (v2020). Machine-Learning Workload Trace with Per-Task GPU Counts and Durations. 2020. Available online: https://github.com/alibaba/clusterdata (accessed on 15 September 2026).

27. Shao, S.; Harirchi, F.; Dave, D.; Gupta, A. Preemptive Scheduling of EV Charging for Providing Demand Response Services. *Sustain. Energy Grids Netw.* **2023**, *33*, 100986. [CrossRef]
28. Huangfu, Q.; Hall, J.A.J. Parallelizing the Dual Revised Simplex Method. *Math. Program. Comput.* **2018**, *10*, 119–142. [CrossRef]
29. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberl, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. Correction in *Nat. Methods* **2020**, *17*, 352. [CrossRef] [PubMed]
30. ISO New England. 2019 Real-Time Marginal CO_2 Emission Rates (Load-Weighted and Time-Weighted); Data File 20211102_2019_rt_marginal_co2_emission_rates.xlsx. ISO New England Inc., System Planning, 2021. Available online: https://www.iso-ne.com/static-assets/documents/2021/03/2019_air_emissions_report.pdf (accessed on 15 September 2026).
31. ISO New England. 2020 ISO New England Electric Generator Air Emissions Report. ISO New England Inc., System Planning, 2022. Available online: https://www.iso-ne.com/static-assets/documents/2022/05/2020_air_emissions_report.pdf (accessed on 15 September 2026).
32. Luo, Z.; Peng, J.; Cao, J.; Yin, R.; Zou, B.; Tan, Y.; Yan, J. Demand Flexibility of Residential Buildings: Definitions, Flexible Loads, and Quantification Methods. *Engineering* **2022**, *16*, 123–140. [CrossRef]
33. Oldewurtel, F.; Parisio, A.; Jones, C.N.; Gyalistras, D.; Gwerder, M.; Stauch, V.; Lehmann, B.; Morari, M. Use of Model Predictive Control and Weather Forecasts for Energy Efficient Building Climate Control. *Energy Build.* **2012**, *45*, 15–27. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.