Bridging the Temporal Gap: From Historical Monthly Invoices to Granular Hourly Energy Forecasting for Sustainable Operations

Prathamesh Pawar[1]** **Alec Hewitt**[1] **William Schuerman**[1] **Seyma Gunes**[1]

Will Sorenson[1]

[1]Amazon Web Services

Abstract

Large-scale commercial facilities face significant challenges in sustainability planning due to limited granular energy consumption data. While monthly energy data exists for all facilities in the form of utility invoices, hourly or better resolution is often available for only a subset of locations. This paper presents a machine learning framework that synthesizes low-resolution temporal data to generate high resolution hourly energy forecasts. Our approach combines monthly data from 351 facilities with hourly patterns from 175 instrumented sites using a Bayesian disaggregation model (remaining sites were used as test sites). The system achieves 1-3 month ahead hourly forecasts with a 30% reduction in MAE compared to the utility-bill baseline. This could enable applications in carbon accounting, demand response, and renewable integration planning. As commercial loads continue to grow as a percentage of total grid consumption, forecasting their unique consumption patterns becomes increasingly valuable for grid reliability and efficiency. Grid operators could adopt this methodology to improve the accuracy of their load forecasts, making the entire grid more efficient and reliable.

1 Introduction

Insight into hourly electricity usage is becoming increasingly important. It is necessary for carbon accounting (how much are my operations polluting?), demand response participation (how do I know how much load I can promise to curtail?), and renewable energy integration planning (how much will rooftop solar lower my net peak load, allowing me to save on capacity costs?). It may also be useful for Grid Operators, who already incorporate large loads in their Load Forecasts[6].

As has often been the case in recent years, 2024 was by far the hottest year since records began in 1850[1]. Industrial electrification has been seen as a key component of de-carbonization. According to the IEA [2], electricity's share of total energy consumption must rise from 20% to 27% by 2030 to stay on track for Net Zero by 2050. This means that industrial processes with individually large loads will become an increasing large percentage of electricity grids' load. For example, H2 Steel is connecting a 800MW electrolyzer to the grid in Sweden. [3]. Unfortunately, it's been delayed due in part to the Grid Operator having concerns about grid stability[4]. This could be mitigated if we could have more confidence in hourly electricity usage. Other examples include buildings serving EV charging fleets, desalination, warehouses with large HVAC systems, and indoor agriculture.

**Corresponding author: prathavp@amazon.com

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2025.

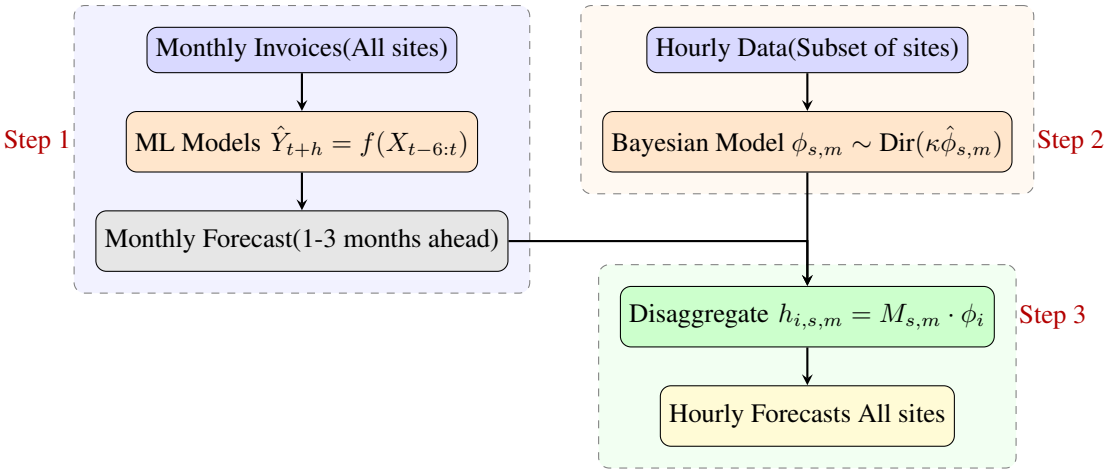

Figure 1: The proposed 3-step methodology

It's often difficult to measure high resolution electricity that a facility is using [5]. If the instrumentation technology (Smart Meters, Power Quality Monitors, and Current Transformers) are not installed when the facility is constructed, adding them after-the-fact will often require expensive outages along with permitting, utility coordination, equipment procurement, and installation hurdles. Sometimes there are technical barriers like limited panel space in electrical rooms that make adding new hardware prohibitively difficult. When installed, they may be installed at the incorrect point such that not all electricity use is measured.

In this paper, we aim to demonstrate that invoices are all you need for at least some sites. While not all industrial sites have Power Quality Monitors and Smart Meters, all electricity customers do have an invoice. We can take the invoices and leverage a small subset of accurately instrumented industrial sites to go from 2-3 month lagged invoices with a monthly granularity to forward-looking forecasts at an hourly granularity. The methods we present here perform well with limited data– 351 sites with invoices and 50% (175) sites with accurate hourly data and with loads with differing load profiles. Our key innovation is demonstrating that hourly data from sparse sites can effectively disaggregate monthly invoices for the remaining sites, solving a major roadblock for operational planning. Other research has tried this before, but then often rely on an additional level of information like time of use as their base [5]. Disaggregation has previously been used to estimate the efficiency of residential solar[7], but even here the method relies on clear diurnal and nocturnal signals which may not always be available for pure building loads.

2 Methods

We use the Electricity / ECL[8] dataset to demonstrate our technique. This has 15 minute data on 370 Portuguese Industrial Customers from 2012-2015. We removed 19 sites for data reliability issues. Figure 1 describes our method. We generate synthetic invoices by aggregating all the customers at a monthly level. We then forecast invoices 3 months forward with standard ML techniques. We use hourly data from a subset of sites to build a disaggregation model that disaggregates monthly invoice data into hourly data. We then apply the disaggregation model to the values forecasted by the invoice.

2.1 Forecasting Monthly Electricity Usage

This forecasting problem differs from many forecasting problems in the literature in that it is "small" data. We only have 37 observations (months) for each of 351 separate industrial customers. As such, we rely on methods that are either (1) efficient on small data or (2) take into account that these 351 time series are related to each other.

We use expanding window cross-validation with 24-month minimum training and 3-month forecast horizons, creating temporally-ordered splits that prevent data leakage. This approach enables robust

evaluation of multi-step forecasting performance while preventing data leakage by ensuring that all training data precedes the test period.

Table 1 shows the results from the models that were most promising given those constraints. The baseline was calculated by taking the invoices lagged `horizon` months as the forecast.

Table 1: Model Performance Comparison Across Different Horizons

Model	Horizon	MAE	RMSE	MdAPE	Model	Horizon	MAE	RMSE	MdAPE
Baseline	1	50	440	8.0	LightGBM[9]	1	69	481	9.2
	2	99	921	13.7		2	69.2	479	9.1
	3	128	1173	18.0		3	68.9	484	9.2
TabPFN2[10]	1	68	524	8.4	AutoGluon[11]	1	73	509	10
	2	65	480	8.3		2	76	607	8.9
	3	65	462	9.7		3	78	615	8.5

Notes: MAE = Mean Absolute Error (MWh per month); RMSE = Root Mean Squared Error; MdAPE = Median Absolute Percentage Error. We prefer this over Mean because meters occassionally have 0 readings, resulting in infinite MAPE. Lower values indicate better performance. Horizon refers to the months ahead the forecast is. Often the most recent invoice available reflects data that is 2 to 3 months old, so forecasting present or month-ahead usage requires forecasting 3 months out.

2.2 From Monthly Data to Hourly Data

We then disaggregate monthly electricity usage $M_{s,m}$ to hourly intervals $h_{s,m,i}$. Our preferred method is via the bayesian probabilistic model specified below, as it extends itself nicely to future work when we can add more metadata about the sites.

We evaluate four disaggregation approaches for allocating monthly consumption M across hourly intervals h_i. With only 351 sites having hourly data, a 50/50 split provides sufficient training data (175 sites) while reserving enough test sites (176) to evaluate performance and it mirrors the real-world conditions. The **Uniform** model distributes M evenly across all hours. The **Template-based** model scales historical average hourly patterns by M. The **Cluster-based** model first groups sites by monthly scale summary consumption features, then applies cluster-specific templates.

Finally, we propose an **Dirichlet–Bayesian Disaggregation Model**, which specifies Dirichlet priors over hourly share vectors and incorporates three refinements: (1) robust prior estimation via trimmed means, (2) day-of-week corrections for systematic weekday–weekend shifts, and (3) cluster-aware template mixing for site-specific adaptation. Predictions are obtained through posterior Monte Carlo sampling, yielding calibrated and uncertainty-aware load profiles. Technical details and mathematical derivations are provided in Appendix A. Figure 2 presents examples of a few industrial sites (test set). For disaggregation using baseline vs. forecasted monthly loads see Appendix B.

2.3 Applying Disaggregation to the Forecasts and Results

We then apply our model from section 2.2 to the best forecasts in section 2.1 (TabPFN). Table 2 presents our overall results for end to end process. For many use cases, the relevant Baseline is H3 Uniform Baseline as that represents only having utility bills but want to know our power usage now. Our method sees a 30% reduction in MAE, with improvements across the board.

3 Conclusion

In this work, we've developed a method for understanding electricity usage using subsets of data that many companies can acquire without comprehensive upgrades to their metering systems. This has broad applicability across a variety of industries.

Beyond individual facilities, this methodology addresses grid-scale challenges. As industrial electrification accelerates, grid operators need granular load forecasts without comprehensive metering. Our approach of requiring only invoices and sparse instrumentation democratizes hourly energy insights, enabling better demand response, renewable integration, and grid stability during the energy transition. Another exciting direction is, given that we now have these forecasts of our load profile, how can we strategically deploy Carbon Free Energy (CFE) in ways that free up capacity on the grid?

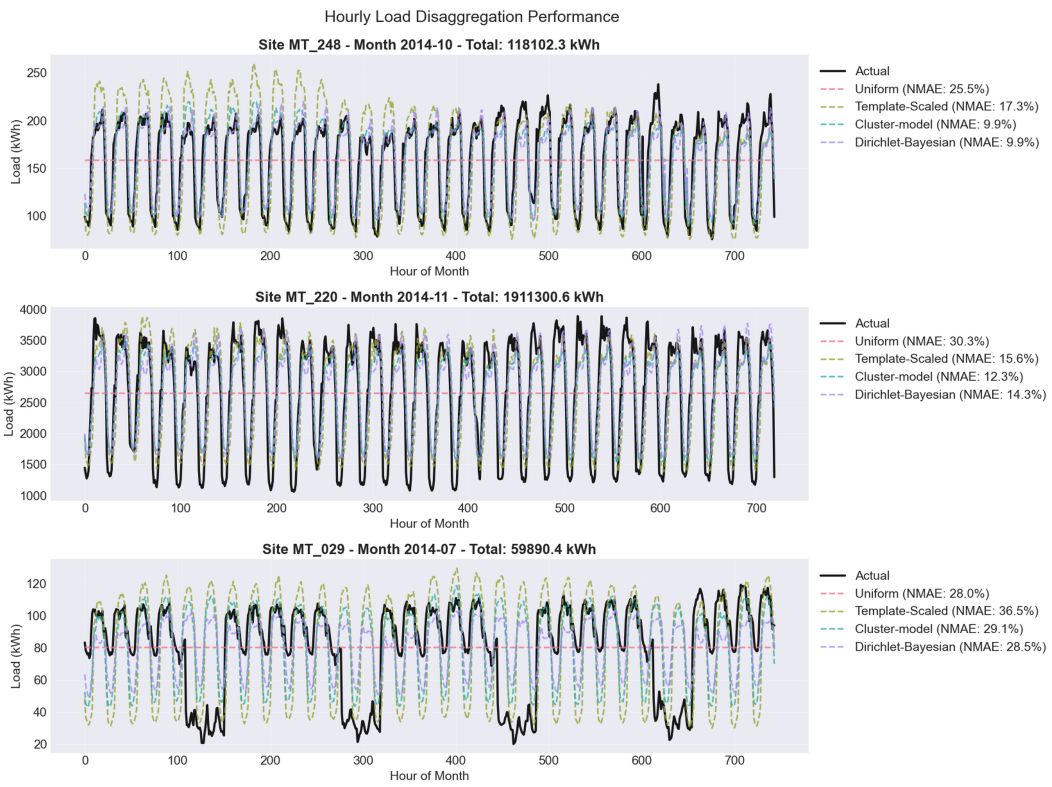

Figure 2: Hourly load disaggregation performance for three UCI commercial sites comparing actual consumption (black) against four disaggregation methods. Normalized mean absolute error (MAE) shown in parentheses, with the Dirichlet-Bayesian and cluster approaches achieving lowest errors.

Table 2: Performance comparison of disaggregation models across forecasting scenarios and horizons

Model	**Baseline**			**Forecast**			**Actual**
	H1	H2	H3	H1	H2	H3	Observed
Uniform	178.8	185.7	189.3	174.2	176.7	178.3	157.6
Template-Scaled	152.4	160.6	164.9	159.2	160.1	159.1	145.6
Cluster-Based	132.2	143.6	150.7	131.4	133.5	134.0	112.7
Dirichlet-Bayesian	**130.5**	**142.5**	**150.2**	**129.7**	**132.1**	**133.1**	**109.0**

Note: H1, H2, H3 represent 1-, 2-, and 3-month ahead forecast horizons. Mean MAE across 1232 test site-month pairs (176 x 7 rolling test windows) in kWh. Baseline uses lagged monthly totals, Forecast uses ML predicted totals, and Actual uses observed monthly totals. Bold values indicate best performance in each column. We choose MAE primarily because it is most relevant for use cases like Demand Response, Integrated Resource Planning, etc.

For example, a Steel plant might be able to better understand the viability of installing solar or BTM BESS. Companies will have a better idea of how much carbon they are emitting while they are emitting it rather than being lagged by 3 months, leading to increased demand to offset that carbon. Since our results reduce the MAE of our forecasts by 30%, even with this diverse, uninformative dataset, we believe there will be a wide variety of potential applications. We expect errors will be lower in most real-world applications, where there is both less heterogeneity in load behavior and more informative features we can use to make our forecasts.

Acknowledgments and Disclosure of Funding

The authors would like to thank the NeurIPS 2025 Tackling Climate Change with Machine Learning workshop for providing the platform to share this research. We are grateful to Amazon Science for supporting this work. This research was conducted as part of Amazon Web Services' commitment to enabling sustainable operations through data-driven approaches.

Funding: This work was supported by Amazon Web Services.

Competing Interests: All authors are employees of Amazon Web Services.

References

[1] R. Lindsey and L. Dahlman, "Climate change: global temperature," Climate.gov, May 2025. [Online]. Available: https://www.climate.gov/news-features/understanding-climate/climate-change-global-temperature

[2] International Energy Agency, "Net Zero by 2050: A Roadmap for the Global Energy Sector," IEA, Paris, Oct. 2021. [Online]. Available: https://iea.blob.core.windows.net/assets/deebef5d-0c34-4539-9d0c-10b13d840027/NetZeroby2050-ARoadmapfortheGlobalEnergySector_CORR.pdf

[3] J. Burgess, "Sweden's H2 Green Steel signs 14-TWh PPA to power planned electrolyzer," *S&P Global Commodity Insights*, Jun. 8, 2022. [Online]. Available: https://www.spglobal.com/commodityinsights/en/market-insights/latest-news/electric-power/060822-swedens-h2-green-steel-signs-14-twh-ppa-to-power-planned-electrolyzer

[4] A. Tigerstedt, "Swedish miner postpones 5 TWh green pilot scheme to 2028," *Montel News*, Oct. 1, 2024. [Online]. Available: https://montelnews.com/news/ebc381f4-dc05-4ce8-ad10-59e6cb8c9416/swedish-miner-postpones-5-twh-green-pilot-scheme-to-2028

[5] Lamagna, M., Nastasi, B., Groppi, D. et al. Hourly energy profile determination technique from monthly energy bills. Build. Simul. 13, 1235–1248 (2020). https://doi.org/10.1007/s12273-020-0698-y.

[6] PJM Interconnection, "PJM Manual 11: Energy & Ancillary Services Market Operations, Revision 129," PJM Interconnection, Feb. 22, 2024.

[7] Bu, F., Dehghanpour, K., Yuan, Y., Wang, Z., and Guo, Y. (2021). Disaggregating Customer-Level Behind-the-Meter PV Generation Using Smart Meter Data and Solar Exemplars. IEEE Transactions on Power Systems, 36(6):5417-5427. DOI: 10.1109/TPWRS.2021.3074614.

[8] Trindade, A. (2015). ElectricityLoadDiagrams20112014 [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C58C86.

[9] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). LightGBM: a highly efficient gradient boosting decision tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), pages 3149–3157, Long Beach, California, USA.

[10] Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. (2023). TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. arXiv preprint arXiv:2207.01848. https://arxiv.org/abs/2207.01848.

[11] Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., and Smola, A. (2020). AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. arXiv preprint arXiv:2003.06505. https://arxiv.org/abs/2003.06505.

A Appendix A: Enhanced Dirichlet–Bayesian Disaggregation Model

We denote the monthly total consumption for site s in month m as $M_{s,m}$, and the corresponding hourly share vector as

$$\boldsymbol{\phi}_{s,m} = (\phi_1, \ldots, \phi_{24D_m}), \quad \sum_i \phi_i = 1.$$

, where D_m is the number of days in that month.

Our generative model specifies:

$$\boldsymbol{\phi}_{s,m} \sim \text{Dirichlet}(\kappa\, \hat{\boldsymbol{\phi}}_{s,m}), \quad h_{i,s,m} = M_{s,m} \cdot \phi_i,$$

where $\hat{\boldsymbol{\phi}}_{s,m}$ is an empirical Bayes prior constructed from historical reference patterns, and κ is a concentration parameter governing dispersion.

Refinements. Three refinements are incorporated:

1. **Robust prior estimation.** The base prior $\hat{\boldsymbol{\phi}}$ is estimated using trimmed means of reference sites to mitigate sensitivity to outliers.
2. **Day-of-week corrections.** A multiplicative adjustment matrix $\mathbf{C}_{dow}$ modulates hourly shares depending on weekday vs. weekend patterns:

   $$\hat{\boldsymbol{\phi}}'_{s,m} = \hat{\boldsymbol{\phi}}_{s,m} \odot \mathbf{C}_{dow},$$

   followed by normalization.
3. **Cluster-aware mixing.** If site s belongs to cluster $c(s)$, the effective prior is a convex combination:

   $$\hat{\boldsymbol{\phi}}''_{s,m} = \lambda \hat{\boldsymbol{\phi}}_{c(s),m} + (1-\lambda)\hat{\boldsymbol{\phi}}'_{s,m},$$

   where λ is a mixing weight learned from validation data.

Posterior inference is performed via Monte Carlo sampling of $\boldsymbol{\phi}_{s,m}$, yielding uncertainty-quantified predictions $\{h_{i,s,m}\}$ across hours.

B Appendix B: Sample hourly disaggregated graphs baseline vs. forecast

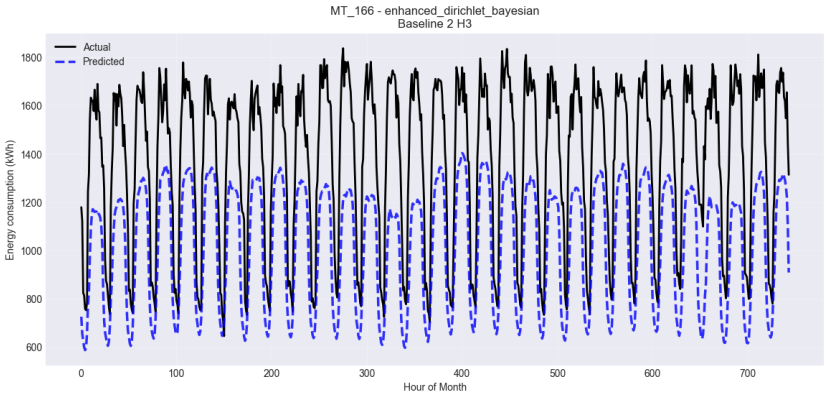

Figure 3a: Sample hourly disaggregation using the Bayesian model for a random test site MT_166. The black solid line is the actual observed hourly load for the month of May 2014. And the blue dotted line is the baseline monthly forecast (plot is for horizon H3, so the monthly baseline is the actual monthly consumption from T-3 months.

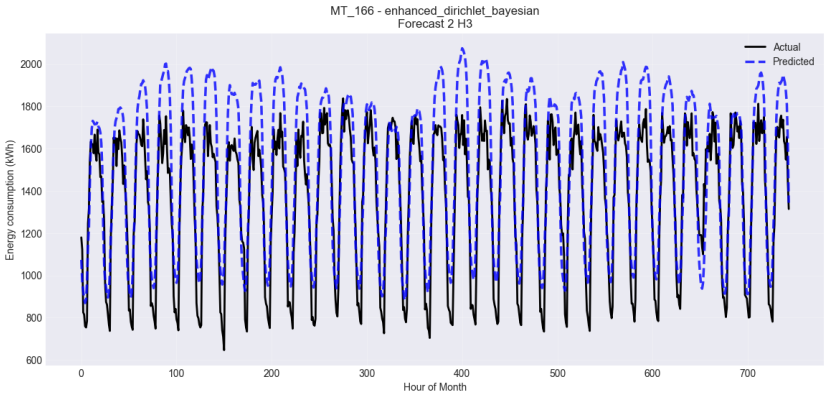

Figure 3b: Sample hourly disaggregation using the Bayesian model for a random test site MT_166. The black solid line is the actual observed hourly load for the month of May 2014. And the blue dotted line is the ML monthly forecast from section 2.1 (plot is for horizon H3, so the ML forecasting model was trained on data until T-3 months. And the monthly features that are used for the hourly disaggregation model also respect the horizon limits.