A Multi-source Data Repository and Profit-robust Framework for Energy Storage Planning

Dongjoo Kim, Subir Majumder
Department of Electrical and Computer Engineering
Texas A&M University
College Station, TX, USA
{djkim, subir.majumder}@tamu.edu

Le Xie
Harvard John A. Paulson
School of Engineering and Applied Sciences
Allston, MA, USA
xie@seas.harvard.edu

***Abstract*—This paper introduces a data-driven framework and comprehensive repository for profit-robust planning of Energy Storage Systems (ESS) in electricity markets. Utilizing nearly a decade of granular nodal price data, which we have made accessible for public use, we develop a rigorous methodology — a scenario-based optimization approach combined with clustering techniques — to uncover regions where ESS investments can maximize returns, even in the face of price volatility. By analyzing both spatial and temporal real-time price patterns, we show that profit-robust clusters often correspond to areas with abundant renewable energy resources and high demand, such as key urban and industrial zones. While the case study is based in Texas, the data repository and computational tools developed are publicly accessible, enabling broader application and further research into energy storage planning across diverse regions.**

***Index Terms*—Energy storage system (ESS), profit-robust planning, energy arbitrage, spectral clustering, storage deployment**

I. Introduction

The rapid expansion of renewable energy resources, particularly solar and wind, has reshaped electricity grids across the world, with Texas serving as a prominent example of this transformation. As shown in Fig. 1, the state's power grid has seen a dramatic increase in renewable capacity, particularly alongside the decline of coal-fired generation, with many coal plants decommissioned between 2017 and 2018 [1]. This shift has led to increased generation variability, creating significant opportunities for energy arbitrage—purchasing electricity during low-price periods and selling it during high-price intervals. Between 2020 and 2023, the deployment of energy storage systems (ESS) in Texas surged from 275 MW to over 3,500 MW [2], a trend mirrored in other regions undergoing similar transitions toward renewable dominance. However, this growth also presents challenges in determining the optimal locations and capacity for ESS, as deployment decisions must balance energy price volatility, storage costs, and grid stability. This paper addresses a critical question: *What are the key, profit-robust locations for integrating ESS into electricity grids, both in Texas and beyond?* To answer this, we propose the creation of a cross-source data hub to support profit-robust ESS planning, allowing for more informed, data-driven decisions that maximize returns from energy arbitrage while improving grid resilience in regions undergoing rapid renewable integration.

In this study, we examine Texas as a case study and develop a cross-source data repository critical for profit maximization of ESS operators engaging in diverse revenue streams from day-ahead, real-time, and Ancillary Services (AS) markets.

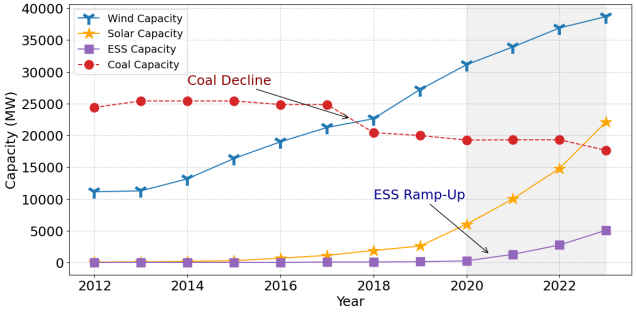


Fig. 1: Evolution of installed capacity of generating resources in Texas (2012 - 2023).

These datasets exhibit variability by location, source, and spatial resolution. For instance, day-ahead and real-time market prices are location-specific, while ERCOT's ancillary services are uniform across locations. As large-scale ESS deployments in Texas have expanded (see Fig. 1), operators have increasingly targeted dominance in the AS market, particularly in segments with lower power-to-energy requirements [3]. This strategic focus likely stems from the infrequent activation of AS [4]. However, intensified competition has led to a decline in anticipated AS profits [5]. Meanwhile, ERCOT reports rising participation in the ERCOT Contingency Reserve Service (ECRS), with recent analysis from Modo Energy showing that ECRS clearing prices have now stabilized relative to their peak in 2023 [6]. Given that real-time prices exhibit higher volatility than day-ahead prices and newer ESS installations are built for extended durations compared to earlier generations, operators are increasingly targeting the real-time market to capitalize on additional profit opportunities [7]. Considering the location-dependent nature of real-time prices, it is pertinent to identify high-profit locations for ESS operators focusing exclusively on real-time market participation.

While ESS participation in the ERCOT market has become increasingly profitable, significant potential remains for further improvement. The U.S. Department of Energy (DOE)-funded Labs are developing tools in this regard to optimize ESS sizing under varying market conditions to enhance system reliability and economic returns [8]. Academic research is also advancing by refining ESS optimization strategies for sizing and siting, focusing on minimizing operational and investment costs in complex grid models like the Western Electricity Coordinating Council (WECC) interconnection [9]. This paper contributes

to this growing body of literature by addressing two main challenges: (i) methodological issues and (ii) data accessibility. We adopt a data-driven, scenario-based approach, using historical price data to assess ESS profitability. While ERCOT provides some public data, accessing it can be cumbersome. To address this, we have curated a streamlined, publicly accessible dataset of ERCOT's real-time prices for around the past 10 years, now available on GitHub. This resource aims to remove access barriers and empower researchers to conduct more efficient analysis, fostering innovation in ESS optimization and grid economics.

We suggest the contribution of this paper as follows:

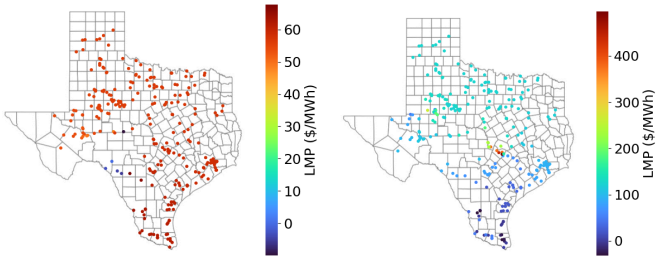
(i) We release a cross-source dataset of over 10 years of historical ERCOT nodal real-time electricity price data, curated and hosted for open access. This dataset allows researchers and stakeholders to analyze price dynamics across Texas's electricity markets with unprecedented ease.

(ii) Utilizing these datasets, we introduce an innovative optimization framework for ESS deployment in energy arbitrage. Through advanced filtering, we generate profit-robust clusters that identify optimal ESS siting locations, maximizing investment recovery by enhancing data compactness and clustering precision.

The paper is organized as follows. Section II presents a detailed overview of the data repository developed for ESS planning. Section III introduces a new filtering and Principal Component Decomposition (PCA)-based clustering techniques to generate scenarios, enabling a focused approach to ESS sizing based on spatial and temporal price dynamics. In Section IV, we perform a locational-based clustering based on the calculated profitability information to improve the robustness of optimal ESS locations. Finally, Section V concludes with key findings, implications for ESS deployment strategies, and potential directions for future research.

II. Cross-source Data Repository for Energy Storage Planning

In this section, we introduce a data repository created to support robust ESS planning, which is publicly accessible for the research community [10]. Our analysis leverages three main datasets: (i) Locational Marginal Prices (LMPs) for all ERCOT nodes at a 5-minute interval resolution, (ii) geographic locations of ERCOT nodes across Texas, and (iii) locations of existing solar, wind, and ESS installations in Texas. The first two datasets are publicly available from the ERCOT website, while the third, detailing renewable and storage sites, is accessible through the Energy Information Administration (EIA). Although ERCOT provides streaming LMP data at a 5-minute resolution, long-term access to this dataset can be a challenge. To address this, we compiled and made available over eight years of historical real-time price data from ERCOT's public archives [11], democratizing this dataset. Our repository, organized by node and year, spans from 2010 to 2022 and supports longitudinal studies at specific nodes. Each day occupies a unique column, with each 5-minute interval across 24 hours represented by one of 288 rows, enabling detailed temporal analysis. The repository also includes metadata on each node's substation and load zone, adding critical spatial context for location-based analyses.



(a) LMP at 10:35, July 25, 2022 (b) LMP at 17:35, July 25, 2022

Fig. 2: Differences in locational marginal price over time.

For the second aspect, we have extracted the locational information about the ERCOT nodes and mapped coordinates aligned to a 600-by-600-pixel representation of Texas. The map and conversion tools are provided to enable spatially accurate visualizations like that shown in Fig. 2. Lastly, the repository includes locational data for solar farms, wind power facilities, and ESS sites in Texas, with coordinates to support precise spatial analysis and mapping. The availability of such a large-scale cross-source data repository is definitely a first in the electricity market research.

III. Scenario-based Optimization for Maximizing Energy Storage Profits

As discussed earlier, determining the optimal location of ESS critically depends on generating scenarios or the availability of accurate forecasts. Clustering techniques, particularly those utilizing cluster means or centroids, can be employed to identify representative scenarios that could be used to determine optimal ESS choice. By categorizing actual data points into these clusters, one can determine the most likely scenarios the system might encounter. In this section, we first describe the representative clusters, and next, we provide a methodology to perform scenario-based optimization to calculate the profit.

A. Scenario Reduction through Clustering

Electricity price data presents inherent challenges for clustering due to its high volatility, seasonality, and dependence on multiple external factors. Unlike stationary data, electricity prices fluctuate sharply based on demand, generation constraints, fuel costs, weather conditions, and grid reliability, often introducing significant noise and irregularities that complicate the clustering process. We first utilized a backward scenario reduction approach [12], clustering data points with a classical centroid-based distance metric and assessing performance via silhouette scores. However, we observed that this method failed to accurately capture similarities within time-series data, probably due to its non-stationary nature – highlighting the need for a more effective strategy. By applying PCA before clustering, we can reduce dimensionality significantly, enhance pattern recognition, and improve the accuracy of clustering, thereby overcoming the challenges faced with backward scenario reduction. As a part of methodological innovation, we have also tested applying a low pass filter to remove the spiky components in the historical data, applying Inverse Quantile Transformation (IQT) to remove

heteroskedasticity, and subsequently removing the seasonality within the dataset by applying normalization. As shown in Fig. 3, we found that integrating IQT with filtering before applying PCA enhances dimensionality significantly, allowing for a compact representation of the data.

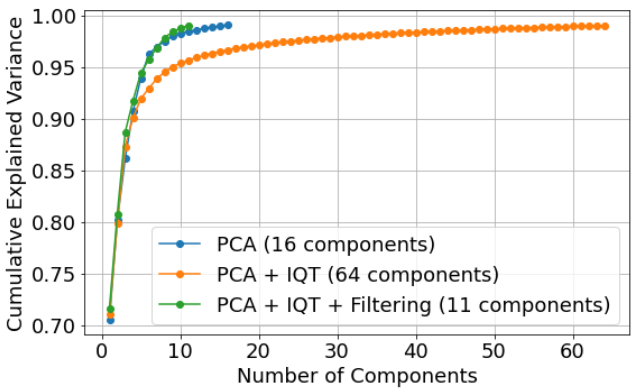


Fig. 3: Comparative analysis of PCA transformations on explained variance retention.

We performed the impacts of varying degrees of data preprocessing and found that they significantly impact the compactness of the dataset within a cluster. We observe that the dataset is more condensed after applying inverse quantile transform, with tighter clustering. With the application of filters, we observe minimal outliers with well-defined clusters. This progression demonstrated the necessity of aggressive preprocessing to generate compact clusters, which likely addresses challenges imposed by the backward reduction method. If we consider filtered transformed data, we see that the data have similar patterns within a given cluster– this impact is visible from Fig. 4. The filtered data reveals slow-varying trends by smoothing out volatility, improving interpretability and suitability for clustering while reducing the impact of abrupt price changes.

However, the original real-time prices are retained to generate scenarios based on the median of all data within a cluster. This approach enables accurate scenario reduction without compromising the integrity of the optimization process.

B. Scenario-based Optimization

Assuming the storage operators maximizes profit by only participating in the real-time market, we solve the planning problem based on earlier generated scenarios [13]:

$$\max \sum_{\forall\omega,\forall i} \mathbb{P}_\omega k_h \left(c_{i,\omega} \left(P_{i,\omega}^{B,-} - P_{i,\omega}^{B,+} \right) - \mathcal{C} \left(P_{i,\omega}^{B,-} + P_{i,\omega}^{B,+} \right) \right) \tag{1}$$

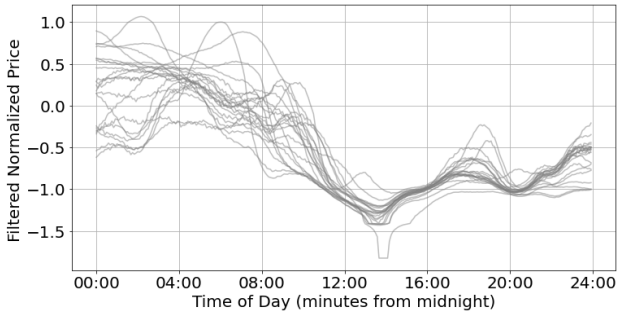
subject to,

$$\mathbb{B}_{i+1,\omega} = \mathbb{B}_{i,\omega} + k_h \left(P_{i,\omega}^{B,+} \eta_{ch} z_{i,\omega} - P_{i,\omega}^{B,-} (1 - z_{i,\omega}) / \eta_{dch} \right);$$

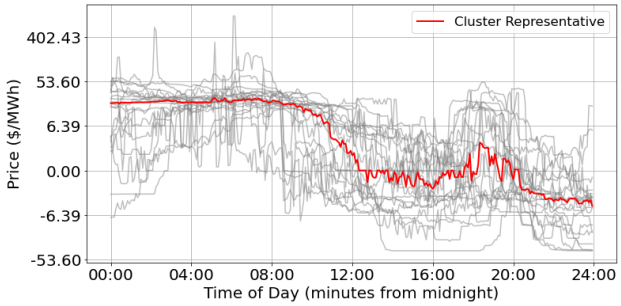
$$\mathbb{B}_{0,\omega} = \mathbb{B}_{N,\omega}; \quad \forall i, \omega \tag{2}$$

$$P_{i,\omega}^{B,+}, P_{i,\omega}^{B,-} \geq 0 \tag{3}$$

$$P_{i,\omega}^{B,+} - P_{i,\omega}^{B,-} \leq P_{MP}, \quad \mathbb{B}_{i,\omega} \leq C_{MP} \tag{4}$$



(a) With filter.



(b) Without filter to generate the cluster representative.

Fig. 4: Filters and transformations assist in difficult-to-cluster price data.

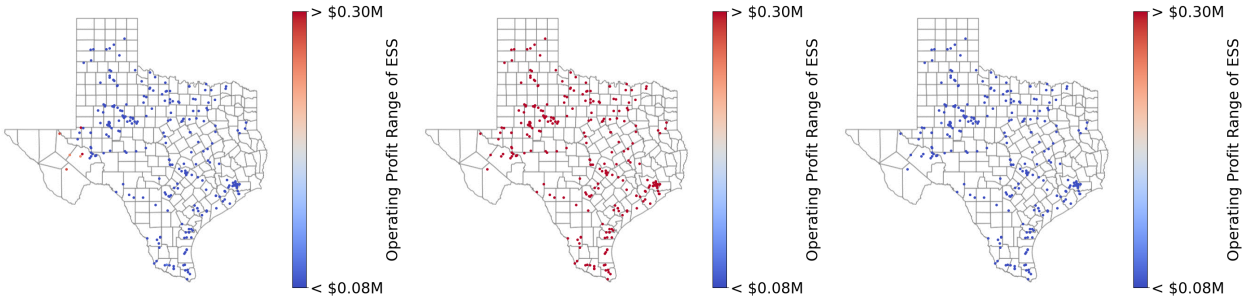
TABLE I: Notations

$c_{i,\omega}$: Cluster representative of real-time locational marginal prices within a day i for scenario ω.
$\mathcal{C}$: Capacity degradation per MWh of cycling.
$P_{i,\omega}^{B,-}$, $P_{i,\omega}^{B,+}$: Power extracted from/injected into battery pack.
$z_{i,\omega}$: Variable driving charging and discharging of BESS.
$\mathbb{B}_{i,\omega}$: Energy stored within the BESS.
$\mathbb{P}_\omega$: Probability of scenario ω.
η_{ch}, η_{dch}: Charging and discharging efficiency.
P_{MP}, C_{MP} : Megapack power and energy rating.
k_h: Hour equivalent for charging and discharging intervals.

The objective function (1) comprises two main components: (i) revenue generated from the real-time energy market, and (ii) capacity degradation costs due to the cycling of storage devices. Equation (2) ensures energy balance across multiple time steps, maintaining constant energy levels at the beginning and end of the storage cycle. Equations (3) and (4) restrict the power output of storage devices to the converter rating and the energy stored within the battery to its storage capacity. This optimization problem was solved for multiple resource nodes in ERCOT, using clustered data from 2020 to 2022 — especially because of the rapid integration of ESS into the Texas grid as shown in Fig. 1. The transitional operating profit of ESS operators is provided in Fig. 5. First, to compare the profitability, we first annualize it to 2023 dollars by accounting for inflation. Second, to prevent extreme values from disproportionately influencing the visualization, we normalize these profitability values by applying a sigmoid transformation:

$$\mathrm{Pr}_i' = \frac{1}{1 + e^{-\alpha\left(\frac{\mathrm{Pr}_i - \mathrm{lower}}{\mathrm{upper} - \mathrm{lower}} - 0.5\right)}} \tag{5}$$

where 'lower' and 'upper' are the bounds of the profitability



(a) 2020 (b) 2021 (c) 2022

Fig. 5: Operational profit transition of ESS (2020 - 2022).

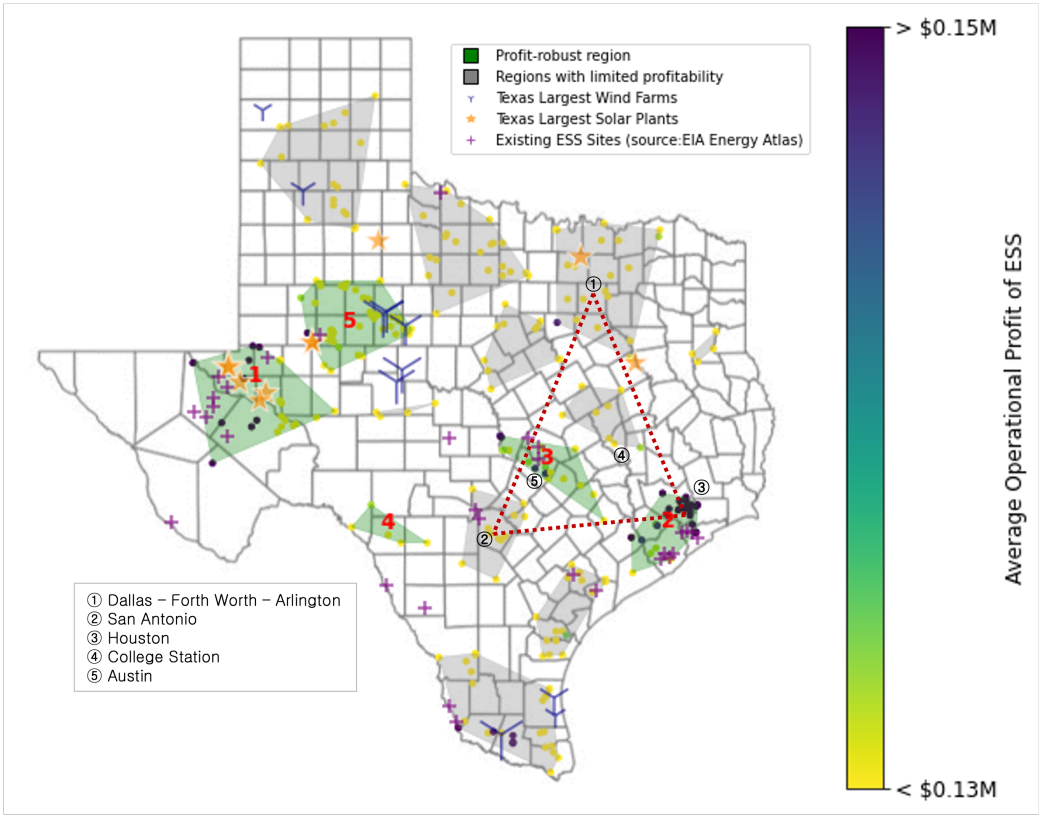


Fig. 6: Correlation of existing ESS locations in Texas with profit-robust clusters, solar and wind sites, and load centers.

range, and α controls the steepness of the transformation. Please note that this transformation is done only for visualization purposes.

IV. Identifying Profit-Robust Locations for ESS

This section explores the identification of *robust* locations for ESS by evaluating high-resolution locational profitability across Texas.

A. BIRCH Clustering Process

To categorize nodes by considering locational profitability impacts, we employed the Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) method [14], which identifies *robust* clusters by minimizing intra-cluster variance. First, we calculate the average profitability across 2020-2022, $\overline{\text{Pr}_i}$ for each node i. We, then, assign coordinates to each profitability score (Pr_i) based on their locational information x_i. Our dataset becomes of the form $\chi_i = [x_i, \overline{\text{Pr}_i}]$. We aim to partition the set of nodes $X = \{\chi_1, \chi_2, \ldots, \chi_n\}$ into ℓ clusters $C = \{C_1, C_2, \ldots, C_\ell\}$ such that nodes within each cluster C_j exhibit similar profitability characteristics. The BIRCH algorithm constructs a Clustering Feature (CF) tree to identify clusters, with each subcluster represented by a CF vector:

$$\text{CF} = (N, \mathbf{LS}, \mathbf{SS}) \tag{6}$$

where N is the number of nodes in the subcluster, $\mathbf{LS} = \sum_{i=1}^{N} \chi_i$ is the linear sum, and $\mathbf{SS} = \sum_{i=1}^{N} \chi_i^2$ is the squared sum of the nodes. Once the clusters are formed, we calculate

the average profitability $\hat{\text{Pr}}_j$ for each cluster C_j to assess their potential for ESS deployment:

$$\hat{\text{Pr}}_j = \frac{1}{|C_j|} \sum_{\chi_i \in C_j} \overline{\text{Pr}_i} \tag{7}$$

where $|C_j|$ is the number of nodes in cluster C_j. Clusters with higher $\hat{\text{Pr}}_j$ values are identified as profit-robust regions, signaling optimal locations for energy storage investment.

B. Profit-robust Clustering Results

The clusters with average annual profitability based on profitability between 2020 and 2022–as demonstrated in Fig. 5–are shown in Fig. 6. Since the clusters include neighboring points of similar profitability, the nodes within a cluster provide robustness. Here, we have marked robust, highly profitable clusters in green – if future prices exhibit similar statistical behavior. Gray regions indicate areas with limited profitability. To understand what factors affect cluster profitability, we have highlighted Texas's largest wind and solar farms, represented by blue wind farm symbols and orange stars. We have included the Texas Triangle, a densely populated region that includes the state's five largest cities—Dallas, Fort Worth, San Antonio, Houston, College Station, and Austin. We have also marked the existing ESS sites across the state of Texas. First, we find that two of the identified profit-robust clusters align with these renewable energy installations. Second, the Texas Triangle houses two profit-robust clusters. This area's diurnal demand fluctuations may drive locational price variability, making ESS deployments particularly profitable in response to residential and commercial energy needs. Third, the existing ESS sites are already positioned within or near profit-robust clusters.

This spatial alignment suggests strategic siting decisions by operators who prioritize proximity to renewable sites and high-volatility load centers to capitalize on real-time market revenues. While AS remains a significant revenue source, its profitability is generally location-agnostic. Therefore, real-time market participation in high-volatility areas makes these locations especially attractive, even as AS margins fluctuate.

V. Conclusions

This research introduces a cross-source data repository and a robust framework for ESS planning. A key innovation of this study is the cross-source data hub, which consolidates and integrates multiple data streams to inform ESS planning decisions effectively. This repository contains almost a decade of ERCOT nodal real-time price data and provides a detailed examination of locational price patterns, facilitating the identification of profit-robust regions for ESS deployment across varying temporal and spatial resolutions. Secondly, by utilizing scenario reduction through clustering techniques, optimization frameworks, and clustering to ensure robustness, this study provides methodological advancements – making it generalizable well across diverse energy markets. Our findings indicate that profit-robust clusters are not only strategically aligned with renewable generation sites and high-demand load centers but also highlight the economic significance of locational price variability on ESS profitability.

This cross-source, data-driven framework offers actionable insights for ESS operators and policymakers, enabling data-informed decisions for optimized ESS siting. Future research could explore integrating market behavior data, applying this cross-source framework to diverse energy markets. Additionally, investigating ESS deployment within distributed energy resource networks and expanding the data hub for cross-regional analysis would further validate and broaden the framework's applicability. These directions will deepen insights into ESS profitability and resilience across varied grid regions.

References

[1] U.S. Energy Information Administration, "Existing Nameplate and Net Summer Capacity by Energy Source, Producer Type and State," 2023, accessed: 2024-11-06. [Online]. Available: https://www.eia.gov/electricity/data/state/existcapacity_annual.xls

[2] S&P Global Commodity Insights, "Battery Storage Expected to Make Up 21% of US Capacity Additions in 2024," 2023, accessed: 2024-06-22. [Online]. Available: https://www.spglobal.com/commodityinsights/en/market-insights/latest-news/electric-power/122923-battery-storage-expected-to-make-up-21-of-us-capacity-additions-in-2024.

[3] Potomac Economics, "2023 State of the Market Report for the ERCOT Electricity Markets," Tech. Rep., May 2024, accessed: 2024-11-07. [Online]. Available: https://www.potomaceconomics.com/wp-content/uploads/2024/05/2023-State-of-the-Market-Report_Final.pdf

[4] ERCOT, "Ancillary services study final white paper," Tech. Rep., October 2024, accessed: 2024-11-07. [Online]. Available: https://www.ercot.com/files/docs/2024/10/07/ERCOT-Ancillary-Services-Study-Final-White-Paper.pdf

[5] Modo Energy, "ERCOT Battery Energy Storage Systems Annual Revenues 2023: BESS Index, Ancillary Services, Arbitrage, and ECRS," 2023, [Accessed: 2024-11-07]. [Online]. Available: https://modoenergy.com/research/ercot-battery-energy-storage-systems-annual-revenues-2023-bess-index-ancillary-services-arbitrage-ecrs

[6] ——, "ERCOT: How has the ECRS market evolved since its launch," 2024, [Accessed: 2024-11-07]. [Online]. Available: https://modoenergy.com/research/ercot-battery-energy-storage-systems-ecrs-ancillary-services-2024-revenues-capacity-allocation-strategy-cycles-saturation

[7] ——, "ERCOT Battery Energy Storage Revenues H1 January–June 2024: Key Insights on Owners, Ancillary Services, Arbitrage, and the BESS Index," 2024, [Accessed: 2024-11-07]. [Online]. Available: https://modoenergy.com/research/ercot-battery-energy-storage-revenues-h1-january-june-2024-owners-ancillary-services-arbitrage-bess-index-jupiter-power-key-capture

[8] National Renewable Energy Laboratory, "EMIS: Electricity Markets Investment Suite Model," Tech. Rep., 2024, [Accessed: 2024-11-07]. [Online]. Available: https://www.nrel.gov/grid/emis.html

[9] R. Fernández-Blanco, Y. Dvorkin, B. Xu, Y. Wang, and D. S. Kirschen, "Optimal energy storage siting and sizing: A wecc case study," *IEEE Transactions on Sustainable Energy*, vol. 8, no. 2, pp. 733–743, 2017.

[10] D. Kim, S. Majumder, and L. Xie, "A Cross-Source Data Repository for Energy Storage Planning." [Online]. Available: https://github.com/tamu-edu/tx-ess-planning

[11] ERCOT, "ERCOT Data Access Portal." [Online]. Available: https://www.ercot.com/mp/data-products/data-product-details?id=NP6-788-CD

[12] N. M. M. Razali and A. H. Hashim, "Backward reduction application for minimizing wind power scenarios in stochastic programming," in *2010 4th International Power Engineering and Optimization Conference (PEOCO)*, 2010, pp. 430–434.

[13] J. Faraji, A. Ketabi, and H. Hashemi-Dezaki, "Optimization of the scheduling and operation of prosumers considering the loss of life costs of battery storage systems," *Journal of energy storage*, vol. 31, p. 101655, 2020.

[14] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: an efficient data clustering method for very large databases," in *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '96. New York, NY, USA: Association for Computing Machinery, 1996, p. 103–114.