Forecasting Large Loads in the Age of AI and Data Centers

December 2025

Energy+Environmental Economics

Authors and Acknowledgements

Project Team

Energy and Environmental Economics, Inc. (E3) is a leading economic consultancy focused on the power sector in North America. For over 30 years, E3's data-driven analysis and unbiased recommendations have been utilized across the power industry by the utilities, regulators, government agencies, project developers, investors, and non-profit entities. E3 has offices in San Francisco, Boston, New York, Denver, and Calgary.

E3 Study Team

Stephen Bessasparis

Shana Ramirez

Emily Zhao

Ye Zheng

Kush Patel

Isabelle Riu

Table of Contents

Executive Summary

The U.S. electric power sector is currently facing the prospect of explosive demand growth driven by large, energy-intensive loads. Emerging industries, such as artificial intelligence (AI) data centers, cryptocurrency operations, and advanced manufacturing, are reshaping the scale and location of electricity demand in ways that traditional forecasting methods cannot capture. This has led to a patchwork of forecasting approaches across the U.S., making it challenging to align on best practices.

This whitepaper outlines a framework to forecast and manage large load growth in this evolving landscape. Effective forecasting is not about predicting a single outcome but about outlining a credible, data-driven range of possibilities that supports flexible investment, proactive risk management, and informed regulatory decisions.

E3's analysis finds that while AI and data center electricity demand is accelerating, its pace, scale and ultimate durability remain uncertain. Overestimating growth can lead to overbuilding and stranded costs; while underestimating it can cause reliability shortfalls and congestion. A disciplined approach built on verified baselines, diverse scenarios, adaptive mechanisms, and continuous performance feedback is essential, especially as large and long-term planning and investment decisions are being made based on these forecasts. Moreover, forecasting should be integrated with demand response (DR) and rate design to effectively manage risk. When forecasting is treated as a dynamic system of learning rather than a static task, it can enhance resilience, protect customers, and enable utilities and system operators to adapt confidently to rapid changes.

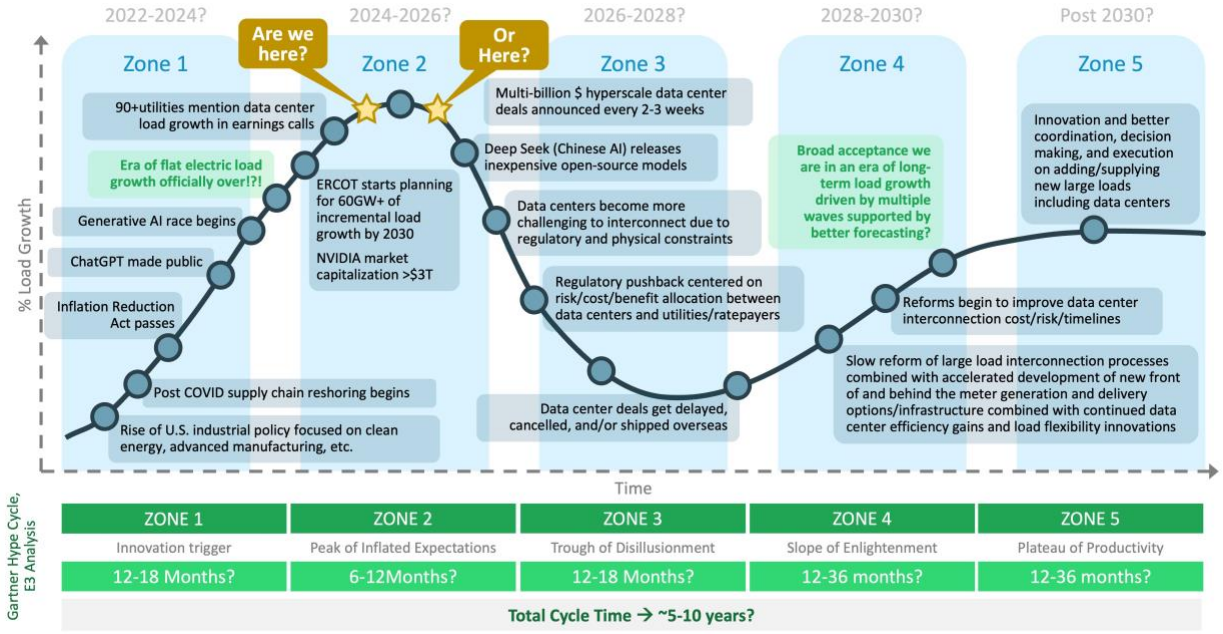
Introduction

The U.S. power sector is entering a new era of forecasting complexity. Emerging industries and the electrification of buildings and transportation are reshaping electricity demand at a pace unseen since the post-war industrial boom. Unlike historical growth tied to population and economic activity, today's drivers are concentrated, volatile, and technology-based.

These loads can appear rapidly, cluster geographically, and evolve faster than regulatory or investment processes can adapt. For utilities, this creates a planning challenge: over-forecasting risks over-building and higher costs for customers, while under-forecasting threatens reliability and missed economic opportunities. The objective is not to predict a single outcome but to delineate a credible range of futures that guide prudent and flexible planning and investment decisions.

E3's July 2024 whitepaper *"Load Growth Is Here to Stay, but Are Data Centers?"* [a] found that while load growth is accelerating, it remains highly uncertain and uneven across regions. Reliable forecasting starts with a clear baseline, yet even this is difficult without a comprehensive national inventory of existing and planned data centers. Furthermore, E3 assessed that data center load may fluctuate significantly over the next decade as the AI industry matures beyond its initial investment boom.

[a] *Riu, Isabelle, Dieter Smiley, Stephen Bessasparis, and Kush Patel. "Load Growth Is Here to Stay, but Are Data Centers?" E3, 17 July 2024, https://www.ethree.com/data-center-load-growth/.*



This paper builds on E3's findings and outlines key considerations for forecasting in this new and evolving landscape, with a particular focus on data centers, by strengthening baselines, applying scenario-based modeling, and leveraging demand response and rate design to enhance flexibility. Together, these practices create a resilient framework for forecasting amid rapid technological change and uncertainty.

Challenges in Forecasting Data Center Loads

Current utility forecasting approaches are individualized to each utility, making it difficult from the outside to compare, contrast, and aggregate forecasts. When combined with the size of individual large loads, this can result in multi-GW swings in utility forecasts that complicate decision-making.

Georgia Power serves as an example of these load oscillations. Their expected large load additions decreased by 6 GW between Q2 and Q3 of 2025 due to projects exiting the pipeline; their near-term look of projects energizing between Winter 2028 and Winter 2029 alone dropped 1.4 GW.[b] In filed testimony, Georgia Power warned that customer rates could increase if new generators finish procurement but the loads they intend to serve do not materialize, with staff characterizing data centers as "primarily underperforming expectations due to a mixture of lower materialization rates, project cancellations, and delays."[c]

AEP Ohio showed even more dramatic losses, possibly in response to new regulatory changes. In a September 2025 filing, AEP Ohio stated its total volume of interconnection requests fell from 30 GW to 13 GW.[d] This drop followed a new data center tariff approved by the Public Utilities Commission of Ohio that requires data centers to pay for 85%

[b] *DiGangi, Diana. "Georgia Power's Large Load Pipeline Shrinks by 6 GW." Utility Dive, 24 Nov. 2025, https://www.utilitydive.com/news/georgia-power-large-load-data-centers/806300/.*

[c] *DiGangi, "Georgia Power's Large Load Pipeline."*

[d] *Skidmore, Zachary. "AEP Ohio Slashes Data Center Pipeline by More Than Half – Report." Data Center Dynamics, 1 Oct. 2025, https://www.datacenterdynamics.com/en/news/aep-ohio-slashes-data-center-pipeline-by-more-than-half-report/.*

of their requested energy even if the electricity is not needed at that time, over a 12-year period. This illustrates the potential for expected loads to suddenly disappear, posing significant risks to utilities and ratepayers.

Even when data center loads materialize, their actual energy consumption can underperform expectations. Many planning assumptions for data center load factors use values upwards of 90% of the expected interconnection request. However, E3 analysis of real-world load telemetry from data centers in a major utility territory indicates that real load factors can be significantly lower. Only two data centers had load factors above 90% of their reported peak loads, and fewer than half had load factors above 80%. Similarly, PG&E has published analysis of their data center load indicating that, on average, data center peak loads are only about 67% of their "nameplate load." [e] Small changes in these planning assumptions can have large impacts on resource procurement and transmission construction plans.

Historical Forecasting Errors and their Consequences

Forecasting has always been an exercise in managing uncertainty. Even the most advanced models rely on assumptions about future behavior, and when those assumptions fail, the consequences can be severe. The greatest forecasting errors in the energy sector have rarely been technical; they have arisen from misplaced confidence in the permanence of past trends. When forecasters mistake momentum for inevitability, systemic failures often follow.

This dynamic was clearly seen in the U.S. electricity sector of the 1970s. For two decades, demand grew in step with industrial and economic expansion, leading utilities and regulators to assume the trend would continue indefinitely. Forecasts based on this assumption justified rapid investment in nuclear and fossil generation. As O'Neill and Desai (2003) observed, "assumption drag" prevented models from recognizing signs of slower GDP growth, industrial restructuring, and rising energy prices.[f] When energy costs quadrupled and conservation efforts took hold, electricity demand growth slowed sharply. Utilities, having already committed billions to new capacity, were left with stranded assets and massive overcapacity. Moreover, the Washington Public Power Supply System default became the second-largest municipal bond failure in U.S. history.[g]

The failure was not only analytical but also cultural. Forecasters assumed that industrial expansion, population growth, and electrification were immutable forces. Demand was treated as largely price-inelastic, and each small assumption error compounded over time, creating an illusion of inevitability.

Today's surge in AI, data centers, and digital infrastructure presents similar risks. Utilities are projecting record load growth and investing heavily in new generation and transmission, yet the longevity of this demand remains uncertain. Advances in chip efficiency, cooling technologies, and computing optimization could sharply reduce energy intensity. A market slowdown or correction in AI investment could flatten demand just as abruptly as in the 1970s. Both eras share the same institutional tendencies: belief in uninterrupted progress and confidence that expansion represents a permanent transformation rather than a temporary surge. Embedding such optimism in long-lived infrastructure creates risks of stranded costs, financial strain, and public mistrust.

---

[e] *California Energy Commission. "Data Center Load Forecasts, 2024-2040". 21 Oct. 2024, https://www.energy.ca.gov/filebrowser/download/6686?fid=6686.*

[f] *O'Neill, B.C. and Desai, M. The Accuracy of Past Projections of U.S. Energy Consumption. IIASA, 2003, IIASA Interim Report IR-03-053, http://pure.iiasa.ac.at/7030/*

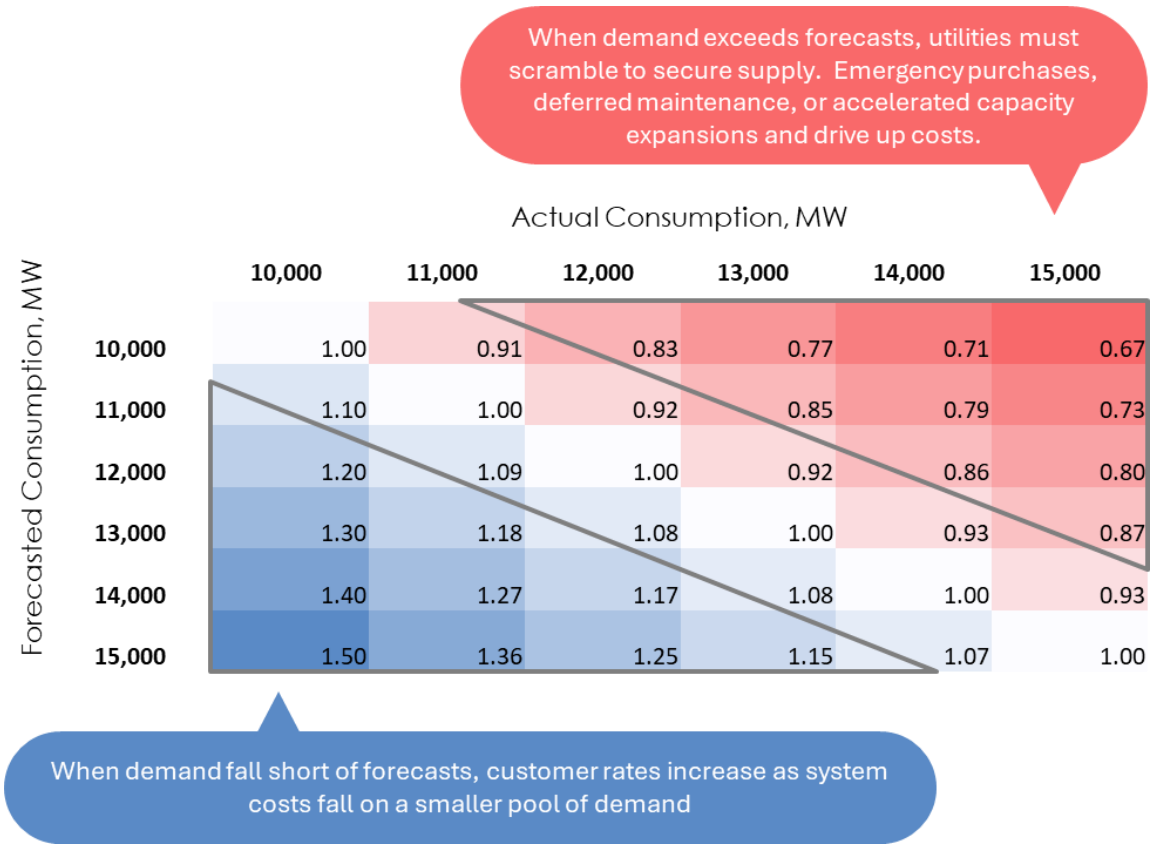
[g] *Bezdek, Roger. "Washington's Power Supply Collapse." Nature, vol. 317, 1985, pp. 309–313. https://doi.org/10.1038/317309a0.*

Forecasting errors carry significant financial and operational consequences. Over-forecasting leads to overcapacity and higher retail rates as fixed costs are spread over fewer kilowatt-hours, while under-forecasting results in shortages, emergency purchases, and reliability challenges. Either outcome undermines system efficiency and public confidence.

$$\text{Retail Rates} \propto \frac{\text{Fixed Costs} + \text{Variable Costs}}{\text{System Consumption (MWh)}}$$

When consumption falls short of projections, fixed costs remain constant while the denominator shrinks, driving rates upward even when total costs do not increase. This was the legacy of the 1970s overbuild, when customers paid for idle plants while regulatory trust eroded. The same pattern could reemerge if today's data center boom fails to sustain long-term growth. Large-scale projects with multi-decade recovery periods could become financial liabilities within years, leaving ratepayers to absorb the stranded costs.

Underestimating demand carries its own risks. When load exceeds expectations, utilities must rely on emergency procurement, deferred maintenance, and peaking resources, all of which are costly and emissions-intensive. Persistent under-forecasting can also deter industrial development by signaling limited system capacity. In rapidly growing regions dominated by data center construction, utilities face both short-term strain and long-term overbuild risk. Rapid development can stress existing infrastructure before new capacity is available, while long-term overconfidence can leave excess capacity once growth levels off. This dual exposure is among the most expensive outcomes of forecasting error.



Forecasting mistakes ripple beyond balance sheets and can include:

- Diverted investment: Capital locked in unused infrastructure limits resources for grid modernization, distributed energy, and resilience.
- Delayed innovation: Under-forecasting forces utilities into crisis management rather than strategic calculation.
- Eroded trust: Chronic inaccuracy undermines credibility with regulators, investors, and the public.

These consequences underscore the broader challenge of traditional forecasting methods in today's uncertain and rapidly evolving landscape. Instead of treating forecasts as fixed predictions, planners must continuously incorporate new data, revisit assumptions and maintain flexibility, especially when long-term commitments carry high risks. The overconfidence of the 1970s offers a powerful warning for today's AI-driven expansion, and the goal is not to abandon forecasting but to prepare for a range of futures, while safeguarding against the costs of misplaced optimism.

Utility Forecasting Methodology Examples

Utilities are adopting varied approaches to data center load forecasting, reflecting different treatments of uncertainty. Four examples: Georgia Power (GP), Arizona Public Service (APS), American Electric Power Ohio (AEP), and Dominion Energy illustrate contrasting methodologies in response to rapid data center–driven load growth.

Georgia Power uses a Load Realization Model (LRM) built on Monte Carlo simulations to generate a probability distribution of potential large load outcomes. Each interconnection project receives probabilistic weights for factors, such as location selection, provider choice, timing, and load magnitude. Thousands of iterations yield a P50 (median) forecast adjustment to GP's traditional regression-based forecast. This explicit modeling of uncertainty parallels E3's treatment of attrition, completion windows, and ramp rates within its scenario framework.

Arizona Public Service employs a customer-specific approach for Extra High Load Factor customers, developing individual forecasts based on interconnection queue data and contractual milestones, such as Letters of Agreement and Energy Service Agreements. APS benchmarks ramp-up assumptions against peer utilities and applies discount factors to account for potential variability.

Dominion Energy integrates scenario-based and multi-model forecasting into its resource planning. Customer segments are modeled under high- and low-growth cases using historical data, customer intelligence, and statistical fits. Outputs are blended into a moderate view and enhanced with spatial forecasting to identify where within the service territory load growth is likely to occur, supporting transmission and generation planning.

AEP Ohio applies a queue-based forecast similar to APS, including only loads with executed Letters of Agreement and Electric Service Agreements to ensure forecasts reflect committed rather than speculative demand.

These utilities represent three main forecasting approaches: probabilistic modeling that quantifies uncertainty through simulations, scenario-based modeling that explores a range of possible growth outcomes, and contractual filtering that includes only committed projects to manage risk.

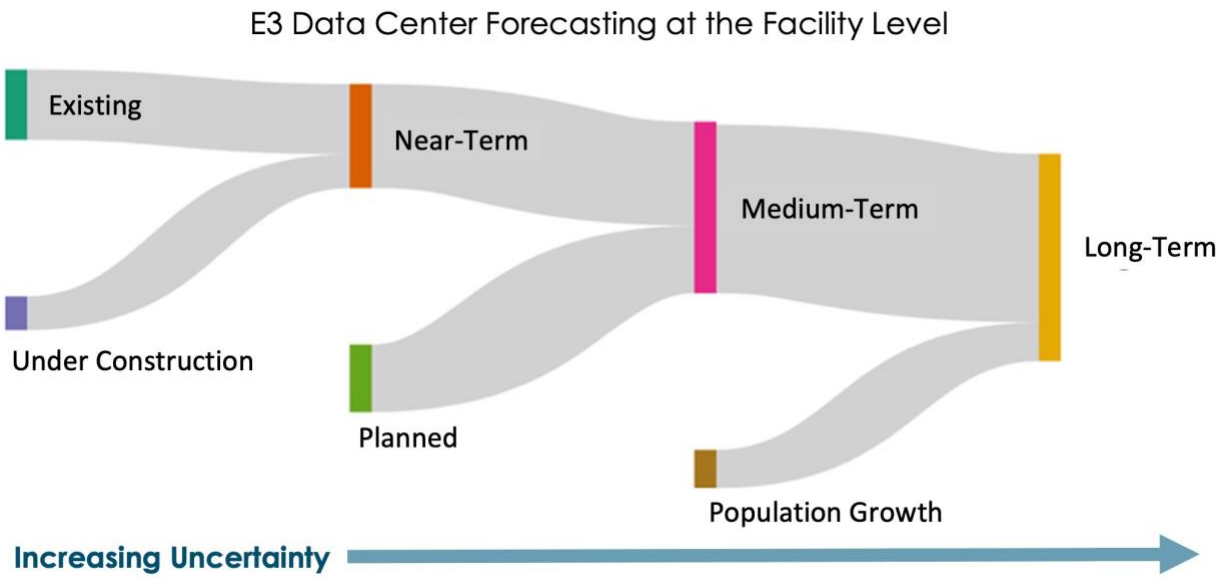
E3's Baseline for Large Load Forecasting

Forecasts, inherently, will never be "right", and E3's approach is not the only approach to large load forecasting. However, E3 has found that the following principles and guidelines protect from the risk of over- or under-forecasting.

Forecasting usually begins with a snapshot in time assessment of existing load, and large load or data center forecasting is no different. However, no complete national inventory of U.S. data centers and their electricity use currently exists. Many facilities do not disclose load data, making it difficult to model hourly or seasonal variations. Others withhold peak load information, requiring estimates based on physical footprint. Some are not even identified as data centers by their host balancing authority (BA). Before 2023, most BAs outside major data center hubs did not track these facilities separately, leaving forecasters to work with fragmented and incomplete datasets.

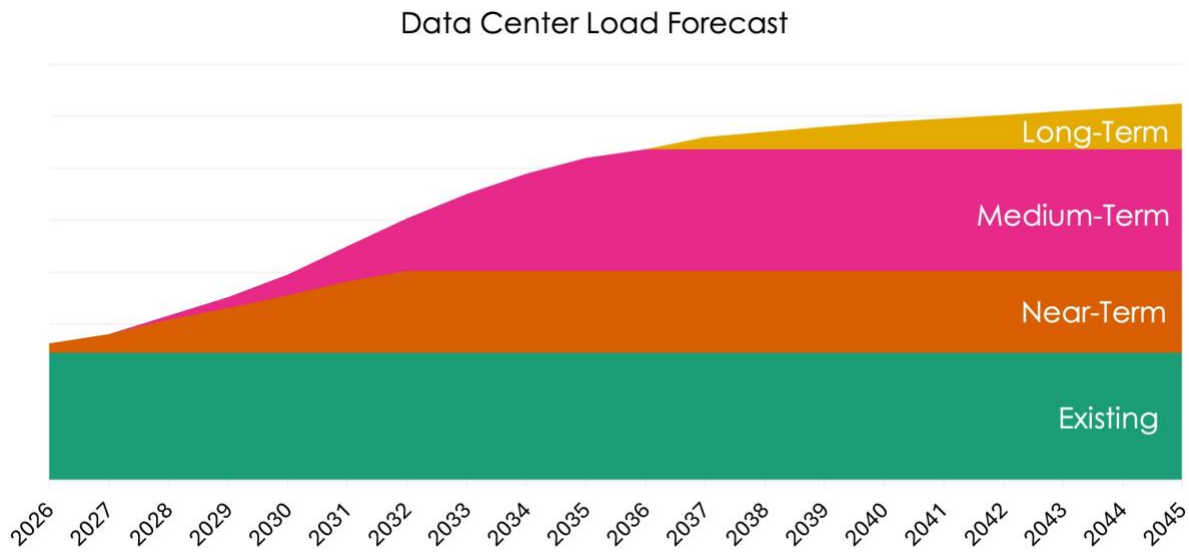
Selecting and refining these datasets is a core forecasting task. Utility-based telemetry provides high-quality data but may omit unreported loads, while market surveys capture a broader range of facilities without offering actual load measurements. Reliable forecasts combine both sources to create a balanced view of current demand.

Projecting future data center growth introduces additional uncertainty. Until recently, many BAs did not distinguish data center interconnection requests from other large load applications, and some developers engage in "queue shopping," submitting multiple interconnection requests to shorten connection timelines. These practices inflate apparent demand. Actual data center load also diverges from nameplate capacity, as facilities typically ramp up gradually after interconnection, taking several years to reach full operation.



E3's forecasting framework utilizing a bottom-up approach addresses these challenges by emphasizing analytical rigor and conservative assumptions. The foundation is a facility-level dataset that includes listed power draw, physical footprint, ownership, grid coordinates, and other attributes. This dataset is scrubbed, validated against utility-published data in key markets, and supplemented with regional inventory estimates to create a verified baseline.

E3's core forecast relies primarily on publicly announced projects, which establish a conservative floor for expected growth. Given limited data on interconnection queue completion rates, E3 uses this foundation to support alternative scenarios and sensitivity analyses. Facilities are categorized into "Existing", under construction ("Near-Term"), planned ("Medium-Term"), and "Long-Term" growth, which is inferred from demographic and economic trends. Together, these categories form E3's base case for U.S. and regional forecasts.



Data center growth over time can also be influenced by changes in regional policy that are either meant to attract or detract large load customers, which underscores the need to dynamically incorporate policies into forecasting. Because states can adjust their policies to attract load growth and economic development, changes in incentives or regulations can quickly redirect development across state and/or utility lines, shifting expected growth trajectories dramatically. Key siting factors include low-cost electricity, short interconnection timelines, land availability, fiber connectivity, and favorable tax policies. Regions scoring well across these dimensions tend to experience the fastest growth.

Siting factor	What "better" means	How this factor affects data center siting
Power availability and cost	More firm capacity and lower cost are better	The ability to secure large, long-term, reliable power blocks at competitive, predictable prices is a primary driver of total cost of ownership and schedule risk. Grid capacity, interconnection timing, curtailment risk, and tariff/market exposure are all critical.
Land price	Lower prices are better	Upfront capex for land can be material for large campuses; inexpensive but well-located land enables future expansion and lowers entry cost.
Water availability	More water is better	Cooling for certain designs requires dependable water sources; physical availability, drought risk, and competing uses drive technology choice and siting risk.
Fiber network	Closer and stronger connections are better	Proximity to high-capacity, diverse fiber routes reduces latency, increases resilience, and lowers connectivity costs; key for hyperscale and AI workloads.
Proximity to demand	Closer is better	Being near major user bases or interconnection hubs can reduce latency and improve performance, especially for real-time and enterprise workloads.

Siting factor	What "better" means	How this factor affects data center siting
Corporate tax rate	Lower is better	Jurisdiction-wide tax regimes (corporate income, property, and sales/use taxes) influence after-tax returns and can tilt decisions between similar sites.
Tax incentives	More are better	Targeted incentives (tax credits, abatements, grants, discounted tariffs) directly improve project economics and often drive site competition among states/regions.
Human capital	Stronger is better	Availability of skilled labor for construction, operations, networking, and security affects execution risk, ramp-up speed, and long-term operating quality.
Community opposition	Lower is better	Local acceptance, permitting risk, and potential litigation or political pushback can delay or block projects, increase mitigation costs, or constrain expansion.

Relative Importance of Siting Factors Over Time (Illustrative)

Siting Factor	Historical Importance	Importance Today	Importance in ~5 years
Power Availability and Cost	3	1	3
Land Price	4	7	8
Water Availability	8	3	2
Fiber Network	2	4	5
Proximity to Demand	6	8	6
Corporate Tax Rate	5	9	9
Tax Incentives	1	5	7
Human Capital	7	6	4
Community Opposition	9	2	1

As the illustrative matrix demonstrates, historically, data center siting has been driven primarily by tax incentives, proximity to fiber networks, access to low-cost power. In particular, E3 has found that tax exemptions and credits for IT equipment are among the most powerful incentives for development. Since servers and processors represent one of the largest capital expenses and are replaced every three to five years, exemptions on sales and property taxes can significantly reduce lifecycle costs. States with stronger incentives generally attract more and larger facilities, while those with weaker incentives see slower growth.

In recent years, however, growing grid constraints and extended interconnection timelines have elevated power availability to the dominant siting factor, diminishing the relative weight of other criteria, such as financial considerations.

Over the next several years, as power supply is expected to gradually catch up to demand, its relative influence on siting decisions is expected to moderate. Instead, water availability, amid growing resource constraints, and rising community opposition are expected to carry greater importance. Together, these trends may shift future development toward regions that offer both ample power and water resources and lower levels of community opposition.

Another challenge in developing a baseline is that "data center" is not a homogeneous load category. There are many types of facilities, each with a distinct load profile that can vary by factors such as geography or underlying business model. Understanding these differences is essential for accurate, high-fidelity modeling.

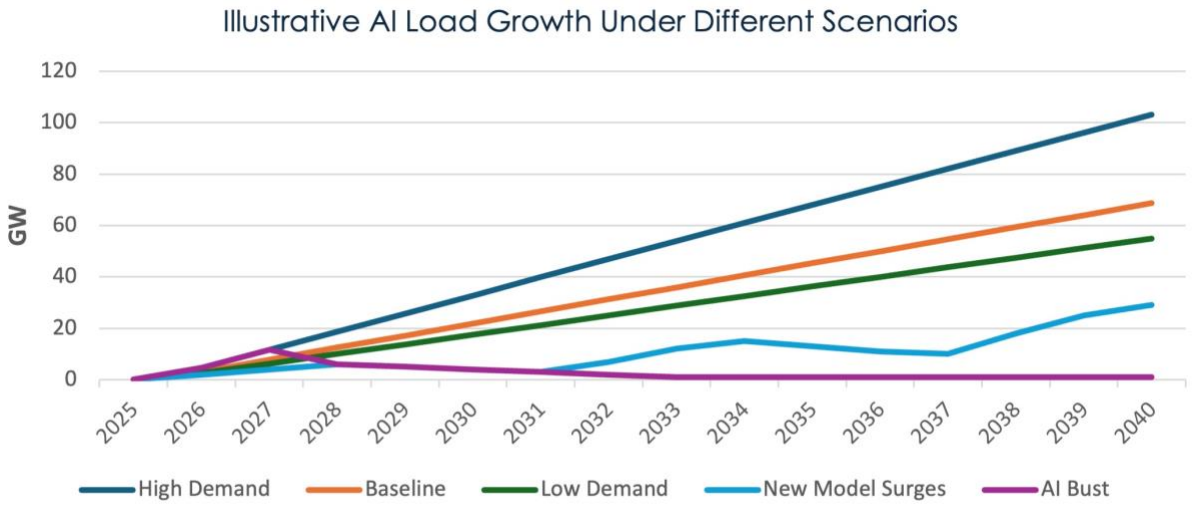
For example, facility size strongly influences load behavior. Edge facilities (less than 5 MW) are small, geographically distributed, and exhibit high variability tied to user activity. Standard facilities (5–100 MW) serve a wide range of cloud and enterprise functions with moderate variability. Hyperscale facilities (over 100 MW) operate as large, stable point loads with dedicated substations and minimal variation.

Operational purpose can also shape load profiles. Customer-facing facilities host web and cloud services with demand peaks that follow human activity patterns and can further vary depending on whether the primary use is business or personal. In contrast, modeling facilities focus on compute-intensive tasks, such as AI training, and operate with high load factors and minimal variability. Cryptocurrency mining facilities are more price-sensitive, often responding to electricity market conditions by curtailing load when costs rise or coin values fall and are more likely to participate in demand response programs.

E3's Approach to Modeling Future Large Load Growth

E3 uses scenario-based modeling to evaluate potential data center load growth across a wide range of possible futures. This approach helps utilities prepare for the most likely development patterns while identifying strategies to manage higher-risk outcomes. Given the high uncertainty surrounding data center expansion, these scenarios represent explorations of what could occur rather than precise forecasts.

E3 typically models a baseline case with accompanying high- and low-growth sensitivities to capture near-term uncertainty. For example, the "New Model Surge" scenario reflects rapid AI innovation that drives cyclical spikes in data center demand, while the "AI Bust" scenario illustrates a market contraction leading to widespread near- and medium-term underutilization of new infrastructure. These contrasting futures highlight both the opportunities and risks of long-term investment in large load growth.

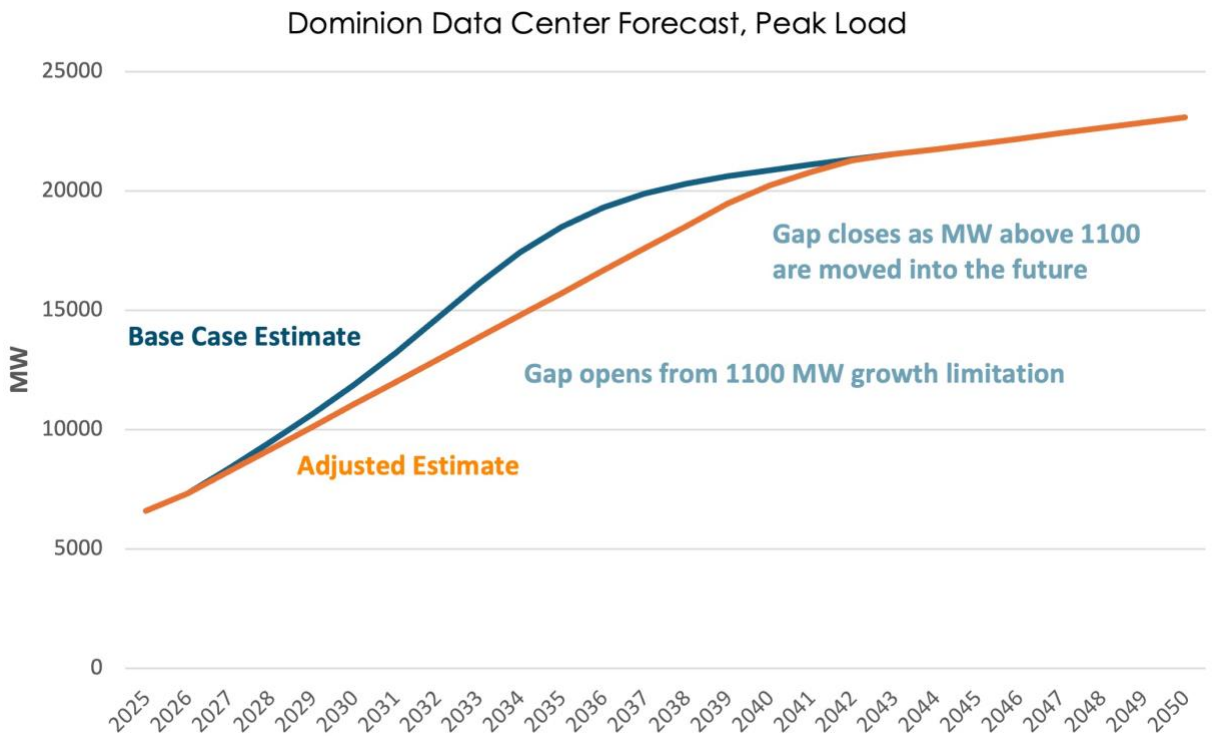


To test alternative growth scenarios, E3 applies a set of dynamic adjustment levers to its base-case load curve. These levers reflect key assumptions about project completion, timing, and operations:

- **Attrition Rates**: Share of planned and under-construction projects expected to reach completion
- **Completion Windows**: Timeline for new projects to achieve energization
- **Boom and Decay Factor**: Rate at which current rapid growth stabilizes over time
- **Ramp Rates**: Duration from initial energization to full operational load
- **Load Shapes**: Hourly variation in data center electricity use throughout the year
- **Peak Factor**: Portion of interconnection capacity used during system peak periods
- **Load Factor**: Ratio of annual energy use to theoretical maximum consumption

Adjusting these levers allows the examination of how timing, scale, and utilization affect overall load growth. For instance, higher attrition rates reduce total peak demand, while longer completion or ramp periods delay growth without changing total capacity. Variations in load shape can alter annual energy consumption without affecting peak demand.

Systematically testing alternative combinations of these parameters helps reveal the full range of possible grid impacts. For example, in one utility territory, E3 compared its base forecast with a constrained case that limited new large load additions to 1,100 MW. Although total load by 2045 remained the same, the constrained case resulted in nearly 3 GW less deployed capacity by 2035, underscoring how different development timelines can dramatically influence grid planning and investment needs.



Applying E3's Approach at the Utility Level

Utilities are uniquely positioned to apply a similar framework for forecasting large load growth. Unlike external analysts, utilities have direct access to customer interconnection data, construction timelines, and operational telemetry within their service territories. By combining this granular insight with E3's structured scenario-based methods, utilities can create more accurate, flexible forecasts that reflect both near-term realities and long-term uncertainty.

A utility-led approach could use verified interconnection queues, customer milestones, and load telemetry as the foundation for baseline forecasts, then apply scenario modeling and dynamic adjustment levers to explore alternative futures. This integration would help identify risks of over- or under-building, test system resilience under varying demand trajectories, and guide more adaptive investment strategies.

Through disciplined data validation and iterative scenario testing, utilities can strengthen their forecasting processes, reduce exposure to forecasting errors, and better align planning decisions with the evolving landscape of large load growth.

Leveraging Forecasts for Strategy: Managing Risk through DR and Rate Design

Building on E3's modeling approach, utilities can strengthen their planning by connecting forecasting, demand response (DR), rate design, and risk management. DR (i.e., customer load flexibility in response to grid needs) and rate

design can serve as critical risk mitigation tools that help utilities manage uncertainty and bridge gaps between forecasted and actual load. Together, these elements create a unified framework that supports new load growth, maintains reliability, and protects customers from financial exposure.

Forecasting is an essential first step, as it provides the analytical foundation for understanding large load behavior and identifying opportunities for flexibility. It allows planners to characterize major customer types, such as data centers, and assess their potential for DR participation. As previously discussed, load profiles can vary widely and offer differing levels of flexibility, such as AI training facilities being more cyclical versus inference or cloud-hosting centers with steadier, less flexible load profiles. Recognizing these differences is essential for quantifying the role and potential for DR, whether load shed (reducing usage during times of grid stress) or load shift (moving usage to lower demand periods).

Accurate large load forecasting provides the foundation for determining:

- **Available Demand Response Potential:** By identifying typical peak demand periods and operational profiles, utilities can estimate how much discretionary load can be curtailed or shifted.
- **Trigger Thresholds and Event Frequency:** Forecast precision helps define when and how often DR events might be called, avoiding both overreliance and underutilization of participating facilities.
- **Infrastructure Requirements:** Forecasts inform whether additional metering, telemetry, or automation investments are needed to support DR participation from speculative or rapidly scaling customers.

This understanding sets the stage for DR and other flexibility tools to play a meaningful role in maintaining reliability without overbuilding.

The Potential Role of Load Flexibility

The relationship between forecasting and DR is reciprocal. Accurate forecasts inform DR program design, and operational data from DR events enhance forecast precision. For example, some computing tasks, such as AI training or batch processing, can be deferred or shifted without major operational impacts, while real-time services like website hosting cannot.

To date, data centers have had low DR participation, given their load profiles and economic cost of downtime, driven by the high value of IT equipment and rapid depreciation. But certain computing tasks, such as AI training or batch processing, are inherently more flexible than real-time inference or storage, and some customers are willing to flex load for certain incentives such as faster interconnection timelines. For forecasters, these behavioral distinctions across load types and demand elasticity are essential. They allow models to move beyond treating data center demand as a uniform block and instead reflect the diversity of technologies, business models, and operating philosophies emerging across the sector.

Recent examples suggest that under the right conditions, portions of data center load can be managed more dynamically. In 2021, Google shared it could shift compute tasks across data centers according to hourly carbon-free

energy availability[h] and in 2025, Google has cemented DR participation in two new utility agreements with Indiana Michigan Power and the Tennessee Valley Authority to participate in DR by targeting machine learning workloads[i].

From a technical demonstration standpoint, a collaboration among DCFlex, Emerald AI, Oracle, and Nvidia displayed a 25% reduction in a facility's power use over a three-hour period through real-time orchestration of GPU workloads without compromising performance. The platform achieved this by slowing non-critical tasks, pausing non-time-sensitive batch processes, and rescheduling flexible workloads to lower-demand periods[j].

Data centers can also participate in load shedding programs without necessarily reducing IT load if supported by on-site generation. For example, Microsoft's San Jose data center uses its RNG microgrid to power facility operations when it participates in California's Base Interruptible Protocol (BIP) events[k].

These examples are a growing trend and can provide valuable insights, as over time, data from DR participation enables utilities to refine their forecasts with greater precision. Real-world observations of curtailment depth, duration, and frequency help ground assumptions about load elasticity and peak-period behavior, reducing uncertainty about actual load flexibility, which can otherwise lead to overstated peak forecasts and unnecessary capacity procurement. Although broad participation remains limited, the expanding portfolio of pilots underscores the value of experimentation. Each event and tariff negotiation yields new insight into how large customers engage with the grid. As these lessons are integrated into planning frameworks, utilities develop a clearer and more empirical view of potential.

Integrating Forecasting, DR and Rate Design for Risk Management

Forecasting, DR, and rate design each play distinct roles, but they are most powerful when integrated into a single, coordinated framework to manage risk and uncertainty.

Forecasting establishes the analytical foundation by defining the range of possible outcomes and highlighting where uncertainty is greatest. Rather than seeking a single "correct" forecast, planners can develop multiple scenarios to test exposure (e.g., examining what happens if load materializes faster, slower, or not at all) and identify the points at which DR activation or financial safeguards become necessary.

Demand response serves as a critical hedge against forecast error. If load grows faster than anticipated, flexible demand can be curtailed to preserve reliability. Conversely, if growth slows or fails to materialize, DR commitments can be scaled back (with stranded asset risk mitigated via rate design). This flexibility transforms DR from a short-term reliability tool into a dynamic risk mitigation mechanism that helps utilities balance uncertainty and infrastructure

---

[h] *Koningstein, Ross. "We Now Do More Computing Where There's Cleaner Energy." Google, 18 May 2021, https://blog.google/outreach-initiatives/sustainability/carbon-aware-computing-location/.*

[i] *Terrell, Michael. "How We're Making Data Centers More Flexible to Benefit Power Grids." Google, 4 August 2025, https://blog.google/inside-google/infrastructure/how-were-making-data-centers-more-flexible-to-benefit-power-grids/.*

[j] *Allsup, Maeve. "Nvidia and Oracle Tapped This Startup to Flex a Phoenix Data Center." Latitude Media, 1 July 2025, https://www.latitudemedia.com/news/nvidia-and-oracle-tapped-this-startup-to-flex-a-phoenix-data-center/.*

[k] *Judge, Peter. "Microsoft's San Jose Data Center Will Use Food Waste Gas for Back-up Power." Data Center Dynamics, 12 Dec. 2023, https://www.datacenterdynamics.com/en/news/microsofts-san-jose-data-center-will-use-food-waste-gas-for-back-up-power/*

efficiency. Moreover, DR can serve as an enabling mechanism for load growth, by allowing utilities to integrate more new customers by leveraging flexible demand as curtailment-enabled headroom.[l]

Rate design can play multiple roles in managing uncertainty, both by incentivizing or requiring DR and by mitigating financial and operational risk. For example, key incentives could be reduced rates (such as in non-firm rates) or could include accelerated interconnection timelines, which are increasingly valuable given the current focus on time-to-power amidst mounting delays. Rate designs with required load flexibility are emerging as well, such as Texas Senate Bill 6, which establishes both mandatory load-shed participation and voluntary DR programs for loads of 75 MW or more.[m]

Rate design can also be a powerful tool for mitigating financial risks stemming from forecasting uncertainty and load variability. To help reduce forecasting error, specifically around speculative and/or duplicative projects, utilities can implement financial safeguards, such as upfront deposits or milestone-based payment frameworks. To reduce risk of default, utilities can employ a menu of credit and collateral requirements and can go beyond the traditional use of credit ratings. E3 took a deep dive into these best practices in the July 2025 whitepaper "Balancing Risk and Growth: Best Practices for Utility Credit and Collateral Requirements for Large Load Customers".[n] For example, DR participation requirements or incentives could be linked to creditworthiness or collateral tiers to align system benefits with financial exposure. To help reduce stranded asset risk, utilities can use rate design mechanisms (e.g., contract minimums, "take-or-pay" provisions, exit fees) that secure baseline commitments, align financial responsibility, and minimize potential ratepayer impacts of deviations between forecasted and actual load.

Together, these elements create a resilient planning framework — where forecasting provides the analytical foundation, DR offers the operational buffer, and rate design equitably allocates costs and mitigates risk — to enable utilities to adapt dynamically to both over- and under-forecasting scenarios.

Conclusion and Key Takeaways

Utilities face a dual challenge: meeting the electricity needs of new industries and protecting customers from the financial risks of forecasting uncertainty. Both over-building and under-building carry long-term consequences that can affect system performance, affordability, and public trust.

The goal is not to eliminate uncertainty but to manage it intelligently. The pace and scale of load growth driven by AI and data centers is testing the limits of traditional planning methods, and addressing this challenge requires forecasting that is adaptive, transparent, and grounded in real-world data. Forecasting must evolve from a static prediction exercise into a continuous process of measurement, testing, and recalibration that integrates operational and financial feedback.

When forecasting becomes part of a broader risk management framework aligned with DR, rate design, and financial safeguards, it shifts from being a one-time analysis to an active management tool. Utilities that adopt this approach

---

[l] *Norris, Tyler H., Tim Profeta, Dalia Patiño-Echeverri, and Adam Cowie-Haskell. Rethinking Load Growth: Assessing the Potential for Integration of Large Flexible Loads in US Power Systems. Nicholas Institute for Energy, Environment & Sustainability, Feb. 2025, https://nicholasinstitute.duke.edu/publications/rethinking-load-growth*

[m] *Martucci, Brian. "Texas Law Gives Grid Operator Power to Disconnect Data Centers During Crisis." Utility Dive, 25 June 2025, https://www.utilitydive.com/news/texas-law-gives-grid-operator-power-to-disconnect-data-centers-during-crisi/751587/.*

[n] *Riu, Isabelle, Shana Ramirez, Melina Bartels, Sophia Greszczuk, and Kush Patel. Balancing Risk and Growth: Best Practices for Utility Credit and Collateral Requirements for Large Load Customers. E3, July 2025. https://www.ethree.com/wp-content/uploads/2025/08/E3_Utility-Credit-and-Collateral-for-Large-Load_Whitepaper.pdf.*

will be better positioned to make durable investments, maintain reliability, and demonstrate to regulators and customers that their decisions are guided by evidence and prudence rather than speculation.

History underscores the stakes. The overconfidence of the 1970s left customers paying for stranded assets and excess capacity, and the same risk now looms as large load growth accelerates. Without disciplined forecasting, utilities could again overbuild infrastructure or fail to prepare for surging demand. Either error would impose unnecessary costs and erode public trust. Applying the structured, scenario-based methods described in this paper allows utilities to strengthen forecasting practices while balancing ratepayer protection with economic opportunity.

Key Principles for a Resilient Forecasting Framework

The following practices summarize the approach described throughout this whitepaper:

1. **Start with verified data:** Build forecasts from measured baselines, not speculative interconnections.

Build forecasts from verified, geolocated data center inventories and historical performance data when available, reconciling interconnection queues with actual energization timelines to ground projections in reality.

Plan proactively for large load growth and leverage market context, such as distinguishing between different data center types and factoring in regional competitiveness, permitting, transmission, and policy constraints.

2. **Use scenario diversity**: Explore a range of scenarios to understand the realm of possibility (e.g., AI "bust" vs. AI "boom" and the impact of key levers).

Given the unpredictable nature of this load growth, scenario diversity is critical to stress-test forecasts and understand implications. Utilities should model a core base case anchored in verified project data and then explore a set of complementary cases that can help bracket uncertainty.

Scenarios can be designed by applying a series of adjustment levers (including attrition rates, boom-and-decay factors, ramp rates, and load or peak factors) to test the sensitivity of the forecasts to different underlying assumptions. Potential scenarios could include:

- High-growth or acceleration scenarios, where rapid AI adoption and/or new industrial electrification push demand beyond expectations.
- Efficiency or plateau scenarios, where technology improvements or operational changes moderate demand intensity.
- Market-correction or deflation scenarios, which capture the risk of broader market cycles, speculative overbuild, failures to ramp, and/or investment slowdowns.

The goal is not to predict the most likely outcome but to understand the consequences of variance in either direction and the biggest drivers that can alter the load forecast's magnitude and shape. Scenario diversity builds resilience by revealing throughlines across scenarios and highlighting which strategies can be applicable under a wide range of outcomes.

3. **Implement adaptive planning**: Regularly recalibrate forecasts using real-time data from utility interconnections and AI industry trends especially from other jurisdictions that provide valuable insights and signposts.

Forecasting should function as a living process that evolves with changing conditions, bolstered by regular updates and discrete revisions triggered by clear operational or market signals. Utilities can strengthen forecast resilience by using real-time data and operational feedback to refine assumptions as new information emerges and benchmark

against peer utilities and broader industry trends. Each forecasting cycle becomes an opportunity for learning by investigating causes of deviations and adaptively integrating those insights into future planning efforts.

4. **Leverage Demand Response to Improve Forecast Credibility**: Integrate DR into forecasting to provide an operational buffer against volatility and uncertainty.

Integrate DR into forecasting to create a vital feedback loop between planning and operations. As utilities observe real-world curtailment depth, duration, and responsiveness, they should refine assumptions about flexibility potential, validate load assumptions, and improve forecast accuracy.

5. **Use Rate Design to Incentivize Load Flexibility**: Reward customers who modulate demand, including faster interconnection access and demand credits.

Meaningfully incentivize load flexibility, such as by offering faster interconnection to customers who agree to curtail and/or shift load. Well-structured tariffs can turn load flexibility into a cost-effective system asset and enabler of additional load growth, instead of posing risks.

6. **Align financial risk to maintain "symmetry"**: Require load commitments and credit support for speculative customers.

Protect ratepayers from speculative growth by using risk mitigation tools tailored to each stage of the project lifecycle to reduce stranded asset risk. During early study phases, use financial mechanisms like upfront deposits, milestone-based payments, or credit requirements proportional to project maturity and creditworthiness. After projects are energized, safeguards like contract minimums, "take-or-pay" provisions, or exit fees become relevant to secure baseline commitments.

7. **Share lessons transparently**: Publish forecast-to-actual tracking to improve model credibility over time and promote institutional learning.

Transparency transforms forecasting from a technical exercise into a collective learning process. Utilities can strengthen credibility and institutional knowledge by comparing forecasts to actual outcomes and sharing results with regulators, stakeholders, and peer organizations as permitted and/or in the aggregate/anonymized. Maintaining documentation of how methodologies evolve over time ensures that model improvements are traceable and grounded in evidence. By engaging stakeholders early and fostering collaboration across utilities, the sector can build a shared foundation of data, experience, and best practices that continually improve forecasting quality and trust.

The future of forecasting will not depend on perfect prediction but on adaptive precision and the ability to respond effectively as conditions change. Utilities that view forecasting as an evolving, data-driven discipline will be best equipped to navigate the next decade of energy transformation. By integrating forecasting with operational flexibility and financial discipline, the sector can support technological growth without repeating the costly overconfidence of the past. A resilient forecasting framework turns uncertainty into a strategic advantage, enabling the power sector to grow not only larger but also smarter, stronger, and more responsive to the needs of the customers and communities it serves.