Rethinking Load Growth: Assessing the Potential for Integration of Large Flexible Loads in US Power Systems

Tyler H. Norris, Tim Profeta, Dalia Patino-Echeverri, and Adam Cowie-Haskell

Webinar Presentation

February 19, 2025

Duke NICHOLAS INSTITUTE *for* ENERGY, ENVIRONMENT & SUSTAINABILITY

Agenda

Introduction

Who we are

Duke NICHOLAS INSTITUTE *for* ENERGY, ENVIRONMENT & SUSTAINABILITY

The Nicholas Institute at Duke University accelerates solutions to critical energy and environmental challenges, advancing a more just, resilient, and sustainable world.

The Nicholas Institute conducts and supports actionable research and undertakes sustained engagement with policymakers, businesses, and communities—in addition to delivering transformative educational experiences to equip future leaders. The Nicholas Institute's work is aligned with the Duke Climate Commitment (climate.duke.edu).

GRACE Lab

A Grid that is Risk-Aware for Clean Electricity

Duke | NICHOLAS SCHOOL *of the* ENVIRONMENT

GRACE Lab's research explores, assesses, and proposes technological, policy, and market approaches to contribute to the pursuit of sustainability, affordability, reliability, and justice in the energy sector.

Primary research areas:

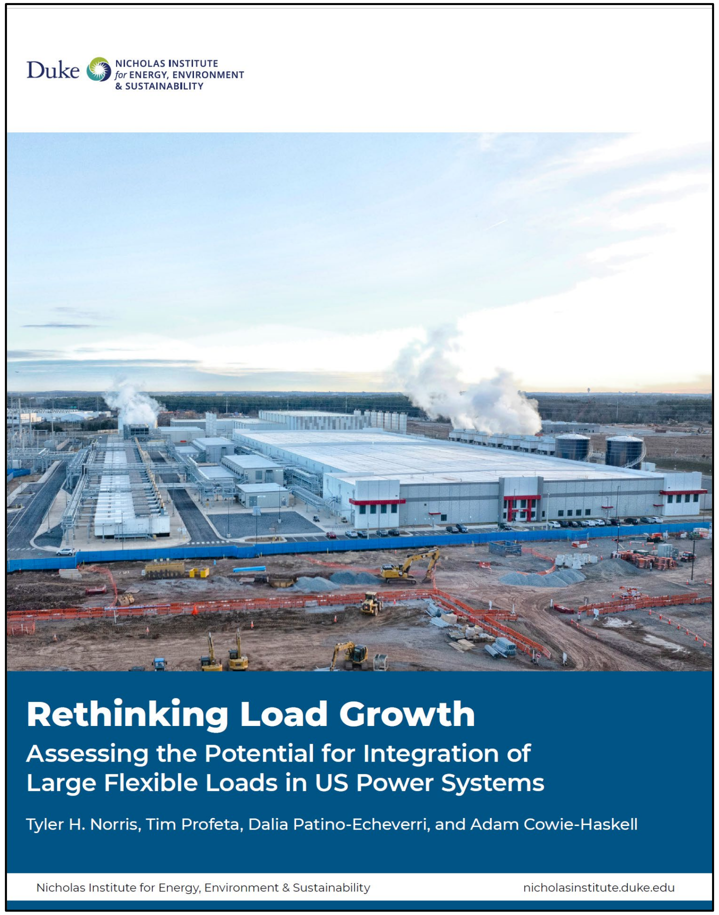
- Characterizing sources of uncertainty that increase the financial and reliability risk of power systems, and designing risk management strategies
- Examining the possibilities and advantages of designing flexible policy mechanisms
- Assessing the economic, environmental, and reliability potential of low-emissions energy technologies

Goals for *Rethinking Load Growth*

- Support regulators and stakeholders in identifying strategies to accommodate load growth without compromising reliability, affordability, or progress on decarbonization

- Provide informational resources and a first-order estimate of the potential for accommodating new loads while mitigating or deferring capacity expansion

- Motivate additional analysis to more precisely quantify headroom in each balancing authority



Load growth is colliding with resource constraints

- Supply chain constraints, protracted interconnection queues, and extended permitting processes are limiting the development of new electricity infrastructure

- The scale of the challenge underscores the importance of deploying every available tool, especially those that can more swiftly, affordably, and sustainably integrate large loads

- The time-sensitivity for solutions is amplified by the market pressure for many of these loads to interconnect as quickly as possible

Credit: IEEE

Load flexibility offers a near-term solution

The US Secretary of Energy Advisory Board (SEAB) and the Electrical Power Research Institute (EPRI) have highlighted a solution: load flexibility.

U.S. DEPARTMENT OF
ENERGY
Secretary of Energy Advisory Board

Recommendations on
Powering Artificial Intelligence and Data Center Infrastructure

Presented to the Secretary of Energy on July 30, 2024

"For immediate impact, the Secretary should direct relevant offices across DOE to explore opportunities for temporal and spatial ***flexibility in AI training and inference****..."*

SEAB Recommendations (PDF)

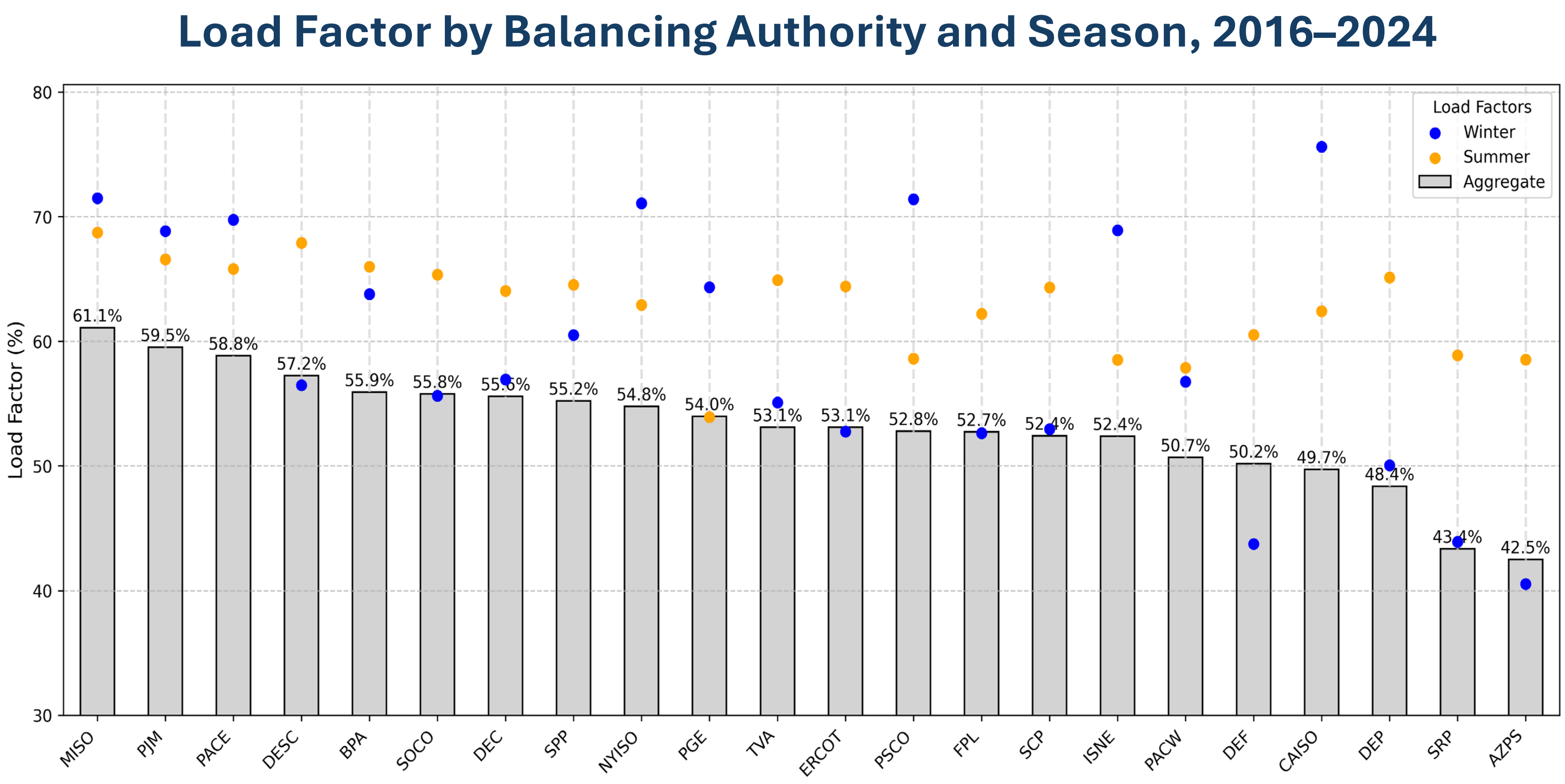
"EPRI's DCFlex Initiative will demonstrate how data centers can ***support and stabilize*** *the electric grid while improving interconnection and efficiency."*

https://msites.epri.com/dcflex

Background

US power systems operate at 53% avg. load factor

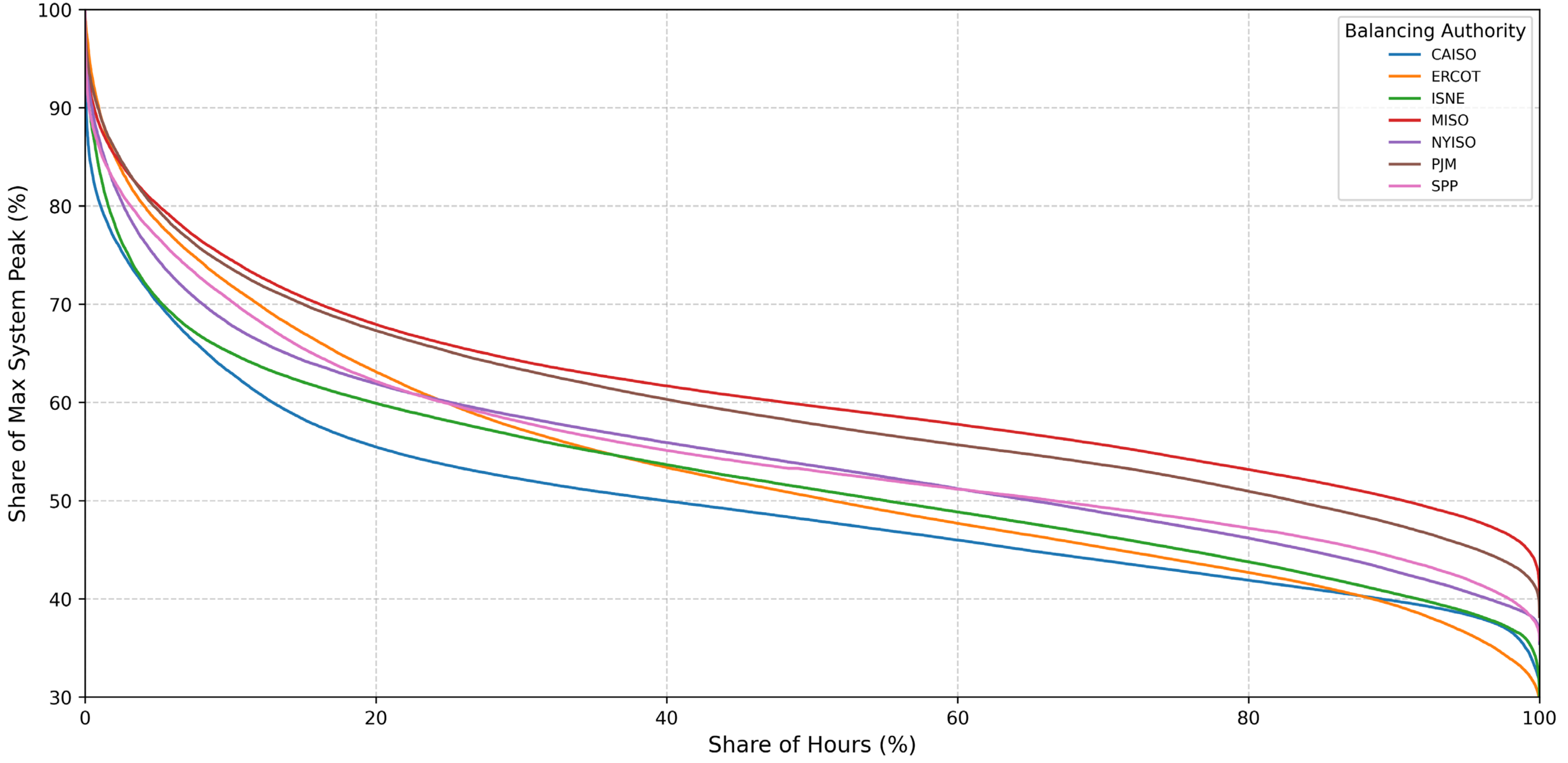
- Load factor is the ratio of average demand to peak demand and is an indication of system utilization
- Aggregate load factors range between 43% to 61%, with an average and median value of 53%
- Winter load factors were notably lower than summer (59% vs. 63% average)



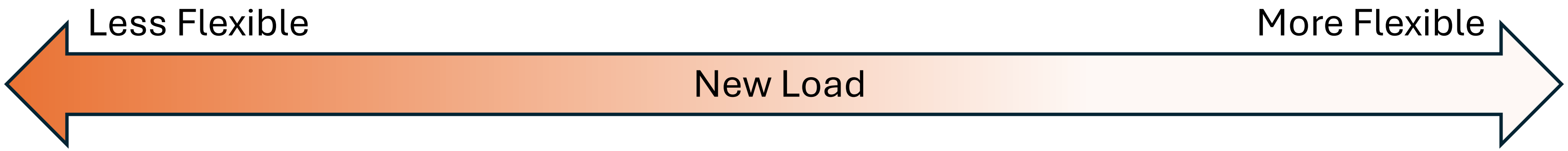
Power systems are designed to meet occasional peaks

- The load duration curve illustrates system utilization by ranking demand from highest to lowest over a given period
- A steep LDC suggests high demand variability, with peaks significantly exceeding typical loads, while a flatter LDC indicates more consistent usage

Load Duration Curve for US RTO/ISOs, 2016–2024



Implications for system planning and interconnection



Inflexible Loads

- More burden on system peaks
- More likely to trigger need for additional generation and transmission capacity
- More likely to require longer interconnection

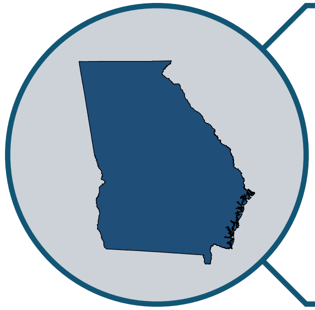
Flexible Loads

- Less burden on system peaks
- Less likely to trigger need for additional generation and transmission capacity
- Potential for accelerated grid interconnection

Data center growth is challenging regulators today

Ohio

- After confirming 5 GW of new data centers and receiving 30 GW in requests, American Electric Power (AEP) issued a temporary moratorium on data center service requests in 2023
- Settlement between AEP and consumer advocates calls for longer contracts, load ramping schedules, min. demand charge, and collateral for service from data centers >25 MW



Georgia

- After 7.3 GW of large load customers committed to receive electric service from Georgia Power, the Georgia PSC implemented changes to customer contract provisions
- Changes mandate a GPSC review and allow the utility to seek longer contracts and minimum billing for cost recovery

Indiana

- Data center service requests represent a 157% increase in peak load for Indiana Michigan Power over the next six years
- Stakeholders there have proposed "firewalling" the associated cost of service for data centers from the rest of the rate base

Types of load flexibility

On-site power and storage
Utilizing co-located storage, renewables, or other generators

Temporal flexibility
Scheduling computational loads before or after periods of high system stress

Spatial flexibility
Distributing workloads across one or multiple data centers in different geographic locations

Reduced operations
Planning for reduced workload during defined periods

Trends enabling flexibility

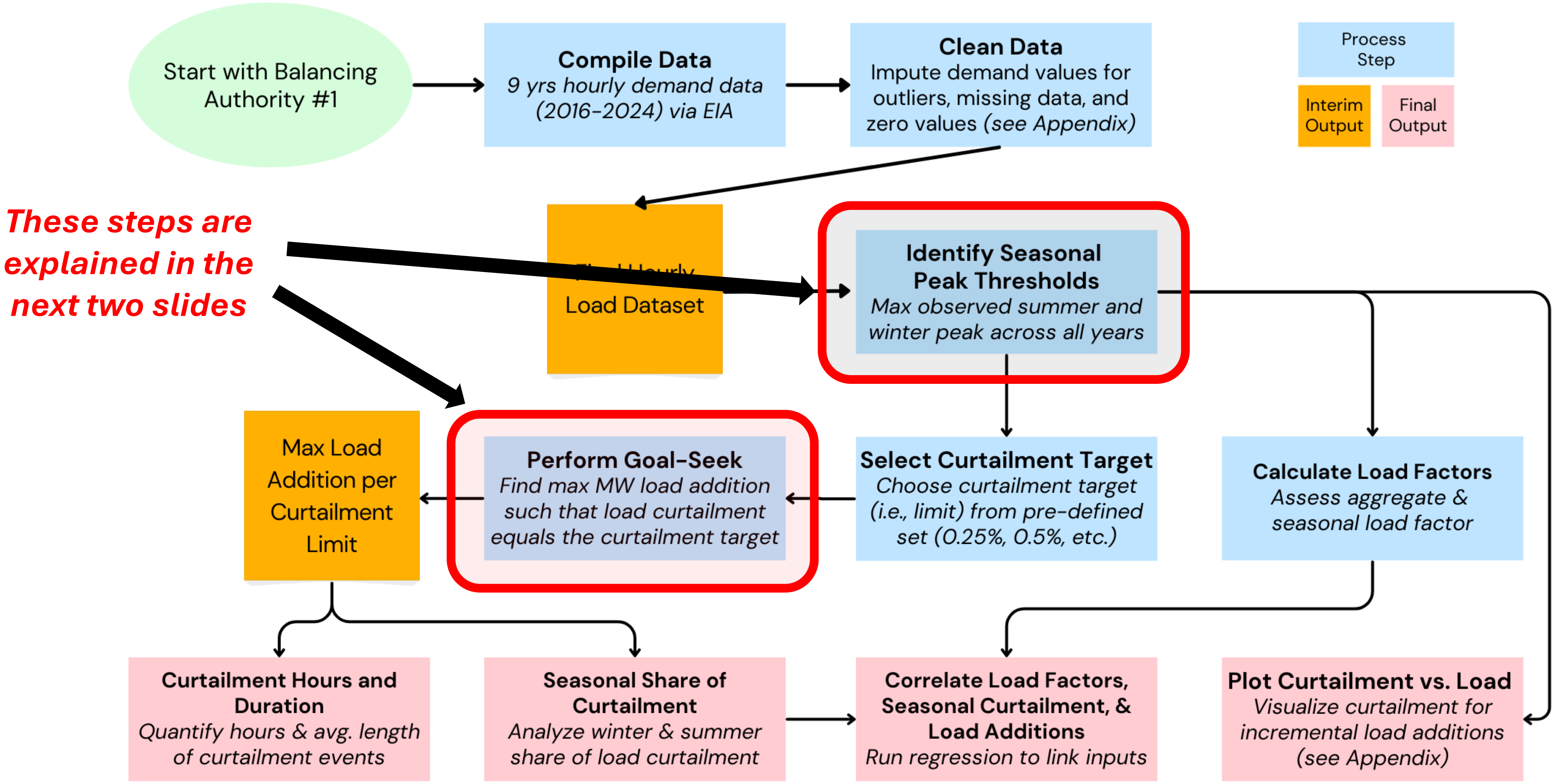
Category	Legacy Computational Loads	Future Computational Loads
Load profile	• Majority real-time, delay-intolerant processing (e.g., cloud services)	• Greater portion of delay-tolerant and scheduled machine learning workloads (model training, non-interactive services)
Operational capabilities	• Minimal workload shifting (spatial or temporal) • On-site power typically Tier 2 diesel and restricted due to pollution concerns	• Commercial adoption of temporal workload shifting • Development in spatial workload migration and other flexible processes • On-site power diversified with cleaner resources
Market Conditions	• Minimal load growth and generally high available capacity • Standard interconnection queues and supply chain readiness	• High load growth and tight available capacity • Long interconnection queues and costly supply chain bottlenecks • Lower cost of cleaner, on-site power • Data center operations concentrating in large-scale operators

Select examples of implementation

Market Participants	System Operators and Utilities
• **Google** data centers have participated in demand response, "carbon aware" temporal workload-shifting, and geospatial workload shifting • **Enel X** supports data center demand response participation, including use of on-site batteries and generators to enable islanding within minutes • **Verrus** is developing flexible data centers with different electrical and cooling systems architecture with distributed energy resources • **Enchanted Rock** supported Microsoft to install on-site natural gas generators for a data center in San Jose, CA • Startups like **Emerald AI** are developing software for advanced computational resource management	• **ERCOT** established a Large Flexible Task Force and implemented an interim interconnection process proposing to allow loads to be studied as flexible, "Controllable Load Resources" to interconnect within a two-year timeframe • **PG&E** debuted Flex Connect, a pilot that provides quicker interconnection to large loads in return for flexibility when the system is constrained

Analysis

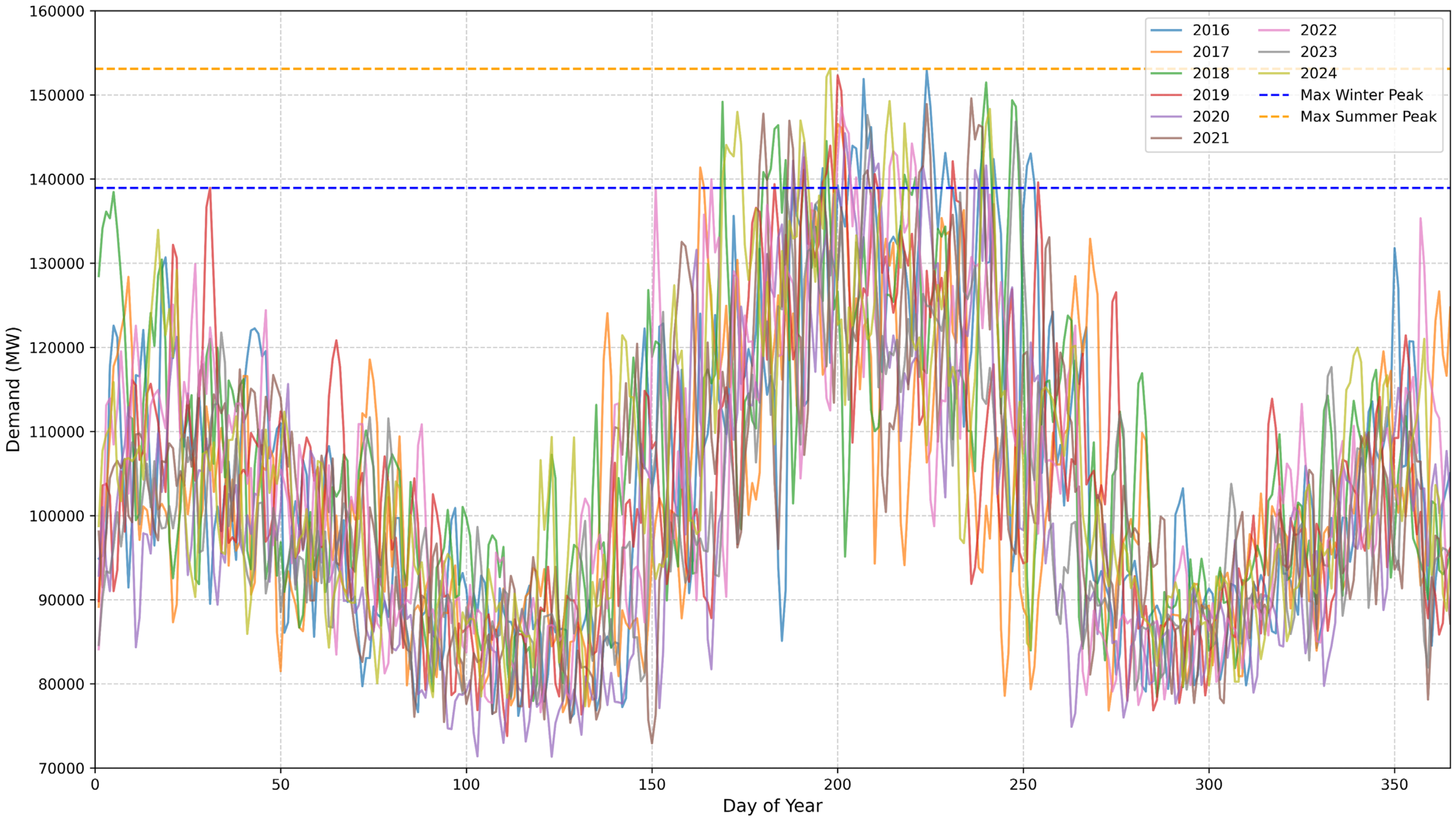
Method Summary



Method: Identify Seasonal Peak Thresholds

- Balancing authorities develop resource expansion plans to support different peak loads in winter and summer
- To account for variation, we identified the max winter and summer peak observed for each balancing authority
- These thresholds serve as the upper limits for system demand during their respective seasons

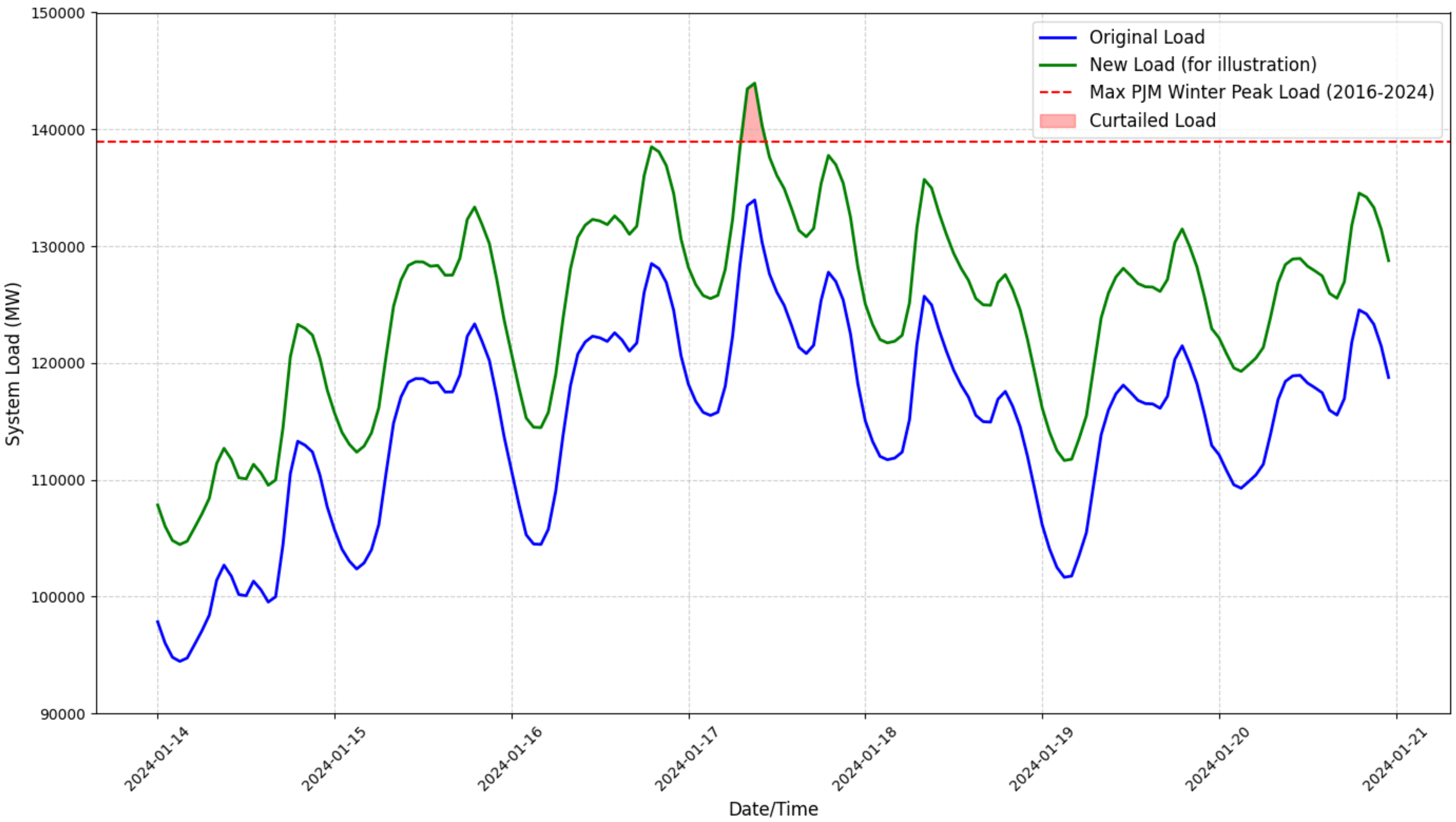
Daily and Seasonal Maximum Demand in PJM (2016-2024), MW



Method: Calculate Load Additions for Curtailment Limits

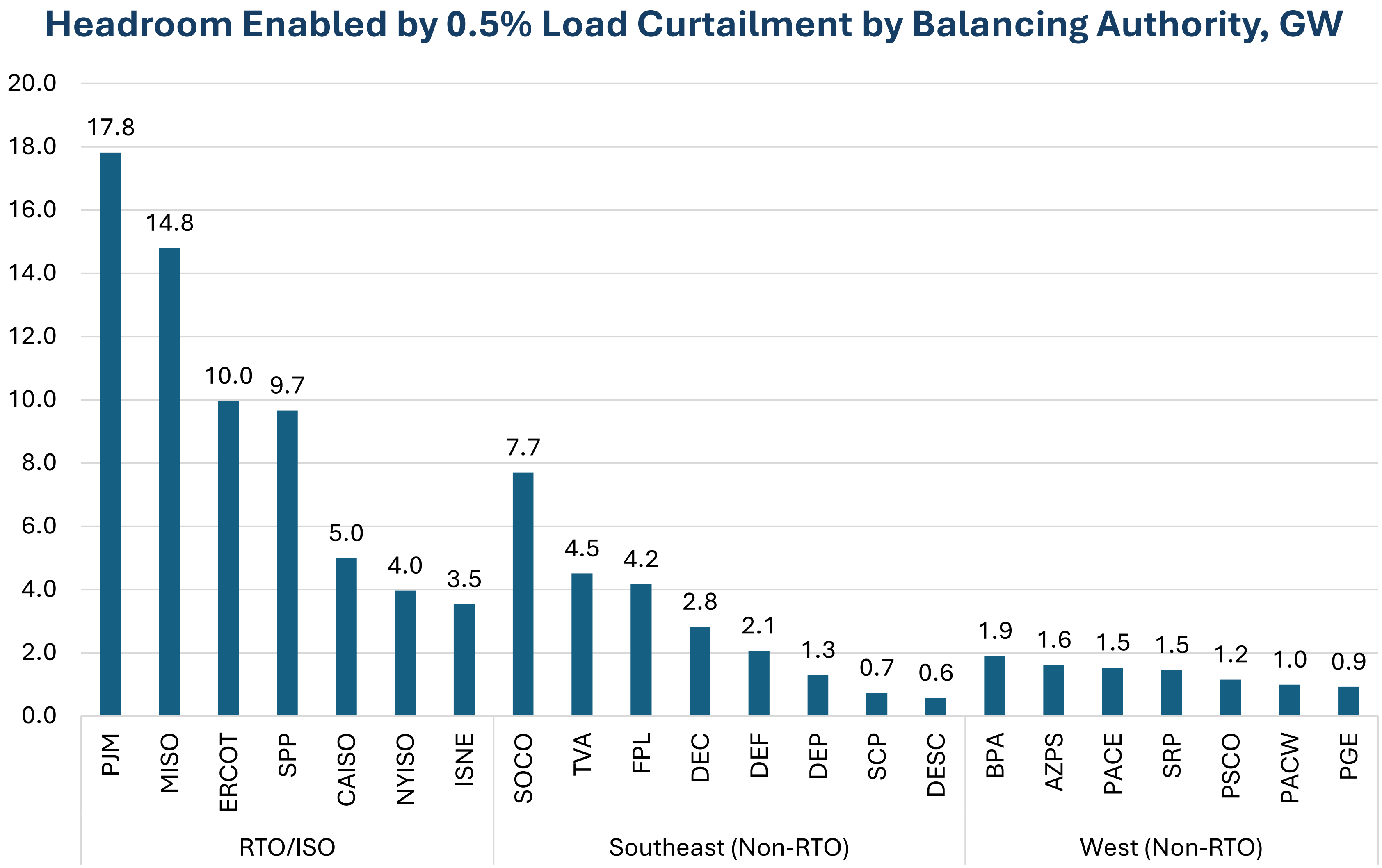
- New, constant load was added in all hours
- Curtailment was calculated as the difference between the new load and the seasonal peak threshold in each hour, summed across all hours in a year
- The curtailment rate for each load increment was defined as the total annual curtailed MWh divided by the new load's max potential annual consumption

Illustrative Load Curtailment in PJM (Jan. 14–21, 2024)



Results: Curtailment-Enabled Headroom

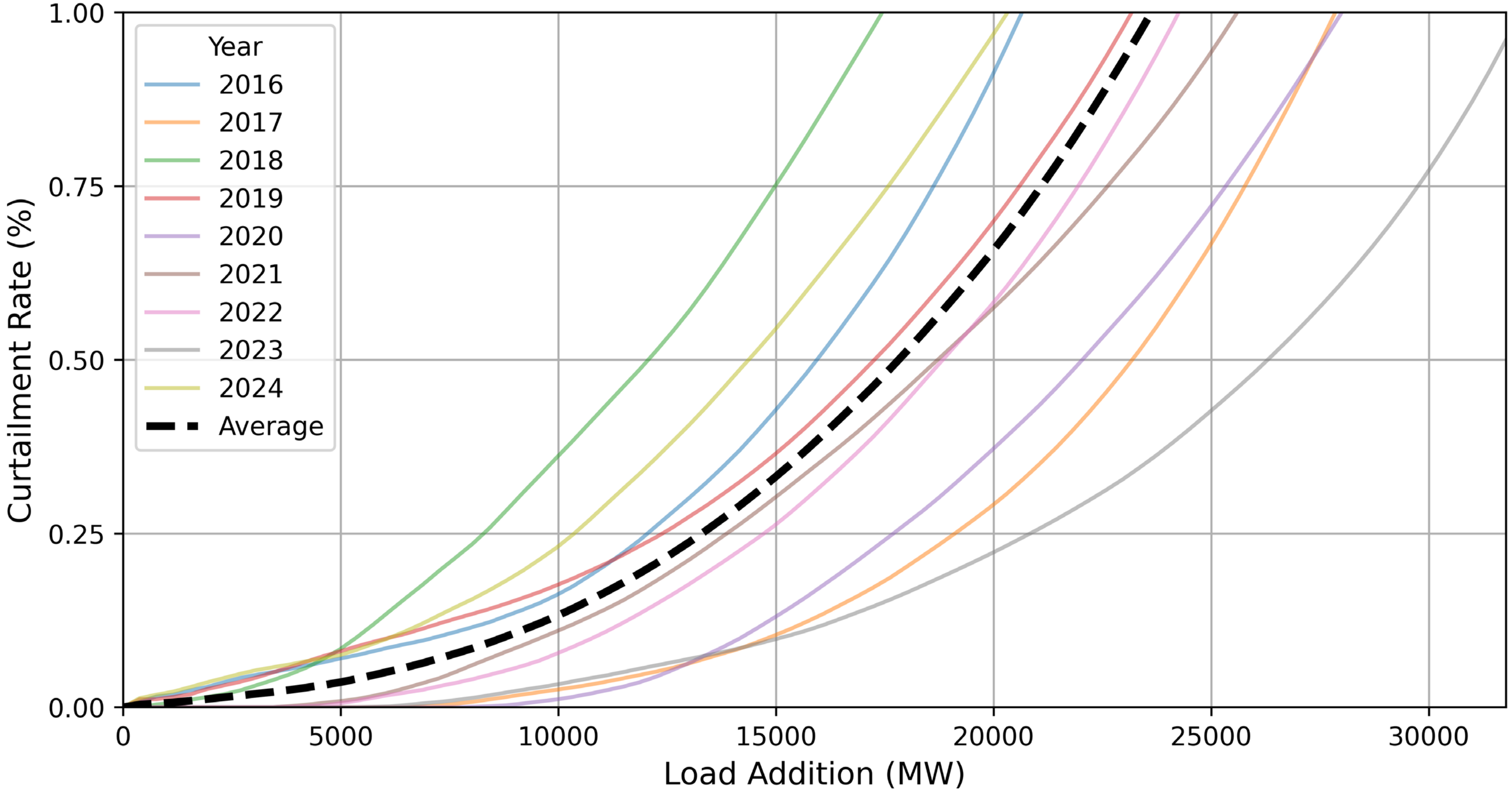
- Headroom across the 22 analyzed balancing authorities is between 76 to 215 GW, depending on the applicable load curtailment limit
- 76 GW of headroom is available at an expected load curtailment rate of 0.25%
- This headroom increases to 98 GW at 0.5% curtailment, 126 GW at 1.0%, and 215 GW at 5.0% curtailment



Results: Curtailment-Enabled Headroom

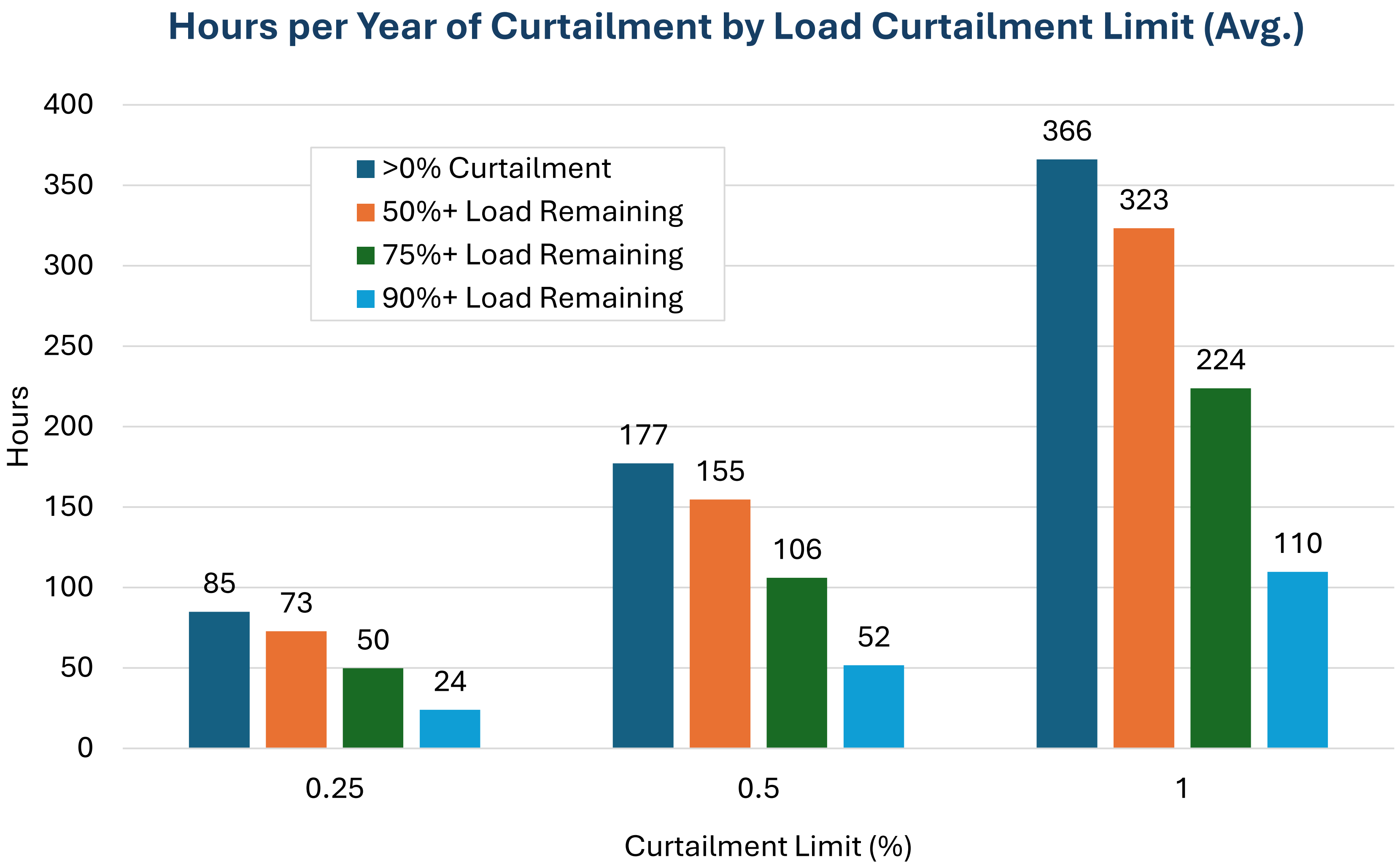
- To visualize the relationship between load additions and curtailment, curtailment was calculated for small incremental load additions
- This plot captures year-by-year variability in demand patterns, including the effects of extreme weather and economic conditions
- A similar plot is presented for each balancing authority in Appendix A of the report

Curtailment Rate vs. Load Addition, MW – PJM



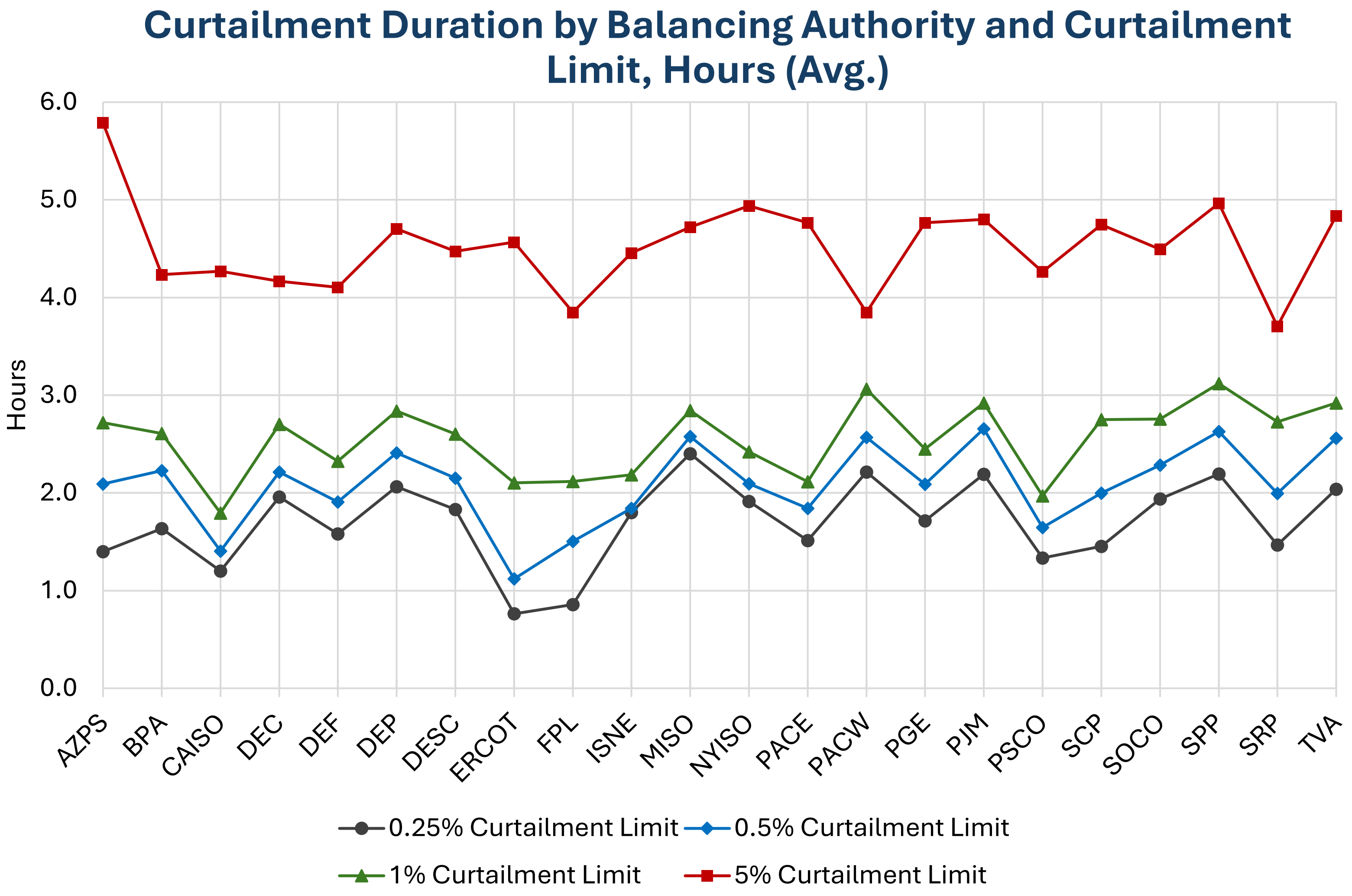
Results: Annual Hours of Curtailment

- A large majority of curtailment hours retain most of the new load
- 88% of hours during which load curtailment is required retain at least half of the new load
- 60% of the hours retain at least 75% of the load, and 29% retain at least 90% of the load



Results: Duration of Curtailment Events

- Curtailment duration is the length of time the new load is curtailed during curtailment events
- All hours in which any curtailment occurred were included, regardless of magnitude
- Average duration was 1.7 hours at the 0.25% curtailment limit, 2.1 hours at 0.5%, 2.5 hours at 1.0%, and 4.5 hours at 5.0%



Next steps

Limitations

This analysis provides a first-order assessment of available capacity for serving new curtailable loads

Factors that may **reduce** available headroom estimate	Factors that may **increase** available headroom estimate
• Network constraints • Generators Intertemporal constraints • Additional variability and uncertainty from new load • Non-constant load increases (will increase curtailment as % of served load)	• Accounting for reserve capacity • Lower expected system peaks • Non-constant load shape (may reduce absolute MWh of curtailed load)

Research plan

Address limitations of existing analysis and re-calculate *curtailment enabled headroom* by:

Simulating scheduling and dispatch of the following:

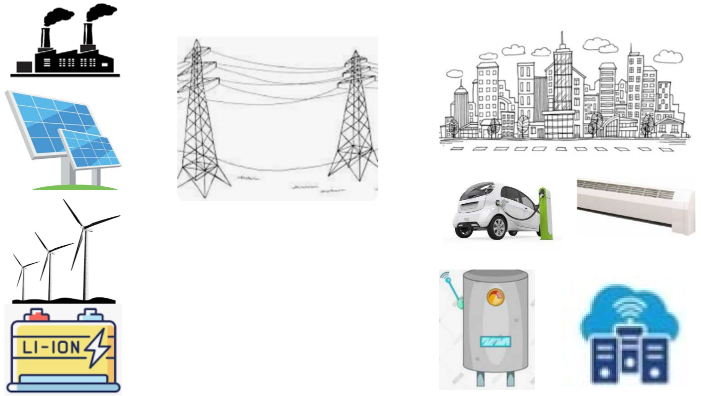
- Electrical Generating Units (EGUs)
- Flexible demands
- Energy Storage

Accounting for the following:

- EGU's technical and inter-temporal constraints
- Electric Power Transmission Network
- Scenarios of new load variability, uncertainty, and responsiveness/flexibility

GRACE modeling framework

Balancing Authority Representation



New Loads Characterization

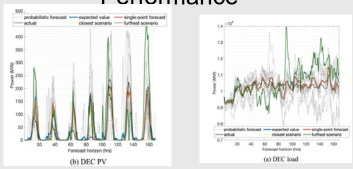
- Temporal pattern
- Geographic profile
- Responsiveness (technical capabilities and costs)

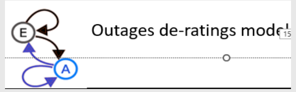
Uncertainty Characterization

Weather based risk

Intrinsic risk

Probabilistic Forecasts of Demand, Renewables and Assets' Performance





Unit Commitment

Deterministic / Stochastic / Risk adjusted

Daily – Weekly Horizon

$$\min_{\Xi}$$

$(1-\beta)$ Expected Cost $+\beta$ CVaR of Cost:

$$(1-\beta)\left[\sum_{t\in T}\left(\sum_{j\in J^{\mathrm{coal}}\cup J^{\mathrm{gas}}} (c_j(p_j(t)) + c_j^{\mathrm{U}} y_j(t) + c_j^{\mathrm{D}} z_j(t)) + \sum_{j\in J^{\mathrm{ng}}} c_j(p_j(t)) + \sum_{j\in J^{\mathrm{w}}} c_j(p_j^{\mathrm{W}}(t))\right) + \sum_{\omega\in\Omega}\pi_\omega \sum_{t\in T}\left(\sum_{j\in J}(c_j^{\mathrm{RU}} r_{j\omega}^{\mathrm{U}}(t) - c_j^{\mathrm{RD}} r_{j\omega}^{\mathrm{D}}(t)) + \sum_{i\in\Lambda^D} c_i^{\mathrm{LOL}} D_{i\omega}^{\mathrm{shed}}(t)\right)\right] + \beta\left(\zeta + \frac{1}{1-\alpha}\sum_{\omega\in\Omega}\pi_\omega\eta_\omega\right)$$

subject to:

GRACE Libraries for fast solution

Economic Dispatch

Deterministic / Stochastic

Myopic / With Look-ahead

$$\min_{\Xi}$$

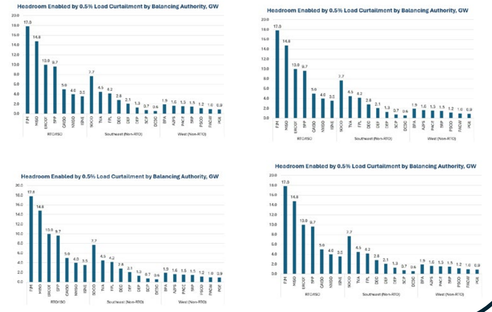
$(1-\beta)$ Expected Cost $+\beta$ CVaR of Cost:

$$(1-\beta)\left[\sum_{t\in T}\left(\sum_{j\in J^{\mathrm{coal}}\cup J^{\mathrm{gas}}} (c_j(p_j(t)) + c_j^{\mathrm{U}} y_j(t) + c_j^{\mathrm{D}} z_j(t)) + \sum_{j\in J^{\mathrm{ng}}} c_j(p_j(t)) + \sum_{j\in J^{\mathrm{w}}} c_j(p_j^{\mathrm{W}}(t))\right) + \sum_{\omega\in\Omega}\pi_\omega \sum_{t\in T}\left(\sum_{j\in J}(c_j^{\mathrm{RU}} r_{j\omega}^{\mathrm{U}}(t) - c_j^{\mathrm{RD}} r_{j\omega}^{\mathrm{D}}(t)) + \sum_{i\in\Lambda^D} c_i^{\mathrm{LOL}} D_{i\omega}^{\mathrm{shed}}(t)\right)\right] + \beta\left(\zeta + \frac{1}{1-\alpha}\sum_{\omega\in\Omega}\pi_\omega\eta_\omega\right)$$

to:

Results

Per scenario / Per BA



Conclusions

1

New Load, Modest Curtailment

The existing US power system could accommodate significant new load additions with relatively modest load curtailment

2

Near-Term Strategy

Load flexibility offers a promising near-term strategy to expedite load interconnection, mitigate costs, and improve system utilization

3

Headroom Estimate Refinement

Additional analysis can refine this study's first-order estimate of system headroom for purposes of system planning

NICHOLAS INSTITUTE *for* ENERGY, ENVIRONMENT & SUSTAINABILITY

Fueling ambitions. Powering networks. Accelerating solutions.

SIGN UP FOR EMAIL:
bit.ly/niemail

@NichInstitute

bit.ly/nilinkedin

@NichInstitute

@NichInstitute

@NicholasInstitute